

## 4. La traducción automática

### Índice

|   |    |
|---|----|
| 4.1. Introducción.....  | 1  |
| 4.2. Historia de la traducción automática.....                            | 3  |
| 4.3. Tipos de sistemas de traducción automática.....                      | 6  |
| 4.3.1. Traducción directa.....  | 7  |
| 4.3.2. Sistemas de transferencia.....                                     | 8  |
| 4.3.3. Sistemas de interlingua.....                                       | 9  |
| 4.3.4. Sistemas estadísticos.....   | 13 |
| $\operatorname{argmax} p(e f) = \operatorname{arg} \max p(f e)p(e)$ ..... | 14 |
| 4.3.5. Sistemas de traducción automática basada en ejemplos.....          | 19 |
| Ano akai kasa wa ikura desu ka.....                                       | 20 |
| Ano chiisai kamera wa ikura desu ka.....                                  | 20 |
| 4.3.6. Sistemas híbridos.....   | 20 |
| 4.4. Usos de la traducción automática.....                                | 22 |
| 4.5. La traducción automática en Internet.....                            | 25 |
| 4.6. Combinación de traducción automática y traducción asistida.....      | 28 |
| 4.7. Post-edición de traducción automática.....                           | 29 |
| 4.8. Conclusiones.....  | 31 |
| 4.9. Para ampliar conocimientos.....                                      | 32 |
| 4.9.1. Historia de la traducción automática.....                          | 32 |
| 4.9.2. Moses: un sistema de traducción automática estadística.....        | 32 |
| 4.9.3. Otros sistemas de traducción automática.....                       | 33 |
| Bibliografía.....   | 33 |
| Licencia de este documento.....   | 34 |

### 4.1. Introducción

En este capítulo ofrecemos una panorámica general sobre los sistemas de traducción automática orientada a las necesidades de los traductores humanos. Durante las pasadas décadas los traductores a menudo han tenido la percepción de que la traducción automática era una amenaza para su profesión. Se pensaba que en un futuro próximo los sistemas de traducción automática conseguirían un nivel de calidad suficiente como para hacer innecesaria la intervención humana en el proceso de traducción. Esta percepción exageradamente optimista no era exclusiva de los profesionales de la traducción, sino que otros sectores de la sociedad también han creído que el traductor humano era una especie en extinción.

Aunque es cierto que la calidad de los sistemas de traducción automática ha mejorado muchísimo en los últimos años, y dado que estos sistemas están actualmente al alcance de cualquier usuario, el público general es más consciente de los límites de este tipo de sistemas. Este mejor conocimiento de los sistemas de traducción automática hace que se valore más la tarea del traductor humano y que se tenga conciencia de qué tipos de trabajos de traducción requieren contar con un profesional de la traducción.

La percepción que tienen los traductores humanos también está cambiando desde una amenaza para la profesión hacia una herramienta de trabajo que puede ser de utilidad en algunas situaciones (pares de lenguas, tipos de texto y nivel de calidad requerido concretos). En este sentido, la práctica totalidad de los sistemas de traducción asistida permiten una conexión con sistemas de traducción automática. De esta manera, cuando la herramienta de traducción asistida no encuentra ninguna coincidencia con un índice de similitud suficiente en la memoria de traducción, se presenta el resultado de traducir el segmento actual

con un sistema de traducción automática. El traductor podrá aceptar, editar o rechazar esta propuesta automática.

Para algunos pares de lenguas y tipo de texto es habitual trabajar con post-edición de textos traducidos automáticamente. En algunas situaciones esta estrategia permite alcanzar unos niveles de productividad muy elevados.

Para que un traductor pueda trabajar de manera efectiva con un sistema de traducción automática es importante que conozca a fondo la tecnología que hay detrás de estos sistemas. De esta manera el traductor podrá decidir en cada situación (par de lenguas, tipos de texto y calidad final requerida) qué uso hace de los sistemas de traducción automática. Este uso se puede concretar en tres:

- Ningún uso, es decir, no usar ningún sistema de traducción automática.
- Combinación de traducción asistida y traducción automática dentro de un sistema de traducción asistida por ordenador.
- Post-edición de traducción automática, es decir, traducir automáticamente todo el texto y post-editarlo posteriormente.

Este capítulo está dividido en seis apartados, En primer lugar hablaremos de la historia de la traducción automática. Esta historia es interesante debido a que se pasa de un gran optimismo inicial a una situación más realista a partir del informe ALPAC, que marca las pautas de la investigación posterior, que a largo plazo dio sus frutos. Después veremos las diferentes estrategias que se utilizan en los sistemas de traducción automática: directa, transferencia, interlingua, estadística y basada en ejemplos, además de los sistemas híbridos, que combinan varias de las estrategias anteriores. Para un profesional de la traducción que utiliza un sistema de traducción automática es importante conocer la estrategia en que se basa el sistema que está utilizando. De esta manera podrá revisar con más atención aquellos aspectos en los que el sistema empleado tenga más probabilidades de producir errores.

Seguirá el capítulo explicando los diferentes usos que pueden tener los sistemas de traducción automática, que van desde tener una idea de lo que habla un texto, hasta producir una traducción con una calidad suficiente para ser publicada directamente. Después analizaremos los sistemas de traducción automática que están disponibles en Internet y las diferentes maneras de acceder a estos sistemas. Nos fijaremos especialmente en las posibilidades de acceso directo y automático desde alguna aplicación de traducción asistida. Veremos también los diferentes niveles posibles de combinación de traducción automática y asistida, que va desde el caso más simple, que consiste en presentar el resultado de la traducción automática sólo cuando no se encuentra nada significativo en la memoria de traducción; hasta la combinación de segmentos provenientes de la memoria de traducción con segmentos traducidos automáticamente.

Por último hablaremos de la tarea de post-edición de traducción automática, en la que se fijará el traductor-post-editor y qué herramientas y estrategias puede utilizar para llevar a cabo esta tarea de una manera más efectiva.

En el apartado *Para ampliar conocimientos* presentaremos una lectura complementaria para saber más sobre la historia de la traducción automática. También presentaremos Moses, un *toolkit* para la creación de sistemas de traducción estadísticos. Por último, proporcionaremos un par de enlaces a recopilaciones de sistemas de traducción automática.

## 4.2. Historia de la traducción automática

La traducción automática es una de las primeras tareas computacionales que se intentaron desarrollar una vez estuvieron disponibles los primeros ordenadores. Según señala Trujillo (1999) la historia de la traducción automática ha sido influenciada por la política, la ciencia y la economía de los diferentes periodos de la historia moderna y distingue una serie de períodos en su historia. También Hutchins (2007) hace una secuenciación similar. Podemos distinguir pues, los siguientes períodos en la historia de la traducción automática:

- **Precursores o etapa anterior a la aparición de los ordenadores.** Desde el siglo XVII varios científicos y filósofos han propuesto representaciones del significado independientes de la lengua para superar las barreras lingüísticas. Leibniz y Descartes propusieron una serie de códigos que relacionaban las palabras entre diferentes lenguas. En 1933 el francés Georges Artsrouni y el ruso Petr Smirnov-Troyanski patentaron procedimientos mecánicos para llevar a cabo traducciones.
- **Pioneros o esfuerzos iniciales.** A partir del uso de los ordenadores para descifrar mensajes cifrados durante la segunda guerra mundial se empezaron a utilizar técnicas numéricas para abordar la tarea de traducción automática. Andrew Booth y Warren Weaver en 1946 y 1947 hicieron los primeros intentos de uso de los ordenadores para la traducción. En esta época el término que se utilizaba era el de traducción mecánica (mechanical translation). En 1948 Andrew Booth trabajó en el análisis morfológico para un diccionario mecánico. En 1949 Warren Weaver puso las bases para el tratamiento del problema de la ambigüedad semántica. A partir de este momento comienza la investigación en traducción automática tanto en Estados Unidos como en otros países del mundo. En 1954 se lleva a cabo la primera demostración pública de un sistema de traducción automática ruso-inglés. En 1952 se organizó la primera conferencia sobre traducción automática en el MIT. En la conferencia surgieron una serie de ideas respecto a la pre-edición y post-edición, el uso de micro-glosarios para evitar los problemas de ambigüedad, y algún tipo de análisis de la estructura sintáctica. Se propuso también la organización de una demostración pública para poder atraer fondos para la investigación en traducción automática. De esta manera la primera demostración pública de un sistema de traducción automática se llevó a cabo en 1952. La demostración consistía en la traducción de 49 oraciones bien seleccionadas del ruso al inglés, utilizando un diccionario muy restringido de sólo 250 palabras y 6 reglas gramaticales. El interés científico real no fue muy notable, pero fue suficientemente impresionante para estimular la gran inversión que inició por la investigación en traducción automática, especialmente en Estados Unidos y para inspirar el inicio de esta investigación en otros países, especialmente la URSS.
- **La década de las grandes expectativas y la desilusión (1956/66).** Durante esta década aparecieron muchos grupos de investigación especialmente en los Estados Unidos la Unión Soviética, y en general los métodos de investigación eran una mezcla entre las aproximaciones empíricas y teóricas. A mediados de los años 1960 había grupos de investigación en muchos países europeos ( Hungría, Checoslovaquia, Bulgaria, Bélgica, Alemania, Francia, etc.) y también en China, México y Japón. Muchos de estos grupos tuvieron una vida corta, con la excepción de un grupo de la Universidad de Grenoble en Francia. De hecho este periodo fue importante para la investigación en traducción automática ya que permitió investigar en ámbitos que hoy llamaríamos lingüística computacional o ingeniería del lenguaje.
- **El informe ALPAC (1966).** El exagerado optimismo inicial llega a su fin en los Estados Unidos con este informe, que afirmaba que la traducción automática no era efectiva a nivel de costes. Se elimina la financiación pública en los Estados Unidos para proyectos relacionados con la traducción automática y la investigación en esta área se continúa principalmente fuera de este

país. El informe ofrecía una serie de recomendaciones e indicaba que la investigación en esta área debía centrarse en:

