

2. Les memòries de traducció

Índex

2.1. Introducció

En aquest capítol parlarem en detall de les memòries de traducció, el principal recurs en què es basen els sistemes de traducció assistida per ordinador.

> *Una memòria de traducció és un repositori de segments de text en una determinada llengua amb les traduccions a una o més llengües.*

D'aquesta manera tenim una relació directa entre un segment de text en una llengua i la seva traducció a una altra llengua. Aquests segments de text acostumen a ser oracions, tot i que no sempre ho són des del punt de vista gramatical. Per aquest motiu es parla de *segments* i no d'oracions.

En les memòries de traducció no es relacionen unitats més grans, com per exemple paràgrafs, ja que la probabilitat de trobar paràgrafs iguals o semblants en dos textos és molt baixa. Tampoc es fan relacions entre unitats molt petites, com per exemple, paraules o sintagmes, ja que el traductor humà no treballa tractant de manera aïllada aquestes unitats[1].

1 Això no vol dir que no hi hagi eines de traducció assistida que intentin combinar unitats més petites de diversos segments que es troben a la memòria per provar de formar una proposta vàlida.

La funció principal de les memòries de traducció és oferir al traductor suggeriments de traducció del segment que està traduint. Aquest suggeriment pot provenir d'una *coincidència exacta* (*exact match*) quan el segment en la llengua de partida que hi ha a la memòria és exactament igual al que s'està traduint; o bé d'una *coincidència parcial* (*fuzzy match*), quan el segment en la llengua de partida no és exactament igual al que s'està traduint. L'índex de similitud mínim per a que apareguin suggerències és totalment configurable. Si s'escullen índexos molt alts, per exemple 99%, apareixeran molt poques propostes. Si s'opta per un índex molt baix, per exemple, el 60%, apareixeran moltes més propostes, però apareixeran moltes que no seran útils. Cal tenir en compte que si s'accepta una coincidència parcial, caldrà dur a terme alguna edició (canviar alguna paraula, per exemple). Si l'índex de similitud és molt baix, l'esforç d'edició serà molt alt i probablement valgui més la pena traduir el segment manualment des de zero. Un bon compromís pot ser configurar la similitud mínima entre el 65 i el 85%. Quan hi ha més d'una coincidència parcial, el sistema de les mostra endreçades de més similitud a menys similitud.

Aquesta funció de recuperació de segments similars d'una base de dades que anomenem memòria de traducció presenta dos reptes importants:

- Trobar de manera ràpida els segments més semblants.
- Fer servir una mesura de similitud per endreçar els segments recuperats de manera que es presentin en primer lloc els més similars. Aquesta mesura ha de ser indicativa del grau de dificultat i temps necessari per editar la traducció del segment recuperat i deixar-la com a traducció del segment original que estem traduint. Tot això tenint en compte que només podem comparar els segments en la llengua de partida ja que la traducció del segment que busquem encara no la tenim.

Els sistemes de traducció assistida també permeten fer cerques d'unitats més petites (per exemple paraules o termes) dins de les memòries actives. D'aquesta manera podem buscar si s'ha traduït anteriorment un determinat terme. Cal no confondre aquesta funcionalitat amb la cerca automàtica en bases de dades terminològiques. En aquest segon cas tenim termes i les seves traduccions en una base de dades. En el cas de fer servir memòries de traducció el que tenim són oracions originals i traduïdes que poden contenir el terme. Si fem una cerca el sistema ens mostrarà les oracions originals que contenen el terme i les traduccions d'aquestes oracions (on es suposa que l'usuari podrà trobar la traducció del terme)[2].

2 Alguns sistemes van més enllà i intenten també inferir la traducció del terme original a partir dels segments traduïts.

2.2. Indexació i recuperació de segments

Hem definit les memòries de traducció com un repositori de segments de text en més d'una llengua. Per accedir de manera eficient a aquest repositori aquest ha d'estar contingut en una base de dades i s'ha hagut de dur a terme algun tipus d'indexació de les dades. Aquesta indexació és important, ja que la cerca a la memòria s'ha de fer en un temps molt petit, des del moment que el usuari introdueix un segment fins el moment que apareix el següent segment a traduir. La sensació per a l'usuari ha de ser que el pas d'un segment a un altre és immediat.

Imaginem-nos que tenim una memòria de traducció d'una mida mitjana, per exemple, 10.000 segments. La cerca no pot ser seqüencial, és a dir, mirar si tenim un segment igual o semblant al que estem traduint començant pel primer i anant comparant un a un. La cerca de segments iguals pot arribar a ser molt ràpida, però la de semblants triga una mica més. Més endavant dediquem un subapartat al tema del càlcul de la similitud entre segments. Si en comptes de 10.000 segments la memòria fos de 100.000 una cerca seqüencial trigaria també molt més. Com que ens interessa treballar amb memòries molt grans per augmentar la probabilitat de trobar segments interessants, és imprescindible trobar un mecanisme d'indexació i recuperació eficients per evitar una resposta massa lenta del sistema de traducció assistida.

En la resta d'aquest apartat explicarem tècniques genèriques ja que cada eina incorpora variacions en la indexació i el càlcul de similituds.

2.2.1. Indexació de memòries de traducció

La indexació d'una memòria de traducció consisteix a realitzar un índex invers de les paraules (o fragments de paraules, o almenys d'algunes paraules) que apareixen a la memòria de traducció. L'índex ens dóna l'identificador de tots els segments en els que apareix una determinada paraula (o fragment de paraula).

Imaginem-nos que tenim una memòria de traducció amb els següents segments:

ID	Segment original	Segment traduït
1	Search the Legal framework	Cercar en Marc jurídic
2	Legal framework of the ESCB	Règim jurídic del SEBC
3	ECB institutional provisions	Disposicions institucionals del BCE
4	Monetary policy and Operations	Política monetària i operacions
5	Payment and settlement systems	Sistemes de pagament i liquidació
6	Banknotes and coins, means of payment and currency matters	Bitllets de banc i monedes, mitjans de pagament i qüestions de moneda
7	Foreign exchange and Foreign reserves	Divises i reserves exteriors
8	The approach was successful: in volume terms, close to 80 % of the initial banknote demand and 97 % of the total coin needs for the changeover had been distributed before 1 January 2002.	Va resultar en un plantejament encertat ja que, en termes de volum, gairebé el 80 % de la demanda inicial de bitllets i el 97 % de les monedes per a l'introducció de l'euro van ser distribuïts abans de l'1 de gener de 2002.
9	Employment, conduct , fraud prevention and transparency	Contractació, conducta, prevenció del frau i transparència
10	Financial market stability	Estabilitat dels mercats financers

L'índex invers ens indicaria l'identificador de tots els segments on apareix una determinada paraula (en el nostre exemple no s'indexen els números). Per a aquest exemple tindria el següent aspecte.

and 4:5:6:7:8:9	financial 10	payment 5:6
approach 8	for 8	policy 4
banknote 8	foreign 7	prevention 9
banknotes 6	framework 1:2	provisions 3
been 8	fraud 9	reserves 7
before 8	had 8	search 1
changeover 8	in 8	settlement 5
close 8	initial 8	stability 10
coin 8	institutional 3	successful 8
coins 6	january 8	systems 5
conduct 9	legal 1:2	terms 8
currency 6	market 10	the 1:2:8
demand 8	matters 6	to 8
distributed 8	means 6	total 8
ecb 3	monetary 4	transparency 9
employment 9	needs 8	volume 8
escb 2	of 2:6:8:8	was 8
exchange 7	operations 4	

Aquesta taula ens proporciona informació sobre en quins segments apareixen cada una de les paraules. Per exemple, *payment* apareix als segments 5 i 6 i *market* en el 10.

Imaginem-nos que volem trobar segments semblants a aquest:

Banknotes, coins, and types of payment

Agafaríem els índexos dels segments i l'índex que aparegués més vegades seria probablement el més semblant, ja que contindria més paraules comunes. Depenent de l'algorisme de càlcul de similitud entre segments, l'ordre de les paraules ens pot jugar males passades, així que sovint es pren no només el més semblant, si no els primers més semblants i es calcularia la similitud, fins que aquesta estigués per sota de la similitud mínima donada per l'usuari. Sobre el càlcul de similitud, parlarem en el següent subapartat. Segons això, quedaria:

```
banknotes 6
coins 6
types
payment 5:6
```

i per tant els segments més semblants seria el 6 (*Banknotes and coins, means of payment and currency matters*), ja que té 3 paraules coincidents. El segon segment més semblant seria el 5 (*Payment and settlement systems)*.

Fixem-nos en un parell d'aspectes de l'índex invers d'aquesta memòria de traducció. Les paraules molt curtes tendeixen a ser paraules funcionals que apareixen en molts segments (fixem-nos en *and*, que apareix en 6 segments i *the*, que apareix en 3 segments, en l'exemple). Sovint, no es tenen en compte les paraules molt curtes (per exemple de menys de 2 o 3 caràcters) en els índexos inversos[3]. Les paraules molt llargues acostumen a ser paraules corresponents a categories obertes i en moltes llengües aquestes estan sotmeses a flexió. Fixem-nos per exemple en *banknote* y banknotes, que són en realitat la mateixa paraula flexionada. En el cas de paraules llargues sovint es prenen els primers *n* caràcters, per exemple els 5 o 6 primers. Tenint en compte aquests aspectes, l'índex invers ens quedaria de la següent manera (fixem-nos que per les paraules de més de 6 caràcters s'han indexat només els 6 primers) :

3 Això és per a llengües que facin servir caràcters alfabètics i no es pot aplicar a llengües que fan servir ideogrames.

approa 8	financ 10	policy 4
bankno 6:8	foreig 7:7	preven 9
been 8	framew 1:2	provis 3
before 8	fraud 9	reserv 7
change 8	initia 8	search 1
close 8	instit 3	settle 5
coin 8	januar 8	stabil 10
coins 6	legal 1:2	succes 8
conduc 9	market 10	system 5
curren 6	matter 6	terms 8
demand 8	means 6	total 8
distri 8	moneta 4	transp 9
employ 9	needs 8	volume 8
escb 2	operat 4	
exchan 7	paymen 5:6	

El fet d'indexar només els 6 primers caràcters de les paraules de més de 6 caràcters ens ha permès indexar amb el mateix índex les formes *banknote* i *banknotes* (sota l'índex *bankno*); no ha permès ajuntar les formes *coin* i *coins*. Cal tenir en compte que en general es busquen formes d'indexació generals, que serveixin per a moltes llengües i que no incorporin coneixement lingüístic sobre una determinada llengua.

Ara, per cercar els segments més semblants a

Banknotes, coins, and types of payment

seguiríem la mateixa estratègia i també retallaríem les paraules de més de 6 lletres en les seves 6 primeres lletres.

```
bankno 6
coins 6
types
paymen 5:6
```

i el resultat seria el mateix, és a dir, que el segment més semblant seria el 6 (*Banknotes and coins, means of payment and currency matters*) i el segon més semblant el 5 (*Payment and settlement systems*), és a dir, exactament igual que en el cas anterior.

Recordem que els desenvolupadors de programes de traducció assistida intenten que les seves eines funcionin per un gran nombre de llengües i en general s'eviten fer servir mètodes d'indexació i recuperació que requereixin d'informació lingüística. Si afegim coneixement lingüístic, la indexació de memòries de traducció pot millorar notablement.

