

2. Las memorias de traducción

Índice

2.1. Introducción....1
2.2. Indexación y recuperación de segmentos....3
2.2.1. Indexación de memorias de traducción....3
2.2.2.Cálculo de la similitud de segmentos....10
2.3. Coincidencia exacta y parcial....16
2.4.Combinación de unidades subsegmentales....17
2.5. Formato de intercambio de memorias de traducción: TMX....18
2.6.Creación de memorias de traducción....21
2.6.a. Traducción con sistemas de traducción asistida....21
2.6.b. Memorias de traducción y corpus paralelos....21
2.6.c. Alineación manual de documentos....21
2.6.d. Alineación automática de documentos....23
2.7. Memorias de traducción remotas compartidas y públicas....25
2.8. Trabajo con memorias de traducción....31
2.9. Análisis de proyectos y tarificación....33
2.8. Nuevas funcionalidades....35
2.8.1.Sistema de traducción automática integrado....35
2.8.2.Autocompletado inteligente....35
2.8.3. Medidas de confianza....36
2.9. Conclusiones....37
2.10. Para ampliar conocimientos....38
2.10.1. Corpus paralelos y memorias de traducción disponibles públicamente....38
2.10.2. Etiquetadores morfosintácticos....39
2.9.3. Propiedad de las memorias de traducción....40
Bibliografía....41

2.1. Introducción

En este capítulo hablaremos en detalle de las memorias de traducción, el principal recurso en que se basan los sistemas de traducción asistida por ordenador.

Una memoria de traducción es un repositorio de segmentos de texto en una determinada lengua con las traducciones a una o más lenguas.

De esta manera tenemos una relación directa entre un segmento de texto en una lengua y su traducción a otra lengua. Estos segmentos de texto suelen ser oraciones, aunque no siempre lo son desde el punto de vista gramatical. Por este motivo se habla de *segmentos* y no de oraciones.

En las memorias de traducción no se relacionan unidades más grandes, como por ejemplo párrafos, ya que la probabilidad de encontrar párrafos iguales o similares en dos textos es muy baja. Tampoco se hacen

relaciones entre unidades muy pequeñas, como por ejemplo, palabras o sintagmas, ya que el traductor humano no trabaja tratando aisladamente estas unidades[1].

La función principal de las memorias de traducción es ofrecer al traductor sugerencias de traducción del segmento que está traduciendo. Esta sugerencia puede provenir de una *coincidencia exacta (exact match)* cuando el segmento en la lengua de partida que hay en la memoria es exactamente igual al que se está traduciendo; o bien de una coincidencia parcial (*fuzzy match*), cuando el segmento en la lengua de partida no es exactamente igual al que se está traduciendo. El índice de similitud mínimo para que aparezcan sugerencias es totalmente configurable. Si se escogen índices muy altos, por ejemplo 99%, aparecerán muy pocas propuestas. Si se opta por un índice muy bajo, por ejemplo, el 60%, aparecerán muchas más propuestas, pero aparecerán muchas que no serán útiles. Hay que tener en cuenta que si se acepta una coincidencia parcial, se tendrá que llevar a cabo alguna edición (cambiar alguna palabra, por ejemplo). Si el índice de similitud es muy bajo, el esfuerzo de edición será muy alto y probablemente valga más la pena traducir el segmento manualmente desde cero. Un buen compromiso puede ser configurar la similitud mínima entre el 65 y el 85%. Cuando existen más de una coincidencia parcial, el sistema de las muestra ordenadas de más similitud a menos similitud.

Esta función de recuperación de segmentos similares de una base de datos que llamamos memoria de traducción presenta dos retos importantes:

- Encontrar de forma rápida los segmentos más similares.
- Utilizar una medida de similitud para ordenar los segmentos recuperados de forma que se presenten en primer lugar los más similares. Esta medida debe ser indicativa del grado de dificultad y del tiempo necesario para editar la traducción del segmento recuperado y dejarla como traducción del segmento original que estamos traduciendo. Todo ello teniendo en cuenta que sólo podemos comparar los segmentos en la lengua de partida ya que la traducción del segmento que buscamos todavía no la tenemos.

Los sistemas de traducción asistida también permiten realizar búsquedas de unidades más pequeñas (por ejemplo palabras o términos) dentro de las memorias activas. De esta manera podemos buscar si se ha traducido anteriormente un determinado término. No hay que confundir esta funcionalidad con la búsqueda automática en bases de datos terminológicas. En este segundo caso tenemos términos y sus traducciones en una base de datos. En el caso de utilizar memorias de traducción lo que tenemos son oraciones originales y traducidas que pueden contener el término. Si hacemos una búsqueda el sistema nos mostrará las oraciones originales que contienen el término y las traducciones de estas oraciones (donde se supone que el usuario podrá encontrar la traducción del término[2].

1 Esto no quiere decir que no existan herramientas de traducción asistida que intente combinar unidades más pequeñas provenientes de diversos segmentos que se encuentren en la memoria para intentar formar una propuesta válida.

2 Algunos sistemas van más allá e intentan también inferir la traducción del término original a partir de los segmentos traducidos.

2.2. Indexación y recuperación de segmentos

Hemos definido las memorias de traducción como un repositorio de segmentos de texto en más de una lengua. Para acceder de manera eficiente a este repositorio éste debe estar contenido en una base de datos y se ha tenido que llevar a cabo algún tipo de indexación de los datos. Esta indexación es importante, ya que la búsqueda en la memoria se debe realizar en un tiempo muy corto, desde el momento en que el usuario introduce un segmento hasta el momento en que aparece el siguiente segmento a traducir. La sensación para el usuario debe ser que el paso de un segmento a otro es inmediato.

Imaginemos que tenemos una memoria de traducción de un tamaño medio, por ejemplo 10.000 segmentos. La búsqueda no puede ser secuencial, es decir, mirar si tenemos un segmento igual o similar al que estamos traduciendo comenzando por el primero y comparando uno a uno los segmentos. La búsqueda de segmentos iguales puede llegar a ser muy rápida, pero la de segmentos semejantes es mucho más lenta. Más adelante dedicamos un subapartado al tema del cálculo de la similitud entre segmentos. Si en vez de 10.000 segmentos la memoria fuera de 100.000 una búsqueda secuencial tardaría también mucho más. Como nos interesa trabajar con memorias muy grandes para aumentar la probabilidad de encontrar segmentos interesantes, es imprescindible encontrar un mecanismo de indexación y recuperación eficientes para evitar una respuesta demasiado lenta del sistema de traducción asistida.

En el resto de este apartado explicaremos técnicas genéricas ya que cada herramienta incorpora variaciones en la indexación y en el cálculo de similitudes.

2.2.1. Indexación de memorias de traducción

La indexación de una memoria de traducción consiste en realizar un índice inverso de las palabras (o fragmentos de palabras, o al menos de algunas palabras) que aparecen en la memoria de traducción. El índice nos da el identificador de todos los segmentos en los que aparece una determinada palabra (o fragmento de palabra).

Imaginemos que tenemos una memoria de traducción con los siguientes segmentos:

ID	**Segmento original**	**Segmento traducido**
1	Search the Legal framework	Buscar en Marco jurídico
2	Legal framework of the ESCB	Régimen jurídico del SEBC
3	ECB institutional provisiones	Disposiciones institucionales del BCE
4	Monetary policy and Operations	Política monetaria y operaciones
5	Payment and settlement systems	Sistemas de pago y Liquidación
6	Banknotes and coins, means of payment and currency matters	Billetes de banco y monedas, mitjans de pago y Cuestiones de moneda
7	Foreign exchange and Foreign reservas	Divisas y reservas exteriores
8	The approach was successful: in volume terms, close to 80% of the initial Banknote demand and 97% of the	Resultando un Planteamiento acertada ya que, en tÃ © rminos de volumen, casi el 80% de la demanda inicial de

	total coin needs for the changeover had been distributed before 1 January 2002.	palanquillas y el 97% de las monedas necesarias para la introducción del euro habian sidos distribuidos tas del 1 de enero de 2002.
9	Employment, conduct, fraud prevention and transparency	Contratación, conducta, Prevención del fraude y transparencia
10	Financial market stability	Estabilidad de los Mercados Financieros

El índice inverso nos indicaría el identificador de todos los segmentos donde aparece una determinada palabra (en nuestro ejemplo no se indexan los números). Para este ejemplo tendría el siguiente aspecto.

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

Esta tabla nos proporciona información sobre en qué segmentos aparecen cada una de las palabras. Por ejemplo, *payment* aparece en los segmentos 5 y 6 y *market* en el 10.

Imaginemos que queremos encontrar segmentos parecidos a este:

Banknotes, coins, and types of payment

Tomaríamos los índices de los segmentos y el índice que apareciera más veces sería probablemente el más parecido, ya que contendría más palabras comunes. Dependiendo del algoritmo de cálculo de similitud entre segmentos, el orden de las palabras nos puede jugar malas pasadas, así que a menudo se toma no sólo el más parecido, si no los primeros más parecidos y se calcularía la similitud, hasta que ésta estuviera por debajo de la similitud mínima dada por el usuario. Sobre el cálculo de similitud, hablaremos en el siguiente subapartado. Según esto, quedaría:

```
Banknotes 6
coins 6
types
payment 5: 6
```

y por tanto los segmentos más similares sería el 6 (*Banknotes and coins, means of payment and currency matters*), ya que tiene 3 palabras coincidentes. El segundo segmento más parecido sería el 5 (*Payment and settlement systems*).

Fijémonos en un par de aspectos del índice inverso de esta memoria de traducción. Las palabras muy cortas tienden a ser palabras funcionales que aparecen en muchos segmentos (fijémonos en *and*, que aparece en 6

segmentos y *the*, que aparece en 3 segmentos, en el ejemplo).A menudo, no se tienen en cuenta las palabras muy cortas (por ejemplo de menos de 2 ó 3 caracteres) en los índices inversos[3].Las palabras muy largas suelen ser palabras correspondientes a categorías abiertas y en muchas lenguas éstas están sometidas a flexión. Fijémonos por ejemplo en *Banknote* y *Banknotes*, que son en realidad la misma palabra flexionada. En el caso de palabras largas menudo se toman los primeros *n* caracteres, por ejemplo los 5 o 6 primeros. Teniendo en cuenta estos aspectos, el índice inverso nos quedaría de la siguiente manera (fijémonos que para las palabras de más de 6 caracteres se han indexado sólo los 6 primeros):

approa 8	financ 10	policy 4
bankno 6: 8	foreig 7: 7	preven 9
been 8	framew 1: 2	pruebes 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	suce 8
conduc 9	market 10	system 5
curran 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	operado 4	
exchan 7	Payment 5: 6	

El hecho de indexar sólo los 5 primeros caracteres de las palabras de más de 6 caracteres nos ha permitido indexar con el mismo índice las formas *Banknote* y *Banknotes* (bajo el índice *bankno);* no ha permitido juntar las formas *coin* y *coins.* Hay que tener en cuenta que en general se buscan formas de indexación generales, que sirvan para muchas lenguas y que no incorporen conocimiento lingüístico sobre una determinada lengua.

