TESIS DOCTORAL

PROGRAMA DE DOCTORADO EN LENGUAS, TEXTOS Y CONTEXTOS

TRADUCCIÓN E INTERPRETACIÓN

UNIVERSIDAD
DE GRANADA

Traducción al inglés y al español de textos árabes en las Naciones Unidas: transformaciones discursivas mediante división de oraciones

Juan José Roldán Verdejo

Director: Manuel Feria García

Granada, 2022

Editor: Universidad de Granada. Tesis Doctorales
Autor: Juan José Roldán Verdejo
ISBN: 978-84-1117-555-5
URI: https://hdl.handle.net/10481/77668

Agradecimientos

Me gustaría expresar mi más sincero agradecimiento a todas las personas que de una u otra forma han contribuido a que esta tesis pueda completarse y, en especial, al padre de la idea, el Dr. Manuel Feria, sin cuyo esfuerzo y dedicación no hubiera podido terminar este trabajo.

Me gustaría asimismo dar las gracias a mis padres y, en general, a mis dos familias, la española y la tunecina, por el apoyo moral y material que me han prestado, y de manera especial a Mouna, mi mujer, y a mi hija Sara, por haber sido tan pacientes conmigo durante estos años.

Muchas gracias a todos.

Índice de contenidos

Índice de contenidos ..... **iii**
Índice de tablas ..... **vi**
Índice de gráficos ..... **viii**

Capítulo I: Introducción **1**
1. Objetivos 1
2. Justificación 4
3. Contextualización: la lengua árabe en las Naciones Unidas 4
4. Preguntas de investigación, hipótesis nula y aplicaciones 7
5. Metodología 8
6. Estructura de la tesis 8
7. Consideraciones ortotipográficas y formales 9

Capítulo II: Estado de la cuestión **10**
1. Traductología 10
2. Lingüística Computacional 16
3. Conclusiones 20

Capítulo III: Corpus y metodología **23**
1. Corpus 23
1.1. Corpus 1 24
a) Origen, originalidad y representatividad 24
b) Accesibilidad, formato y confianza 25
c) Delimitación 26
d) Compilación 28
e) Distribución de lenguas 29
1.2. Corpus 2 34
1.3. Corpus 3 36
1.4. Corpus 4 38
2. Metodología 39
2.1. Traducción del árabe al español y al inglés 39

a) Consideraciones preliminares 39
b) Selección de la muestra 40
c) Eliminación de los elementos de distorsión 44
d) Segmentación 47
2.2. Traducción del español y del inglés al árabe 48

Capítulo IV: Datos y análisis de los datos **50**
1. Datos 50
a) Número de oraciones 50
b) Número de palabras gráficas 52
c) Media de palabras por oración 54
d) *Stopwords* 56
2. Análisis de datos 68
a) Representatividad 68
b) Número de oraciones 69
c) Número de palabras 70
d) Media de palabras por oración 71
e) Direccionalidad 73
Número de oraciones 74
Número de palabras 75
Media de palabras por oración 76
Conclusiones 77
f) Stopwords 78

Capítulo V: Conclusiones **83**
1. Principales resultados y aportaciones 83
a) Refutación de la hipótesis nula 83
b) Determinación de las stopwords 86
2. Líneas futuras de investigación 87

Bibliografía **89**

Anexo I: Composición del Corpus 1 con datos desglosados por signatura, Estado, número de palabras y lengua original. 99

Anexo II: Composición del Corpus 2 con datos desglosados por signatura, Estado, año y número de palabras en árabe, español e inglés 110

Anexo III: Relación de las 996 stopwords localizadas en las traducciones del árabe al español con datos desglosados por documentos y segmentos.....115

Anexo IV: Relación de las 983 stopwords localizadas en la traducción del árabe al inglés, con datos desglosados por documentos y segmentos (documento_segmento). 125

Anexo V: Relación de las 647 stopwords comunes localizadas en las traducciones del árabe al español y al inglés con datos desglosados por documentos y segmentos.....135

Anexo VI: Relación de las 100 stopwords más ocurrentes en la muestra segmentada. 142

Índice de tablas

Tabla 1: Número máximo de palabras en una oración en inglés y árabe, sus ratios máximas y su ratio media en el corpus no procesado de Salameh et al. (2011) *19*

Tabla 2: Tabla 2. Distribución por lenguas en el Corpus 1, con datos desglosados por número de palabras y documentos *29*

Tabla 3: Tabla 3. Distribución por lenguas en el Corpus 1, con datos desglosados por Estado y número de palabras *30*

Tabla 4: Contribución al Corpus 1, con datos desglosados por Estado y palabras (en términos absolutos y porcentuales) *32*

Tabla 5: Evolución cronológica de la aportación por lenguas al Corpus 1, en términos porcentuales *33*

Tabla 6: Aportación de cada Estado al Corpus 2, con datos desglosados por número de palabras árabes y porcentaje correspondiente del total de palabras árabes. *36*

Tabla 7: Composición del Corpus 3 ..... *37*

Tabla 8: Composición del Corpus 4 ..... *38*

Tabla 9: Composición de la muestra, con datos desglosados por número de palabras gráficas del original y sus traducciones y por aportación a la muestra en palabras gráficas por informe y Estado *41*

Tabla 10: Documentos con posible traducción intermediada al español que fueron descartados de la muestra *42*

Tabla 11: Número de oraciones y media de palabras por oración en las tres fases de preprocesamiento de la muestra *46*

Tabla 12: Desviaciones típicas de las medias de palabra por oración y su variación en las tres fases de preprocesamiento de la muestra *47*

Tabla 13: Número de segmentos por documento *48*

Tabla 14: Número de oraciones en árabe, español e inglés en la muestra del Corpus 2 (recuento automático) *50*

Tabla 15: Número de oraciones en árabe, español e inglés en la muestra segmentada del Corpus 2 (recuento manual) *51*

Tabla 16: Número de oraciones en árabe y español en la muestra del Corpus 3 (recuento automático).....51

Tabla 17: Número de oraciones en árabe e inglés en la muestra del Corpus 4 (recuento automático).....52

Tabla 18: Número de palabras gráficas en árabe, español e inglés en el Corpus 2, y sus respectivas ratios.....52

Tabla 19: Número de palabras gráficas en árabe, español e inglés en la muestra del Corpus 2, y sus respectivas ratios.....53

Tabla 20: Número de palabras gráficas en la muestra segmentada del Corpus 2, y sus respectivas ratios.....53

Tabla 21: Número de palabras gráficas en la muestra del Corpus 3.....54

Tabla 22: Número de palabras gráficas en la muestra del Corpus 4 54

Tabla 23: Media de palabras por oración en la muestra del Corpus 2.....55

Tabla 24: Media de palabras por oración en la muestra del Corpus 2 segmentada.....55

Tabla 25: Media de palabras por oración en la muestra del Corpus 3.....56

Tabla 26: Media de palabras por oración en la muestra del Corpus 4.....56

Tabla 27: Número de segmentos por documento, número de stopwords localizadas en la traducción al español y al inglés y ratios de STW E/SEG, STW I/SEG y STW E/STW I.....57

Tabla 28: Número total de stopwords comunes por documento y sus porcentajes respecto al total de stopwords del documento 58

Tabla 29: Stopwords con PA >50 y OA >1, con mención de su OT, OA y D, en las traducciones al español 59

Tabla 30: Stopwords con PA >50 y OA >1, con mención de su OT, OA y D, en las traducciones al ingles.....60

Tabla 31: Stopwords con D >50 y OA >1 en las traducciones al español, con mención de sus OT, OA, PA y posición en la lista de palabras más frecuentes en la muestra segmentada ...62

Tabla 32: Stopwords con D >50 y OA >1 en las traducciones al inglés, con mención de sus OT, OA, PA y posición en la lista de palabras más frecuentes en la muestra segmentada.....63

Tabla 33: Stopwords con D>20 y PA>20 en las traducciones al español 64

Tabla 34: Stopwords con D>20 y PA>20 en las traducciones al inglés.....64

*Tabla 35: Stopwords comunes con datos desglosados por número de ocurrencias comunes (OC), ocurrencias activas en español (OA E) y en inglés (OA I) y porcentaje de coincidencia respecto al número de ocurrencias activas de la stopword en cada lengua (% C E y % C I).*65

Tabla 36: Stopwords comunes a las traducciones al español y al inglés con D>20 y PA>20...66

Tabla 37: Stopwords introducidas por wa en las traducciones al español, con datos desglosados por types y tokens 67

Tabla 38: Stopwords introducidas por wa en las traducciones al inglés, con datos desglosados por types y tokens 67

Tabla 39: Stopwords comunes introducidas por wa, con datos desglosados por types y tokens 68

Índice de gráficos

Gráfico 1: Distribución por lenguas en el Corpus 1, con datos desglosados por Estado y número de palabras en términos porcentuales 31

Gráfico 2: Evolución cronológica del número de palabras por lengua 34

Capítulo I: Introducción

1. Objetivos

El interés por la traducción del y al árabe se ha incrementado muy considerablemente en las últimas décadas. Entre las razones se cuenta el mayor protagonismo de esa lengua en términos económicos, políticos y geoestratégicos en la comunidad internacional, inclusive en la Organización de las Naciones Unidas (Sánchez-Ratia, 2002; Arias y Feria, 2012, p. 421, nota 234; Avión, 2013), así como su mayor presencia en la vida diaria española (Ilhami, 2015, 9).

Fruto de ese interés fue la creación de instituciones como la Escuela de Traductores de Toledo (1994) y Casa Árabe (2006) y la inclusión del árabe como lengua B en la Licenciatura en Traducción e Interpretación en la Universidad de Granada (2002). Sin embargo, la falta de investigación en la materia y la necesidad de seguir mejorando la formación de los traductores del árabe (Lafeber, 2012, p. 265), en particular de los traductores del árabe al español (Ilhami, 2015, pp. 313-314), siguen siendo acuciantes.

Uno de los motivos que explican el escaso avance en la investigación en la materia es el escaso desarrollo de la Lingüística Computacional árabe. Ese escaso desarrollo se refleja negativamente en la eficiencia de las herramientas para la traducción automática o asistida del árabe, incluso de las estadísticas, que son muy dependientes de la eficiencia de las herramientas de segmentación.

Según la literatura previa, una diferencia discursiva notable entre el árabe y las otras dos lenguas analizadas radica en la extensión de las oraciones, lo que se refleja en el plano sintáctico. Como señala Keskes,

> *The abundance of coordination in written Arabic texts makes short sentences very rare to exist. Arab writers tend to write very long sentences, some of which could be a paragraph long with one full stop at the end. (Keskes, 2015, p. 38).*

El fenómeno induce a menudo a los traductores a dividir esas oraciones para mejorar la legibilidad en la lengua de llegada. De este recurso, que conocíamos por nuestra experiencia como traductores, es buen ejemplo la traducción del siguiente párrafo extraído de CRC/C/OPSC/IRQ/1, documento presentado por el Iraq al Comité de los Derechos del Niño:

71- فيما يتعلق بالسكن تسكن العوائل المهجرة في مخيمات كما يوجد (12) مخيماً آخر موزعة على بقية المحافظات ويتراوح عدد الخيم في المخيم الواحد ما بين (45-100) خيمة، وتبعاً لإحصاء منظمة الهجرة الدولية فإن (11 في المائة) من النازحين في بغداد و(22 في المائة) من النازحين في محافظة القادسية و(30 في المائة) في الأنبار يعيشون في بنايات عامة مهجورة أو مهدمة، ويعيش الأطفال مع أسرهم في أماكن غير لائقة بلغت نسبة الاكتظاظ (88 في المائة) اعتماداً على معيار (أكثر من ثلاثة أشخاص في الغرفة الواحدة) مع افتقارها إلى الخدمات ولسوء هذه الظروف المعيشية أدى إلى تسرب الأطفال من المدارس ونزولهم إلى سوق العمل مبكراً بالإضافة إلى تعرض بعضهم إلى التشرد.

71. En la esfera de la vivienda conviene recordar que las familias desplazadas viven en campamentos, y que hay otros 12 campamentos repartidos por las diferentes provincias del país. Cada campamento dispone de entre 45 y 100 tiendas de campaña. Según la Organización Internacional para las Migraciones (OIM), el 11% de los refugiados que se encuentran en Bagdad, el 22% de los que se encuentran en la provincia de Qadisiya y el 30% de los que se encuentran en al-Anbar habitan edificios públicos abandonados o derruidos. Los niños viven junto a sus familias en lugares inadecuados y carentes de servicios. El 88% de estas viviendas sufre una situación de hacinamiento (hogares con más de tres personas por habitación). Estas condiciones de vida han conducido a los niños al abandono escolar y a la integración precoz en el mercado laboral, y a algunos incluso los ha arrastrado a la mendicidad callejera.

71. With respect to housing, displaced families are accommodated in camps and there are also 12 other camps dispersed throughout the remaining governorates. The number of tents per camp ranges from 45 to 100. According to statistics of the International Organization for Migration (IOM), 11 per cent of displaced persons are in Baghdad, 22 per cent are in Qadisiyah governorate and 30 per cent are in Anbar, living in abandoned or demolished public buildings. Children live with their families in unsuitable places, where overcrowding stands at 88 per cent, based on the criterion of more than three individuals per room, and services are lacking. On account of these poor living

> standards, children drop out of school and move into the job market at an early age. Some are also at risk of homelessness.

Como podemos observar, el párrafo original consiste en una sola oración en árabe de 109 palabras gráficas que en las traducciones al español y al inglés se ha dividido, respectivamente, en seis y cinco oraciones.

La tendencia a la división de oraciones en la traducción del árabe se había destacado sin pruebas empíricas en la literatura previa (Zantout y Guessoum, 2015, p. 248; Dickins, Hervey y Higgins, 2017, p. 183). Ante ese nicho de investigación determinamos los objetivos de investigación siguientes:

a) Testar empíricamente la supuesta tendencia del árabe a formar oraciones extensas que tienden a ser divididas cuando se traduce al español o al inglés, y ello mediante un análisis comparativo (cuantitativo y cualitativo) de originales árabes y sus traducciones realizadas por profesionales de primer nivel;
b) Testar si la direccionalidad de la traducción tiene un impacto en esa tendencia, es decir, si además de dividirse oraciones al traducir del árabe se tiende a fusionar oraciones al traducir al árabe;
c) Si en efecto quedara demostrado a), detectar las principales marcas lingüísticas elegidas por los traductores para dividir las oraciones del original (*stopwords*) o, dicho de otro modo, aquellas que probablemente generarían punto en un texto que se deseara traducir del árabe.

Nuestra investigación, por tanto, pretende contribuir al conocimiento del proceso en la traducción árabe-español-inglés mediante el análisis empírico del producto, en particular al conocimiento de las estrategias de adaptación discursiva basadas en la introducción de puntos. Ello redundaría en beneficio de la Lingüística Computacional árabe al contribuir a la mejora de las herramientas automáticas de segmentación (y con ello, de las herramientas de traducción asistida y de traducción automática), así como tendría aplicaciones para la enseñanza de la traducción del árabe.

2. Justificación

La mayoría de la literatura previa sobre la traducción del árabe se orienta al análisis de los elementos culturales. Apenas existen estudios empíricos sobre el producto. Nuestra investigación es la primera que proporciona y analiza datos cuantitativos sobre transformaciones discursivas en la traducción árabe-español-inglés, en concreto sobre el número de oraciones y palabras y la media de palabras por oración, así como sobre sus respectivas ratios, en los originales árabes y sus traducciones a las lenguas analizadas, así como en los originales en español y en inglés y sus traducciones al árabe. También determinamos las marcas lingüísticas de los originales árabes que con mayor frecuencia originan puntos en las traducciones y que, por tanto, tienen mayores posibilidades de generar puntos en nuevos textos que se deban traducir. Esos datos empíricos se extrajeron de diversos corpus confeccionados *ex profeso* y que cumplen todos los requisitos para su análisis mediante técnicas de Lingüística de Corpus.

La elección de la combinación lingüística responde a que se trata de lenguas de gran relevancia internacional y que generan un gran volumen de traducciones. El análisis se basó en textos de las Naciones Unidas, ya que la organización proporciona un corpus fiable de originales y traducciones. Además, el tipo de texto elegido presenta características muy determinadas de registro, género y situación traslativa que permite interpretar las potenciales diferencias interlingüísticas de forma muy controlada.

3. Contextualización: la lengua árabe en las Naciones Unidas

El árabe, entendido como una macrolengua con variedades regionales, cuenta con unos 350 millones de hablantes nativos y no nativos, lo que la convierte en la quinta lengua del mundo por número de hablantes (Ethnologue, 24th ed., 2021).

En la actualidad, el árabe es lengua oficial o cooficial en 24 Estados miembros de Naciones Unidas y en 2 Estados con reconocimiento limitado (Estado de Palestina y República Árabe Saharaui Democrática). El árabe es una de las seis lenguas oficiales de las Naciones Unidas y lengua oficial en otros organismos internaciones, como la Unión Africana, la Organización para la Cooperación Islámica y la Liga Árabe.

El 18 de diciembre de 1973, la Asamblea General de las Naciones Unidas aprobó incluir el árabe entre las lenguas oficiales y de trabajo de la Asamblea General y de sus

comisiones principales. La Asamblea General reconocía así la importancia de una lengua que ya era lengua de trabajo en la UNESCO, la FAO, la OMS y la OIT, además de lengua oficial y de trabajo de la Organización para la Unidad Africana (predecesora de la Unión Africana)[1].

En 1980, la Asamblea General aprobó hacer extensivo el uso del árabe como lengua oficial y de trabajo al resto de comisiones y subcomisiones de las Naciones Unidas −igualándose de este modo a las otras cinco lenguas oficiales− a partir del 1 de enero de 1982. Por medio de la misma resolución[2], la Asamblea General solicitó al Consejo de Seguridad que incluyera el árabe entre sus lenguas oficiales y de trabajo y al Consejo Económico y Social que lo incluyera entre sus lenguas oficiales antes del 1 de enero de 1983. El Consejo Económico y Social adoptó el árabe como lengua oficial el 15 de abril de 1982[3] y el Consejo de Seguridad lo incluyó entre sus lenguas oficiales y de trabajo el 21 de diciembre del mismo año[4].

Pese al estatus del árabe como lengua oficial de las Naciones Unidas desde 1973 y de las diferentes campañas para la promoción del multilingüismo (Tafalla, 2010, p. 154), los Estados árabes no siempre utilizan esa lengua en los documentos que presentan a las Naciones Unidas. Como tendremos ocasión de comprobar, algunos solo recurren al francés y otros conjugan el uso del árabe con el del francés y/o el inglés.

Pese a la importancia del multilingüismo para las Naciones Unidas y al ingente volumen de traducción que la organización genera, las investigaciones en la materia son muy escasas, como podremos comprobar. Particularmente escasa es la investigación relativa a la traducción del árabe, pese a que en los últimos años esa lengua se ha convertido en la segunda más importante para la Organización, tras el inglés, por el volumen de documentación para traducir que genera (Sánchez Ratia, 2002). Los pocos estudios existentes han sido realizados por traductores que trabajan o han trabajado en las secciones de español de Naciones Unidas, como los realizados por Sánchez-Ratia (2002), García Verdugo (en Pouliquen et al., 2012 y Pouliquen et al., 2013) y Feria (2013; Sainz-Quinn y Feria, 2020).

Por otro lado, la demanda de traductores de árabe en las Naciones Unidas es acuciante. Sabela Avión Martínez, traductora en el Servicio de Traducción desde 2007 y revisora

[1] Resolución 3190 (XXVIII) http://www.un.org/ga/search/view_doc.asp?symbol=a/res/3190(XXVIII).
[2] Resolución 35/219 del 17 de diciembre de 1980 http://www.un.org/documents/ga/res/35/a35r219e.pdf.
[3] Decisión 1982/147 http://www.un.org/en/ga/search/view_doc.asp?symbol=E/1982/82, p. 48.
[4] Resolución 528 (1982) http://unscr.com/en/resolutions/528.

desde 2014, por ejemplo, recomienda encarecidamente a los estudiantes españoles que tienen como meta trabajar en las Naciones Unidas que estudien ruso, chino y árabe (Avión, 2013, minuto 18).

Un claro indicio de la escasez de traductores de árabe en Naciones Unidas es el recurso a la intermediación a través del inglés o el francés (*relay*, en adelante, *relé*). Sánchez-Ratia, traductor de árabe y revisor de la Sección de Traducción de la ONU en Ginebra, afirma que "la escasez de traductores de árabe, combinada con el crecimiento exponencial de los documentos originales en este idioma, han obligado a muchos traductores del servicio a realizar traducciones por *relay*, es decir, a través de un tercer idioma" (Sánchez Ratia, 2002).

Anne Lafeber, revisora del Servicio de Traducción de Inglés en Nueva York, por su parte, afirma que los traductores de la Sección de Árabe son los únicos que no se limitan a traducir hacia su lengua principal habida cuenta la escasez de traductores de árabe (Lafeber, 2012, p. 97). Cao y Zhao también incluyen a los traductores al chino entre los traductores de las Naciones Unidas que realizan traducción inversa (Cao y Zhao, 2008, p. 43).

Tanto la traducción inversa como el *relé* pueden tener un efecto negativo en la calidad de las traducciones. La UNESCO, "con miras a mejorar la calidad de las traducciones", afirma que, "por regla general, la traducción debe hacerse a partir del original, recurriéndose a la retraducción solamente en caso de que sea absolutamente necesario", y que "en la medida de lo posible, el traductor debe traducir a su lengua materna o a un idioma que domine como su lengua materna" [5].

La literatura previa que demostrado que el recurso al *relé* en la traducción del árabe entraña problemas de homogeneidad terminológica (Sainz-Quinn y Feria, 2020) y de traducción de elementos culturales particularmente sensibles, como conceptos idiosincráticamente islámicos de gran relevancia para los derechos humanos (Feria 2013, p. 47), que debieran evitarse. Nuestra investigación también ha contribuido a conocer mejor el fenómeno del *relé* en la traducción del árabe al español en las Naciones Unidas.

[5] Recomendación de la UNESCO aprobada el 22 de noviembre de 1976 sobre la protección jurídica de los traductores y de las traducciones y sobre los medios prácticos de mejorar la situación de los traductores, p.165. http://unesdoc.unesco.org/images/0011/001140/114038s.pdf#page=157

4. Preguntas de investigación, hipótesis nula y aplicaciones

En definitiva, nuestro trabajo pretende responder a las siguientes preguntas de investigación:

a) ¿Las traducciones del árabe al español y al inglés comprenden más oraciones que los originales?
b) ¿Las traducciones del español y el inglés al árabe comprenden menos oraciones que sus originales?
c) ¿Cuáles son las marcas lingüísticas del texto árabe que suelen elegir con mayor frecuencia los traductores al español y al inglés para dividir las oraciones?

La hipótesis nula que se plantea en la presente investigación es que al traducir del árabe al español y al inglés la introducción de puntos para la división de oraciones no es un procedimiento estadísticamente significativo o, dicho con otras palabras, que las traducciones del árabe al español y al inglés no comprenden un número de oraciones significativamente mayor que el de los originales árabes.

Si se rechaza la hipótesis nula, analizaremos el posible papel de la direccionalidad sobre la división de oraciones y determinaremos las principales marcas lingüísticas o *stopwords* del texto árabe elegidas con mayor frecuencia por los traductores para introducir los puntos.

Desde una perspectiva teórica, la refutación de la hipótesis nula contribuiría a desarrollar, entre otras disciplinas, la Lingüística Contrastiva, la Lingüística del Discurso, la Lingüística Computacional árabe y la Traductología.

Desde un punto de vista práctico, la refutación de la hipótesis nula podría tener aplicaciones en la formación de traductores árabe-español-inglés al permitir determinar empíricamente objetivos específicos de formación relacionados con la transformación discursiva. De ese modo contribuiríamos a paliar la escasez de materiales de aprendizaje de calidad diseñados específicamente para traductores del árabe al español y derivados de investigaciones lingüísticas basadas en datos cuantitativos bien fundados (Feria, 2014, 214-215).

Refutar la hipótesis nula también podría tener aplicaciones prácticas para la Lingüística Computacional dado que, si pudiéramos determinar estadísticamente las palabras gráficas que originan divisiones con mayor frecuencia, podríamos mejorar el

rendimiento de las herramientas de segmentación, y con ellas, la alineación automática y la calidad de las memorias de traducción.

5. Metodología

Con objeto de testar nuestra hipótesis nula y responder a las preguntas de investigación compilamos y analizamos varios corpus de textos de las Naciones Unidas y sus traducciones institucionales.

En primer lugar se compiló y analizó un corpus paralelo trilingüe (Corpus 2) formado por originales en árabe y sus traducciones institucionales al español y al inglés. Para ello se emplearon criterios y métodos propios de la Lingüística de Corpus.

A continuación, para comprobar la potencial influencia de la direccionalidad, compilamos y analizamos dos corpus paralelos (Corpus 3 y 4) formados, respectivamente, por originales en español y sus traducciones al árabe y por originales en inglés y sus traducciones al árabe.

Para la extracción de datos se empleó la herramienta WordSmith y algunas herramientas integradas en Word. Con todo, la mayor parte de los datos tuvieron que ser extraídos de forma manual, para lo que fue necesario segmentar una muestra representativa del Corpus 2.

Finalmente, mediante la comparación de los originales y sus traducciones, se determinaron de forma manual las principales marcas lingüísticas o *stopwords* empleadas por los traductores del árabe al español y al inglés para dividir las oraciones de los originales.

6. Estructura de la tesis

La presente tesis consta de cinco capítulos:

- En el Capítulo I (“Introducción”) se presenta, justifica y describe de forma resumida nuestra investigación;

- En el Capítulo II (“Estado de la cuestión”) se revisa la literatura previa que, de manera directa o tangencial, ha abordado el fenómeno de la división de oraciones en el proceso de traducción;

- En el Capítulo III (“Corpus y metodología”) se describen los corpus en los que se basa la investigación, así como la metodología empleada para la selección de la muestra, la eliminación de los elementos de distorsión y la extracción de los datos;

- En el Capítulo IV (“Datos y análisis de los resultados”) se presentan los datos extraídos de los tres corpus y se analizan los datos obtenidos;

- En el Capítulo V (“Conclusiones”) se resumen las principales conclusiones de nuestra investigación, se subrayan sus aportaciones al estado de la cuestión y se presentan potenciales líneas futuras de investigación.

Tras esos cinco capítulos se incluyen las referencias bibliográficas y algunos anexos.

7. Consideraciones ortotipográficas y formales

En esta tesis seguiremos la ortografía recomendada por UNTERM para denominar a los Estados cuando exista discrepancia entre esta y las normas de la Real Academia Española de la Lengua.

El sistema de transliteración del árabe empleado es el propuesto por *Journal of Islamic Studies*.

En cuanto a las citas y referencias bibliográficas, seguiremos las normas de la American Psychological Association (APA).

Capítulo II: Estado de la cuestión

En este capítulo revisaremos la literatura previa que, de manera directa o tangencial, ha abordado el fenómeno de la división de oraciones en el proceso de traducción. El primer apartado está dedicado a los estudios traductológicos; en el segundo revisaremos las investigaciones en Lingüística Computacional que, sin analizar específicamente el fenómeno, han realizado aportaciones de interés para nuestra investigación.

1. Traductología

Hasta donde se nos alcanza, no existe estudio traductológico previo que aborde la división de oraciones en la traducción del árabe. Sin embargo, algunos trabajos abordan el fenómeno en otras combinaciones lingüísticas, todas ellas del entorno europeo, cuya descripción nos permitirá contextualizar nuestros hallazgos en el marco de la disciplina.

En las investigaciones previas se observan tres tendencias a la hora de explicar la división de oraciones en el producto de la traducción: unos la consideran un fenómeno unidireccional impuesto por diferencias interlingüísticas estructurales; otros, la expresión de un universal de la traducción que, como tal, se registrará siempre y con independencia del par de lenguas implicado y de la direccionalidad; y otros, por último, consideran el fenómeno desde la concurrencia de ambas causas.

Fabricius-Hansen (1999), basándose en el análisis manual de un ensayo de Konrad Lorenz (*Das sogenannte Böse*) y sus traducciones al noruego y el inglés, registra dos procedimientos de mutación discursiva: transformación de una oración simple a una oración compuesta con una o más cláusulas subordinadas (*clausalization*) y transformación de una oración en varias otras (*sententialization*), con o sin aparición de cláusulas subordinadas. La autora concluye que, por sus características estructurales, el alemán tiende a “empaquetar” más la información, por lo que al traducir a las otras dos lenguas es necesario “desempaquetarla” dividiendo oraciones o introduciendo cláusulas

subordinadas. La autora señala asimismo que el inglés y el noruego difieren en la forma de solucionar ese problema de traducción.

Solfjeld (2008), basándose en un corpus de 10 textos en alemán (biografías y divulgación científica) de unas 5 000 palabras cada uno y sus traducciones al noruego, observa también que en la versión noruega ciertas oraciones con cláusulas subordinadas se dividen en varias oraciones o, incluso si se mantiene una sola oración, los traductores introducen cláusulas coordinadas (no subordinadas, debido al escaso recurso a la subordinación en noruego).

También Ramm (2004) estudia la división de oraciones en la traducción del alemán al noruego mediante un corpus paralelo. La principal aportación de esta autora consiste en la determinación de los adverbios que introducen en alemán las cláusulas de relativo y el cálculo de la frecuencia con que en las traducciones al noruego esas cláusulas se transforman en oraciones independientes.

Así pues, para Fabricius-Hansen (1999), Solfjeld (2008) y Ramm (2004), la tendencia a la división de oraciones se explica por diferencias interlingüísticas estructurales, en particular por la mayor "densidad informativa" (*informational density*) del alemán respecto al noruego y el inglés.

En la misma línea, Bloch (2005) aborda la división de oraciones en la traducción del inglés al francés, italiano, alemán y neerlandés partiendo de la hipótesis de que la complejidad sintáctica es menor en las lenguas germánicas que en las romances y, en concreto, de mayor a menor, en francés, italiano, alemán y neerlandés. Como consecuencia, la mayor frecuencia de división debería registrarse en las traducciones al neerlandés, seguida de las traducciones al alemán, el italiano y el francés.

Para probar su hipótesis, Bloch (2005) utiliza un corpus paralelo para cada par de lenguas conformado por textos técnicos redactados en inglés y sus traducciones. Los cuatro corpus suman en total algo menos de 300 000 palabras. Las traducciones fueron realizadas con herramientas de traducción asistida. Los resultados arrojaron los siguientes porcentajes de división de oraciones: neerlandés (4,5 %) > alemán (3,4 %) > francés (2,7 %) > italiano (0,6 %). Boch confirma, pues, su hipótesis de manera parcial: en todos los casos se registra división de oraciones y solo el porcentaje de división de oraciones en italiano, que es inferior al del francés, contradice sus previsiones. Bloch

concluye que la combinación lingüística tiene un impacto significativo en la división de oraciones, aunque la presencia ubicua del fenómeno revela que se trata de un universal de traducción (Baker, 1993 y 1996).

Musacchio (2005) analiza un corpus paralelo formado por artículos sobre economía redactados en inglés y sus traducciones al italiano. En este caso, los traductores tendían a unir, y no a dividir, las oraciones del texto original mediante conjunciones coordinantes. Según la autora, la causa de ello radica en que el italiano muestra una mayor preferencia por las oraciones extensas que el inglés. No se trataría, pues, de un universal de traducción.

Merkel (2001) y Ahrenberg (2017) estudian las diferencias entre la traducción manual y la traducción automática del inglés al sueco. Merkel (2001) analiza un corpus paralelo de originales ingleses (dos novelas, cinco manuales de informática y una colección de diálogos, en total 805 277 palabras) y sus traducciones al sueco (732 628 palabras). El corpus incluye tres tipos de traducciones: humanas, asistidas con memoria de traducción y automáticas. Concluye que, en general, en la traducción al sueco se registran más oraciones que en el original inglés. Entre las posibles explicaciones a este fenómeno, Merkel (2001) destaca las diferencias en el uso de los signos de puntuación: en sueco, el punto y coma del inglés y las comas que en ocasiones preceden en inglés a un entrecomillado suelen transformarse en punto.