Una primera estratègia pot ser l'ús de la tècnica coneguda com a *stemming* que consisteix a eliminar els afixos morfològics de les paraules. Aquest procés es pot dur a terme fent servir diversos algorismes, com el de Porter (1980),

Veiem a continuació el resultat de fer servir l'algorisme de Porter al nostre exemple:

ID	Segment original	Segment original stemmer Porter
1	Search the Legal framework	search the legal framework
2	Legal framework of the ESCB	legal framework of the ESCB
3	ECB institutional provisions	ECB institut provis
4	Monetary policy and Operations	monetari polici and operation
5	Payment and settlement systems	payment and settlement system
6	Banknotes and coins, means of payment and currency matters	banknot and coin, mean of payment and currenc matter
7	Foreign exchange and Foreign reserves	foreign exchang and foreign reserv
8	The approach was successful: in volume terms, close to 80 % of the initial banknote demand and 97 % of the total coin needs for the changeover had been distributed before 1 January 2002.	the approach wa successful: in volum term, close to Z % of the initi banknot demand and Z % of the total coin need for the changeov have be distribut befor Z januari Z.
9	Employment, conduct , fraud prevention and transparency	employ, conduct , fraud preventi and transpar
10	Financial market stability	financi market stabil

Els indexos queden de la següent manera (eliminem també els *stems* de tres o menys lletres):

```
approach 8
banknot 6:8
befor 8
changeov 8
close 8
coin 6:8
conduct 9
currenc 6
demand 8
distribut 8
employ 9
escb 2
exchang 7
financi 10
foreign 7:7
framework 1:2
fraud 9
have 8
initi 8
institut 3
januari 8
legal 1:2
market 10
matter 6
mean 6
monetari 4
need 8
operation 4
payment 5:6
polici 4
preventi 9
provis 3
reserv 7
search 1
settlement 5
stabil 10
successful 8
system 5
term 8
total 8
transpar 9
volum 8
```

Per fer ara la cerca del segment més semblant a:

Banknotes, coins, and types of payment

hauríem de fer servir el mateix *stemmer* a l'oració que cerquem, que quedaria:

banknote , coin , and type of payment

i consultant els índexos:

```
banknot 6:8
coin 6:8
type
payment 5:6
```

Resulta en què el segment que recupera en primer lloc és el 6 (*Banknotes and coins, means of payment and currency matters*) , seguit del 8 (*The approach was successful: in volume terms, close to 80 % of the initial banknote demand and 97 % of the total coin needs for the changeover had been distributed before 1 January 2002.*). Veiem que ara el segment 5 *(Payment and settlement systems*) que amb les estratègies anteriors es recuperava en segon lloc ara es recupera en tercer lloc.

Una segona aproximació, que requereix encara de més informació lingüística específica de la llengua de partira, consistiria en disposar d'un lematitzador per a la llengua de partida (l'anglès en aquest exemple). El lematitzador és capaç de substituir cada una de les paraules pel seu lema, és a dir, la seva forma base. Així, la taula del nostre exemple quedaria de la següent manera (lematitzem només la llengua de partida, ja que és la que fem servir per fer les cerques). Fixem-nos també que podem substituir les xifres per una etiqueta que ens indiqui simplement que es tracta d'una xifra.

ID	**Segment original**	**Segment original lematitzat**
1	Search the Legal framework	search the legal framework
2	Legal framework of the ECB	legal framework of the ECB
3	ECB institutional provisions	ECB institutional provision
4	Monetary policy and Operations	monetary policy and operation
5	Payment and settlement systems	payment and settlement system
6	Banknotes and coins, means of payment and currency matters	banknote and coin, mean of payment and currency matter
7	Foreign exchange and Foreign reserves	foreign exchange and foreign reserve
8	The approach was successful: in volume terms, close to 80 % of the initial banknote demand and 97 % of the total coin needs for the changeover had been distributed before 1 January 2002.	the approach be successful: in volume term, close to Z % of the initial banknote demand and Z % of the total coin need for the changeover have be distribute before Z january Z.
9	Employment, conduct , fraud prevention and transparency	employment, conduct , fraud prevention and transparency
10	Financial market stability	financial market stability

Ara els índexos quedarien de la següent manera (eliminem també les lemes de 3 o menys caràcters):

```
approach 8
banknote 6:8
before 8
changeover 8
close 8
coin 6:8
conduct 9
currency 6
demand 8
distribute 8
employment 9
exchange 7
financial 10
foreign 7:7
framework 1:2
fraud 9
have 8
initial 8
institutional 3
january 8
legal 1:2
market 10
matter 6
mean 6
monetary 4
need 8
operation 4
payment 5:6
policy 4
prevention 9
provision 3
reserve 7
search 1
settlement 5
stability 10
successful: 8
system 5
term 8
total 8
transparency 9
volume 8
```

Per fer ara la cerca del segment més semblant a:

Banknotes, coins, and types of payment

hauríem de lematitzar també l'oració, que quedaria:

banknote , coin , and type of payment

i consultant els índexos:

```
banknote 6:8
coin 6:8
type
payment 5:6
```

el segment que es recuperaria en primer lloc seria el 6 (amb 3 lemes coincidents), el 8 (amb 2) i el 5 (amb 1); aquests resultats són iguals que en el cas de fer servir l'*stemmer*.

Si disposem d'un analitzador més potent per a la llengua de partida podem millorar encara més els índexos. Per exemple, si el nostre analitzador és capaç de lematitzar i indicar la categoria gramatical de cada paraula podem afegir aquesta informació als índexos i fer després la cerca amb aquest criteri. Per exemple, en el segment 8 la paraula *approach* pot ser tant un nom com un verb. El nostre etiquetador l'etiqueta correctament com a nom. Després podrem fer servir aquesta informació per cercar les paraules per una determinada categoria. També podem fer servir la informació de l'etiquetador per indexar únicament les paraules que pertanyin a categories obertes (noms, verbs, adjectius i adverbis). En la següent taula podem veure la versió lematitzada i etiquetada amb un conjunt d'etiquetes molt reduït (n: nom; v: verb; a: adjectiu; r: adverbi; x: qualsevol altra categoria).

ID	**Segment original**	**Segment original lematitzat i etiquetat**
1	Search the Legal framework	search_n the_x legal_a framework_n
2	Legal framework of the ESCB	legal_a framework_n of_x the_x escb_n
3	ECB institutional provisions	ecb_n institutional_a provision_n
4	Monetary policy and Operations	monetary_a policy_n and_x operation_n
5	Payment and settlement systems	payment_n and_x settlement_n system_n
6	Banknotes and coins, means of payment and currency matters	banknote_n and_x coin_n ,_x **mean_v** of_x payment_n and_x currency_n matter_n
7	Foreign exchange and Foreign reserves	foreign_a exchange_n and_x foreign_a reserve_n
8	The approach was successful: in volume terms, close to 80 % of the initial banknote demand and 97 % of the total coin needs for the changeover had been distributed before 1 January 2002.	the_x approach_n be_v successful_a :_x in_x volume_n term_n ,_x close_x to_x Z_z %_x of_x the_x initial_a banknote_n demand_n and_x Z_x %_x of_x the_x total_a coin_n **need_v** for_x the_x changeove_n have_v be_v distribute_v before_x Z_x january_n Z_x ._x
9	Employment, conduct , fraud prevention and transparency	employment_n ,_x conduct_n ,_x fraud_n prevention_n and_x transparency_n
10	Financial market stability	financial_a market_n stability_n

El fet d'afegir més informació pot millorar la indexació i la recuperació de segments similars, però també fa que el sistema sigui més vulnerable a errors. Per exemple, en el segment 7 la paraula *means* s'ha etiquetat errròniament com a verb i el en segment 8 la paraula *needs* també s'ha etiquetat errròniament com a verb. Els índexos fent servir aquests lemes i etiquetes, i evitant la indexació de les paraules no pertanyents a categories obertes, quedaria de la següent manera:

approach_n 8	foreign_a 7:7	payment_n 5:6
banknote_n 6:8	framework_n 1:2	policy_n 4
be_v 8:8	fraud_n 9	prevention_n 9
changeove_n 8	have_v 8	provision_n 3
coin_n 6:8	initial_a 8	reserve_n 7
conduct_n 9	institutional_a 3	search_n 1
currency_n 6	january_n 8	settlement_n 5
demand_n 8	legal_a 1:2	stability_n 10
distribute_v 8	market_n 10	successful_a 8
ecb_n 3	matter_n 6	system_n 5
employment_n 9	mean_v 6	term_n 8
escb_n 2	monetary_a 4	total_a 8
exchange_n 7	need_v 8	transparency_n 9
financial_a 10	operation_n 4	volume_n 8

Per fer ara la cerca del segment més semblant a:

Banknotes, coins, and types of payment

hauríem de lematitzar i etiquetar també l'oració, que quedaria:

banknote_n ,_x coin_n ,_x and_x type_n of_x payment_x

i com que només indexem les categories obertes els índexos quedarien:

```
banknote_n 6:8
coin_n 6:8
type_n
payment_n 5:6
```

El segment que es recuperaria en primer lloc seria el 6 (amb 3 coincidències), seguit del 8 (amb 2 coincidències) i finalment el 5 (amb una coincidència)

Fixem-nos que aplicant tècniques que fan servir coneixement lingüístic (*stemming*, lematització i informació de categoria gramatical) hem obtingut resultats diferents que amb les tècniques basades en paraules senceres o bé en fragments de paraules sense motivació lingüística. El que ens queda per determinar ara és quin dels segments recuperats és el més adient per mostrar en primer lloc al traductor.

Tots aquests canvis, per a una memòria de 10 segments no són gaire significatius. Imaginem-nos, però, una memòria amb 1.000.000 de segments i veurem que la mida dels índexs es podria reduir notablement.

El que hem explicat fins aquí sobre la indexació de memòries de traducció és una aproximació a com es fa realment. Cada eina de traducció assistida pot fer servir una estratègia o una altra d'indexació per aconseguir una millor eficiència. També es poden fer servir tècniques més avançades provinents del camp de la recuperació de la informació (Frakes and Baeza-Yates, 1992).

Un cop el programa de traducció assistida ha recuperat una sèrie de segments de la memòria de traducció fent servir un sistema d'indexació, haurà de calcular un índex de similitud entre l'oració que cerquem i tots els segments recuperats de la memòria. Expliquem aquest aspecte en el següent apartat.

2.2.2. Càlcul de la similitud de segments

En aquest apartat veurem diferents maneres de calcular la similitud entre dos segments. El primer que hem de tenir en ment és què volem que signifiqui aquest tant per cent de similitud: en quin grau s'assemblen? quin esforç s'ha de fer per passar del segment obtingut al desitjat? Per a un traductor probablement serà el

segon aspecte, és a dir, obtenir una mena de mesura que ens indiqui l'esforç de canvi. Veurem un parell d'estratègies genèriques i observarem quina de les dues estratègies s'aproxima millor a aquest objectiu.

Càlcul de paraules coincidents

La primera idea que ens ve al cap per calcular la similitud entre dos segments és mirar quantes paraules tenen en comú. Si totes les paraules són iguals, els segments tindran un 100% de similitud (això pot fallar per l'ordre de les paraules). Anem a calcular la similitud segons això pels segments de l'exemple anterior. Primer calcularem el % tenint en compte el nombre de paraules iguals respecte al nombre de paraules total.

Segment que cerquem:	Banknotes, coins, and types of payment
1r segment trobat:	Banknotes and coins, means of payment and currency matters
Paraules coincidents:	3 Total paraules (segment a cercar): 6 Similitud: 50%
2n segment trobat:	Payment and settlement systems
Paraules coincidents:	1 Total paraules (segment a cercar): 6 Similitud: 16.6%

Ara fem el càlcul tenint en compte el nombre de caràcters de les paraules que són iguals respecte al nombre total de caràcters.

Segment que cerquem:	Banknotes, coins, and types of payment
1r segment trobat:	Banknotes and coins, means of payment and currency matters
Caràcters coincidents:	22 Total caràcters (segment a cercar): 34 Similitud: 64.7%
2n segment trobat:	Payment and settlement systems
Caràcters coincidents:	7 Total caràcters (segment a cercar): 34 Similitud: 20.6%

Fixem-nos però, que alguns canvis d'ordre de paraules podrien fer que l'esforç d'edició per mantenir el significat fons molt superior al % calculat.

Càlcul de la distància d'edició

La *distància d'edició* o *distància de Levenshtein* és el nombre mínim d'edicions requerides (inserció, supressió o substitució d'un caràcter) per a transformar una cadena de caràcters en una altra.

Aquest càlcul ens pot donar una idea molt aproximada de l'esforç real que pot suposar editar una coincidència parcial d'una memòria de traducció en la traducció real del segment original. Per aquest motiu es pot fer servir amb èxit per al càlcul de la similitud entre dos segments.