- Métodos prácticos para evaluar las traducciones
  - Medios para acelerar el proceso de traducción humana
  - Evaluación de la calidad y costes de diversas fuentes de traducciones
  - Investigación sobre la utilización de las traducciones, para evitar la producción de traducciones que luego no se leen.
  - Estudio de las fuentes de retrasos en el proceso de traducción
  - Evaluación de la velocidad relativa y costes de varios tipos de traducción asistida por ordenador
  - Adaptación a la traducción de los procesos de edición y producción existentes
  - Estudio de los procesos de traducción en su globalidad.
  - Producción de material de referencia para el traductor, incluyendo la adaptación de glosarios existentes para la consulta automática en sistemas de traducción automática
- 
- **Los años 1970 y los sistemas de traducción automática operativos.** En los Estados Unidos la investigación se centra en la traducción al inglés de textos científicos rusos. En cambio, en Canadá y en Europa las necesidades eran muy diferentes (inglés-francés en Canadá y entre las lenguas de la Comunidad Europea). En Canadá hay que destacar el proyecto TAUM (*Traduction Automatique* de la Université de Montreal), que tiene como resultado el sistema Météo para traducir previsiones del tiempo, que se ha estado utilizando con éxito desde 1976. Los experimentos más innovadores de esta época se centraban en la estrategia de interlingua. Hacia mediados de esta década se empieza a dudar de esta aproximación y se piensa que la aproximación de transferencia ofrecía mejores perspectivas. Durante esta época también se instalan las primeras versiones de Systran y la Comunidad Europea compra una versión inglés-francés en 1976.
  - **El renacimiento de principios de los años 1980.** Hacia finales de los años 1970 y principios de los 1980 creció el interés hacia la traducción automática. En 1982 se inicia el proyecto Eurotra de la Comunidad Europea, el mismo año se inicia el proyecto Mu en Japón. El principal rival de Systran es Logos, desarrollado para traducir manuales de aviones. En 1982 aparece la versión alemán-inglés de Systran y durante los años 1980 van apareciendo otros pares de lenguas. Los primeros sistemas de traducción para ordenadores personales son los sistemas American Weidner (1981) y ALPS (1983).
  - **Los finales de los 1980 y los principios los 1990.** A finales de los 1980 aparece el par alemán-inglés del sistema comercial METAL. Este sistema sigue una estrategia de transferencia y pronto aparecen otros pares de lenguas para el holandés, francés, español así como de inglés y alemán. Empiezan a aparecer sistemas diseñados para lenguajes especializados, como el de la PAHO (Pan American Health Organization). Las grandes compañías de electrónica japonesas, tales como Fujitsu, Hitachi, NEC, Sharp, Toshiba empiezan a comercializar programas de ayuda a la traducción, especialmente para japonés-inglés e inglés-japonés. En este período también comienza a trabajar en traducción de voz y en la traducción automática con aproximación estadística.
  - **Los finales de los 1990 y MAT (*Machine Aided Translation*).** Aparecen sistemas de traducción potentes para ordenadores personales, la traducción automática en Internet y se generaliza mucho el uso de memorias de traducción y programas de ayuda para traductores. Crece el interés por la traducción automática basada en ejemplos.

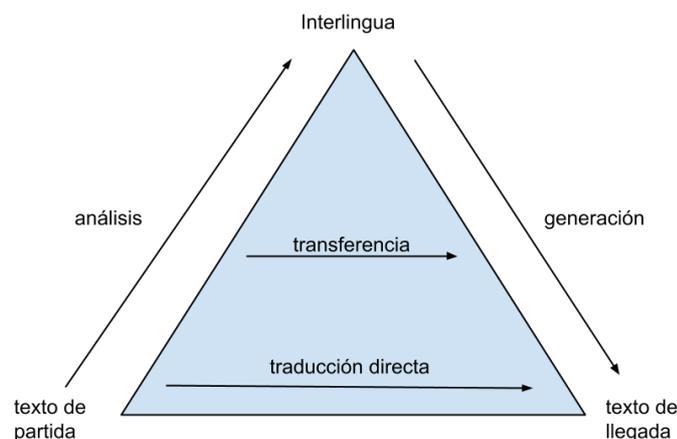
Actualmente hay disponibles en Internet sistemas de traducción automática que ofrecen buena calidad (como por ejemplo Google Translate (<https://translate.google.com/>) o Microsoft Bing Translation (<http://www.bing.com/translator/>)). Estos sistemas, además de poderse consultar por Internet, ofrecen una API para una conexión directa con sistemas de traducción asistida (algunos lo ofrecen de forma gratuita pero con un volumen de texto determinado, o bien como servicio de pago) . También se populariza muchísimo el uso de sistemas de traducción automática entre los traductores y aparecen sistemas TAO completos gratuitos y con una licencia libre (por ejemplo OmegaT - <http://www.omegat.org/> ). También se popularizan conjuntos de herramientas que permiten entrenar tus propios sistemas de traducción automática estadística, como por ejemplo Moses (<http://www.statmt.org/moses/>).

### 4.3. Tipos de sistemas de traducción automática

Tradicionalmente los sistemas de traducción automática se han dividido en tres tipos:

- **Traducción directa:** en estos sistemas la traducción se lleva a cabo de una manera directa a partir de la consulta en diccionarios bilingües y con unas pocas reglas que permiten hacer algunos cambios, como por ejemplo el cambio del orden de las palabras. Los primeros sistemas de traducción automática funcionaban de esta manera y todavía hay sistemas comerciales que se basan en esta estrategia.
- **Transferencia:** estos sistemas realizan un análisis del texto de entrada que da como resultado una estructura que se transfiere a partir de unas reglas a una estructura propia de la lengua de salida. A partir de la estructura de salida se genera la oración en la lengua de llegada. Normalmente estos sistemas realizan un análisis sintáctico, que puede ser completo o superficial y por ello se habla de **sistemas de transferencia sintáctica**.
- **Interlingua:** en esta estrategia las oraciones en la lengua de partida se analizan para obtener una representación independiente de la lengua. A partir de esta representación independiente se genera la oración en la lengua de llegada.

Para representar las diferencias entre estas tres estrategias a menudo se ha utilizado el llamado *triángulo de Vauquois*, que recibe el nombre del matemático e informático francés Bernard Vauquois (1929-1985). En la siguiente imagen podemos observar este triángulo:



En el eje vertical se representa el esfuerzo de análisis y generación necesario para cada estrategia y en el eje horizontal el esfuerzo en la transferencia. En la estrategia de traducción directa el esfuerzo de análisis y generación es muy bajo, ya que el análisis y generación se limitan a la morfología. En cambio el esfuerzo de transferencia en esta estrategia es muy elevado. Para la estrategia de transferencia los esfuerzos en el análisis, generación y transferencia son medios. En el extremo, en la estrategia de interlingua los esfuerzos de análisis y generación son máximos, mientras que el esfuerzo en la transferencia es nulo, ya que la representación es independiente de la lengua y coincide tanto por el texto de partida como por el texto de llegada.

Aparte de las estrategias presentadas hasta ahora y representadas en el triángulo de Vauquois cabe mencionar dos más:

- **Sistemas estadísticos:** en esta estrategia las traducciones se generan a partir de modelos estadísticos. Los parámetros de estos modelos estadísticos se calculan a partir de corpus textuales bilingües. Estos sistemas, pues, pueden *aprender* a traducir a partir de traducciones ya realizadas.
- **Sistemas basados en ejemplos:** esta estrategia también utiliza corpus textuales bilingües para deducir cómo traducir nuevas oraciones. En este caso no se generan modelos estadísticos, sino que la estrategia se basa en la traducción por analogía. El sistema busca oraciones ya traducidas que puedan servir como ejemplos para traducir las nuevas oraciones.

Estas son las estrategias principales para la construcción de sistemas de traducción automática. Estas estrategias se pueden combinar para producir sistemas **híbridos**. Por ejemplo, un sistema de transferencia sintáctica se puede combinar con un sistema estadístico. Cuando el análisis sintáctico no se ha podido producir satisfactoriamente porque la gramática no es capaz de analizar la oración, o bien el sistema no dispone de reglas de transferencia para llevarla a cabo, se puede dar paso a solucionar los fragmentos necesarios mediante el modelo estadístico. Algunos sistemas híbridos funcionan tomando la salida de varios sistemas de traducción automática e intentan determinar cuál de las salidas es de más calidad para tomar ésta como la resultante, o bien seleccionar fragmentos con buenos índices de calidad para combinarlos. En estos casos, la dificultad radica en saber determinar cuál de las salidas o fragmentos son los de mejor calidad.

#### 4.3.1. Traducción directa

En los sistemas de traducción directa la traducción se basa en la consulta a diccionarios bilingües para determinar la traducción de las palabras o expresiones multipalabra del texto de partida. Los procesos de análisis y generación suelen limitarse al análisis morfológico y la lematización. La lematización permitirá realizar la búsqueda más fácilmente el diccionario bilingüe. Una vez se genera la oración de llegada se aplican una serie de reglas para poder tratar algunos fenómenos, como por ejemplo, el cambio de orden de las palabras. Podemos ver el proceso de traducción directa con el siguiente ejemplo:

Queremos traducir la oración castellana:

*El niño come un helado grande.*

El análisis morfológico y lematización podría dar el siguiente resultado:

|                        |                            |                             |                        |                                |                                |                      |
|------------------------|----------------------------|-----------------------------|------------------------|--------------------------------|--------------------------------|----------------------|
| <b>El</b><br><i>el</i> | <b>niño</b><br><i>niño</i> | <b>come</b><br><i>comer</i> | <b>un</b><br><i>un</i> | <b>helado</b><br><i>helado</i> | <b>grande</b><br><i>grande</i> | <b>.</b><br><i>.</i> |
|------------------------|----------------------------|-----------------------------|------------------------|--------------------------------|--------------------------------|----------------------|

|        |         |         |        |         |        |    |
|--------|---------|---------|--------|---------|--------|----|
| DA0MS0 | NCMS000 | VMIP3S0 | DI0MS0 | NCMS000 | AQ0CS0 | Fp |
|--------|---------|---------|--------|---------|--------|----|

Esto nos permitiría consultar el diccionario bilingüe (a partir de los lemas) y obtener las traducciones de cada una de las palabras:

|     |      |       |    |           |        |   |
|-----|------|-------|----|-----------|--------|---|
| el  | niño | comer | un | helado    | grande | . |
| the | boy  | eat   | an | ice cream | big    | . |

Las etiquetas morfosintácticas del análisis nos permitirán generar la forma correcta del verbo *eat*, que al ser 3ª persona singular del presente de indicativo pasa a *eats*.

Ahora, en el conjunto de reglas tenemos una que dice:

NC \* AQ \* -> AQ \* NC \*

Que afecta a *icecream big* y que hace que se cambie el orden para *big ice cream*, resultando en la frase traducida

*The boy eats a big ice cream.*

La explicación expuesta con este ejemplo es una simplificación. Quedan muchos aspectos a tratar como por ejemplo la ambigüedad, que hace que una palabra del texto de partida se pueda traducir por más de una palabra del texto de llegada, así como el tratamiento de fenómenos lingüísticos más complejos.

Esta estrategia es la que empleaban los primeros sistemas de traducción automática y todavía se utiliza en algunos sistemas comerciales, como por ejemplo:

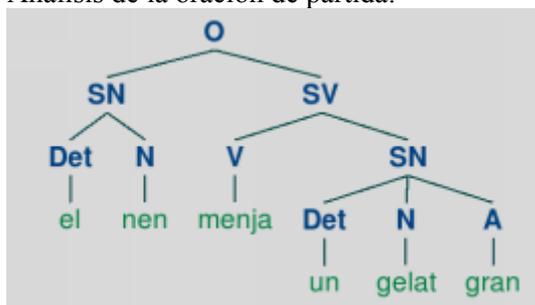
- Systran ( <http://www.systransoft.com/> )
- Logos ( <http://logos-os.dfki.de/> ).Este sistema se ha liberado recientemente y ha pasado a ser Open Logos.

