e-ISSN: 2695-8465
ISSN: 2255-3703

Análisis contrastivo de las coincidencias parciales en español, francés y árabe mediante transformaciones lingüísticas

Souhila Djabri
Universidad de Alicante
sd89@alu.ua.es

Fecha de recepción: 27.09.2020
Fecha de aceptación: 11.12.2020

Resumen: Los sistemas de traducción asistida por ordenador (TAO) han revolucionado el mundo de la traducción almacenando automáticamente las traducciones anteriores, por lo que ofrecen cada vez más soluciones. No obstante, la persona que utiliza estos sistemas observa una serie de deficiencias en particular en la recuperación de los datos por medio de las coincidencias o *matching* considerando que ciertos sistemas funcionan con un algoritmo basado en la distancia de edición o distancia de Levenshtein. El presente estudio ofrece la posibilidad de mejorar estas coincidencias gracias a la identificación de los límites o las dificultades sintácticas y semánticas que afronta una memoria de traducción, aplicando diez tipos de transformación en distintas combinaciones lingüísticas: español-francés/francés-español/árabe-español. Los resultados muestran ciertas limitaciones tanto en las combinaciones ES-FR/FR-ES y aún más entre el AR-ES que se pueden reducir mediante la integración de conocimientos lingüísticos de modo que aumente el grado de las coincidencias para el mayor beneficio del traductor.

Palabras clave: coincidencias parciales, corpus paralelo, lingüística contrastiva, memoria de traducción, transformación.

A contrastive analysis of Fuzzy Matches in Spanish, French and Arabic using linguistic transformation

Abstract: Computer-aided translation (CAT) tools have revolutionized the world of translation by automatically storing previous translations and are increasingly offering solutions. However, the person using these systems observes a number of deficiencies particularly in the retrieval of the data matching considered that some work with an algorithm based on the editing distance or distance of Levenshtein. This study offers a possibility to improve the matching by the identification of the syntactic and semantic difficulties and the limitation faced by a translation memory with ten types of transformations in different linguistic combinations: Spanish-French/French-

Spanish/Arabic-Spanish. The results show certain limitation in both ES-FR/FR-ES combinations and even more between AR-ES that can be reduced by integrating linguistic knowledge and increasing the matching level for the greatest benefit of the translator.

Key words: fuzzy matches; parallel corpora; comparative linguistics; translation memory; transformation.

Sumario: 1. Introducción. 2. Metodología. 2.1 Definición del corpus. 2.2 Sistemas TAO utilizados y *Pipeline* de trabajo. 3. Transformaciones. 3.1 Transformación de los segmentos en idioma español. 3.2 Transformación de los segmentos en idioma francés. 3.3. Transformación de los segmentos en idioma árabe. 4. Presentación y análisis de los resultados. 4.1. Español-francés. 4.2. Francés-español. 4.3. Árabe-español. 5. Conclusiones y trabajos futuros.

1. Introducción

Las memorias de traducción son el mayor componente de un sistema de traducción asistida por ordenador (Simard 2020:79), almacenan las traducciones anteriores de manera que, al traducir un nuevo segmento, se puede recuperar automáticamente su equivalente en el idioma de destino a partir de una base de datos evitando tener que traducir un segmento ya guardado en la memoria del sistema. Previamente, el sistema fragmenta el texto de origen en segmentos generalmente formados por frases u oraciones (Antoni González 2016:17) y busca su traducción mediante coincidencias de varios grados. Bowker (2002: 95-106) ofrece una clasificación amplia de las coincidencias destacando seis tipos de coincidencias: (i) la coincidencia exacta o perfecta, 100% idéntica al segmento de origen; (ii) la coincidencia completa, que aparece cuando el segmento de origen es distinto al segmento almacenado en la memoria de traducción por unos elementos variables como los nombres de entidades; (iii) coincidencias parciales, que recuperan segmentos de la memoria de traducción similares pero no idénticos, suelen oscilar entre 1 y 99 %. Los tres otros tipos son las coincidencias terminológicas, es decir, recuperan la traducción de un término a partir de una base de datos terminológica; coincidencia mediante subsegmentos ubicada entre una coincidencia parcial y termológica; y, por último, las coincidencias cero, al no existir ningún segmento o subsegmento que puede ser recuperado de la memoria de traducción.

La presente investigación se centra en el tercer tipo de coincidencias o las coincidencias parciales (*fuzzy matches*) y ofrece un estudio cuantitativo mediante el análisis de estas mismas aplicando diez

transformaciones semántico-sintácticas en tres combinaciones lingüísticas español-francés/francés-español/árabe-español. El trabajo tiene como principales objetivos estudiar las posibilidades de mejorar el grado de estas coincidencias y aumentar el rendimiento de las memorias de traducción mediante la detección de los límites y deficiencias que impactan estas mismas.