Ahora, para buscar los segmentos más parecidas a

Banknotes, coins, and types of payment

seguiríamos la misma estrategia y también recortaríamos las palabras de más de 6 letras en sus 6 primeras letras.

```
bankno 6
coins 6
types
Payment 5: 6
```

y el resultado sería el mismo, es decir, que el segmento más parecido sería el 6 *(Banknotes and coins, means of payment and currency matters)* y el segundo más parecido el 5 *(Payment and settlement systems)*, es decir, exactamente al igual que en el caso anterior.

Recordemos que los desarrolladores de programas de traducción asistida intentan que sus herramientas funcionen por un gran número de lenguas y en general evitan utilizar métodos de indexación y recuperación que requieran de información lingüística. Si añadimos conocimiento lingüístico, la indexación de memorias de traducción puede mejorar notablemente.

Una primera estrategia puede ser el uso de la técnica conocida como *stemming* que consiste en eliminar los afijos morfológicos de las palabras. Este proceso se puede llevar a cabo utilizando varios algoritmos, como el de Porter (1980).

3 Esto es para lenguas que utilicen caracteres alfabéticos y no se puede aplicar a lenguas que usan ideogramas.

Vemos a continuación el resultado de utilizar el algoritmo de Porter en nuestro ejemplo:

ID	Segmento original	Segmento 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 provisiones	ECB instituto pruebes
4	Monetary policy and Operations	monetario poli and operation
5	Payment and settlement systems	payment and settlement system
6	Banknotes and coins, means of payment and currency matters	Banknotes and coin, mean of payment and concurrencia matter
7	Foreign exchange and Foreign reservas	foreign Exchanges 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 volumen term, close to Z% of the initio Banknotes demand and Z% of the total coin need for the changeov have bien distribuido befor Z januari Z.
9	Employment, conduct, fraud prevention and transparency	Employ, conduct, fraud preventiva and transparencia
10	Financial market stability	financie market Stabil

Los índices quedan de la siguiente manera (eliminamos también los *stems* de tres o menos letras):

```
approach 8
Banknotes 6: 8
befor 8
changeov 8
close 8
coin 6: 8
conduct 9
concurrencia 6
demand 8
distribuido 8
Employ 9
escb 2
Exchanges 7
financie 10
foreign 7: 7
framework 1: 2
fraud 9
have 8
initio 8
instituto 3
januari 8
legal 1: 2
market 10
matter 6
mean 6
monetario 4
need 8
operation 4
payment 5: 6
poli 4
preventiva 9
pruebes 3
reserv 7
search 1
settlement 5
Stabil 10
successful 8
system 5
term 8
total 8
transparencia 9
volumen 8
```

Para hacer ahora la búsqueda del segmento más parecido a:

Banknotes, coins, and types of payment

tenemos que usar el mismo *Stemmer* a la oración que buscamos, que quedaría:

Banknote, coin, and type of payment

y consultando los índices:


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

Resulta en que el segmento que recupera en primer lugar es el 6 *(Banknotes and coins, means of payment and currency matters)*, seguido 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).*Vemos que ahora el segmento 5 *(Payment and settlement systems)* que con las estrategias anteriores se recuperaba en segundo lugar ahora se recupera en tercer lugar.

Una segunda aproximación, que requiere aún de más información lingüística específica de la lengua de partida, consistiría en disponer de un lematizador para la lengua de partida (del inglés en este ejemplo). El lematizador es capaz de sustituir cada una de las palabras por su lema, es decir, su forma base. Así, la tabla de nuestro ejemplo quedaría de la siguiente manera (lematizamos sólo la lengua de partida, ya que es la que utilizamos para hacer las búsquedas). Fijémonos también que podemos sustituir las cifras por una etiqueta que nos indique simplemente que se trata de una cifra.

ID	**Segmento original**	**Segmento original lematizado**
1	Search the Legal framework	search the legal framework
2	Legal framework of the ECB	legal framework of the ECB
3	ECB institutional provisiones	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 reservas	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

Ahora los índices quedarían de la siguiente manera (eliminamos también las lemas de 3 o menos caracteres):

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

Para hacer ahora la búsqueda del segmento más parecido a:

Banknotes, coins, and types of payment

deberíamos lematizar también la oración, que quedaría:

Banknote, coin, and type of payment

y consultando los índices:

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

el segmento que se recuperaría en primer lugar sería el 6 (con 3 lemas coincidentes), el 8 (con 2) y el 5 (con 1); estos resultados son iguales que en el caso de utilizar el *Stemmer*.

Si disponemos de un analizador más potente para la lengua de partida podemos mejorar aún más los índices. Por ejemplo, si nuestro analizador es capaz de lematizar e indicar la categoría gramatical de cada palabra podemos añadir esta información a los índices y hacer después la búsqueda con este criterio. Por ejemplo, en el segmento 8 la palabra *approach* puede ser tanto un nombre como un verbo. Nuestro etiquetador la etiqueta correctamente como nombre. Después podremos usar esta información para buscar las palabras para una determinada categoría. También podemos utilizar la información del etiquetador para indexar únicamente las palabras que pertenezcan a categorías abiertas (nombres, verbos, adjetivos y adverbios). En la siguiente tabla podemos ver la versión lematizada y etiquetada con un conjunto de etiquetas muy reducido (n: nombre; v: verbo; en: adjetivo; r: adverbio; x: cualquier otra categoría).

ID	Segmento original	Segmento original lematizado y etiquetado
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 provisiones	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 reservas	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 hecho de añadir más información puede mejorar la indexación y la recuperación de segmentos similares, pero también hace que el sistema sea más vulnerable a errores. Por ejemplo, en el segmento 7 la palabra *means* ha etiquetado erróneamente como verbo y el en segmento 8 la palabra *needs* también ha etiquetado erróneamente como verbo. Los índices utilizando estos lemas y etiquetas, y evitando la indexación de las palabras no pertenecientes a categorías abiertas, quedaría de la siguiente manera:

```
approach_n 8
banknote_n 6: 8
be_v 8: 8
changeove_n 8
coin_n 6: 8
conduct_n 9
currency_n 6
demand_n 8
distribute_v 8
ecb_n 3
employment_n 9
escb_n 2
exchange_n 7
financial_a 10
```

```
foreign_a 7: 7
framework_n 1: 2
fraud_n 9
have_v 8
initial_a 8
institutional_a 3
january_n 8
legal_a 1: 2
market_n 10
matter_n 6
mean_v 6
monetary_a 4
need_v 8
operation_n 4
```

```
payment_n 5: 6
policy_n 4
prevention_n 9
provision_n 3
reserve_n 7
search_n 1
settlement_n 5
stability_n 10
successful_a 8
system_n 5
term_n 8
total_a 8
transparency_n 9
volume_n 8
```

Para hacer ahora la búsqueda del segmento más parecido a:

Banknotes, coins, and types of payment

tendríamos que lematizar y etiquetar también la oración, que quedaría:

banknote_n, _x coin_n, _x and_x type_n of_x payment_x

y como sólo indexamos las categorías abiertas los índices quedarían:

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

El segmento que se recuperaría en primer lugar sería el 6 (con 3 coincidencias), seguido del 8 (con 2 coincidencias) y finalmente el 5 (con una coincidencia)

Fijémonos en que aplicando técnicas que utilizan conocimiento lingüístico *(stemming*, lematitzación e información de categoría gramatical) hemos obtenido resultados diferentes que con las técnicas basadas en palabras enteras o bien en fragmentos de palabras sin motivación lingüística. Lo que nos queda por determinar ahora es cuál de los segmentos recuperados es el más adecuado para mostrar en primer lugar al traductor.

Todos estos cambios, para una memoria de 10 segmentos no son muy significativos. Imaginemos, sin embargo, una memoria con 1.000.000 de segmentos y veremos que el tamaño de los índices se podría reducir notablemente.

Lo que hemos explicado hasta aquí sobre la indexación de memorias de traducción es una aproximación a cómo se lleva a cabo realmente. Cada herramienta de traducción asistida puede utilizar una estrategia u otra de indexación para lograr una mejor eficiencia. También se pueden utilizar técnicas más avanzadas provenientes del campo de la recuperación de la información (Frakes and Baeza-Yates, 1992).

Una vez el programa de traducción asistida ha recuperado una serie de segmentos de la memoria de traducción usando un sistema de indexación, deberá calcular un índice de similitud entre la oración que buscamos y todos los segmentos recuperados de la memoria. Explicamos este aspecto en el siguiente apartado.

2.2.2.Cálculo de la similitud de segmentos

En este apartado veremos diferentes maneras de calcular la similitud entre dos segmentos. Lo primero que debemos tener en mente es qué queremos que signifique este tanto por ciento de similitud: ¿en qué grado se parecen?, ¿qué esfuerzo se debe hacer para pasar del segmento obtenido al deseado? Para un traductor probablemente será el segundo aspecto, es decir, obtener una especie de medida que nos indique el esfuerzo de cambio. Veremos un par de estrategias genéricas y observaremos cuál de las dos estrategias se aproxima mejor a este objetivo.

Cálculo de palabras coincidentes

La primera idea que se nos ocurre para calcular la similitud entre dos segmentos es mirar cuántas palabras tienen en común. Si todas las palabras son iguales, los segmentos tendrán un 100% de similitud (esto puede fallar por el orden de las palabras). Vamos a calcular la similitud según esto por los segmentos del ejemplo anterior. Primero calcularemos el % teniendo en cuenta el número de palabras iguales respecto al número de palabras total.

Segmento que buscamos:	Banknotes, coins, and types of payment
1º segmento encontrado:	Banknotes and coins, means of payment and currency matters
Palabras coincidentes:	3 Total palabras (segmento a buscar): 6 Similitud: 50%
2º segmento encontrado:	Payment and settlement systems
Palabras coincidentes:	1 Total palabras (segmento a buscar): 6 Similitud: 16.6%

Ahora hacemos el cálculo teniendo en cuenta el número de caracteres de las palabras que son iguales respecto al número total de caracteres.

Segmento que buscamos:	Banknotes, coins, and types of payment
1º segmento encontrado:	Banknotes and coins, means of payment and currency matters
Caracteres coincidentes:	22 Total caracteres (segmento a buscar): 34 Similitud: 64.7%
2º segmento encontrado:	Payment and settlement systems
Caracteres coincidentes:	7 Total caracteres (segmento a buscar): 34 Similitud: 20.6%

Fijémonos sin embargo, que algunos cambios de orden de palabras podrían hacer que el esfuerzo de edición para mantener el significado fondo muy superior al % calculado.

Cálculo de la distancia de edición

La *distancia de edición* o *distancia de Levenshtein* es el número mínimo de ediciones requeridas (inserción, supresión o sustitución de un carácter) para transformar una cadena de caracteres en otra.

Este cálculo nos puede dar una idea muy aproximada del esfuerzo real que puede suponer editar una coincidencia parcial de una memoria de traducción en la traducción real del segmento original. Por este motivo se puede utilizar con éxito para el cálculo de la similitud entre dos segmentos.