#### 4.3.2. Sistemas de transferencia

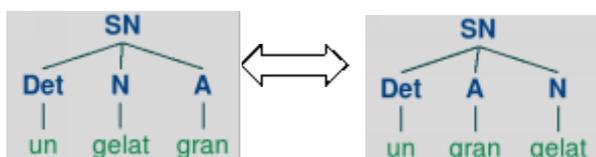
En los sistemas de transferencia se genera una representación de la oración de partida, generalmente en forma de análisis sintáctico, aunque también puede incluir información semántica. Esta representación se transforma en una representación en la lengua de llegada a partir de un conjunto de reglas. Una vez se tiene la representación en la lengua de llegada se hace la transferencia léxica, es decir, se traducen las palabras mediante un diccionario. Por último, se generarán las formas adecuadas de las palabras en la lengua de llegada.

Si continuamos con el mismo ejemplo del apartado anterior, donde queremos traducir la oración castellana *El niño come un helado grande.* al inglés se llevarán a cabo los siguientes pasos (consideramos que la transferencia es sintáctica):

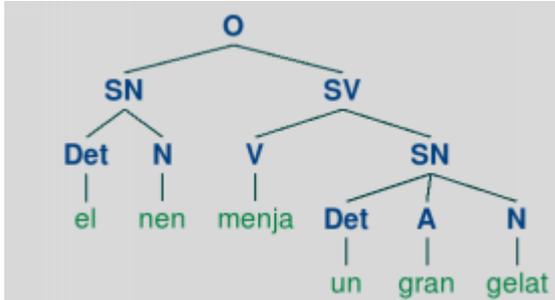
Análisis de la oración de partida:



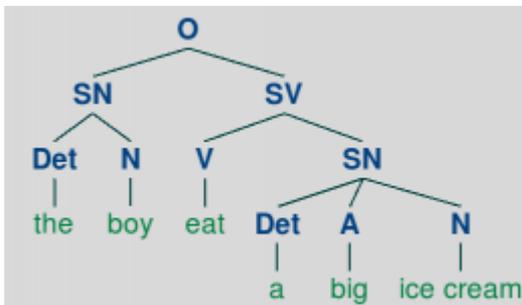
Las reglas de transferencia se referirán a partes del análisis y harán por ejemplo:



Y dado que no se puede aplicar ninguna otra transformación, la representación de la oración de llegada queda de la siguiente manera:



Que una vez se hace la transferencia léxica del árbol queda:



Y que al hacer la generación morfológica se pone el verbo (*eat*) en tercera persona singular (*eats*) resultando en la frase traducida:

*The boy eats a big ice cream.*

De nuevo esto es una simplificación del funcionamiento de este tipo de sistemas, pero expresa los fundamentos básicos de su funcionamiento.

#### 4.3.3. Sistemas de interlingua

En esta estrategia el texto original se transforma en una representación abstracta, llamada interlingua, que es independiente de la lengua y el texto traducido se genera directamente a partir de esta generación abstracta. Esta estrategia tiene una serie de ventajas, entre las que se pueden destacar:

- Necesita menos componente para relacionar cada lengua de partida con cada lengua de llegada.
- Se requieren menos componentes para añadir una nueva lengua.
- El componentes de análisis y generación pueden ser desarrollados con conocimiento de una sola lengua.
- Se pueden desarrollar sistemas de traducción automática para pares de lenguas muy diferentes (por ejemplo inglés y árabe)

Y tiene un inconveniente principal e importante:

- La dificultad de definir la representación abstracta independiente de la lengua, la interlingua. La dificultad es aún mayor si se pretende desarrollar sistemas para dominios no restringidos o amplios.

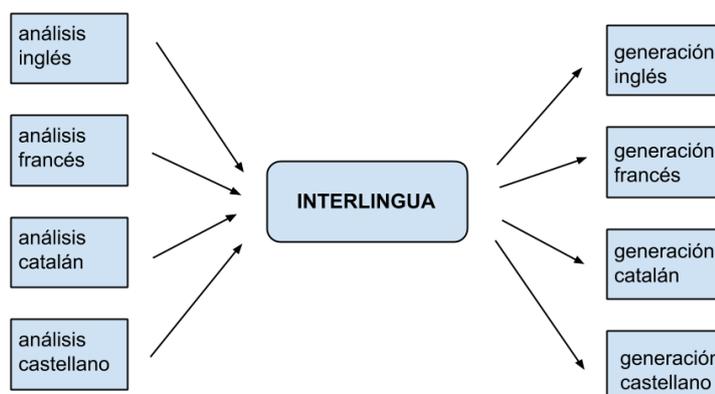
Vemos ahora con más detalle el hecho de que en los sistemas de Interlingua se necesitan menos componentes para desarrollar sistemas de traducción automática para varios pares de lenguas. Supongamos que queremos desarrollar un sistema de traducción automática para los siguientes pares de lengua (y en ambas direcciones): inglés, francés, catalán y castellano. Es decir, disponemos de los siguientes pares de lenguas:

- inglés-francés
- francés-inglés
- inglés-catalán
- catalán-inglés
- inglés-castellano
- castellano-inglés
- francés-catalán
- catalán-francés
- catalán-castellano
- castellano-catalán

En los sistemas interlingua, si consideramos que los módulos de análisis y generación no son reversibles, necesitaríamos los siguientes módulos:

- Sistema de análisis para el inglés
- Sistema de generación para el inglés
- Sistema de análisis para el francés
- Sistema de generación para el francés
- Sistema de análisis para el catalán
- Sistema de generación para el catalán
- Sistema de análisis para el castellano
- Sistema de generación para el castellano

Es decir, necesitamos un total de 8 módulos. Estos módulos se pueden observar en la siguiente figura:

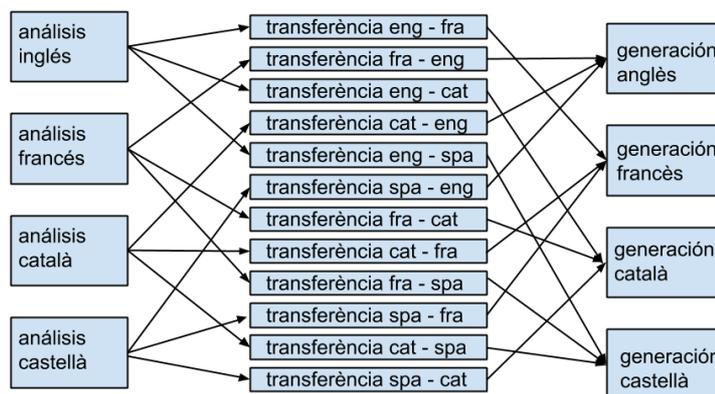


Si queremos añadir una nueva lengua, por ejemplo, el italiano, sólo será necesario añadir dos módulos más, el módulo de análisis del italiano y el módulo de generación del italiano y tendríamos un total de 10. En general, en los sistemas de interlingua, si queremos tener un sistema de traducción automática para  $n$  lenguas necesitaremos un total de  $2n$  módulos (considerando que los módulos de análisis y generación no son reversibles).

La situación para los sistemas de transferencia es muy diferente. Siguiendo el mismo ejemplo, para disponer del sistema de traducción automática necesitaríamos:

- Sistema de análisis para el inglés
- Sistema de generación para el inglés
- Sistema de análisis para el francés
- Sistema de generación para el francés
- Sistema de análisis para el catalán
- Sistema de generación para el catalán
- Sistema de análisis para el castellano
- Sistema de generación para el castellano
- Sistema de transferencia inglés-francés
- Sistema de transferencia francés-inglés
- Sistema de transferencia inglés-catalán
- Sistema de transferencia catalán-inglés
- Sistema de transferencia inglés-castellano
- Sistema de transferencia castellano-inglés
- Sistema de transferencia francés-catalán
- Sistema de transferencia catalán-francés
- Sistema de transferencia francés-castellano
- Sistema de transferencia castellano-francés
- Sistema de transferencia catalán-castellano
- Sistema de transferencia castellano-catalán

Es decir, para  $n$  lenguas necesitamos un total de  $n$  sistemas de análisis,  $n$  de generación y  $n(n-1)$  de transferencia, lo que hace un total de  $2n + n(n-1)$ , es decir,  $n(n+1)$  o lo que es lo mismo,  $n^2 + n$  módulos, que en nuestro ejemplo es de 20 módulos. Esta situación la podemos observar en la siguiente figura:



Si ahora queremos añadir una lengua más, el italiano por ejemplo, tendremos que añadir 1 módulo de análisis, 1 de generación y 8 de transferencia (inglés-italiano, italiano-inglés, francés-italiano, italiano-

francés, catalán- italiano, italiano-catalán, castellano-italiano e italiano-castellano), pasando el sistema a tener 30 módulos.

Si la tecnología que utilizamos permite que los módulos de análisis, generación y transferencia sean reversibles el número de módulos necesarios, tanto para la interlingua como para la transferencia se reduce notablemente, quedando de la siguiente manera:

Interlingua (inglés, francés, catalán y castellano):

- Sistema de análisis y generación para el inglés
- Sistema de análisis y generación para el francés
- Sistema de análisis y generación para el catalán
- Sistema de análisis y generación para el castellano

Es decir, un total de 4 módulos, es decir, de  $n$  módulos. Si queremos añadir una nueva lengua, sólo será necesario añadir un nuevo módulo.

Transferencia (inglés, francés, catalán y castellano):

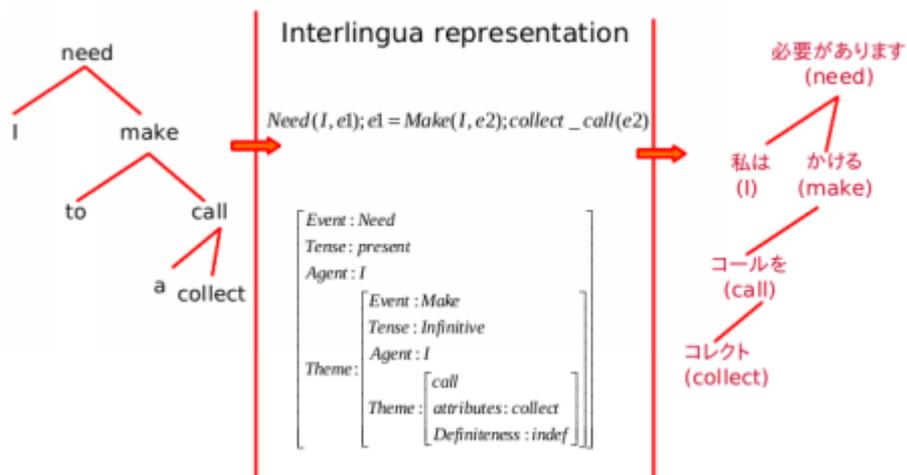
- Sistema de análisis y generación para el inglés
- Sistema de análisis y generación para el francés
- Sistema de análisis y generación para el catalán
- Sistema de análisis y generación para el castellano
- Sistema de transferencia inglés ↔ francés
- Sistema de transferencia inglés ↔ catalán
- Sistema de transferencia inglés ↔ castellano
- Sistema de transferencia francés ↔ catalán
- Sistema de transferencia francés ↔ castellano
- Sistema de transferencia catalán ↔ castellano

Es decir, un total de 10 módulos (4 de análisis-generación y 6 de transferencia). Si ahora añadimos una nueva lengua se añadiría un nuevo sistema de análisis-generación y 4 de transferencia, haciendo un total de 15. Es decir, necesitamos  $n$  sistemas de análisis-generación y  $\frac{n(n-1)}{2}$  sistemas de transferencia, o

lo que es lo mismo, un total de  $\frac{(n^2+n)}{2}$  módulos.

