>>> revista tradumàtica 03
núm.3 desembre 2005 la indústria de la traducció
http://www.fti.uab.es/tradumatica/revista : ISSN 1578-7559

Ús d'un sistema productiu de traducció automàtica

Juan Alberto Alonso, Anna Civil
Comprendium España SL
juan.alonso@comprendium.es,
anna.civil@comprendium.es

Resum: En aquest article expliquem alguns aspectes de la tecnologia de traducció automàtica Comprendium, els diferents usos del nostre sistema de traducció automàtica, la relació entre les necessitats dels usuaris i els nostres productes i uns exemples d'instal·lacions actuals.

Paraules clau: traducció automàtica, servei de traducció, memòries de traducció

---

1 La tecnologia de traducció automàtica Comprendium

1.1 Antecedents històrics

L'actual sistema de traducció automàtica de Comprendium és el successor de l'antic sistema METAL. Aquest sistema es va començar a desenvolupar els anys 70 a la Universitat d'Austin (Texas, EUA) i els anys 80 va ser adquirit per l'empresa Siemens. Aleshores, el sistema funcionava sobre màquines LISP (Symbolics), només hi havia un parell de llengües (anglès-alemany) i la traducció resultava molt lenta. Posteriorment, durant la primera meitat dels anys 90, el sistema es va portar al sistema operatiu UNIX sobre plataformes SUN, la velocitat de traducció va augmentar i es van desenvolupar nous parells de llengües, entre els quals hi havia l'anglès-català, l'anglès-castellà i el castellà-català.

A finals dels anys 90, dins de l'empresa GMS, el sistema es va portar a l'entorn Windows i el nucli es va reprogramar en C/C++. El resultat va ser un increment dràstic de la velocitat de traducció i una integrabilitat molt més alta amb d'altres eines informàtiques. A més, es van incloure nous parells de llengües. Durant aquesta primera part dels anys 2000, ja com a

Comprendium, hem continuat desenvolupant altres parells de llengües, com el català-francès, i hem consolidat una gamma de productes adreçats a diferents tipus d'usuaris.