Ahrenberg (2017), por su parte, compara la traducción humana al sueco de un artículo de opinión en inglés (2 555 palabras) con su traducción automática realizada por Google Translate. El texto original contiene 86 oraciones, las mismas que el texto generado por Google Translate, mientras que la traducción manual consta de 95. El traductor humano dividió el 9 % de las oraciones del original en dos o tres oraciones más cortas.

Merkel (2001) y Ahrenber (2017) coinciden en señalar que en las traducciones del inglés al sueco realizadas por humanos el número de oraciones tiende a ser mayor que en los originales, mientras que en la traducción automática es igual o incluso inferior.

Hareide y Hofland (2012) comparan 31 textos noruegos, principalmente literarios, y sus traducciones al español. Las traducciones contienen un 8 % más de palabras que los originales noruegos (1 564 347 frente a 1 455 021) y un 4 % menos de oraciones (118 439 frente a 123 621). En opinión de los autores, el mayor número de palabras en

las traducciones podría explicarse considerando que el noruego, al igual que otras lenguas germánicas, emplea sustantivos compuestos que se "descomponen" en español (como "*isbilen*" = "el furgón de hielo").

Según Hareide y Hofland (2012), la disminución del número de oraciones podría deberse a factores relacionados con el proceso de traducción y con las preferencias estilísticas de los traductores y editores, así como a diferencias interlingüísticas. Los autores prestan particular atención al grado de intervención de los traductores. La correspondencia en el número de oraciones entre diferentes traductores oscila entre el 98,69 % y el 75,08 % y son los traductores con menor porcentaje de correspondencia los que en mayor medida intervienen también en otros planos.

Kunilovskaya y Morgoun (2013) estudian los errores de traducción asociados al proceso de división de oraciones en traducciones del inglés al ruso realizadas por estudiantes de traducción. Para ello analizan un corpus formado por 164 textos originales en inglés (230 966 palabras y 13 353 oraciones) y un total de 700 traducciones de esos textos al ruso (409 415 palabras). En los resultados se observa que aproximadamente el 3% de las oraciones del original se dividieron. Las autoras también localizan las palabras del original elegidas por los estudiantes para dividir las oraciones (*stopwords*). Entre ellas destacan las conjunciones coordinantes "*and*" y "*but*".

Para Kunilovskaya y Morgoun (2013), al traducir del inglés al ruso es necesario dividir las oraciones complejas, en particular las coordinadas, para evitar pérdidas en la cohesión textual. El fenómeno se explicaría, por tanto, considerando diferencias interlingüísticas.

Serbina (2014) analiza la división de oraciones en la traducción del inglés al alemán y viceversa con un corpus de aproximadamente un millón de palabras. Aunque no proporciona información cuantitativa concreta, la autora observa que, en general, el número de oraciones aumenta al traducir al alemán y disminuye al traducir al inglés. Por tanto, el fenómeno estaría condicionado por la direccionalidad.

Para Serbina (2014), la división de oraciones tiene como objetivo reducir la densidad de información. Se trataría de un procedimiento aplicado por los traductores de forma consciente para hacer el texto meta más asequible al lector. Serbina señala que el

procedimiento se ha relacionado con la explicitación, uno de los universales de la traducción, aunque, como señalan sus datos, está ligado a diferencias interlingüísticas.

Bisiada (2013 y 2016), con un corpus paralelo de aproximadamente un millón de palabras, también estudia la división de oraciones en la traducción del inglés al alemán. En Bisiada (2013) se analiza la variación diacrónica del fenómeno con un corpus de textos fechados entre 1982 y 2008. Sus datos confirman que la tendencia a dividir oraciones es mucho más pronunciada que la tendencia a unirlas y se incrementa con el tiempo (Bisiada 2013, 127).

En Bisiada (2016) se rebaten las conclusiones de Fabricius-Hansen (1999) y Solfjeld (2008), para quienes el fenómeno deriva de diferencias interlingüísticas relacionadas con la densidad informativa. Para Bisiada, sin embargo, refleja la aplicación de una estrategia de explicitación inherente a los textos traducidos en general y que no necesariamente está ligada a diferencias interlingüísticas (Bisiada 2016, 374). En sus conclusiones, Bisiada también destaca la influencia de los editores, que favorecen la división de oraciones al considerar que las oraciones cortas mejoran la legibilidad.

Nadvornikova (2017) plantea dos hipótesis. La primera, que la densidad informativa del checo es menor que la del inglés y el francés. En consecuencia, al traducir del inglés o el francés al checo la tendencia a la división de oraciones será mayor que con la direccionalidad inversa y, viceversa, la tendencia a unir oraciones será mayor al traducir del checo al inglés o francés. La segunda de sus hipótesis afirma, que, a pesar de la diferencia en la densidad informativa, habrá división de oraciones con independencia de la combinación lingüística y la direccionalidad, por tratarse de un universal de traducción.

Para comprobar sus hipótesis, Nadvornikova (2017) analiza un corpus paralelo trilingüe bidireccional formado principalmente por novelas. Los resultados confirman las hipótesis de la autora, que atribuye la aplicación de estas estrategias a diferencias interlingüísticas (densidad informativa) y al reflejo de universales de traducción, en particular de la normalización, la simplificación y la explicitación. Como ejemplo de la normalización, la autora cita la fusión de oraciones intencionadamente breves de los originales franceses o ingleses en oraciones de mayor longitud con objeto de hacerlas más legibles en checo, a pesar de que como contrapartida se pierda el estilo original.

Frankenberg-Garcia (2019) estudia la fusión y la división de oraciones en la traducción del inglés al portugués, y viceversa, con un corpus paralelo bidireccional inglés-portugués de textos de ficción compuesto por 3 millones de palabras y algo más de 90 000 oraciones.

Según Frankenberg-Garcia (2019), en su corpus se respetan los puntos del original en el 94,6 % y el 90,1% de las ocasiones al traducir, respectivamente, al portugués y al inglés. En la traducción al portugués se fusiona el 2,71 % de las oraciones y se divide el 1,81 % y en la traducción al inglés el 2,77 % y el 3,23 %, respectivamente. A pesar de que los traductores al portugués tienden más a fusionar que a dividir y los traductores al inglés tienden a lo contrario, las diferencias no son estadísticamente significativas. La autora concluye que el factor determinante no es la direccionalidad sino el influjo de universales de traducción como la normalización y la explicitación.

Frankenberg-Garcia (2019) comprueba también que los traductores emplean los mismos procedimientos para unir o dividir oraciones con independencia de la direccionalidad y que el más frecuente de ellos es la sustitución de un signo de puntuación por otro (por ejemplo, una coma del original se convierte en un punto en la traducción).

Como hemos señalado previamente, no existen trabajos previos que aborden de manera específica la división de oraciones en la traducción desde o hacia el árabe. En la literatura sobre Lingüística Computacional que describiremos en el siguiente apartado, sin embargo, encontramos referencias aisladas y sin apoyo empírico a la cuestión. Por ejemplo, Zantout y Guessoum (2015) afirman que los textos en árabe

> *almost always consists of long sentences which, when translated into other languages, would result into several, shorter sentences. This makes the automatic correspondence between Arabic and non-Arabic sentences and words more difficult compared to other languages.* (Zantout y Guessoum, 2015, p. 248).

Keskes (2015) señala la tendencia del árabe a formar oraciones muy extensas en comparación con otras lenguas ("*Arab writers tend to write very long sentences, some of which could be a paragraph long with one full stop at the end*" (Keskes 2015, p. 37) y Dickins et al. (2017, p. 183) afirman que, "*because sentences in Arabic tend to be longer than sentences in English, it is not infrequently necessary to split up one Arabic sentence into several English ones*".

2. Lingüística Computacional

En este apartado revisaremos las investigaciones que, desde el ámbito de la Lingüística Computacional árabe, han aportado datos de interés para nuestra investigación. Como veremos, se trata de trabajos dirigidos fundamentalmente a mejorar los sistemas de segmentación y alineación automáticas. Como afirman Zantout y Guessoum (2015, p. 230), "a major problem facing AMT is the building of parallel corpora, that is, the selection of corpora content, their cleanup, preprocessing, tagging, and alignment".

Que sepamos, en Choueka et al. (2000) se concretan por primera vez desde una perspectiva empírica las dificultades de la alineación automática hebreo-inglés y se concluye que no es posible lograr una alineación eficiente sin tokenizar el texto hebreo. Según los autores, esta conclusión es trasladable a otras lenguas semíticas, incluido el árabe.

Samy y otros llevaron a cabo distintos experimentos (Samy et al. 2004 y 2006, y Samy y González 2008) con un corpus paralelo de 2 millones de palabras compuesto por documentos de la ONU redactados en inglés y traducidos al español y el árabe. La investigación comenzó con el par de lenguas árabe-español (Samy et al., 2004) y se amplió posteriormente al inglés.

En los análisis de Samy y otros se registró un número muy similar de oraciones para las tres lenguas, con una tendencia prácticamente irrelevante a unir, no a separar, oraciones en los textos traducidos; en concreto, 1 165 en árabe y 1 168 oraciones en español (Samy et al. 2004, sin paginación); 308 en inglés, 307 en árabe y 300 en español (Samy 2005), y 1 182 en inglés, 1 173 en árabe y 1 179 en español (Samy y González 2008, p. 3302). La interpretación de esos datos por los autores fue evolucionando. En principio, Samy et al. (2004) concluyeron que:

> *It is obvious that we can perform the alignments with minimal linguistic resources and thus contrasting the opinion of Choueka et al., which states that when dealing with Semitic languages, no statistical procedures, especially alignment is possible without some normalizing pre-processing; mainly lemmatization (2000).*

Más adelante, Samy (2005) matizaba esas conclusiones y subrayaba que esa escasa diferencia podía deberse a que los documentos en español y en árabe son traducciones

del inglés. En cualquier caso, en esas investigaciones se ignora el potencial impacto de la direccionalidad y del entorno institucional de la traducción.

Touir et al. (2008), en la senda de trabajos anteriores (Al-Sanie et. al. 2005a y 2005b, y Mathkour et al. 2005) dedicados a un analizador retórico (*rhetorical parsing*) y un sintetizador de contenidos (*automatic text summarization*), testan un segmentador automático de textos árabes no estructurados capaz de preservar la coherencia semántica de los segmentos resultantes.

Los autores aplican su segmentador a un corpus de 10 artículos en árabe de entre 500 y 700 palabras y comparan los resultados con los obtenidos tras la segmentación manual de los mismos textos, que constituye su *gold standard*. Puesto que los signos de puntuación en árabe se emplean de forma errática, su eficacia para determinar los límites entre segmentos (*segments boundaries*) es escasa. Por ello, los autores basan su herramienta solo en el emparejamiento de conectores.

La lista de potenciales conectores definida en sus trabajos previos (Al-Sanie et. al. 2005a y 2005b, y Mathkour et al. 2005) se completa con otros tomados de obras de lingüistas árabes contemporáneos. Los conectores se clasificaron como "activos" (aquellos que por sí solos podrían indicar el principio o final de un segmento) y "pasivos" (aquellos que por sí solos no indicarían el principio o final de un segmento pero, en concurrencia con un conector activo, ayudan a determinar los límites del segmento).

Sin embargo, los propios autores reconocen que la diferencia entre conectores activos y pasivos se diluye en la práctica, de modo que, según el contexto, el conector puede actuar activa o pasivamente. Concluyen que, en efecto, los conectores son más efectivos en árabe como referentes para la segmentación que los signos de puntuación, aunque el sistema precisa ser testado con un corpus mayor.

La importancia de la tokenización para el procesamiento automático del árabe, apuntada por Choueka et al. (2000), quedó finalmente demostrada en Alotaiby et al. (2010), trabajo trascendental a nuestro juicio. Los autores emplean un corpus de 600 millones de palabras formado por textos de agencias de prensa árabes y tokenizado mediante el programa AMIRA 2.0. Tras la tokenización, el número de *tokens* aumentó un 41,33 % y el de *types* se redujo un 24,54 %.

Pese al potencial error sistemático introducido por AMIRA 2.0, la posibilidad de tokenizar automáticamente un corpus tan enorme permitió extraer conclusiones que marcaron un hito en la Lingüística Computacional árabe. Esta fuerte dependencia de la tokenización explica en parte por qué las herramientas de alineación automática basadas en la extensión de las oraciones –un criterio eficaz para las lenguas europeas que fue establecido por Gale and Church (1993)- no son efectivas en árabe.

En los años posteriores, las potencialidades de AMIRA 2.0 se incrementaron vertiginosamente con el desarrollo de MADAMIRA (Pasha et al., 2014), que combina las utilidades de MADA, un analizador morfológico y desambiguador (Habash et al., 2010), y AMIRA. MADAMIRA supuso un salto cualitativo trascendental e inspiró el desarrollo de otras herramientas (Guellil et al., s. f.), como QATARA (Darwish et al., 2014), Standford Arabic Segmenter (Monroe et al., 2014), FARASA (Abdelali et al., 2016), YAMAMA (Khalifa et al., 2016) y CamelParser (Shahrour et al., 2016).

Salameh et al. (2011) aplican el alineador automático GIZA++, muy efectivo con pares de lenguas del entorno europeo, a un corpus de textos de la ONU formado por 2 154 pares de oraciones traducidas del inglés al árabe. En primer lugar se empleó el corpus sin preprocesamiento y seguidamente se compararon los resultados con los obtenidos tras el procesamiento manual del corpus.

El procesamiento consistió en convertir en minúsculas las mayúsculas del inglés (recordemos que en árabe no existen las mayúsculas) y tokenizar manualmente el texto árabe (fundamentalmente en la separación de los clíticos). En el nivel oracional, se procedió a segmentar las oraciones más largas, ya que GIZA++ había resultado más eficiente en otros pares de lenguas con oraciones cortas.

Para dividir las oraciones, los autores tomaron como referencia las comas del original y cierto número de *stopwords* en inglés (como *particularly*, *therefore* o *if*). Las *stopwords* se eligieron manualmente tras determinar aquellas que también en árabe correspondían a una *stopword* y siempre que, una vez dividida la oración, las oraciones resultantes en ambas lenguas fueran similares en contenido y la correspondencia en su número de palabras se mantuviera en un umbral de 0,21.

Salameh et al. (2011) añadieron también etiquetas para señalar los límites entre oraciones y otras (<EnLen>, <ArLen>, <Ratio>) que permitían mostrar información

sobre el número de palabras por oración en cada lengua y su ratio. Conviene señalar que los autores determinan también los datos de su corpus no procesado que figuran en la tabla 1.

Number of Pairs	2154
Maximum Number of Words in an English Sentence	132
Maximum Number of Words in an Arabic Sentence	175
Maximum Ratio of English to Arabic Sentence	1.5
Minimum Ratio of English to Arabic Words	0.291
Average Ratio of English to Arabic Words	0.706

Tabla 1. Número máximo de palabras en una oración en inglés y árabe, sus ratios máximas y su ratio media en el corpus no procesado de Salameh et al. (2011)

Los autores concluyen también que el preoprocesamiento del corpus mejora en un 100 % los resultados de la alineación automática con GIZA++.

Keskes (2015), por su parte, desarrolla un analizador discursivo para la lengua árabe con miras a desarrollar un sintetizador automático de contenidos. En lo que interesa a nuestra investigación, el autor concluye una vez más la ineficacia de los signos de puntuación como referente para la segmentación en árabe y afirma: "*The punctuation marks are not widely used in current Arabic texts (i.e., at least not regularly) and when they are used, they do not respect the typography rules*" (Keskes 2015, p. 55). Keskes (2015) también señala la tendencia del árabe a formar oraciones muy extensas en comparación con otras lenguas.

Por último, Zantout y Guessoum (2015), con el corpus utilizado por Salameh et al. (2011), analizan las dificultades de la traducción automática árabe-inglés, destacan la tendencia a la división de oraciones al traducir del árabe y proponen, sin apoyo empírico, una lista de *stopwords* en inglés que generan puntos en la traducción al árabe, así como determinan la ratio entre el número de palabras en inglés y árabe (p. 238).

3. Conclusiones

En la literatura previa se observa una marcada separación entre dos áreas principales de investigación: la Lingüística Computacional y los Estudios de Traducción. A nuestro juicio, esta división no ha repercutido positivamente en el desarrollo del estado de la cuestión.

En la literatura pertinente de Lingüística Computacional se siguen dos líneas: desarrollo de segmentadores automáticos y tokenizadores, y desarrollo de alineadores automáticos para corpus multilingües.

En la primera de esas líneas de investigación, desarrollada por Touir y su equipo y por Alotaiby y el suyo, se utilizan corpus unilingües en árabe formados por textos periodísticos; en la segunda, corpus bilingües o trilingües formados por documentos de la ONU redactados en inglés y sus traducciones, al árabe o al árabe y al español, realizadas en o para la ONU.

Los mayores avances se han logrado en el desarrollo de tokenizadores automáticos. Estos avances han culminado con el desarrollo de MADAMIRA, que permite preprocesar corpus árabes muy voluminosos. Las conclusiones extraídas en la literatura orientada al desarrollo o la mejora de alineadores se legitiman por el gran volumen de palabras analizadas gracias a los tokenizadores. Sin embargo, los tokenizadores introducen un margen de error sistemático que se evita cuando los resultados se contrastan con el *gold standard* del corpus procesado manualmente.

Ante el escaso rendimiento de las técnicas de alineación automática aplicadas a otras lenguas, la literatura previa enfatiza la necesidad de tokenizar el texto árabe y aplicar técnicas multicriteriales, entre otras cosas mediante la determinación de la ratio de palabras por oración, de los *stopwords* en inglés y/o en árabe y la localización de palabras comunes (en particular entidades nombradas) y su posición en la oración (Samy et al., 2006; Semmar y Fluhr, 2007; Ghaly , 2014).

Sin embargo, las técnicas multicriteriales padecen las dificultades de rendimiento de las herramientas asociadas a cada criterio. No funcionan de forma satisfactoria, por ejemplo, los diccionario electrónicos (*machine-readable dictionaries*) (Salameh et al., 2011, p. 173) y la localización de palabras comunes se ve obstaculizada por la escasa

efectividad de las herramientas de Reconocimiento de Entidades Nombradas en árabe (*Named Entity Recognition*) (Farghaly y Shaalan, 2009; Habash, 2010; Darwish y Gao, 2014; Shaalan, 2014; Sainz-Quinn, 2022). Por ello, los puntos de anclaje, al igual que las *stopwords*, se establecen en inglés, no en árabe (Samy, 2005; Salameh et al., 2011; Sainz-Quinn y Feria, 2020, entre otros).

A nuestro juicio, el principal problema metodológico de todas estas investigaciones es que se ignora la naturaleza de los corpus, que están compuestos por traducciones, la direccionalidad de esas traducciones y el entorno institucional en que fueron realizadas. Como consecuencia, las conclusiones están sesgadas. Superar ese escollo exige un acercamiento multidisciplinar que tenga en cuenta los avances en Traductología.

Los estudios traductológicos disponibles, por su parte, analizan el fenómeno de la división y la fusión de oraciones con combinaciones lingüísticas europeas. En general se registra una clara tendencia a la división de oraciones, pero no existe unanimidad sobre las causas del fenómeno.

Algunos autores defienden que la división de oraciones es consecuencia de diferencias interlingüísticas, por lo que la direccionalidad tiene un gran impacto (Fabricius-Hansen, 1999; Solfjeld, 2008), inclusive por el diferente uso de los signos de puntuación (Merkel, 2001; Hareide y Hofland, 2012; Frankenberg-Garcia, 2019).

Otros autores la explican como un universal de traducción independiente, por tanto, de la combinación lingüística y la direccionalidad (Bisiada, 2016). Otros, con una posición más sincrética, subrayan la concurrencia de ambos factores (Bloch, 2005; Serbina, 2014; Frankenberg-Garcia, 2019), y otros, finalmente, destacan el impacto adicional de las preferencias estilísticas y del grado de intervención de traductores y editores (Hareide y Hofland, 2012; Bisiada, 2013 y 2016).

La literatura previa sobre división de oraciones como estrategia en la traducción del árabe al inglés es escasa y casi inexistente en la traducción del árabe al español. Solo Samy et al. (2004, 2006) y Samy y González (2008) abordan la cuestión, aunque de forma tangencial y a partir de originales ingleses traducidos al árabe y al español.

Por otro lado, las conclusiones de la literatura traductológica previa se basan en datos recopilados a partir de pequeños corpus paralelos y analizados desde una perspectiva lingüística apropiada, aunque su base empírica es escasa o nula.

La falta de evidencia empírica y análisis de datos estadísticos es particularmente acentuada en los estudios de traducción del árabe. Aunque se ha señalado repetidamente la tendencia de los textos árabes a mostrar un uso irregular de los signos de puntuación (Touir et al., 2008; Keskes, 2015) y a generar oraciones extensas que deben dividirse al traducir a otras lenguas (Keskes, 2015; Zantout y Guessoum, 2015; Dickins et al., 2017), esas afirmaciones no se demuestran con datos estadísticos sólidos. Entre las razones que explican esta falla conviene señalar la inexistencia de alineadores automáticos efectivos. Esos alineadores no existen, irónicamente, por la dificultad que para el alineador implica la tendencia a dividir las oraciones al traducir del árabe.

No es de extrañar, por tanto, que la literatura previa subraye la necesidad de mejorar las herramientas de alineación y segmentación automática del árabe, entre otras cosas determinando una lista de potenciales *stopwords* en árabe. Salameh et al. (2011, p. 175) produjeron manualmente una lista de ese tipo mediante información filológica, no mediante datos empíricos. En nuestra opinión, sería más adecuado generar una lista similar mediante datos empíricos extraídos de un corpus de traducciones cuya direccionalidad y entorno social estén bien definidos.

A la vista de lo señalado, y como conclusión, podemos afirmar que los objetivos de la presente investigación son relevantes, contribuyen a cubrir un nicho de investigación y pueden tener importantes aplicaciones prácticas.

Capítulo III: Corpus y metodología

Este capítulo consta de dos partes. En la primera describimos los corpus empleados en la investigación. En la segunda detallamos la metodología utilizada para la selección de la muestra, la eliminación de los elementos de distorsión y la extracción de los datos.

1. Corpus

En este apartado se describe la compilación de los cuatro corpus empleados en nuestra investigación.

Hasta donde se nos alcanza, los corpus previamente disponibles no cumplían los requerimientos exigidos por esta investigación. Algunos no incluyen el español, como SauLTC (Al-Harthi y Alsaif, 2019) o ULTC (Al-Raisi et al., 2018; Taji et al., 2018; Altammami et al., 2019; Alfuraih, 2020). Otros, no especifican la direccionalidad de las traducciones (como, entre otros, Xu et al., 2001). Y otros, por último, comprenden textos no contemporáneos (Altammami et al., 2019) o están formados por traducciones realizadas por estudiantes de traducción, como SauLTC (Al-Harthi y Alsaif, 2019) y ULTC (Alfuraih, 2020) o por traducciones automáticas (Taji et al., 2018 y Al-Raisi et al., 2018).

Nuestra investigación comenzó con la compilación de un corpus (en adelante, Corpus 1) que comprende todos los documentos originales presentados entre 2000 y 2018, ambos años incluidos, por los Estados árabes a varios órganos de las Naciones Unidas. Seguidamente se compiló un segundo corpus (en adelante, Corpus 2) formado por los textos originales en árabe del Corpus 1 y sus traducciones al español y al inglés. Por último, y con miras a testar el impacto potencial de la direccionalidad, se compilaron el Corpus 3 (textos originales en español y sus traducciones al árabe) y el Corpus 4 (originales en inglés y sus traducciones al árabe).

1.1. Corpus 1

El Corpus 1 comprende las aportaciones documentales originales (en árabe o en otras lenguas) presentadas entre 2000 y 2018, ambos años incluidos, por los miembros de la Liga de los Estados Árabes a los Comités de Derechos Humanos y el Grupo de Trabajo sobre el Examen Periódico Universal (Oficina de la ONU en Ginebra).

Los comités de las Naciones Unidas que supervisan la aplicación de los principales tratados internacionales de derechos humanos son los siguientes:

- Comité de Derechos Humanos (CCPR)
- Comité de Derechos Económicos, Sociales y Culturales (CESCR)
- Comité para la Eliminación de la Discriminación Racial (CERD)
- Comité para la Eliminación de la Discriminación contra la Mujer (CEDAW)
- Comité contra la Tortura (CAT)
- Comité de los Derechos del Niño (CRC)
- Comité para la Protección de los Derechos de todos los Trabajadores Migratorios y de sus Familiares (CMW)
- Comité sobre los derechos de las personas con discapacidad (CRPD)
- Comité contra las Desapariciones Forzadas (CED)

El objetivo de la compilación del Corpus 1 era verificar la disponibilidad de material suficiente para compilar el Corpus 2.

A continuación exponemos los principales criterios de selección (origen, originalidad, representatividad, accesibilidad, formato y confianza) considerados para la compilación del Corpus 1.

a) Origen, originalidad y representatividad

El Corpus 1 está compuesto por los textos presentados por todos los Estados árabes (origen) a los órganos señalados. Teóricamente, los textos fueron redactados en la lengua que figura como original (originalidad), con independencia de cuál sea esa

lengua. Como veremos, el Corpus 1 comprende principalmente textos escritos en árabe y, en menor número, en francés y en inglés.

La representatividad del Corpus 1 está garantizada por el hecho de que abarca todos los Estados árabes. A efectos de esta investigación se consideran Estados árabes los pertenecientes a la Liga de los Estados Árabes.

Para denominar a los Estados árabes empleamos la nomenclatura siguiente de las Naciones Unidas (entre paréntesis, el acrónimo utilizado en las signaturas de los documentos): Arabia Saudita (SAU), Argelia (DZA), Bahrein (BHR), Comoras (COM), Djibouti (DJI), Egipto (EGY), Emiratos Árabes Unidos (ARE), Iraq (IRQ), Jordania (JOR), Kuwait (KWT), Líbano (LBN), Libia (LBY), Marruecos (MAR), Mauritania (MRT), Omán (OMN), Estado de Palestina (PSE), Qatar (QAT), República Árabe Siria (SYR), Sudán (SDN), Túnez (TUN) y Yemen (YEM). En texto corrido portarán artículos los nombres de Estado siguientes: la Arabia Saudita, las Comoras, los Emiratos Árabes Unidos, el Iraq, el Líbano, el Estado de Palestina, la República Árabe Siria, el Sudán y el Yemen.

b) Accesibilidad, formato y confianza

Los textos, al haber sido presentados a las Naciones Unidas y figurar en sus repertorios documentales, cumplen los requisitos de accesibilidad y fiabilidad por haber superado los filtros de calidad de las Naciones Unidas. La accesibilidad en línea, por su parte, asegura la reproducibilidad de los experimentos.

Existen varias webs para la consulta y descarga de documentos de las Naciones Unidas. La documentación presentada a los comités de derechos humanos se localizó y descargó del portal del Alto Comisionado de las Naciones Unidas para los Derechos Humanos (http://tbinternet.ohchr.org/_layouts/TreatyBodyExternal/Countries.aspx). Su buscador proporciona información específica por países sobre el estado de ratificación de los convenios internacionales de derechos humanos y permite descargar la documentación pertinente.

Para descargar la documentación presentada ante el Grupo de Trabajo sobre el Examen Periódico Universal recurrimos a la web del Consejo de Derechos Humanos de las

Naciones Unidas (https://www.ohchr.org/SP/HRBodies/UPR/Pages/Documentation.aspx).

Cuando los archivos descargados se encontraban dañados o en formato pdf, se recurrió a la web https://undocs.org/, que permite descargar los documentos una vez conocida su signatura.

Téngase en cuenta que los documentos anteriores al año 2000 solo se encuentran en formato pdf. Este formato complica sobremanera el procesamiento de los textos árabes. El rendimiento de los sistemas de OCR en árabe presenta tasas de precisión inferiores al 75 % (Alghamdi y Teahan 2017, 239). Esto demandaría tiempos de corrección manual imposibles de asumir para un corpus como el nuestro. Por ello, el Corpus 1 solo comprende textos posteriores al 2000, que en su gran mayoría se encuentran disponibles en archivos Word.

El año 2000 también se ha tomado como punto de partida para la compilación de otros corpus de textos de las Naciones Unidas por las mismas razones, confesas o no. Así, Rafalovitch y Dale compilan las resoluciones de la Asamblea General de las Naciones Unidas emitidas a partir del año 2000 porque entre 1946, año de la primera sesión de la Asamblea General de la ONU, y 2000, las versiones electrónicas de las resoluciones son documentos escaneados (Rafalovitch y Dale, 2009, p. 294). Salameh et al. (2011, p. 172) y Chen y Eisele (2012, p. 2501) coinciden en señalar este mismo motivo como razón para compilar solo textos fechados a partir del año 2000. Aunque sin mencionar explícitamente la causa, también MultiUn (Eisele y Chen, 2010, p. 2869) compila documentos de las Naciones Unidas fechados a partir del año 2000.

c) Delimitación

Si bien la documentación de las Naciones Unidas disponible en línea es muy diversa, la presentada por los Estados árabes no lo es tanto. Por otro lado, para los fines de esta investigación era imprescindible determinar en la mayor medida posible la lengua original de esos documentos y la direccionalidad de las traducciones. Estos factores son mucho más importantes en nuestro caso que en el de los otros corpus de textos de las Naciones Unidas existentes.

La determinación de la lengua original de nuestros textos y de la lengua desde la que se traducen tropieza con varios obstáculos:

i. Los informes presentados como originales árabes pueden haber sido escritos en inglés o francés, y traducidos y presentados, por razones políticas, en árabe. En ocasiones, los traductores afirman que intuyen que ese es el caso porque la traducción les resulta inusualmente fácil[6];
ii. Las traducciones al español del Corpus 2 puedan estar intermediadas del inglés o del francés, pese a que oficialmente no se haga constar así;
iii. Las traducciones del español al árabe del Corpus 3 pueden estar intermediadas del inglés o del francés.

Pese a ello, los textos de las Naciones Unidas y sus traducciones presentan ventajas fundamentales:

i. La calidad de los originales está garantizada por los Estados;
ii. La calidad de las traducciones está garantizada por las Naciones Unidas;
iii. Tanto los originales como las traducciones son accesibles en línea y de libre acceso;
iv. Los originales tienen unas características de registro y género tan definidas que las variaciones estadísticamente relevantes difícilmente podrían atribuirse a variación intergénero y, en escasa medida, a variación idiolectal (definida como "*the set of options that writers take from the linguistic repertoire available to them as users of a specific language*", Turell, 2010, p. 217). Por tanto, si el volumen y la representatividad del corpus son suficientes, las variaciones intralingüísticas podrían considerarse principalmente de carácter diatópico y las interlingüísticas podrían atribuirse interlingüísticas.

Nuestro corpus, por tanto, puede considerarse homogéneo en cuanto al registro y en los planos temático, cronológico y geográfico.

[6] José García Verdugo, traductor la Sección de Español de la ONU en Nueva York (comunicación personal de 08/01/2014) y Manuel Feria García, traductor de la Sección de Español de la ONU en Ginebra (comunicación personal de 09/01/2014).

d) Compilación

Como quedó dicho, se accedió a los documentos en formato Word gracias a los buscadores de los portales del Alto Comisionado de las Naciones Unidas para los Derechos Humanos y del Consejo de Derechos Humanos de las Naciones Unidas. Si los archivos se encontraban dañados o en formato pdf, se recurrió a la web https://undocs.org/, que permite descargar los documentos una vez conocida su signatura.

Junto con los documentos originales se descargaron sus traducciones al español y al inglés con miras a incluirlas en el Corpus 2. Cada documento se identifica en las Naciones Unidas mediante una signatura única, común también a sus traducciones, que suele aparecer en la portadilla. En nuestro corpus, los archivos se clasificaron por Estado y se identificaron mediante la signatura original del documento y la terminación _AR (árabe), _SP (español), _EN (inglés) o _FR (francés).

Conviene señalar que, con frecuencia, en la portadilla de los documentos figura el inglés como lengua original y pocas líneas más abajo se especifica que en realidad es el árabe, ya que la portadilla, y solo ella, se redactó en inglés. Por lo tanto, hemos considerado esos documentos originales árabes.

Por otro lado, en la portadilla de algunos documentos figuran dos lenguas como lenguas de redacción; en estos casos, se descargaron los dos informes y se contabilizaron como si se tratara de dos documentos diferentes.