A continuación podemos observar el código Python de una función para calcular la distancia de edición de Levenshtein:

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

Si aplicamos este algoritmo en el ejemplo que nos ocupa obtenemos las siguientes cifras:

Segmento que buscamos:	Banknotes, coins, and types of payment
1º segmento encontrado:	Banknotes and coins, means of payment and currency matters
Distancia de edición:	31
2 º segmento encontrado:	Payment and settlement systems
Distancia de edición:	29

Este resultado nos puede sorprender un poco, ya que el segundo segmento encontrado, a pesar de tener menos palabras coincidentes, necesita menos esfuerzo de edición para poder formar el segmento original buscado.

En Somers (2003) se presenta un ejemplo muy claro de cómo las técnicas simples basadas en la distancia de edición pueden no funcionar correctamente en la selección del segmento más parecido (reproducimos aquí la parte inglesa del ejemplo). Consideramos que tenemos que traducir la siguiente oración:

Select 'Symbol' in the Insert menu.

y que disponemos de una memoria de traducción con los siguientes segmentos

S	Source	Target	Distancia edición
1	Select 'Symbol' in the Insert menu to enter a character from the symbol septiembre	Seleccione 'Símbolo' en el menú Insertar para introducir un carácter del conjunto de símbolos.	41
2	Select 'Paste' in the Edit menu.	Seleccione 'Pegar' en el menú Edición.	11
3	Select 'Paste' in the Edit menu to enter some texto from the clipboard.	Seleccione 'Pegar' en el menú Editar para introducir el texto del portapapeles. .	46

La mayoría de métricas de similitud basadas en la distancia de edición seleccionarían como más parecido el segmento 2, ya que sólo cambian dos palabras y la distancia de edición es menor. Pero intuitivamente el segmento 1 es una mejor coincidencia ya que aunque la distancia de edición es mucho mayor, en cambio incluye de manera exacta el texto que buscamos. Si recuperamos la traducción de la memoria de traducción sólo tendremos que borrar la parte final (*para introducir un carácter del conjunto de símbolos*).

Fijémonos también que los segmentos:

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

que tienen una distancia de edición de 30 y 8 palabras en común y 6 palabras diferentes.

Son más parecidos entre sí que los segmentos:

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

que tienen una distancia de edición de 11 y 4 palabras en común y 2 palabras diferentes.

Así pues, una métrica de similitud debería tener en cuenta no sólo la distancia de edición, sino también el número de palabras en común y palabras diferentes e incluso la longitud del segmento. Para tener en cuenta también la longitud de los segmentos que se comparan se puede utilizar el coeficiente de Dice (Trujillo, 1999), que se define como:

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

donde A y B es el número de palabras en el segmento de entrada y en el segmento recuperado de la memoria de traducción, respectivamente. C es el número de palabras comunes en los dos segmentos. A pesar de tener en cuenta el número total de palabras, esta medida no tiene en cuenta el orden de las palabras. También hay

que tener en cuenta que se debe hacer con las palabras duplicadas, si se tienen que considerar como una o como dos palabras. En algunas implementaciones de esta medida, en lugar de computar por palabras se computa por *bigramas*, es decir, por grupos de dos palabras adyacentes. Hay muchas más medidas que intentan determinar cuál es la similitud entre dos cadenas. En la siguiente tabla veremos el resultado de aplicar las siguientes medidas (no proporcionamos un definición de cada una de ellas) :

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

En los segmentos del apartado anterior:

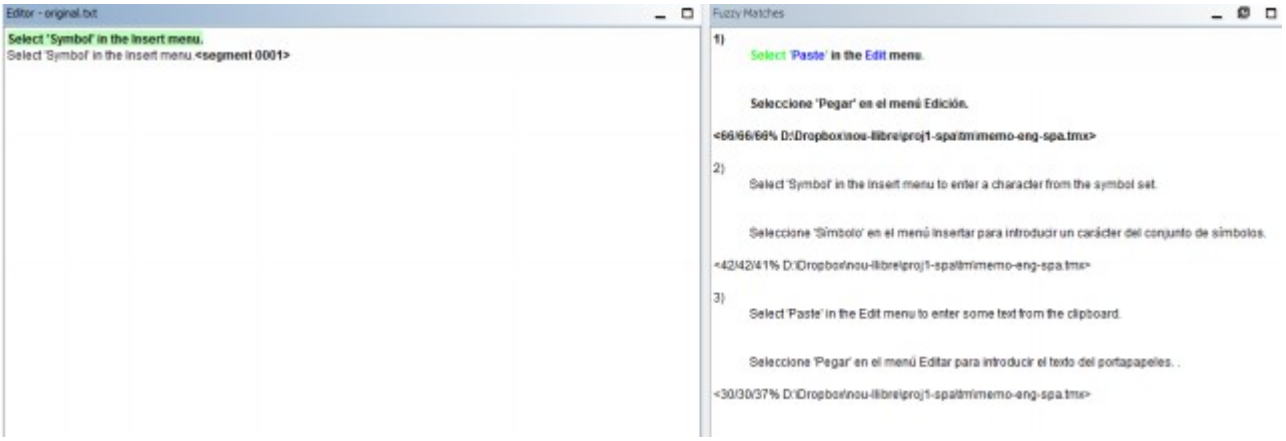
S: *Banknotes, coins, and types of payment*

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

Para cada medida marcamos en negrita el par de segmentos más similares según esta medida. La mayoría de medidas coinciden en determinar que los segmentos más similares son el S - M6. En cambio, según la *Levenshtein distance* los más parecidos son el S - M5, y según la *Longest common subsequence* las más parecidas son la S – M8.

Vemos ahora en la práctica qué segmento selecciona como más similar la herramienta OmegaT:


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.
Seleccione 'Pegar' en el menú Edición.
<66/66/66% D:\Dropbox\nou-llibre\proj1-spa\tm\memo-eng-spa.tmx>
2)
Select 'Symbol' in the Insert menu to enter a character from the symbol set.
Seleccione 'Símbolo' en el menú Insertar para introducir un carácter del conjunto de símbolos.
<42/42/41% D:\Dropbox\nou-llibre\proj1-spa\tm\memo-eng-spa.tmx>
3)
Select 'Paste' in the Edit menu to enter some text from the clipboard.
Seleccione 'Pegar' en el menú Editar para introducir el texto del portapapeles. .
<30/30/37% D:\Dropbox\nou-llibre\proj1-spa\tm\memo-eng-spa.tmx>

Vemos que en primer lugar selecciona el segmento 2 (asignándole una similitud del 66/66/66%); el segundo seleccionado es el segmento 1 (42/42/41%) y en tercer lugar el segmento 3 (30/30/37%). ¿Qué significan exactamente estos porcentajes que asigna OmegaT?

- La primera cifra indica el porcentaje de similitud utilizando el módulo de *tokenización* que permite calcular la raíz de las palabras y detectar las formas flexionadas. En este ejemplo las dos primeras cifras coinciden porque no tenemos este módulo activado y para que todas las palabras del original están sin flexionar.
- La segunda cifra nos indica el porcentaje del número de palabras coincidentes (ignorando los números y las etiquetas) dividido entre el número total de palabras.
- La tercera cifra es igual que la anterior pero teniendo en cuenta las cifras y las etiquetas.

Veamos ahora cómo responde a este ejemplo la herramienta Tikal (de Okapi Tools) si indexamos los segmentos de nuestro ejemplo en una memoria de traducción de tipo Pensieve (la propia de Okapi Tools).

```
tikal.sh -q "Select 'Symbol' in the Insert menu." -pen memoPengspa.pentm / -SL en -tl se -Optar 0
------------------------------------------------ ----------------------------
Okapi Tikal - Localization Toolset
Version: 2.0.23
------------------------------------------------ ----------------------------

= From net.sf.okapi.connectors.pensieve.PensieveTMConnector (entre> s)
  Threshold = 0, Maximum hits = 25
resultado: 54, origin: ''
  Source: "
             Select 'Symbol' in the Insert menu to enter a character from the symbol septiembre
          "
  Target: "
              Seleccione 'Símbolo' en el menú Insertar para introducir un carácter del conjunto de
símbolos.
          "
```

Vemos pues, que comparando sólo dos herramientas concretas, ya se producen discrepancias importantes en las coincidencias que se obtienen.

La recuperación de segmentos y el cálculo de la similitud entre el segmento a buscar y los recuperados desde la memoria de traducción es un aspecto muy importante en la técnica de traducción automática llamada

traducción automática basada en ejemplos (*Example-Based Machine Translation* – EBMT). En los siguientes párrafos describimos algunas

Algunos autores hacen propuestas diferentes para el cálculo de la similitud entre segmentos (Somers y Fernández, 2004). Por ejemplo, Denet (1995) propone por un lado tener en cuenta la significancia relativa de las palabras que cambian, basándose en datos estadísticos, y por otro lado identificar fragmentos de segmentos que sean significativos desde el punto de vista sintáctico. Cranias et al (1991) propone considerar lemas en vez de cadenas de caracteres y hacer uso de las palabras funcionales así como de las etiquetas morfosintácticas.

Planas y Furuse (1999) proponen un esquema de búsqueda de coincidencias multi-capa y flexible. Los autores proponen 8 capas: (1) caracteres de texto, (2) palabras, (3) lemas, (4) categorías gramaticales (POS), (5) etiquetas XML de contenido, (6) etiquetas XML vacías (que representan por ejemplo imágenes), (7) entradas de glosario y (8) estructuras de análisis lingüístico (esta capa dependerá del nivel de análisis que proporcione el analizador lingüístico disponible). La similitud entre estas estructuras se calcula a partir de la distancia de edición.

Macklovitch y Russell (2000) tienen en cuenta otros aspectos como la flexión de las palabras y las categorías gramaticales y también consideran el reconocimiento de ciertas entidades con nombre, el reconocimiento de nombres propios y el análisis sintáctico superficial.

Rapp (2002) utiliza un etiquetador morfosintáctico para procesar las memorias de traducción y utilizar la información de categoría gramatical para mejorar la búsqueda de segmentos similares.

Indicación de las palabras coincidente y diferentes con un código de colores

Es interesante que el programa de traducción asistida muestre las coincidencias parciales de las memorias con indicación de las palabras que son coincidentes y las que son diferentes. A menudo esto se hace mediante un código de colores.

Siguiendo con el mismo ejemplo que en los apartados anteriores, observamos en la siguiente imagen como OmegaT marca con colores la coincidencia más parecida de la memoria de traducción (en el ejemplo, el segmento que estamos traduciendo es: *Select 'Symbol' in the Insert menu.)*

El programa es capaz de detectar las palabras iguales (en verde) y diferentes (azul) del texto correspondiente a la lengua de partida, pero no marca de ningún color el segmento correspondiente en la lengua de llegada, ya que no tiene suficiente información lingüística para determinar qué palabra de la lengua de partida corresponde con qué palabra de la lengua de llegada.