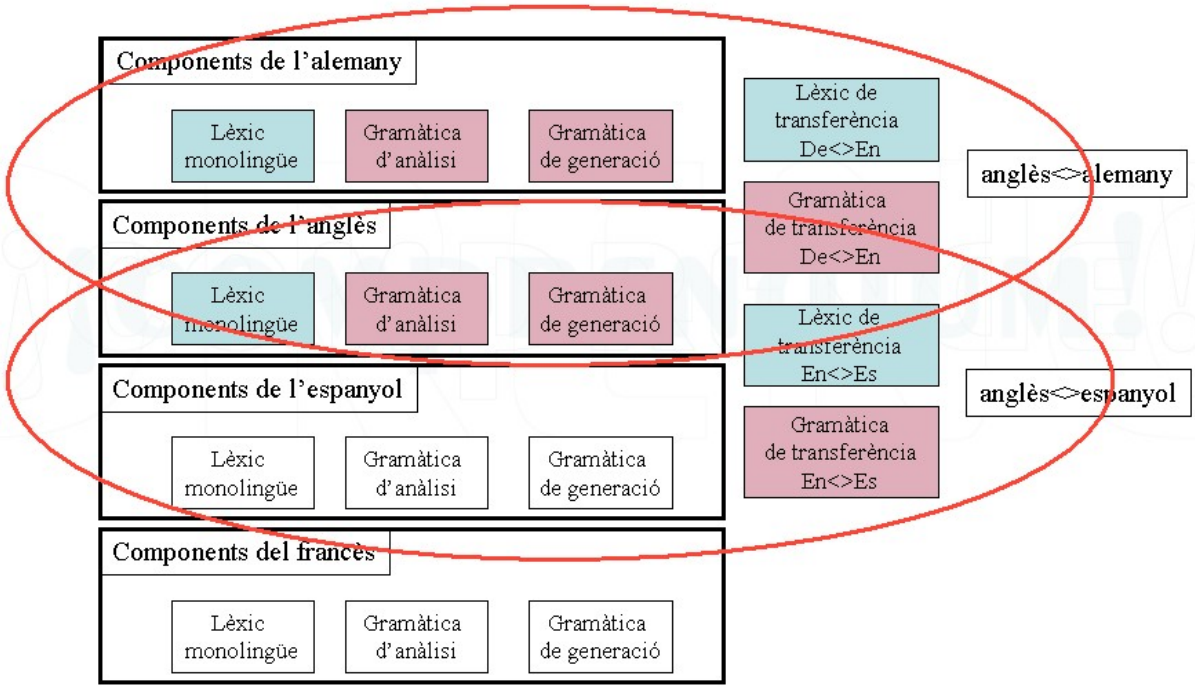
1.2 Arquitectura del sistema

Els dos principals blocs de components del sistema de TA són els lèxics i les gramàtiques.

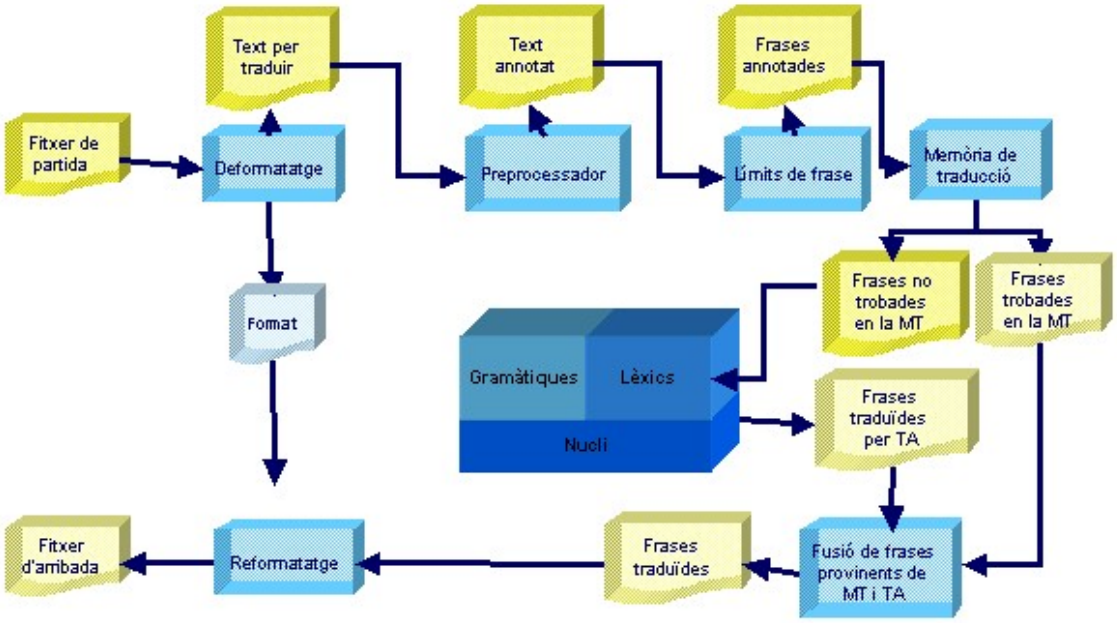
En qualsevol direcció de parell de llengües (per exemple anglès→català) el sistema de TA de Comprendium utilitza dos lèxics monolingües (un per la llengua de partida, l'anglès, i un per la llengua d'arribada, el català) i un lèxic bilingüe de transferència (l'anglès→català). Els lèxics monolingües contenen informació morfològica, sintàctica i semàntica rellevant per a totes les paraules conegudes pel sistema. Els lèxics bilingües contenen totes les traduccions per una paraula determinada, en funció del context sintàctic i/o semàntic.

Al mateix temps, el sistema utilitza una gramàtica d'anàlisi per dur a terme l'anàlisi sintàctica de les frases d'origen, una gramàtica de transferència per transformar els arbres d'anàlisi en l'estructura sintàctica de la llengua d'arribada, i una gramàtica de generació per generar les frases en la llengua d'arribada.