Varios estudios se interesaron por las memorias de traducción como la segmentación, la alineación de los textos o la recuperación del mayor número posible de datos almacenados. Investigaciones recientes en el procesamiento del lenguaje natural realizadas por Tharindu et al. (2020:175-184) posibilitarían mejorar la recuperación de datos como son las coincidencias parciales mediante el uso de un codificador de oraciones universal como alternativa al algoritmo clásico de las memorias de traducción, un planteamiento innovador de memorias de traducción inteligentes y de nueva generación. Otras investigaciones (Silvestre Baquero y Mitkov 2017: 44-51) demostraron mediante transformaciones lingüísticas la necesidad de integrar tecnologías de procesamiento del lenguaje natural al detectar que las coincidencias parciales del español como lengua de destino presentan más dificultades léxicas y sintácticas que el inglés. Varios trabajos realizados por los miembros del Grupo de Investigación en Lingüística Computacional de la Universidad de Wolverhampton (Gupta et al. 2015:35-42) experimentaron técnicas de paráfrasis en las memorias de traducción. Los resultados, que son prometedores, se centran en el idioma inglés y alemán. Chatzitheodorou (2015:24-30) utiliza la herramienta de tratamiento del lenguaje NooJ para crear paráfrasis de los términos compuestos de textos de origen para mejorar las coincidencias. Sus primeros resultados para el inglés-italiano muestran mejoras significativas en las coincidencias de diferentes tipos de textos.

Nuestro planteamiento es distinto, aunque tiene como principal objetivo mejorar el grado de las coincidencias parciales. Se emplean por primera vez tres combinaciones lingüísticas con un análisis comparativo entre el español, el francés y el árabe, partiendo de una hipótesis que supone que las transformaciones aplicadas influyen en el comportamiento de las memorias de traducción entre las combinaciones lingüísticas ES-FR/FR-ES y lo harán significadamente entre el AR-ES dado que las primeras combinaciones derivan del latín y comparten ciertas similitudes lingüísticas, lo que permite obtener un grado de coincidencia con descenso menos significativo que la combinación AR-ES, al ser el idioma árabe lengua semítica con un sistema lingüístico lejano a los sistemas latinos ES/FR.

El resto del artículo se presenta como sigue: el segundo apartado describe la metodología de nuestra investigación con un primer subapartado que define el corpus empleado y un segundo subapartado que presenta los sistemas TAO utilizados y el pipeline de trabajo. Posteriormente, detallamos en el apartado 3 las diez transformaciones aplicadas para cada idioma. El apartado 4 presenta y analiza los datos empíricos de las tres combinaciones lingüísticas y el apartado 5, las conclusiones y los trabajos futuros.

2. Metodología

En este estudio analizamos las coincidencias parciales de tres memorias de traducción: español, francés y árabe, para lo que seleccionamos 500 segmentos en cada idioma (250 originales con sus respectivos 250 segmentos transformados), un total de 1500 segmentos de los cuales 750 se transforman según se detalla en la tabla 2, 3 y 4. Las unidades de análisis de nuestra investigación son: (i) segmentos con una transformación para los tipos 1, 2, 3,4 y 7; (ii) segmentos con dos transformaciones para los tipos 5, 8,9 y 10; (iii) segmentos con tres transformaciones para el tipo 6. Por consiguiente, obtenemos 30 tablas en hojas Excel aplicando transformaciones sintácticas y semánticas para cada segmento y de forma manual sin el uso de ninguna herramienta lingüística. Las memorias de traducción están generadas mediante operaciones previas, es decir, a partir de la importación de documentos paralelos originales y transformados o la alineación de textos de origen con su destino (véase 2.2) aunque existen otros métodos para crear una memoria de traducción como por ejemplo durante la tarea traductora, aplicando al mismo tiempo las transformaciones para cada segmento.

2.1 Definición del corpus

El corpus utilizado en esta investigación es multilingüe y paralelo: multilingüe porque se trata del reglamento de la Asamblea General de las Naciones Unidas en su versión española (A), francesa (B) y árabe (C) y paralelo (McEnery, 2016:3) al utilizar el mismo documento en el análisis de las coincidencias entre los pares de idiomas (A-B) (B-A) (C-A), generando tres memorias de traducción para cada idioma. A continuación, una tabla resumen de unas estadísticas del corpus:

Denominación del documento	Idioma	N. ° de palabras
Reglamento de la Asamblea General	Español	36 642
Règlement Intérieur de l'Assemblée Générale	Francés	34 890
النظام الداخلي للجمعية العامة	Árabe	29 530

Tabla 1: Algunas estadísticas del corpus

2.2 Sistemas TAO utilizados y Pipeline *de trabajo*

Para calcular las coincidencias parciales de todos los segmentos transformados, se utilizan dos sistemas TAO: SDL Trados 2019 y MemoQ V 8.7. La primera etapa del proceso de trabajo consiste en deshabilitar la traducción automática y no asignar a las tres memorias de traducción generadas ninguna herramienta como las bases de datos terminológicas o los glosarios o toda herramienta que pueda interferir en el cálculo de las coincidencias. Se describe a continuación el proceso de trabajo y el *pipeline* elaborado para cada sistema TAO:

SDL Trados

- Crear una nueva memoria de traducción denominada "corpus paralelo no transformado";
- Seleccionar los siguientes pares de idioma de forma separada: ES-FR/ FR-ES /AR-ES;
- Importar el corpus original sin transformaciones y generar así una memoria de traducción para cada idioma;
- Importar de nuevo el documento con todas las transformaciones efectuadas.