A continuació podem observar el codi Python d'una función per a calcular la distància d'edició de Levenshtein:

```
def distance(str1, str2):
    d=dict()
    for i in range(len(str1)+1):
        d[i]=dict()
        d[i][0]=i
    for i in range(len(str2)+1):
        d[0][i] = i
    for i in range(1, len(str1)+1):
        for j in range(1, len(str2)+1):
            d[i][j] = min(d[i][j-1]+1, d[i-1][j]+1, d[i-1][j-1]+(not str1[i-1] == str2[j-1]))
    return d[len(str1)][len(str2)]
```

Si apliquem aquest algorisme a l'exemple que ens ocupa obtenim les següents xifres:

Segment que cerquem:	Banknotes, coins, and types of payment
1r segment trobat:	Banknotes and coins, means of payment and currency matters
Distància d'edició:	31
2n segment trobat:	Payment and settlement systems
Distància d'edició:	29

Aquest resultat ens pot sorprendre una mica, ja que el segon segment trobat, tot i tenir menys paraules coincidents, necessita menys esforç d'edició per poder formar el segment original cercat.

En Somers (2003) es presenta un exemple molt clar de com les tècniques simples basades en la distància d'edició poden no funcionar correctament en la selecció del segment més semblant (reproduïm aquí la part anglesa de l'exemple). Considerem que hem de traduir la següent oració:

Select 'Symbol' in the Insert menu.

i que disposem d'una memòria de traducció amb els següents segments

S	Source	Target	Distància edició
1	Select 'Symbol' in the Insert menu to enter a character from the symbol set.	Seleccioni 'Símbol' en el menú Insereix per introduir un caràcter del conjunt de símbols.	41
2	Select 'Paste' in the Edit menu.	Seleccioni 'Enganxa' en el menú Edita.	11
3	Select 'Paste' in the Edit menu to enter some text from the clipboard.	Seleccioni 'Engaxa' en el menú Edita per introduir el text del porta-retalls.	46

La majoria de mètriques de similitud basades en la distància d'edició seleccionarien com a més semblant el segment 2, ja que només canvien dues paraules i la distància d'edició és més petita. Però intuïtivament el segment 1 és una millor coincidència ja que tot i que la distància d'edició és molt més gran, en canvi inclou de manera exacta el text que cerquem. Si recuperem la traducció de la memòria de traducció només haurem d'esborrar la part final (*per introduir un caràcter del conjunt de símbols*).

Fixem-nos també que els segments:

Select 'Symbol' in the Insert menu to enter a character from the symbol set.
Select 'Paste' in the Edit menu to enter some text from the clipboard.

que tenen una distància d'edició de 30 i 8 paraules en comú i 6 paraules diferents.

Són més semblants entre sí que els segments:

Select 'Symbol' in the Insert menu.
Select 'Paste' in the Edit menu.

que tenen una distància d'edició de 11 i 4 paraules en comú i 2 paraules diferents.

Així doncs, una métrica de similitud hauria de tenir en compte no només la distància d'edició, sinó també el nombre de paraules en comú i paraules diferents i fins i tot la llargada del segment. Per tenir en compte també la longitud dels segments que es comparen es pot fer servir el coeficient de Dice (Trujillo, 1999), que es defineix com a:

$$S = \frac{2C}{A+B} = \frac{2|A \cap B|}{|A| + |B|}$$

on A i B és el nombre de paraules en el segment d'entrada i en el segment recuperat de la memòria de traducció, respectivament. C és el nombre de paraules comunes en els dos segments. Tot i tenir en compte el nombre total de paraules, aquesta mesura no té en compte l'ordre de les paraules. També cal tenir en compte què es fa amb les paraules duplicades, si es tenen en compte com a una o com a dues paraules. En algunes implementacions d'aquesta mesura, en lloc de computar per paraules es computa per bígrams, és a dir, per grups de dues paraules adjacents. Hi ha moltes més mesures que intenten determinar quina és la similitud entre dues cadenes. En la següent taula veurem el resultat d'aplicar les següents mesures (no proporcionem un definició de cada una de elles)::

- Levenshtein distance
- Jaccard distance
- Jaro distance
- Jaro Wiknkler distance
- Dice Coefficient
- Longest common subsequence

Als segments de l'apartat anterior:

S: *Banknotes, coins, and types of payment*

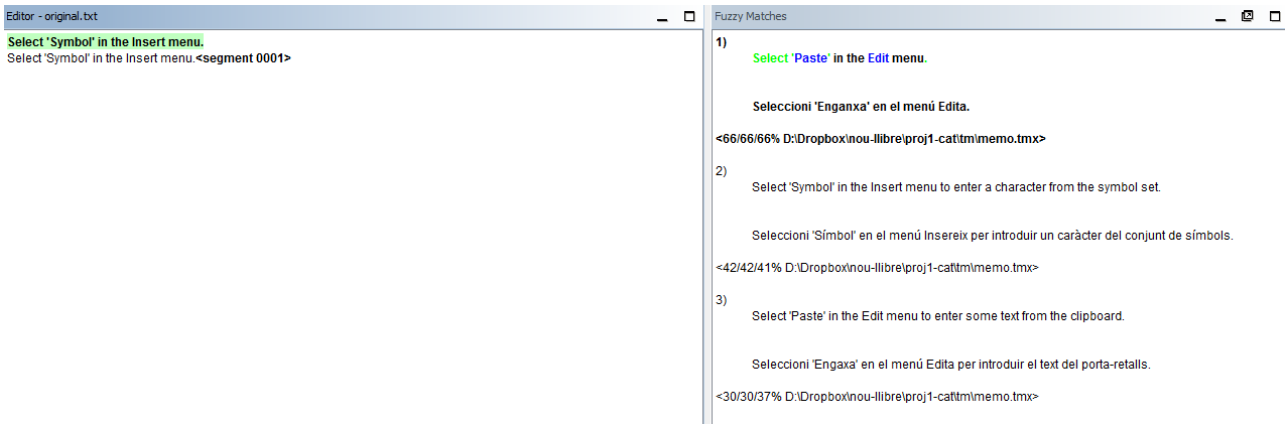
M6: *Banknotes and coins, means of payment and currency matters*
M8: *The approach was successful: in volume terms, close to 80 % of the initial banknote demand and 97 % of the total coin needs for the changeover had been distributed before 1 January 2002.*
M5: *Payment and settlement systems*

	S - M6	S - M8	S - M5
Levenshtein distance	31	158	**27**
Jaccard distance	**0.1053**	0.5556	0.5263
Jaro distance	**0.8326**	0.6439	0.7082
Jaro Winkler distance	**0.8995**	0.6439	0.7082
Dice Coefficient	**0.617**	0.2072	0.303
Longest common subsequence	31	**27**	14

Per a cada mesura marquem en negreta el parell de segments més similars segons aquesta mesura. La majoria de mesures coincideixen en determinar que els segments més semblants són el S - M6. En canvi, segons la Levenshtein distance els més semblants són el S - M5, i segons la Longest common subsequence les més semblants són la S - M8.

Veiem ara a la pràctica quin segment selecciona com a més similar l'eina OmegaT:

[per la versió catalana]


Editor - original.txt
Select 'Symbol' in the Insert menu.
Select 'Symbol' in the Insert menu.<segment 0001>
Fuzzy Matches
1)
Select 'Paste' in the Edit menu.
Seleccioni 'Enganxa' en el menú Edita.
<66/66/66% D:\Dropbox\nou-llibre\proj1-cat\tm\memo.tmx>
2)
Select 'Symbol' in the Insert menu to enter a character from the symbol set.
Seleccioni 'Símbol' en el menú Insereix per introduir un caràcter del conjunt de símbols.
<42/42/41% D:\Dropbox\nou-llibre\proj1-cat\tm\memo.tmx>
3)
Select 'Paste' in the Edit menu to enter some text from the clipboard.
Seleccioni 'Engaxa' en el menú Edita per introduir el text del porta-retalls.
<30/30/37% D:\Dropbox\nou-llibre\proj1-cat\tm\memo.tmx>

Veiem que en primer lloc selecciona el segment 2 (assignant-li una similitud del 66/66/66%); el segon seleccionat és el segment 1 (42/42/41%) i en tercer lloc el segment 3 (30/30/37%). Què signifiquen exactament aquests percentatges que assigna OmegaT?

- La primera xifra ens indica el percentatge de similitud fent servir el mòdul de tokenització que permet calcular l'arrel de les paraules i detectar les formes flexionades. En aquest exemple les dues primeres xifres coincideixen perquè no tenim aquest mòdul activat i perquè totes les paraules de l'original estan sense flexionar.
- La segona xifra ens indica el percentatge del nombre de paraules coincidents (ignorant els números i les etiquetes) dividit entre el nombre total de paraules.
- La tercera xifra és igual que l'anterior però tenint en compte les xifres i les etiquetes.

Veiem ara com respon a aquest exemple l'eina Tikal (d'Okapi Tools) si indexem els segments del nostre exemple en una memòria de traducció de tipus Pensieve (la pròpia d'Okapi Tools).

[per la versió catalana]

```
tikal.sh -q "Select 'Symbol' in the Insert menu." -pen memoPengcat.pentm/ -sl en -tl ca -opt 0
------------------------------------------------------------------------------
Okapi Tikal - Localization Toolset
Version: 2.0.23
------------------------------------------------------------------------------

= From net.sf.okapi.connectors.pensieve.PensieveTMConnector (en->ca)
  Threshold=0, Maximum hits=25
score: 53, origin: ''
  Source: "
           Select 'Symbol' in the Insert menu to enter a character from the symbol set.
         "
  Target: "
                Seleccioni 'Símbol' en el menú Insereix per introduir un caràcter del conjunt de
símbols.
         "
```

Veiem doncs, que comparant només dues eines concretes, ja es produeixen discrepàncies importants en les coincidències que s'obtenen.

La recuperació de segments i el càlcul de la similitud entre el segment a cercar i els recuperats des de la memòria de traducció és un aspecte molt important en la tècnica de traducció automàtica anomenada *traducció automàtica basada en exemples* (*Example-Based Machine Translation* - EBMT). En els següents paràgrafs en descrivim algunes.

Alguns autors fan propostes diferents per al càlcul de la similitud entre segments (Somers i Fernández, 2004). Per exemple, Denet (1995) proposa per una banda tenir en compte la significància relativa de les paraules que canvien, basant-se en dades estadístiques, i per una altra banda identificar fragments de segments que siguin significatius des del punt de vista sintàctic. Cranias et al (1991) proposa considerar lemes en comptes de cadenes de caràcters i fer ús de les paraules funcionals així com de les etiquetes morfosintàctiques.

Planas i Furuse (1999) proposen un esquema de cerca de coincidències multi-capa i flexible. Els autors proposen 8 capes: (1) caràcters de text, (2) paraules, (3) lemes, (4) categories gramaticals (POS), (5) etiquetes XML de contingut, (6) etiquetes XML buides (que representen per exemple imatges), (7) entrades de glossari i (8) estructures d'anàlisi lingüística (aquesta capa dependrà del nivell d'anàlisi que proporcioni l'analitzador lingüístic disponible). La similitud entre aquestes estructures es calcula a partir de la distància d'edició.

Macklovitch i Russell (2000) tenen en compte altres aspectes com la flexió de les paraules i les categories gramaticals i també consideren el reconeixement de certes entitats amb nom, el reconeixement de noms propis i l'anàlisi sintàctica superficial.

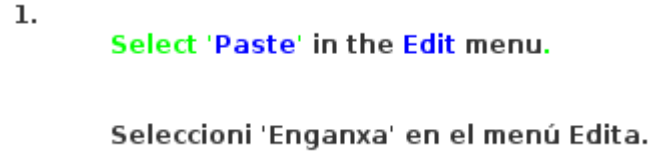
Rapp (2002) fa servir un etiquetador morfosintàctic per processar les memòries de traducció i fer servir la informació de categoria gramatical per millorar la cerca de segments similars.

Indicació de les paraules coincident i diferents amb un codi de colors

És interessant que el programa de traducció assistida mostri les coincidències parcials de les memòries amb indicació de les paraules que són coincidents i les que són diferents. Sovint això es fa mitjançant un codi de colors.

Seguint amb el mateix exemple que en els apartats anteriors, observem en la següent imatge com OmegaT marca amb colors la coincidència més semblant de la memòria de traducció (en l'exemple, el segment que estem traduïnt és: *Select 'Symbol' in the Insert menu*.)