Tots aquests components es poden reutilitzar per altres parells de llengües. És a dir, per exemple, els sistemes anglès→català i anglès→alemany comparteixen el mateix lèxic monolingüe anglès i la mateixa gramàtica d'anàlisi de l'anglès. Vegeu la figura següent:



1.3 Com funciona el procés de traducció

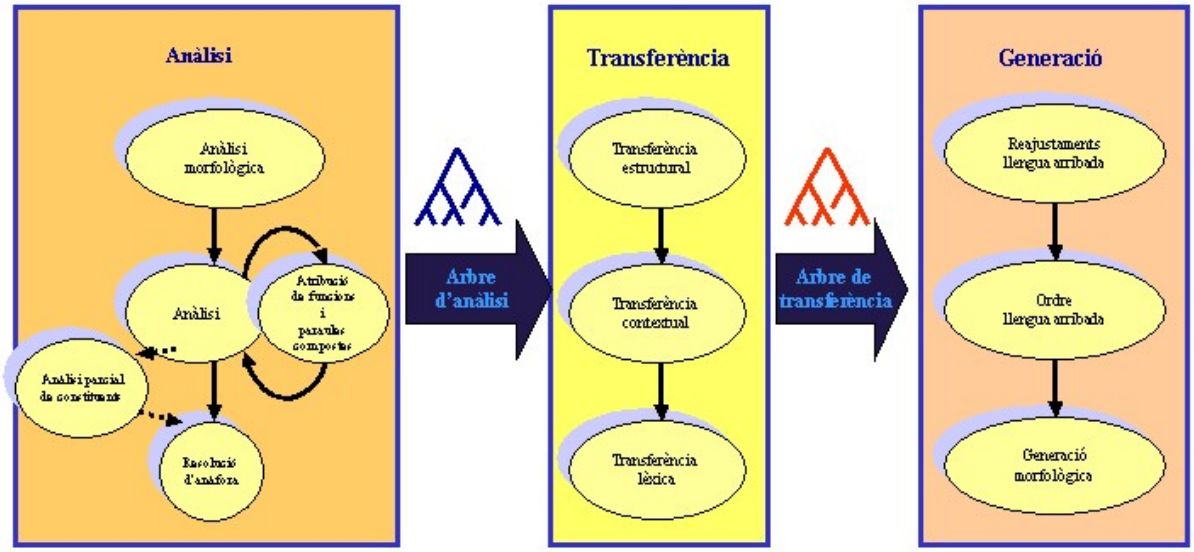


El procés de traducció comença amb un fitxer en la llengua de partida i continua en les fases següents:

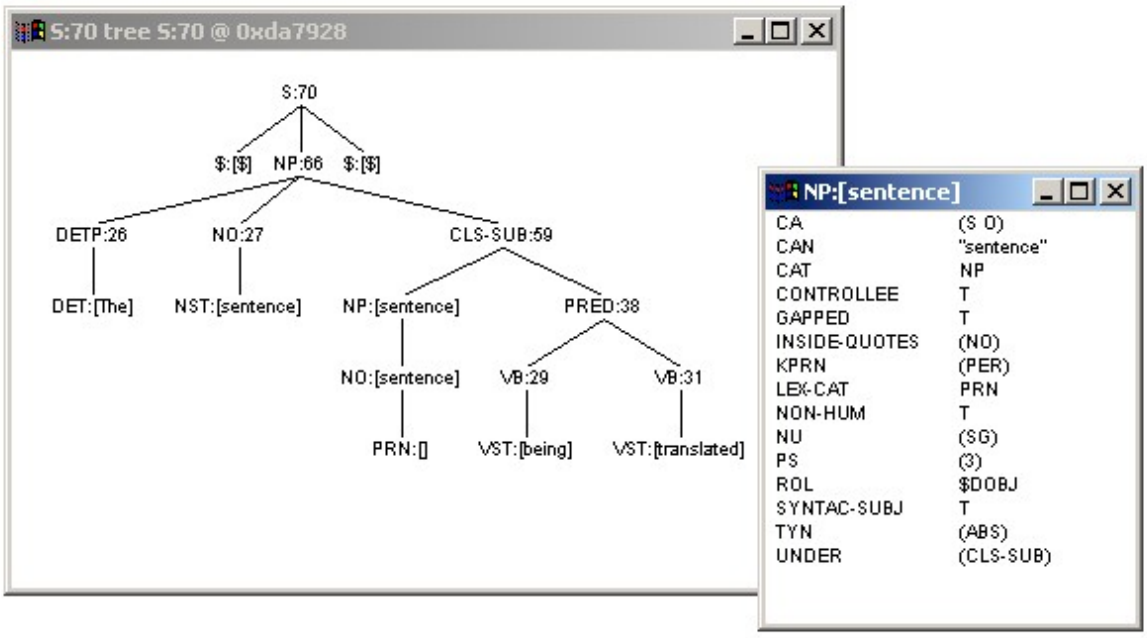
- Primer, es **deformata** el text. Es separen els fragments de text de la informació de composició i s'emmagatzema el format de composició. Això es fa perquè, al final, cal que el text traduït tingui la mateixa composició que el document original. Una recomposició del text a mà requeriria més temps que la traducció en si.
- S'envia el text a un **preprocessador**, que identifica les parts que no s'haurien de traduir (noms, adreces de correu electrònic, noms de fitxers, etc.) i les marca.
- Aleshores el text és segmentat en frases perquè la majoria d'eines de traducció treballen en l'àmbit de la frase, així com també ho fa el motor de TA.
- Després es consulten les **memòries de traducció**, si n'hi ha. Les memòries de traducció són una gran base de dades de frases on per cada frase de la llengua de partida s'emmagatzema la frase equivalent en la llengua d'arribada, que és el resultat d'esforços de traducció anteriors. Les memòries de traducció són eines de suport molt útils en el cas de textos repetitius. Si es troba una frase a les memòries, no cal que s'enviï al motor de TA.
- Si una frase no és a les memòries de traducció, s'envia al **motor de TA**, que consisteix en un nucli de programari que dirigeix la traducció. La direcció de traducció, la determinen els recursos lingüístics (gramàtiques i lèxics) que es carreguen en el nucli. El motor de TA tradueix frase per frase.
- La sortida del motor de TA (és a dir les frases en la llengua d'arribada) **s'ajunten** amb els resultats de la cerca a les memòries de traducció, de manera que totes les frases del document queden traduïdes en la llengua d'arribada.
- L'últim pas és unir el text resultant amb la informació de composició emmagatzemada i **reformatar** tot el document. Així es crea el fitxer en la llengua d'arribada.

1.3.1 Les fases de traducció

Tal com hem esmentat més amunt, el sistema de TA de Comprendium utilitza l'enfocament de transferència, que vol dir que utilitza el paradigma d'anàlisi → transferència → generació.



Els arbres de partida i d'arribada són representacions de l'estructura sintàctica de la frase que es tradueix. Al mateix temps, els arbres contenen una informació lingüística àmplia en forma de parells tret-valor:



A la **fase d'anàlisi**, primer es fa l'anàlisi morfològica, és a dir es busquen les paraules al lèxic i s'interpreta la seva posició específica a la frase. En segon lloc, es fa l'anàlisi sintàctica, és a dir s'apliquen una sèrie de regles gramaticals que creen sintagmes i combinen sintagmes en parts de frase i finalment en una interpretació completa. El repte més gran aquí és la desambiguació. En aquesta fase també es fa l'atribució de funcions sintàctiques, és a dir l'assignació a alguns

sintagmes de funcions gramaticals (com *subjecte* o *complement indirecte*) i l'anàlisi de multimots, que és la interpretació de grups de paraules com a termes (com *"car park"* en anglès).