A pesar de esta ventaja respecto al número de módulos necesarios, los sistemas Interlingua enfrentan a la gran dificultad que supone el diseño de esta representación intermedia independiente de la lengua, es decir, la definición de la propia interlingua. Además, es mucho más complejo diseñar los sistemas de análisis hacia esta interlingua y los de generación desde el interlingua hasta la lengua de llegada.

Es difícil encontrar sistemas interlingua operativos y disponibles de forma abierta y por tanto no podemos mostrar un ejemplo real de traducción con estos tipos de sistemas. Sin embargo, presentaremos un ejemplo extraído de los materiales docentes de los profesores Dr. Srinivas Bangalore y Juan Carlos Niebla<sup>1</sup>.



Si queremos traducir la frase inglesa *I need to make a collect call* (Necesito hacer una llamada a cobro revertido) al japonés, lo primero que se hace es analizar la oración inglesa para llegar a la representación interlingua que vemos en la parte central de la imagen. A partir de esta representación se podrá generar la oración japonesa.

Algunos sistemas de interlingua, como el descrito en Schubert (1988) usan el esperanto como representación intermedia, ya que es una lengua con una sintaxis extremadamente regular y una semántica muy clara. No hay que confundir esta estrategia de usar otra lengua como interlingua, con el hecho de utilizar una lengua como *pivote* o *lengua puente*. Las lenguas pivote se pueden utilizar en cualquiera de las estrategias de traducción automática y consiste en traducir de una lengua A a una B a través de una lengua P. La ventaja de utilizar lenguas pivote es que se requieren la construcción de menos módulos, pero por el contrario en cada retraducción se van introduciendo errores que perjudican a la calidad final.

#### 4.3.4. Sistemas estadísticos

La *traducción automática estadística* (SMT- *Statistical Machine Translation*) es una estrategia de traducción automática que se basa en el uso de modelos estadísticos con modelos obtenidos a partir del análisis de corpus bilingües. Es decir, el sistema es capaz de traducir a partir de una serie de parámetros que han sido calculados a partir de una gran cantidad de textos originales y sus traducciones.

Aunque las primeras ideas datan de 1949, esta estrategia no se desarrolla plenamente hasta principios de los años 1990.

<sup>1</sup> Corresponsales al curso COS401/TRA301 - *Introduction to Machine Translation* de la Universidad de Princeton ([http://www.cs.princeton.edu/courses/archive/spring09/cos401/slides/Interlingua-based\\_MT.ppt](http://www.cs.princeton.edu/courses/archive/spring09/cos401/slides/Interlingua-based_MT.ppt))

En general, los sistemas de traducción automática estadística trabajan con dos probabilidades:

- La probabilidad de que una palabra o conjunto de palabras de la lengua de partida (SL- *source language*) se traduzca por una palabra o conjunto de palabras de la lengua de llegada (TL- *target language*).
- La probabilidad de que una cadena en la lengua de llegada sea una oración válida en esta lengua.

Veremos la notación que se utiliza habitualmente, que está influenciada por los primeros ejemplos que consideraban un sistema de traducción automática entre el francés y el inglés. Lo que nos interesa es calcular la probabilidad de que una frase en la lengua de llegada (*e*: ya que se considera el inglés como lengua de llegada) sea la traducción de una frase en la lengua de partida (*f*: ya que se considera el francés como lengua de partida), esto se escribe como  $p(e|f)$ . Esta misma probabilidad se puede calcular como:

$$p(e|f) = p(f|e) p(e)$$

Dónde:

- $p(f|e)$  es el *modelo de traducción* que expresa la probabilidad de que la cadena en la lengua de partida sea la traducción de la cadena en la lengua de llegada.
- $p(e)$  es el *modelo de lengua* y expresa la probabilidad de que la cadena en la lengua de llegada sea una cadena válida en esta lengua.

Esta descomposición es interesante ya que divide el problema en dos subproblemas. La cantidad de posibles traducciones que se pueden deducir de los parámetros del sistema estadístico es enorme, y lo que se tendrá que hacer será encontrar la traducción que tenga una mayor probabilidad, es decir:

$$\operatorname{argmax} p(e|f) = \operatorname{arg} \max p(f|e)p(e)$$

La búsqueda de la mejor traducción, es decir, la que maximiza esta probabilidad no se puede hacer calculando la probabilidad de todos los posibles candidatos de traducción, ya que la cantidad de candidatos es enorme y el cálculo tardaría mucho tiempo. Por este motivo, la búsqueda del mejor candidato se hace a partir de diversas técnicas heurísticas para reducir el espacio de búsqueda pero manteniendo al mismo tiempo un nivel de calidad aceptable. El módulo encargado de realizar esta búsqueda es el *decodificador* y asegura que la traducción elegida tenga una alta probabilidad, pero no puede asegurar que sea la más probable de todas.

Los sistemas de traducción automática estadística tienen una serie de ventajas, entre las que se pueden destacar:

- No es necesario el desarrollo de reglas lingüísticas, que en general requieren un gran esfuerzo humano, lo que encarece notablemente el desarrollo.
- Los sistemas de traducción automática estadístico pueden entrenarse y utilizarse para un gran abanico de lenguas, ya que no contienen módulos específicos para un determinado par de lenguas.
- Hay disponibles recursos lingüísticos en formato electrónico para muchas lenguas (pero no para todas) y muchos de estos recursos son multilingües (al menos entre algunos pares de lenguas). Existen algoritmos eficientes para el alineamiento automático de textos y sus traducciones.

Y como inconvenientes:

- La última ventaja que hemos comentado es un inconveniente para ciertos pares de lenguas. Cuando no están disponibles los recursos necesarios para entrenar un sistema estadístico para un

determinado por de lenguas, la elaboración de estos recursos puede ser tan costosa como la generación de sistemas de traducción basados en reglas.

- En cierto modo los resultados son inesperados, pudiendo obtenerse tanto traducciones de muy buena calidad como traducciones realmente decepcionantes.
- Estos sistemas de traducción no funcionan bien para lenguas que tengan un orden de palabras muy diferente.

Los sistemas de traducción automática estadística se pueden clasificar en los siguientes tipos principales:

Traducción a partir de palabras  
 Traducción a partir de frases  
 Traducción a partir de sintaxis  
 Modelos de traducción factorizados

Veamos brevemente cada uno de estos tipos.

### Traducción a partir de palabras (*word-based translation*)

Este es el modelo más sencillo y se basa únicamente en la traducción léxica, es decir, en la traducción de palabras de manera aislada. Para llevar a cabo la traducción con esta estrategia necesitamos un diccionario bilingüe entre la lengua de partida y la lengua de llegada. Imaginemos que queremos traducir la misma frase que hemos utilizado en otros ejemplos:

*El niño come un helado grande.*

En un diccionario convencional tendríamos la siguiente información sobre la traducción de la palabra *niño*.

*niño*: *child*, *boy*

Para nuestro sistema de traducción lo que necesitamos es un diccionario que además de las traducciones, aporte la probabilidad asociada a cada una de las traducciones. Esta probabilidad se puede calcular a partir de un corpus paralelo. Si en nuestro corpus paralelo la palabra *niño* aparece 160 veces en la parte castellana, y aparece traducida como *boy* 100 veces y como *child* 60 veces podemos calcular las probabilidades de la siguiente manera:

| Palabra castellano | Frecuencia | Traducción inglesa | Frecuencia | Probabilidad      |
|--------------------|------------|--------------------|------------|-------------------|
| niño               | 160        | boy                | 100        | $100/160 = 0,625$ |
| niño               | 160        | child              | 60         | $60/160 = 0,375$  |

Para poder traducir la frase del ejemplo, necesitaríamos disponer del diccionario probabilístico para cada una de las palabras.

| El                                 | niño                         | come  | un  | helado                                       | grande                                |
|------------------------------------|------------------------------|---|---|--|---------------------------------------|
| the - 0,8<br>a - 0,1<br>this - 0,1 | boy - 0,625<br>child - 0,375 | eats - 0,8<br>consumes - 0,1<br>ingests - 0,1 | a - 0,3<br>an - 0,2<br>one - 0,3<br>the - 0,1<br>this - 0,1 | ice_cream - 0,6<br>frozen - 0,3<br>ice - 0,1 | big - 0,4<br>large - 0,4<br>old - 0,2 |

El número de posibles traducciones según esta tabla es de  $3 \times 2 \times 3 \times 5 \times 3 \times 3 = 810$  y por cada una de ellas podemos calcular una probabilidad (veamos algunos ejemplos):

|                                 |  |
|---------------------------------|--|
| The boy eats a ice_ cream big   | $0,8 \times 0,625 \times 0,8 \times 0,3 \times 0,6 \times 0,4 = 0,0288$    |
| The boy eats a ice_ cream large | $0,8 \times 0,625 \times 0,8 \times 0,3 \times 0,6 \times 0,4 = 0,0288$    |
| This child ingests this ice old | $0,1 \times 0,375 \times 0,1 \times 0,1 \times 0,1 \times 0,2 = 0,0000075$ |

Como vemos el sistema asigna una probabilidad más alta a las traducciones que parecen más correctas (los dos primeros ejemplos tienen la misma probabilidad). Ahora bien, quedan algunos aspectos por resolver. Lo primero que veremos será los posibles cambios de orden de las palabras. Los modelos que se aplican permiten ciertos cambios en los órdenes de las palabras, de esta manera se acabarían evaluando muchas más opciones, como por ejemplo:

Boy the eats a ice\_ cream big  
 The boy eats big ice\_ cream  
 The boy eats a large ice\_ cream

Entre muchísimas otras. Ahora, a partir del modelo de lengua de llegada, lo que indica si una oración es probable en la lengua de llegada. Podemos simular este modelo haciendo consultas en la interfaz del British National Corpus (<http://www.natcorp.ox.ac.uk/>) y obtendríamos los siguientes resultados:

| Combinación      | Frecuencia en el corpus |
|------------------|-------------------------|
| the boy          | 3296                    |
| boy the          | 27                      |
| big ice_ cream   | 2                       |
| large ice_ cream | 0                       |

Combinando las diferentes probabilidades (los del modelo de traducción y del modelo de lengua) el decodificador probablemente elegiría como traducción más probable:

*The boy eats big ice cream.*

Otros problemas que hay que solucionar es el tema de la *fertilidad*, que se da porque en una determinada lengua una palabra puede ser traducida por más de una palabra en otro idioma. Cuando una palabra de la lengua de partida se traduce por más de una palabra en la lengua de llegada, esto no supone ningún problema, ya que se soluciona a nivel de diccionario (obsérvese helado - ice cream). Cuando ocurre al revés, la cosa cambia y es más complicado. Diversas palabras en la lengua de partida pueden traducirse por una única palabra en la de llegada, e incluso algunas palabras del original no aparecen en la traducción. La solución pasa por introducir palabras nulas en la traducción.