MemoQ

- Agregar pares de alineación;
- Seleccionar los tres pares de idioma de forma separada;
- Agregar el documento de origen y el documento de destino;
- MemoQ realiza la importación del corpus paralelo original no transformado para guardarlo como nueva memoria de traducción;
- Importar de nuevo el documento con los segmentos transformados y asignar la memoria de traducción antes creada.

Al término de estas operaciones, se calcula el grado de coincidencia de cada segmento transformado y en cada combinación lingüística según se detalla en el apartado 4.

3. Transformaciones

Detallamos a continuación las características de las diez transformaciones aplicadas para cada idioma: a saber, español, francés y árabe (véase unos ejemplos en las tablas 2, 3 y 4). La primera columna indica el número de la transformación que utilizamos más adelante para presentar y analizar los datos empíricos, la segunda columna refleja el número de segmentos para cada transformación (25 segmentos transformados para cada una y en cada idioma),la tercera columna indica el tipo o la denominación de la transformación con un ejemplo del segmento original (SO) en la cuarta columna y su respectiva transformación en la quinta columna (ST).

3.1 Transformación de los segmentos en idioma español

La primera transformación aplicada permite cambiar la voz activa en pasiva donde el sujeto de la voz activa se convierte en el agente precedido de la voz pasiva y el verbo se convierte en participio pasado precedido por el verbo 'ser', entre otras transformaciones (Zulma y Dezier 2014:271). La segunda transformación cambia las voces pasivas en activas, aunque el español tiende a evitar la pasiva, según señala García Yebra (1990:348) con unos usos exclusivos por razones de ritmo. Sin embargo, por necesidades de la presente investigación, aplicamos esta transformación al igual que en las memorias de traducción francés y árabe. La tercera transformación cambia el orden de las palabras, enunciados, frases u oraciones. La cuarta transformación se aplica para la sinonimia, es decir, al sustituir una palabra por su sinónimo (se emplea el diccionario digital proporcionado por la Editorial Santillana). Elegimos para estos casos un sinónimo del mismo género y número para que el segmento no experimente otras modificaciones, también prestamos mayor atención al sinónimo en su contexto ya que nos enfrentamos a muchos casos de polisemia. Para la quinta trasformación, sustituimos dos palabras por sus respectivos sinónimos sin cambio de orden de las palabras mientras mantengamos la sustitución por dos sinónimos para la sexta transformación, pero cambiando de orden de las palabras**.** En la séptima transformación sustituimos una palabra por un pronombre personal sin cambio de orden y volvemos a sustituir una palabra por un pronombre en la octava transformación cambiando el orden de las palabras. Las dos últimas transformaciones, cada una con 25 segmentos, cambian las voces activas en pasivas

sustituyendo una palabra por su sinónimo y por un pronombre personal para la décima transformación.

N.º	N. ° ST	Tipo	SO	ST
1	25	Voz activa-pasiva	La Asamblea General **establecerá** el reglamento para la administración financiera de las Naciones Unidas.	El reglamento para la administración financiera de las Naciones Unidas **será establecido** por la Asamblea General.
2	25	Voz pasiva-activa	En virtud de las cuales los artículos 51 a 59 **fueron sustituidos por** los nuevos artículos 51 a 57 y los artículos 60 a 165 fueron renumerados en consecuencia.	Los nuevos artículos 51 a 57 **sustituyeron** los artículos 51 a 59 y los artículos 60 a 165 fueron renumerados en consecuencia.
3	25	Cambio de orden de las palabras, enunciado, frases o/y oraciones	La Asamblea General, por mayoría de los miembros **presentes** y votantes, podrá modificar o suprimir temas de s u programa.	La Asamblea General, por mayoría de los miembros votantes y **presentes,** podrá modificar o suprimir temas de su programa.

Tabla 2: Segmentos originales y transformados/español

3.2 Transformación de los segmentos en idioma francés

Al igual que los segmentos en español, se aplican las mismas diez transformaciones antes mencionadas, obviamente el genio de cada idioma impone respetar las características de cada sistema lingüístico, dado que el francés tiene sus propias características gramaticales como por ejemplo en el uso de la voz pasiva. El francés no parece estar sometido a ninguna restricción y la voz pasiva es muy frecuente en francés tanto en la lengua oral como escrita, mientras que es menos frecuente en español (Weber 2014: 370). Por lo tanto, seleccionamos varias construcciones pasivas en la transformación de los segmentos: pasiva refleja o pasiva perifrástica. Con relación a la sustitución por sinónimos, utilizamos el diccionario digital francés *Larousse*. En cuanto a la sustitución por pronombre, optamos por el uso de los pronombres personales de la tercera persona del singular,

masculino y femenino: *il/elle* (él/ella) además de la tercera persona del plural, masculino y femenino: *ils /elles* (ellos/ellas).