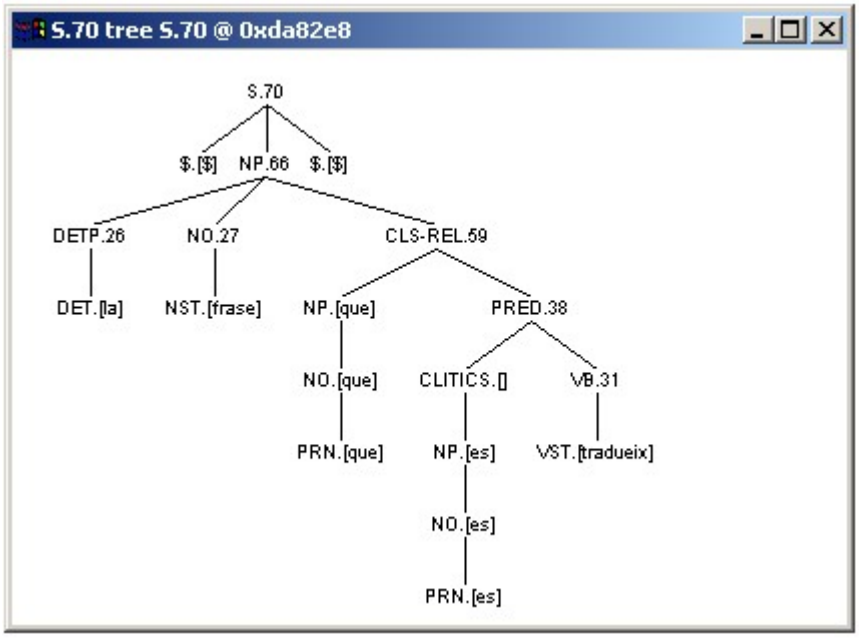
Si l'anàlisi sintàctica falla (perquè la frase d'origen és massa complexa o incorrecta), es desencadena un mecanisme especial, anomenat anàlisi parcial de constituents, que prova d'identificar sintagmes significatius en la frase original.

El resultat és un arbre d'anàlisi per frase de la llengua de partida.

A la **fase de transferència**, s'hi fa la transferència estructural, amb canvis d'estructura com per exemple de l'anglès "three years ago" al català "fa tres anys"; la transferència contextual, on per exemple es canvien les funcions sintàctiques dels constituents (el verb anglès *"to like"* és en català *"agradar"*, però el subjecte anglès passa a complement indirecte en català), i la transferència lèxica, que substitueix els termes i els mots de la llengua de partida pels de la llengua d'arribada (per exemple, el mot anglès *"airbag"* es tradueix en català per *"coixí de seguretat").*

El resultat és un arbre intermedi entre la llengua de partida i la llengua d'arribada.

A la **fase de generació**, es fan reajustaments de l'arbre sintàctic específics de la llengua d'arribada, com per exemple en català la inserció de pronoms febles (*I have four → <u>En</u> tinc quatre);* es fan transformacions per ordenar la frase en l'ordre específic de la llengua d'arribada, i finalment es flexionen les paraules i es creen les formes correctes.



2 L'ús del sistema de traducció automàtica

2.1 Expectatives actuals de la traducció automàtica

2.1.1 Factors que influeixen en la qualitat de la traducció automàtica

Bàsicament, hi ha tres factors que influeixen en la qualitat de la traducció d'un sistema de traducció automàtica:

- La potència lingüística del sistema de traducció: traduir no és simplement substituir les paraules del document original per la seva traducció. En una traducció entre dues llengües intervenen factors lèxics, morfològics, sintàctics, semàntics i pragmàtics i, per tant, quant més coneixement lingüístic tingui el sistema de traducció, més alta serà la qualitat.
- La proximitat lingüística de les llengües entre les quals es tradueix: òbviament la qualitat d'un sistema castellà-català o català-francès serà sempre més alta que la d'un sistema entre el català i el rus, posem per cas.
- El tipus de documents que s'han de traduir: el cas ideal de document per a un sistema de traducció automàtica hauria de tenir frases curtes, ben redactades, sense cap falta d'ortografia o tipogràfica, sense jocs de paraules o diàlegs. És per això que textos molt col·loquials, com ara la majoria de correus electrònics o de xats, o bé textos literaris, per exemple, són poc adequats a les capacitats dels sistemes de traducció automàtica.

2.1.2 Usos de la traducció automàtica

Hi ha dos usos principals que es poden donar als sistemes de traducció automàtica:

- **Ús informatiu**, pensat per a entendre textos escrits en llengües que l'usuari no coneix. L'exemple típic seria la traducció automàtica de pàgines web. Per aquest ús informatiu, cal que el sistema de traducció tingui un ventall ampli de parells de llengües disponibles encara que la qualitat de traducció d'alguns d'aquests parells de llengües no sigui molt elevada.
- **Ús productiu**, pensat per a entorns de traducció professional, amb l'objectiu d'estalviar temps i esforç en la traducció de grans volums de documentació. En aquest cas, sí que és necessari que, a més de la rapidesa, la qualitat de traducció del parell de llengües que es faci servir sigui prou elevada. Un cas típic d'aquest ús seria la traducció en temps real d'articles per a un diari (vegeu l'apartat 4.1).