Inicialmente descargamos 512 textos. Una vez descartados los no editados conforme a los estándares habituales (fotocopias escaneadas, documentos sin portadilla o sin signatura y documentos defectuosos o incompletos), ese número se redujo a 469. La relación completa de documentos que forman el Corpus 1 figura en el Anexo I.

Estos textos, redactados en árabe, francés o inglés, comprenden 7 154 608 palabras gráficas. Este cálculo no fue tan sencillo como hubiera sido de esperar. A pesar de tratarse de documentos publicados a partir del año 2000, 10 de ellos solo se encontraban disponibles en formato pdf, lo que obstaculizaba el recuento de palabras. En estos casos el cálculo se realizó por aproximación tras estimar la media de palabras por página para las tres lenguas en los documentos en formato Word. Los resultados fueron los siguientes:

- árabe: 288,48 palabras/página;
- francés: 420,72 palabras/página;
- inglés: 409,64 palabras/página.

En consideración a esos datos, los 10 documentos en pdf sumaron 274 334 palabras gráficas, es decir, en torno al 3 % del Corpus 1.

e) Distribución de lenguas

Los Estados emplearon tres lenguas para la redacción de sus originales: árabe, francés e inglés. La tabla 2 muestra la distribución por lenguas en el Corpus 1 con datos desglosados por número de palabras y documentos.

	Nº palabras	*% palabras respecto al total*	*Nº documentos*	*% documentos respecto al total*
Árabe	4 955 575	69,3 %	332	70,8 %
Francés	1 871 236	26,2 %	118	25,2 %
Inglés	3 277 97	4,6 %	19	4,1 %
TOTAL	**7 154 608**		**469**	

Tabla 2. Distribución por lenguas en el Corpus 1, con datos desglosados por número de palabras y documentos

Por tanto, el árabe fue la principal lengua de redacción escogida por los Estados árabes durante el periodo considerado (en torno al 70 % de las palabras y documentos), seguido por el francés (en torno al 25 %, unas tres veces menos que el árabe) y el inglés (en torno al 5 %).

Conviene señalar que, de los 19 documentos presentados en inglés, únicamente 8 fueron redactados exclusivamente en esa lengua, mientras que los 11 restantes son originales bilingües presentados también en árabe (10) o francés (1). Así pues, si consideramos los documentos presentados exclusivamente en inglés, la aportación de esta lengua al Corpus 1 quedaría reducida a 8 documentos y 81 284 palabras, lo que supone el 1,1 %

de las palabras. La aportación del inglés como lengua original exclusiva es, por tanto, insignificante.

Sin embargo, los datos anteriores deben matizarse considerando diferencias interestatales. En la tabla 3 se presenta la distribución por lenguas en el Corpus 1 con datos desglosados por Estado y número de palabras. En el gráfico 1 se representa la misma distribución en términos porcentuales.

Estado	*Árabe*	*Francés*	*Inglés*	*Total*
Arabia Saudita	308 969	0	11 445	320 414
Argelia	0	506 233	0	506 233
Bahrein	369 275	0	0	369 275
Comoras	0	46 706	0	46 706
Djibouti	0	188 087	0	188 087
EAU	216 895	0	0	216 895
Egipto	336 694	0	8 236	344 930
Iraq	298 655	0	15 673	314 328
Jordania	388 067	0	0	388 067
Kuwait	328 714	0	0	328 714
Líbano	413 501	11 293	98 973	523 767
Libia	81 222	0	0	81 222
Marruecos	187 422	353 715	2 662	543 799
Mauritania	0	342 131	0	342 131
Omán	222 997	0	30 764	253 761
Palestina	76 151	0	0	76 151
Qatar	361 160	0	4 260	365 420
Siria	347 155	0	0	347 155
Somalia	0	0	17 431	17 431
Sudán	253 531	0	53 770	307 301
Túnez	160 371	423 071	0	583 442
Yemen	604 796	0	84 583	689 379
TOTAL	**4 955 575**	**1 871 236**	**327 797**	**7 154 608**

Tabla 3. Distribución por lenguas en el Corpus 1, con datos desglosados por Estado y número de palabras

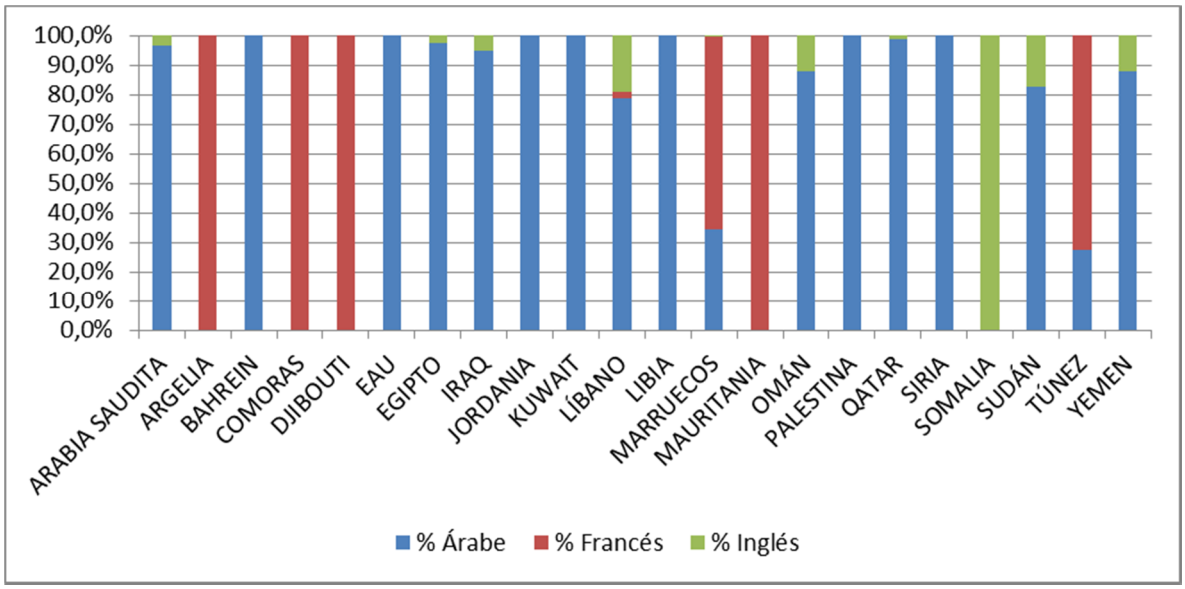

Gráfico 1. Distribución por lenguas en el Corpus 1, con datos desglosados por Estado y número de palabras en términos porcentuales

Los Estados árabes, por tanto, se clasifican en seis grupos en función de la lengua o lenguas en las que presentaron su documentación a los órganos de las Naciones Unidas en cuestión durante el período concernido:

i. En árabe: Bahrein, Emiratos Árabes Unidos, Jordania, Kuwait, Libia, Palestina y Siria;
ii. En árabe e inglés: Arabia Saudita, Egipto, Iraq, Omán, Qatar, Sudán y Yemen;
iii. En árabe y francés: Túnez;
iv. En árabe, francés e inglés: Líbano y Marruecos;
v. En francés: Argelia, Mauritania, Comoras y Yibuti;
vi. En inglés: Somalia.

De los 22 Estados concernidos, 17 emplearon el árabe, 7 de ellos como única lengua de redacción. De los 5 Estados que no emplearon el árabe, 4 utilizaron exclusivamente el francés y 1 empleó el inglés como única lengua de redacción.

Por otro lado, los Estados contribuyen de forma muy dispar al Corpus 1 (tabla 4). Yemen aporta 48 veces más documentación que Somalia, y Túnez 7,5 veces más que Libia o Palestina. Como era de esperar, no existe correlación alguna entre el volumen de

población de un Estado y la cantidad de documentación aportada (Baréin, por ejemplo, aporta más documentación que Egipto a pesar de tener 66 veces menos población).

Estado	*Nº de palabras*	*% Corpus 1*
Yemen	689 379	9,6 %
Túnez	583 442	8,2 %
Marruecos	543 799	7,6 %
Líbano	523 767	7,3 %
Argelia	506 233	7,1 %
Jordania	388 067	5,4 %
Bahrein	369 275	5,2 %
Qatar	365 420	5,1 %
Siria	347 155	4,9 %
Egipto	344 930	4,8 %
Mauritania	342 131	4,8 %
Kuwait	328 714	4,6 %
Arabia Saudita	320 414	4,5 %
Iraq	314 328	4,4 %
Sudán	307 301	4,3 %
Omán	253 761	3,5 %
EAU	216 895	3 %
Djibouti	188 087	2,6 %
Libia	81 222	1,1 %
Palestina	76 151	1,1 %
Comoras	46 706	0,7 %
Somalia	17 431	0,2 %
TOTAL	7 154 608	100 %

Tabla 4. Contribución al Corpus 1, con datos desglosados por Estado y palabras (en términos absolutos y porcentuales)

Finalmente, en la tabla 5 se presenta la evolución cronológica del porcentaje de aportación por lengua y en el gráfico 2, el volumen de documentación en número de palabras por lengua.

Año	*Lengua*		
	Árabe	*Francés*	*Inglés*
2000	46,7 %	30,3 %	23 %
2001	67,7 %	32,3 %	0 %
2002	92,7 %	7,3 %	0 %
2003	64,8 %	30,2 %	5 %
2004	61,3 %	21 %	17,7 %
2005	57,1 %	24,6 %	18,3 %
2006	79,6 %	19,3 %	1,1 %
2007	54,6 %	36,7 %	8,8 %
2008	47,5 %	48,9 %	3,6 %
2009	75,1 %	24,9 %	0 %
2010	69,6 %	26,1 %	4,2 %
2011	58,9 %	35,2 %	5,8 %
2012	55,1 %	44,9 %	0 %
2013	75,1 %	21,4 %	3,4 %
2014	83,3 %	15,5 %	1,2 %
2015	77,6 %	20,2 %	2,2 %
2016	82,7 %	17,3 %	0 %
2017	79,5 %	20,5 %	0 %
2018	75,8 %	24,2 %	0 %

Tabla 5. Evolución cronológica de la aportación por lenguas al Corpus 1, en términos porcentuales

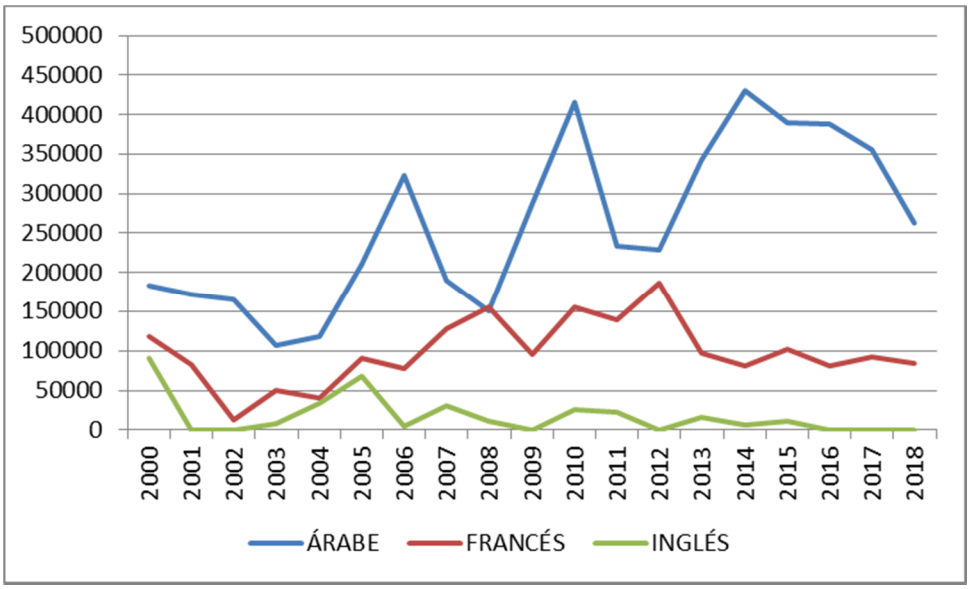

Gráfico 2. Evolución cronológica del número de palabras por lengua

Los datos anteriores se analizan en detalle en el Capítulo IV.

1.2. Corpus 2

Para refutar o corroborar la hipótesis nula se precisaba contar con un corpus paralelo (Corpus 2) que comprendiera los textos originalmente redactados en árabe del Corpus 1 y sus traducciones al inglés y español. De él se extraería una muestra y de ella, el grueso de nuestras conclusiones.

Una vez analizado el Corpus 1 se comprobó que no todos sus 332 documentos redactados en árabe podían incluirse en el Corpus 2, ya que algunos no contaban con traducciones al español y al inglés o sus traducciones no podían descargarse en formato Word, y otros tenían una extensión tan reducida que no podían considerarse representativos: los *corrigenda*, por ejemplo, consisten únicamente en la portadilla, que está estandarizada, y una o dos oraciones.

En consecuencia, se decidió incluir en el Corpus 2 solo los originales árabes con más de 10 000 palabras gráficas y cuyas versiones al español y al inglés pudieran descargarse en formato Word. La cifra anterior se determinó tras realizar diversas pruebas en las que se comprobó que los documentos menores (generalmente adendas o corrigendas) presentaban datos estadísticos extremadamente heterogéneos.

Finalmente, el Corpus 2 comprendió 171 documentos en árabe (3 541 670 palabras) y sus correspondientes traducciones al español (5 139 440 palabras) y al inglés (4 424 882 palabras), en total 513 documentos y 13 105 992 palabras. El listado completo de esos documentos figura en el Anexo II.

Este corpus, que ya fue utilizado por Sainz-Quinn y Feria-García (2020) y por Sainz-Quinn (2022), cumple los requisitos de autenticidad, representatividad, accesibilidad, fiabilidad (en particular de direccionalidad en la traducción) y calidad (tipográfica y de traducción) por las razones antedichas, lo que, entre otras cosas, asegura la reproducibilidad de los experimentos.

En el Corpus 2 se encuentran representados 17 Estados: Arabia Saudita, Bahrein, Egipto, Emiratos Árabes Unidos, Iraq, Jordania, Kuwait, Líbano, Libia, Marruecos, Omán, Palestina, Qatar, Siria, Sudán, Túnez y Yemen, es decir, los mismos Estados incluidos en el Corpus 1 salvo los que no presentaron documento alguno en árabe durante el periodo considerado por utilizar exclusivamente el francés (Argelia, Comoras, Djibouti y Mauritania) o el inglés (Somalia).

En la tabla 6 figura la aportación de cada Estado al Corpus 2, con datos desglosados por número de palabras árabes y porcentaje correspondiente del total de palabras árabes.

Estado	*Nº de palabras*	*% en el corpus*
Yemen	433 762	12,2 %
Bahrein	304 920	8,6 %
Líbano	288 893	8,2 %
Jordania	279 295	7,9 %
Qatar	261 727	7,4 %
Siria	258 216	7,3 %
Egipto	258 026	7,3 %
Arabia Saudita	252 870	7,1 %
Kuwait	251 713	7,1 %
Iraq	243 023	6,9 %
EAU	146 528	4,1 %
Omán	145 413	4,1 %
Sudán	137 257	3,9 %
Túnez	100 422	2,8 %
Marruecos	91 765	2,6 %
Palestina	76 151	2,2 %
Libia	11 689	0,3 %

Tabla 6. Aportación de cada Estado al Corpus 2, con datos desglosados por número de palabras árabes y porcentaje correspondiente del total de palabras árabes

En comparación con el Corpus 1, además de la desaparición de los Estados que no hicieron aportaciones en árabe destaca que Túnez y Marruecos, por su uso preferente del francés, descienden de los puestos 2º y 3º, respectivamente, a los puestos 14º y 15º.

El Corpus 2, además de ser necesario para la selección de la muestra, permite extraer las conclusiones sobre la ratio de palabras gráficas en árabe, español e inglés que se presentan en el Capítulo IV.

1.3. Corpus 3

El Corpus 3 comprende documentos presentados originalmente en español a los mismos órganos de las Naciones Unidas y en el mismo período de tiempo, así como sus traducciones al árabe. Este corpus permitirá analizar los posibles efectos de la

direccionalidad, en particular las variaciones en el número de oraciones. Para ello, debía presentar características similares a los Corpus 1 y 2, en particular a la muestra extraída del Corpus 2.

Los documentos que forman el Corpus 3 fueron descargados de las mismas webs institucionales de las Naciones Unidas empleadas para descargar los documentos de los Corpus 1 y 2 y cumplen los mismos requisitos, inclusive el de número de palabras gráficas y formato. Además, el Corpus 3 comprende un número de informes y palabras similar al de la muestra extraída del Corpus 2 para asegurar la comparabilidad.

El Corpus 3 comprende 12 documentos y 196 968 palabras: 6 originales en español (108 976 palabras) y sus traducciones al árabe (87 992 palabras). En el Corpus 3 están representados los 6 Estados más poblados cuya lengua oficial es el español: Argentina, Colombia, España, México, Perú y Venezuela. Cada Estado aporta un solo documento que, con objeto de facilitar el manejo de los datos, se designa con la inicial de ese Estado.

Los seis informes que componen el Corpus 3 (signatura de la ONU y, entre paréntesis, código de identificación en esta investigación) se relacionan en la tabla 7 con datos desglosados por número de palabras de los originales en español y de sus traducciones al árabe.

Estado	Documento	Nº de palabras	
		Español	Árabe
Argentina	CED/C/ARG/1 (A)	17 011	13 396
Colombia	A/HRC/WG.6/30/COL/1 (C)	10 719	9 425
España	CEDAW/C/ESP/7-8 (E)	24 036	20 484
México	CRPD/C/MEX/1 (M)	22 145	16 034
Perú	CEDAW/C/PER/7-8 (P)	14 862	12 320
Venezuela	CEDAW/C/VEN/7-8 (V)	20 203	16 333
TOTAL		108976	87 992

Tabla 7. Composición del Corpus 3

1.4. Corpus 4

Este corpus es similar al Corpus 3, salvo que sus documentos fueron redactados originalmente en inglés, cumple sus mismos requisitos y permitirá igualmente analizar los posibles efectos de la direccionalidad. Está formado por seis documentos originales en inglés (95 409 palabras) y sus traducciones al árabe (89 546 palabras); por tanto, comprende 12 documentos y 184 955 palabras.

En el Corpus 4 se encuentran representados seis Estados que cuentan con el inglés entre sus lenguas oficiales. Entre ellos se encuentran los cinco Estados angloparlantes más poblados de sus respectivos continentes: Australia, Estados Unidos, India, Nigeria y Reino Unido. El sexto Estado con representación en el corpus es Canadá. Cada Estado aporta un solo documento que, con objeto de facilitar el manejo de los datos, se designa con la inicial de ese Estado.

Los seis informes que componen el Corpus 4 (signatura de la ONU y, entre paréntesis, código de identificación en esta investigación) se relacionan en la tabla 8 con datos desglosados por número de palabras de los originales en inglés y sus traducciones al árabe.

Estado	Documento	Nº de palabras	
		Inglés	Árabe
Australia	CRPD/C/AUS/1 (AU)	20 609	19 347
Canadá	A/HRC/WG.6/16/CAN/1 (CA)	10 476	99 86
Estados Unidos	A/HRC/WG.6/22/USA/1 (EU)	10 979	11 145
India	CEDAW/C/IND/4-5 (IN)	11 082	10 075
Nigeria	A/HRC/WG.6/31/NGA/1 (NI)	10 555	91 13
Reino Unido	CEDAW/C/GBR/7 (RU)	31 708	29 880
TOTAL		95 408	89 546

Tabla 8. Composición del Corpus 4

2. Metodología

Este apartado se divide en dos subapartados. En el primero se describe la metodología empleada para analizar la división de oraciones en la traducción del árabe al español y al inglés; en el segundo, la empleada para analizar el fenómeno en la traducción del español y el inglés al árabe.

2.1. Traducción del árabe al español y al inglés

a) Consideraciones preliminares

Para analizar la división de oraciones en la traducción del árabe al español y al inglés cuantificaremos en el Corpus 2 el número de oraciones de los textos originales y sus traducciones. A este respecto, debemos realizar algunas precisiones relacionadas con la lengua árabe y con el tipo de texto analizado.

En árabe no existen letras mayúsculas. El recuento de oraciones, por tanto, no puede basarse en ese elemento. Por otro lado, dadas las características del género (informes de derechos humanos), las oraciones de nuestro corpus son enunciativas y completas: no se registran oraciones interrogativas o exclamativas, ni oraciones incompletas propias del registro oral y del tono insinuativo. Además, las siglas y los acrónimos con puntos, de uso muy limitado en árabe, son prácticamente inexistentes en los textos árabes de las Naciones Unidas. Como señala Samy (2005, sin paginación),

> *the Acronyms are equally distributed in the Spanish and the English corpora. However, the Arabic corpus shows a very little frequency of this type of Named Entities (only 1%). This is due to the fact that the Arabic language rarely adopts the acronyms. Instead it uses the full form of the name, or in very few cases it might opt for a phonetic transcription of the acronym. This is the case of "KFOR", which was phonetically transcripted into "كفور".*

En consecuencia, los criterios tipográficos para la determinación de los límites entre oraciones se reducen drásticamente, lo que nos hacía prever una gran fiabilidad en el recuento de oraciones basado en puntos que realiza WordSmith 5 (Scott, 2008). En nuestra investigación hacemos un uso extensivo de esta herramienta informática.

b) Selección de la muestra

Varias pruebas piloto realizadas con WordSmith revelaron que, contrariamente a lo que cabía esperar, el número de oraciones calculado automáticamente no coincidía con el obtenido mediante recuento manual. En consecuencia, seleccionamos una muestra representativa del Corpus 2 para identificar y eliminar aquellos elementos que pudieran haber causado esta distorsión. Una vez procesada, esta muestra permitiría refutar o no las hipótesis.

Con objeto de asegurar la representatividad de esa muestra aplicamos los criterios siguientes:

i. Se eliminaron los textos presentados por Libia, dado que su contribución al corpus es meramente testimonial (0,3 %). Para comprobar la aportación de cada Estado al Corpus 2, consúltese la tabla 5;
ii. Los Estados restantes se dividieron en dos grupos: los que contribuyeron en mayor medida (de Yemen a Iraq) y los que lo hicieron en menor medida (de Emiratos Árabes a Palestina). A la hora de incluir a los Estados en uno de esos dos grupos se tomó como referencia el porcentaje medio de aportación (6,23 %);
iii. Aplicando criterios cuantitativos y geográficos, se seleccionaron cinco de los Estados más representativos: Yemen (Península Arábiga), Bahrein (Golfo Pérsico), Líbano (Gran Siria), Egipto (región árabe central) e Iraq (extremo oriental de la región árabe);
iv. Marruecos fue elegido dentro del grupo de los Estados menos representativos (extremo occidental de la región árabe).

Finalmente, se procedió a seleccionar de forma aleatoria un documento presentado por cada uno de esos seis Estados, por lo que se obtuvo una muestra que comprendía seis informes. Para facilitar el análisis, los seis documentos seleccionados se designaron empleando la inicial del Estado en cuestión: B (Bahrein), E (Egipto), I (Iraq), L (Líbano), M (Marruecos) e Y (Yemen).

Los informes que componen la muestra se relacionan en la tabla 9 (signatura de la ONU y, entre paréntesis, código de identificación en esta investigación) con datos desglosados por número de palabras gráficas del original árabe y sus traducciones al

español y al inglés y por aportación a la muestra en palabras gráficas por informe y Estado.

Estado	Documento	Nº de palabras gráficas			Contribución
		Árabe	Español	Inglés	
Bahrein	CAT/C/BHR/3 (B)	21 073	29 138	25 174	18,6 %
Egipto	E/C.12/EGY/Q/2-4_Add.1 (E)	18 266	29 932	22 903	16,2 %
Iraq	CRC/C/OPSC/IRQ/1 (I)	17 985	24 832	22 291	15,9 %
Líbano	CRC/C/LBN/4-5 (L)	21 062	32 823	27 533	18,6 %
Marruecos	E/C.12/MAR/Q/2/Add.2 (M)	17 737	27 572	21 781	15,7 %
Yemen	CRC/C/OPSC/YEM/1 (Y)	16 877	25 417	21 455	14,9 %
TOTAL		113 000	169 714	141 137	100%

Tabla 9. Composición de la muestra, con datos desglosados por número de palabras gráficas del original y sus traducciones y por aportación a la muestra en palabras gráficas por informe y Estado

En total, la muestra comprende 423 851 palabras, de las que 113 000 son palabras gráficas en árabe, algo más del 3 % de las palabras en árabe del Corpus 2. Una muestra seleccionada al azar siete veces menor (16 563 palabras gráficas) ya hubiera sido estadísticamente representativa de la población (con un 99 % de intervalo de confianza y un margen de error = 1). La ampliación de la muestra y la determinación de criterios geográficos para su selección responden al deseo de testar posibles variaciones interestatales e interregionales en investigaciones posteriores.

El promedio porcentual de contribución de los informes a la muestra es del 16,67 % con una desviación típica (σ) de 1,58; por tanto, el 50 % de los valores se ubica a no más de una desviación típica de la media y el 100 % a no más de dos. La muestra, por tanto, presenta una distribución normal y es representativa del Corpus 2, en términos estadísticos y geopolíticos.

Conviene señalar que ocho de los diez primeros informes primeramente seleccionados para la muestra fueron descartados al detectarse indicios de que sus traducciones al

español habían sido realizadas, al menos parcialmente, a partir de las traducciones inglesas o francesas. Esos ocho documentos figuran en la tabla 10.

Documento	*Estado*	*Año*
CCPR/C/EGY/2001/3	EGIPTO	2002
CMW/C/EGY/1	EGIPTO	2006
CEDAW_C_IRQ_4-6	IRAQ	2013
CEDAW_C_LBN_2	LÍBANO	2005
CEDAW_C_LBN_3	LÍBANO	2006
CERD_C_MAR_17-18	MARRUECOS	2009
E_1990_5_Add.54	YEMEN	2002
CERD_C_362_Add.8	YEMEN	2002

Tabla 10. Documentos con posible traducción intermediada al español que fueron descartados de la muestra

El primer indicio de la existencia de traducción intermediada lo constituyó la coincidencia casi plena en el número de oraciones de las versiones española e inglesa o francesa. Ese primer indicio nos condujo a comparar cualitativamente las traducciones y, en efecto, se hallaron evidencias de intermediación.

Las evidencias más claras de intermediación radican en los errores materiales propios de la traducción humana, en particular en relación con números. A modo de ejemplo, en el párrafo 534 de CCPR/C/EGY/2001/3, 37 se convirtió en 38; en el párrafo 1 de CEDAW/C/LBN/2, 11 se convirtió en 12; en su párrafo 3, 562 se convirtió en 563; en el párrafo 22 de CEDAW/C/LBN/3, 1951 se convirtió en 1952, y en su párrafo 42, 10 se convirtió en 0. En todos estos casos la discrepancia con el original se halla tanto en español como en inglés.

En otras ocasiones, los indicios apuntan a decisiones de traducción para las que no hallamos justificación. Por ejemplo, en el párrafo 170 de CERD/C/362/Add.8, la expresión الجهاز التنفسي [“*al-jihāz al-tanaffusī*” = “el sistema respiratorio”] se tradujo como “*genital apparatus*” y “aparato genital”.

No es fácil, solo a la vista del producto, determinar con seguridad que la intermediación fue a través del inglés, y no en inglés a través del español. Sin embargo, como afirma Sánchez-Ratia (2002), "la escasez de traductores de árabe, combinada con el crecimiento exponencial de los documentos originales en este idioma, han obligado a muchos traductores del servicio a realizar traducciones por *relay*, es decir, a través de un tercer idioma". Los textos intermediarios, como también señala Sánchez Ratia (2018), son siempre las traducciones al inglés o al francés. Cabe descartar, por tanto, la intermediación del español.

De hecho, se registran indicios de que, en efecto, es el inglés la lengua de intermediación. Por ejemplo, en alguna ocasión, la indefinición de género en inglés, inexistente en árabe, se traslada al español. Así, en el párrafo 38 de CEDAW/C/LBN/3, la versión inglesa traduce المسنّات والمعّوقات ["*al-musinnāt wa al-muʿawwaqāt*" = "las mujeres de edad avanzada y las mujeres con discapacidad"] como "*the aged and the disabled*", en español "los ancianos y los discapacitados", y en francés "*les femmes âgées, les femmes handicapées*". Tan solo la traducción al español de CERD/C/MAR/17-18 parece estar intermediada del francés.

Por otro lado, en los documentos finalmente elegidos se registran pruebas positivas de que nos encontramos ante traducciones directas del árabe al español, por ejemplo:

- Documento I: Las versiones inglesa y francesa traducen يتيم ضمن قيد الأرملة أي يتيم الأب ["*yatīmun ḍimna qaydi l-ʾarmalati ʾay yatīmu l-ʾabi*" = "huérfano a cargo de la supérstite, es decir, huérfano de padre"] como "*who had lost either a mother or a father*" y "*orphelins*" respectivamente, mientras que en español figura "huérfanos de padre a cargo de la madre";

- Documento M: Las versiones inglesa y francesa traducen غالباً ما تحل بطريقة حبية عن طريق تدخل مفتشية الشغل ["*ghāliban mā taḥullu bi-ṭarīqatin ḥubbiyyatin ʿan ṭarīqi tadakhkhuli mufattishiyyati sh-shughli*" = "generalmente se resuelven por vía amistosa mediante la intervención de la inspección de trabajo"] como "*were usually settled amicably with the mediation of the labour inspectorate*" y "*généralement réglés de manière ponctuelle par l'intervention de l'Inspection*

du travail", mientras que en español figura "se solucionan la mayor parte de los casos por vía amistosa y no mediante la intervención de la Inspección de Trabajo". En este mismo documento, إعادة إسكان ["*ʾiʿādatu ʾiskānin*" = "realojamiento"] se tradujo al inglés y al francés como "*rehousing*" y "*reloger*", y al español como "rehabilitar las viviendas";

- Documento E: Las versiones inglesa y francesa traducen المشروع الوطني للإسكان ابني بيتك ["*al-mashrūʿu l-waṭaniyyu li-l-ʾiskāni ibni baytaka*" = "el proyecto nacional de vivienda construye tu casa"] como "*the 'Build your own house' housing project*" y "*projet national «Construis ton chez toi»*", mientras que en español figura "el Proyecto Nacional de Vivienda *`Ibni Baitak´* ("Hijo mío, tu casa")".

Por todo ello se puede considerar probado, siempre bajo la sombra de la duda metódica, que las traducciones analizadas son traducciones directas del árabe al español.

c) Eliminación de los elementos de distorsión

La falta de coincidencia entre el número de oraciones calculado automáticamente por WordSmith y el calculado de forma manual indicaba que era necesario preprocesar los documentos para evitar la distorsión de los datos.

Las pruebas realizadas con la muestra permitieron determinar los siguientes elementos de distorsión: portadillas, índices, tablas, títulos, referencias bibliográficas y enumeraciones en forma de viñetas o párrafos numerados, alfabetizados o separados con guiones.

Las enumeraciones son el elemento de distorsión con mayor diversidad ortotipográfica en los planos intradocumental, interdocumental e interlingüístico: en un mismo documento, la versión original y su traducción pueden presentar las enumeraciones separadas por punto y coma y en renglón aparte, por punto y coma en texto corrido o por coma en texto corrido. Además, en un único documento, las enumeraciones pueden separarse en una versión con punto y aparte y en otra mediante coma o punto y coma, por lo que WordSmith consideraría oraciones las primeras y partes de una sola oración las segundas.