2.3. Coincidencia exacta y parcial

Así pues, las coincidencias que se recuperan de la memoria de traducción pueden ser:

- Coincidencia exacta (*exact match):* cuando los texto del segmento que recuperamos de la memoria de traducción es exactamente igual al texto del segmento que buscamos.
- Coincidencia parcial (f*uzzy match):* cuando los texto del segmento que recuperamos de la memoria de traducción no es exactamente igual al texto del segmento que buscamos, pero su índice de similitud es superior o igual al índice fijado por el usuario.

A esta definición simplista hay que añadir algunas explicaciones:

- Una coincidencia que sólo difiera en algunas cifras se puede considerar exacta si el programa de traducción asistida es capaz de sustituir-las por las cifras presentes en el segmento que buscamos (muchas de las herramientas actuales son capaces de llevar a cabo esta sustitución). Consideremos que estamos traduciendo el segmento *"An example is shown in figure 3*". En nuestra memoria tenemos el segmento *"An example is shown in figure 1"* con su correspondiente traducción *"Se muestra un ejemplo en la figura 1".* Muchas de las herramientas actuales son capaces de recuperar el segmento y mostrar el texto traducido con la cifra sustituida *"Se muestra un ejemplo en la figura 3".*
- Algunas herramientas son capaces de hacer estas sustituciones para cadenas alfanuméricas, no sólo para cifras y serían capaces de hacer lo mismo para un segmento como *"This is shown as A in the diagram".* (Ejemplo tomado de Somers (2004)). Si en nuestra memoria tenemos un segmento como *"This is shown as B in the diagram"* con su correspondiente traducción *"Esto se indica como B en el diagrama"*, el programa propondría la traducción cambiando la B por A y con una coincidencia exacta *"Esto se indica como A en el diagrama".*
- Otro caso a tener en cuenta es aquel en el que los textos del segmento a buscar y el recuperado de la memoria son exactamente iguales pero difieren en algún tipo de formato o marca especial. Pongamos por caso que el texto a buscar es "*Press **OK** to continue*" (con OK en negrita) y en la memoria tenemos un par "*Press OK to continue*" y su traducción "*Haga clic en Aceptar para continuar*". En este caso el programa podrá recuperar el texto pero no será capaz de poner Aceptar en negrita.

En este sentido, Bowker (2002) introduce la distinción entre *coincidencia exacta (exact match)* y *coincidencia completa (full match).* Según este autor:

> *Una coincidencia exacta es 100% idéntica al segmento que está traduciendo el traductor tanto desde el punto de vista lingüístico, como desde el punto de vista del formato. Esto significa que las dos cadenas tienen que ser idénticas en todos los sentidos, incluyendo la ortografía, la puntuación, la flexión, cifras e incluso el formato (cursiva, negrita, etc.)*

Una *coincidencia total* se produce cuando el segmento que se está traduciendo difiere del segmento almacenado en la memoria de traducción sólo en lo que se denominan *elemento variables,* que a menudo en inglés reciben el nombre de *placeables.* Estos elementos variables incluyen cifras, fechas, horas y unidades monetarias. Algunas herramientas de traducción asistida son capaces de detectar y hacer cambios de algunos de estos elementos variables. El caso más habitual es el de las cifras. Por ejemplo: si estamos traduciendo un fragmento como "*The papel tray has a capacity of 200 sheets*" y en la memoria tenemos un segmento "*The papel tray has a capacity of 150 sheets*" y su traducción "*La bandeja de papel tiene una capacidad de 150 hojas*" muchas de las herramientas actuales podrán recuperar el segmento como coincidencia exacta cambiando la cifra y mostrar "*La bandeja de papel tiene una capacidad de 200 hojas*".

2.4.Combinación de unidades subsegmentales

En muchos casos en la memoria de traducción no tenemos ningún segmento completo similar al que estamos traduciendo, pero en cambio disponemos de uno o más fragmentos de segmentos que contienen información interesante. Veamos un ejemplo (adaptado de Bowker 2002)

Segmento a traducir	First, **check for disk space on the drive** that contaré the Temp folder.
Segmento de la memoria de traducción	(eng) Close other programs, **check for disk space on the drive** you are saving to, and then save again (spa) Cierre los otros programas, verifique el espacio de disco en la unidad donde está guardando y luego vuelva a guardar.

Algunos programas son capaces de buscar en la memoria coincidencias a nivel sub-segmental. Una tarea más compleja, pero que algunas herramientas pueden llegar a hacer, es deducir que la traducción de *check for disk space on the drive* al castellano es *verifique el espacio de disco en la unidad.* El programa puede llegar a deducir esto de una manera totalmente estadística si este subsegmento aparece en varios segmentos de la memoria de traducción.

Un paso más allá aún es la capacidad de algunos programas para componer una nueva traducción a partir de coincidencias subsegmentales. Veamos el siguiente ejemplo (también adaptado de Bowker 2002).

Segmento a traducir	**The file operation can not be completed because the disk is** full.
Segmento de la memoria de traducción	(eng) There is not enough memory to perform **the file** operation. (spa) No hay suficiente memoria para llevar a cabo **la operación de** archivos.
Segmento de la memoria de traducción	(eng) This action **can not be completed because the** program is busy. (spa) Esta acción **no se puede completar porque el** programa está ocupado.
Segmento de la memoria de traducción	(eng) **Disk is full** (spa) **El disco está lleno**
Traducción compuesta a partir de las coincidencias subsegmentals	La operación de archivos no se puede completar porque el disco está lleno

Simard (2001) presenta un sistema capaz de trabajar a nivel sub-segmental que trabajaba a partir de la consideración de todas las subsecuencias posibles (n-gramas) del segmento a traducir y recupera también todos los pares de subsecuencias posibles de la memoria de traducción . Langas (2001) y Colominas (2008) presentan propuestas donde las subsecuencias tienen una motivación lingüística (se trata de *chunks:* es decir un sintagma no recursivo correspondiente a una categoría léxica principal (nombre, adjetivo, preposición y verbo), que admiten que junto con el núcleo pueden incluir tanto premodificadores como postmodificadores). En el ejemplo anterior no se han tenido en cuenta *chunks*, sino unidades arbitrarias. En la siguiente tabla podemos observar el mismo ejemplo si trabajamos con chunks (calculados con Freeling).

Segmento a traducir	[The file operation] [can not be completed] because [the disk] is hoja.
Segmento de la memoria de traducción	(eng) There is [not enough memory] to perform [the file operation]. (spa) No hay [suficiente memoria] para llevar a cabo [la operación de archivos].
Segmento de la memoria de traducción	(eng) [This action] [can not be completed] because [the program] is busy. (spa) [Esta acción] [no se puede completar] porque [el programa está ocupado].
Segmento de la memoria de traducción	(eng) Disk is full (spa) El disco está lleno
Traducción compuesta a partir de las coincidencias subsegmentals	La operación de archivos no se puede completar porque el disco está lleno

Recordemos que el sistema de traducción asistida sólo podrá determinar la traducción de una subsecuencia si esta aparece muchas veces en la memoria de traducción. Como no siempre las memorias de traducción que se asignan a un proyecto tienen el tamaño suficiente, algunos autores (Bicicat 2008) proponen utilizar un sistema de traducción automática estadística basado en frases entrenado con un corpus del mismo dominio.

2.5. Formato de intercambio de memorias de traducción: TMX

Como hemos visto en los apartados anteriores, los diferentes sistemas de traducción asistida indexan y almacenan en bases de datos las memorias de traducción de formas muy diferentes. Por este motivo, las memorias de traducción no son compatibles entre las diferentes herramientas. Para poder compartirlas ha creado un lenguaje de intercambio basado en XML llamado *Translation Memory eXchange* (TMX). Todas las herramientas de traducción asistida son capaces de guardar sus memorias en este formato y cargar archivos TMX en sus bases de datos.

El TMX se define en dos partes:

- Una especificación del formato del contenedor, es decir, los elementos de nivel superior que proporcionan información sobre el archivo en conjunto y sobre las entradas. En TMX una entrada consistente en segmentos alineados de texto en dos o más lenguas se denomina *unidad de traducción* (el elemento <tu>).
- Una especificación para el formato de meta-marcado de bajo nivel para el contenido de un segmento de texto de la memoria de traducción. En TMX, un segmento individual del texto de la memoria de traducción en una lengua determinada se denota con el elemento <seg>.

A continuación podemos observar un ejemplo de memoria de traducción en formato TMX consistente en un único segmento en castellano y catalán:


```
<?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 puede tener dos niveles de implementación:

- Nivel 1. Únicamente texto plano: Soporte sólo para el contenedor. Los datos entre los elementos <seg> contienen únicamente información textual, sin marcas de formato.
- Nivel 2. Marcado del contenido: Soporte tanto para el contenedor como para el contenido. Se utiliza el marcado de contenido propio del TMX para permitir que otras herramientas que sean compatibles con TMX nivel 2 puedan recrear la versión traducida de un documento original usando únicamente el archivo TMX.

El ejemplo anterior de TMX correspondería a un Nivel 1. Si el segmento original tuviera el siguiente formato:

EDICTO de 14 de febrero de 2000, sobre un acuerdo de la *Comisión de Urbanismo* de Tarragona referente al municipio de Reus.

la memoria en TMX de Nivel 2 tendría el siguiente aspecto:

```
<?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>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>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 nos fijamos, este nivel incluye información sobre el formato del segmento. El nivel 2 de TMX es muy útil para traducir documentación con formato (negritas, colores, etc) variado, ya que en muchos casos podrá recuperar también las marcas de formato y ahorrará tiempo de edición al traductor.

2.6.Creación de memorias de traducción

2.6.a. Traducción con sistemas de traducción asistida

Si trabajamos habitualmente con sistemas de traducción asistida la creación de memorias de traducción es directa, ya que todos los sistemas TAO son capaces de generar las memorias para cada proyecto de traducción.

2.6.b. Memorias de traducción y corpus paralelos

El concepto de memoria de traducción y corpus paralelo se puede considerar equivalente. Actualmente hay una gran cantidad de corpus paralelos disponibles en Internet. Se puede echar un vistazo a la Colección Opus (http://opus.lingfil.uu.se/) (Tiedemann 2012), de la que hablaremos con más detalle en el apartado *Para ampliar conocimientos* de este mismo capítulo.

En general los corpus paralelos son de tamaños relativamente grandes. Si hay disponible un corpus paralelo para nuestro par de lenguas y especialidad, puede resultar de utilidad descargarlo y usarlo como memoria de traducción dentro de nuestros proyectos de traducción.

2.6.c. Alineación manual de documentos

La alineación de documentos es un proceso por el que se toman un documento original y su traducción y se genera un archivo que relaciona los segmentos originales con los correspondientes segmentos traducidos, es decir, una memoria de traducción. Este proceso es útil para crear memorias de traducción a partir de documentos originales y sus traducciones. Es importante tener en cuenta, como hemos comentado anteriormente, que si los documentos los hemos traducido con una herramienta de traducción asistida, no será necesario llevar a cabo este proceso, ya que podremos generar la memoria de traducción directamente desde la herramienta de traducción asistida.