N. °	N.º ST	Tipo	SO	ST
4	25	Sustituir una palabra por su sinónimo sin cambiar el orden de las palabras.	Sur recommandation du Bureau, l'Assemblée générale fixe, au début de chaque session, une date pour **la clôture** de la session.	Sur recommandation du Bureau, l'Assemblée générale fixe, au début de chaque session, une date pour **la fin** de la session.
5	25	Sustituir dos palabras por sus sinónimos sin cambiar el orden de las palabras.	**Les termes** de ces recommandations seraient **incorporés**, sous forme d'annexe, au Règlement intérieur ;	**Les dispositions** de ces recommandations seraient **associées**, sous forme d'annexe, au Règlement intérieur ;
6	25	Sustituir dos palabras por sus sinónimos y cambiar el orden de las palabras.	Les chiffres **indiqués** entre crochets après le numéro des articles relatifs aux **séances** plénières renvoient aux articles identiques ou correspondants relatifs aux séances de commissions, et vice versa.	Les chiffres **précisés** après le numéro des articles entre crochets relatifs aux **sessions** plénières renvoient aux articles identiques ou correspondants relatifs aux séances de commissions, et vice versa.

Tabla 3: Segmentos originales y transformados/francés

3.3. *Transformación de los segmentos en idioma árabe*

La tercera combinación lingüística permite crear una memoria de traducción para el idioma árabe aplicando las diez trasformaciones antes mencionadas para un total de 250 segmentos (véase algunos ejemplos en la tabla 4). En primer lugar, se transforma la voz activa en pasiva y viceversa. La voz activa árabe se denomina *maaloum* ('conocido') mientras la pasiva *machhul* ('desconocido') refiriéndose al agente de la acción, el sujeto de esta misma voz no se denomina agente sino sustituto del agente, es decir, el que hace la acción, término equivalente al de sujeto paciente en

español. Se obtiene por cambio de la vocalización del verbo con una *damma(ُ)* en la primera sílaba del verbo, una *Kasra (ِ)* en la penúltima para el perfectivo y una *fatha (َ)* en el imperfectivo (Alubudi 2008:290). En la tercera transformación, cambiamos el orden de las palabras, enunciados, frases u oraciones con menos restricciones y más libertad dada la naturaleza del idioma árabe, que ofrece una morfología más rica y flexible en cuanto al orden de las palabras, de modo que se puede cambiar el orden según distintas estructuras: VSO, SVO y VOS (Hammo et al, 2017:612).Para la sustitución por sinónimos, se utilizan los diccionarios digitales árabes *Mu'jam al-Wasīṭ* y *Mu'jam al-Maany*. En la séptima y la octava transformación, sustituimos las palabras por un pronombre personal singular o plural de género masculino o femenino, que pueden ser separados o conjuntos, no cambiamos el orden de las palabras en la séptima mientras que sí lo cambiamos en la octava. Por último, transformamos la voz activa en pasiva sustituyendo en la novena una palabra por su sinónimo y por un pronombre en la décima transformación.

N. °	N.º ST	Tipo	SO	ST
7	25	Sustituir una palabra por un pronombre sin cambiar el orden de las palabras.	تنتخب **كل لجنة أخرى** رئيسا لها، ونائبا للرئيس أو أكثر، ومقررا؛	**هي** تنتخب رئيسا لها، ونائبا للرئيس أو أكثر، ومقررا؛
8	25	Sustituir una palabra por un pronombre y cambiar el orden de las palabras, enunciados, frases u oraciones.	لدى تصويت **الجمعية العامة** بواسطة الجهاز الآلي، يحل التصويت غير المسجل محل التصويت برفع الأيدي أو بالوقوف ويحل التصويت المسجل محل التصويت بنداء الأسماء.	يحل التصويت غير المسجل محل التصويت برفع الأيدي أو بالوقوف لدى **تصويتها** بواسطة الجهاز الآلي، ويحل التصويت المسجل محل التصويت بنداء الأسماء.
9	25	Transformar la voz activa en pasiva y sustituir una palabra por su sinónimo.	يُبلغ الأمين العام القرارات التي تتخذها الجمعية العامة إلى أعضاء الأمم المتحدة في غضون خمسة عشر يوما من **اختتام** الدورة.	تُبلغ القرارات التي تتخذها الجمعية العامة إلى أعضاء الأمم المتحدة في غضون خمسة عشر يوما من **نهاية** الدورة.
10	25	Transformar la voz activa en pasiva y sustituir una palabra por un pronombre.	ولا يجوز للرئيس أن يأذن **لصاحب اقتراح** أو تعديل بأن يُعلل تصويته على الاقتراح أو التعديل الذي قدمه.	ولا يجوز أن يُأذن **لصاحبه** بأن يُعلل تصويته على الاقتراح أو التعديل الذي قدمه.