La estrategia de traducir por palabras prácticamente no se utiliza hoy en día y se utilizan mucho más la traducción a partir de frases, que veremos a continuación.

### **Traducción a partir de frases (*phrase-based translation*)**

Esta estrategia de traducción automática estadística es la que se utiliza más actualmente y la que obtiene mejores resultados y se basa en la traducción de pequeñas secuencias de palabras. Estas secuencias no tienen porque estar motivadas lingüísticamente, sino que son secuencias arbitrarias. Para explicar esta estrategia utilizaremos la misma frase del ejemplo anterior, la traducción de la frase castellana *El niño*

*come un helado grande* al inglés. Ahora nuestra tabla de traducción (la misma función que el diccionario bilingüe en el caso de traducción por palabras) tendría un aspecto como el siguiente (la expresaremos en el formato Moses, un sistema de traducción automática estadística que explicaremos brevemente más adelante). Fijémonos en que la tabla está construida por combinación de 3 palabras, 2 palabras y 1 palabra. Es un modelo de trigramas (donde el orden máximo de los n-gramas, o combinaciones de palabras, es 3).

```

el niño come ||| the boy eats ||| 0.60 ||| |||
el niño come ||| this boy eats ||| 0.1 ||| |||
el niño come ||| the child eats ||| 0.3 ||| |||
niño come un ||| boy eats a ||| 0.6 ||| |||
niño come un ||| boy eats an ||| 0.4 ||| |||
come un helado ||| eats an ice cream ||| 1 ||| |||
un helado grande ||| big ice cream ||| 0.75 ||| |||
un helado grande ||| a large helado ||| 12:25 ||| |||
el niño ||| the boy ||| 0.5 ||| |||
el niño ||| the child ||| 0.3 ||| |||
el niño ||| a boy ||| 0.1 ||| |||
el niño ||| a child ||| 0.1 ||| |||
niño come ||| boy eats ||| 0.6 ||| |||
niño come ||| child eats ||| 0.4 ||| |||
come un ||| eats a ||| 0.5 ||| |||
come un ||| eats an ||| 0.4 ||| |||
come un ||| eats the ||| 0.1 ||| |||
un helado ||| an ice cream ||| 0.85 ||| |||
un helado ||| one helado ||| 12:15 ||| |||
helado gran ||| big ice cream ||| 0.5 ||| |||
helado gran ||| large helado ||| 0.3 ||| |||
helado gran ||| great ice cream ||| 0.2 ||| |||
el ||| the ||| 0.8 ||| |||
el ||| a ||| 0.1 ||| |||
el ||| this ||| 0.1 ||| |||
niño ||| boy ||| 0.625 ||| |||
niño ||| child ||| 0.375 ||| |||
come ||| eats ||| 0.9 ||| |||
come ||| consumas ||| 0.1 ||| |||
come ||| ingesta ||| 0.1 ||| |||
un ||| a ||| 0.3 ||| |||
un ||| an ||| 0.2 ||| |||
un ||| one ||| 0.3 ||| |||
un ||| the ||| 0.1 ||| |||
un ||| this ||| 0.1 ||| |||
helado ||| helado ||| 0.6 ||| |||
helado ||| frozen ||| 0.3 ||| |||
helado ||| ice ||| 0.1 ||| |||
gran ||| big ||| 0.4 ||| |||
gran ||| large ||| 0.4 ||| |||
gran ||| old ||| 0.2 ||| |||

```

Aparte de este componente, también necesitamos un modelo de la lengua de llegada (en este caso el inglés) que nos de la probabilidad de que una frase sea una frase correcta de la lengua de llegada. Supondremos que disponemos de un corpus del inglés que está formado por 9 oraciones y 79 palabras (evidentemente, en los casos reales los corpus contienen cientos de miles de oraciones y millones de palabras). Veamos el corpus:

```

the baby was a boy
she made the boy brush his teeth every night
he remained a child in practical matters as long as he Lived
what did you eat for dinner last night?
Y did not eat yet, sonido Y gladly accept your invitation
this dog does not eat certain kinds of meat
the girl eats an ice cream
my father eats big ice cream
we also saw Abundance of large whales
had a great time at the party

```

Calculamos el modelo de lenguaje con algunas de las herramientas disponibles y obtenemos un archivo que tiene el siguiente aspecto (el formato final puede depender de la herramienta empleada; mostramos únicamente un fragmento):

```
\1-grams:
-1.614897 The -0.011136
-2.012837 Baby -0.011136
-1.836746 Boy -0.011136
-2.012837 Child -0.011136
-2.012837 An -0.011136
-1.836746 Ice -0.312166
-1.836746 Cream -0.011136
-2.012837 My -0.011136
-2.012837 Father -0.011136
-2.012837 Big -0.011136
-2.012837 We -0.011136
-2.012837 Also -0.011136
-2.012837 Saw -0.011136
-2.012837 Abundance -0.011136
-2.012837 Large -0.011136
-2.012837 Whales -0.011136
-2.012837 Had -0.011136
-2.012837 Great -0.011136
-2.012837 Time -0.011136
-2.012837 At -0.011136
-2.012837 Party 0.000000
....
-0.507687 <UNK>
\ 2-grams:
-2.198657 The baby 0.000000
-2.198657 The boy 0.000000
-2.198657 The girl 0.000000
-2.198657 The party 0.000000
-1.596597 Baby was 0.000000
-1.596597 Was a 0.000000
-2.198657 A boy 0.000000
-2.198657 A child 0.000000
-2.198657 Big 0.000000
-2.198657 A great 0.000000
....
```

Con toda esta información podemos calcular las traducciones más probables de nuestra frase de ejemplo (*el niño come un helado grande*). En este punto es importante establecer qué pesos damos a cada uno de los modelos (el de traducción y el de lengua). Si damos un peso de 3 al modelo de traducción y un 1 a de lengua obtenemos la traducción satisfactoria *the boy eats a big ice cream* en primera posición. El programa también nos puede dar una lista de traducciones ordenadas por probabilidad e indicando los valores de probabilidad según el modelo de lengua y el modelo de traducción. Veamos esta información para las 10 mejores traducciones:

```
0 ||| the boy eats big ice cream ||| LM = -24.4035 Distortion0 = 0 WordPenalty0 = -7
PhraseDictionaryMemory0 = -0.798508 ||| -26.799
0 ||| the boy eats big ice cream ||| LM = -24.4035 Distortion0 = 0 WordPenalty0 = -7
PhraseDictionaryMemory0 = -1.02165 ||| -27.4685
0 ||| the boy eats big ice cream ||| LM = -24.4035 Distortion0 = 0 WordPenalty0 = -7
PhraseDictionaryMemory0 = -1.08619 ||| -27.6621
0 ||| the boy eats big ice cream ||| LM = -24.4035 Distortion0 = 0 WordPenalty0 = -7
PhraseDictionaryMemory0 = -1.08619 ||| -27.6621
0 ||| the boy eats big ice cream ||| LM = -24.4035 Distortion0 = 0 WordPenalty0 = -7
PhraseDictionaryMemory0 = -1.42712 ||| -28.6849
0 ||| the boy eats an ice cream old ||| LM = -21.92 Distortion0 = 0 WordPenalty0 = -7
PhraseDictionaryMemory0 = -2.28278 ||| -28.7683
0 ||| the boy eats an ice cream old ||| LM = -21.92 Distortion0 = 0 WordPenalty0 = -7
PhraseDictionaryMemory0 = -2.30259 ||| -28.8277
0 ||| the boy eats an ice cream old ||| LM = -21.92 Distortion0 = 0 WordPenalty0 = -7
PhraseDictionaryMemory0 = -2.30259 ||| -28.8277
0 ||| the boy eats a frozen old ||| LM = -18.7077 Distortion0 = 0 WordPenalty0 = -6
PhraseDictionaryMemory0 = -3.54738 ||| -29.3498
0 ||| the boy eats an ice cream old ||| LM = -21.92 Distortion0 = 0 WordPenalty0 = -7
PhraseDictionaryMemory0 = -2.50593 ||| -29.4377
```

Los pesos de los diferentes modelos, y otros parámetros del sistema se pueden calcular automáticamente. Este proceso se conoce como *ajuste (tuning)*. Una parte del corpus paralelo disponible se reserva para esta etapa (un pequeño fragmento es suficiente). Entonces, una vez calculados los modelos, el sistema traduce el corpus varias veces con diferentes combinaciones de los diferentes parámetros. Como se trata de un corpus paralelo y la traducción es conocida, el sistema podrá determinar cuál es la mejor combinación de parámetros. Esta combinación de parámetros se utilizará en el sistema de traducción automática estadística resultante.

### **Traducción a partir de sintaxis (*syntax-based translation*)**

La traducción a partir de sintaxis se basa en la traducción de unidades sintácticas en vez de palabras aisladas o de cadenas de palabras. Para poder llevar a cabo esta estrategia es indispensable disponer de *parsers* potentes capaces de llevar a cabo el análisis sintáctico, tanto en el momento del entrenamiento del sistema como en el momento de llevar a cabo la traducción.

### **Modelos de traducción factorizados (*factored translation models*)**

En estos modelos de traducción se trabaja con información morfológica, sintáctica o semántica integrándolas en diferentes niveles. La motivación para trabajar de esta manera es doble:

- Los modelos de traducción contruidos a partir de representaciones más generales, como por ejemplo a partir de lemas en lugar de formas pueden basarse en estadísticas más ricas y evitar los problemas derivados de la escasez de datos (*data sparseness*) debidos a las limitaciones de tamaño de los datos de entrenamiento.
- Muchos aspectos de la traducción se pueden explicar mejor en el nivel morfológico, sintáctico o semántico. Por ejemplo, la concordancia se puede modelar mejor a partir de información morfológica y la reordenación de palabras a partir de principios sintácticos.
- 

Los modelos factorizados se basan en los modelos a partir de frases pero donde los tokens ya no son simplemente la forma de la palabra, sino un vector que contiene información sobre la forma, el lema, la categoría gramatical, morfología, semántica, etc.

Por ejemplo, un sistema que trabajara con formas, lemas, categoría gramatical e información morfológica funcionaría de la siguiente manera:

- Traduciría los lemas de la lengua de partida en lemas de la lengua de llegada
- Traduciría los factores morfológicos y de categoría gramatical
- Generaría las formas correspondientes a partir de los lemas y la información morfológica y de categoría gramatical

### **4.3.5. Sistemas de traducción automática basada en ejemplos**

Los sistemas de traducción automática basada en ejemplos (EBMT - *Example Based Machine Translation*) se basan en explotar traducciones anteriores similares para poder hacer la traducción de una nueva oración. En cierto modo se puede ver como un caso extremo de uso de memorias de traducción. En el caso de las herramientas de traducción asistida el sistema busca traducciones de oraciones similares a la que estamos traduciendo y deja la tarea de generar la nueva traducción al traductor humano. En el caso de los sistemas de traducción automática basada en ejemplos también se buscan traducciones de oraciones similares a la que se quiere traducir a unas memorias de traducción, pero la composición de la nueva traducción la hace totalmente el ordenador.