[versió catalana]



El programa és capaç de detectar les paraules iguals (en verd) i diferents (en blau) del text corresponent a la llengua de partida, però no marca de cap color el segment corresponent a la llengua d'arribada, ja que no té prou informació lingüística per determinar quina paraula de la llengua de partida correspon amb quina paraula de la llengua d'arribada.

2.3. Coincidència exacta i parcial

Així doncs, les coincidències que es recuperen de la memòria de traducció poden ser:

- Coincidència exacta (*exact match*): quan els text del segment que recuperem de la memòria de traducció és exactament igual al text del segment que cerquem.
- Coincidència parcial (*fuzzy match)*: quan els text del segment que recuperem de la memòria de traducció no és exactament igual al text del segment que cerquem, però el seu índex de similitud és superior o igual a l'índex fixat per l'usuari.

A aquesta definició simplista cal afegir algunes explicacions:

- Una coincidència que només difereixi en algunes xifres es pot considerar exacta si el programa de traducció assistida és capaç de subtituir-les per les xifres presents en el segment que cerquem (moltes de les eines actuals són capaces de dur a terme aquesta substitució). Considerem que estem traduint el segment "*An example is shown in figure 3"*. A la nostra memòria tenim el segment *"An example is shown in figure 1"* amb la seva corresponent traducció *"Es mostra un exemple en la figura 1"*. Moltes de les eines actuals són capaces de recuperar el segment i mostrar el text traduït amb la xifra substituïda *"Es mostra un exemple en la figura 3"*.
- Algunes eines són capaces de fer aquestes substitucions per cadenes alfanumèriques, no només per a xifres i serien capaces de fer el mateix per a un segment com *"This is shown as A in the diagram"*. (exemple pres de Somers (2004)). Si en la nostra memòria tenim un segment com *"This is shown as B in the diagram"* amb la seva corresponent traducció *"Això s'indica com a B en el diagrama"*, el programa proposaria la traducció canviant la B per A i amb una coincidència exacta *"Això s'indica com a A en el diagrama"*.
- Un altre cas a tenir en compte és aquell en el que els textos del segment a cercar i el recuperat de la memòria són exactament iguals però difereixen en algun tipus de format o marca especial. Posem per cas que el text a cercar és "Press **OK** to continue" (amb OK en negreta) i a la memòria tenim un parell "Press OK to continue" i la seva traducció "Premeu D'acord per continuar" [versió castellana: "Haga clic en Aceptar para continuar."]. En aquest cas el programa podrà recuperar el text però no serà capaç de posar D'acord en negreta.

En aquest sentit, Bowker (2002) introdueix la distinció entre *coincidència exacta* (*exact match*) i *coincidència completa* (*full match*). Segons aquest autor:

> *Una coincidència exacta és 100% idèntica al segment que està traduint el traductor tant des del punt de vista lingüístic, com des del punt de vista del format. Això significa que les dues cadenes han de ser idèntiques en tots els sentits, incloent l'ortografia, la puntuació, la flexió, xifres i inclòs el format (cursiva, negreta, etc.)*

Una *coincidència total* es produeix quan el segment que s'està traduint difereix del segment emmagatzemat a la memòria de traducció només en el que s'anomenen *element variables*, que sovint en anglès reben el nom de *placeables*. Aquests elements variables inclouen xifres, dates, hores i unitats monetàries. Algunes eines de traducció assistida són capaces de detectar i fer canvis d'alguns d'aquests elements variables. El cas més habitual és el de les xifres. Per exemple: si estem traduint un fragment com "The paper tray has a capacity of 200 sheets" i en la memòria tenim un segment "The paper tray has a capacity of 150 sheets" i la seva traducció "La safata de paper té una capacitat de 150 fulls" moltes de les eines actuals podran recuperar el segment com a coincidència exacta canviant la xifra i mostrant "La safata de paper té una capacitat de 200 fulls".

2.4. Combinació d'unitats subsegmentals

En molts casos a la memòria de traducció no tenim cap segment complet semblant al que estem traduint, però en canvi disposem d'un o més fragments de segments que contenen informació interessant. Veiem un exemple (adaptat de Bowker 2002)

Segment a traduir	First, **check for disk space on the drive** that contain the Temp folder.
Segment de la memòria de traducció	(eng) Close other programs, **check for disk space on the drive** you are saving to, and then save again (cat) Tanqui els altres programes, verifiqui l'espai de disc en la unitat on està guardant i després torni a guardar.

Alguns programes són capaços de cercar a la memòria coincidències a nivell sub-segmental. Una tasca més complexa, però que algunes eines poden arribar a fer, és deduir que la traducció de *check for disk space on the drive* al català és *verifiqui l'espai de disc en la unitat*. El programa pot arribar a deduir això d'una manera totalment estadística si aquest subsegment apareix a diversos segments de la memòria de traducció.

Un pas més enllà encara és la capacitat d'alguns programes per composar una nova traducció a partir de coincidències subsegmentals. Veiem el següent exemple (també adaptat de Bowker 2002)

Segment a traduir	**The file operation cannot be completed because the disk is full**.
Segment de la memòria de traducció	(eng) There is not enough memory to perform **the file operation**. (cat) No hi ha prou memòria per dur a terme **l'operació d'arxius**.
Segment de la memòria de traducció	(eng) This action **cannot be completed because the** program is busy. (cat) Aquesta acció **no es pot completar perquè el** programa està ocupat.
Segment de la memòria de traducció	(eng) **Disk is full** (cat) **El disc està ple**
Traducció composada a partir de les coincidències subsegmentals	L'operació d'arxius no es pot completar perquè el el disc està ple

Simard (2001) presenta un sistema capaç de treballar a nivell sub-segmental que funciona a partir de la consideració de totes les subseqüències possibles (n-grames) del segment a traduir i recupera també tots els parells de subseqüències possibles de la memòria de traducció. Langais (2001) i Colominas (2008) presenten propostes on les subseqüències tenen una motivació lingüística (es tracta de *chunks*: és a dir un sintagma no recursiu corresponent a una categoria lèxica principal (nom, adjectiu, preposició i verb), que admeten que junt amb el nucli poden incloure tant premodificadors com postmodificadors. En l'exemple anterior no s'han tingut en compte *chunks*, sinó unitats arbitràries. En la taula següent podem observar el mateix exemple si treballem amb chunks (calculats amb Freeling)

Segment a traduir	[The file operation] [cannot be completed] because [the disk] is full.
Segment de la memòria de traducció	(eng) There is [not enough memory] to perform [the file operation]. (cat) No hi ha [prou memòria] per dur a terme [l'operació d'arxius].
Segment de la memòria de traducció	(eng) [This action] [cannot be completed] because [the program] is busy. (cat) [Aquesta acció] [no es pot completar] perquè [el programa està ocupat].

Segment de la memòria de traducció	(eng) Disk is full (cat) El disc està ple
Traducció composada a partir de les coincidències subsegmentals	L'operació d'arxius no es pot completar perquè el disc està ple

Recordem que el sistema de traducció assistida només podrà determinar la traducció d'una subseqüència si aquesta apareix força vegades en la memòria de traducció. Com que no sempre les memòries de traducció que s'assignen a un projecte tenen la mida suficient, alguns autors (Biçici 2008) proposen fer servir un sistema de traducció automàtica estadística basat en frases entrenat amb un corpus del mateix domini.

2.5. Format d'intercanvi de memòries de traducció: TMX

Com hem vist en els apartats anteriors, els diferents sistemes de traducció assistida indexen i emmagatzemen en bases de dades les memòries de traducció de formes molt diferents. Per aquest motiu, les memòries de traducció no són compatibles entre les diferents eines. Per poder compartir-les s'ha creat un llenguatge d'intercanvi basat en XML anomenat *Translation Memory eXchange* (TMX). Totes les eines de traducció assistida són capaces de guardar les seves memòries en aquest format i carregar fitxers TMX en les seves bases de dades.

El TMX es defineix en dues parts:

- Una especificació del format del contenidor, és a dir, els elements de nivell superior que proporcionen informació sobre el arxiu en conjunt i sobre les entrades. En TMX una entrada consistent en segments alineats de text en dos o més llengües s'anomena *unitat de traducció* (l'element <tu>).
- Una especificació per al format de meta-marcatge de baix nivell per al contingut de un segment de text de la memòria de traducció. En TMX, un segment individual del text de la memòria de traducció en una llengua determinada se denota amb l'element <seg>.

A continuació podem observar un exemple de memòria de traducció en format TMX consistent en un únic segment en castellà i català:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE tmx SYSTEM "tmx11.dtd">
<tmx version="1.1">
  <header creationtool="OmegaT" o-tmf="OmegaT TMX" adminlang="EN-US" datatype="plaintext"
creationtoolversion="2.6.3" segtype="sentence" srclang="CA"/>
  <body>
<!-- Default translations -->
    <tu>
      <tuv lang="CA">
        <seg>EDICTE de 14 de febrer de 2000, sobre un acord de la Comissió d'Urbanisme de Tarragona
referent al municipi de Reus.</seg>
      </tuv>
      <tuv lang="ES" changeid="aoliverg" changedate="20140508T150609Z">
        <seg>EDICTO de 14 de febrero de 2000, sobre un acuerdo de la Comisión de Urbanismo de
Tarragona referente al municipio de Reus.</seg>
      </tuv>
    </tu>
<!-- Alternative translations -->
  </body>
</tmx>
```

El TMX pot tenir dos nivells d'implementació:

- Nivell 1. Únicament text pla: Suport només per al contenidor. Les dades entre els elements <seg> contenen únicament informació textual, sense marques de format.

- Nivell 2. Marcatge del contingut: Suport tant per al contenidor com per al contingut. Es fa servir el marcatge de contingut propi del TMX per permetre que altres eines que siguin compatibles amb TMX nivell 2 puguin recrear la versió traduïda d'un document original fent servir únicament el fitxer TMX.

L'exemple anterior de TMX correspondria a un Nivell 1. Si el segment original tingués el següent format:

```
EDICTE de 14 de febrer de 2000, sobre un acord de la Comissió d'Urbanisme de Tarragona referent al
municipi de Reus.
```


la memòria en TMX de Nivell 2 tindria el següent aspecte:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE tmx SYSTEM "tmx14.dtd">
<tmx version="1.4">
  <header creationtool="OmegaT" o-tmf="OmegaT TMX" adminlang="EN-US" datatype="plaintext"
creationtoolversion="2.6.3" segtype="sentence" srclang="CA"/>
  <body>
<!-- Default translations -->
    <tu>
      <tuv xml:lang="CA">
        <seg><bpt i="0" x="0">&lt;f0&gt;</bpt>EDICTE<ept i="0">&lt;/f0&gt;</ept><bpt i="1"
x="1">&lt;f1&gt;</bpt> de 14 de febrer de 2000, sobre un acord de la <ept
i="1">&lt;/f1&gt;</ept><bpt i="2" x="2">&lt;f2&gt;</bpt>Comissió d'Urbanisme<ept
i="2">&lt;/f2&gt;</ept><bpt i="3" x="3">&lt;f3&gt;</bpt> de Tarragona referent al municipi de
Reus.<ept i="3">&lt;/f3&gt;</ept></seg>
      </tuv>
      <tuv xml:lang="ES" changeid="aoliverg" changedate="20140508T150609Z">
        <seg><bpt i="0" x="0">&lt;f0&gt;</bpt>EDICTO<ept i="0">&lt;/f0&gt;</ept><bpt i="1"
x="1">&lt;f1&gt;</bpt> de 14 de febrero de 2000, sobre un acuerdo de la <ept
i="1">&lt;/f1&gt;</ept><bpt i="2" x="2">&lt;f2&gt;</bpt>Comisión de Urbanismo<ept
i="2">&lt;/f2&gt;</ept><bpt i="3" x="3">&lt;f3&gt;</bpt> de Tarragona referente al municipio de
Reus.<ept i="3">&lt;/f3&gt;</ept></seg>
      </tuv>
    </tu>
<!-- Alternative translations -->
  </body>
</tmx>
```

Si ens fixem, aquest nivell inclou informació sobre el format del segment. El nivell 2 de TMX és molt útil per traduir documentació amb format (negretes, colors, etc.) variat, ja que en molts casos podrà recuperar també les marques de format i estalviarà temps d'edició al traductor.