A modo de ejemplo de lo anterior, en la página 16 del informe E/C.12/MAR/Q/2/Add.2 (M) leemos lo siguiente:

وتتلخص أهم المبادئ التي تتعلق بمسطرة الطلاق فيما يلي:

- ضرورة إجراء محاولة الصلح عند مسطرة الطلاق أو التطليق، باستثناء حالة التطليق للغيبة (المادة 113) وإمكانية تعيين حكمين ومجلس العائلة؛
- إسناد الإذن بالطلاق أمام العدول للمحكمة (قضاء الأسرة) (قبل المدونة الحالية كان مسنداً لقاضي التوثيق)؛
- تحديد أجل البت في دعاوى التطليق في ستة أشهر ما لم تكن هناك ظروف خاصة (المادة 113)؛
- عدم قابلية الأحكام الصادرة بالتطليق لأي طعن (المادة 1/128)؛
- سلطة المحكمة في اتخاذ التدابير المناسبة لفائدة الزوجة والأطفال أثناء النظر في النزاع بين الزوجين (المادة 121)؛
- تحديد المحكمة المختصة بالإذن بالطلاق ويمتد هذا الاختصاص المكاني إلى أربع حالات (المادة 79) أو أربع محاكم (رباعية الاختصاص)؛
- تحديد التدابير الرامية إلى ضمان أداء مستحقات المطلقة والأطفال في مواجهة الزوج المطلق (المادة 83) (إيداع مبلغ مسبقاً بالصندوق)؛
- تحديد المحكمة مستحقات الزوجة والأولاد فور توصلها بنسخة الطلاق وكذا مستحقاتهما عند الحكم بالتطليق (المادة 85 و87 و88)؛
- تحديد صلاحية الزوجة في حالة الطلاق المملك (المادة 89)؛
- النص على أسباب جديدة للتطبيق (الشقاق - الاتفاق - الإخلال بشرط في عقد الزواج) بالإضافة إلى الأسباب الواردة صراحة في المدونة السابقة (المادة 94-97 و99 و114)؛
- النص على حلول واضحة في بعض الاختلافات حول الخلع (المادة 120)؛
- وضع معايير واضحة لتذييل الأحكام الأجنبية الصادرة بالطلاق أو بالتطليق أو بالخلع أو بالفسخ - بالصيغة التنفيذية (2/128).

El análisis de ese fragmento mediante Wordsmith arroja los resultados siguientes:

- *Sentences: 1*

- *Mean in words: 189,00*

Para WordSmith, por tanto, todo el párrafo constituye una sola oración.

Una posible solución hubiera sido conservar las enumeraciones y concluir cada elemento con punto. Sin embargo, la transformación manual necesaria para ello hubiera implicado un esfuerzo muy considerable sin una mejora clara en la calidad de los resultados, ya que a menudo las enumeraciones no constituyen oraciones completas. Finalmente se adoptó como criterio eliminar las enumeraciones, mantener la oración precedente y sustituir los dos puntos por punto.

Una vez establecido el criterio aplicable a las enumeraciones, y para determinar el grado mínimo de intervención necesario que asegurara la calidad de los datos, se realizaron mediciones con WordSmith del número de oraciones (NO) y la media de palabras por oración (MPO) en tres fases: informes sin modificar (Fase 1), informes sin portadilla ni índice (Fase 2) e informes sin ninguno de los elementos de distorsión detectados (portadillas, índices, tablas, títulos, referencias bibliográficas y enumeraciones) (Fase 3). En las tablas 11 y 12 se muestran, respectivamente, los resultados de dichas mediciones y la variación de la desviación típica de la MPO en cada fase.

	Fase 1		*Fase 2*		*Fase 3*	
	NO	*MPO*	*NO*	*MPO*	*NO*	*MPO*
B	474	44,46	472	44,38	343	39,73
E	301	60,68	300	60,39	220	56,90
I	292	61,59	287	61,09	204	45,33
L	634	33,22	633	32,45	601	29,56
M	352	50,39	350	50,26	306	39,31
Y	307	54,97	306	53,56	232	35,05

Tabla 11. Número de oraciones y media de palabras por oración en las tres fases de preprocesamiento de la muestra

	σ (Fase 1)	*σ (Fase 2)*	*σ (Fase 2) - σ (Fase 1)*	*σ (Fase 3)*	*σ (Fase 3) - σ (Fase 2)*
B	46,09	46,09	0	27,53	-18,56
E	55,19	55,28	0,09	36,62	-18,66
I	50,48	50,5	0,02	38,75	-11,75
L	24,29	18,39	-5,9	15,46	-2,93
M	43,69	43,34	-0,35	32,68	-10,66
Y	53,51	49,47	-4,04	20,82	-28,65
Σ			-10,18		-91,21

Tabla 12. Desviaciones típicas de las medias de palabra por oración y su variación en las tres fases de preprocesamiento de la muestra

Los valores de la MPO obtenidos en la fase 2 y, sobre todo, en la fase 3 son más fiables y homogéneos a la vista de la disminución de la desviación típica. De hecho, esa disminución es 9 veces mayor al pasar de la fase 2 a la 3 que al pasar de la 1 a la 2. Era necesario, por tanto, eliminar todos los elementos de distorsión.

Por último, la alta desviación típica media de los documentos, incluso en la fase 3 indica la coexistencia en la muestra, incluso tras eliminar los elementos de distorsión, de oraciones excepcionalmente largas y cortas.

d) Segmentación

Una vez eliminados los elementos de distorsión y con objeto de facilitar la detección de las *stopwords* se procedió a dividir los originales que componen la muestra en 114 segmentos de entre 600 y 650 palabras gráficas (342 segmentos con las correspondientes traducciones al español y al inglés). Ese número de palabras se determinó en aras de la uniformidad metodológica (Touir et al. 2008, p. 1011; García Barrero, Feria y Turell, 2012, p. 45) y con un intervalo que permitiera mantener oraciones completas.

Los segmentos obtenidos comprenden 270 300 palabras gráficas: 71 257 en árabe, 109 287 en español y 89 756 en inglés, lo que supone un 64 % de la muestra sin procesar (423 851 palabras). En la tabla 13 se presenta el número de segmentos por documento.

	Nº de segmentos
B	21
E	20
I	14
L	28
M	18
Y	13
TOTAL	114

Tabla 13. Número de segmentos por documento

Tras la segmentación se calculó con WordSmith el número de oraciones en árabe, español e inglés (OA, OE y OI respectivamente) y las ratios OE/OA y OI/OA. Adelantamos que los valores de OE/OA oscilaron entre 3,6 y 0,9, con un valor promedio ($\bar{x}$) de 1,68 y una desviación típica (σ) de 0,63, y los de OI/OA entre 3,7 y 0,9, con $\bar{x} = 1{,}67$ y $\sigma = 0{,}6$. No se detectaron valor atípicos mediante el test de Grubbs. Por tanto, en torno al 74 % de los valores de OE/OA y el 59,6 % de los valores de OI/OA se encuentran a no más de una desviación típica de la media y, respectivamente el 96 % y el 95,6 % a no más de dos. Los datos, por tanto, siguen presentando una distribución normal.

2.2. Traducción del español y del inglés al árabe

Para analizar el efecto de la direccionalidad sobre la división de oraciones necesitamos datos sobre el número de oraciones de los originales en español e inglés y en sus correspondientes traducciones al árabe. Para ello emplearemos, respectivamente, los

corpus 3 y 4. En ambos casos, por tanto, no se precisa determinar las *stopwords* de los originales y tampoco, por consiguiente, segmentar las muestras.

Para asegurar la comparabilidad, los corpus 3 y 4 comprenden seis documentos originales en español o inglés y sus traducciones al árabe, y sus características son similares a la muestra extraída del Corpus 2, inclusive en número de palabras gráficas. En ambos corpus se eliminaron los elementos de distorsión.

Los datos obtenidos se presentan y analizan en el Capítulo IV.

Capítulo IV: Datos y análisis de los datos

Este capítulo consta de dos partes. En la primera presentaremos los datos extraídos de los corpus y en la segunda procederemos a analizar dichos datos.

En adelante, por "muestra del Corpus 2" debe entenderse esa muestra una vez preprocesada (fase 3) y por "muestra segmentada", esa misma muestra preprocesada tras su segmentación.

1. Datos

a) Número de oraciones

En la tabla 14 se presentan los datos relativos al número de oraciones en árabe (OA), español (OE) e inglés (OI) en la muestra del Corpus 2, así como sus respectivas ratios. Los datos se obtuvieron mediante WordSmith.

	OA	*OE*	*OI*	*OE/OA*	*OI/OA*	*OE/OI*
B	343	549	562	1,60	1.64	0.98
E	220	499	478	2,27	2.17	1.04
I	204	430	366	2,11	1.79	1.17
L	601	659	591	1,10	0.98	1.12
M	306	419	557	1,37	1.82	0.75
Y	232	335	353	1,44	1.52	0.95
TOTAL	1 906	2 891	2 907	1,52	1.53	0.99
			σ	0,45	0,40	0,15

Tabla 14. Número de oraciones en árabe, español e inglés en la muestra del Corpus 2 (recuento automático)

En la tabla 15 se presentan los datos relativos al número de OA, OE y OI, y sus respectivas ratios, en la muestra segmentada del Corpus 2. En este caso, los datos se extrajeron mediante recuento manual. Estos datos son prácticamente iguales que los que figuran en la tabla 14, aunque en estos se evita el error leve introducido por la herramienta informática. Ofrecemos los datos de la tabla 14 con miras a asegurar la comparabilidad con los datos relativos a la traducción al árabe, que se extrajeron únicamente de las muestras (recordemos, no segmentadas) de los Corpus 3 y 4.

	OA	*OE*	*OI*	*OE/OA*	*OI/OA*	*OE/OI*
B	334	529	538	1,58	1.61	0,98
E	219	498	477	2,27	2.18	1,04
I	194	407	343	2,10	1.77	1,19
L	585	640	585	1,09	1.00	1,09
M	275	388	494	1,41	1.80	0,79
Y	229	332	351	1,45	1.53	0,95
TOTAL	1 836	2 794	2 788	1,52	1.52	1,00
			σ	0,45	0,39	0,14

Tabla 15. Número de oraciones en árabe, español e inglés en la muestra segmentada del Corpus 2 (recuento manual)

En las tablas 16 y 17 se presentan, respectivamente, los datos relativos al número de OA, OE y OI, y sus respectivas ratios, en las muestras del Corpus 3 y del Corpus 4.

	OE	*OA*	*OE/OA*
A	408	388	1,05
C	294	297	0,99
E	666	646	1,03
M	436	457	0,95
P	366	358	1,02
V	439	458	0,96
TOTAL	2 609	2 604	1,00
		σ	0,04

Tabla 16. Número de oraciones en árabe y español en la muestra del Corpus 3 (recuento automático)

	OI	*OA*	*OI/OA*
AU	577	576	1,00
CA	304	306	0,99
EU	435	364	1,20
IN	415	401	1,03
NI	271	267	1,01
RU	973	984	0,99
TOTAL	2 975	2 898	1,03
		σ	0,08

Tabla 17. Número de oraciones en árabe e inglés en la muestra del Corpus 4 (recuento automático)

b) Número de palabras gráficas

En todos los casos, el recuento del número de palabras gráficas en árabe (PA), en español (PE) y en inglés (PI) se realizó con el contador de palabras de Word, que resultó ser la herramienta más fiable.

Los datos relativos a las traducciones del árabe al español y al inglés se obtuvieron de tres fuentes: el Corpus 2, la muestra del Corpus 2 y la muestra segmentada del Corpus 2. En las tablas 18, 19 y 20 se presentan los datos relativos al número de PA, PE y PI obtenidos de esas tres fuentes, así como sus respectivas ratios.

PA	*PE*	*PI*	*PE/PA*	*PI/PA*	*PE/PI*
3 541 670	5 139 440	4 424 882	1,45	1,25	1,16
		σ	0,11	0,08	0,08

Tabla 18. Número de palabras gráficas en árabe, español e inglés en el Corpus 2, y sus respectivas ratios

	PA	*PE*	*PI*	*PE/PA*	*PI/PA*	*PE/PI*
B	13 627	18 990	16 387	1,39	1,20	1,16
E	12 517	20 862	15 826	1,67	1,26	1,32
I	92 48	13 267	11 717	1,43	1,27	1,13
L	17 766	27 925	23 396	1,57	1,32	1,19
M	12 029	18 874	15 303	1,57	1,27	1,23
Y	8 132	12 325	10 384	1,52	1,28	1,19
TOTAL	73 319	112 243	93 013	1,53	1,27	1,21
			σ	0,10	0,04	0,07

Tabla 19. Número de palabras gráficas en árabe, español e inglés en la muestra del Corpus 2, y sus respectivas ratios

	Nº de segmentos	*PA*	*PE*	*PI*	*PE/PA*	*PI/PA*	*PE/PI*
B	21	13 094	18 371	15 784	1,40	1,21	1,16
E	20	12 517	20 813	15 793	1,66	1,26	1,32
I	14	8 821	12 633	11 061	1,43	1,25	1,14
L	28	17 456	27 467	23 145	1,57	1,33	1,19
M	18	11 245	17 678	13 746	1,57	1,22	1,29
Y	13	8 124	12 325	10 227	1,52	1,26	1,21
TOTAL	114	71 257	109 287	89 756	1,53	1,26	1,22
				σ	0,10	0,04	0,07

Tabla 20. Número de palabras gráficas en la muestra segmentada del Corpus 2, y sus respectivas ratios

En las tablas 21 y 22 se presentan los datos relativos al número de palabras gráficas en las traducciones del español y el inglés al árabe. Estos datos se extrajeron de las muestras correspondientes a los Corpus 3 y 4.

	PE	*PA*	*PE/PA*
A	15 343	12 111	1,27
C	9 649	8 297	1,16
E	22 816	19 066	1,20
M	17 179	13 513	1,27
P	13 586	11 626	1,17
V	18 026	14 553	1,24
TOTAL	96 599	79 166	1,22
		Σ	0,05

Tabla 21. Número de palabras gráficas en la muestra del Corpus 3

	PI	*PA*	*PI/PA*
AU	15 128	13 979	1,08
CA	7 812	7 371	1,06
EU	10 184	10 637	0,96
IN	10 342	9 599	1,08
NI	7 229	6 144	1,18
RU	25 518	23 669	1,08
TOTAL	76 213	71 399	1,07
		Σ	0,07

Tabla 22. Número de palabras gráficas en la muestra del Corpus 4

c) Media de palabras por oración

En las tablas 23 y 24 se muestran, respectivamente los valores de la media de palabras gráficas por oración (MPO) para cada una de las tres lenguas (MPO A, MPO E y MPO I) en la muestra del Corpus 2 y en la muestra segmentada. En el primer caso, el MPO se calculó con WordSmith y el número de palabras gráficas, con Word; en el

segundo, el número de oraciones se determinó manualmente y el número de palabras gráficas, con Word.

	MPO A	*MPO E*	*MPO I*
B	39,73	34,59	29,16
E	56,90	41,81	33,11
I	45,33	30,85	32,01
L	29,56	42,37	39,59
M	39,31	45,05	27,47
Y	35,05	36,79	29,42
TOTAL	38,47	38,82	32,00
σ	9,40	5,39	4,33

Tabla 23. Media de palabras por oración en la muestra del Corpus 2

	MPO A	*MPO E*	*MPO I*
B	39,20	34,73	29,34
E	57,16	41,79	33,11
I	45,47	31,04	32,25
L	29,84	42,92	39,56
M	40,89	45,56	27,83
Y	35,48	37,12	29,14
TOTAL	38,81	39,11	32,19
σ	9,37	5,50	4,27

Tabla 24. Media de palabras por oración en la muestra del Corpus 2 segmentada

En las tablas 25 y 26 se presentan los valores de MPO A, MPO E y MPO I en los Corpus 3 y 4, respectivamente. El número de oraciones se calculó con WordSmith y el de palabras, con Word.

	MPO E	*MPO A*
A	37,61	31,21
C	32,82	27,94
E	34,26	29,51
M	39,40	29,57
P	37,12	32,47
V	41,06	31,78
TOTAL	37,03	30,40
σ	3,09	1,70

Tabla 25. Media de palabras por oración en la muestra del Corpus 3

	MPO I	*MPO A*
AU	26,22	24,27
CA	25,70	24,09
EU	23,41	29,22
IN	24,92	23,94
NI	26,68	23,01
RU	26,23	24,05
TOTAL	25,62	24,64
σ	1,20	2,23

Tabla 26. Media de palabras por oración en la muestra del Corpus 4

d) *Stopwords*

La comparación entre la muestra segmentada del Corpus 2 y sus correspondientes traducciones al español y al inglés nos permitió localizar de forma manual 996 *stopwords* en las traducciones del árabe al español y 983, en las del árabe al inglés. En los Anexos III y IV figura la relación completa de esas *stopwords*.

En la tabla 27 se presentan datos relativos a la distribución de esas *stopwords*, desglosados por número de segmentos del documento en cuestión (SEG), número de

stopwords localizadas en la traducción al español (STW E) y al inglés (STW I) de ese documento, y ratios de STW E/SEG, STW I/SEG y STW E/STW I.

	SEG	*STW E*	*STW E/SEG*	*STW I*	*STW I/SEG*	*STW E/STW I*
B	21	201	9,6	215	10,2	0,93
E	20	278	13,9	258	12,9	1,08
I	14	213	15,2	155	11,1	1,37
L	28	85	3	7	0,3	12,14
M	18	114	6,3	220	12,2	0,52
Y	13	105	8,1	128	9,8	0,82
TOTAL	114	996	8,7	983	8,6	1,01

Tabla 27. Número de segmentos por documento, número de stopwords localizadas en la traducción al español y al inglés y ratios de STW E/SEG, STW I/SEG y STW E/STW I

Con frecuencia, las *stopwords* son coincidentes en las traducciones al español y al inglés. En el Anexo V figura la relación completa de ellas. En la tabla 28, además, se presentan los datos relativos al número total de *stopwords* comunes (STW C) por documento, así como su porcentaje respecto al total de *stopwords* del documento.

	STW C	*% STW CE*	*% STW CI*
B	144	72 %	67 %
E	196	71 %	76 %
I	129	61 %	83 %
L	4	5 %	57 %
M	95	83 %	43 %
Y	79	75 %	62 %
TOTAL	647	65 %	66 %

Tabla 28. Número total de stopwords comunes por documento y sus porcentajes respecto al total de stopwords del documento

En las traducciones al español se localizaron 996 *stopwords* (*tokens* divisores) y 400 *types* divisores diferentes, es decir, un promedio de 2,49 *tokens* por *type*. Esto supone la existencia de numerosas *stopwords* con una sola ocurrencia activa: el 29,02 % de los *tokens* y el 72,25 % de los *types*.

En las traducciones al inglés se localizaron 983 *stopwords* y 415 *types* divisores, con una media de 2,37 *tokens* por *type*. En este caso, las *stopwords* con una sola ocurrencia activa constituyen el 30,72 % de los *tokens* y el 72,77 % de los *types*.

Las *stopwords* comunes presentan 647 *tokens* y 271 *types*, es decir, 2,39 *tokens* por *type*, y las *stopwords* con una sola ocurrencia activa supusieron el 29,83 % de los *tokens* y el 71,22 % de los *types*.

Por otro lado, se asignaron cuatro valores numéricos a cada *stopword*: ocurrencias totales (OT), ocurrencias activas (OA), probabilidad de activación (PA) y distribución (D). Estas variables se definen como sigue:

- OT: Número total de ocurrencias de la *stopword*, con independencia de que en ellas se divida o no oración. El recuento se realizó con la función WordList de WordSmith;

- OA: Número de divisiones originadas por la *stopword*, es decir, de veces que la palabra gráfica se eligió para dividir oración. El recuento en este caso se realizó manualmente;

- PA: Probabilidad (expresada en términos porcentuales) de que la *stopword* origine una división. Para calcular esa posibilidad dividimos OA entre OT y multiplicamos por 100;

- D: Distribución porcentual de las ocurrencias de la *stopword* en la muestra, con independencia de que la ocurrencia origine o no división. El recuento se realizó con WordSmith y el cálculo se realizó mediante una simple regla de tres (si una *stopword* aparece en los 114 segmentos, D = 100; si lo hace en uno solo, D = 0,88).

En las tablas 29 y 30 se presentan las *stopwords* con PA>50 y OA>1 que figuran en las traducciones, respectivamente, al español y al inglés, con mención de sus OT, OA y D.

	PA (%)	*OT*	*OA*	*D (%)*
ويشترط	100	3	3	2,63
وتتواصل	100	2	2	1,75
وتحدد	100	2	2	1,75
وتضع	100	2	2	1,75
وتواصل	100	2	2	1,75
وتوالي	100	2	2	1,75
ولكنه	100	2	2	1,75
ونشير	100	2	2	0,88
ويتضح	100	2	2	1,75
ويرجى	100	2	2	1,75
وبسبب	83	6	5	3,51
ويجوز	80	5	4	3,51
وعقب	75	4	3	2,63
ويجب	75	4	3	2,63
فضلا	69	16	11	9,65
وجاري	67	12	8	7,01
ففي	67	6	4	5,26
وتعد	67	6	4	5,26
وتلقي	67	3	2	2,63
وللنزيل	67	3	2	2,63
ولذلك	67	3	2	2,63
ويصدر	67	3	2	2,63
فإذا	60	10	6	6,14
وتتمثل	60	5	3	4,39

وسيتم	60	5	3	4,39
وهذه	57	7	4	6,14
وإذا	56	9	5	6,14

Tabla 29. Stopwords *con PA >50 y OA >1, con mención de su OT, OA y D, en las traducciones al español*

	PA (%)	*OT*	*OA*	*D (%)*
ويجب	100	4	4	2,63
ويحظر	100	3	3	2,63
والخروج	100	3	3	1,75
وانتهت	100	2	2	1,75
وأيضاً	100	2	2	1,75
وتتواصل	100	2	2	1,75
وتضع	100	2	2	1,75
وتواصل	100	2	2	1,75
وتوالي	100	2	2	1,75
وصرفت	100	2	2	1,75
ويتولى	100	2	2	1,75
ويرجى	100	2	2	1,75
وأنه	100	2	2	0,88
وبسقوط	100	2	2	0,88
وقاموا	100	2	2	0,88
ونشير	100	2	2	0,88
وبسبب	83	6	5	3,51

ويجوز	80	5	4	3,51
وهما	75	4	3	3,51
وتختص	75	4	3	2,63
وتكون	75	4	3	2,63
وعقب	75	4	3	2,63
فإذا	70	10	7	6,14
وإحالة	67	3	2	2,63
وإعطاء	67	3	2	2,63
وتلقي	67	3	2	2,63
ولذلك	67	3	2	2,63
وللنزيل	67	3	2	2,63
ويشترط	67	3	2	2,63
ويصدر	67	3	2	2,63
ويكون	67	3	2	2,63
وتتمثل	60	5	3	4,39
وجاري	58	12	7	7,01
وهذه	57	7	4	6,14

Tabla 30. Stopwords *con PA >50 y OA >1, con mención de su OT, OA y D, en las traducciones al ingles*

En las tablas 31 y 32 figuran las *stopwords* con D >50 y OA >1 en las traducciones al español y al inglés, respectivamente, con mención de su OT, OA, PA y posición en la lista de palabras gráficas más frecuentes en la muestra segmentada. La lista de frecuencia se confeccionó utilizando WordSmith. En el Anexo VI figuran las 100 palabras más frecuentes en la muestra segmentada.

	D (%)	*OT*	*OA*	*PA (%)*	*Posición*
في	100	2255	6	0,3	1
من	100	1979	8	0,4	2
على	99,1	994	10	1	3
التي	96,5	455	2	0,4	5
مع	94,7	440	6	1,4	7
كما	92,1	317	95	30	11
هذا	79,0	249	3	1,2	16
وقد	79,0	211	90	42,7	22
تم	67,5	241	7	2,9	18
الذي	63,2	139	2	1,4	37
حيث	58,8	139	54	38,8	38
وفي	57,0	110	20	18,2	49
غير	56,1	112	2	1,8	48
وذلك	53,5	103	23	22,3	51

Tabla 31. Stopwords *con D >50 y OA >1 en las traducciones al español, con mención de sus OT, OA, PA y posición en la lista de palabras más frecuentes en la muestra segmentada*

	D (%)	*OT*	*OA*	*PA (%)*	*Posición*
في	100	2255	6	0,3	1
من	100	1979	4	0,2	2
على	100	995	6	0,6	3
كما	100	318	76	23,9	11
مع	94,7	440	7	1,6	7
هذا	79	249	2	0,8	16
وقد	79	212	79	37,3	22
تم	67,5	241	3	1,2	18
الذي	63,2	139	4	2,9	37
كل	60,5	134	2	1,5	41
حيث	58,8	139	66	47,5	38
وفي	57,0	110	20	18,2	49
وذلك	53,5	103	22	21,4	51

Tabla 32. Stopwords *con D >50 y OA >1 en las traducciones al inglés, con mención de sus OT, OA, PA y posición en la lista de palabras más frecuentes en la muestra segmentada*

En las tablas 33 y 34 figuran las *stopwords* con D>20 y PA>20 en las traducciones al español y al inglés, respectivamente.

	D %	*PA %*
كما	92,11	30,0
وقد	78,95	42,7
حيث	58,77	38,8
وذلك	53,51	22,3
ومن	45,61	22,2
وهي	40,35	28,6
وعلى	34,21	23,7
بالإضافة	22,81	20,6
وتم	21,93	27,8
وهو	21,05	23,1

Tabla 33. Stopwords con D>20 y PA>20 en las traducciones al español

	D %	*PA %*
كما	92,11	23,9
وقد	78,95	37,3
حيث	58,77	47,5
وذلك	53,51	21,4
وهي	40,35	28,6
بالإضافة	24,56	22,2
وتم	22,81	27
إضافة	22,81	20,6
وهو	21,05	20,5
إذ	20,18	29,6

Tabla 34. Stopwords *con D>20 y PA>20 en las traducciones al inglés*

En la tabla 35 figuran las 10 *stopwords* comunes a las traducciones al español y al inglés con datos desglosados por número de ocurrencias comunes (OC), ocurrencias activas en español (OA E) y en inglés (OA I) y porcentaje de coincidencia respecto al número de ocurrencias activas de la *stopword* en cada lengua (% C E y % C I).

STW	OC	OA E	OA I	% C E	% CI
وقد	70	90	79	78 %	89 %
كما	69	95	76	73 %	91 %
حيث	41	54	66	76 %	62 %
وذلك	12	23	22	52 %	55 %
وفي	12	20	20	60 %	60 %
ومن	12	18	16	67 %	75 %
ويتم	11	12	13	92 %	85 %
وعلى	10	14	11	71 %	91 %
وتم	8	10	10	80 %	80 %
وهي	8	16	16	50 %	50 %

Tabla 35. Stopwords *comunes con datos desglosados por número de ocurrencias comunes (OC), ocurrencias activas en español (OA E) y en inglés (OA I) y porcentaje de coincidencia respecto al número de ocurrencias activas de la* stopword *en cada lengua (% C E y % C I).*

En la tabla 36 figuran las *stopwords* comunes a las traducciones al español y al inglés con D>20 y PA>20.

	D %	*PA %*
كما	92,11	21,8
وقد	78,95	33,2
حيث	58,77	29,5
وتم	21,93	22,2

Tabla 36. Stopwords *comunes a las traducciones al español y al inglés con D>20 y PA>20*

El gran número de *stopwords* con una sola ocurrencia activa y la fuerte correlación lineal negativa entre PA y D para *types* con OA >1 (r [coeficiente de correlación muestral de Pearson] = -0,58 para la traducción al español y r = -0,59 para la traducción al inglés) se explican por la aglutinación gráfica de algunos clíticos en árabe, en particular la conjunción copulativa *wa* (و).

Aunque inicialmente consideramos limitar nuestro análisis a las palabras gráficas con objeto de facilitar la labor y la eficiencia de las herramientas de segmentación, los datos anteriores nos revelaron la necesidad de tokenizar los datos, no el corpus. El proceso se realizó manualmente para evitar la considerable tasa de error de herramientas como MADAMIRA y FARASA (Pasha et al., 2014; Neme y Paumier, 2020). La tokenización manual fue perfectamente factible gracias al número limitado de elementos a tokenizar: hubiera sido inasumible en caso de tokenizar, como sugiere la literatura previa, la totalidad del corpus como método de preprocesamiento.

La tokenización mostró que en las traducciones al español el 61,8 % de los *tokens* divisores y el 68% de los *types* comenzaban por la conjunción *wa* (و) aglutinada gráficamente. En las traducciones al inglés, esas cifras fueron, respectivamente, el 63,4 % y el 68,4 %. También comenzaban con *wa* el 67,7 % de los *tokens* comunes y el 75,6 % de los *types*.

Las *stopwords* que comienzan con *wa* pueden dividirse en dos grupos en función del elemento que siga a *wa*:

- Aquellas a las que sigue un verbo o partículas que preceden a verbo, como *qad* (قد), *laqad* (لقد), *sawfa* (سوف), *sa-* (س), *idha* (إذا), *lam* (لم) y *la* (لا) seguida de verbo;
- Aquellas a las que sigue cualquier otra categoría gramatical.

En las tablas 37 y 38 figuran los porcentajes de esos dos grupos de *stopwords* introducidas por *wa*, con datos desglosados por *types* y *tokens*, en las traducciones al español y al inglés. En la tabla 39 figuran los mismos datos para las *stopwords* comunes.

	Tokens *introducidos por* wa *(%)*	Types *introducidos por* wa *(%)*	*Total de* tokens *(%)*	*Total de* types *(%)*
Wa + *verbo o partícula verbal*	57	58,5	35,2	39,8
Wa + *otras categorías*	43	41,5	26,6	28,2
		TOTAL	61,8	68

Tabla 37. Stopwords *introducidas por* wa *en las traducciones al español, con datos desglosados por* types *y* tokens

	Tokens *introducidos por* wa *(%)*	Types *introducidos por* wa *(%)*	*Total de* tokens *(%)*	*Total de* types *(%)*
Wa + *verbo o partícula verbal*	54,1	56,3	34,3	38,5
Wa + *otras categorías*	45,9	43,7	29,1	29,9
		TOTAL	63,4	68,4

Tabla 38. Stopwords *introducidas por* wa *en las traducciones al inglés, con datos desglosados por* types *y* tokens

	Tokens *introducidos por* wa *(%)*	Types *introducidos por* wa *(%)*	*Total de* tokens *(%)*	*Total de* types *(%)*
Wa + *verbo o partícula verbal*	60,7	57,1	41,1	43.2
Wa + *otras categorías*	39,3	42,9	26,6	32.5
		TOTAL	67,7	75,6

Tabla 39. Stopwords *comunes introducidas por* wa*, con datos desglosados por* types *y* tokens

2. Análisis de datos

En este apartado analizaremos los datos presentados en la primera parte del capítulo. Los datos sobre las traducciones del español y el inglés al árabe serán analizados de forma separada en el apartado 2.e).

a) Representatividad

A nuestro juicio, los documentos originales analizados son representativos de los textos informativos en lengua árabe producidos en el primer cuarto del siglo XXI. Nuestros datos y análisis no deben generalizarse a otros tipos de texto en árabe estándar contemporáneo.

Las traducciones de los documentos analizados se produjeron en un marco institucional de gran prestigio. Podemos, pues, conceder un alto nivel de confianza a esas traducciones. También podemos considerar probado que las traducciones al español analizadas no son traducciones intermediadas (véase III 2.1b)).

No obstante, conviene tener en cuenta que el carácter institucional de esas traducciones puede haber introducido un elemento de conservadurismo ausente, o presente pero en menor grado, en las traducciones de textos informativos producidas en otros contextos sociales. Por tanto, algunas de las tendencias estadísticas registradas en esta investigación podrían matizarse en otros contextos comunicativos.

b) Número de oraciones

Para analizar las ratios de número de oraciones empleamos los datos recogidos en la tabla 15, dado que se obtuvieron mediante recuento manual y, aunque prácticamente son idénticos a los obtenidos con WordSmith y recogidos en la tabla 14, son más fiables.

El número de oraciones en español (OE) e inglés (OI) es superior al número de oraciones en árabe (OA) en todos los documentos salvo L, en que OI y OA son coincidentes. La primera conclusión, pues, es que, considerada la muestra en su conjunto, los traductores tienden a dividir las oraciones del original, o en todo caso las mantienen, pero no las fusionan.

De hecho, los traductores a ambas lenguas tienden a dividir las oraciones en la misma medida y con similar regularidad. Adviértase que los valores de OE/OA y OI/OA para el conjunto de la muestra son idénticos (1,52), con desviaciones típicas también similares (σ = 0,45 y 0,39) y con el 100 % de los valores a no más de dos desviaciones típicas de la media. Dicho de otro modo, cada oración en árabe generó de media 1,5 oraciones en español y en inglés.

Por otro lado, los valores más altos de OE/OA y OI/OA corresponden a documentos cuyos originales árabes presentan también valores de MPO más altos, como demuestra la elevada correlación lineal positiva entre OE/OA y MPO A, por un lado, y OI/OA y MPO A, por otro. Esa correlación es prácticamente idéntica en las dos lenguas (respectivamente, r = 0,93 y 0,94). De igual modo, cuanto más cortas son las oraciones de un original, más bajos son los valores de OE/OA y OI/OA.