Tabla 4: Segmentos originales y transformados/árabe

4. Presentación y análisis de los resultados

Al término de todas las operaciones de transformación, obtenemos los resultados de las coincidencias de cada par de idiomas, analizamos los resultados de dos formas: (i) resultados de las coincidencias según las combinaciones lingüísticas; (ii) resultados de las coincidencias según las combinaciones lingüísticas y por categorías de coincidencias.

4.1. Español-francés

El objetivo de esta investigación es ver el comportamiento de los sistemas TAO ante las transformaciones semántico-sintácticas para cada idioma con el fin de detectar las carencias y aportar propuestas de mejora, sea para pares de idiomas con estructuras lingüísticas suficientemente cercanas como es el caso del español y el francés o lejanas (árabe/español). Los resultados de las coincidencias ES-FR de la tabla 5, donde la primera columna indica el número de la trasformación, y la segunda, los resultados de los segmentos transformados (ST) para SDL según dos clases: inferior y superior a 75 % y la tercera columna indica los resultados para MemoQ. En primer lugar, los resultados muestran que el grado de coincidencia baja de forma significativa al utilizar MemoQ especialmente con la trasformación 9 donde se registra un 32 % de los segmentos por debajo de 75 % y 28 % para la sexta transformación o 24 % para las transformaciones 1, 5, 10, indicando claramente que el sistema afronta ciertas dificultades debido a los cambios aplicados. En suma, cuando se trata de múltiples operaciones como por ejemplo sustituir dos palabras por sus respectivos sinónimos y cambiar el orden de las palabras (tres transformaciones en un segmento) o transformar las voces activas en pasivas sustituyendo una palabra por su sinónimo (dos transformaciones en un segmento). SDL parece soportar mejor las transformaciones en ciertos casos, aunque también registra descensos, pero con grados de coincidencia más elevados, incluso nos ofrece 100 % de segmentos transformados > 75 % para el segundo tipo de transformación (voz pasiva-activa). También registramos grados elevados o con poco descenso para la mayoría de los segmentos transformados mediante los tipos 1, 3, 4, 5, 9 y 10, casi todos superiores a los resultados obtenidos con MemoQ. No obstante, el grado más bajo es de 44 % de los segmentos transformados que se sitúan por debajo de 75 % para la sexta transformación mediante SDL en comparación con 28 % para MemoQ, es decir para este caso donde sustituimos dos palabras por sus respectivos sinónimos y cambiamos de orden de las palabras, MemoQ soporta mejor los cambios pese al número de operaciones. De igual modo, los grados de coincidencia siguen descendiendo para la séptima y la octava transformación con 72 % y 68 %

de los segmentos > 75 % comparando con resultados no tan bajos para MemoQ: 92 % y 88 % respectivamente, demostrando que SDL tiene dificultades con la sustitución por pronombre sobre todo cuando se cambia el orden de las palabras, mientras MemoQ soporta mejor y no parece experimentar grandes problemas.

	SDL		MEMOQ	
	ST		ST	
N. °	<75 %	>75 %	<75 %	>75 %
1	12 %	**88 %**	24 %	**76 %**
2	0 %	**100 %**	12 %	**88 %**
3	8 %	**92 %**	16 %	**84 %**
4	16 %	**84 %**	8 %	**92 %**
5	8 %	**92 %**	24 %	**76 %**
6	44 %	**56 %**	28 %	**72 %**
7	28 %	**72 %**	8 %	**92 %**
8	32 %	**68 %**	12 %	**88 %**
9	20 %	**80 %**	32 %	**68 %**
10	12 %	**88 %**	24 %	**76 %**

Tabla 5: Resultados de las coincidencias ES-FR

Los primeros resultados permiten detallar más los grados de coincidencia para cada sistema TAO. Para ello, se establecen cálculos según distintas categorías, intentamos en la primera columna de la tabla 6 ofrecer una categorización más amplia dividiendo los grados de coincidencia que se ubican entre 99 % y 60 %. Como se aprecia no incluimos el grado 100 % (coincidencia total o perfecta) ya que no es posible obtenerlo después de la transformación. La segunda columna, desde la transformación 1 hasta la 10, indica los resultados obtenidos por número de segmento (un total de 25 segmentos para cada transformación) mediante SDL y la tercera columna (también de 1 a 10) indica los resultados obtenidos mediante MemoQ. Las cifras indicadas en negrita son el número de segmentos ubicados en cada categoría de coincidencia, el mayor número de segmentos se ubica entre 90 %-99 % para SDL después de aplicar las transformaciones 2, 3, 4, 5, 7, 8 y 10, lo que significa que, a pesar de los cambios sintáctico-semánticos, SDL nos ofrece un grado de coincidencia con un leve descenso como, por ejemplo, para la segunda transformación, donde 20 de los 25 segmentos transformados se ubican

entre 90 %-99 %. Se confirman así los resultados de la tabla anterior y muestran que el sistema no experimenta muchas complicaciones. Sin embargo, no parece soportar los cambios de la transformación 6 con 12 de 25 segmentos ubicados entre 70 %-79 %, cuando sustituimos dos palabras por sus respectivos sinónimos y cambiamos el orden de las palabras. No obstante, MemoQ parece tener dificultades significativas al notar que la mayoría de los segmentos transformados bajan de grado de coincidencia situándose entre 80 %-89 % (transformación 1, 2, 3, 4, 6, 8 y 10), registrando incluso más segmentos situados entre 60 %-69 %, al contrario de SDL con tan sólo un segmento para las transformaciones 4, 5, 6 y 8.