2.2 El flux de feina des del punt de vista de l'usuari

Si parlem d'usuaris professionals, els passos per traduir un document amb el nostre sistema de traducció automàtica serien els següents:

1. Triar el document a traduir.
2. Triar la direcció de traducció desitjada (p.ex. català→castellà).
3. Triar l'àrea temàtica a la qual pertany el document.
4. Seleccionar els paràmetres de traducció adequats:
    a. Marques per a paraules desconegudes, compostos, alternatives, segments traduïts amb memòries de traducció, etc.
    b. Ús o no de mòduls de memòries de traducció (i si se'n fan servir, seleccionar quins mòduls)
    c. Ús o no de filtres de preedició i de postedició.
    d. Triar els paràmetres lingüístics desitjats (dependents de la direcció de traducció triada).
5. Enviar el document a traduir
6. Un cop traduït, corregir la traducció oferta pel sistema
7. Idealment, realimentar el sistema amb el resultat de la correcció, per exemple codificant terminologia. Per més informació, vegeu el punt 2.3.

2.3 Realimentació del sistema de traducció automàtica

En un entorn de traducció professional i massiva, és convenient realimentar el sistema de traducció amb els resultats de les successives correccions de les traduccions ofertes pel sistema. D'aquesta manera s'aprofita en part la feina de correcció per anar millorant la qualitat de traducció del sistema.

Hi ha tres àrees en les quals l'usuari pot dur a terme aquesta feina de realimentació:

- Mòduls de memòries de traducció
- Filtres de preedició i postedició
- Lèxics del sistema

2.3.1 Memòries de traducció

Tots els productes de traducció automàtica Comprendium ofereixen la possibilitat de fer servir mòduls de memòries de traducció. A més, existeix un producte adreçat a entorns professionals de traducció, des del qual es poden crear aquests mòduls de memòries de traducció a partir de dos documents en diferents llengües (normalment el document original i la seva traducció ja corregida). És també possible importar mòduls de memòries de traducció ja existents, en format TXT o TMX.

Això fa que, un cop corregits els documents traduïts pel sistema, aquests es puguin incorporar a mòduls de memòries de traducció, de manera que, en posteriors traduccions, el sistema pugui agafar les frases ja traduïdes i corregides (i per tant, correctes) en comptes de tornar-les a traduir amb el motor de traducció automàtica.

2.3.2 Filtres de preedició i postedició

Durant el procés de traducció automàtica es poden activar els anomenats filtres de preedició i de postedició. Aquests filtres permeten fer dues coses bàsiques:

- Marcar paraules o grups de paraules (o més exactament, seqüències de caràcters) com a constants, de manera que el motor de traducció automàtica no els tradueixi (evitant, per exemple, que es tradueixin noms propis com ara "Zapatero", "Bush", etc.
- Fer substitucions d'unes seqüències de caràcters específiques per unes altres. Aquest procés de substitució es pot fer abans o després de la traducció. Un cas típic d'aplicació d'aquest filtre de substitució seria la correcció d'errors ortogràfics o problemes tipogràfics repetitius que poden aparèixer als documents d'entrada (p.ex. "T I T O L" $\rightarrow$ "TÍTOL").

2.3.3 Manteniment del lèxic

Si s'han activat les marques de paraules desconegudes abans d'enviar el document a traduir (vegeu punt 4.a de l'apartat 2.2), es podran veure a la traducció quines paraules no s'han traduït correctament perquè no hi són als lèxics del sistema, o perquè hi són amb una traducció incompleta o incorrecta.

El LexShop és una eina de codificació de lèxic que permet a l'usuari de codificar noves entrades lèxiques en el sistema, o modificar-ne d'existents.

3 Diferents tipus d'usuaris, diferents productes de Comprendium

Després d'anys d'experiència en les necessitats de traducció automàtica, sabem que una de les coses més difícils quan una persona intenta comprar/fer servir eines de lingüística computacional és trobar la solució específica que pot donar resposta a les seves necessitats.