2.6. Creació de memòries de traducció

2.6.a. Traducció amb sistemes de traducció assistida

Si treballem habitualment amb sistemes de traducció assistida la creació de memòries de traducció és directa, ja que tots els sistemes TAO són capaços de generar les memòries per cada projecte de traducció.

2.6.b. Memòries de traducció i corpus paral·lels

El concepte de memòria de traducció i corpus paral·lel es pot considerar equivalent. Actualment hi ha una gran quantitat de corpus paral·lels disponibles a Internet. Podeu donar una ullada a la Col·lecció Opus (http://opus.lingfil.uu.se/) (Tiedemann 2012), de la que parlarem amb més detall en l'apartat *Per ampliar coneixements* d'aquest mateix capítol.

En general els corpus paral·lels són de mides relativament grans. Si hi ha disponible un corpus paral·lel per al nostre parell de llengües i especialitat, pot resultar d'utilitat descarregar-lo i fer-lo servir com a memòria de traducció dins dels nostres projectes de traducció.

2.6.c. Alineació manual de documents

L'alineació de documents és un procés pel qual es prenen un document original i la seva traducció i es genera un fitxer que relaciona els segments originals amb els corresponents segments traduïts, és a dir, una memòria de traducció. Aquest procés és útil per crear memòries de traducció a partir de documents originals i les seves traduccions. És important tenir en compte, com hem comentat més a dalt, que si els documents els hem traduït amb una eina de traducció assistida, no serà necessari portar a terme aquest procés, ja que podrem generar la memòria de traducció directament des de l'eina de traducció assistida.

El procés genèric d'alineació de documents es pot dividir en dos passos:

- Segmentació dels documents originals i traduïts
- Relacionar els segments originals amb els segments traduïts corresponents

La segmentació consisteix a dividir el text dels documents en segments a partir d'un conjunt de regles de segmentació. Les regles de segmentació ens indiquen on acaba un segment i on comença un altre. Una regla de segmentació ens podria indicar que un punt, seguit d'un espai en blanc i seguit d'una paraula que comença per majúscula indica un límit de segment. Aquesta regla ens podria segmentar correctament el text:

Avui he dinat a casa. Demà dinaré a la feina.

en els segments:

Avui he dinat a casa.

Demà dinaré a la feina.

Ara bé, aquesta regla no funcionaria correctament per segmentar el text:

El sr. Martínez no ha assistit a la reunió.

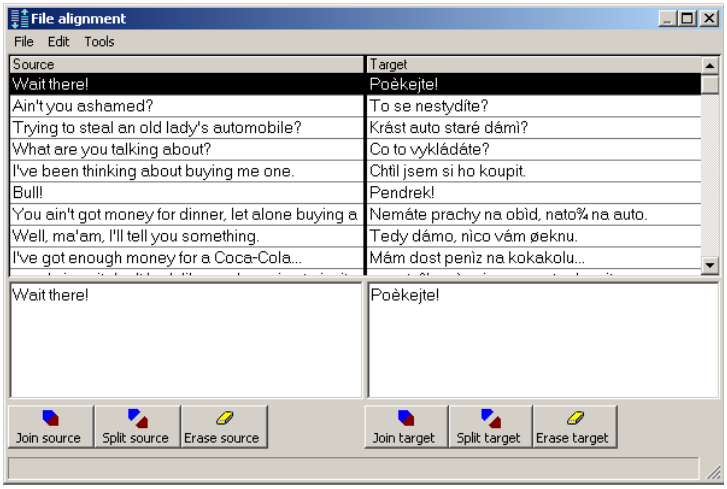
ja que resultaria la segmentació:

El sr.

Martínez no ha assistit a la reunió.

La majoria de sistemes de traducció assistida ofereixen la possibilitat d'especificar les regles de segmentació que fan servir. Per treure el màxim profit d'una determinada memòria de traducció convé fer servir les mateixes regles de segmentació en la creació del projecte que les que es van fer servir en la creació de la memòria de traducció. Per aquest motiu s'ha creat un format estàndard d'intercanvi de regles de segmentació basat en XML que s'anomena SRX (Segmentation Rule eXchange).

Les eines d'alineació manual de documents disposen d'una interfície gràfica que ens permet relacionar manualment els segments originals amb els corresponents segments traduïts.



Interfície d'alineació del Déja Vu 3

Si els documents original i traduït s'assemblen en quant a format i puntuació i la majoria de segments originals tenen una relació 1:1 (es a dir, cada segment original es correspon amb un segment traduït) l'alineació obtinguda únicament a partir de la segmentació serà prou acurada i es requerirà poca intervenció humana per completar l'alineació. Ara bé, això no sempre passa. Molt sovint un únic segment original es tradueix per dos segments (relació 1:2) o bé dos segments originals es tradueixen per un de sol (relació 2:1). Fins i tot a vegades passa que un segment original simplement no apareix a la traducció (relació 1:0) o que a la traducció apareixen nous segments (relació 0:1). Això fa que l'alineació manual de documents arribi a ser una ser una tasca realment feixuga i que requereixi una gran intervenció humana. Per aquest motiu s'han

desenvolupat diverses metodologies i eines d'alineació automàtica de documents, que veurem en el següent apartat.

2.6.d. Alineació automàtica de documents

L'alineació manual de documents pot suposar una càrrega de feina important. Quan la traducció respecta molt el format, els paràgrafs i el nombre d'oracions de l'original, l'alineació manual pot resultar efectiva. En altres casos l'alineació manual pot suposar una càrrega de feina més elevada que la que ens estalviarem traduint fent servir la memòria que podem generar a partir de l'alineació.

Hi ha una sèrie d'algorismes que permeten portar a terme l'alineació automàtica de documents. Aquests algorismes ens permetran alinear conjunts grans de documents sense pràcticament cap esforç i generar memòries de traducció.

L'alineació automàtica de documents segueix els passos genèrics de segmentació i relació de segments, però la relació de segments es fa de manera automàtica i sense intervenció de l'usuari.

Es poden distingir tres metodologies d'alineació automàtica:

- Basada en la longitud dels segments (en caràcters o paraules)
- Basada en un diccionari bilingüe
- Basada en tècniques gràfiques

La primera de les metodologies es basa en el fet que normalment els segments originals més llargs es tradueixen per segments més llargs. A partir de la segmentació dels documents es computen paràmetres estadístics basats en la longitud dels segments i es calculen aquests mateixos paràmetres estadístics de diverses variacions de la segmentació original. S'escull com a millor segmentació aquella que presenta una distribució més uniforme de la relació longitud del segment original respecte la longitud del segment traduït. Aquesta estratègia va ser emprada per primera vegada per Kay i Röscheisen (1993) mitjançant una aproximació que no va resultar gaire efectiva per a corpus de gran mida. Per una altra banda, i de manera força simultània Brown(1993) i Gale i Church (1993) van desenvolupar altres algorismes basats en aquesta mateixa idea.

La segona metodologia es basa en el fet de conèixer la traducció de certs mots o grups de mots. Si aquests mots apareixen en el segment original s'espera que en el segment traduït aparegui la corresponent traducció. A partir de la segmentació original es va modificant aquesta per aconseguir que el nombre de segments originals que presenten mots del diccionari i que el segment traduït corresponent contingui la traducció dels mots sigui màxim. Aquestes aproximacions es poden trobar en Chen (1993) i Wu (1994).

La tercera de les metodologies fa servir tècniques gràfiques (representant gràficament diversos paràmetres dels documents originals i traduïts) per trobar l'alineació més probable (Melamed 1997).

Moore (2002) ha desenvolupat una estratègia híbrida que fa servir tant la metodologia basada en la longitud de segments com la basada en diccionaris bilingües. El sistema presenta la particularitat de no necessitar d'un diccionari bilingüe ja que funciona en dos pasos. En el primer pas es du a terme una alineació automàtica basada en la longitud dels segments i el sistema pren únicament aquells parells de segments que

s'han pogut alinear amb molta seguretat. A partir d'aquestes alineacions segures el sistema aprèn automàticament un diccionari bilingüe estadístic que es fa servir per dur a terme una segona alineació basada en diccionari i complementarà a la primera alineació intentant alinear aquells segments no segurs. L'eina Hunalign (Varga el al. 2005) fa servir una estratègia híbrida molt semblant. En aquesta eina es pot fer servir també un diccionari bilingüe proporcionat per l'usuari. En el cas que no es proporcioni un diccionari el programa aprèn un amb una primera alineació basada en longitud.

2.7. Memòries de traducció remotes compartides i públiques

La manera tradicional de treballar amb memòries de traducció ha estat fins fa poc treballar amb les memòries locals, a la pròpia màquina o en una xarxa local. Com que els equips de traductors habitualment estan en diferents ubicacions, sovint a molta distància el fet de treballar amb memòries locals ha provocat alguns problemes:

- Si les memòries de traducció són molt grans, enviar-les als diferents traductors ha estat un problema per la mida dels arxius.
- A més, fer arribar tota una memòria de traducció a un col·laborador puntual pot crear problemes importants de confidencialitat
- Hi ha la possibilitat d'enviar només els segments de la memòria útils per al projecte en que està treballant actualment el traductor
- Tot i així, els nous segments traduïts per un altre col·laborador del mateix projectes no estaran disponibles per a la resta de traductors

Les tecnologies relacionades amb Internet han permès el desenvolupament de memòries de traducció remotes. En Simões (2004) es descriu un sistema de memòries de traducció distribuïdes implementades mitjançant serveis web.

Hi ha diverses eines que ofereixen implementacions eficients de memòries de traducció remotes, entre les que podem destacar:

- TM Server de Translate Toolkit (http://docs.translatehouse.org/projects/translate-toolkit)
- amaGama (http://amagama.translatehouse.org/)

Aquestes dues eines són de programari lliure.

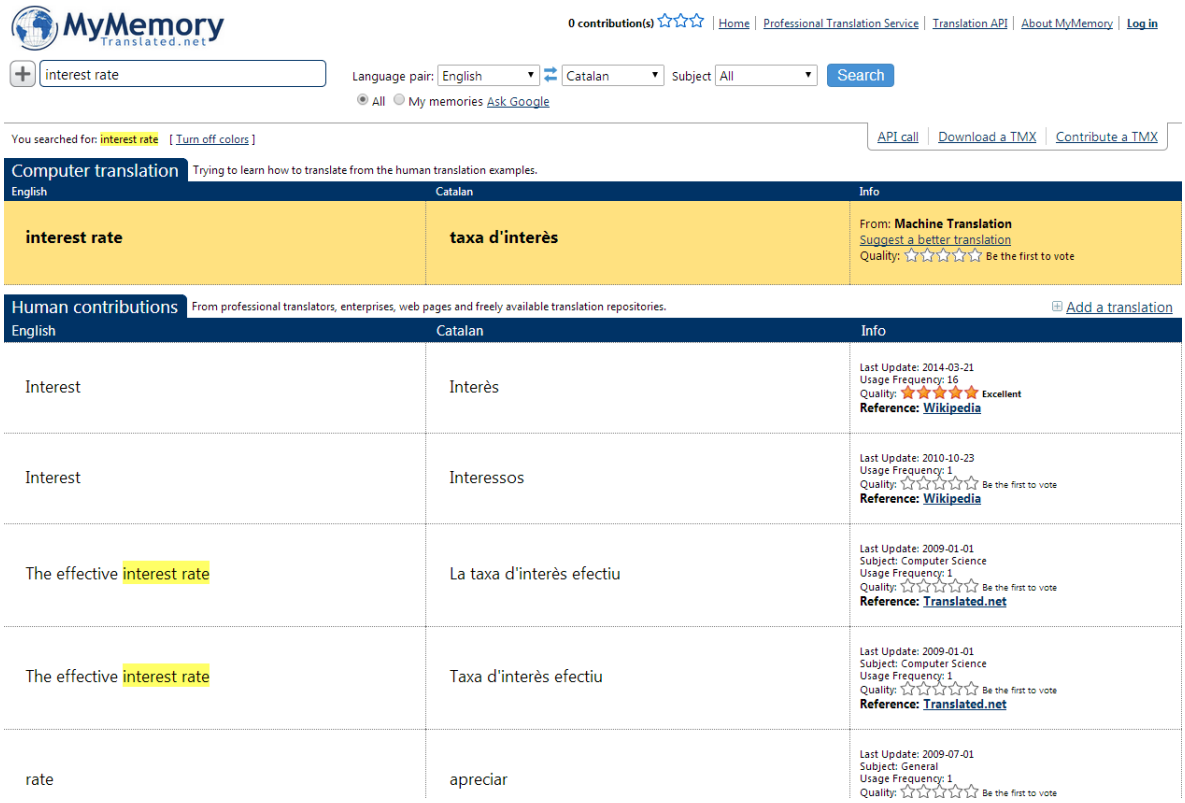
Algunes aplicacions que incorporen directament la funcionalitat de memòries remotes:

- Google Translator Toolkit (http://translate.google.com/toolkit)
- WordFast Anywhere (http://www.freetm.com/)

Aquestes dues eines, tot i no ser de programari lliure, són d'ús gratuït. Totes dues funcionen directament des del navegador d'Internet i són una molt bona opció per treballar amb una eina de traducció assistida sense haver d'instal·lar res al nostre ordinador.