	SDL										MEMOQ									
GRADO	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
90 %-99 %	7	**20**	**17**	**15**	**16**	9	**16**	**11**	5	**11**	0	8	1	8	**12**	0	**13**	1	0	2
80 %-89 %	**10**	3	5	4	4	3	2	4	9	7	**10**	**11**	**12**	**11**	5	**15**	9	**20**	10	**11**
70 %-79 %	8	2	3	5	4	**12**	7	9	**11**	7	9	4	11	4	8	10	2	2	**11**	8
60 %-69 %	0	0	0	1	1	1	0	1	0	0	6	2	1	2	0	0	1	2	4	4

Tabla 6: Resultado de las coincidencias ES-FR por segmentos

4.2. Francés-español

La segunda operación abarca la combinación lingüística FR-ES, se presentan en la tabla 7 los resultados de las coincidencias parciales, primero con SDL: 100 % de los segmentos transformados, o sea un total de 25 indican una coincidencia parcial superior a 75 % para las transformaciones 4 y 7, seguidos por 96 % para las transformaciones 5 y 8, baja el grado de coincidencia de forma significativa para el resto de las transformaciones, especialmente en la novena y la décima con unos 28 % y 36 %, respectivamente, por debajo de 75 %, indicando que el sistema experimenta dificultades ante los cambios aplicados. En suma, cuando se trata de voces activas y pasivas en francés o al substituir una palabra por un pronombre personal o un sinónimo además de cambiar el orden de las oraciones. MemoQ, vuelve a presentar grados de coincidencias parciales más bajos a los de SDL salvo para las transformaciones 6, 8 y 9, con resultados similares y para la trasformación 10, donde registramos el grado más bajo con 54 % de los segmentos transformados < 75%, indicando que MemoQ afronta más dificultades ante los cambios, peor aun cuando son dos o tres cambios a la vez, como por ejemplo en la sustitución por dos

sinónimos, un pronombre, cambio de orden, transformación de activa a pasiva además de una sustitución por un pronombre.

	SDL		MEMOQ	
	ST		ST	
N. °	<75 %	> 75 %	<75 %	>75 %
1	12 %	**88 %**	32 %	**68 %**
2	12 %	**88 %**	36 %	**64 %**
3	4 %	**96 %**	12 %	**88 %**
4	0 %	**100 %**	4 %	**96 %**
5	4 %	**96 %**	12 %	**88 %**
6	24 %	**76 %**	24 %	**76 %**
7	0 %	**100 %**	4 %	**96 %**
8	4 %	**96 %**	4 %	**96 %**
9	28 %	**72 %**	28 %	**72 %**
10	36 %	**64 %**	**54 %**	46 %

Tabla 7: Resultados de las coincidencias FR-ES

SDL presenta resultados moderados para las coincidencias FR-ES, registrando un número cambiante según cada categoría pero muy elevado con 25 segmentos ubicados entre 90 %-99 % en la cuarta transformación(es decir, el total de los segmentos transformados), seguidos por 18, 15 y 13 segmentos para las transformaciones 5, 7, 8 y 1, confirmando una vez más que el sistema se comporta mejor pese a los leves descensos cuando se trata de sustitución sin cambio de orden de las palabras, enunciados, frases u oraciones, pero menos bien al cambiar de orden como para la tercera transformación con 13 segmentos ubicados entre 80 %-89 % y 12 segmentos entre 70 %-79 % al tratarse de una operación de múltiples cambios. MemoQ presenta a su vez resultados aún más dispares, donde la mayoría de los segmentos se sitúan en la segunda y la tercera categoría con grados de coincidencia más bajos; incluso ciertos segmentos registran un descenso significativo en comparación con los resultados de SDL, con 6 segmentos ubicados entre 60 %-69 % para la transformación 10 y 4 para las transformaciones 2 y 9.

	SDL										MEMOQ									
GRADO	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
90 %-99 %	**13**	7	10	**25**	**18**	5	**15**	**13**	5	**11**	1	1	3	4	0	0	7	3	1	0
80 %-89 %	7	**10**	**13**	0	4	**11**	7	4	8	5	7	**10**	**18**	**18**	**20**	9	**13**	**18**	2	4
70 %-79 %	5	8	2	0	3	9	3	8	**12**	9	**14**	**10**	3	3	4	**15**	5	4	**18**	**15**
60 %-69 %	0	0	0	0	0	0	0	0	0	0	3	4	1	0	1	1	0	0	4	6