El proceso genérico de alineación de documentos se puede dividir en dos pasos:

- Segmentación de los documentos originales y traducidos
- Relacionar los segmentos originales con los segmentos traducidos correspondientes

La segmentación consiste en dividir el texto de los documentos en segmentos a partir de un conjunto de reglas de segmentación. Las reglas de segmentación nos indican dónde termina un segmento y dónde empieza otro. Una regla de segmentación nos podría indicar que un punto, seguido de un espacio en blanco y seguido de una palabra que empieza por mayúscula indica un límite de segmento. Esta regla podría segmentar correctamente el texto:

Hoy he comido en casa. Mañana comeré en el trabajo.

en los segmentos:

Hoy he comido en casa.

Mañana comeré en el trabajo.

Ahora bien, esta regla no funcionaría correctamente para segmentar el texto:

El sr. Martínez no ha asistido a la reunión.

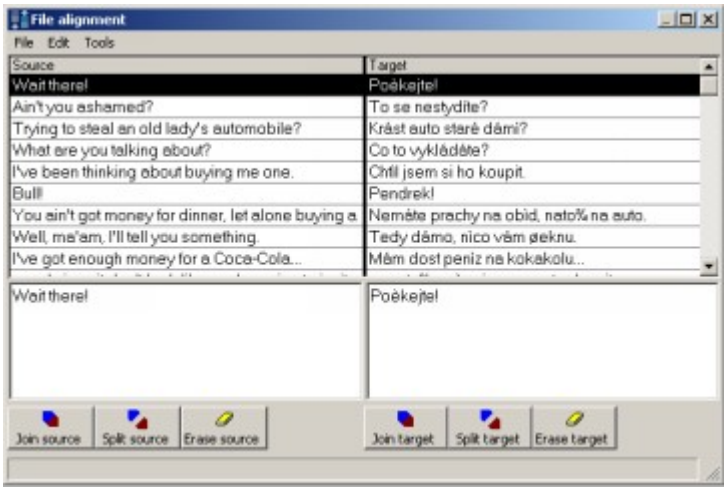
ya que resultaría la segmentación:

El sr.

Martínez no ha asistido a la reunión.

La mayoría de sistemas de traducción asistida ofrecen la posibilidad de especificar las reglas de segmentación que utilizan. Para sacar el máximo provecho de una determinada memoria de traducción conviene utilizar las mismas reglas de segmentación en la creación del proyecto que las que se utilizaron en la creación de la memoria de traducción. Por este motivo se ha creado un formato estándar de intercambio de reglas de segmentación basado en XML que se llama SRX (*Segmentation Rule eXchange*).

Las herramientas de alineación manual de documentos disponen de una interfaz gráfica que nos permite relacionar manualmente los segmentos originales con los correspondientes segmentos traducidos.



Interfaz de alineación del Déja Vu 3

Si los documentos original y traducido parecen en cuanto a formato y puntuación y la mayoría de segmentos originales tienen una relación 1:1 (es decir, cada segmento original se corresponde con un segmento traducido) la alineación obtenida únicamente a partir de la segmentación será suficientemente precisa y se requerirá poca intervención humana para completar la alineación. Ahora bien, esto no siempre sucede. Muy a menudo un único segmento original se traduce por dos segmentos (relación 1:2) o bien dos segmentos originales se traducen por uno solo (relación 2:1). Incluso a veces sucede que un segmento original simplemente no aparece en la traducción (relación 1:0) o que en la traducción aparecen nuevos segmentos (relación 0:1). Esto hace que la alineación manual de documentos llegue a ser una ser una tarea realmente ardua y que requiere una gran intervención humana. Por este motivo se han desarrollado diversas metodologías y herramientas de alineación automática de documentos, que veremos en el siguiente apartado.

2.6.d. Alineación automática de documentos

La alineación manual de documentos puede suponer una carga de trabajo importante. Cuando la traducción respeta mucho el formato, los párrafos y el número de oraciones del original, la alineación manual puede resultar efectiva. En otros casos la alineación manual puede suponer una carga de trabajo más elevada que la que nos ahorraremos traduciendo utilizando la memoria que podemos generar a partir de la alineación.

Existen una serie de algoritmos que permiten llevar a cabo la alineación automática de documentos. Estos algoritmos nos permitirán alinear conjuntos grandes de documentos sin prácticamente ningún esfuerzo y generar memorias de traducción.

La alineación automática de documentos sigue los pasos genéricos de segmentación y relación de segmentos, pero la relación de segmentos se hace de manera automática y sin intervención del usuario.

Se pueden distinguir tres metodologías de alineación automática:

- Basada en la longitud de los segmentos (en caracteres o palabras)
- Basada en un diccionario bilingüe
- Basada en técnicas gráficas

La primera de las metodologías se basa en el hecho de que normalmente los segmentos originales más largos se traducen por segmentos más largos. A partir de la segmentación de los documentos se computan parámetros estadísticos basados en la longitud de los segmentos y se calculan estos mismos parámetros estadísticos de diversas variaciones de la segmentación original. Se elige como mejor segmentación aquella que presenta una distribución más uniforme de la relación longitud del segmento original respecto a la longitud del segmento traducido. Esta estrategia fue empleada por primera vez por Kay y Röscheisen (1993) mediante una aproximación que no resultó muy efectiva para corpus de gran tamaño. Por otra parte, y de manera bastante simultánea Brown (1993) y Gale y Church (1993) desarrollaron otros algoritmos basados en esta misma idea.

La segunda metodología se basa en el hecho de conocer la traducción de ciertas palabras o grupos de palabras. Si estas palabras aparecen en el segmento original se espera que en el segmento traducido aparezca

la correspondiente traducción. A partir de la segmentación original se va modificando esta para conseguir que el número de segmentos originales que presentan palabras del diccionario y que el segmento traducido correspondiente contenga la traducción de las palabras sea máximo. Estas aproximaciones se pueden encontrar en Chen (1993) y Wu (1994).

La tercera de las metodologías utiliza técnicas gráficas (representando gráficamente diversos parámetros de los documentos originales y traducidos) para encontrar la alineación más probable (Melamed 1997).

Moore (2002) ha desarrollado una estrategia híbrida que utiliza tanto la metodología basada en la longitud de segmentos como la basada en diccionarios bilingües. El sistema presenta la particularidad de no necesitar de un diccionario bilingüe ya que funciona en dos pasos. En el primer paso se lleva a cabo una alineación automática basada en la longitud de los segmentos y el sistema toma únicamente aquellos pares de segmentos que se han podido alinear con mucha seguridad. A partir de estas alineaciones seguras el sistema aprende automáticamente un diccionario bilingüe estadístico que se utiliza para llevar a cabo una segunda alineación basada en diccionario y complementará a la primera alineación intentando alinear aquellos segmentos no seguros. La herramienta Hunalign (Varga el al. 2005) utiliza una estrategia híbrida muy parecida. En esta herramienta se puede utilizar también un diccionario bilingüe proporcionado por el usuario. En caso de que no se proporcione un diccionario el programa aprende uno con una primera alineación basada en longitud.

2.7. Memorias de traducción remotas compartidas y públicas

La manera tradicional de trabajar con memorias de traducción ha sido hasta hace poco trabajar con las memorias locales, en la propia máquina o en una red local. Como los equipos de traductores habitualmente están en diferentes ubicaciones, a menudo a mucha distancia, el hecho de trabajar con memorias locales ha provocado algunos problemas:

- Si las memorias de traducción son muy grandes, enviarlas a los diferentes traductores ha sido un problema por el tamaño de los archivos.
- Además, enviar una memoria de traducción completa a un colaborador puntual puede crear problemas importantes de confidencialidad
- Existe la posibilidad de enviar sólo los segmentos de la memoria útiles para el proyecto en que está trabajando actualmente el traductor
- Sin embargo, los nuevos segmentos traducidos por otro colaborador del mismo proyectos no estarán disponibles para el resto de traductores

Las tecnologías relacionadas con Internet han permitido el desarrollo de memorias de traducción remotas. En Simões (2004) se describe un sistema de memorias de traducción distribuidas implementadas mediante servicios web.

Existen diversas herramientas que ofrecen implementaciones eficientes de memorias de traducción remotas, entre las que podemos destacar:

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

Estas dos herramientas son de software libre.

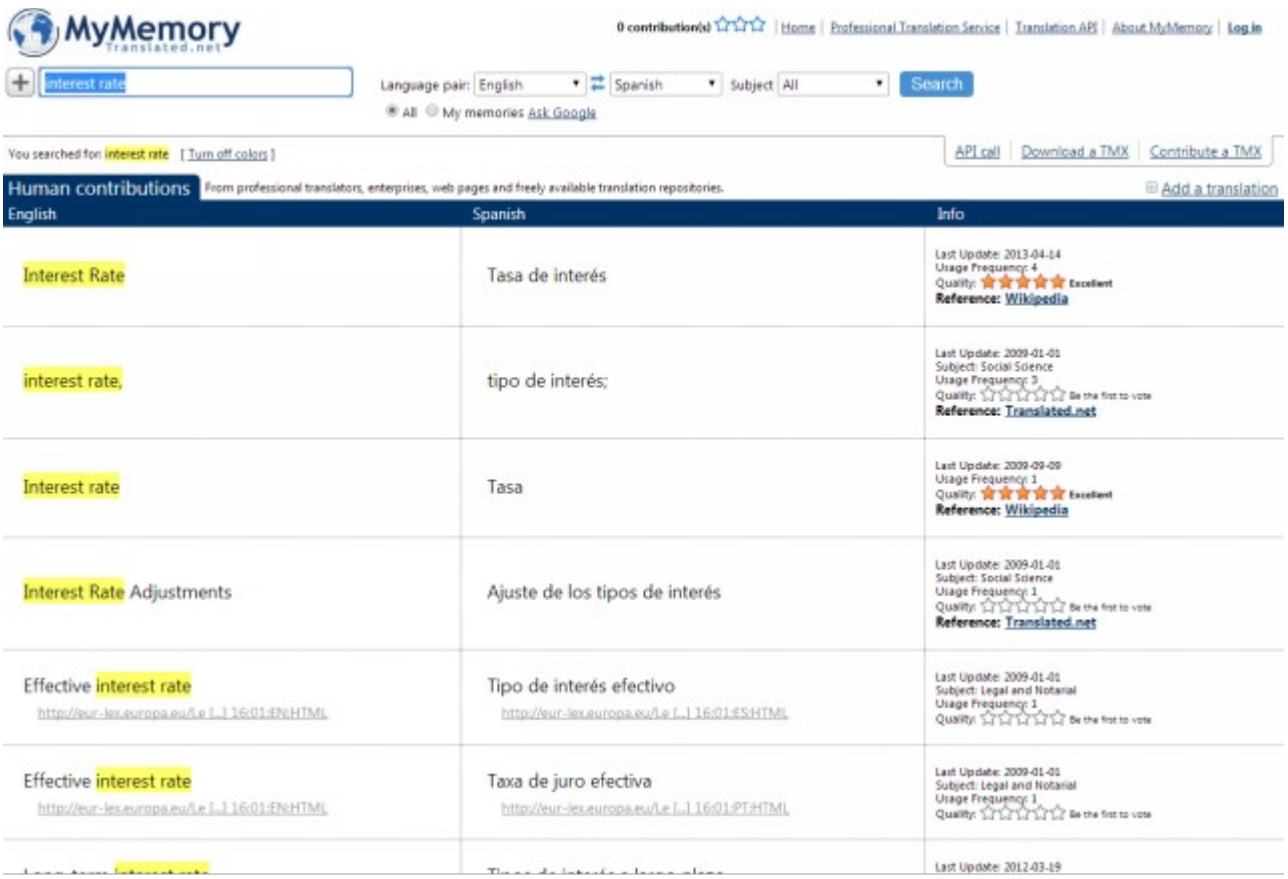
Algunas aplicaciones que incorporan directamente la funcionalidad de memorias remotas:

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

Estas dos herramientas, a pesar de no ser de software libre, son de uso gratuito. Ambas funcionan directamente desde el navegador de Internet y son una muy buena opción para trabajar con una herramienta de traducción asistida sin tener que instalar nada en nuestro ordenador.