Veamos el siguiente ejemplo entre el inglés y el japonés<sup>2</sup>:

| Inglés                                 | Japonés                                     |
|--|---|
| How much is that <b>red umbrella</b> ? | Ano <b>akai kasa</b> wa ikura desu ka.      |
| How much is that <b>small camera</b> ? | Ano <b>chiisai kamera</b> wa ikura desu ka. |

A partir de estos ejemplos el sistema podría aprender que:

|                             |                                |
|-----------------------------|--------------------------------|
| How much is that <b>X</b> ? | Ano <b>X</b> wa ikura desu ka. |
|-----------------------------|--------------------------------|

Esta metodología de traducción fue propuesta en primer lugar por Nagao (1984) donde presentaba un sistema de traducción entre el japonés y el inglés. La traducción automática basada en ejemplos es especialmente atractiva para pares de lenguas bastante alejados, donde el resto de metodologías de traducción automática presentan problemas difíciles de resolver.

El funcionamiento general de estos sistemas se puede explicar en los siguientes pasos:

- Segmentación de la oración de entrada en fragmentos. Estos fragmentos por regla general están motivados lingüísticamente, a diferencia de la traducción automática estadística basada en frases, donde los fragmentos son arbitrarios.
- Se buscan los ejemplos adecuados para traducir estos fragmentos.
- Se compone la oración completa traducida a partir de los fragmentos.

#### 4.3.6. Sistemas híbridos

En los apartados anteriores hemos visto varias metodologías para la traducción automática. Cada una de estas metodologías tiene sus ventajas e inconvenientes y son capaces de solucionar algunos fenómenos lingüísticos de manera satisfactoria y en cambio tienen problemas para solucionar otros aspectos del lenguaje. Estos hechos hacen pensar que combinar de alguna manera dos o más sistemas de estrategias diferentes pueden hacer mejorar la calidad final de la traducción. La manera de combinar los sistemas pueden ser diferentes y se pueden distinguir tres estrategias principales<sup>3</sup>:

- Combinación de varios sistemas de traducción automática (*multi-engine approach*)
- Generación estadística de reglas
- Multi-etapas (*multi-pass approach*)

#### Combinación de varios sistemas de traducción automática (*multi-engine approach*)

En esta estrategia la traducción se lleva a cabo en paralelo con varios sistemas de traducción automática. La salida final se genera a partir de las salidas de los diferentes sistemas. En algunos casos se aplican técnicas estadísticas para determinar cuál de las salidas es la mejor y se toma la mejor como salida del sistema híbrido. En sistemas más avanzados la salida final se genera mediante la combinación de las diferentes salidas. Esta estrategia se utiliza habitualmente para combinar sistemas basados en reglas y sistemas estadísticos, aunque también se han explorado otras combinaciones.

#### Generación estadística de reglas

- 2 Ejemplo obtenido de la Wikipedia: [http://en.wikipedia.org/wiki/Example-based\\_machine\\_translation](http://en.wikipedia.org/wiki/Example-based_machine_translation)
- 3 Fuente: [http://en.wikipedia.org/wiki/Hybrid\\_machine\\_translation](http://en.wikipedia.org/wiki/Hybrid_machine_translation)

Esta estrategia implica el uso de datos estadísticos para generar reglas tanto léxicas como sintácticas. Una vez generadas estas reglas el sistema funciona como un sistema basado en reglas. El objetivo de esta estrategia es evitar la generación de reglas lingüísticas de forma manual, ya que es una tarea que requiere un gran esfuerzo humano. El sistema intenta generar las reglas de manera automática a partir del corpus de entrenamiento. Esta estrategia ha resultado especialmente útil para dominios restringidos.

### **Multi-etapas (*multi-pass approach*)**

Esta estrategia supone el procesamiento del texto de entrada varias veces en serie, es decir, la entrada se procesa por un sistema y la salida de este primer sistema se procesa por otro sistema, y así sucesivamente para todos los sistemas que conforman el sistema, hasta producir la salida final. El caso más habitual de sistema multi-etapa es el procesamiento del texto de entrada con un sistema de traducción automática basado en reglas y el procesamiento posterior de esta salida con un sistema de traducción automática estadístico que produce la salida final. La ventaja de esta aproximación es que se reduce la cantidad de información que necesita procesar el sistema estadístico y el hecho de que el sistema basado en reglas no necesite ser un sistema de traducción completo, reduciendo de esta manera el esfuerzo humano necesario para construir el sistema.

## 4.4. Usos de la traducción automática

### 4.4.a. Usos tradicionales de la traducción automática

Tradicionalmente se han distinguido dos usos principales de la traducción automática (Hutchins, 2009): la *diseminación* y la *asimilación*:

- **Asimilación:** el uso de sistemas de traducción automática para producir traducciones con el objetivo de tener una idea de sobre qué habla un texto.
- **Diseminación:** el uso de sistemas de traducción automática para producir traducciones de suficiente calidad para ser publicadas. En la mayoría de los casos, el resultado de la traducción automática debe ser revisado o post-editado.

En cuanto a la diseminación, para producir un texto de calidad suficiente para ser publicado prácticamente siempre es necesario una revisión o post-edición por parte de traductores humanos. En muchos casos, esta revisión es tan importante que la traducción automática se puede considerar simplemente como un borrador de la traducción. Para evitar o al menos reducir el esfuerzo de revisión y post-edición se usan *lenguajes controlados*, es decir, se parte de textos de entrada que se han producido a partir de unas normas que restringen la gramática y el vocabulario con el objetivo de disminuir la complejidad y eliminar la ambigüedad. Partiendo de estos lenguajes controlados se pueden desarrollar sistemas de traducción automática que proporcionen una salida de suficiente calidad. Esta práctica se utiliza en empresas que producen productos que requieren mucha documentación y que tengan la necesidad de producir la documentación en varias lenguas.

En cuanto a la asimilación el sistema de traducción automática debe ser capaz de producir una salida que permita hacerse una idea del contenido del texto y entender los aspectos fundamentales. Este uso es habitual en Internet, cuando un usuario visita una página en un idioma que no conoce y quiere saber las ideas principales de su contenido. Este uso también ha sido propiciado por agencias gubernamentales de inteligencia para poder filtrar ciertos textos sospechosos de contener información importante. Ante un texto en una lengua totalmente desconocida, el sistema de traducción automática nos permitirá saber si habla de terrorismo u otros temas sensibles, o si simplemente es una crónica de un partido de fútbol. En caso de confirmarse que se trata de un documento sensible, éste podrá ser traducido posteriormente por un traductor humano.

A estos usos tradicionales actualmente se añaden muchos otros, que resumiremos a continuación siguiendo el artículo de Hutchins (2009).

### 4.4.b. La traducción automática como ayuda al traductor humano

En un primer momento, la traducción automática se veía como una verdadera amenaza para la profesión de traductor y no se consideraba que esta tecnología pudiera resultar de ayuda al traductor. Una vez se ha tomado conciencia de las limitaciones de la traducción automática y a la vez los sistemas han mejorado mucho la calidad de salida, se considera ya que la traducción automática puede ser una ayuda a la profesión. Un poco más adelante en este capítulo veremos la posibilidad de combinar traducción automática y asistida dentro de una herramienta de traducción asistida por ordenador. Esta combinación puede resultar muy productiva para algunos pares de lenguas y tipos de textos concretos. También veremos más adelante el tema de la post-edición de traducción automática, que abre la puerta a nuevas posibilidades profesionales. De nuevo, para ciertos pares de lenguas y tipos de textos hay sistemas de traducción automática capaces de proporcionar una calidad suficiente para que la revisión y corrección de estos textos (su post-edición) hasta alcanzar la calidad final necesaria sea más productiva que la traducción humana.

#### 4.4.c. La traducción automática en Internet

A esta cuestión también le dedicaremos un apartado entero un poco más adelante. En Internet hay disponibles una serie de sistemas de traducción automática que ofrecen buena calidad y que en la mayoría de casos permiten un acceso gratuito. Este hecho ha popularizado el uso de sistemas de traducción automática entre usuarios generales. Los usos habituales pasan por la navegación en páginas escritas en lenguas no suficientemente conocidas por el usuario y en la traducción de pequeños textos para uso personal. Esta popularización de los sistemas de traducción automática ha hecho que la sociedad en general sea más consciente de la calidad que se puede alcanzar y para qué usos esta calidad es suficiente y también para qué usos es imprescindible la participación de un traductor humano.

#### 4.4.c. La traducción automática y los dispositivos móviles

La disponibilidad actual de dispositivos móviles potentes y con conexión a Internet ha propiciado la aparición una serie de aplicaciones relacionadas con la traducción automática que ha popularizado aún más el uso de estos sistemas. Un buen ejemplo es la integración de técnicas de OCR y traducción automática en aplicaciones que permiten traducir pequeños textos fotografiados directamente desde la cámara del dispositivo. La posibilidad de fotografiar textos y traducirlos es especialmente interesante para lenguas que usan alfabetos diferentes al del usuario. Las aplicaciones que utilizan estas técnicas lo que hacen es conectarse a un servidor que lleva a cabo las tareas de OCR y traducción y que envía el resultado al dispositivo.

#### 4.4.d. La traducción automática de voz

Los sistemas de traducción automática más conocidos son los que traducen un texto de una lengua a otra. Hay sistemas, sin embargo, capaces de traducir mensajes de voz en una lengua y producir el mensaje de voz traducido en otra lengua. En la estrategia más sencilla, un sistema de traducción automática de voz se compone de los siguientes módulos:

- Sistema de reconocimiento de voz (*speech recognition*) capaz de transformar el mensaje hablado en la lengua de partida a un texto.
- Sistema de traducción automática que traduce el texto en la lengua de partida a un texto en la lengua de llegada
- Sistema de síntesis de voz (*speech synthesis*) que transforma el texto traducido en voz en la lengua de llegada

Hay sistemas de traducción automática de voz que no son la simple concatenación de estos módulos, sino que combinan los modelos estadísticos de las diferentes etapas para poder mejorar el resultado final.

La combinación de estos sistemas con la disponibilidad de dispositivos móviles ofrece grandes posibilidades en el ámbito de la comunicación entre personas que hablan diferentes idiomas. También son importantes las aplicaciones militares de estos tipos de sistemas.

#### 4.4.e. Traducción automática y subtítulos

Actualmente existen sistemas capaces de subtítular automáticamente un vídeo. Estas herramientas se basan en el reconocimiento de voz, que convierte la señal de voz en texto y este texto se inserta automáticamente en el vídeo como subtítulo. Muchos de estos servicios ofrecen también la traducción automática de estos subtítulos. Estas funcionalidades pueden ser interesantes pero hay que tener en cuenta que se pueden producir errores tanto en la generación automática de los subtítulos originales como en su traducción.

#### 4.4.f. Traducción automática y lenguaje de signos

También se está experimentando en el área del lenguaje de signos para personas sordas. Estos sistemas, como por ejemplo el Kinect de Microsoft<sup>4</sup> o el Sisi de IBM<sup>5</sup>, transforman un mensaje de voz en signos mediante el uso de avatares.