Como consecuencia de lo anterior, la ratio OE/OI es prácticamente 1 (con σ = 0.14 y el 100 % de los valores a no más de dos desviaciones típicas), es decir, que en la muestra considerada en su conjunto, hay prácticamente el mismo número de oraciones en español y en inglés. Sin embargo, a esa media subyacen variaciones interdocumentales que van de 1,19 en I a 0,79 en M. Recordemos que la ausencia de esas variaciones concretas, y la plena concordancia en el número de oraciones, fue precisamente lo que nos indujo a realizar el análisis cualitativo que finalmente nos permitió desechar los documentos en los que todo parecía indicar la presencia de *relé* (véase III 2.1b)).

Esos resultados, tan similares y regulares, podrían interpretarse como conservadores debido al contexto institucional de las traducciones y, por tanto, reflejarían el grado

mínimo de intervención necesario para asegurar la legibilidad. Este punto es concordante con la elevada correlación lineal positiva entre OE/OA y MPO A, por un lado, y OI/OA y MPO A, por otro.

Como conclusión, podemos afirmar que los datos demuestran que, en general, los traductores al español y al inglés, debido al contexto institucional al que sirven, dividen las oraciones solo cuando lo consideran necesario y solo en la medida imprescindible y que, pese a ello, cada oración árabe produjo de media 1,5 oraciones en español y en inglés, cifra ciertamente elevada.

c) Número de palabras

Los datos revelan que en todos los casos PE > PI > PA. Los valores medios de PE/PA, PI/PA y PE/PI son, respectivamente, 1,53; 1,27 y 1,21 (σ = 0,10, 0,04 y 0,07, con el 100 % de los valores a no más de dos desviaciones típicas de la media) (véase tabla 19). Esos valores son similares a los del corpus paralelo principal sin procesar (PE/PA = 1,45 con σ = 0.11; PI/PA = 1,25 con σ = 0,08 y PE/PI = 1,16 con σ = 0,08). Ese menor número de palabras en árabe se explicaría por la aglutinación gráfica (véase II.2). Analizaremos más adelante las posibles explicaciones a estos datos.

La prácticamente nula correlación lineal entre OE/OA y PE/PA (r = -0,07) indica que el aumento en el número de oraciones en español no conlleva un aumento proporcional del número de palabras. Sin embargo, en inglés se observa una correlación línea negativa acusada entre OI/OA y PI/PA (r = -0,62). En el caso del inglés, por tanto, una mayor intervención en términos de división de oraciones supone claramente una tendencia a la disminución en el número de palabras.

Se observa, pues, una diferencia acusada entre los traductores al español y al inglés: los primeros dividen más las oraciones cuanto más largas son en árabe, pero cuando lo hacen no parecen intervenir de ningún otro modo que afecte al número de palabras; los segundos, sin embargo, tienden también a dividir más las oraciones cuanto más largas son en árabe, pero cuanto más dividen las oraciones en mayor medida parecen recurrir también a otras estrategias (presumiblemente de simplificación) que disminuyen el número de palabras.

En lo que se refiere a la correlación lineal entre PE y PI, como era lógico suponer, es cercana a 1 (0,98).

d) Media de palabras por oración

Las MPO en el conjunto de la muestra segmentada son MPO E = 39,11, MPO I = 32,19 y MPO A = 38,81. Los valores medios de MPO en español y árabe son muy similares: es lógico, ya que las traducciones al español comprendían un 53 % más de palabras que los originales y un 52 % más de oraciones. MPO I, sin embargo, es ligeramente más bajo.

La correlación lineal entre MPO A y MPO E es prácticamente nula (r = -0,06) y entre MPO A y MPO I es moderadamente negativa (r = -0.220). En consecuencia, se registra gran variabilidad interdocumental: B, E, I presentan una MPO A mucho mayor que en español e inglés; L, M, Y, una MPO E mayor que en inglés y árabe, e I, una MPO I mayor que en español. Dicho de otro modo, un valor elevado de MPO A no conlleva necesariamente un valor elevado de MPO E, pero si una ligera disminución de MPO I. Ello constituye otro indicio de que ciertos documentos exigen un grado mayor de intervención, ya sea mediante la división de oraciones o mediante otros mecanismos que supongan una disminución del número de palabras.

Por otro lado, MPO E y MPO I presentan una correlación r = 0,170, muy bajo para lo que cabría esperar si el grado de intervención de los traductores dependiera únicamente de las características de los originales. Los valores de MPO de I y M son de nuevo reveladores: el primero es el único en el que el valor de MPO I es superior al de MPO E, mientras que el segundo es el que mayor diferencia positiva registra entre MPO E y MPO I.

El traductor al español del informe I se encontró ante un documento con MPO A = 45,47 (el segundo valor más alto) y produjo una traducción con MPO E de 31,04 (el valor más bajo) y 2,1 oraciones por oración árabe (el segundo valor más alto). El valor de PE/PA (1,43) es también el segundo más bajo y se encuentra muy por debajo de la media (1,53). El grado de intervención en este caso es, pues, elevado tanto en lo que respecta a la división de oraciones como a la disminución del número de palabras.

En el extremo contrario, el traductor al español de M se encontró ante un original con MPO A = 40,89, es decir, cercano a la media (38,81), y produjo un informe con MPO E = 45,56 (el valor más elevado), 1,41 oraciones por oración árabe (el segundo valor más bajo) y una PE/PA por encima de la media.

Las importantes diferencias entre las versiones españolas de los documentos I y M parecen indicar un mayor grado de intervención en el primero y una traducción más literal en el segundo. En la versión inglesa de M, el traductor produce un documento con el valor de MPO I más bajo de todas las versiones inglesas (27,83) mientras que el español, ante el mismo original, hacía justamente lo contrario. Asimismo el valor de OI/OA es el segundo más alto, mientras que el de OE/OA para el mismo documento es, como dijimos, el segundo más bajo. Además, el valor de PI/PA es el segundo más bajo (1,22), cuatro puntos por debajo de la media, mientras que en la versión española PE/PA está cinco puntos por encima de la media.

Por su parte, la versión inglesa del documento L es un ejemplo perfecto de grado de intervención nulo, ya que el traductor mantiene todos los puntos del original y ni divide ni fusiona oración alguna. El autor de la versión española, por su parte, introduce 1,09 oraciones por oración árabe, el valor mínimo de OE/OA registrado. En este caso, la MPO A (29,84), el más bajo de todos, permite que el grado de intervención de ambos traductores sea nulo en la versión inglesa y muy reducido en la española.

Como hemos señalado con anterioridad, la tendencia a la división de oraciones se acentúa en los documentos con valores más altos de MPO A, que exigirían un mayor grado de intervención para asegurar la legibilidad en español y en inglés. Sin embargo, como también hemos podido observar, para un mismo original árabe se registran grados de intervención muy diferentes en español y en inglés. Por tanto, la tendencia a la división de oraciones depende tanto de las características del original como de las preferencias del traductor, pese a las restricciones impuestas por el carácter institucional de las traducciones analizadas.

Por otro lado, como demuestran sus valores de desviación típica, el grado de homogeneidad de los valores de MPO es diferente para las tres lenguas. En árabe, con una media de 38,81 y una desviación típica de 9,37, la MPO oscila entre 57,16 (E) y 29,84 (L). En español, los valores extremos oscilan entre 45,56 (M) y 31,04 (I), con una

media de 39,11 y una desviación típica de 5,5. En inglés, los valores oscilan entre 39,56 (L) y 29,14 (Y), con una media de 32,19 y una desviación típica de 4,27.

Esas diferencias en las MPO A confirma el uso errático de los signos de puntuación (véase II.2), incluso en textos tan institucionalizados como los de nuestro corpus. En las traducciones en inglés, frente a la flexibilidad del árabe en este registro, las oraciones son muy uniformes en cuanto a longitud. En las traducciones al español registramos una posición intermedia. En definitiva, con mayor grado de uniformidad en español e inglés, y con tendencias más o menos acusadas en unos traductores y otros, con independencia de la lengua, nuestros datos reflejan en las traducciones una clara tendencia a la estandarización.

Además del uso errático de los signos de puntuación, caben otras dos posibles explicaciones a las importantes diferencias en los valores de MPO A: 1) la existencia de diferencias interregionales o interestatales, y 2) la posibilidad de que los valores más bajos de MPO A correspondan a falsos originales árabes, es decir, a textos escritos originalmente en inglés o francés por consultores privados y traducidos al árabe para, por razones políticas, ser presentados como originales en esta última lengua (lo que afirman los traductores que sospechan cuando la traducción de un texto árabe les resulta inusualmente fácil).

Para verificar la posibilidad 1) se precisaría un análisis que escapa a los intereses de esta investigación. La explicación 2), por desgracia, es imposible de verificar. Por lo tanto, nuestra asunción metodológica es que los textos de muestra fueron escritos originalmente en árabe.

e) Direccionalidad

En este apartado analizaremos el efecto de la direccionalidad. Para ello compararemos los datos y el análisis anteriores con las ratios de número de oraciones y palabras y las medias de palabras por oración en las traducciones del español y el inglés al árabe.

Número de oraciones

El número total de oraciones en los originales en español es prácticamente idéntico al de sus traducciones al árabe (2609 y 2604, respectivamente). Los valores de OE/OA oscilan entre 0,95 y 1,05, con un valor promedio de 1,00 y una desviación típica de 0,04. El grado de intervención de los traductores del español al árabe en materia de fusión o división de oraciones es, por tanto, mínimo, con una tendencia a la fusión estadísticamente no significativa.

Esas cifras en las traducciones del inglés al árabe son muy similares. Tras desechar el documento EU, que en este valor resultó ser un valor atípico según el test de Grubbs, el número total de oraciones en inglés y árabe es 2540 y 2534, respectivamente, con un promedio de OI/OA de 1,00 y una desviación típica de 0,02 (incluso menor que en las traducciones al español), con valores que oscilan entre 0,99 y 1,03. También en este caso observamos una tendencia a la fusión estadísticamente no significativa.

Si comparamos los valores medios de OE/OA y OI/OA para ambas direcciones (aproximadamente 1,5 para la traducción del árabe al español y al inglés y 1 en la dirección inversa) concluimos que la direccionalidad tiene una influencia decisiva (50 % de diferencia) sobre el fenómeno de la división de oraciones en la traducción del y al árabe. De hecho, en las traducciones al árabe se registra una ligerísima tendencia a la fusión (0,2 % en ambos casos).

La cuestión fundamental, pues, es por qué al traducir al árabe no es necesario o conveniente, en términos generales, fusionar oraciones. Para responder debemos considerar las importantes diferencias registradas entre los valores de MPO A de los originales árabes, cuya desviación típica es muy pronunciada (9,37). Esas diferencias parecen indicar que en el conjunto de los documentos redactados en árabe no supone un problema la coexistencia de oraciones muy breves y muy largas. En consecuencia, es lógico suponer que los traductores al árabe en este contexto comunicativo y registro no necesitan fusionar, y mucho menos dividir, las oraciones originales, dado que su mayor o menor extensión no afecta a la legibilidad o las características de registro.

Número de palabras

Se registran más palabras en español que en árabe tanto en las traducciones del español al árabe como en las del árabe al español. Sin embargo, la ratio PE/PA, que en la traducción del árabe al español era de 1,53 con $\sigma = 0,10$, desciende en el sentido inverso hasta 1,22 con $\sigma = 0,05$ y oscilaciones de 1,16 a 1,27.

El mismo fenómeno se registra en las traducciones del inglés al árabe, con la excepción una vez más del documento EU. Al igual que en el caso del español, la ratio PI/PA desciende sustancialmente con respecto a la de las traducciones del árabe al inglés y pasa de 1,27 a 1,07, con $\sigma = 0,07$ y oscilaciones de 0,96 y 1,18.

El documento EU es el único en el que, aun no llegando a ser valor atípico, se registró menor número de palabras en inglés que en árabe (PI/PA = 0,96). Esto, junto con su elevada ratio de OI/OA (1,20), indica que el pronunciado grado de intervención de su traductor al árabe se revela tanto en la fusión de oraciones como en el recurso a otras estrategias que resultaron en un mayor número de palabras.

En conclusión, en la traducción del español y el inglés al árabe disminuyen los valores de PE/PA y PI/PA respecto a la direccionalidad contraria, es decir, al traducir al árabe desde esas lenguas se incrementa la ratio media de palabras en mayor medida que al traducir del árabe. Estas comparaciones, debido a la aglutinación gráfica propia del árabe, solo pueden alcanzarse considerando las ratios, no las cifras absolutas de número de palabras.

Así pues, para todas las lenguas y direccionalidades consideradas se cumple como norma general que la ratio de número de palabras gráficas de una lengua A y una lengua B es mayor en las traducciones de B a A que en las traducciones de A a B. Si quedara demostrado que el fenómeno se debe al empleo de estrategias de explicitación, ello abundaría en la idea de la existencia de un universal de la traducción consistente en "*an overall tendency to spell things out rather than leave them implicit in translation*" (Baker, 1996, p. 180). En cualquier caso, demostrar que se ha recurrido a ese tipo de estrategias escapa a los objetivos de esta investigación.

Media de palabras por oración

En las traducciones del español al árabe se observa una tendencia estadísticamente significativa a la disminución de MPO A frente MPO E (30,40 frente a 37,03). Los valores, además, son bastante homogéneos en ambas lenguas (en MPO E oscilan entre 32,82 y 41,06 y en MPO A, entre 27,94 y 32,47). En la direccionalidad inversa, recordemos, no se registraba tendencia significativa alguna en este sentido.

Llama poderosamente la atención la gran diferencia en las desviaciones típicas de MPO A en las traducciones del español al árabe ($\sigma = 1,70$) y en los originales árabes ($\sigma = 9,40$). Esta importante homogeneización de los valores de MPO A en las traducciones se explica parcialmente si consideramos que los valores de MPO E eran bastante homogéneos en los originales ($\sigma = 3,09$), aunque podría haber contribuido al fenómeno la tendencia general a la normalización que se registra en todas las traducciones analizadas.

En las traducciones del inglés al árabe se observa un comportamiento análogo. Una vez más, EU resultó ser un valor atípico (MPO A = 29,22), con oraciones mucho más extensas que las otras traducciones del inglés como resultado de su mayor tendencia a la fusión de oraciones. Eliminado ese valor atípico, el promedio de MPO A, que era claramente superior al de MPO I en las traducciones del árabe al inglés (38,47 frente a 32,00), cae ahora hasta situarse por debajo del de los originales ingleses (MPO A = 23,98 frente a MPO I = 32,00). Los valores de MPO I oscilan en las traducciones al árabe entre 24,92 y 26,68, con $\sigma = 0,67$, mientras que MPO A oscila entre 23,01 y 24,27, con $\sigma = 0,50$.

Por otro lado, si en las traducciones del árabe al inglés tan solo un documento registraba un valor de MPO I mayor que el de su original árabe, en la direccionalidad inversa, una vez eliminado el valor atípico de EU, todos los documentos presentan valores de MPO I superiores a los de MPO A. La importante reducción del valor medio de MPO A (que pasa de 38,47 a 23,98) va acompañada de una acentuada disminución de la desviación típica de los valores de MPO A, que pasa de 9,40 en los originales árabes a tan solo 0,50 en las traducciones al árabe. Esto podría deberse tanto a la extraordinariamente baja desviación típica de los originales ingleses (1,20) como a una tendencia a la homogeneización de las traducciones.

La disminución de los valores de MPO A en las traducciones al español y al inglés con respecto a los de los originales árabes respaldaría la idea de que los traductores al árabe, en general, tienden más a respetar las oraciones del original dado que la legibilidad de los documentos en árabe no se vería tan influida por la longitud de las oraciones como ocurre en español e inglés. Al mantenerse un número similar de oraciones y un número de palabras menor en árabe, los valores de MPO A en las traducciones del español (30,40) y del inglés (23,98) son notablemente inferiores a los de los originales árabes (38,47).

En la direccionalidad inversa, los valores de MPO E y MPO I son más altos en las traducciones al árabe (38,82 y 32) que en los documentos originales en español e inglés (37,03 y 25,62). Una posible explicación al aumento de la MPO en español e inglés en las traducciones del árabe sería precisamente la presencia de oraciones más extensas en los originales, lo que provocaría que las de las traducciones también lo fueran.

Por otro lado, para todas las combinaciones lingüísticas consideradas e independientemente de la direccionalidad, la desviación típica de la MPO de las traducciones es menor que la de la MPO de los originales. Así, en la traducción del árabe al español y al inglés, los valores de desviación típica son 9,40 (árabe), 5,39 (español) y 4,33 (inglés), mientras que en la traducción del español al árabe pasamos de 3,09 (español) a 1,70 (árabe) y en la traducción del inglés al árabe de 0,67 (inglés) a 0,50 (árabe).

Por tanto, podemos afirmar que nuestros datos indican una tendencia a la homogeneización de las MPO en las traducciones respecto a los originales.

Conclusiones

La direccionalidad de la traducción influye decisivamente en la tendencia a la división de las oraciones, que es un 50 % más acentuada en la traducción del árabe que al árabe. Por el contrario, dos fenómenos son independientes de la direccionalidad:

i. La tendencia a la homogeneización (disminución de la desviación típica) de los valores de MPO de las traducciones con respecto a los documentos originales, y
ii. La ratio de número de palabras gráficas de una lengua A y una lengua B es mayor en las traducciones de B a A que en las traducciones de A a B.

f) Stopwords

En las traducciones del árabe al español y al inglés se contabilizaron, respectivamente, 996 *stopwords* y 1 002 divisiones de oraciones y 983 *stopwords* y 984 divisiones de oraciones. La diferencia entre el número de *stopwords* y de divisiones se explica porque en ocasiones, por el alto grado de reformulación, no fue posible relacionar la introducción de un punto con una marca lingüística concreta.

La diferencia entre el número total de oraciones en español (OE = 2 794) y en árabe (OA = 1 836) fue de 958. El número de divisiones (1 002) y la diferencia entre OE y OA no son coincidentes porque 44 puntos que aparecen en el original no fueron reproducidos por los traductores al español por dos razones:

i. En 14 casos el traductor al español ignoró un punto final del original árabe, presumiblemente por considerarlo una errata, y
ii. en 30 casos, el traductor al español optó por fusionar dos oraciones del original árabe en una sola oración en español.

Se concluye que en las traducciones del árabe al español se produjo una fusión por cada 33,4 divisiones (1002/30).

En la traducción del árabe al inglés, la diferencia entre OI y OA (952) y el número de divisiones (984) se debe a que en 8 casos se ignoró un punto del original y en 24 casos se fusionaron dos oraciones del original. Por tanto, se produjo una fusión por cada 41 divisiones (984/24).

Por otro lado, en las traducciones del árabe al español solo el 2,5 % de los 400 *types* registraron simultáneamente valores de probabilidad de activación (PA) y distribución (D) superiores al 20 %; solo 4 (حيث [*ḥaythu*], وقد [*wa-qad*], كما [*kamā*] y وذلك [*wa-dālika*]) presentaban D >50, y solo 2 (وقد [*wa-qad*] y حيث [*ḥaythu*]) registraron un valor de PA > 33. *Ḥaythu* (حيث) es particularmente interesante, dado que su elevado valor de PA (38,8 %) puede ser indicativo de un proceso de desemantización (Heine y Kuteva, 2002).

En la traducción al inglés los resultados son prácticamente idénticos: el 2,4 % de los *types* presentaron simultáneamente valores de PA y D superiores al 20 %, solo 4

registraron un valor de D > 50 (حيث [*ḥaythu*]], وقد [*wa-qad*], كما [*kamā*] y وذلك [*wa-dālika*]) y solo 2 (وقد [*wa-qad*] y حيث [*ḥaythu*]) presentaron un valor de PA > 33. El valor de PA de حيث [*ḥaythu*] es incluso más alto que en la traducción al español (47,5 %).

Entre las *stopwords* comunes, tan solo cuatro (كما [*kamā*], وقد [*wa-qad*], حيث [*ḥaythu*] y وتم [*wa-tamma*], esta última a bastante distancia con valores de D y PA muy próximos a 20) presentaron simultáneamente valores de D>20 y PA>20.

Estos datos reflejan la fuerte correlación negativa (r = -0,58) entre PA y D en los *types* con más de una ocurrencia activa (r = -0,58 en la traducción al español y r = -0,59 en la traducción al inglés).

En consecuencia, (1) ningún *type* con PA > 50 y OA > 1 presenta un valor de D > 10, y (2) los *types* con valores de PA = 100 tienen una D media extremadamente baja (1,8 y 1,6 en las traducciones al español y al inglés respectivamente, lo que significa que aparecen en una media de 2 de los 114 segmentos). Esto es consecuencia lógica de la definición de parámetros, ya que D depende del número total de ocurrencias y PA es inversamente proporcional a ese número.

También existe una elevadísima correlación negativa (r = -0,97 para las traducciones a ambas lenguas) entre los valores de D más altos de la muestra y la posición que ocupa cada *type* en la lista de frecuencia de *tokens* del corpus. Esto indica que el valor de D de un *type* es solo reflejo de su ocurrencia general.

En nuestra opinión, estos resultados son en gran medida consecuencia de la aglutinación gráfica del árabe. La aglutinación gráfica (1) aumenta el número total de *types* (palabras gráficas); en consecuencia, (2) aumenta el número de palabras con una única aparición en el corpus, incluidos los *types* con OA = 1; (3) intensifica la correlación negativa entre D y la posición de cada *type* en la lista de frecuencia de *tokens* del corpus; y (4) intensifica la correlación negativa entre PA y D.

En otras palabras, la palabra gráfica árabe tomada como unidad de análisis magnifica el papel del azar. Por lo tanto, la palabra gráfica es ineficaz para mejorar el rendimiento de

las herramientas de segmentación o alineación, ya que cuando D es estadísticamente significativo PA cae por debajo de 50.

Los resultados del análisis basado en palabras gráficas, si bien tienen poco valor desde el punto de vista computacional, como hemos podido comprobar, son muy útiles para la formación de traductores. En particular, la atención del traductor en formación debe dirigirse a tres palabras gráficas: *ḥaythu* [حيث], *wa-qad* [وقد] y *kamā* [كما], y ello porque, a diferencia de lo que ocurre con los modelos informáticos, sus capacidades humanas sí le permiten tomar decisiones no aleatorias sobre si la división de oraciones es apropiada o no, incluso con valores de PA por debajo de 50.

Por otro lado, como muestran los resultados, una parte muy importante de las *stopwords* con D>20 y AP>20 comienza con la partícula aglutinada *wa* (el 70 % y el 50 % en las traducciones al español y al inglés, respectivamente). La importancia de dicha partícula fue evidente a primera vista y mostró la necesidad de tokenizar las *stopwords*.

La tokenización de las *stopwords* mejoró enormemente la validez de nuestros hallazgos. En la traducción al español el 61,8 % de la divisiones fueron introducidas por la partícula *wa* y, de ellas, más de la mitad (57 %) por la partícula *wa* seguida de un verbo o una partícula verbal. En la traducción al inglés, la partícula *wa* introdujo el 63,4 % de las divisiones y, de ellas, en el 54 % iba seguida de verbo o partícula verbal.

El destacado papel de *wa*, en general, y de *wa* seguida de verbo, en particular, no es sorprendente. Al-Khuli (1998, p. 199), a partir de un corpus compuesto por 88 muestras de 50 palabras cada una (4 400 palabras en total), concluye que la partícula *wa* representa el 7,96 % del número total de palabras en árabe. Según Buckwalter y Parkinson (2011), además, la partícula *wa* es la segunda palabra más frecuente en árabe, tan solo superada por el artículo definido, que también se aglutina gráficamente (ال [al]).

Por otro lado, en árabe estándar, el 64,21 % de las oraciones son verbales (Al-Khuli, 1998, p. 179), siendo el número total de oraciones verbales igual al número total de verbos. Además, el 9,63 % son oraciones que incluyen *kāna wa-akhawātu-hā* (كان وأخواتها) (Ryding, 2005, p. 634-40), es decir, oraciones atributivas o asimiladas, que en nuestro recuento hemos considerado oraciones verbales. Por tanto, en nuestro recuento,

las oraciones verbales constituirían en torno al 74 %. Además, el árabe estándar, aunque con variaciones relacionadas con el género y el autor, tiende a ser un idioma VSO (Parkinson, 1981 y Abdul-Raof, 1998).

Así, coincidimos con Alazzawie cuando afirma que

> *Wa is frequently used, always sentence-initially, and its function is to textually relate sentences to each other. It is a cohesive device or a text-building device (Halliday & Hasan, 1976) widely used by writers and speakers and expected by hearers to establish ties between sentences and to indicate continuity of discourse. However, this functional item makes no semantic or syntactic predictions about the internal structure of the following sentence it is prefaced to. In fact, the sentence would be propositionally complete without wa. This does not, however, mean that it is superfluous or redundant and therefore dispersible because it does signal some aspect of meaning, such as slight topic shifting or slight refocusing of discourse* (2014, pp. 2010-2011).

Por tanto, si tenemos en cuenta la abundancia de oraciones verbales en árabe, la tendencia a las oraciones VSO, el uso tan frecuente de *wa* y su función y, al mismo tiempo, el uso de puntos en español, nuestros datos eran más que esperables. Como bien afirma Ryding,

> *sentences within an expository text […] are often initiated with wa- 'and' and/or another connective expression. […] As a sentence-starter, wa- is considered good style in Arabic, but it is not usually translated into English because English style rules normally advise against starting sentences with 'and.'"* (2005: 409).

A lo afirmado por Ryding, que nuestros datos confirman también para la traducción al español, podemos añadir que una palabra no siempre se traduce, antes bien puede generar un cambio discursivo, por ejemplo una división de oraciones.

Finalmente, según Al-Khuli (1998, p. 192), el 6,9 % de las palabras en árabe son *ḍamīr al-ghāʾib* (ضمير الغائب) (pronombre personal en tercera persona). Un pronombre precedido por la partícula *wa* puede realizar numerosas funciones, como enfatizar el sujeto de un verbo, introducir el sujeto de una oración nominal (Ryding, 2005, pp. 299-301) o simplemente aclarar información previa. Por tanto, es comprensible que *wa-huwa* (وهو = "y él") y *wa-hiya* (وهي = "y ella") sean 2 de las 10 *stopwords* con D >20 y AP >20 en las traducciones al español y al inglés.

Tampoco resulta sorprendente que *wa-dhalika* (وذلك) sea una de esas diez primeras *stopwords*. En una cláusula, el pronombre demostrativo *dhalika*, a menudo precedido por *wa*, "serves as the subject of the clause, followed by a complement or predicate. There is therefore a syntactic boundary between the demonstrative and the rest of the clause." (Ryding, 2005, p. 318).

Capítulo V: Conclusiones

En este capítulo resumiremos las principales conclusiones de nuestra investigación, las relacionaremos con nuestras hipótesis y objetivos, destacaremos sus aportaciones al estado de la cuestión, subrayaremos algunas limitaciones de nuestras conclusiones y determinaremos potenciales líneas futuras de investigación.

1. Principales resultados y aportaciones

a) Refutación de la hipótesis nula

Nuestra investigación se basó en un corpus de 171 documentos en árabe (3 541 670 palabras) y sus correspondientes traducciones al español (5 139 440 palabras) y al inglés (4 424 882 palabras), en total 513 documentos y 13 105 992 palabras, del que se extrajo una muestra representativa (3 % del corpus).

Para analizar el impacto de la direccionalidad, por otra parte, se analizaron dos corpus paralelos. El primero comprende 6 documentos originales en español y sus traducciones al árabe (12 documentos y 196 968 palabras) y el segundo, 6 documentos originales en inglés y sus traducciones al árabe (en total, 12 documentos y 184 955 palabras).

A nuestro juicio, las conclusiones de esta investigación son válidas y representativas de la traducción del árabe al español y al inglés de textos informativos producidos en el primer cuarto del siglo XXI. Debemos destacar que las traducciones del árabe al español analizadas pueden ser consideradas traducciones directas con un grado de fiabilidad notable, ya que aquellas traducciones que presentaban indicios de intermediación fueron descartadas. Sin embargo, nuestras conclusiones no deben generalizarse a otros géneros textuales en árabe estándar contemporáneo.

Los datos extraídos en nuestro análisis refutaron empíricamente la hipótesis nula y demostraron que cada oración en árabe originó 1,5 oraciones en español e inglés.

Nuestros datos confirmaron también la tendencia del árabe a formar oraciones extensas y su uso errático de los signos ortotipográficos, que en otras lenguas, inclusive el español y el inglés, constituyen las principales marcas de división de oraciones. Esto explicaría la marcada tendencia a la división de oraciones al traducir del árabe a otras lenguas, que la literatura previa había sugerido sin apoyo empírico (véase II 1. y II 2.). Nuestra investigación demuestra empíricamente esa tendencia.

Por otro lado, el análisis de los dos corpus paralelos demostró que en las traducciones al árabe disminuye ligeramente el número de oraciones respecto a los originales en español e inglés, lo que indica una tendencia poco significativa a la fusión. Esto contrasta con la fuerte tendencia a la división de oraciones en las traducciones del árabe y demuestra empíricamente el fuerte impacto de la direccionalidad en las combinaciones lingüísticas analizadas, incluso en un contexto tan institucionalizado como la ONU.

Por otro lado, las ratios de número de palabras se incrementan en las traducciones respecto a los originales en todas las combinaciones y direccionalidades consideradas, con valores muy bajos de desviación típica que indican diferencias interdocumentales mínimas. Se trata, por tanto, de un fenómeno universal y con un respaldo estadístico muy sólido.

Los valores de desviación típica de la media de palabras por oración en los originales en árabe, por su parte, sugieren que la alternancia de oraciones muy dispares en extensión no perjudica la legibilidad en árabe. El fenómeno se explica por el uso errático de los signos de puntuación en esa lengua, previamente señalado por la literatura sin base empírica (Touir et al., 2008; Keskes, 2015; Sánchez-Ratia, 2002). Esta mayor flexibilidad de la lengua árabe explicaría el escaso recurso a la fusión de oraciones al traducir del español o el inglés al árabe.

Nuestros resultados contrastan marcadamente con las tesis de quienes consideran la división de oraciones un reflejo de universales de traducción y minimizan el impacto de la direccionalidad, y están más en consonancia con quienes consideran las diferencias interlingüísticas un factor fundamental, lo que permite prever un impacto de la direccionalidad (véase II 1.). En nuestras combinaciones lingüísticas, el impacto de la direccionalidad es muy acusado.

Por otro lado, las diferencias intertextuales y los valores atípicos hallados en nuestros datos avalan la tesis de Hareide y Hofland (2012) de que las preferencias individuales de los traductores también pueden influir en la división de oraciones, aunque las tendencias estadísticas demuestran que, en un contexto marcadamente institucional como la ONU, esas preferencias desempeñan un papel muy limitado.

Así pues, en nuestro caso, la mayor o menor tendencia a la división de oraciones responde, sobre todo, a la direccionalidad. La necesidad de estandarización, tan acusada en la traducción institucional, tiene en la ONU un mayor impacto en la traducción del árabe que en la traducción al árabe, ya que los originales árabes son mucho más heterogéneos en este aspecto que los originales en español o inglés.

Por otro lado, nuestros datos parecen confirmar que, con independencia de las matizaciones que conlleva la direccionalidad, el contexto comunicativo y las preferencias personales, existen tendencias comunes a las traducciones analizadas. Esas tendencias, que definen los textos como traducciones frente a los originales, se nos antojan reflejos "matizados" de universales de traducción. Fundamentalmente hemos encontrado dos tendencias comunes a todas las traducciones analizadas y que las diferencian de los originales.

En primer lugar, los valores de desviación típica de las medias de palabra por oración son siempre más elevados en los originales que en las traducciones. Esto sugiere un proceso de estandarización que, aunque muy propio de la traducción institucional, también podemos interpretar como manifestación de un afán de "normalización", entendido como "*the tendency to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them*" (Baker, 1996, p. 176).

En segundo lugar, el número de palabras es siempre mayor en las traducciones que en los documentos originales, lo que podría considerarse una manifestación de fenómenos de explicitación, entendida esta como "*an overall tendency to spell things out rather than leave them implicit in translation*" (Baker, 1996, p. 180).