Tabla 8: Resultado de las coincidencias FR-ES por segmentos

4.3. Árabe-español

Los resultados de las coincidencias como aparecen indicados en la tabla 9 descienden para todos, pero no tanto para SDL, especialmente para las transformaciones 4, 7 y 8, donde registramos 4 % de los segmentos transformados por debajo de 75 %, lo que indica que el sistema se comporta bastante bien ante la sustitución por sinónimo o por pronombre personal sin o con cambio de orden de las palabras. De igual manera, se registran unos grados de coincidencia más bajos como, por ejemplo, para la tercera trasformación con 16 % del total de los segmentos trasformados por debajo de 75 %, indicando que SDL afronta ciertas dificultades al cambiar el orden de las palabras, frases, enunciados u oraciones. Tampoco parece soportarla transformación de voz activa-pasiva acompañada de sustitución por sinónimo. MemoQ presenta un descenso significativo en los grados de coincidencias parciales con 44 % de los segmentos transformados <75% para la décima transformación de contraste con 8 % para SDL, demostrando que el sistema sufre más dificultades cuando se trata de una operación múltiple (voz activa-pasiva y sustitución por un pronombre personal). Asimismo, afronta dificultades ante el cambio de orden de las palabras con 40 % de los segmentos <75% pese a que el idioma árabe es flexible y permite estructurar una frase con menos restricciones mediante estructuras VSO/SVO/VOS, para el resto de las transformaciones, siguen descendiendo en comparación con SDL excepto para la quinta y la sexta con resultados similares a los de SDL, concluyendo que MemoQ tiene más dificultades ante los cambios sintáctico-semánticos.

	SDL		MEMOQ	
	ST		ST	
N. °	<75 %	>75 %	<75 %	>75 %
1	12 %	**88 %**	20 %	**80 %**
2	8 %	**92 %**	24 %	**76 %**
3	16 %	**84 %**	40 %	**60 %**
4	4 %	**96 %**	32 %	**68 %**
5	16 %	**84 %**	4 %	**96 %**
6	20 %	**80 %**	16 %	**84 %**
7	4 %	**96 %**	24 %	**76 %**
8	4 %	**96 %**	16 %	**84 %**
9	16 %	**84 %**	28 %	**72 %**
10	8 %	**92 %**	44 %	**56 %**

Tabla 9: Resultados de las coincidencias AR-ES

Los resultados por categorización de las coincidencias (véase tabla 10) permiten determinar dónde se sitúan los segmentos más afectados por las transformaciones. SDL, por tercera vez, presenta un leve descenso en contraste con MemoQ para la cuarta transformación, donde la mayoría de las coincidencias parciales se sitúan entre 90 %-99 % para SDL, con un total de 18 segmentos en contraste con 15 para MemoQ. El grado de coincidencia sigue bajando con ambos SDL y MemoQ, donde la mayoría de los segmentos se sitúan entre 70 %-79 %, confirmando que estos sistemas afrontan dificultades lingüísticas. Sin embargo, se observa que no hay ningún segmento que se sitúe entre los grados 60 %-69 % para SDL con todas las transformaciones, pero sí que hay varios segmentos en esta misma categoría para MemoQ, especialmente para la décima transformación con 5 segmentos o 4 para ambas (primera y tercera transformación), indicando que el descenso se debe a las dificultades lingüísticas de transformación activa-pasiva, sustitución por un pronombre personal y cambio de orden de las palabras. Se observa por primera vez la integración de una nueva categoría para el idioma árabe, esta categoría representada por un cero significa que el sistema no ofrece ninguna coincidencia como es el caso de las transformaciones 7, 8, 9 y 10 con uno y dos segmentos sin ninguna coincidencia con MemoQ pero no con SDL, por lo que el primer sistema experimenta obviamente más dificultades, con grados aún más bajos.

	SDL										MEMOQ									
GRADO	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
90 %-99 %	**13**	**17**	**13**	**18**	**13**	**11**	**12**	4	8	10	0	**10**	1	**15**	7	2	3	0	0	0
80 %-89 %	6	5	7	4	8	3	10	**17**	**11**	**12**	10	7	**10**	4	**15**	**11**	**15**	**14**	**11**	**9**
70 %-79 %	6	3	5	3	4	**11**	3	4	6	3	**11**	7	**10**	6	2	**11**	6	9	10	**9**
60 %-69 %	0	0	0	0	0	0	0	0	0	0	4	1	4	0	1	1	0	1	2	5
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	2

Tabla 10: Resultado de las coincidencias AR-ES por segmentos

5. Conclusiones y trabajos futuros

Hemos presentado en esta investigación un estudio cuantitativo comparando la traducción de tres pares de idiomas mediante el uso de dos sistemas de TAO. La aplicación de diez tipos de transformaciones semántico-sintácticas indica que ambos sistemas sufren dificultades y tienen un comportamiento distinto ante los cambios pese a nuestro intento de guardar siempre el mismo sentido. Para el español como idioma de origen, el grado de coincidencia registra más descenso con MemoQ y baja significativamente cuando se cambia la voz activa en pasiva y se sustituye una palabra por su sinónimo, mientras que SDL registra su mayor descenso con la transformación 6 al tratarse de una operación múltiple de tres cambios. Las coincidencias parciales presentadas para las transformaciones 4, 6, 7 y 8 registran un descenso menor con MemoQ, indicando que estas mismas (sustitución por un sinónimo con o sin cambio de orden de las palabras, sustitución por un pronombre con o sin cambio de orden de las palabras) no afectan tanto al grado de las coincidencias parciales como lo hacen con SDL.