Serveis de memòries compartides públiques:

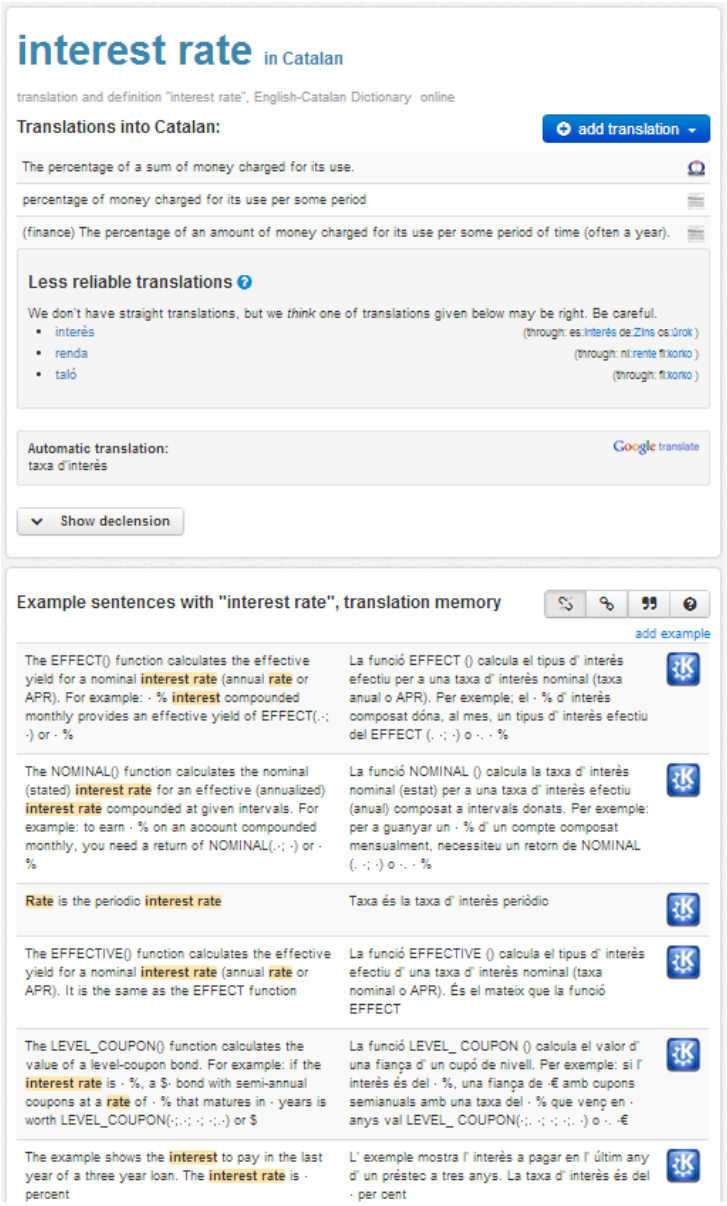
- **MyMemory TM** (http://mymemory.translated.net/): aquest recurs s'ha creat a partir de les memòries de traducció de la Unió Europea i de les Nacions Unides, així com alineant diversos llocs web multilingües. També permet pujar memòries de traducció pròpies. El sistema es pot consultar automàticament des d'altres aplicacions i també permet descarregar memòries de traducció a partir de documents a traduir. També es poden fer consultes directament a la seva interfície web. Aquesta funcionalitat pot resultar d'utilitat per cercar equivalents de traducció de termes (vegeu la següent figura). El sistema també proporciona una traducció automàtica a partir d'uns sistema de traducció automàtica estadística.


MyMemory
Translated.net
0 contribution(s)
Home | Professional Translation Service | Translation API | About MyMemory | Log in
interest rate
Language pair: English Catalan Subject All Search
All My memories Ask Google
You searched for: interest rate [Turn off colors]
API call | Download a TMX | Contribute a TMX
Computer translation
Trying to learn how to translate from the human translation examples.
English
Catalan
Info
interest rate
taxa d'interès
From: Machine Translation
Suggest a better translation
Quality: Be the first to vote
Human contributions
From professional translators, enterprises, web pages and freely available translation repositories.
Add a translation
English
Catalan
Info
Interest
Interès
Last Update: 2014-03-21
Usage Frequency: 16
Quality: Excellent
Reference: Wikipedia
Interest
Interessos
Last Update: 2010-10-23
Usage Frequency: 1
Quality: Be the first to vote
Reference: Wikipedia
The effective interest rate
La taxa d'interès efectiu
Last Update: 2009-01-01
Subject: Computer Science
Usage Frequency: 1
Quality: Be the first to vote
Reference: Translated.net
The effective interest rate
Taxa d'interès efectiu
Last Update: 2009-01-01
Subject: Computer Science
Usage Frequency: 1
Quality: Be the first to vote
Reference: Translated.net
rate
apreciar
Last Update: 2009-07-01
Subject: General
Usage Frequency: 1
Quality: Be the first to vote

- **TDA Translation Repository** (http://www.tausdata.org/index.php/taus-search): la TAUS Data Association (TDA) ofereix una interfície de cerca pública a un gran corpus de traduccions. Algunes aplicacions (com per exemple les eines d'Okapi) ofereixen la consulta automàtica a aquest recurs.

- **Linguee** (http://www.linguee.com/): Combina un diccionari amb un repositori de memòries de traducció. En el moment d'escriure aquest capítol no disposava d'una API per a consulta automàtica des d'una altre eina.

- **Glosbe** (http://glosbe.com/): Una eina similar a les anteriors però que intenta incloure el major nombre de llengües possible. A més de la cerca a memòries proporciona una traducció automàtica feta amb Google Translate. També permet que l'usuari aporti noves traduccions.



Algunes eines, com per exemple MemoQ (http://kilgray.com/products/memoq) permeten tres tipus de memòries de traducció:

- memòries de traducció locals
- memòries de traducció remotes
- memòries de traducció remotes sincronitzades: que són un híbrid de les dues anteriors. Es tracta de memòries de traducció remotes però que es descarreguen i sincronitzen amb una còpia local, de manera que si en algun moment no disposem de connexió a Internet puguem continuar treballant amb el projecte. En el moment que recuperem de nou la connexió a Internet la còpia local i remota es tornaran a sincronitzar.

2.8. Treball amb memòries de traducció

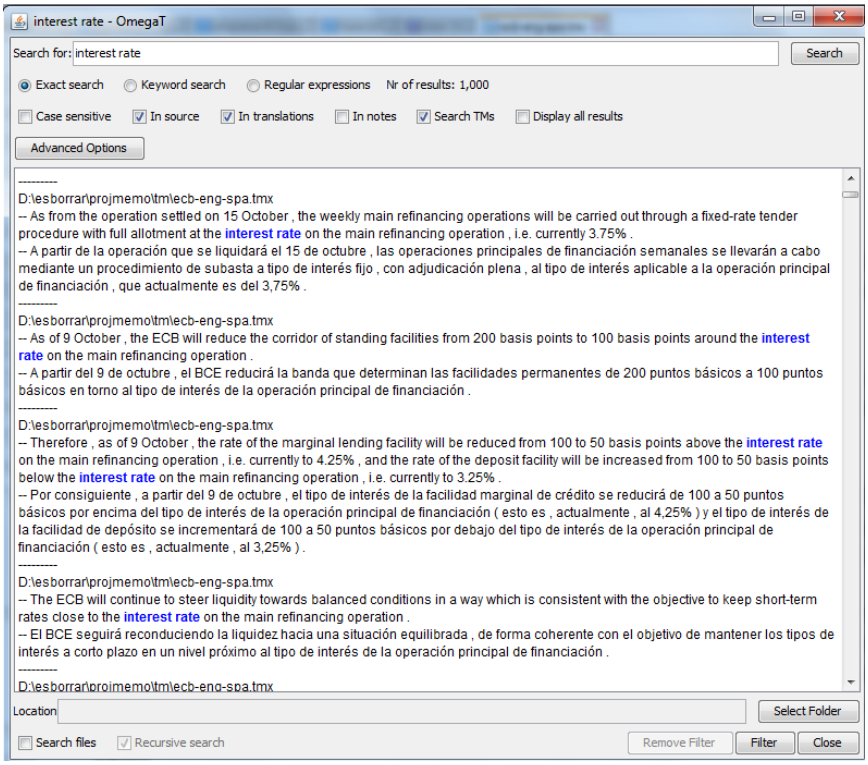
Bowker (2002) distingeix dues maneres de treballar amb memòries de traducció dins d'una eina de traducció assistida:

- Mode interactiu (interactive mode): A mesura que el traductor va treballant amb l'eina de traducció assistida, cada cop que canvia d'un segment a un altre el sistema cerca coincidències dins de la memòria de traducció i mostra les coincidències exactes o les que tinguin un índex de similitud superior o igual a l'indicat per l'usuari.
- Mode per lots (*batch mode*): Aquest mode sovint s'anomena **pretraducció** i consisteix a consultar a la memòria de traducció tots els segments del projecte i posar en la traducció el segment més semblant que superi un índex mínim de similitud indicat per l'usuari.

En tots dos casos el traductor haurà de revisar les propostes donades per la memòria de traducció. La pre-traducció és útil en els casos que vulguem enviar un projecte de traducció a un col·laborador sense haver d'enviar tota una memòria de traducció.

Un punt important a tenir en compte quan es treballa amb memòries de traducció és l'establiment de la similitud mínima per recuperar segments de la memòria. Si treballem amb similituds molt altes, per exemple del 95%, serà molt difícil trobar coincidències a la memòria i el programa mostrarà molt poques suggerències. En canvi, si treballem amb similituds molt baixes, de per exemple el 10%, el sistema ens mostrarà moltes suggerències, però probablement seran de poca utilitat. Un bon compromís pot ser establir la similitud mínima entre el 65 i el 85%.

Cal tenir en compte, però, que tot i que la nostra memòria de traducció no contingui segments semblants al que estem traduint, la majoria d'eines de traducció permeten la cerca de fragments de text (típicament unitats termiològiques) dins de la memòria de traducció, de manera que ens mostri tots els segments originals de la memòria de traducció que contenen aquest fragment i les corresponents traduccions. En la següent imatge podem veure la funció *Search project* d'OmegaT que permet fer cerques als projectes i les memòries de traducció:



És important no confondre el concepte de pretraducció amb el concepte de **pseudotraducció**. La pseudotraducció consisteix a fer una primera traducció del projecte simulant una llengua d'arribada fictícia: pot composar-se de caràcters aleatoris, de canvis de certs caràcters de l'original, etc. L'objectiu és verificar si el procés d'importació dels fitxers originals i de la creació dels fitxers traduïts finals funciona correctament. A continuació podem veure un exemple de cadena pseudotraduïda (exemple extret de http://en.wikipedia.org/wiki/Pseudolocalization).