A nuestro juicio, todos estos hallazgos representan una contribución significativa a la Lingüística Contrastiva, la Lingüística del Discurso y la Traductología.

b) Determinación de las stopwords

Refutada la hipótesis nula, procedimos a determinar las *stopwords* que originaron las divisiones.

Nuestros datos han demostrado, en primer lugar, que la aglutinación gráfica genera en árabe un notable aumento del número total de *types* y, en consecuencia, del número de *types* divisores con NAO = 1. Dicho de otro modo, la palabra gráfica tomada como unidad no permitiría mejorar el rendimiento de las herramientas automáticas de segmentación o alineación en árabe ya que, cuando D es estadísticamente significativa, PA cae por debajo de 50. Nuestra investigación, por tanto, ha aportado pruebas empíricas de que la tokenización es imprescindible para cualquier operación computacional relacionada con la lengua árabe, en línea con las tesis de Choueka et al. (2000) y Salameh et al. (2011).

Los datos extraídos del análisis basado en las palabras gráficas árabes, aun con esas limitaciones computacionales, son muy valiosos para la formación de traductores. Téngase en cuenta que el segmentador automático, si PA desciende de 50 o se mantiene en ese umbral, activaría o no la división al azar. Los traductores humanos, sin embargo, pueden discernir cuándo deben o no dividir una oración una vez que han cobrado conciencia de cuáles son las marcas lingüísticas que podrían conducir a ello.

Nuestros datos demuestran que se debe llamar la atención de los traductores en formación sobre tres palabras gráficas que constituyen marcas claras de una potencial división de oraciones: *ḥaythu* [حيث], *wa-qad* [وقد] y *kamā* [كما]. Esto es igualmente aplicable a la docencia de la traducción del árabe al inglés y del árabe al español. De esas tres marcas, *ḥaythu* [حيث] es particularmente interesante, dado que su alto PA es indicativo de que ha culminado un proceso de desemantización (Heine y Kuteva, 2002).

Así pues, ante la indeterminación de los datos desde una perspectiva computacional, procedimos a tokenizarlos. El resultado fue que más de 6 de cada 10 divisiones habían sido originadas por la partícula *wa*. De ellas, en torno a la mitad lo fueron por esa partícula seguida de un verbo o una partícula verbal, incluido *wa-qad* [وقد], y el resto por la partícula *wa* seguida de cualquier otra categoría de palabra. En este último grupo destacan *wa-huwa* [وهو], *wa-hiya* [وهي] y *wa-dhalika* [وذلك].

Estos hallazgos no sorprenden si consideramos la frecuencia de estas palabras en árabe y sus funciones sintácticas y discursivas, así como la función discursiva del punto en español e inglés (véase IV 2.f)). Se trata, pues, de datos con sostén empírico y que son perfectamente lógicos desde un punto de vista lingüístico. Estos datos pueden mejorar la eficiencia de cualquier segmentador automático que preprocese los textos mediante tokenización.

Además, una vez tokenizados, los datos cobran aún mayor valor para la docencia de la traducción del árabe al español y al inglés. Se trataría, en particular, de diseñar ejercicios de reformulación discursiva mediante la introducción de puntos en los que, principalmente, se buscarían las siguientes marcas potenciales de división:

- *ḥaythu* [حيث],
- *kamā* [كما]
- conjunción *wa* más verbo o partícula verbal (en particular, *wa-qad* [وقد]), y
- conjunción *wa* más pronombre personal de tercera persona o pronombre demostrativo *dhalika* [وذلك].

A nuestro juicio, estos datos también suponen una contribución significativa a la Lingüística Computacional árabe y a la Traductología.

2. Líneas futuras de investigación

Entre las líneas futuras de investigación conviene señalar las siguientes:

i. Testar la validez de nuestras conclusiones en otros géneros textuales en árabe estándar contemporáneo y en otros contextos de traducción;
ii. Incorporar a la investigación la traducción del árabe al francés y viceversa de los mismos textos para encontrar diferencias interlingüísticas y determinar si se registran las mismas tendencias;
iii. Diseñar actividades para la formación de traductores basadas en nuestros resultados y que comprendan ejemplos auténticos de división de oraciones seleccionados de nuestra muestra;

iv. Forjar relaciones de colaboración interdisciplinares que permitan aplicar nuestros hallazgos a las herramientas de segmentación en árabe ya existentes o diseñar otras mejoradas.

Bibliografía

Abdelali, A., Darwish, K., Durrani, N., & Mubarak, H. (2016). Farasa: A Fast and Furious Segmenter for Arabic. En J. DeNero, M. Finlayson, & S. Reddy (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 11-16). Association for Computational Linguistics. Recuperado de http://alt.qcri.org/~ndurrani/pubs/farasa-fast-furious.pdf

Abdul-Raof, H. (1998). *Subject, Theme and Agent in Modern Standard Arabic*, Surrey: Curzon Press.

Ahrenberg, L. (2017). Comparing Machine Translation and Human Translation: A Case Study. En I. Temnikova, C. Orasan, G. Corpas & S. Vogel (Eds.), *Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT)* (pp. 21.28). Shoumen, Bulgaria: Association for Computational Linguistics. Recuperado de https://www.acl-bg.org/proceedings/2017/RANLP_W3%202017/pdf/HiT-IT003.pdf

Alazzawie, A. (2014). The Discourse Marker *wa* in Standard Arabic—A Syntactic and Semantic Analysis. *Theory and Practice in Language Studies*, 4(10), 2008-2015.

Alfuraih, R. (2020). The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics. *Language Resources & Evaluation*, 54, 801-830.

Alghamdi, M., & Teahan, W. (2017). Experimental evaluation of Arabic OCR systems. *PSU Research Review*, 1(3), 229-241.

Al-Harthi, M., & Alsaif, A. (2019). The Design of the SauLTC application for the English-Arabic Learner Translation Corpus. En M. El-Haj, P. Rayson, E. Atwell & L. Alsudias (Eds.), *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics* (pp.

80-88). Association for Computational Linguistics. Recuperado de https://www.aclweb.org/anthology/W19-5610.pdf

Al-Khuli, M. (1998). *Al-tārakīb al-shāʾiʿa fi l-lugha al-ʿarabiyya. Dirāsa iḥṣāʾiyya* [*Las estructuras más comunes en lengua árabe. Un estudio estadístico*]. Ammán: Dār Al-Falāḥ.

Alotaiby, F., Foda S., & Alkharashi I. (2010). Clitics in Arabic Language: A Statistical Study. *Proceedings of Pacific Asia Conference on Language, Information and Computation* (PACLIC), 24, 595-602.

Al-Raisi, F., Lin, W., & Bourai, A. (2018). A Monolingual Parallel Corpus of Arabic. *Procedia Computer Science*, 142, 334-338.

Al-Sanie, W., Touir, A., & Mathkour, H. (2005a). Towards a suitable representation of Arabic text summarization. In *Proceedings of the 7th International Conference on Information Integration and Web-based Applications and Services, septiembre 2005* (pp. 535-542). Kuala Lumpur, Malasia.

Al-Sanie, W., Touir, A., & Mathkour, H. (2005b). Towards a rhetorical parsing of Arabic text. In *Proceedings of the International Conference on Intelligent Agents, Web Technology and Internet Commerce, noviembre 2005* (pp. 1086-1091). Viena, Austria.

Altammami, S., Atwell, E., & Alsalka A. (2019). Text Segmentation Using N-grams to Annotate Hadith Corpus. En M. El-Haj, P. Rayson, E. Atwell & L. Alsudias (Eds.), *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics* (pp. 31-39). Association for Computational Linguistics. Recuperado de https://www.aclweb.org/anthology/W19-5605.pdf

Arias, J. P., & Feria, M. (2012). *Los traductores de árabe del Estado español: del Protectorado a nuestros días*. Barcelona: Bellaterra.

Avión Martínez, S. (2013). *El trabajo de traductor en la Naciones Unidas*. Conferencia. 18/03/2013. UVigoTV. Recuperado de http://tv.uvigo.es/gl/video/mm/17945.html

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. En M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology*: *In Honour of John Sinclair* (pp. 233-250). Amsterdam/Philadelphia: John Benjamins.

Baker, M. (1996). Corpus-based Translation Studies: the challenges that lie ahead. En H. Sommers (Ed.), *Terminology, LSP and Translation Studies in Language engineering in honour of J. C. Sager* (pp. 175-187). Amsterdam/Philadelphia: John Benjamins.

Bisiada, M. (2013). *From hypotaxis to parataxis: An investigation of English–German syntactic convergence in translation*. Tesis doctoral. University of Manchester. Recuperado de https://www.research.manchester.ac.uk/portal/files/54546816/FULL_TEXT.PDF

Bisiada, M. (2016). *Lösen Sie Schachtelsätze möglichst auf*: The Impact of Editorial Guidelines on Sentence Splitting in German Business Article Translations. *Applied Linguistics* 37(3), 354–376.

Bloch, I. (2005). Sentence Splitting as an Expression of Translationese: Seminar Paper. In *Black Box Seminar*, Bar Ilan University. Recuperado de https://www.biu.ac.il/hu/stud-pub/tr/tr-pub/bloch-split.htm

Buckwalter, T., & Parkinson, D. (2011). *A Frequency Dictionary of Arabic: Core Vocabulary for Learners*. Londres/Nueva York: Routledge.

Cao D. & Zhao X. (2008). Translation at the United Nations as Specialized Translation. *The Journal of Specialised Translation*, 9, 39-54.

Chen, Y., & Eisele, A. (2012). MultiUN v2: UN documents with multilingual alignments. En N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidiset (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2500-2504). Estambul: European Language Resources Association (ELRA). Recuperado de http://www.lrec-conf.org/proceedings/lrec2012/pdf/641_Paper.pdf

Choueka, Y., Conley E., & Dagan I. (2000). A comprehensive bilingual word alignment system. Application to disparate languages: Hebrew and English. En J. Véronis (Ed.), *Parallel Text Processing. Alignment and Use of Translation Corpora* (pp. 69-96). Dordrecht/Boston/Londres: Kluwer Academic Publishers.

Darwish, K., & Gao W. (2014). Simple Effective Microblog Named Entity Recognition: Arabic as an Example. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidiset (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*

(pp. 2513-2517). Islandia: European Languages Resources Association (ELRA). Recuperado de http://www.lrec-conf.org/proceedings/lrec2014/pdf/186_Paper.pdf

Dickins, J., Sándor H., & Higgins, I. (2017). *Thinking Arabic Translation. A course in translation method: Arabic to English.* Londres/Nueva York: Routledge.

Eisele, A., & Chen, Y. (2010). MultiUnited Nations: A Multilingual Corpus from United Nation Documents. En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation* (pp. 2868-2872). European Language Resources Association (ELRA). Recuperado de http://www.lrec-conf.org/proceedings/lrec2010/pdf/686_Paper.pdf

Ethnologue (2021). En G. F. Simons & C. D. Fennig (eds.) 'Ethnologue: Languages of the World, 24th edition'.

Fabricius-Hansen, C. (1999). Information packaging and translation: aspects of translational sentence splitting (German-English/Norwegian). En M. Doherty (Ed.), *Sprachspezifissche Aspekte der Informationsverteilung* (pp. 175-214). Berlín: Akademie Verlag.

Farghaly, A., & Shaalan, K. (2009). Arabic Natural Language Processing: Challenges and solutions. *ACM TraSActions on Asian Language Information Processing (TALIP)*, 8(4), 1–22.

Feria, M. (2014). Planning the Acquisition and Enhancement of Language Skills for Translation and Interpreting Trainees: The Case of Arabic. En V. Aguilar, W. Saleh, M. A. Manzano, L. M. Pérez Cañada, & P. Santillán Grimm (Eds.), *Arabele 2012: enseñanza y aprendizaje de la lengua árabe* (pp. 197-221). Murcia: Universidad.

Frankenberg-Garcia, A. (2019). A corpus study of splitting and joining sentences in translation. *Corpora*, 14(1), 1-30.

Gale, W., & Kenneth C. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics*, 19(1), 75-102.

García Barrero, D., Feria García M., & Turell, M. (2012). Using function words and punctuation marks in Arabic forensic authorship attribution. En R. Sousa-Silva, R.

Faria, N. Gavaldà & B. Maia (Eds.), *Proceedings of the 3rd European Conference of the International Association of Forensic Linguists* (pp. 42-56). Oporto: Universidade.

Ghaly, H. (2014). *Canvas: A fast and accurate geometric sentence alignment system using lexical cues within complex misalignment settings.* Nueva York: CUNY Academic Works.

Guellil, I., Saâdane, H., Azouaou, F., Gueni, B., & Nouvel, D. (s. f.). Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*. https://doi.org/10.1016/j.jksuci.2019.02.006

Habash, N. (2010). Introduction to Arabic Natural Language Processing. *Synthesis Lectures on Human Language Technologies*, 3(1), 1-187.

Hareide, L., & Hofland, K. (2012). Compiling a Norwegian-Spanish parallel corpus. Methods and challenges. En M. Oakes & J. Meng (Eds.), *Quantitative Methods in Corpus-Based Translation Studies* (pp. 75-114). Amsterdam/Philadelphia: John Benjamins.

Heine, B., & Kuteva, T. (2002). *World Lexicon of Grammaticalization*. Nueva York: Cambridge University Press.

Ilhami, N. (2015). *La formación de traductores e intérpretes árabe-español: adecuación del diseño curricular en la Universidad de Granada*. Tesis doctoral. Granada: Universidad de Granada. Tesis Doctorales.

Keskes, I. (2015). *Discourse Analysis of Arabic Documents and Application to Automatic Summarization* Tesis doctoral en cotutela. Université de Sfax/Université Toulouse III. Recuperado de https://core.ac.uk/download/pdf/42969051.pdf

Khalifa, S., Zalmout, N., & Habash, N. (2016). YAMAMA: Yet another multi-dialect Arabic morphological analyzer. *Proceedings of COLING 2016 - 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 223-227). https://www.aclweb.org/anthology/C16-2047.pdf

Kunilovskaya, M., & Morgoun, N. (2013). Gains and pitfalls of sentence-splitting in translation. *Perm National Research Polytechnic University Herald. Issues in Linguistics and Pedagogy*, 8(50), 152–166.

Lafeber, A. (2012). *Translation at inter-governmental organizations: the set of skills and knowledge required and the implications for recruitment testing.* Tesis doctoral. Universitat Rovira i Virgili.

Mathkour, H., Touir, A., & Al-Sanie, W. (2005). Automatic information classifier using rhetorical structure theory. In *Proceedings of the Internation Conference on Intelligent Information Processing and Web Mining* (pp. 229-236). Gdansk, Polonia.

Merkel, M. (2001). Comparing source and target texts in a translation corpus. En A. S. Hein (Ed.), *Proceedings of the 13th Nordic Conference of Computational Linguistics, NODALIDA* (pp. 81-85). Association for Computational Linguistics. Recuperado de https://www.aclweb.org/anthology/W01-1716.pdf

Monroe, W., Green, S., & Manning, C. D. (2014). Word segmentation of informal Arabic with domain adaptation. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, 2,* (pp. 206-211). https://doi.org/10.3115/v1/p14-2034

Musacchio, M. T. (2005). Chapter 5. The Influence of English on Italian: The Case of Translations of Economics Articles. En G. Anderman & M. Rogers (Eds.), *In and Out of English: For Better, For Worse. Blue Ridge Summit* (pp. 71-96). Bristol: Multilingual Maters. https://doi.org/10.21832/9781853597893-008

Nadvornikova, O. (2017). Parallel Corpus in Translation Studies: Analysis of Shifts in the Segmentation of Sentences in the Czech-English-French Part of the InterCorp Parallel Corpus. En J. Emonds & M. Janebová (Eds.), *Language Use and Linguistic Structure. Proceedings of the Olomouc Linguistics Colloquium 2016* (pp. 445-460). Olomouc: Palacký University Olomouc.

Neme, A., & Paumier, S. (2020). Restoring Arabic vowels through omission-tolerant dictionary lookup. *Language Resources and Evaluation*, 54, 487-551.

Parkinson, D. (1981). VSO to SVO in Modern Standard Arabic: A study in diglossia syntax. *Al-Arabiyya*, 14, 24-37.

Pasha, A., Al-Badrashiny M., Diab, M., & El Kholy, A., Eskander, R., Habash, N., Pooleery, M, Rambow, O., & Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. En N. Calzolari, K.

Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *LREC 2014, Ninth International Conference on Language Resources and Evaluation* (pp. 1094-1101). European Language Resources Association. Recuperado de http://www.lrec-conf.org/proceedings/lrec2014/pdf/593_Paper.pdf

Pouliquen, B., Mazenc, C., Elizalde, C., & Garcia-Verdugo, J. (2012). Statistical Machine Translation prototype using UN parallel documents. En *Proceedings of the 16th Annual conference of the European Association for Machine Translation*, (pp. 12-19). Trento, Italia: European Association for Machine Translation.

Pouliquen, B., Elizalde, C., Junczys-Dowmunt, M., Mazenc, C., & Garcia-Verdugo, J. (2013). Large-scale multiple language translation accelerator at the United Nations. En *Proceedings of the 14th Machine Translation Summit* (pp.345-352) Niza.

Ramm, W. (2004). Sentence-boundary adjustment in Norwegian-German and German-Norwegian translations: first results of a corpus-based study. En K. Aijmer & H. Hasselgard (Eds.), *Translation and Corpora* (pp. 129-147)). Gothenburg: Acta Universitatis Gothoburgensis.

Rafalovitch, A., & Dale, R. (2009). United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. En *Proceedings of the MT Summit XII* (pp. 292-299). International Association of Machine Translation. Recuperado de http://www.mt-archive.info/MTS-2009-Rafalovitch.pdf

Ryding, K. (2005). *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Sainz-Quinn, C. & Feria-García, M. (2020). Translating Arabic named entities into English and Spanish: translation consistency at the United Nations. En S. Hanna, H. El-Farahaty & A. W. Khalifa (Eds.), Routledge Handbook of Arabic Translation (pp. 381-396). Manchester: Routledge.

Sainz-Quinn, C. (2022). *Traducción al inglés y al español de textos árabes en las Naciones Unidas: homogeneidad de entidades nombradas.* Tesis doctoral. Universidad de Granada.

Salameh, M., Zantout R., & Mansour N. (2011). Improving the Accuracy of English-Arabic Statistical Sentence Alignment. *The International Arab Journal of Information Technology*, 8(2), 171-177.

Samy, D., Moreno-Sandoval A., & Guirao, J. M. (2004). An Alignment Experiment of a Spanish-Arabic Parallel Corpus. En *Proceedings of the International Conference on Arabic Language Resources and Tools* (pp. 85-89). El Cairo: NEMLAR 2004. Recuperado de http://elvira.lllf.uam.es/ESP/Publicaciones/AlignmentPaper04.pdf

Samy, D. (2005). Named Entities: Structure and Translation. A study based on a Parallel Corpus (Arabic-English-Spanish). En *Proceedings from the Corpus Linguistics Conference Series.* Birmingham. Recuperado de http://www.lllf.uam.es/ESP/Publicaciones/NamedEntitiesParallelCorpus.pdf

Samy, D., Moreno-Sandoval, A., Guirao J. M., & Alfonseca, E. (2006). Building a Parallel Multilingual Corpus (Arabic-Spanish-English). En N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (Eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC'06).* GeNAO, Italia. Recuperado de http://www.lllf.uam.es/~doaa/Publications/SamyMultilingualLREC06.pdf

Samy, D., & González Ledesma, A. (2008). Pragmatic Annotation of Discourse Markers in a Multilingual Parallel Corpus (Arabic-Spanish-English). En N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008* (pp. 3299-3305). Recuperado de http://www.lrec-conf.org/proceedings/lrec2008/

Sánchez-Ratia, J. (2002). El árabe en la traducción al español de las Naciones Unidas: algunas consideraciones y directrices para su tratamiento coherente y homogéneo. *Anexo de la versión 2002 del Manual del Traductor del Servicio de Traducción al Español de la ONU.* Recuperado de

https://conf-dts1.unog.ch/1%20SPA/Tradutek/Varios/00-Tratamiento_arabe_Anexo_ManualTraductor.htm

Scott, M. (2008). *WordSmith Tools 5.0*. Liverpool: Lexical Analysis Software.

Semmar, N., & Fluhr, C. (2007). Arabic to French Sentence Alignment: Exploration of A Cross-language Information Retrieval Approach'. In V. Cavalli-Sforza & I. Zitouni (Eds.), *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources* (pp. 73-80). Recuperado de https://www.aclweb.org/anthology/W07-0810.pdf

Serbina, T. (2014). Sentence splitting in the translation pair English-German. En *4th Using Corpora in Contrastive and Translation Studies Conference. Abstrac Book* (pp. 61-62). Lancaster University. Recuperado de http://ucrel.lancs.ac.uk/uccts4/doc/UCCTS4-abstract-book.pdf

Shaalan, K. (2014). A Survey of Arabic Named Entity Recognition and Classification. *Computational Linguistics*, 40(2), 469-510.

Shahrour, A., Khalifa, S., Taji, D., & Habash, N. (2016). CamelParser: A system for Arabic syntactic analysis and morphological disambiguation. *Proceedings of COLING 2016 - 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 228-232). Recuperado de https://www.aclweb.org/anthology/C16-2048.pdf

Solfjeld, K. (2008). Sentence splitting and discourse structure in translations. *Languages in Contrast*, 8(1), 21-46.

Tafalla Plana, M. (2010). El multilingüismo en la Organización de las Naciones Unidas. *Revista de Llengua i Dret*, 53, 137-162.

Taji, D., El Gizuli J., & Habash, N. (2018). An Arabic dependency treebank in the travel domain. En N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA). Recuperado de http://lrec-conf.org/workshops/lrec2018/W30/pdf/14_W30.pdf

Touir, A., Mathkour, H. & Al-Sanea, W. (2008). Semantic-Based Segmentation of Arabic Texts. *Information Technology Journal*, 7, 1009-1015.

Turell, M.T. (2010). The use of textual, grammatical and sociolinguistic evidence in forensic text comparison. *The International Journal of Speech, Language and the Law,* 17(2), 211-250. Equinox Publishing.

Xu, J., Fraser, A. & Weischedel, R. (2001). TREC 2001 Cross-lingual Retrieval at BBN. En *NIST TREC 2001 Proceedings* (pp. 68-77). Recuperado de https://trec.nist.gov/pubs/trec10/papers/BBNTREC2001.pdf

Zantout, R., & Guessoum, A. (2015). Obstacles Facing Arabic Machine Translation: Building a Neural Network-based Transfer Module. En Sattar Izwaini (Ed.), *Papers in Translation Studies* (pp. 229-251). Cambridge Scholars Publishing.

Anexo I

Composición del Corpus 1 con datos desglosados por signatura, Estado, número de palabras y lengua original.

SIGNATURA	ESTADO	AÑO	PALABRAS	LENGUA
CRC/C/61/Add.2	ARABIA SAUDITA	2000	23823	ÁRABE
CAT/C/42/Add.2	ARABIA SAUDITA	2001	4355	ÁRABE
CERD/C/370/Add.1	ARABIA SAUDITA	2001	3445	ÁRABE
CERD/C/370/Add.1/Corr.1	ARABIA SAUDITA	2002	47	ÁRABE
CRC/C/136/Add.1	ARABIA SAUDITA	2005	21543	ÁRABE
CEDAW/C/SAU/2	ARABIA SAUDITA	2007	17289	ÁRABE
A/HRC/WG.6/17/SAU/1	ARABIA SAUDITA	2013	11367	ÁRABE
CAT/C/SAU/2	ARABIA SAUDITA	2015	15210	ÁRABE
CRC/C/SAU/3-4	ARABIA SAUDITA	2015	42774	ÁRABE
CRPD/C/SAU/1	ARABIA SAUDITA	2015	26367	ÁRABE
CAT/C/SAU/Q/2/Add.2	ARABIA SAUDITA	2016	11036	ÁRABE
CEDAW/C/SAU/3-4	ARABIA SAUDITA	2016	36904	ÁRABE
CERD/C/SAU/4-9	ARABIA SAUDITA	2016	20987	ÁRABE
CRC/C/SAU/Q/3-4/Add.1	ARABIA SAUDITA	2016	8592	ÁRABE
CEDAW/C/SAU/Q/3-4/Add.1	ARABIA SAUDITA	2017	10345	ÁRABE
CRC/C/OPAC/SAU/1	ARABIA SAUDITA	2017	11181	ÁRABE
CRC/C/OPSC/SAU/1	ARABIA SAUDITA	2017	15495	ÁRABE
A/HRC/WG.6/31/SAU/1	ARABIA SAUDITA	2018	12372	ÁRABE
CRC/C/OPAC/SAU/Q/1/Add.1	ARABIA SAUDITA	2018	2808	ÁRABE
CRC/C/OPSC/SAU/Q/1/Add.1	ARABIA SAUDITA	2018	4524	ÁRABE
A/HRC/WG.6/4/SAU/1	ARABIA SAUDITA	2008	8505	ÁRABE-INGLÉS
A/HRC/WG.6/4/SAU/1	ARABIA SAUDITA	2008	11445	INGLÉS-ÁRABE
CERD/C/362/Add.6	ARGELIA	2000	13269	FRANCÉS
E/1990/6/Add.26	ARGELIA	2000	17625	FRANCÉS
CEDAW/C/DZA/2	ARGELIA	2003	33657	FRANCÉS

CRC/C/93/Add.7	ARGELIA	2005	27993	FRANCÉS
CAT/C/DZA/3	ARGELIA	2006	13275	FRANCÉS
CCPR/C/DZA/3	ARGELIA	2006	29110	FRANCÉS
CCPR/C/DZA/CO/3/Add.1	ARGELIA	2007	1296	FRANCÉS
CCPR/C/DZA/Q/3/Add.1	ARGELIA	2007	5006	FRANCÉS
CCPR/C/DZA/Q/3/Add.1/Corr.1	ARGELIA	2007	115	FRANCÉS
A/HRC/WG.6/1/DZA/1	ARGELIA	2008	10184	FRANCÉS
CAT/C/DZA/Q/3/Add.1	ARGELIA	2008	5269	FRANCÉS
CMW/C/DZA/1	ARGELIA	2008	21180	FRANCÉS
CAT/C/DZA/CO/3/Add.1	ARGELIA	2009	3227	FRANCÉS
E/C.12/DZA/4	ARGELIA	2009	32867	FRANCÉS
CEDAW/C/DZA/3-4	ARGELIA	2010	40324	FRANCÉS
CMW/C/DZA/Q/1/Add.1	ARGELIA	2010	8278	FRANCÉS
E/C.12/DZA/Q/4/ADD.1	ARGELIA	2010	16074	FRANCÉS
CRC/C/DZA/3-4	ARGELIA	2011	43798	FRANCÉS
CRC/C/DZA/Q/3-4/ADD.1	ARGELIA	2012	9256	FRANCÉS
A/HRC/WG.6/13/DZA/1	ARGELIA	2012	10306	FRANCÉS
CEDAW/C/DZA/Q/3-4/Add.1	ARGELIA	2012	12217	FRANCÉS
CERD/C/DZA/15-19	ARGELIA	2012	15378	FRANCÉS
CEDAW/C/DZA/CO/3-4/Add.1	ARGELIA	2015	3765	FRANCÉS
CRPD/C/DZA/1	ARGELIA	2015	26235	FRANCÉS
CERD/C/DZA/20-21	ARGELIA	2016	19683	FRANCÉS
CMW/C/DZA/2	ARGELIA	2016	21377	FRANCÉS
A/HRC/WG.6/27/DZA/1	ARGELIA	2017	10511	FRANCÉS
CCPR/C/DZA/4	ARGELIA	2017	21542	FRANCÉS
CRC/C/OPAC/DZA/1	ARGELIA	2017	3539	FRANCÉS
CCPR/C/DZA/Q/4/Add.1	ARGELIA	2018	10767	FRANCÉS
CMW/C/DZA/Q/2/Add.1	ARGELIA	2018	7167	FRANCÉS
CRC/C/OPAC/DZA/Q/1/Add.1	ARGELIA	2018	2832	FRANCÉS
CRPD/C/DZA/Q/1/Add.1	ARGELIA	2018	9111	FRANCÉS
CRC/C/11/Add.24	BAHREIN	2001	25536	ÁRABE
CAT/C/47/Add.4	BAHREIN	2004	11091	ÁRABE
CERD/C/443/Add.1	BAHREIN	2004	17744	ÁRABE
CAT/C/BHR/CO/1/Add.1	BAHREIN	2007	1061	ÁRABE
CEDAW/C/BHR/2	BAHREIN	2007	45634	ÁRABE
CERD/C/BHR/CO/7/Add.1	BAHREIN	2007	1598	ÁRABE
A/HRC/WG.6/1/BHR/1	BAHREIN	2008	9019	ÁRABE
CEDAW/C/BHR/2/ADD.1	BAHREIN	2008	21010	ÁRABE
CEDAW/C/BHR/Q/2/Add.1	BAHREIN	2008	6129	ÁRABE
CAT/C/BHR/CO/1/Add.2	BAHREIN	2010	284	ÁRABE
CEDAW/C/BHR/CO/2/ADD.1	BAHREIN	2010	1029	ÁRABE
CRC/C/BHR/2-3	BAHREIN	2010	40098	ÁRABE
CEDAW/C/BHR/3	BAHREIN	2011	19102	ÁRABE
CRC/C/BHR/Q/2-3/ADD.1	BAHREIN	2011	5303	ÁRABE
A/HRC/WG.6/13/BHR/1	BAHREIN	2012	9364	ÁRABE

CEDAW/C/BHR/Q/3/Add.1	BAHREIN	2013	10143	ÁRABE
CAT/C/BHR/2	BAHREIN	2015	21464	ÁRABE
CAT/C/BHR/3	BAHREIN	2016	21073	ÁRABE
CEDAW/C/BHR/CO/3/Add.1	BAHREIN	2016	1281	ÁRABE
A/HRC/WG.6/27/BHR/1	BAHREIN	2017	10648	ÁRABE
CCPR/C/BHR/1	BAHREIN	2017	26019	ÁRABE
CAT/C/BHR/CO/2-3/Add.1	BAHREIN	2018	3751	ÁRABE
CCPR/C/BHR/Q/1/Add.1	BAHREIN	2018	10754	ÁRABE
CEDAW/C/BHR/4	BAHREIN	2018	17840	ÁRABE
CRC/C/BHR/4-6	BAHREIN	2018	21575	ÁRABE
CRC/C/BHR/Q/4-6/Add.1	BAHREIN	2018	10725	ÁRABE
CEDAW/C/COM/1-4	COMORAS	2011	20402	FRANCÉS
CEDAW/C/COM/Q/4/Add.1	COMORAS	2011	8072	FRANCÉS
A/HRC/WG.6/18/COM/1	COMORAS	2013	10344	FRANCÉS
A/HRC/WG.6/32/COM/1	COMORAS	2018	7888	FRANCÉS
CRC/C/DJI/2	DJIBOUTI	2007	22375	FRANCÉS
A/HRC/WG.6/4/DJI/1	DJIBOUTI	2008	8497	FRANCÉS
CRC/C/DJI/Q/2/Add.1	DJIBOUTI	2008	2998	FRANCÉS
CEDAW/C/DJI/1-3	DJIBOUTI	2010	40201	FRANCÉS
CAT/C/DJI/1	DJIBOUTI	2011	20438	FRANCÉS
CCPR/C/DJI/1	DJIBOUTI	2012	17480	FRANCÉS
E/C.12/DJI/1-2	DJIBOUTI	2012	13961	FRANCÉS
A/HRC/WG.6/16/DJI/1	DJIBOUTI	2013	8097	FRANCÉS
CCPR/C/DJI/Q/1/Add.1	DJIBOUTI	2013	6100	FRANCÉS
E/C.12/DJI/Q/1-2/Add.1	DJIBOUTI	2013	8589	FRANCÉS
CCPR/C/DJI/CO/1/Add.1	DJIBOUTI	2015	5662	FRANCÉS
CERD/C/DJI/1-2	DJIBOUTI	2016	8255	FRANCÉS
CRPD/C/DJI/1	DJIBOUTI	2017	15726	FRANCÉS
A/HRC/WG.6/30/DJI/1	DJIBOUTI	2018	9708	FRANCÉS
CRC/C/78/Add.2	EAU	2001	20954	ÁRABE
A/HRC/WG.6/3/ARE/1	EAU	2008	10018	ÁRABE
CEDAW/C/ARE/1	EAU	2008	19158	ÁRABE
CEDAW/C/ARE/Q/1/Add.1	EAU	2009	13367	ÁRABE
CERD/C/ARE/12-17	EAU	2009	12019	ÁRABE
CERD/C/ARE/Q/12-17/ADD.1	EAU	2009	7970	ÁRABE
A/HRC/WG.6/15/ARE/1	EAU	2012	10508	ÁRABE
CRPD/C/ARE/1	EAU	2014	20375	ÁRABE
CEDAW/C/ARE/2-3	EAU	2014	10502	ÁRABE
CRC/C/ARE/2	EAU	2014	23069	ÁRABE
CEDAW/C/ARE/Q/2-3/Add.1	EAU	2015	7927	ÁRABE
CRC/C/ARE/Q/2/Add.1	EAU	2015	9391	ÁRABE
CERD/C/ARE/18-21	EAU	2016	17077	ÁRABE
CRPD/C/ARE/Q/1/Add.1	EAU	2016	6744	ÁRABE
A/HRC/WG.6/29/ARE/1	EAU	2017	10435	ÁRABE
CAT/C/ARE/1	EAU	2018	17381	ÁRABE