La tecnologia de Comprendium consta d'un motor de traducció que pot estar connectat a mòduls de memòries de traducció, així com a una eina de codificació lèxica i una eina de marcatge i substitució d'elements. Es pot accedir a aquest motor de traducció automàtica des de diferents interfícies i dins de diferents configuracions (ordinador individual, arquitectures client-servidor dins d'una intranet, o per internet).

A molt grans trets, els productes bàsics de què disposem són:

- Comprendium Translator Desktop, que és el traductor professional monousuari.
- Comprendium Translator Portal, que és una solució multiusuari en xarxa que permet la traducció de text, documents i pàgines web.
- el LexShop, que és l'eina d'administració i codificació del lèxic i permet la codificació i modificació d'entrades.
- el servei TRANSLENDIUM, que és el servei de traducció automàtica en línia al web www.comprendium.es i està adreçat sobretot a usuaris individuals o a usuaris esporàdics. Vegeu-ne més informació al punt 4.2.

A partir d'aquests productes bàsics, el traductor de Comprendium ofereix una àmplia gamma de solucions personalitzades, des de les pensades per a usuaris individuals que treballin a casa, passant per solucions de gamma mitjana client-servidor per a petites i mitjanes empreses fins a solucions distribuïdes d'alt rendiment per a administracions locals, portals i proveïdors de serveis lingüístics o grans empreses. Aquestes solucions es configuren d'acord amb les necessitats de cada client pel que fa a volum de traducció, nombre d'usuaris concurrents, parells de llengües, necessitat d'eina de codificació de lèxic, necessitat de memòries de traducció, etc.

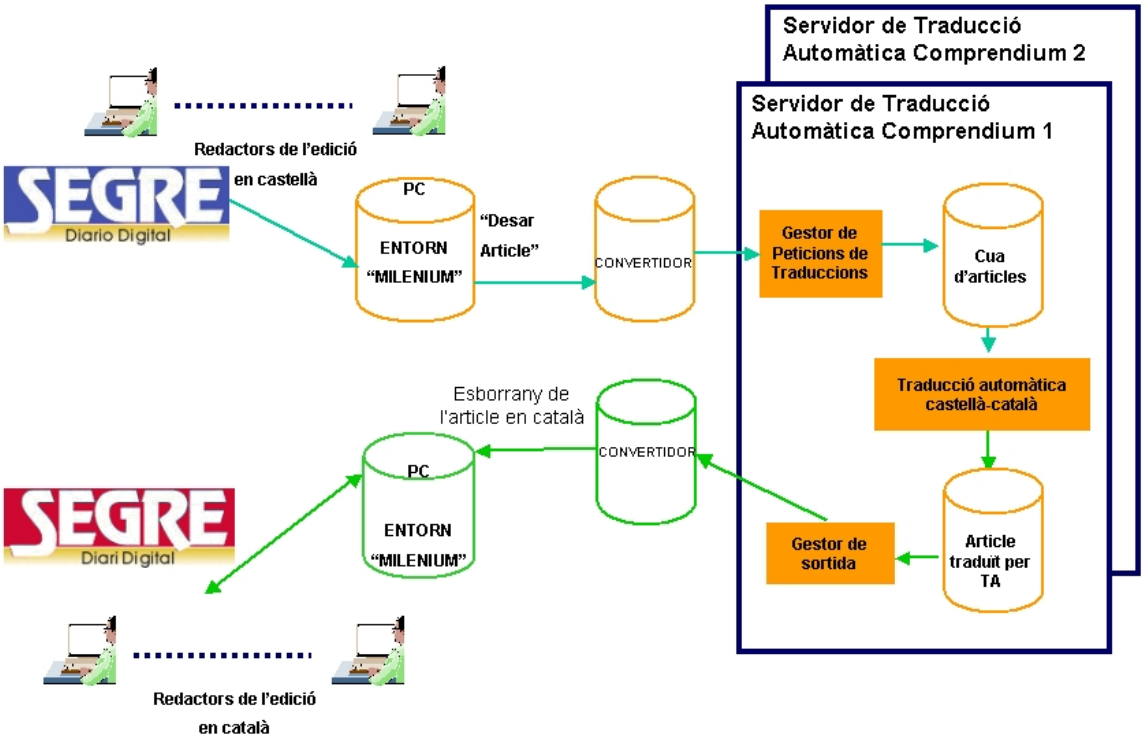
Les 23 direccions de traducció disponibles actualment són: català↔castellà, català↔anglès, català↔francès (prototipus), castellà↔alemany, castellà↔anglès, castellà↔francès, anglès↔alemany, anglès↔francès, anglès↔rus, alemany↔francès, alemany↔rus i anglès→italià.