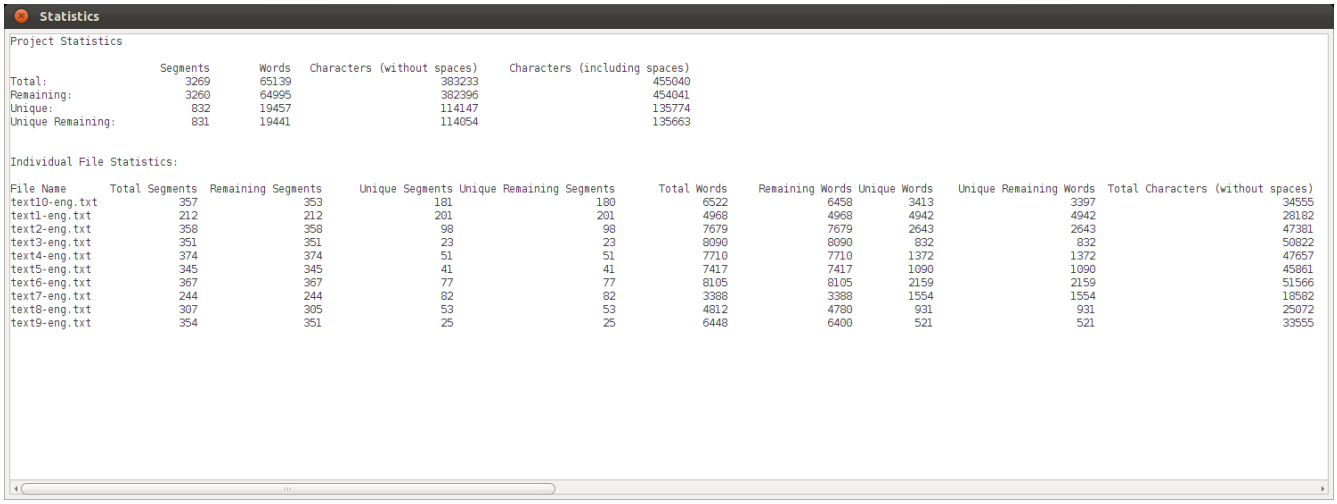
Account Settings	[!!! Àççôûñţ Šéţţîñĝš !!!]

2.9. Anàlisi de projectes i tarifació

Abans de començar a treballar amb un projecte de traducció és molt important conèixer en detall la feina que portarà. Quan es treballa sense l'ajut de programes de traducció assistida habitualment es compten paraules o caràcters dels fitxers a traduir. Aquests comptatges serveixen també per pressupostar o facturar al nostre client.

Quan treballem amb sistemes de traducció assistida cal tenir en compte també les repeticions internes del projecte i els segments que es recuperaran de la memòria de traducció. Aquesta informació convé tenir-la també per diferents marges de similitud, ja que no és el mateix una repetició exacta que una al 75%. La majoria de sistemes de traducció assistida compten amb funcions d'anàlisi de projectes.

A la següent imatge podem observar una anàlisi d'un projecte realitzada amb l'opció *Project statistics* d'OmegaT:



Statistics

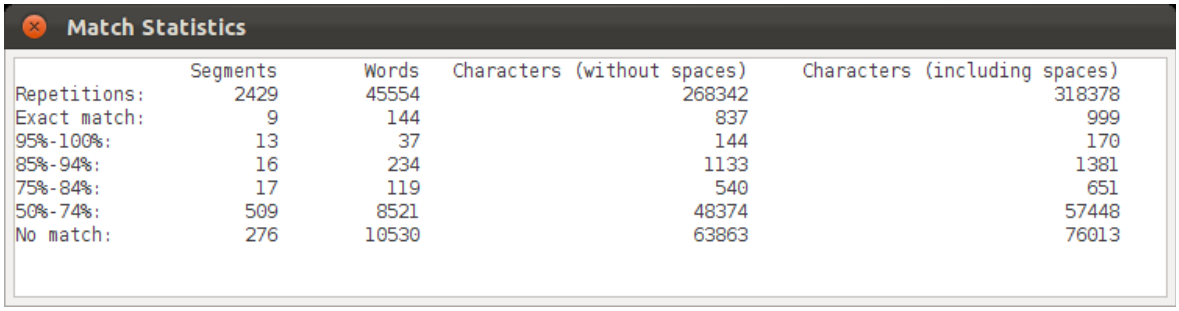
Project Statistics

	Segments	Words	Characters (without spaces)	Characters (including spaces)
Total:	3269	65139	383233	455040
Remaining:	3260	64995	382396	454041
Unique:	832	19457	114147	135774
Unique Remaining:	831	19441	114054	135663

Individual File Statistics:

File Name	Total Segments	Remaining Segments	Unique Segments	Unique Remaining Segments	Total Words	Remaining Words	Unique Words	Unique Remaining Words	Total Characters (without spaces)
text10-eng.txt	357	353	181	180	6522	6458	3413	3397	34555
text1-eng.txt	212	212	201	201	4968	4968	4942	4942	28182
text2-eng.txt	358	358	98	98	7679	7679	2643	2643	47381
text3-eng.txt	351	351	23	23	8090	8090	832	832	50822
text4-eng.txt	374	374	51	51	7710	7710	1372	1372	47657
text5-eng.txt	345	345	41	41	7417	7417	1090	1090	45861
text6-eng.txt	367	367	77	77	8105	8105	2159	2159	51566
text7-eng.txt	244	244	82	82	3388	3388	1554	1554	18582
text8-eng.txt	307	305	53	53	4812	4780	931	931	25072
text9-eng.txt	354	351	25	25	6448	6400	521	521	33555

Si volem també disposar de les estadístiques de coincidències amb les memòries de traducció farem servir l'opció Tools > Match Statistics, que ens oferirà la següent informació:



Match Statistics

	Segments	Words	Characters (without spaces)	Characters (including spaces)
Repetitions:	2429	45554	268342	318378
Exact match:	9	144	837	999
95%-100%:	13	37	144	170
85%-94%:	16	234	1133	1381
75%-84%:	17	119	540	651
50%-74%:	509	8521	48374	57448
No match:	276	10530	63863	76013

Aquesta informació es pot fer servir per a pressupostar o facturar projectes de traducció: es poden cobrar tarifes diferents per a les paraules corresponents a segments nous que cal traduir des de zero, altres tarifes per les coincidències exactes provinents de memòries de traducció o de repeticions internes, i tarifes diferents segons els graus de similitud. El que no és recomanable és no cobrar res per les coincidències exactes, ja que

de fet porten també una feina de verificació de si la traducció proposada és la més escaient en el nou context.

2.8. Noves funcionalitats

En aquest apartat explicarem algunes de les funcionalitats interessants que se estan desenvolupant dins del projecte CASMACAT (Alabau, 2013) i que de ben segur en un futur proper veurem implementades en els sistemes habituals del mercat.

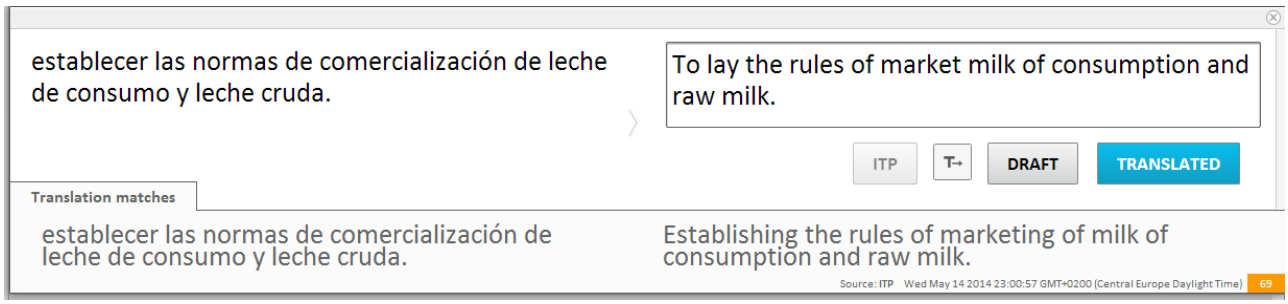
2.8.1. Sistema de traducció automàtica integrat

Molts sistemes de traducció assistida disposen d'alguna mena de connexió a sistemes de traducció automàtica externs que el permeten recuperar una traducció automàtica del segment que s'està traduint en aquell moment. La connexió no va més enllà d'aquesta recuperació, i el sistema de traducció automàtica per regla general no sap què fa l'usuari amb la traducció ni es capaç de donar-li altres traduccions alternatives.

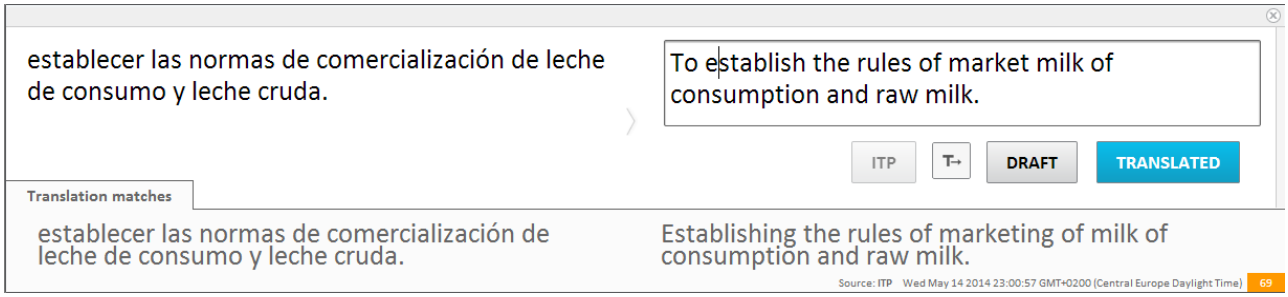
CASMACAT integra un sistema de traducció automàtica estadística. Aquesta integració implica per una banda que pot proporcionar a l'usuari més d'una proposta de traducció, i per altra banda, que el sistema de traducció automàtica pot aprendre sabent si una proposta ha estat acceptada íntegrament o bé ha estat modificada. Aquesta integració també permet la funcionalitat que presentem a continuació, l'*autocompletat intel·ligent*.

2.8.2. Autocompletat intel·ligent

Molts processadors de textos disposen de la funció d'autocompletat que tenen en compte les paraules més probables que poden seguir a les paraules que estem escrivint i que mostra un cop escrivim unes poques lletres. Els sistemes de traducció assistida poden disposar també de la mateixa funcionalitat, però a més poden tenir més evidències del que volem escriure ja que saben l'original que estem traduint i disposen d'evidències provinents de la memòria de traducció i del sistema de traducció automàtica.



Si comencem a escriure "es…" just darrera de "To" el sistema autocompletarà amb la continuació més probable donada tant per les memòries de traducció com del sistema de traducció automàtica estadística.



2.8.3. Mesures de confiança

Aquestes mesures implementades en CASMACAT informen l'usuari sobre quines parts de la traducció tenen més probabilitat de ser incorrectes. Per una banda es marquen en vermell les paraules que tenen molta probabilitat de ser incorrectes i per l'altra banda es marquen en taronja les paraules dubtoses per al sistema.

2.9. Conclusions

En aquest capítol hem presentat amb detall un dels recursos principals de les eines de traducció assistida: les memòries de traducció. S'ha analitzat el procés de recuperació de segments similars i el càlcul de la similitud entre el segment que estem traduint i els recuperats de la memòria.

Les eines de traducció assistida, per regla general, treballen amb poc coneixement lingüístic específic d'una determinada llengua. L'objectiu és que l'eina pugui funcionar per a un gran ventall de llengües. Hem vist com la inclusió d'informació especifica i la capacitat de fer una anàlisi lingüística (procés que és depenent de la llengua) pot millorar l'ús de les memòries de traducció. Aquesta millora es pot obtenir tant en el procés de recuperació de segments similars, com en la combinació d'unitats subsegmentals.

Les memòries de traducció són un recurs que es pot combinar amb els sistemes de traducció automàtica (que veurem a fons en el capítol 4). Per ara, la majoria d'eines limiten aquesta combinació a fer una proposta provinent d'un sistema de traducció automàtica si no es troba cap segment similar a la memòria de traducció. Hem vist també en aquest capítol com aquesta combinació pot anar molt més enllà i pot passar per la retroalimentació dels sistemes de traducció automàtica amb els nous segments de la memòria, el suport del sistema de traducció automàtica per a la combinació d'unitats subsegmentals i fins a la possibilitat de fer un autocompletat intel·ligent.