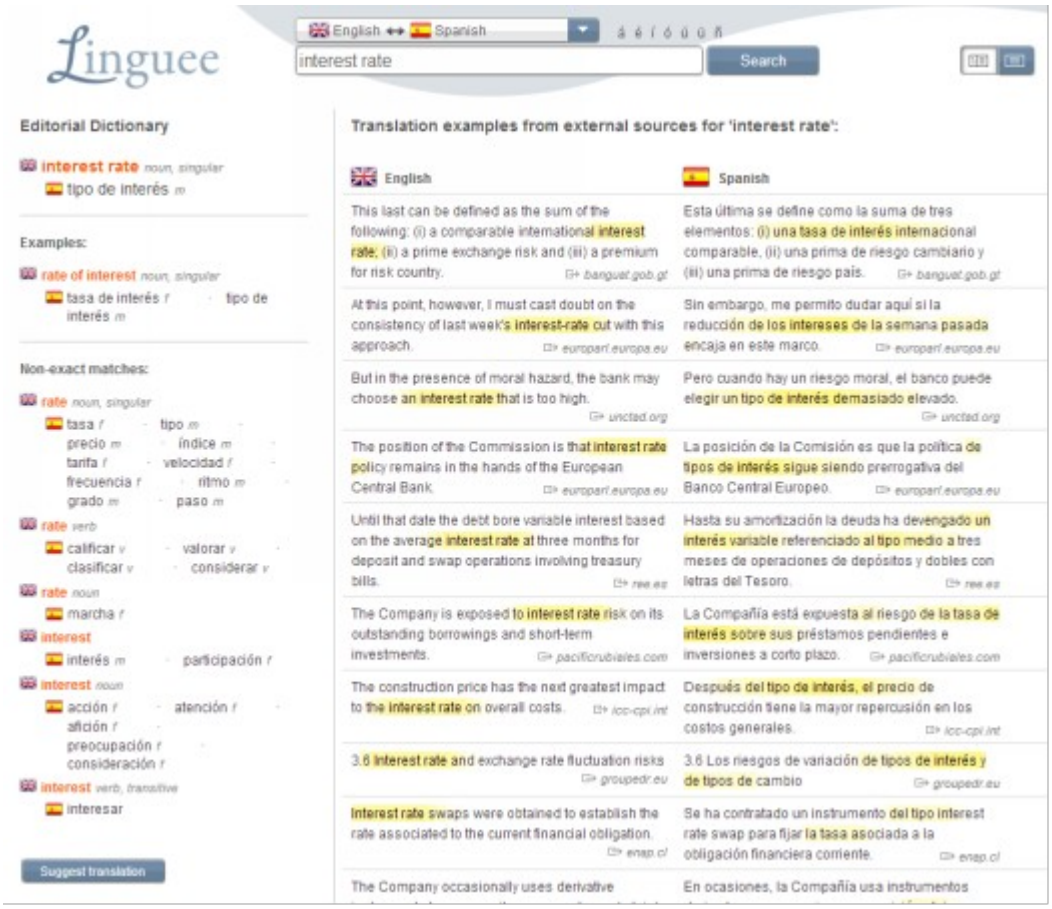
Servicios de memorias compartidas públicas:

- **MyMemory TM** (http://mymemory.translated.net/): este recurso se ha creado a partir de las memorias de traducción de la Unión Europea y de las Naciones Unidas, así como alineando varios sitios web multilingües. También permite subir memorias de traducción propias. El sistema se puede consultar automáticamente desde otras aplicaciones y también permite descargar memorias de traducción a partir de documentos a traducir. También se pueden hacer consultas directamente a su interfaz web. Esta funcionalidad puede resultar de utilidad para buscar equivalentes de traducción de términos (véase la siguiente figura). El sistema también proporciona una traducción automática a partir de unos sistema de traducción automática estadística.

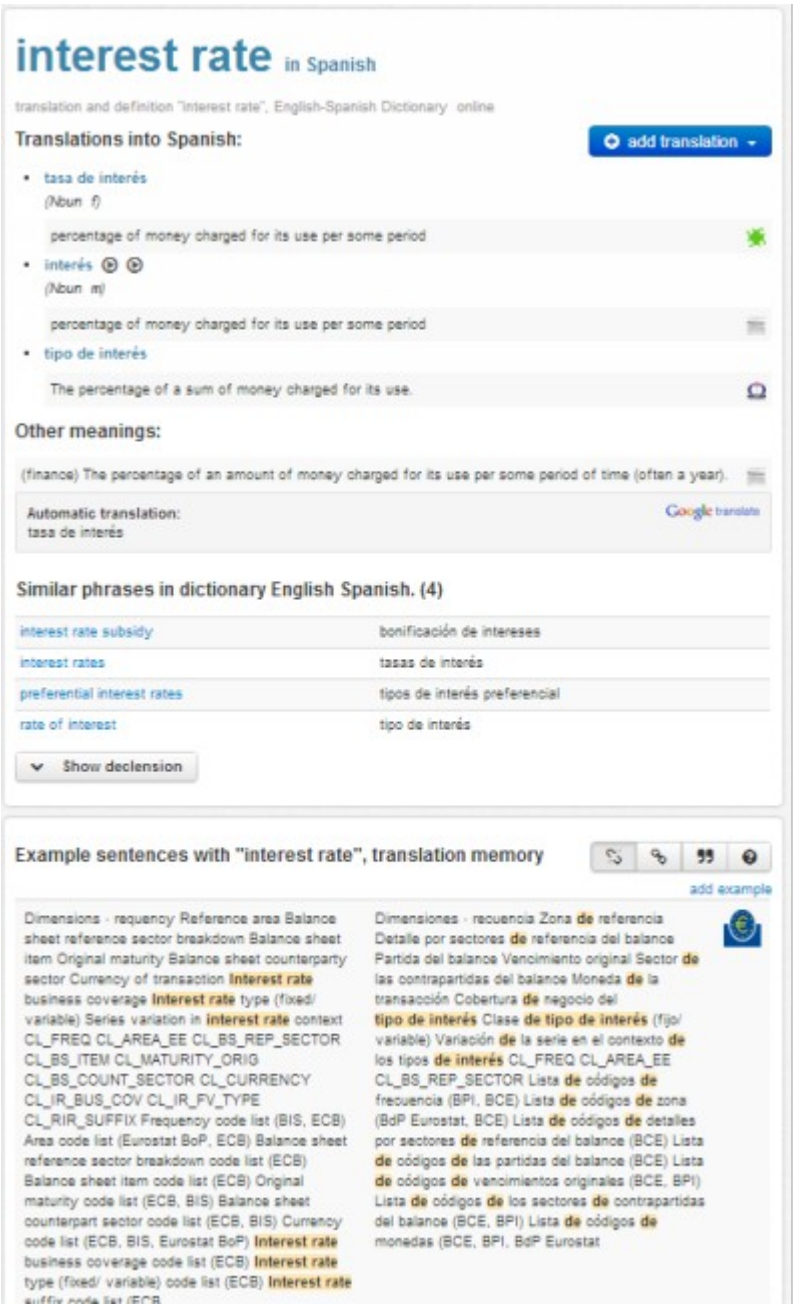

MyMemory
translated.net
0 contribution(s)
Home | Professional Translation Service | Translation API | About MyMemory | Log in
interest rate
Language pair: English Spanish Subject: All Search
All My memories Ask Google
You searched for: interest rate [Turn off colors]
API call | Download a TMX | Contribute a TMX
Human contributions
From professional translators, enterprises, web pages and freely available translation repositories.
Add a translation
English
Spanish
Info
Interest Rate
Tasa de interés
Last Update: 2013-04-14
Usage Frequency: 4
Quality: Excellent
Reference: Wikipedia
interest rate,
tipo de interés;
Last Update: 2009-01-01
Subject: Social Science
Usage Frequency: 3
Quality: Be the first to vote
Reference: Translated.net
Interest rate
Tasa
Last Update: 2009-09-09
Usage Frequency: 1
Quality: Excellent
Reference: Wikipedia
Interest Rate Adjustments
Ajuste de los tipos de interés
Last Update: 2009-01-01
Subject: Social Science
Usage Frequency: 1
Quality: Be the first to vote
Reference: Translated.net
Effective interest rate
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:116:0001:0001:EN:HTML
Tipo de interés efectivo
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:116:0001:0001:ES:HTML
Last Update: 2009-01-01
Subject: Legal and Notarial
Usage Frequency: 1
Quality: Be the first to vote
Effective interest rate
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:116:0001:0001:EN:HTML
Taxa de juro efectiva
http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:116:0001:0001:PT:HTML
Last Update: 2009-01-01
Subject: Legal and Notarial
Usage Frequency: 1
Quality: Be the first to vote
Last Update: 2012-03-19

- **TDA Translation Repository** (http://www.tausdata.org/index.php/taus-search): la TAUS Data Association (TDA) ofrece una interfaz de búsqueda pública a un gran corpus de traducciones. Algunas aplicaciones (como por ejemplo las herramientas de Okapi) ofrecen la consulta automática a este recurso.

- **Linguee** (http://www.linguee.com/): Combina un diccionario con un repositorio de memorias de traducción. En el momento de escribir este capítulo no disponía de una API para consulta automática desde otra herramienta.



- **Glosbe** (http://glosbe.com/): Una herramienta similar a las anteriores pero que intenta incluir el mayor número de lenguas posible. Además de la búsqueda en memorias proporciona una traducción automática realizada con Google Translate. También permite que el usuario aporte nuevas traducciones.



Algunas herramientas, como por ejemplo MemoQ (http://kilgray.com/products/memoq) permiten tres tipos de memorias de traducción:

- memorias de traducción locales
- memorias de traducción remotas
- memorias de traducción remotas sincronizadas: que son un híbrido de las dos anteriores. Se trata de memorias de traducción remotas pero que se descargan y sincronizan con una copia local, de modo que si en algún momento no disponemos de conexión a Internet podamos seguir trabajando con el proyecto. En el momento que recuperamos de nuevo la conexión a Internet la copia local y remota se volverán a sincronizar.