CEDAW/C/EGY/4-5	EGIPTO	2000	22162	ÁRABE
CAT/C/55/Add.6	EGIPTO	2001	10596	ÁRABE
CERD/C/384/Add.3	EGIPTO	2001	27609	ÁRABE
CCPR/C/EGY/2001/3	EGIPTO	2002	44008	ÁRABE
CCPR/CO/76/EGY/Add.2	EGIPTO	2004	408	ÁRABE
CCPR/CO/76/EGY/Add.3	EGIPTO	2004	101	ÁRABE
CMW/C/EGY/1	EGIPTO	2006	22084	ÁRABE
CMW/C/EGY/Q/1/Add.1	EGIPTO	2007	8463	ÁRABE
CEDAW/C/EGY/7	EGIPTO	2008	25238	ÁRABE
A/HRC/WG.6/7/EGY/1	EGIPTO	2009	10466	ÁRABE
CEDAW/C/EGY/Q/7/Add.1	EGIPTO	2009	2654	ÁRABE
CRC/C/EGY/3-4	EGIPTO	2010	44002	ÁRABE
CRC/C/OPAC/EGY/1	EGIPTO	2010	5532	ÁRABE
CRC/C/OPSC/EGY/1	EGIPTO	2010	14621	ÁRABE
E/C.12/EGY/2-4	EGIPTO	2011	41136	ÁRABE
CEDAW/C/EGY/CO/7/Add.1	EGIPTO	2013	4176	ÁRABE
E/C.12/EGY/Q/2-4/Add.1	EGIPTO	2013	18266	ÁRABE
A/HRC/WG.6/20/EGY/1	EGIPTO	2014	9205	ÁRABE
CERD/C/EGY/17-22	EGIPTO	2014	25967	ÁRABE
CCPR/CO/76/EGY/Add.1	EGIPTO	2003	8236	INGLÉS
A/HRC/WG.6/7/IRQ/1	IRAQ	2009	10361	ÁRABE
CCPR/C/IRQ/5	IRAQ	2013	26199	ÁRABE
CEDAW/C/IRQ/4-6	IRAQ	2013	27340	ÁRABE
CERD/C/IRQ/15-21	IRAQ	2013	9546	ÁRABE
CRC/C/OPAC/IRQ/1	IRAQ	2013	5740	ÁRABE
CRC/C/OPSC/IRQ/1	IRAQ	2013	17985	ÁRABE
E/C.12/IRQ/4	IRAQ	2013	24749	ÁRABE
A/HRC/WG.6/20/IRQ/1	IRAQ	2014	10983	ÁRABE
CAT/C/IRQ/1	IRAQ	2014	9465	ÁRABE
CED/C/IRQ/1	IRAQ	2014	14858	ÁRABE
CRC/C/IRQ/2-4	IRAQ	2014	30296	ÁRABE
CRC/C/IRQ/Q/2-4/Add.1	IRAQ	2014	4450	ÁRABE
CRC/C/OPAC/IRQ/Q/1/Add.1	IRAQ	2014	1054	ÁRABE
CRC/C/OPSC/IRQ/1/Add.1	IRAQ	2014	696	ÁRABE
CCPR/C/IRQ/Q/5/Add.1	IRAQ	2015	10490	ÁRABE
CED/C/IRQ/Q/1/Add.1	IRAQ	2015	7400	ÁRABE
E/C.12/IRQ/Q/4/Add.1	IRAQ	2015	8657	ÁRABE
CEDAW/C/IRQ/CO/4-6/Add.1	IRAQ	2016	3479	ÁRABE
CCPR/C/IRQ/CO/5/Add.1	IRAQ	2017	4019	ÁRABE
CED/C/IRQ/CO/1/Add.1	IRAQ	2017	1126	ÁRABE
CERD/C/IRQ/22-25	IRAQ	2017	19958	ÁRABE
CRPD/C/IRQ/1	IRAQ	2017	16164	ÁRABE
CEDAW/C/IRQ/7	IRAQ	2018	20892	ÁRABE
CEDAW/C/IRQ/Q/4-6/Add.1	IRAQ	2013	12748	ÁRABE-INGLÉS
CEDAW/C/IRQ/Q/4-6/Add.1	IRAQ	2013	15673	INGLÉS-ÁRABE

CRC/C/JOR/Q/3/Add.1	JORDANIA	2006	14424	ÁRABE
CEDAW/C/JOR/3-4	JORDANIA	2006	39857	ÁRABE
CRC/C/JOR/3	JORDANIA	2006	40459	ÁRABE
CEDAW/C/JOR/Q/4/Add.1	JORDANIA	2007	2200	ÁRABE
A/HRC/WG.6/4/JOR/1	JORDANIA	2009	8052	ÁRABE
CAT/C/JOR/2	JORDANIA	2009	8865	ÁRABE
CCPR/C/JOR/3	JORDANIA	2009	10057	ÁRABE
CAT/C/JOR/Q/2/Add.1	JORDANIA	2010	9685	ÁRABE
CCPR/C/JOR/Q/4/ADD.1	JORDANIA	2010	7671	ÁRABE
CEDAW/C/JOR/5	JORDANIA	2010	37385	ÁRABE
CERD/C/JOR/13-17	JORDANIA	2011	7990	ÁRABE
CEDAW/C/JOR/Q/5/Add.1	JORDANIA	2012	10278	ÁRABE
A/HRC/WG.6/17/JOR/1	JORDANIA	2013	9464	ÁRABE
CRC/C/JOR/4-5	JORDANIA	2013	22661	ÁRABE
CRC/C/OPAC/JOR/1	JORDANIA	2013	6908	ÁRABE
CRC/C/OPSC/JOR/1	JORDANIA	2013	12215	ÁRABE
CAT/C/JOR/3	JORDANIA	2014	15765	ÁRABE
CEDAW/C/JOR/CO/5/Add.1	JORDANIA	2014	593	ÁRABE
CRC/C/JOR/Q/4-5/Add.1	JORDANIA	2014	9736	ÁRABE
CEDAW/C/JOR/6	JORDANIA	2015	20945	ÁRABE
CRPD/C/JOR/1	JORDANIA	2015	18555	ÁRABE
CCPR/C/JOR/5	JORDANIA	2016	17952	ÁRABE
CERD/C/JOR/18-20	JORDANIA	2016	11405	ÁRABE
CCPR/C/JOR/Q/5/Add.1	JORDANIA	2017	11262	ÁRABE
CEDAW/C/JOR/Q/6/Add.1	JORDANIA	2017	7953	ÁRABE
CRPD/C/JOR/Q/1/Add.1	JORDANIA	2017	13860	ÁRABE
A/HRC/WG.6/31/JOR/1	JORDANIA	2018	10499	ÁRABE
CRPD/C/JOR/CO/1/Add.1	JORDANIA	2018	1371	ÁRABE
CEDAW/C/KWT/1-2	KUWAIT	2003	29719	ÁRABE
CRC/C/OPAC/KWT/1	KUWAIT	2007	3930	ÁRABE
CRC/C/OPAC/KWT/Q/1/Add.1	KUWAIT	2007	1641	ÁRABE
CRC/C/OPSC/KWT/1	KUWAIT	2007	8630	ÁRABE
CRC/C/OPSC/KWT/Q/1/Add.1	KUWAIT	2007	3489	ÁRABE
CCPR/C/KWT/2	KUWAIT	2009	16694	ÁRABE
A/HRC/WG.6/8/KWT/1	KUWAIT	2010	9064	ÁRABE
CAT/C/KWT/2	KUWAIT	2010	11102	ÁRABE
CEDAW/C/KWT/3-4	KUWAIT	2010	9813	ÁRABE
CCPR/C/KWT/Q/2/Add.1	KUWAIT	2011	9949	ÁRABE
E/C.12/KWT/2	KUWAIT	2011	12797	ÁRABE
CCPR/C/KWT/CO/2/Add.1	KUWAIT	2012	851	ÁRABE
CRC/C/KWT/2	KUWAIT	2012	14626	ÁRABE
CRC/C/KWT/Q/2/Add.1	KUWAIT	2013	15810	ÁRABE
E/C.12/KWT/Q/2/Add.1	KUWAIT	2013	10625	ÁRABE
A/HRC/WG.6/21/KWT/1	KUWAIT	2014	10110	ÁRABE
CCPR/C/KWT/3	KUWAIT	2014	20580	ÁRABE

CAT/C/KWT/3	KUWAIT	2015	18902	ÁRABE
CEDAW/C/KWT/CO/3-4/Add.1	KUWAIT	2015	3213	ÁRABE
CCPR/C/KWT/Q/3/Add.1	KUWAIT	2016	10747	ÁRABE
CEDAW/C/KWT/5	KUWAIT	2016	11334	ÁRABE
CERD/C/KWT/21-24	KUWAIT	2016	21112	ÁRABE
CAT/C/KWT/CO/3/Add.1	KUWAIT	2017	6948	ÁRABE
CCPR/C/KWT/CO/3/Add.1	KUWAIT	2017	2837	ÁRABE
CEDAW/C/KWT/Q/5/Add.1	KUWAIT	2017	10454	ÁRABE
CRPD/C/KWT/1	KUWAIT	2017	19309	ÁRABE
CERD/C/KWT/CO/21-24/Add.1	KUWAIT	2018	826	ÁRABE
E/C.12/KWT/3	KUWAIT	2018	16469	ÁRABE
E/1990/5/Add.57	KUWAIT	2003	17133	ÁRABE
CRC/C/70/Add.8	LÍBANO	2000	59297	ÁRABE
CEDAW/C/LBN/2	LÍBANO	2005	30102	ÁRABE
CEDAW/C/LBN/3	LÍBANO	2006	35687	ÁRABE
CRC/C/LBN/Q/3/Add.1	LÍBANO	2006	14681	ÁRABE
CEDAW/C/LBN/Q/3/Add.1	LÍBANO	2007	4015	ÁRABE
A/HRC/WG.6/9/LBN/1	LÍBANO	2010	8575	ÁRABE
CEDAW/C/LBN/4-5	LÍBANO	2014	57717	ÁRABE
A/HRC/WG.6/23/LBN/1	LÍBANO	2015	11133	ÁRABE
CEDAW/C/LBN/Q/4-5/Add.1	LÍBANO	2015	6949	ÁRABE
CERD/C/LBN/18-22	LÍBANO	2015	11338	ÁRABE
E/C.12/LBN/2	LÍBANO	2015	13166	ÁRABE
CAT/C/LBN/1	LÍBANO	2016	32112	ÁRABE
CCPR/C/LBN/3	LÍBANO	2016	10800	ÁRABE
CRC/C/LBN/4-5	LÍBANO	2016	21062	ÁRABE
E/C.12/LBN/Q/2/Add.1	LÍBANO	2016	4481	ÁRABE
CRC/C/LBN/Q/4-5/Add.1	LÍBANO	2017	6492	ÁRABE
CCPR/C/LBN/Q/3/ADD.1	LÍBANO	2018	7098	ÁRABE
CEDAW/C/LBN/CO/4-5/Add.1	LÍBANO	2018	4003	ÁRABE
CEDAW/C/LBN/1	LÍBANO	2004	23698	ÁRABE-INGLÉS
CRC/C/129/Add.7	LÍBANO	2005	51095	ÁRABE-INGLÉS
CERD/C/383/Add.2	LÍBANO	2003	4995	FRANCÉS
CERD/C/475/Add.1	LÍBANO	2004	6298	FRANCÉS-INGLÉS
CEDAW/C/LBN/1	LÍBANO	2004	28540	INGLÉS-ÁRABE
CRC/C/129/Add.7	LÍBANO	2005	64632	INGLÉS-ÁRABE
CERD/C/475/Add.1	LÍBANO	2004	5801	INGLÉS-FRANCÉS
CRC/C/93/Add.1	LIBIA	2002	31362	ÁRABE
CERD/C/431/Add.5	LIBIA	2003	6561	ÁRABE
CCPR/C/LBY/4	LIBIA	2007	2602	ÁRABE
CCPR/C/LBY/CO/4/ADD.1	LIBIA	2009	718	ÁRABE
CEDAW/C/LBY/Q/2/Add.1	LIBIA	2009	11493	ÁRABE
A/HRC/WG.6/9/LBY/1	LIBIA	2010	9385	ÁRABE
A/HRC/WG.6/22/LBY/1	LIBIA	2015	7412	ÁRABE
E/1990/6/Add.38	LIBIA	2005	11689	ÁRABE

CEDAW/C/MOR/2	MARRUECOS	2000	23255	ÁRABE
CRC/C/93/Add.3	MARRUECOS	2003	40139	ÁRABE
CCPR/CO/82/MAR/Add.1	MARRUECOS	2005	1385	ÁRABE
CRC/C/OPSC/MAR/Q/1/ADD.1	MARRUECOS	2005	1496	ÁRABE
E/C.12/MAR/Q/2/Add.1	MARRUECOS	2006	8285	ÁRABE
E/C.12/MAR/Q/2/Add.2	MARRUECOS	2006	17737	ÁRABE
CAT/C/MAR/4	MARRUECOS	2009	16194	ÁRABE
CERD/C/MAR/17-18	MARRUECOS	2009	23151	ÁRABE
CAT/C/MAR/Q/4/Add.1	MARRUECOS	2011	18894	ÁRABE
CRPD/C/MAR/1	MARRUECOS	2015	24270	ÁRABE
CRPD/C/MAR/Q/1/Add.1	MARRUECOS	2017	10413	ÁRABE
CRC/C/OPSC/MAR/Q/1/ADD.2	MARRUECOS	2006	2203	ÁRABE-FRANCÉS
CERD/C/430/Add.1	MARRUECOS	2002	4992	FRANCÉS
CAT/C/66/Add.1	MARRUECOS	2003	10997	FRANCÉS
CAT/C/66/Add.1/Corr.1	MARRUECOS	2003	241	FRANCÉS
CCPR/C/MAR/2004/5	MARRUECOS	2004	19497	FRANCÉS
CRC/C/OPSA/MAR/1	MARRUECOS	2005	11482	FRANCÉS
E/1994/104/Add.29	MARRUECOS	2005	25550	FRANCÉS
CEDAW/C/MAR/4	MARRUECOS	2006	33022	FRANCÉS
CEDAW/C/MAR/Q/4/Add.1	MARRUECOS	2007	7168	FRANCÉS
A/HRC/WG.6/1/MAR/1	MARRUECOS	2008	12417	FRANCÉS
A/HRC/WG.6/13/MAR/1	MARRUECOS	2012	8285	FRANCÉS
CERD/C/MAR/CO/17-18/Add.1	MARRUECOS	2012	1622	FRANCÉS
CMW/C/MAR/1	MARRUECOS	2012	34830	FRANCÉS
CRC/C/OPAC/MAR/1	MARRUECOS	2012	3896	FRANCÉS
CAT/C/MAR/CO/4/Add.1	MARRUECOS	2013	13132	FRANCÉS
CMW/C/MAR/Q/1/Add.1	MARRUECOS	2013	13577	FRANCÉS
CRC/C/MAR/3-4	MARRUECOS	2013	27761	FRANCÉS
CRC/C/MAR/Q/3-4/Add.1	MARRUECOS	2014	16574	FRANCÉS
CRC/C/OPAC/MAR/Q/1/Add.1	MARRUECOS	2014	1700	FRANCÉS
E/C.12/MAR/4	MARRUECOS	2014	50323	FRANCÉS
CCPR/C/MAR/6	MARRUECOS	2015	21199	FRANCÉS
E/C.12/MAR/Q/4/Add.1	MARRUECOS	2015	11362	FRANCÉS
CCPR/C/MAR/Q/6/Add.1	MARRUECOS	2016	10662	FRANCÉS
A/HRC/WG.6/27/MAR/1	MARRUECOS	2017	10599	FRANCÉS
CRC/C/OPSC/MAR/Q/1/ADD.2	MARRUECOS	2006	2827	FRANCÉS-ÁRABE
CAT/C/CR/31/2/Add.1	MARRUECOS	2005	2662	INGLÉS
CRC/C/8/Add.42	MAURITANIA	2001	23481	FRANCÉS
CERD/C/421/Add.1	MAURITANIA	2004	14928	FRANCÉS
CEDAW/C/MRT/1	MAURITANIA	2005	25750	FRANCÉS
CEDAW/C/MRT/Q/1/Add.1	MAURITANIA	2007	12445	FRANCÉS
CRC/C/MRT/2	MAURITANIA	2008	23283	FRANCÉS
CRC/C/MRT/Q/2/Add.1	MAURITANIA	2009	9118	FRANCÉS
A/HRC/WG.6/9/MRT/1	MAURITANIA	2010	9567	FRANCÉS
E/C.12/MRT/1	MAURITANIA	2011	47418	FRANCÉS

CAT/C/MRT/1	MAURITANIA	2012	13609	FRANCÉS
CCPR/C/MRT/1	MAURITANIA	2012	15079	FRANCÉS
CEDAW/C/MRT/2-3	MAURITANIA	2012	24328	FRANCÉS
E/C.12/MRT/Q/1/Add.1	MAURITANIA	2012	6614	FRANCÉS
CCPR/C/MRT/Q/1/Add.1	MAURITANIA	2013	10307	FRANCÉS
CEDAW/C/MRT/Q/2-3/Add.1	MAURITANIA	2014	11621	FRANCÉS
A/HRC/WG.6/23/MRT/1	MAURITANIA	2015	9821	FRANCÉS
CCPR/C/MRT/CO/1/Add.1	MAURITANIA	2015	1924	FRANCÉS
CMW/C/MRT/1	MAURITANIA	2015	18514	FRANCÉS
CAT/C/MRT/2	MAURITANIA	2017	10461	FRANCÉS
CEDAW/C/MRT/CO/2-3/Add.1	MAURITANIA	2017	3868	FRANCÉS
CERD/C/MRT/8-14	MAURITANIA	2017	13675	FRANCÉS
CAT/C/MRT/Q/2/Add.1	MAURITANIA	2018	10702	FRANCÉS
CCPR/C/MRT/2	MAURITANIA	2018	16091	FRANCÉS
CRC/C/MRT/Q/3-5/Add.1	MAURITANIA	2018	9527	FRANCÉS
CRC/C/OMN/Q/2/Add.1	OMÁN	2006	11251	ÁRABE
CERD/C/OMN/1	OMÁN	2006	9271	ÁRABE
CRC/C/OMN/2	OMÁN	2006	34628	ÁRABE
CRC/C/OPAC/OMN/1	OMÁN	2009	849	ÁRABE
CRC/C/OPAC/OMN/Q/1/Add.1	OMÁN	2009	1629	ÁRABE
CRC/C/OPSC/OMN/1	OMÁN	2009	3403	ÁRABE
CRC/C/OPSC/OMN/Q/1/Add.1	OMÁN	2009	6458	ÁRABE
CEDAW/C/OMN/1	OMÁN	2010	19510	ÁRABE
CERD/C/OMN/2-5	OMÁN	2014	16514	ÁRABE
A/HRC/WG.6/23/OMN/1	OMÁN	2015	7535	ÁRABE
CRC/C/OMN/3-4	OMÁN	2015	34455	ÁRABE
CRC/C/OMN/Q/3-4/Add.1	OMÁN	2015	15364	ÁRABE
CEDAW/C/OMN/2-3	OMÁN	2016	17079	ÁRABE
CRPD/C/OMN/1	OMÁN	2016	23227	ÁRABE
CEDAW/C/OMN/CO/1/Add.1	OMÁN	2017	2126	ÁRABE
CEDAW/C/OMN/Q/2-3/Add.1	OMÁN	2017	7855	ÁRABE
CERD/C/OMN/CO/2-5/ADD.1	OMÁN	2018	1092	ÁRABE
CRPD/C/OMN/Q/1/Add.1	OMÁN	2018	10751	ÁRABE
CRC/C/78/Add.1	OMÁN	2000	20397	INGLÉS
A/HRC/WG.6/10/OMN/1	OMÁN	2010	10367	INGLÉS
CEDAW/C/PSE/1	PALESTINA	2017	32924	ÁRABE
CEDAW/C/PSE/Q/1/Add.1	PALESTINA	2018	11022	ÁRABE
CERD/C/PSE/1-2	PALESTINA	2018	32205	ÁRABE
CERD/C/360/Add.1	QATAR	2001	5353	ÁRABE
CRC/C/51/Add.5	QATAR	2001	12901	ÁRABE
CAT/C/58/Add.1	QATAR	2005	6939	ÁRABE
CRC/C/OPSA/QAT/1	QATAR	2005	9682	ÁRABE
CRC/C/OPAC/QAT/1	QATAR	2006	1545	ÁRABE
CRC/C/OPAC/QAT/Q/1/Add.1	QATAR	2007	3163	ÁRABE
CAT/C/QAT/CO/1/Add.1	QATAR	2008	1310	ÁRABE

CRC/C/QAT/2	QATAR	2008	37904	ÁRABE
CRC/C/QAT/Q/2/ADD.1	QATAR	2009	7212	ÁRABE
A/HRC/WG.6/7/QAT/1	QATAR	2009	9845	ÁRABE
CAT/C/QAT/2	QATAR	2011	11068	ÁRABE
CERD/C/QAT/13-16	QATAR	2011	13760	ÁRABE
CAT/C/QAT/Q/2/Add.2	QATAR	2012	19602	ÁRABE
CEDAW/C/QAT/1	QATAR	2012	41883	ÁRABE
CEDAW/C/QAT/Q/1/Add.1	QATAR	2013	12275	ÁRABE
A/HRC/WG.6/19/QAT/1	QATAR	2014	10721	ÁRABE
CAT/C/QAT/CO/2/Add.1	QATAR	2014	4461	ÁRABE
CRPD/C/QAT/1	QATAR	2014	22023	ÁRABE
CRPD/C/QAT/Q/1/Add.1	QATAR	2015	10941	ÁRABE
CEDAW/C/QAT/CO/1/Add.1	QATAR	2016	2818	ÁRABE
CRC/C/QAT/3-4	QATAR	2016	43763	ÁRABE
CAT/C/QAT/3	QATAR	2017	18695	ÁRABE
CERD/C/QAT/17-21	QATAR	2017	17089	ÁRABE
CRC/C/QAT/Q/3-4/Add.1	QATAR	2017	10827	ÁRABE
CEDAW/C/QAT/2	QATAR	2018	21605	ÁRABE
CRC/C/OPSC/QAT/Q/1/Add.1	QATAR	2006	3775	ÁRABE-INGLÉS
CRC/C/OPSC/QAT/Q/1/Add.1	QATAR	2006	4260	INGLÉS-ÁRABE
CCPR/C/SYR/2000/2	SIRIA	2000	26767	ÁRABE
CRC/C/93/Add.2	SIRIA	2002	22308	ÁRABE
CCPR/C/SYR/2004/3	SIRIA	2004	38783	ÁRABE
CEDAW/C/SYR/1	SIRIA	2005	32710	ÁRABE
CCPR/CO/84/SYR/Add.1	SIRIA	2006	1804	ÁRABE
CRC/C/OPSc/SYR/1	SIRIA	2006	22068	ÁRABE
CEDAW/C/SYR/Q/1/Add.1	SIRIA	2007	11282	ÁRABE
CMW/C/SYR/1	SIRIA	2007	10483	ÁRABE
CRC/C/OPAC/SYR/1	SIRIA	2007	2957	ÁRABE
CRC/C/OPAC/SYR/Q/1/Add.1	SIRIA	2007	1536	ÁRABE
CMW/C/SYR/Q/1/Add.1	SIRIA	2008	5994	ÁRABE
CAT/C/SYR/1	SIRIA	2009	12061	ÁRABE
CRC/C/SYR/3-4	SIRIA	2010	29978	ÁRABE
A/HRC/WG.6/12/SYR/1	SIRIA	2011	10870	ÁRABE
CAT/C/SYR/CO/1/Add.1	SIRIA	2011	6300	ÁRABE
CRC/C/SYR/3-4/ADD.1	SIRIA	2011	13371	ÁRABE
CRC/C/SYR/Q/3-4/ADD.1	SIRIA	2011	9934	ÁRABE
CEDAW/C/SYR/2	SIRIA	2012	35927	ÁRABE
CEDAW/C/SYR/Q/2/Add.1	SIRIA	2014	8753	ÁRABE
A/HRC/WG.6/26/SYR/1	SIRIA	2016	13916	ÁRABE
CEDAW/C/SYR/CO/2/Add.1	SIRIA	2016	1971	ÁRABE
CRC/C/SYR/5	SIRIA	2017	21129	ÁRABE
CRC/C/SYR/Q/5/Add.1	SIRIA	2018	6253	ÁRABE
A/HRC/WG.6/11/SOM/1	SOMALIA	2011	6554	INGLÉS
A/HRC/WG.6/24/SOM/1	SOMALIA	2015	10877	INGLÉS

CRC/C/65/Add.17	SUDÁN	2001	47728	ÁRABE
CRC/C/OPSC/SDN/1	SUDÁN	2006	6810	ÁRABE
CCPR/C/SDN/3	SUDÁN	2007	23213	ÁRABE
CCPR/C/SDN/Q/3/Add.1	SUDÁN	2007	10164	ÁRABE
CCPR/C/SDN/CO/3/Add.1	SUDÁN	2009	2688	ÁRABE
CRC/C/OPAC/SDN/1	SUDÁN	2009	9546	ÁRABE
CRC/C/SDN/3-4	SUDÁN	2010	28409	ÁRABE
CCPR/C/SDN/4	SUDÁN	2012	19333	ÁRABE
CERD/C/SDN/12-16	SUDÁN	2013	12475	ÁRABE
E/C.12/SDN/2	SUDÁN	2013	20555	ÁRABE
CRPD/C/SDN/1	SUDÁN	2015	8775	ÁRABE
E/C.12/SDN/Q/2/Add.1	SUDÁN	2015	8415	ÁRABE
A/HRC/WG.6/25/SDN/1	SUDÁN	2016	9021	ÁRABE
CCPR/C/SDN/5	SUDÁN	2017	10167	ÁRABE
CRPD/C/SDN/Q/1/Add.1	SUDÁN	2017	4982	ÁRABE
CCPR/C/SDN/Q/5/Add.1	SUDÁN	2018	6175	ÁRABE
CRC/C/OPAC/SDN/Q/1/ADD.1	SUDÁN	2010	2800	ÁRABE-INGLÉS
CRC/C/SDN/Q/3-4/ADD.1	SUDÁN	2010	9334	ÁRABE-INGLÉS
A/HRC/WG.6/11/SDN/1	SUDÁN	2011	12941	ÁRABE-INGLÉS
CERD/C/334/Add.2	SUDÁN	2000	15860	INGLÉS
CCPR/C/SDN/Q/4/Add.1	SUDÁN	2014	6331	INGLÉS
CRC/C/OPAC/SDN/Q/1/ADD.1	SUDÁN	2010	3274	INGLÉS-ÁRABE
CRC/C/SDN/Q/3-4/ADD.1	SUDÁN	2010	11616	INGLÉS-ÁRABE
A/HRC/WG.6/11/SDN/1	SUDÁN	2011	16689	INGLÉS-ÁRABE
CAT/C/TUN/3	TÚNEZ	2010	33271	ÁRABE
CCPR/C/TUN/CO/5/ADD.2	TÚNEZ	2010	4925	ÁRABE
CRPD/C/TUN/1	TÚNEZ	2010	15914	ÁRABE
CRPD/C/TUN/Q/1/Add.1	TÚNEZ	2011	16776	ÁRABE
A/HRC/WG.6/13/TUN/1	TÚNEZ	2012	7401	ÁRABE
CAT/C/TUN/3/Add.1	TÚNEZ	2014	20573	ÁRABE
CED/C/TUN/1	TÚNEZ	2014	19751	ÁRABE
E/C.12/TUN/3	TÚNEZ	2015	19470	ÁRABE
CAT/C/TUN/Q/3/ADD.1	TÚNEZ	2016	7883	ÁRABE
A/HRC/WG.6/27/TUN/1	TÚNEZ	2017	10913	ÁRABE
CAT/C/TUN/CO/3/Add.1	TÚNEZ	2017	3494	ÁRABE
CEDAW/C/TUN/3-4	TÚNEZ	2000	88084	FRANCÉS
CRC/C/83/Add.1	TÚNEZ	2001	58420	FRANCÉS
CERD/C/431/Add.4	TÚNEZ	2002	7928	FRANCÉS
CCPR/C/TUN/5	TÚNEZ	2007	45050	FRANCÉS
CERD/C/TUN/19	TÚNEZ	2007	30394	FRANCÉS
CRC/C/OPAC/TUN/1	TÚNEZ	2007	3841	FRANCÉS
A/HRC/WG.6/1/TUN/1	TÚNEZ	2008	13671	FRANCÉS
CCPR/C/TUN/Q/5/Add.1	TÚNEZ	2008	17033	FRANCÉS
CRC/C/OPAC/TUN/Q/1/Add.1	TÚNEZ	2008	2116	FRANCÉS
CRC/C/TUN/3	TÚNEZ	2008	38320	FRANCÉS

CCPR/C/TUN/CO/5/Add.1	TÚNEZ	2009	1620	FRANCÉS
CEDAW/C/TUN/6	TÚNEZ	2009	32248	FRANCÉS
CERD/C/TUN/Q/19/ADD.1	TÚNEZ	2009	16464	FRANCÉS
CEDAW/C/TUN/Q/6/Add.1	TÚNEZ	2010	27521	FRANCÉS
CRC/C/TUN/Q/3/Add.1	TÚNEZ	2010	14221	FRANCÉS
CED/C/TUN/Q/1/Add.1	TÚNEZ	2015	3373	FRANCÉS
E/C.12/TUN/Q/3/Add.1	TÚNEZ	2016	21078	FRANCÉS
CED/C/TUN/CO/1/Add.1	TÚNEZ	2017	1689	FRANCÉS
CCPR/C/YEM/2001/3	YEMEN	2001	13215	ÁRABE
CEDAW/C/YEM/5	YEMEN	2002	33176	ÁRABE
CERD/C/362/Add.8	YEMEN	2002	16727	ÁRABE
E/1990/5/Add.54	YEMEN	2002	17443	ÁRABE
CAT/C/16/Add.10	YEMEN	2003	13460	ÁRABE
CRC/C/129/Add.2	YEMEN	2004	27085	ÁRABE
CAT/C/CR/31/4/Add.1	YEMEN	2005	2764	ÁRABE
CCPR/C/YEM/2004/4	YEMEN	2005	41131	ÁRABE
CERD/C/YEM/16	YEMEN	2006	36076	ÁRABE
CEDAW/C/YEM/Q/6/Add.1	YEMEN	2008	5996	ÁRABE
A/HRC/WG.6/5/YEM/1	YEMEN	2009	8766	ÁRABE
CAT/C/YEM/2	YEMEN	2009	26765	ÁRABE
CRC/C/OPSC/YEM/1	YEMEN	2009	16877	ÁRABE
CRC/C/OPSC/YEM/Q/1/Add.1	YEMEN	2009	7802	ÁRABE
E/C.12/YEM/2	YEMEN	2009	21580	ÁRABE
CCPR/C/YEM/5	YEMEN	2010	30868	ÁRABE
CERD/C/YEM/17-18	YEMEN	2010	32937	ÁRABE
E/C.12/YEM/Q/2/ADD.1	YEMEN	2011	24216	ÁRABE
CRC/C/YEM/4	YEMEN	2012	59082	ÁRABE
A/HRC/WG.6/18/YEM/1	YEMEN	2013	9906	ÁRABE
CRC/C/OPAC/YEM/1	YEMEN	2013	21223	ÁRABE
CRC/C/OPAC/YEM/Q/1/Add.1	YEMEN	2013	2724	ÁRABE
CRC/C/YEM/Q/4/Add.1	YEMEN	2013	17859	ÁRABE
CCPR/C/YEM/CO/5/Add.1	YEMEN	2014	3014	ÁRABE
CEDAW/C/YEM/7-8	YEMEN	2014	33580	ÁRABE
E/C.12/YEM/3	YEMEN	2014	15432	ÁRABE
A/HRC/WG.6/32/YEM/1	YEMEN	2018	10281	ÁRABE
CEDAW/C/YEM/4	YEMEN	2000	28271	ÁRABE-INGLÉS
CEDAW/C/YEM/6	YEMEN	2007	26540	ÁRABE-INGLÉS
CEDAW/C/YEM/4	YEMEN	2000	54072	INGLÉS-ÁRABE
CEDAW/C/YEM/6	YEMEN	2007	30511	INGLÉS-ÁRABE

Anexo II

Composición del Corpus 2 con datos desglosados por signatura, Estado, año y número de palabras en árabe (PA), español (PE) e inglés (PI).