#### 4.4.g. Traducción automática y recuperación y extracción de información

La *recuperación de información (information retrieval)* pretende encontrar los documentos interesantes de una colección de documentos a partir de una cadena de búsqueda del usuario. A menudo la colección de documentos está en varios idiomas y entonces se habla de *cross-language information retrieval (CLIR)*, que lo que pretende es encontrar documentos relevantes a partir de una consulta del usuario cuando esta consulta puede estar en una lengua distinta a la de los documentos. Esta técnica implica estrategias relacionadas con la traducción automática.

La *extracción de información (IE- information extracción)* es la tarea de extraer automáticamente información estructurada a partir de documentos no estructurados o semi-estructurados. Los documentos de los que queremos extraer la información estructurada pueden estar en varios idiomas. En estos casos también se combinan estrategias de extracción de información con estrategias de traducción automática.

Otras tareas que atraen la atención de los investigadores es el resumen automático multilingüe y los sistemas de pregunta-respuesta también multilingües. En el primer caso se pretende crear un resumen en una lengua de un documento escrito en otra lengua. Los sistemas de pregunta-respuesta (*question-answering*) son capaces de responder a preguntas formuladas en lenguaje natural a partir de información almacenada en bases de datos. El hecho de poder formular estas preguntas y recibir también la respuesta en varios idiomas es una tarea compleja.

---

4 <http://research.microsoft.com/en-us/collaboration/stories/kinect-sign-language-translator.aspx>

5 Say It, Sign It, <http://mqtt.org/projects/sisi>

## 4.5. La traducción automática en Internet

En Internet se pueden encontrar una gran cantidad de productos y servicios relacionados con la traducción automática. Como hemos comentado en apartados anteriores, el hecho de estar disponible en Internet ha popularizado el uso de este tipo de sistemas, lo que ha hecho que la sociedad sea más conscientes de la utilidad y limitaciones de la traducción automática. En este apartado presentaremos algunos de los sistemas más populares y los clasificaremos atendiendo a 5 variables, Estas variables son:

- **Modo de acceso:** que puede ser el clásico a través una página web o automáticamente mediante una API (*Application Programming Interface*). La conexión automática permite que un programa se conecte automáticamente al sistema de traducción, le envíe una petición de traducción y reciba el resultado. Este tipo de conexión permitirá la conexión de sistemas de traducción asistida con sistemas de traducción automática (como veremos en el apartado siguiente).
- **Utilización:** online o descarga. La mayoría de sistemas de traducción automática que se ofrecen por Internet, al menos de forma gratuita, permiten únicamente su utilización on-line. Una excepción interesante, como veremos, es el caso de Apertium.
- **Precio: gratuito/de pago.** Muchos sistemas comerciales ofrecen servicios gratuitos con una limitación del número de palabras a traducir. Para traducción masiva de documentos habitualmente se requiere algún tipo de pago. También hay empresas que ofrecen la traducción automática gratuita y ofrecen paralelamente un servicio de post-edición de pago.
- **Servicio principal o servicio de valor añadido.** Las principales empresas que desarrollan sistemas de traducción automática tienen presencia en Internet y ofrecen el acceso a sus productos (habitualmente con alguna limitación en cuanto a la cantidad de texto a traducir). Pero las empresas desarrolladoras no son las únicas que ofrecen traducción automática. Muchos de los sistemas de traducción automática disponibles en Internet se ofrecen como valor añadido de otro tipo de servicio. Entre estos cabe destacar las opciones de traducción automática que ofrecen los principales buscadores de Internet.
- **Estrategia de traducción automática:** intentaremos determinar cuál de las estrategias estudiadas utiliza el sistema de traducción: directa, de transferencia, interlingua, estadística o basada en ejemplos.

### Google Translate ( <https://translate.google.com> )

El motor de traducción de Google está disponible para 80 lenguas y en todas las combinaciones. La estrategia de traducción es estadística. A Google Translate se puede acceder de varias maneras:

- Con la clásica interfaz donde se puede escribir o pegar el original y seleccionar la lengua de partida (que también puede ser detectada automáticamente) y la de llegada. En la misma interfaz se pueden escribir direcciones de Internet para poder traducir páginas web y navegar en la versión traducida automáticamente. Permite también subir documentos en varios formatos y traducirlos. Estos accesos son gratuitos.
- También se puede acceder al traductor mediante una API que permite que una aplicación (por ejemplo de traducción asistida) se conecte automáticamente y recupere la traducción. Este acceso actualmente es de pago.
- Google también ofrece el *Translator Toolkit*, que es una herramienta de traducción asistida on-line. El usuario puede subir memorias de traducción y glosarios terminológicos que se utilizarán durante la traducción. También se podrá utilizar Google Translate, ya que cuando pasamos de un segmento a otro se muestra también el resultado de la traducción automática. El acceso a esta herramienta es gratuito.
- Google Translate admite sugerencias de los usuarios para mejorar la calidad de las traducciones.

### Microsoft Bing Translator ( <http://www.bing.com/translator/> )

El traductor automático de Microsoft utiliza también una estrategia estadística y está disponible para 44 lenguas y en todas las direcciones. Ofrece varias formas de acceso:

- La clásica interfaz para traducir frases, páginas web y documentos. El acceso es gratuito.
- El acceso por API, que es gratuito hasta un límite de 2.000.000 de caracteres mensuales<sup>6</sup> y a partir de este límite es de pago.
- Microsoft ofrece también la integración del traductor en muchas de sus aplicaciones, como Word, Excel, etc.
- Microsoft también ofrece muchos servicios a empresas relacionados con el uso de su traductor automático.

### Systran ( <http://www.systransoft.com/> )

Systran es uno de los sistemas de traducción automática más antiguos que todavía está en funcionamiento y ofrece programas de traducción para instalar en el propio ordenador. Utiliza una estrategia de traducción directa y esta disponible para la siguiente lenguas (pero sólo en algunas de las combinaciones): inglés, español, alemán, chino, coreano, francés, griego, italiano, japonés, holandés, polaco, portugués, ruso, sueco y árabe. Desde su web se pueden hacer traducciones gratuitas de pequeños textos. Este sistema de traducción se utilizaba en *Yahoo Babelfish* (servicio que ya no está en funcionamiento) y en los servicios de traducción de Google hasta el año 2007, momento en que fue sustituido por su propio sistema estadístico.

### Apertium ( <http://www.apertium.org/> )

Apertium es un sistema de transferencia superficial basado en reglas. Es de software libre y se distribuye bajo licencia GNU General Public License. Apertium se diseñó inicialmente para traducir entre lenguas cercanas (como catalán y castellano) pero se expandió para tratar lenguas más diferentes (como por ejemplo catalán e inglés). El sistema Apertium se puede descargar e instalar en un ordenador (preferentemente con sistema operativo Linux, aunque también hay disponibles instaladores para Windows) y proporciona los siguientes elementos:

- Un motor de traducción automática independiente de la lengua.
- Herramientas para manipular los datos lingüísticos necesarios para construir un sistema de traducción automática para cualquier par de lenguas
- Datos lingüísticos para una gran cantidad de pares de lenguas.

Actualmente el sistema está disponible para más de 30 pares de lenguas, aunque no todos los pares se encuentran en el mismo estadio de desarrollo. Desde la web se pueden traducir pequeños textos y documentos.

### Open Logos ( <http://logos-os.dfki.de/> )

Open Logos es un sistema de traducción automática de código abierto que se basa en el sistema Logos, que fue, junto con Systran, uno de los primeros sistemas comerciales. Utiliza una estrategia de traducción directa y traduce del inglés y el alemán al francés, italiano, español y portugués.

---

<sup>6</sup> Se tiene que tener en cuenta que las condiciones de acceso al servicio pueden cambiar en cualquier momento.

El sistema se distribuye bajo una doble licencia: por un lado la de software libre GNU-GPL, que permite usar y redistribuir el software libremente; por otro lado se puede adquirir una licencia comercial que permite integrar el sistema de traducción en aplicaciones propietarias.

## 4.6. Combinación de traducción automática y traducción asistida

La mayoría de sistemas de traducción asistida actuales permiten una conexión con sistemas de traducción automática, tanto si estos están instalados en el propio ordenador, como si se trata de sistemas remotos. La conexión entre estos sistemas suele ser muy simple: al usuario se le muestran las coincidencias obtenidas de la memoria de traducción y la proveniente de la traducción automática. Dependiendo del índice de similitud de la recuperada de la memoria, se inserta esta o la proveniente de la traducción automática. Cuando no hay ningún segmento en la memoria con un índice de similitud suficiente, se inserta también el proveniente de la traducción automática.

La integración podría ir más allá y en un futuro se espera que los sistemas de traducción asistida puedan combinar de una manera inteligente fragmentos de segmentos provenientes de la memoria con fragmentos provenientes de traducción automática. En Simard (2009) se presenta un sistema de integración de traducción asistida con un sistema de traducción automática estadístico a partir de frases. Esta integración precisa de sistemas de traducción automática que se comporten más como el módulo de memoria de traducción de los sistemas de traducción asistida. Esto implica por un lado producir traducciones automáticas que sean consistentes con la memoria de traducción cuando se encuentran segmentos con un alto índice de similitud; y por el otro lado desarrollar un módulo del sistema de traducción automática que sea capaz de filtrar las traducciones automáticas que tenga poca probabilidad de ser útiles.

En Forcada (2014) encontramos una interesante propuesta de ampliación del estándar TMX para incluir información proveniente de diferentes fuentes (memorias de traducción, sistemas de traducción automática, bases de datos terminológicas, glosarios y alineadores estadísticos a nivel de palabra). La idea es proporcionar toda esta información a la herramienta de traducción asistida de manera que sea capaz de identificar coincidencias a nivel subsegmental.

## 4.7. Post-edición de traducción automática

La post-edición aplicada a la traducción automática es el proceso de mejorar una traducción generada por un sistema de traducción automática. El proceso de post-edición lo lleva a cabo una persona con una formación específica para esta tarea de manera que el esfuerzo y el tiempo dedicado sea el mínimo posible. Para pares de lenguas próximas para los que hay disponibles sistemas de traducción automática suficientemente maduros la traducción automática con post-edición puede ser una estrategia muy productiva para traducir documentos logrando una muy buena calidad final.

El proceso de post-edición está relacionado con el proceso de pre-edición, es decir, el hecho de modificar el texto de partida con el objetivo de alcanzar una mejor calidad de traducción automática, por ejemplo, aplicando algún tipo de lenguaje controlado. Con la pre-edición se cambian las oraciones con estructuras sintácticas complejas y léxico poco habitual por otras más simples que puedan ser traducidas automáticamente de manera más satisfactoria. El proceso combinado de pre-edición + traducción automática + post-edición es muy habitual para la traducción de documentación técnica.

El proceso de post-edición, sin embargo, está muy a menudo presente en un flujos de trabajo en los que no se lleva a cabo pre-edición. En caso de que se quiera una traducción de un texto original manteniendo cierta fidelidad al original, no conviene pre-editarlo, ya que los cambios de estructuras se verían reflejados en la traducción.

Así pues, la post-edición consiste en la edición de traducciones realizadas automáticamente para asegurar que se alcance el nivel de calidad acordado entre el cliente y el proveedor.