2.8. Trabajo con memorias de traducción

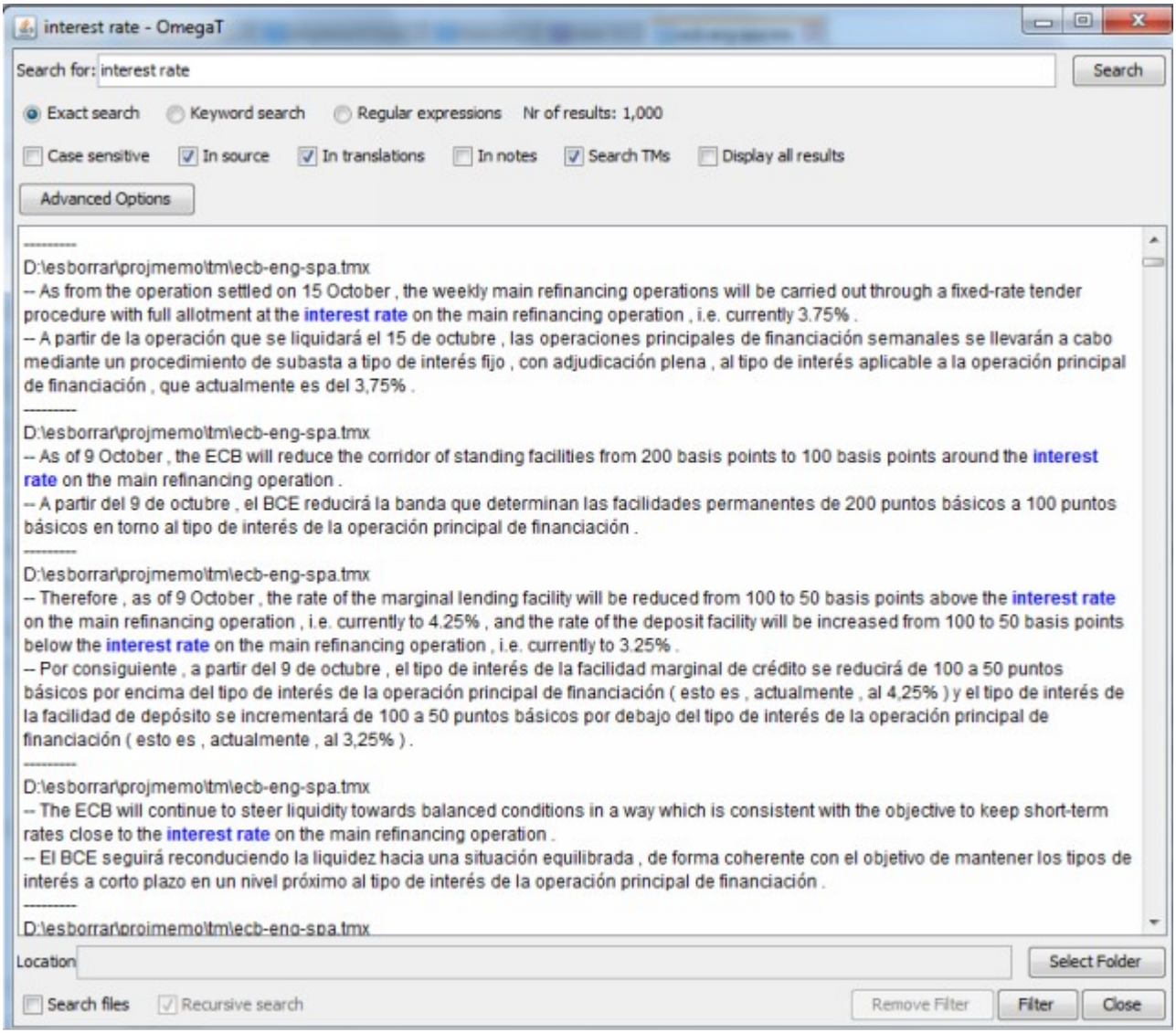
Bowker (2002) distingue dos maneras de trabajar con memorias de traducción dentro de una herramienta de traducción asistida:

- Modo interactivo (*interactive mode*): A medida que el traductor va trabajando con la herramienta de traducción asistida, cada vez que cambia de un segmento a otro el sistema busca coincidencias dentro de la memoria de traducción y muestra las coincidencias exactas o las que tengan un índice de similitud superior o igual al indicado por el usuario.
- Modo por lotes *(batch mode):* Este modo menudo se denomina **pretraducción** y consiste en consultar en la memoria de traducción todos los segmentos del proyecto y poner en la traducción el segmento más parecido que supere un índice mínimo de similitud indicado por el usuario.

En ambos casos el traductor tendrá que revisar las propuestas proporcionadas por la memoria de traducción. La pre-traducción es útil en los casos que queramos enviar un proyecto de traducción a un colaborador sin tener que enviar toda una memoria de traducción.

Un punto importante a tener en cuenta cuando se trabaja con memorias de traducción es el establecimiento de la similitud mínima para recuperar segmentos de la memoria. Si trabajamos con similitudes muy altas, por ejemplo del 95%, será muy difícil encontrar coincidencias en la memoria y el programa mostrará muy pocas sugerencias. En cambio, si trabajamos con similitudes muy bajas, de por ejemplo el 10%, el sistema nos mostrará muchas sugerencias, pero probablemente serán de poca utilidad. Un buen compromiso puede ser establecer la similitud mínima entre el 65 y el 85%.

Hay que tener en cuenta, sin embargo, que a pesar de que nuestra memoria de traducción no contenga segmentos parecidos al que estamos traduciendo, la mayoría de herramientas de traducción permiten la búsqueda de fragmentos de texto (típicamente unidades terminológivas) dentro de la memoria de traducción, de manera que nos muestre todos los segmentos originales de la memoria de traducción que contienen este fragmento y las correspondientes traducciones. En la siguiente imagen podemos ver la función *Search project* de OmegaT que permite realizar búsquedas en los proyectos y las memorias de traducción:



Es importante no confundir el concepto de pretraducción con el concepto de **pseudotraducción.** La pseudotraducción consiste en hacer una primera traducción del proyecto simulando una lengua de llegada ficticia: puede componerse de caracteres aleatorios, de cambios de ciertos caracteres de la original, etc El objetivo es verificar si el proceso de importación de los archivos originales y de la creación de los ficheros traducidos finales funciona correctamente. A continuación podemos ver un ejemplo de cadena pseudotraduïda (ejemplo extraído de http://en.wikipedia.org/wiki/Pseudolocalization).

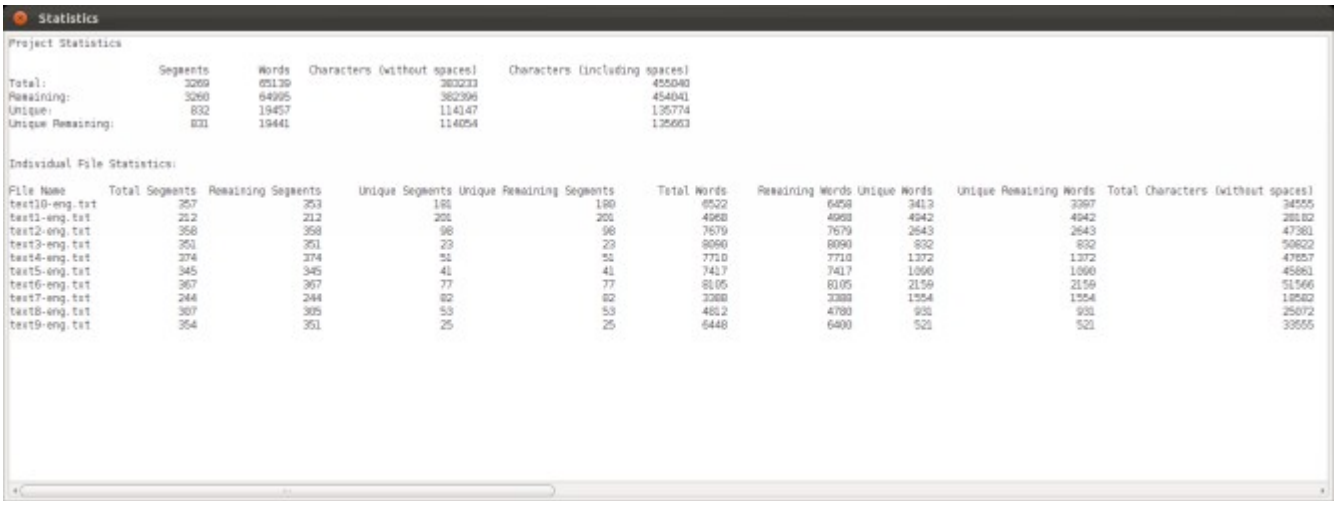
Account Settings	[!!! Account Settings !!!]

2.9. Análisis de proyectos y tarificación

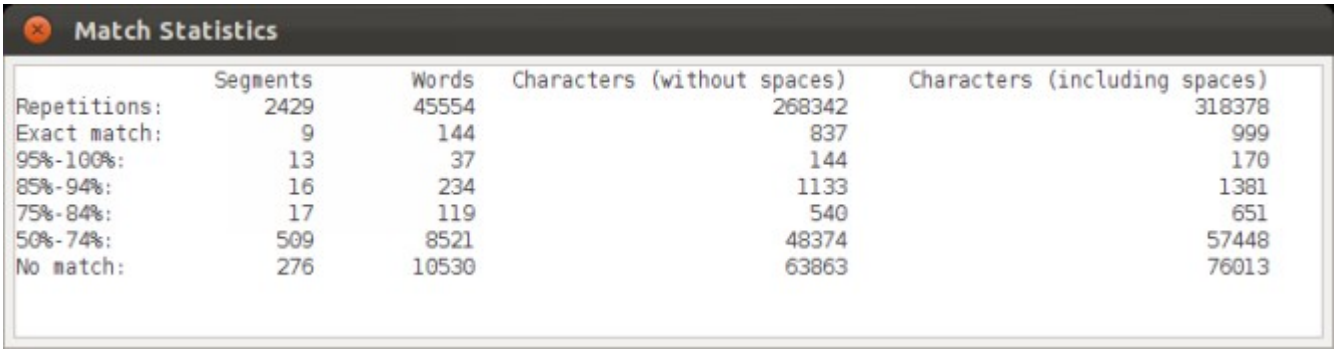
Antes de empezar a trabajar con un proyecto de traducción es muy importante conocer en detalle el trabajo que comportará. Cuando se trabaja sin la ayuda de programas de traducción asistida habitualmente se cuentan palabras o caracteres de los archivos a traducir. Estos recuentos sirven también para presupuestar o facturar a nuestro cliente.

Cuando trabajamos con sistemas de traducción asistida hay que tener en cuenta también las repeticiones internas del proyecto y los segmentos que se recuperarán de la memoria de traducción. Esta información conviene tenerla también para diferentes márgenes de similitud, ya que no es lo mismo una repetición exacta que una al 75%. La mayoría de sistemas de traducción asistida cuentan con funciones de análisis de proyectos.