En cuanto al idioma francés, ambos sistemas registran su grado de coincidencia más bajo con la décima trasformación al cambiar la voz activa en pasiva y sustituir una palabra por un pronombre personal. No obstante, ambos sistemas marcan similitudes en los grados de las coincidencias parciales: sustitución por sinónimo, pronombre personal, transformación activa-pasiva. Para el resto de las transformaciones, los grados de coincidencias parciales no bajan significativamente con SDL, mientras que descienden mucho con MemoQ, indicando que este último afronta dificultades importantes.

Para el árabe como idioma de origen, el mayor descenso se registra para la sexta trasformación con SDL y la décima con MemoQ, es decir, que

el primer sistema experimenta dificultades ante la sustitución por sinónimos y el cambio de orden de las palabras, mientras al segundo sistema le cuesta afrontar la transformación activa-pasiva con sustitución por un pronombre personal. De igual modo, las coincidencias parciales bajan de forma significativa con MemoQ, que no parece soportar el cambio de orden de las palabras en árabe, aunque el idioma semítico es flexible y permite estructurar una frase/oración con menos restricciones que los idiomas español y francés. En definitiva, todos los grados de coincidencia y en todas las combinaciones lingüísticas bajan y difieren según el tipo de la transformación y el sistema empleado en suma con MemoQ, unos segmentos descienden más en árabe como idioma de origen, luego en español y por último en francés, concluyendo que al ser el español y el francés idiomas no tan lejanos como el árabe y por sus sistemas lingüísticos más cercanos se ven menos impactados por las transformaciones semántico-sintácticas que el árabe, aunque se confirma que las coincidencias se ven afectadas al no prestar atención al sentido puesto que estos sistemas funcionan con un algoritmo matemático. Por lo tanto, se necesita integrar más conocimientos lingüísticos o procesamiento del lenguaje natural con el último objetivo de mejorar el grado de las coincidencias y reducir el esfuerzo o el tiempo dedicado a la post- edición.

Nuestra investigación afronta ciertas limitaciones en cuanto al corpus: segmentación y alineación de los segmentos seleccionados, pre- y post-edición, e imposibilidad de calcular el grado de coincidencia para los mismos segmentos en todos los idiomas. Sin embargo, esta primera fase de investigación nos permitió analizar datos empíricos y obtener los primeros resultados. Las investigaciones futuras se centrarán en calificar más datos mediante el uso de otros sistemas de TAO con una evaluación humana más amplia, idealmente por nativos españoles, franceses y árabes para establecer un listado de propuestas y de mejoras propio de cada idioma y sistema de TAO.

Referencias bibliográficas

ALUBUDI, J. (2008): *Árabe culto*. Madrid, España: Liber Factory.

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

CHATZITHEODOROU, K. (2015): “Improving translation memory fuzzy matching paraphrasing”. *Proceedings of the Workshop on Natural Language Processing for Translation Memories*, 24-30.

GARCÍA YERBA, V. (1990): “La voz pasiva francesa y su traducción al español”. *Meta* 35 (3), 510-517.

GUPTA, R., CONSTANTIN, O., ZAMPIERI, M., VELA, M. y VAN GENABITH, J. (2015): "Can Translation Memories afford not to use paraphrasing?" *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*,18, 35-42.

HAMMO, B., MOUBAIDDIN, A., OBEID, N, y TUFFAHA, A. (2014): "Formal Description of Arabic Syntactic Structure in the Framework of the Government and Binding Theory". *Computación y Sistemas*, 18 (3), 611-625.

MCENERY, T. (2016): "Corpora"*. The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 1-19.

OLIVER GONZÁLEZ, A. (2016): *Herramientas tecnológicas para traductores. Capítulo I. La traducción asistida por ordenador*. Barcelona, España: Editorial UOC.

SILVESTRE BAQUERO, A y MITKOV, R. (2017): "Translation Memory Systems Have a Long Way to Go". The *Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)*, 44-51.

SIMARD, M. (2020): "Building and using parallel text for translation". In: Minako O'Hagan (ed)*. The Routledge Handbook of Translation and Technology*. London: Routledge, 78 -90.

THARINDU, R., CONSTANTIN, O., MITKOV, R. (2020): "Intelligent Translation Memory Matching and Retrieval with Sentence Encoders". *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 175-184.

WEBER, E. (2014): "La traducción de la voz pasiva francesa al español: ¿cuestión de lengua o cuestión de traducción?". *La traductología en Brasil (2). Mutatis Mutandis. 7 (2)*, 368 -385.

ZULMA, I. y DOZIER, E. (2014): *Manual de gramática.* Boston, EE. UU.: Heinle Cengage Learning.