Se distinguen dos tipos de post-edición:

- **La post-edición simple** (*light postediting*): el objetivo es lograr una traducción que sea comprensible e implica una intervención mínima del post-editor. La traducción final tendrá una finalidad de asimilación.
- **La post-edición completa** (*full postediting*): el objetivo es lograr una traducción correcta y estilísticamente apropiada que pueda ser utilizada tanto para fines de asimilación como de diseminación. Este tipo de post-edición, en su grado máximo, tiene como objetivo que el nivel de calidad de la traducción la haga indistinguible de la traducción realizada por un traductor humano.

La post-edición de textos traducidos automáticamente pretende aumentar la productividad y eficiencia del proceso de traducción. Aunque hay estudios que intentan establecer este grado de mejora de la productividad, todavía no queda del todo claro si esta mejora depende del par de lenguas ni de qué porcentaje de ahorro de tiempo se logra (que parece que son de entre un 15% y un 40%).

La profesión de post-editor es relativamente nueva y requiere una formación específica. Esta formación específica aún no está totalmente introducida en los planes estudios de las titulaciones relacionadas con la traducción y muy a menudo el profesional se ha formado a partir de una práctica continuada en este sector. Tampoco hay unanimidad en cuanto a los procesos que se llevan a términos ni las herramientas empleadas. En algunos casos, el post-editor se enfrenta a un texto ya totalmente traducido, y tiene acceso también al texto original. En otros casos esta post-edición se lleva a cabo dentro de una herramienta de traducción asistida y el post-editor tiene acceso automático, a la original, la traducción, memorias de traducción y bases de datos terminológicas.

Hay algunas pautas y consejos dados por algunas instituciones, entre las que podemos destacar las que proporciona TAUS<sup>7</sup>. Los puntos que consideramos más importantes para hacer una tarea de post-edición con garantías son:

La persona debe tener una buena formación en las lenguas de partida y de llegada, y preferentemente formación y experiencia en traducción. Como hemos comentado muchos post-editores se han formado mediante la práctica. En estos casos será necesario proporcionar unas pautas y apoyo, así como comentarios sobre las tareas realizadas. De esta manera la formación será más rápida y efectiva.

Es necesario que el post-editor conozca las características principales del sistema de traducción automática que se utiliza (qué estrategia de traducción utiliza, si es un sistema general o específico para la especialidad que se está traduciendo, etc.) De esta manera el post-editor se podrá fijar más en aquellos aspectos que sean más susceptibles de contener errores.

Es imprescindible pactar con el cliente cuál es el resultado que espera: será muy diferente si lo que quiere es tener una traducción que le permita comprender el contenido del texto, o bien si desea una traducción que pueda ser publicada. El cliente también deberá tener claro que el precio de un servicio y otro son diferentes.

Hay que asegurarse de que el texto original sea correcto. En algunos casos específicos también se podrá optar por hacer una pre-edición que asegure que el resultado de la traducción automática sea óptimo.

Es también importante que los post-editores tengan acceso rápido a toda la información necesaria: texto original, traducción, bases de datos terminológicas y memorias de traducción.

Es importante que los post-editores proporcionen información sobre los tipos de errores más habituales. Esta información será interesante para introducir mejoras en el sistema de traducción automática. Ahora bien, si se pide esta información a los post-editores, tendrán que recibir también una compensación económica.

Aunque la mayoría de las herramientas de traducción asistida habituales pueden usarse con éxito en la tarea de post-edición, hay varios estudios que pretenden diseñar herramientas específicas para esta tarea. En Aziv (2012) y Roturier (2013) podemos encontrar algunas propuestas. También es interesante el estudio de Vieira (2011) que muestra una clasificación de los sistemas de traducción automática disponibles desde el punto de vista de la tarea de la post-edición.

---

7 <https://evaluation.taus.net/resources/guidelines/post-editing/machine-translation-post-editing-guidelines-spanish>

## 4.8. Conclusiones

En este capítulo hemos expuesto los conceptos básicos sobre la traducción automática que todo traductor debería conocer. La traducción automática no se ha convertido, como algunos vaticinaban en los primeros años de la historia de estos sistemas, una amenaza para la profesión de traductor, sino que se ha convertido en un aliado a su tarea. En los últimos años la calidad de los sistemas de traducción automática ha mejorado enormemente y se esperan más mejoras así como la disponibilidad de muchos más pares de lenguas en los próximos años. A pesar de esta gran mejora, nadie duda hoy en día que el traductor humano será imprescindible siempre para alcanzar los niveles de calidad óptimos.

## 4.9. Para ampliar conocimientos

### 4.9.1. Historia de la traducción automática

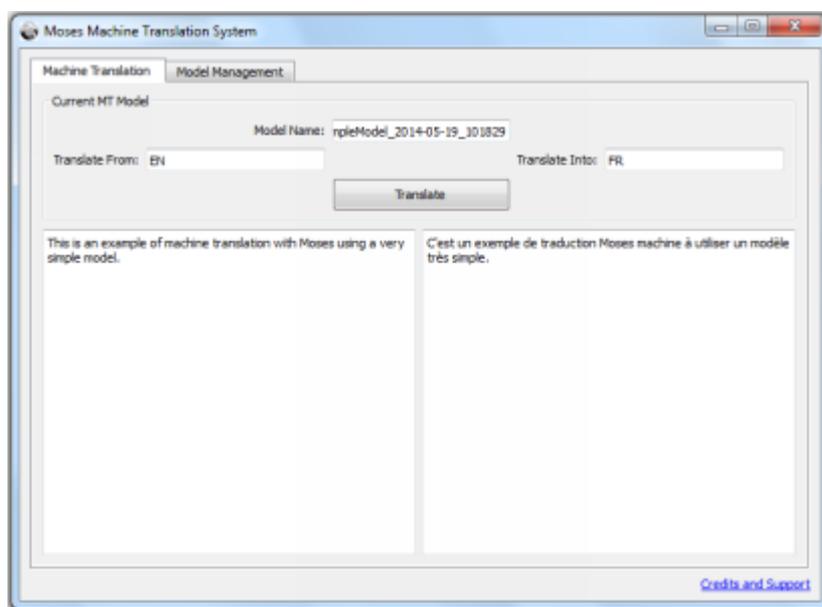
Para ampliar mucho más sobre la historia de la traducción automática, se puede leer:

Hutchins, J. (1986) *Machine Translation: past, present, future* Ellis Horwood Series in Computers and their Applications 382 pp. Chichester (UK): Ellis Horwood, 1986. (ISBN: 0-85312-788-3) New York: Halsted Press, 1986. (ISBN: 0-470-20313-7)

Que está accesible libremente en: <http://www.hutchinsweb.me.uk/PPF-TOC.htm>

### 4.9.2. Moses: un sistema de traducción automática estadística

Moses (<http://www.statmt.org/moses/>) es un sistema de traducción automática estadística que permite entrenar y utilizar sistemas de traducción automática para cualquier par de lenguas. Dispone de un decodificador eficiente que permite encontrar una de las traducciones más probables a partir de los modelos entrenados. El sistema cuenta con una documentación precisa y una serie de tutoriales que permiten introducirse en las tareas de entrenamiento de sistemas de traducción automática estadística de una manera relativamente fácil. Se requiere, sin embargo, una cierta experiencia en el uso de sistemas Linux/Unix. Para usuarios de Windows hay una versión sencilla y con interfaz gráfica que permite probar sistemas ya entrenados. Podemos observar esta interfaz en la siguiente figura:



En la misma web también se pueden descargar sistemas ya entrenados para varios pares de lenguas.

### 4.9.3. Otros sistemas de traducción automática

En el mercado hay una gran cantidad de traducción automática. Se puede encontrar una recopilación de sistemas en: [http://en.wikipedia.org/wiki/Comparison\\_of\\_machine\\_translation\\_applications](http://en.wikipedia.org/wiki/Comparison_of_machine_translation_applications)

Para tener una idea de los sistemas de traducción automática y otro software relacionado con licencia libre se puede consultar el enlace: <http://www.computing.dcu.ie/~mforcada/fosmt.html>

## Bibliografía

Aziz, W.; Sousa, S. C. M.; Specia, L. (2012). *PET: a tool for post-editing and assessing machine translation*. In The Eighth International Conference on Language Resources and Evaluation, LREC '12, Istanbul, Turkey. May 2012.

Forcada, M. (2014) *On the Annotation of TMX Translation Memories for Advanced Leveraging in Computer-aided Translation*. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik (Iceland). Eds. Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Thierry Declerck and Hrafn Loftsson and Bente Maegaard and Joseph Mariani and Asuncion Moreno and Jan Odijk and Stelios Piperidis. European Language Resources Association (ELRA). ISBN 978-2-9517408-8-4.

Green S., Heer J. and Manning c. D. (2013) *The Efficacy of Human Post-Editing for Language Translation*. ACM Human Factors in Computing Systems (CHI). <http://vis.stanford.edu/papers/post-editing>

Hutchins, J. 2007. *Machine translation: A concise history*. In *Computer Aided Translation: Theory and Practice*, C. S. Wai, Ed. Chinese University of Hong Kong.

Hutchins, J. (2009). *Multiple Uses of Machine Translation and Computerised Translation Tools*. In Proceedings of the International Symposium on Data and Sense Mining, Machine Translation and Controlled Languages – ISMTCL 2009. <http://www.hutchinsweb.me.uk/Besancon-2009.pdf>;

Koehn, P. (2010) *Statistical Machine Translation* Cambridge University Press

Makoto Nagao (1984). *A framework of a mechanical translation between Japanese and English by analogy principle* In A. Elithorn and R. Banerji. Artificial and Human Intelligence. Elsevier Science Publishers.

Oliver A. (2006) *La traducció automàtica a Internet* - Revista Tradumàtica – Traducció i Tecnologies de la informació i la Comunicació 04 : Traducció Automàtica : <http://www.fti.uab.cat/tradumatica/revista> ISSN 1578-7559

Roturier J., Mitchell L. and D. Silva (2013) *The ACCEPT Post-Editing Environment: a Flexible and Customisable Online Tool to Perform and Analyse Machine Translation Post-Editing* in Proceedings of the MT Summit XIV Workshop on Post-editing Technology and Practice. Nice (France)

Schubert, K. (1988) *The architecture of DLT – interlingua or double direct?* In Maxwell, Dan, Klaus Schubert and Toon Witkam (eds): (1988). *New Directions in Machine Translation*. Dordrecht: Foris.



Simard M. And Isabelle P. (2009) *Phrase-based Machine Translation in a Computer-assisted Translation Environment*. In The Twelfth Machine Translation Summit (MT Summit XII), pages 120--127, Ottawa, Ontario, Canada.

Trujillo, A. (1999) *Translation Engines: Techniques for Machine Translation*. Springer. ISBN 978-1-85233-057-6

Vieira, L.; Specia, L. (2011). *A Review of Machine Translation Tools from a Post-Editing Perspective*. 3rd Joint EM+/CNGL Workshop Bringing MT to the User: Research Meets Translators (JEC 2011), Luxembourg.

## Licencia de este documento



Traducción y Tecnología by Antoni Oliver

is licensed under a [Creative Commons Attribution - ShareAlike 3.0 Unported License](https://creativecommons.org/licenses/by-sa/3.0/) .