En la siguiente imagen podemos observar un análisis de un proyecto realizado con la opción *Project statistics* de OmegaT:



Si queremos también disponer de las estadísticas de coincidencias con las memorias de traducción usaremos la opción Tools> Match Statistics, que nos ofrecerá la siguiente información:



Esta información se puede utilizar para presupuestar o facturar proyectos de traducción: se pueden cobrar tarifas diferentes para las palabras correspondientes a segmentos nuevos que hay que traducir desde cero,

otras tarifas para las coincidencias exactas provenientes de memorias de traducción o de repeticiones internas, y tarifas diferentes según los grados de similitud. Lo que no es recomendable es no cobrar nada por las coincidencias exactas, ya que de hecho llevan también un trabajo de verificación de si la traducción propuesta es la más adecuada en el nuevo contexto.

2.8. Nuevas funcionalidades

En este apartado explicaremos algunas de las funcionalidades interesantes que se están desarrollando dentro del proyecto CASMACAT (Alabau, 2013) y que seguro que en un futuro próximo veremos implementadas en los sistemas habituales del mercado.

2.8.1.Sistema de traducción automática integrado

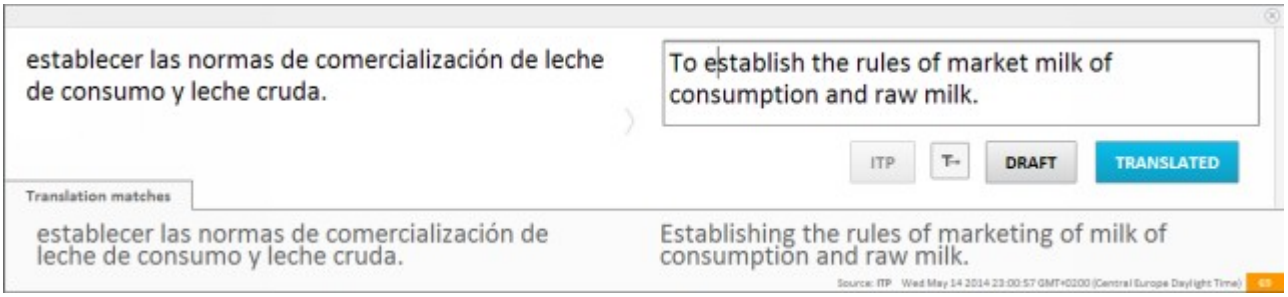
Muchos sistemas de traducción asistida disponen de algún tipo de conexión a sistemas de traducción automática externos que permiten recuperar una traducción automática del segmento que se está traduciendo en ese momento. La conexión no va más allá de esta recuperación, y el sistema de traducción automática por regla general no sabe qué hace el usuario con la traducción ni es capaz de darle otras traducciones alternativas.

CASMACAT integra un sistema de traducción automática estadística. Esta integración implica por una parte que puede proporcionar al usuario más de una propuesta de traducción, y por otra parte, que el sistema de traducción automática puede aprender sabiendo si una propuesta ha sido aceptada íntegramente o bien ha sido modificada. Esta integración también permite la funcionalidad que presentamos a continuación, el *autocompletado inteligente.*

2.8.2.Autocompletado inteligente

Muchos procesadores de textos disponen de la función de autocompletado que tienen en cuenta las palabras más probables que pueden seguir a las palabras que estamos escribiendo y que muestra una vez escribimos unas pocas letras. Los sistemas de traducción asistida pueden disponer también de la misma funcionalidad, pero además pueden tener más evidencias de lo que queremos escribir ya que saben el original que estamos traduciendo y disponen de evidencias provenientes de la memoria de traducción y del sistema de traducción automática .

Si empezamos a escribir "se ..." justo detrás de "To" el sistema autocompletará con la continuación más probable dada tanto por las memorias de traducción como del sistema de traducción automática estadística.



2.8.3. Medidas de confianza

Estas medidas implementadas en CASMACAT informan al usuario sobre qué partes de la traducción tienen más probabilidad de ser incorrectas. Por un lado se marcan en rojo las palabras que tienen mucha probabilidad de ser incorrectas y por el otro lado se marcan en naranja las palabras dudosas para el sistema.

2.9. Conclusiones

En este capítulo hemos presentado en detalle uno de los recursos principales de las herramientas de traducción asistida: las memorias de traducción. Se ha analizado el proceso de recuperación de segmentos similares y el cálculo de la similitud entre el segmento que estamos traduciendo y los recuperados de la memoria.

Las herramientas de traducción asistida, por regla general, trabajan con poco conocimiento lingüístico específico de una determinada lengua. El objetivo es que la herramienta pueda funcionar para un gran abanico de lenguas. Hemos visto como la inclusión de información específica y la capacidad de hacer un análisis lingüístico (proceso que es dependiente de la lengua) puede mejorar el uso de las memorias de traducción. Esta mejora se puede obtener tanto en el proceso de recuperación de segmentos similares, como en la combinación de unidades subsegmentales.

Las memorias de traducción son un recurso que se puede combinar con los sistemas de traducción automática (que veremos a fondo en el capítulo 4). Por ahora, la mayoría de herramientas limitan esta combinación a hacer una propuesta proveniente de un sistema de traducción automática si no se encuentra ningún segmento similar a la memoria de traducción. Hemos visto también en este capítulo como esta combinación puede ir mucho más allá y puede pasar por la retroalimentación de los sistemas de traducción automática con los nuevos segmentos de la memoria, el apoyo del sistema de traducción automática para la combinación de unidades subsegmentals y hasta a la posibilidad de hacer un autocompletado inteligente.

Hoy en día las memorias de traducción son un recurso de gran utilidad para el traductor y es previsible que en un futuro muy próximo lo sean aún más.

No todos los textos son igualmente adecuados para el uso de memorias de traducción. Sin embargo, el uso de memorias de traducción puede ser de utilidad incluso para traducción de textos donde prácticamente no hay repeticiones, como puede ser la traducción literaria. Aunque no es previsible que el sistema nos proporcione prácticamente ninguna coincidencia, podremos hacer búsquedas de expresiones y ver cómo han sido traducidas con anterioridad.

2.10. Para ampliar conocimientos

2.10.1. Corpus paralelos y memorias de traducción disponibles públicamente

Cuando empezamos a trabajar con sistemas de traducción asistida y no disponemos de ninguna memoria de traducción el sistema sólo nos podrá ofrecer propuestas de las llamadas *repeticiones internas*, es decir, segmentos similares que hemos traducido anteriormente en el proyecto de traducción. Esta situación es transitoria, ya que cuando llevamos unos meses trabajando con nuestro sistema de traducción asistida ya dispondremos de un buen volumen de segmentos originales y traducidos en nuestras memorias.

En Internet se pueden encontrar una buena cantidad de corpus paralelos o memorias de traducción disponibles para la descarga. Si alguna de estas memorias son de nuestro ámbito de trabajo las podemos incorporar al sistema de traducción asistida. La mayoría de estas memorias provienen de traducciones oficiales de instituciones multilingües como la Unión Europea, o bien de proyectos de localización de programas de software libre.

Un buen repositorio para empezar a buscar es OPUS (http://opus.lingfil.uu.se/) (Tiedemann 2012). En el momento de escribir este capítulo esta colección estaba compuesta por los siguientes 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)

Recientemente también se ha incorporado a esta colección un corpus paralelo catalán-castellano proveniente de los textos del Diari Oficial de la Generalitat de Catalunya (DOGC).

Otra fuente muy interesante para encontrar corpus es consultar Meta-share (http://metashare.upf.edu/) .

2.10.2. Etiquetadores morfosintácticos

Un *etiquetador morfosintáctico* es un programa informático capaz de etiquetar todas las palabras de un texto con información morfosintáctica. Esta información se da mediante unas etiquetas que expresan la categoría gramatical y una serie de información adicional. En la siguiente figura se puede observar el etiquetado morfosintáctico de la oración castellana *Yo bajo con el hombre bajo a tocar el bajo bajo la escalera* llevada a cabo por el analizador 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

Fijémonos en que el analizador es capaz, al menos en algunos casos, de etiquetar correctamente las palabras aunque éstas sea ambiguas. La palabra *bajo* en esta oración puede ser un verbo (VMIP1S0) un adjetivo (AQ0MS0), un sustantivo (NCMS000) y una preposición (SPS00). La categoría gramatical y las correspondientes subcategoritzacions expresan mediante etiquetas.

Freeling (http://nlp.lsi.upc.edu/freeling) es un analizador lingüístico de código libre desarrollado en la Universitat Politècnica de Catalunya. Puede hacer análisis a varios niveles:

Análisis morfológico *(morphological analysis):* este análisis no lleva a cabo desambiguación y ofrece toda la información posible a cada palabra, sea la correcta por el contexto o no.
Análisis morfosintáctico *(PoS tagging):* pone a cada palabra el lema y la etiqueta que le corresponde
Análisis sintáctico superficial *(shallow parsing):* realiza un análisis sintáctico del texto en la que algunas relaciones pueden no estar presentes.
Análisis sintáctico completo *(full parsing):* hace el análisis sintáctico completo
Análisis de dependencias *(dependency parsing):* un tipo de análisis que marca las dependencias entre las palabras
Análisis semántico: etiqueta los textos con *synsets* de WordNet y es capaz de realizar desambiguación de sentidos

Freeling funciona para muchas lenguas aunque no todas las lenguas disponen de todos los niveles de análisis descritos:

- asturiano
- catalán
- inglés
- francés
- gallego
- portugués
- castellano
- ruso
- galés

Freeling se puede instalar tanto en Linux como en Windows y también dispone de una demo on-line que permite hacerse una idea de sus capacidades.

Otro etiquetador morfosintáctico disponible es el Tree Tagger (Schmid, 1994) (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/) que no tiene una licencia libre pero que puede usarse para fines de investigación, evaluación y educación. Tiene también versiones para Linux y Windows y dispone de modelos para muchas lenguas. A diferencia de Freeling, Tree Tagger sólo ofrece el análisis morfosintáctico y para ciertas lenguas también *chunking* (*análisis fragmental*)

2.9.3. Propiedad de las memorias de traducción

Las memorias de traducción plantean un problema jurídico aún no resuelto del todo respecto a su propiedad intelectual. La memoria se compone de un original (que puede tener sus propios derechos de autor) y una traducción (que genera unos derechos de traducción). La traducción suele estar encargada por un cliente y muchas veces a través de una agencia. En medio se ha generado una traducción que puede estar en manos del traductor, agencia y cliente final. ¿Quién tiene derecho de volver a utilizar esta memoria? ¿El traductor? Pero, ¿sólo para el mismo cliente final? ¿Para todos los clientes? ¿Quién tiene derecho a ceder esta memoria a otros usuarios? ¿Todas las legislaciones opinan lo mismo, o hay diferencias entre los países?

En este apartado no aclararemos estas dudas, pero sí propondremos algunas lecturas que pueden arrojar algo de luz sobre el tema.

Jorge Marcos (2001) *Un enfoque jurídico de las memorias de traducción. Revista Tradumática núm. 0.* http://www.fti.uab.es/tradumatica/revista/num0/articles/jmarcos/art.htm

La presentación *Copyright protection for translation memories* que se puede encontrar en el siguiente enlace: 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)

Bibliografía

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