Avui dia les memòries de traducció són un recurs de gran utilitat per al traductor i és previsible que en un futur molt proper ho siguin encara més.

No tots els textos són igualment adequats per al ús de memòries de traducció. Tot i així, l'ús de memòries de traducció pot ser d'utilitat fins i tot per a traducció de textos on pràcticament no hi ha repeticions, com pot ser la traducció literària. Tot i que no és previsible que el sistema ens proporcioni pràcticament cap coincidència, podrem fer cerques d'expressions i veure com han estat traduïdes amb anterioritat.

2.10. Per ampliar coneixements

2.10.1. Corpus paral·lels i memòries de traducció disponibles públicament

Quan comencem a treballar amb sistemes de traducció assistida i no disposem de cap memòria de traducció el sistema només ens podrà oferir propostes de les anomenades *repeticions internes*, és a dir, segments similars que hem traduït anteriorment en el projecte de traducció. Aquesta situació és transitòria, ja que quan portem uns mesos treballant amb el nostre sistema de traducció assistida ja disposarem d'un bon volum de segments originals i traduïts a les nostres memòries.

A Internet es poden trobar una bona quantitat de corpus paral·lels o memòries de traducció disponibles per a la descàrrega. Si alguna d'aquestes memòries són del nostre àmbit de treball les podem incorporar al sistema de traducció assistida. La majoria d'aquestes memòries provenen de traduccions oficials d'institucions multinlingües com la Unió Europea, o bé de projectes de localització de programes de software libre.

Un bon repositori per començar a cercar és OPUS (http://opus.lingfil.uu.se/) (Tiedemann 2012). En el moment d'escriure aquest capítol aquesta col·lecció estava composada pels següents corpus:

- ECB - European Central Bank corpus
- EMEA - European Medicines Agency documents
- The EU bookshop corpus
- EUconst - The European constitution
- EUROPARL v7 - European Parliament Proceedings
- EUROPARL - European Parliament Proceedings
- The Croatian - English WaC corpus
- KDE4 - KDE4 localization files (v.2)
- KDEdoc - the KDE manual corpus
- MBS - Belgisch Staatsblad corpus
- MultiUN - Translated UN documents
- OO - the OpenOffice.org corpus
- OfisPublik - Breton - French parallel texts
- OpenOffice.org 3 corpus
- OpenSubtitles - the opensubtitles.org corpus
- OpenSubtitles2011 - opensubtitles.org 2011
- OpenSubtitles2012 - opensubtitles.org 2012
- OpenSubtitles2013 - opensubtitles.org 2013
- PHP - the PHP manual corpus
- Regeringsförklaringen - a tiny example corpus
- SETIMES - A parallel corpus of the Balkan languages
- SETIMES2 - A new version of SETIMES
- SPC - Stockholm Parallel Corpora
- Tatoeba - A DB of translated sentences
- TedTalks hr-en
- TEP - The Tehran English-Persian subtitle corpus
- UN - Translated UN documents
- WikiSource (in progress)

Recentment també s'ha incorporat a aquesta col·lecció un corpus paral·lel català-castellà provinent dels textos del Diari Oficial de la Generalitat de Catalunya (DOGC).

Una altra molt bona font per trobar corpus és consultar Meta-share (http://metashare.upf.edu/) .

2.10.2. Etiquetadors morfosintàctics

Un *etiquetador morfosintàctic* és un programa informàtic capaç d'etiquetar totes les paraules d'un text amb informació morfosintàctica. Aquesta informació es dóna mitjançant unes etiquetes que expressen la categoria gramatical i una sèrie d'informació addicional. A la següent figura es pot observar l'etiquetatge morfosintàctic de l'oració castellana *Yo bajo con el hombre bajo a tocar el bajo bajo la escalera* duta a terme per l'analitzador Freeling (padró, 2012).

Yo	bajo	con	el	hombre	bajo	a	tocar	el	bajo	bajo	la	escalera	.
yo	*bajar*	*con*	*el*	*hombre*	*bajo*	*a*	*tocar*	*el*	*bajo*	*bajo*	*el*	*escalera*	*.*
PP1CSN00	VMIP1S0	SPS00	DA0MS0	NCMS000	AQ0MS0	SPS00	VMN0000	DA0MS0	NCMS000	SPS00	DA0FS0	NCFS000	Fp

Fixem-nos que l'analitzador és capaç, almenys en alguns casos, d'etiquetar correctament les paraules tot i que aquestes sigui ambigües. La paraula *bajo* en aquesta oració pot ser un verb (VMIP1S0) un adjectiu (AQ0MS0), un substantiu (NCMS000) i una preposició (SPS00). La categoria gramatical i les corresponents subcategoritzacions s'expressen mitjançant etiquetes.

Freeling (http://nlp.lsi.upc.edu/freeling) és un analitzador lingüístic de codi lliure desenvolupat a la Universitat Politècnica de Catalunya. Pot fer anàlisi a diversos nivells:

- Anàlisi morfològica (*morphological analysis*): aquesta anàlisi no fa desambigüació i posa tota la informació possible a cada paraula, sigui la correcta pel context o no.
- Anàlisi morfosintàctica (*PoS tagging*): posa a cada paraula el lema i l'etiqueta que li correspon
- Anàlisi sintàctica superficial (*shallow parsing*): fa una anàlisi sintàctica del text en la que algunes relacions poden no ser-hi.
- Anàlisi sintàctica completa (*full parsing*): fa l'anàlisi sintàctica completa
- Anàlisi de dependències (*dependency parsing*): un tipus d'anàlisi que marca les dependències entre les paraules
- Anàlisi semàntica: etiqueta els textos amb *synsets* de WordNet i és capaç de fer desambiguació de sentits

Freeling funciona per moltes llengües tot i que no totes les llengües disposen de tots els nivells d'anàlisi descrits:

- asturià
- català
- anglès
- francès
- gallec
- portuguès
- castellà
- rus
- gal·lès

Freeling es pot instal·lar tant en Linux com en Windows i també disposa d'una demo on-line que permet fer-se una idea de les seves capacitats.

Un altre etiquetador morfosintàctic disponible és el Tree Tagger (Schmid, 1994) (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) que no té una llicència lliure però que pot fer-se servir per a finalitats de recerca, avaluació i educació. Té també versions per Linux i Windows i disposa de models per a moltes llengües. A diferència de Freeling, Tree Tagger només ofereix l'anàlisi morfosintàctica i per a certes llengües també *chunking* (anàlisi fragmental).

2.9.3. Propietat de les memòries de traducció

Les memòries de traducció plantegen un problema jurídic encara no resolt del tot respecte a la seva propietat intel·lectual. La memòria es composa d'un original (que pot tenir els seus propis drets d'autor) i una traducció (que genera uns drets de traducció). La traducció acostuma a estar encarregada per un client i moltes vegades a través d'una agència. Entremig s'ha generat una traducció que pot estar en mans del traductor, agència i client final. Qui té dret de tornar a fer servir aquesta memòria? El traductor? Però només per al mateix client final? Per tots els clients? Qui té dret a cedir aquesta memòria a altres usuaris? Totes les legislacions opinen el mateix, o hi ha diferències entre els països?

En aquest apartat no aclarirem aquests dubtes, però sí que proposarem algunes lectures que poden donar una mica de llum sobre el tema.

Jorge Marcos (2001) *Un enfocament jurídic de les memòries de traducció*. Revista Tradumàtica núm. 0. http://www.fti.uab.es/tradumatica/revista/num0/articles/jmarcos/art.htm

La presentació *Copyright protection for translation memories* que es pot trobar al següent enllaç: http://www.fit-europe.org/vault/barcelone/Byrne.pdf

Ross Smith (2009) *Copyright Issues in Translation Memory Ownership* Proceedings of the Thirty-first International Conference on Translating and the Computer 19-20 November 2009, London (http://www.mt-archive.info/Aslib-2009-Smith.pdf)

Bibliografia

Alabau, Vicent, Ragnar Bonk, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Jesús González et al. *CASMACAT: An Open Source Workbench for Advanced Computer Aided Translation.* The Prague Bulletin of Mathematical Linguistics 100, no. 1 (2013): 101-112.

Barrachina S., Bender O., Casacubierta F., Civera J., Cubel E., Khadivi S., Lagarda A., Ney H., Tomás J., Vidal E. and Vilar J.M. (2009) *Statistical approaches to computer-assisted translation.* Computational Linguistics, 35 (1), 3-28.

Biçici, E. and Dymetman, M. (2008) *Dynamic Translation Memory: Using Statistical Machine Translation to improve Translation Memory Fuzzy Matches* Proceedings of the 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2008), Feb , 2008, Haifa, Israel

Bowker, L. (2002). *Computer-Aided Translation Technology. A Practical Introduction.* Ottawa: University of Ottawa Press.

Brown P.F., Lai J.C., Mercer R.L. (1993) *Aligning sentences in parallel corpora*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics. Berkeley. California. pp. 177-184

Chen S.F. (1993) *Aligning Sentences in Bilingual Corpora Using Lexical Information.* In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. Columbus. Ohia. pp. 9-16

Colominas C. (2008) *Towards chunk-based translation memories*. Babel %4:4, pp. 343-354

Cranias, L., H. Papageorgiou and S. Piperidis (1997) *Example Retrieval from a Translation Memory*, Natural Language Engineering 3:255–277.

Dennett, G (1995) *Translation memory: Concept, products, impact and prospects*. MSc dissertation, School of Electrical, Electronic and Information Engineering, South Bank University, London.

Frakes, W.B. and R. Baeza-Yates (Eds.) (1992) *Information Retrieval - Data Structures & Algorithms.* New Jersey. Prentince Hall PTR.

Gale W.A. and Church K.W. (1993) *A Program for Aligning Sentences in Bilingual Corpora*. Computational Linguistics 19 (1). pp. 75-102

Kay M., Röscheisen M. (1993) *Text-Trasnlation Alignment*. Computational Linguistics 19 (1) pp. 121-142

Macklovitch, E./Russell, G. (2000) *What's been forgotten in translation memory*. In: White, J. S. (ed.) Envisioning Machine Translation in the Information Future: 4th Conference of the 750 Association for Machine Translation in the Americas, AMTA 2000, Cuernavaca, Mexico, Berlin: Springer, 137–146.

Melamed I.D. *A portable Algorithm for Mapping Bitext Correspondence.* In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics. Madrid. Spain. pp. 305-312

Lluís Padró and Evgeny Stanilovsky (2012) *FreeLing 3.0: Towards Wider Multilinguality* Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May, 2012.

Planas, E./Furuse, O. (1999) *Formalizing translation memories*. In: Machine Translaiton Summit VII, Singapore, 331-330; repr. in Carl & Way 2003, 157-188.

Porter, M. (1980) *An algorithm for suffix stripping*. Program 14.3 (1980): 130-137

Rapp, R. 2002. *A Part-of-Speech-Based Search Algorithm for Translation Memories.* in LREC 2002, Third International Conference

Simards, M. and /Langlais, P. (2001) *Sub-sentential Exploitation of Translation Memories.* Proceedings of Machine Translation Summit VIII, 335-9. Santiago de Compostela.

Simões A., Gómez-Guinovart X., João J. (2004) *Distributed Translation Memories implementation using WebServices*. Procesamiento del Lenguaje Natural, 38, pp. 89-94.

Schmid, H. (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees*. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Somers, H. (ed.) (2003) *Computers and translation: a translator's guide*. John Benjamins Publishing Company. Philadelphia, PA, USA. ISBN 9789027296696

Somers, H./Fernández Díaz, G. (2004) *Translation memory vs. example-based MT: What is the difference?* In: International Journal of Translation 16(2), 5–33; based on: Diferencias e interconexiones existentes entre los sistemas de memorias de traducción y la EBMT. In:
815 Corpas Pastor, G. & Varela Salinas, M.a-J. (eds) Entornos informáticos de la traducción profesional: las memorias de traducción, Granada (2003): Editorial Atrio, pp. 167–192.

Tiedemann J. (2012) *Parallel Data, Tools and Interfaces in OPUS.* In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). *Parallel corpora for medium density languages.* In Proceedings of the RANLP 2005, pages 590-596.

Wu D. (1994). *Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria.* In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics. Las Cruces, New Mexico. pp. 80-87