4 Exemples d'instal·lacions

4.1 Diari Segre

El diari "Segre" es publica a Lleida amb dues edicions diàries i simultànies, una en castellà i una altra en català.

La figura següent il·lustra el paper que juga el servidor de traducció automàtica de Comprendium dins de la cadena de creació del diari.



El diari s'edita actualment dins de l'entorn "Protec Milenium". Dins d'aquest entorn, els redactors de la versió castellana del diari preparen els articles en castellà. Un cop desada la versió final de cada article en castellà, aquest s'envia automàticament a un convertidor especialment dissenyat per extreure el text de l'article i conservar-ne el format. El text s'envia aleshores a l'entorn de traducció automàtica. Hi ha dos servidors de traducció automàtica, el primari, normalment en funcionament, i el secundari, preparat per entrar en funcionament en cas de fallada del servidor primari.

La interfície entre l'entorn de traducció automàtica i l'editor Milenium és una cua d'entrada/sortida. La cua està implementada en forma de carpetes de fitxers Windows. Les carpetes d'entrada es van monitoritzant constantment i de seguida que s'hi detecta la presència de fitxers d'entrada (el text dels articles que han de ser traduïts) aquests fitxers s'agafen de la cua i s'envien al sistema de traducció automàtica, on són processats pel motor de traducció castellà→català i, un cop traduïts, desats a les carpetes de sortida.

Un altre procés monitoritza la cua de sortida i quan hi detecta la presència de fitxers traduïts, els agafa de la carpeta de sortida, es processen pel convertidor, el qual hi torna a posar el format d'entrada de l'article original, i s'emmagatzemen a les carpetes de l'edició catalana de l'editor Milenium. Els redactors de l'edició catalana reben aleshores un avís que hi ha un esborrany d'article en català llest per ser corregit, el corregeixen i el desen a les carpetes de l'edició final en català.

Tot aquest procés es verifica en temps real, ja que totes dues edicions, la castellana i la catalana, han d'estar enllestides a temps i al mateix temps. Això només es pot assolir si es fa servir un sistema de traducció automàtica que sigui prou ràpid i que lliuri una qualitat de traducció molt bona, de manera que les feines de correcció es puguin reduir a un mínim.

4.2 Translendium

TRANSLENDIUM és un servei de traducció automàtica de documents en línia al web de Comprendium, www.comprendium.es, adreçat a usuaris dins de l'Estat espanyol.

Permet la traducció immediata de documents (en format DOC, ASCII, RTF i HTML) a preus molt assequibles i dóna una traducció de qualitat esborrany, més o menys alta segons el parell de llengües, i que, com tota traducció automàtica, ha de ser revisada i corregida abans de ser utilitzada.

TRANSLENDIUM està pensat per donar servei a usuaris amb necessitats de traducció, a qui no els surt a compte comprar el sistema. Són usuaris amb necessitats esporàdiques de traducció en un parell de llengües o usuaris que han de traduir en molts parells de llengües diferents.

Per accedir al servei l'usuari es registra donant les dades bàsiques. Un cop registrat, l'usuari es connecta a TRANSLENDIUM amb el nom d'usuari. Dins del servei, pot fer les accions bàsiques per mitjà de dues pàgines principals. A la pàgina de traducció de documents, després de seleccionar el document, el parell de llengües i l'àrea temàtica, obté el pressupost i, si n'accepta el preu, l'envia a traduir. A la pàgina de documents traduïts, veu la informació de l'estat de les traduccions, els documents i el preu i pot pagar, baixar o esborrar documents.