SIGNATURA	ESTADO	AÑO	PA	PE	PI
CRC/C/136/Add.1	ARABIA SAUDITA	2005	21543	30085	27292
CEDAW/C/SAU/2	ARABIA SAUDITA	2007	17289	24242	20887
A/HRC/WG.6/17/SAU/1	ARABIA SAUDITA	2013	11367	17703	14934
CAT/C/SAU/2	ARABIA SAUDITA	2015	15210	23560	19724
CRC/C/SAU/3-4	ARABIA SAUDITA	2015	42774	66831	53940
CRPD/C/SAU/1	ARABIA SAUDITA	2015	26367	36353	33194
CAT/C/SAU/Q/2/Add.2	ARABIA SAUDITA	2016	11036	15994	13638
CEDAW/C/SAU/3-4	ARABIA SAUDITA	2016	36904	30659	25474
CERD/C/SAU/4-9	ARABIA SAUDITA	2016	20987	29299	25125
CEDAW/C/SAU/Q/3-4/Add.1	ARABIA SAUDITA	2017	10345	15172	12724
CRC/C/OPAC/SAU/1	ARABIA SAUDITA	2017	11181	17198	14105
CRC/C/OPSC/SAU/1	ARABIA SAUDITA	2017	15495	20089	19544
A/HRC/WG.6/31/SAU/1	ARABIA SAUDITA	2018	12372	17435	14170
CAT/C/47/Add.4	BAHREIN	2004	11091	14552	14490
CERD/C/443/Add.1	BAHREIN	2004	17744	25233	22978
CEDAW/C/BHR/2	BAHREIN	2007	45634	60494	49838
CEDAW/C/BHR/2/ADD.1	BAHREIN	2008	21010	26173	22202
CRC/C/BHR/2-3	BAHREIN	2010	40098	61905	50182
CEDAW/C/BHR/3	BAHREIN	2011	19102	28158	23406
CEDAW/C/BHR/Q/3/Add.1	BAHREIN	2013	10143	13725	11657
CAT/C/BHR/2	BAHREIN	2015	21464	27414	25753
CAT/C/BHR/3	BAHREIN	2016	21073	29138	25174
CCPR/C/BHR/1	BAHREIN	2017	26019	39120	34946
A/HRC/WG.6/27/BHR/1	BAHREIN	2017	10648	16233	12860

CCPR/C/BHR/Q/1/Add.1	BAHREIN	2018	10754	15408	13333
CEDAW/C/BHR/4	BAHREIN	2018	17840	25681	22324
CRC/C/BHR/4-6	BAHREIN	2018	21575	33042	27015
CRC/C/BHR/Q/4-6/Add.1	BAHREIN	2018	10725	14921	12865
A/HRC/WG.6/3/ARE/1	EAU	2008	10018	16666	13227
CEDAW/C/ARE/1	EAU	2008	19158	29561	25717
CEDAW/C/ARE/Q/1/Add.1	EAU	2009	13367	19388	17112
CERD/C/ARE/12-17	EAU	2009	12019	17486	15294
A/HRC/WG.6/15/ARE/1	EAU	2012	10508	17492	13497
CEDAW/C/ARE/2-3	EAU	2014	10502	14279	11621
CRC/C/ARE/2	EAU	2014	23069	32936	27257
CRPD/C/ARE/1	EAU	2014	20375	29902	26332
CERD/C/ARE/18-21	EAU	2016	17077	26149	21178
A/HRC/WG.6/29/ARE/1	EAU	2017	10435	16151	13026
CAT/C/55/Add.6	EGIPTO	2001	10596	15661	14972
CERD/C/384/Add.3	EGIPTO	2001	27609	38434	35843
CCPR/C/EGY/2001/3	EGIPTO	2002	44008	62261	59324
CMW/C/EGY/1	EGIPTO	2006	22084	32298	30823
CEDAW/C/EGY/7	EGIPTO	2008	25238	33570	27914
A/HRC/WG.6/7/EGY/1	EGIPTO	2009	10466	16093	13474
CRC/C/EGY/3-4	EGIPTO	2010	44002	57099	49642
CRC/C/OPSC/EGY/1	EGIPTO	2010	14621	23222	19566
E/C.12/EGY/2-4	EGIPTO	2011	41136	64194	57596
E/C.12/EGY/Q/2-4/Add.1	EGIPTO	2013	18266	29932	22903
A/HRC/WG.6/7/IRQ/1	IRAQ	2009	10361	17561	13142
CCPR/C/IRQ/5	IRAQ	2013	26199	37811	34210
CEDAW/C/IRQ/4-6	IRAQ	2013	27340	37199	33011
CEDAW/C/IRQ/Q/4-6/Add.1	IRAQ	2013	12748	18408	15673
CRC/C/OPSC/IRQ/1	IRAQ	2013	17985	24832	22291
E/C.12/IRQ/4	IRAQ	2013	24749	32863	31716
A/HRC/WG.6/20/IRQ/1	IRAQ	2014	10983	17541	14986
CED/C/IRQ/1	IRAQ	2014	14858	22725	18651
CRC/C/IRQ/2-4	IRAQ	2014	30296	43762	37016
CCPR/C/IRQ/Q/5/Add.1	IRAQ	2015	10490	16190	14574
CERD/C/IRQ/22-25	IRAQ	2017	19958	31064	25393
CRPD/C/IRQ/1	IRAQ	2017	16164	25047	20881
CEDAW/C/IRQ/7	IRAQ	2018	20892	29160	24022
CEDAW/C/JOR/3-4	JORDANIA	2006	39857	59606	50779
CRC/C/JOR/3	JORDANIA	2006	40459	59187	54850
CCPR/C/JOR/3	JORDANIA	2009	10057	16142	12786
CEDAW/C/JOR/5	JORDANIA	2010	37385	56235	51089
CEDAW/C/JOR/Q/5/Add.1	JORDANIA	2012	10278	15941	12692
CRC/C/JOR/4-5	JORDANIA	2013	22661	31784	27790

CRC/C/OPSC/JOR/1	JORDANIA	2013	12215	18690	14748
CAT/C/JOR/3	JORDANIA	2014	15765	21734	19199
CEDAW/C/JOR/6	JORDANIA	2015	20945	34414	26401
CRPD/C/JOR/1	JORDANIA	2015	18555	27558	22793
CCPR/C/JOR/5	JORDANIA	2016	17952	24775	22198
CERD/C/JOR/18-20	JORDANIA	2016	11405	17486	14371
CCPR/C/JOR/Q/5/Add.1	JORDANIA	2017	11262	16598	14857
A/HRC/WG.6/31/JOR/1	JORDANIA	2018	10499	16716	12955
CEDAW/C/KWT/1-2	KUWAIT	2003	29719	41448	36930
E/1990/5/Add.57	KUWAIT	2003	17133	23059	21168
CCPR/C/KWT/2	KUWAIT	2009	16694	23807	22417
CAT/C/KWT/2	KUWAIT	2010	11102	17042	12571
E/C.12/KWT/2	KUWAIT	2011	12797	19130	15713
CRC/C/KWT/2	KUWAIT	2012	14626	23451	18811
E/C.12/KWT/Q/2/Add.1	KUWAIT	2013	10625	16246	13127
A/HRC/WG.6/21/KWT/1	KUWAIT	2014	10110	16555	12936
CCPR/C/KWT/3	KUWAIT	2014	20580	29045	26140
CAT/C/KWT/3	KUWAIT	2015	18902	26181	23606
CCPR/C/KWT/Q/3/Add.1	KUWAIT	2016	10747	16477	13515
CEDAW/C/KWT/5	KUWAIT	2016	11334	16427	14106
CERD/C/KWT/21-24	KUWAIT	2016	21112	30371	25966
CEDAW/C/KWT/Q/5/Add.1	KUWAIT	2017	10454	15064	13032
CRPD/C/KWT/1	KUWAIT	2017	19309	26605	23408
E/C.12/KWT/3	KUWAIT	2018	16469	23617	20906
CEDAW/C/LBN/2	LÍBANO	2005	30102	41442	37261
CRC/C/129/Add.7	LÍBANO	2005	51095	71360	64632
CEDAW/C/LBN/3	LÍBANO	2006	35687	49536	43472
CRC/C/LBN/Q/3/Add.1	LÍBANO	2006	14681	23856	19819
CEDAW/C/LBN/4-5	LÍBANO	2014	57717	76365	67045
A/HRC/WG.6/23/LBN/1	LÍBANO	2015	11133	16586	14411
CERD/C/LBN/18-22	LÍBANO	2015	11338	16148	13286
E/C.12/LBN/2	LÍBANO	2015	13166	18400	16669
CAT/C/LBN/1	LÍBANO	2016	32112	43425	39188
CCPR/C/LBN/3	LÍBANO	2016	10800	17243	14561
CRC/C/LBN/4-5	LÍBANO	2016	21062	32823	27533
E/1990/6/Add.38	LIBIA	2005	11689	15414	13934
E/C.12/MAR/Q/2/Add.2	MARRUECOS	2006	17737	27572	21781
CAT/C/MAR/4	MARRUECOS	2009	16194	22603	17847
CERD/C/MAR/17-18	MARRUECOS	2009	23151	34318	28834
CRPD/C/MAR/1	MARRUECOS	2015	24270	35687	30894
CRPD/C/MAR/Q/1/Add.1	MARRUECOS	2017	10413	15392	13208
CRC/C/OMN/2	OMÁN	2006	34628	47903	45308
CEDAW/C/OMN/1	OMÁN	2010	19510	28240	24375

CERD/C/OMN/2-5	OMÁN	2014	16514	23262	20840
CRC/C/OMN/3-4	OMÁN	2015	34455	54164	42354
CEDAW/C/OMN/2-3	OMÁN	2016	17079	23989	20044
CRPD/C/OMN/1	OMÁN	2016	23227	35876	28519
CEDAW/C/PSE/1	PALESTINA	2017	32924	47042	39296
CEDAW/C/PSE/Q/1/Add.1	PALESTINA	2018	11022	14989	13759
CERD/C/PSE/1-2	PALESTINA	2018	32205	43622	38045
CRC/C/QAT/2	QATAR	2008	37904	58766	44247
CAT/C/QAT/2	QATAR	2011	11068	16720	13386
CERD/C/QAT/13-16	QATAR	2011	13760	20503	16943
CEDAW/C/QAT/1	QATAR	2012	41883	62286	52273
CEDAW/C/QAT/Q/1/Add.1	QATAR	2013	12275	17878	15073
A/HRC/WG.6/19/QAT/1	QATAR	2014	10721	17880	14095
CRPD/C/QAT/1	QATAR	2014	22023	31453	27259
CRPD/C/QAT/Q/1/Add.1	QATAR	2015	10941	15295	13835
CRC/C/QAT/3-4	QATAR	2016	43763	67141	51953
CAT/C/QAT/3	QATAR	2017	18695	29008	24607
CERD/C/QAT/17-21	QATAR	2017	17089	21927	21306
CEDAW/C/QAT/2	QATAR	2018	21605	30515	27231
CCPR/C/SYR/2000/2	SIRIA	2000	26767	39004	36603
CCPR/C/SYR/2004/3	SIRIA	2004	38783	57581	51426
CEDAW/C/SYR/1	SIRIA	2005	32710	50932	43675
CRC/C/OPSc/SYR/1	SIRIA	2006	22068	32178	30427
CEDAW/C/SYR/Q/1/Add.1	SIRIA	2007	11282	15388	13361
CMW/C/SYR/1	SIRIA	2007	10483	14811	13327
CAT/C/SYR/1	SIRIA	2009	12061	16349	15191
CRC/C/SYR/3-4	SIRIA	2010	29978	46352	38251
A/HRC/WG.6/12/SYR/1	SIRIA	2011	10870	19164	14722
CRC/C/SYR/3-4/ADD.1	SIRIA	2011	13371	20779	17058
CEDAW/C/SYR/2	SIRIA	2012	35927	50200	42769
A/HRC/WG.6/26/SYR/1	SIRIA	2016	13916	18839	18006
CCPR/C/SDN/3	SUDÁN	2007	23213	32295	30005
CCPR/C/SDN/Q/3/Add.1	SUDÁN	2007	10164	19073	13610
CRC/C/SDN/3-4	SUDÁN	2010	28409	38615	32440
A/HRC/WG.6/11/SDN/1	SUDÁN	2011	12941	19103	16689
CCPR/C/SDN/4	SUDÁN	2012	19333	32002	25957
CERD/C/SDN/12-16	SUDÁN	2013	12475	18120	16011
E/C.12/SDN/2	SUDÁN	2013	20555	31937	26333
CCPR/C/SDN/5	SUDÁN	2017	10167	15619	13154
CAT/C/TUN/3	TÚNEZ	2010	33271	47552	36507
CRPD/C/TUN/1	TÚNEZ	2010	15914	25166	17989
CAT/C/TUN/3/Add.1	TÚNEZ	2014	20573	32103	25533
CED/C/TUN/1	TÚNEZ	2014	19751	28541	25233

A/HRC/WG.6/27/TUN/1	TÚNEZ	2017	10913	17535	14390
CCPR/C/YEM/2001/3	YEMEN	2001	13215	18973	18501
CERD/C/362/Add.8	YEMEN	2002	16727	24799	22379
E/1990/5/Add.54	YEMEN	2002	17443	24505	22488
CAT/C/16/Add.10	YEMEN	2003	13460	18760	18267
CRC/C/129/Add.2	YEMEN	2004	27085	40065	35530
CCPR/C/YEM/2004/4	YEMEN	2005	41131	62223	59518
CERD/C/YEM/16	YEMEN	2006	36076	50540	46256
CAT/C/YEM/2	YEMEN	2009	26765	38869	34511
CRC/C/OPSC/YEM/1	YEMEN	2009	16877	25417	21455
E/C.12/YEM/2	YEMEN	2009	21580	30611	25817
CCPR/C/YEM/5	YEMEN	2010	30868	45584	42564
CERD/C/YEM/17-18	YEMEN	2010	32937	46841	42705
CRC/C/YEM/4	YEMEN	2012	59082	88185	71685
CRC/C/OPAC/YEM/1	YEMEN	2013	21223	30621	25600
CEDAW/C/YEM/7-8	YEMEN	2014	33580	43833	38770
E/C.12/YEM/3	YEMEN	2014	15432	23888	19783
A/HRC/WG.6/32/YEM/1	YEMEN	2018	10281	15714	13745

Anexo III

Relación de las 996 *stopwords* localizadas en las traducciones del árabe al español con datos desglosados por documentos y segmentos (documento_segmento).

SEG.	STOPW.
B_1	حيث
B_1	على
B_1	واستجابة
B_1	واستمرت
B_1	والتشريع
B_1	وتتعاون
B_1	وتمكن
B_1	وعلى
B_1	وقد
B_1	وقد
B_1	وهي
B_10	إذ
B_10	حيث
B_10	حيث
B_10	فضلاً
B_10	كما
B_10	واشترط
B_10	وذلك
B_10	وعلى
B_10	ونشير
B_10	ونشير
B_10	ويتم
B_10	ويتم
B_10	ويتم

B_10	ويشترط
B_10	يعمل
B_10	يعمل
B_11	حيث
B_11	للتحقيق
B_11	وتحدد
B_11	وتضع
B_11	وتلقي
B_11	وعلى
B_11	وقد
B_11	وللنزيل
B_11	ونص
B_11	ووضعها
B_11	ويخطر
B_12	إضافة
B_12	فضلاً
B_12	فيها
B_12	وتعمل
B_12	وطعنت
B_12	وعلى
B_12	وقد
B_12	ويظهر
B_13	تم
B_13	على
B_13	فصدر

B_13	متضمنا
B_13	وهي
B_14	وقد
B_14	وكم
B_14	ومن
B_15	حيث
B_15	حيث
B_15	حيث
B_15	حيث
B_15	على
B_15	فضلاً
B_15	ففي
B_15	فقد
B_15	كما
B_15	كما
B_15	وأنه
B_15	وعليه
B_15	وفي
B_15	وفي
B_15	وفي
B_15	وهناك
B_16	بهدف
B_16	حيث
B_16	كما
B_16	كما

وإصلاح	B_5
وانتهت	B_5
وباشرت	B_5
وتتمثل	B_5
وتقوم	B_5
وذلك	B_5
حيث	B_6
حيث	B_6
كما	B_6
من	B_6
والأكاديمية	B_6
وامتدت	B_6
وجرى	B_6
أدارها	B_7
التي	B_7
بمشاركة	B_7
بهدف	B_7
كما	B_7
وجاري	B_7
وذلك	B_7
وذلك	B_7
وفي	B_7
وقعت	B_7
ولقد	B_7
ولقد	B_7
ولقد	B_7
بادرت	B_8
بالإضافة	B_8
تم	B_8
فقد	B_8
فقد	B_8
فقد	B_8
كما	B_8
كما	B_8
والمؤسسة	B_8
وأي	B_8
وتختص	B_8

وقد	B_2
وقد	B_2
وللأمانة	B_2
ومن	B_2
وهي	B_2
بما	B_20
حيث	B_20
حيث	B_20
حيث	B_20
وبالإضافة	B_20
وحولت	B_20
وقد	B_20
وقد	B_20
وقد	B_20
فقد	B_21
وتم	B_21
وعلاوة	B_21
وهي	B_21
وورد	B_21
إعداد	B_3
علاوة	B_3
كما	B_3
وذلك	B_3
ولا	B_3
ويصدر	B_3
الذي	B_4
إما	B_4
أولاهما	B_4
بما	B_4
حيث	B_4
وإما	B_4
إضافة	B_5
بينها	B_5
حيث	B_5
حيث	B_5
حيث	B_5
ليكون	B_5

وتحدد	B_16
وتضع	B_16
وتلقي	B_16
وذلك	B_16
وعلى	B_16
وللنزيل	B_16
ونص	B_16
وهذا	B_16
ويرجى	B_16
حيث	B_17
حيث	B_17
حيث	B_17
كما	B_17
وعليه	B_17
وقد	B_17
وقد	B_17
ومن	B_17
ويرجى	B_17
رغم	B_18
علما	B_18
مع	B_18
هل	B_18
وبحضور	B_18
وعلاوة	B_18
ولقد	B_18
إذ	B_19
وأسند	B_19
اءً	B_19
وجاء	B_19
وستمحى	B_19
وقد	B_19
وقد	B_19
كما	B_2
كما	B_2
وتتكون	B_2
وذلك	B_2
وفي	B_2

وذلك	E_12
وسنعرض	E_12
وفي	E_12
ولدعم	E_12
وهي	E_12
ووفرت	E_12
كما	E_13
كما	E_13
كما	E_13
ناهيك	E_13
وتقوم	E_13
وتم	E_13
وجاري	E_13
وجاري	E_13
وذلك	E_13
وسيولة	E_13
وهي	E_13
أما	E_14
كما	E_14
كما	E_14
كما	E_14
والمختلف	E_14
وأن	E_14
وتجري	E_14
وتشكيل	E_14
وتقوم	E_14
وفي	E_14
وقد	E_14
وكنواة	E_14
وللجنة	E_14
ويهدف	E_14
فتعمل	E_15
كما	E_15
كما	E_15
كما	E_15
كما	E_15
لذلك	E_15

وتشجيع	E_10
وتشير	E_10
وتواجه	E_10
وجاري	E_10
وجرم	E_10
ومن	E_10
ومن	E_10
كما	E_11
مع	E_11
من	E_11
وبالتالي	E_11
وتسعي	E_11
وتسير	E_11
وتنتشر	E_11
وذلك	E_11
وذلك	E_11
وذلك	E_11
وسوف	E_11
وفي	E_11
وفي	E_11
وفيما	E_11
وقد	E_11
وقد	E_11
وقد	E_11
ومن	E_11
ويتم	E_11
ويقوم	E_11
تلي	E_12
وتم	E_12
أثمر	E_12
كما	E_12
كما	E_12
هذه	E_12
وامتلاكها	E_12
وتشمل	E_12
وتضطلع	E_12
وجاري	E_12

وتختص	B_8
وتم	B_8
وتمارس	B_8
وشدد	B_8
ونؤكد	B_8
ويلتزم	B_8
تم	B_9
حيث	B_9
واستغرق	B_9
وإضافة	B_9
وتم	B_9
وهي	B_9
ويبلغ	B_9
ويتم	B_9
ويجوز	B_9
ويسمح	B_9
كما	E_1
كما	E_1
وبالتالي	E_1
وبالنسبة	E_1
وتنوه	E_1
وفي	E_1
وقد	E_1
وقد	E_1
وقد	E_1
ونضيف	E_1
وهو	E_1
ويضمن	E_1
حيث	E_10
فضلا	E_10
كما	E_10
كما	E_10
كما	E_10
كما	E_10
لتحديث	E_10
وبالنسبة	E_10
وتتواصل	E_10

وبالنسبة	E_20
وتتشكل	E_20
وتعد	E_20
وتعمل	E_20
وذلك	E_20
ورحلة	E_20
وقد	E_20
إذ	E_3
بالإضافة	E_3
تأسيسا	E_3
ثم	E_3
فضلا	E_3
من	E_3
وتطورت	E_3
وتم	E_3
وحظر	E_3
وحق	E_3
وعلي	E_3
وقد	E_3
ولا	E_3
ولا	E_3
ولا	E_3
وللمواطنين	E_3
ونتيجة	E_3
ويتم	E_3
فإذا	E_4
فان	E_4
والمواد	E_4
وبالتالي	E_4
وتعد	E_4
وتم	E_4
وسنوضح	E_4
وسوف	E_4
وقد	E_4
ولا	E_4
ومن	E_4
ومن	E_4

ولاشك	E_17
ويتضح	E_17
ويثار	E_17
ويجوز	E_17
بالإضافة	E_18
معظمهم	E_18
والوزارة	E_18
وتواصل	E_18
وتوالي	E_18
وذلك	E_18
وفي	E_18
وقد	E_18
ومن	E_18
وهو	E_18
إذ تمت	E_19
حيث	E_19
فضلا	E_19
فضلا	E_19
كما	E_19
كما	E_19
كما	E_19
لهذا	E_19
هذا	E_19
وتقوم	E_19
وقد	E_19
وقد	E_19
وكفلت	E_19
وهو	E_19
والتي	E_2
وأن	E_2
وبسقوط	E_2
وهو	E_2
وهو	E_2
بالتالي	E_20
تأكيدا	E_20
كما	E_20
من	E_20

مع	E_15
مع	E_15
من	E_15
وتشير	E_15
وتشير	E_15
وتعمل	E_15
وجارى	E_15
وجارى	E_15
ورفع	E_15
وقد	E_15
وقد	E_15
ويتم	E_15
تم	E_16
تم	E_16
فضلا	E_16
من	E_16
والتي	E_16
وبالرغم	E_16
وبالنسبة	E_16
وتشير	E_16
وتشير	E_16
وتشير	E_16
وجارى	E_16
وخلال	E_16
وذلك	E_16
وذلك	E_16
وقد	E_16
وقد	E_16
ويتم	E_16
فأجاز ا	E_17
كما	E_17
كما	E_17
كما	E_17
وتقدم	E_17
وتلزم	E_17
وذلك	E_17
وقد	E_17

إن	I_1
حيث	I_1
كل	I_1
والتي	I_1
والذي	I_1
واهتماماً	I_1
وبعد	I_1
وتكلل	I_1
وعلى	I_1
وقد	I_1
ولم	I_1
ومن	I_1
ويمكن	I_1
إلا	I_10
على	I_10
وبما	I_10
وتسهم	I_10
وتعمل	I_10
ولذلك	I_10
ولكن	I_10
ولكن	I_10
ولكن	I_10
ولهذا	I_10
ومن	I_10
وهم	I_10
أما	I_11
باستثناء	I_11
حيث	I_11
كما	I_11
كما	I_11
من	I_11
وإنما	I_11
وبسبب	I_11
وبسبب	I_11
وبسبب	I_11
وتركز	I_11
وتقوم	I_11

وتشديد	E_7
وتعد	E_7
وصدر	E_7
وعقب	E_7
وعقب	E_7
وقد	E_7
وقد	E_7
وقد	E_7
وكذا	E_7
قد	E_8
كما	E_8
كما	E_8
كما	E_8
واستحدث	E_8
واستحدث	E_8
وإلزام	E_8
والفئات	E_8
وإنشاء	E_8
وإنشاء	E_8
وفي	E_8
ومنح	E_8
ونص	E_8
ويدخل	E_8
ويهدف	E_8
اتسعت	E_9
تقوم	E_9
كما	E_9
وتشير	E_9
وقد	E_9
وقد	E_9
وقد	E_9
وقد	E_9
وكذلك	E_9
ويتضح	E_9
ويتضمن	E_9
إلا	I_1
التي	I_1

ويصدر	E_4
ويعد	E_4
فضلا	E_5
فضلا	E_5
كما	E_5
كما	E_5
لا	E_5
وبطبيعة	E_5
وبعد	E_5
وتكفل	E_5
وذلك	E_5
وستتضمن	E_5
وعقب	E_5
وقد	E_5
وقد	E_5
وقد	E_5
ومن	E_5
حيث	E_6
كما	E_6
وبالنسبة	E_6
وتلتزم	E_6
وتوالي	E_6
وسبل	E_6
وقد	E_6
وقد	E_6
وقد	E_6
وقد	E_6
وكذا	E_6
وكذلك	E_6
ولذا	E_6
ويعد	E_6
يتولاه	E_6
حيث	E_7
كما	E_7
كما	E_7
وأجريت	E_7
واشتمل	E_7

مع	I_3
معنية	I_3
وبسبب	I_3
وتفعيل	I_3
وحسب	I_3
وعلى	I_3
ويتم	I_3
فإذا	I_4
فإن	I_4
ففي	I_4
كما	I_4
منها	I_4
والاختلاف	I_4
والعمل	I_4
وأورد	I_4
وتتطلب	I_4
وعلى	I_4
وقد	I_4
ويشترط	I_4
ويشتمل	I_4
ويضطلع	I_4
أن	I_5
بالإضافة	I_5
تتعامل	I_5
حيث	I_5
ففي	I_5
كانت	I_5
لذا	I_5
مما	I_5
و2	I_5
واستناداً	I_5
والذي	I_5
وأن	I_5
وتعد	I_5
وتنوعت	I_5
وذلك	I_5
وذلك	I_5

والذين	I_14
وقد	I_14
وقد	I_14
وقد	I_14
وقد	I_14
وقد	I_14
وقد	I_14
وقد	I_14
وهو	I_14
وهي	I_14
ويعتبر	I_14
ويكون	I_14
أما	I_2
أهم	I_2
بالإضافة	I_2
تتكون	I_2
حيث	I_2
حيث	I_2
كان	I_2
كما	I_2
كما	I_2
لذلك	I_2
من	I_2
وبعد	I_2
وتم	I_2
وشرح	I_2
وقد	I_2
وقد	I_2
وكانت	I_2
وله	I_2
ويعني	I_2
تعمل	I_3
حيث	I_3
فيبقى	I_3
كما	I_3
كما	I_3
لهدف	I_3

وعمل	I_11
وقليل	I_11
وكجزء	I_11
ولذلك	I_11
ومع	I_11
ويقوم	I_11
إلا	I_12
بالرغم	I_12
بمشروع	I_12
تلك	I_12
حيث	I_12
حيث	I_12
حيث	I_12
فيما	I_12
كما	I_12
كما	I_12
كما	I_12
كما	I_12
مضافاً	I_12
نابعة	I_12
والجدول	I_12
وتتمثل	I_12
ولكن	I_12
ومن	I_12
ومنهم	I_12
وهي	I_12
ويمثل	I_12
كما	I_13
والتأكد	I_13
والتي	I_13
وتبذل	I_13
وكما	I_13
ومع	I_13
إذ	I_14
الهدف	I_14
على	I_14
لتعليمهن	I_14

وشملت	L_14
وقد	L_14
وأنشئ	L_15
ويبلغ	L_15
بالإضافة	L_16
بالاضافة	L_16
علماً	L_16
مستهدفة	L_16
وتوفر	L_16
تستفيد	L_17
الذي	L_18
تبعه	L_18
مثل	L_18
وكان	L_18
ويستفيد	L_18
القائم	L_19
تحت	L_19
على	L_19
كما	L_19
إلا	L_2
وبالتالي	L_2
وبنتيجة	L_2
بغية	L_20
غير	L_20
كما	L_20
كما	L_20
ومن	L_20
إلاْ	L_21
وتتولى	L_21
القدامى	L_22
برئاسة	L_22
كما	L_22
كمراكز	L_22
وإذا	L_22
وتم	L_22
وتؤمن	L_23
وقد	L_23

ولسوء	I_7
ويتراوح	I_7
ويعيش	I_7
استمر	I_8
بينما	I_8
ففي	I_8
مساعدة	I_8
منهم	I_8
وبسبب	I_8
وخاصة	I_8
وعلى	I_8
وفي	I_8
وفي	I_8
وكان	I_8
وهذه	I_8
ويعرض	I_8
أبرز	I_9
أما	I_9
باستثناء	I_9
بينما	I_9
تعمل	I_9
ما	I_9
ودار	I_9
وعلى	I_9
ويبلغ	I_9
ويتم	I_9
إضافة	L_10
كانت	L_10
مما	L_10
هدفت	L_10
إستهدفت	L_11
كذلك	L_11
بحيث	L_12
وذلك	L_13
وفي	L_13
في	L_14
وأن	L_14

وقد	I_5
وقد	I_5
وكان	I_5
وكان	I_5
ونسبة	I_5
ونصت	I_5
وهناك	I_5
وهي	I_5
بل	I_6
حيث	I_6
حيث	I_6
على	I_6
على	I_6
وتكون	I_6
ورغم	I_6
ونجد	I_6
وندرج	I_6
وهناك	I_6
ويتبين	I_6
ويشترط	I_6
أدى	I_7
بلغت	I_7
تبين	I_7
حيث	I_7
حيث	I_7
في	I_7
كان	I_7
مع	I_7
وأن	I_7
وأن	I_7
وبيّن	I_7
وتابعت	I_7
وتبعاً	I_7
وتشير	I_7
وتعتبر	I_7
وحسب	I_7
وكذلك	I_7

كما	M_15
كما	M_15
كما	M_15
وحرص	M_15
وقد	M_15
وقد	M_15
أما	M_16
كما	M_16
وقد	M_16
وهذه	M_16
على	M_17
غير	M_17
فعلى	M_17
كما	M_17
وبهذا	M_17
وتتمثل	M_17
وقد	M_17
ولهذه	M_17
وهو	M_17
وهي	M_17
في	M_18
كما	M_18
وبالموازاة	M_18
وبلغة	M_18
وفي	M_18
فالمقتضيات	M_2
كما	M_2
كما	M_2
والمغرب	M_2
وبالمعنى	M_2
وستأتي	M_2
وقد	M_2
وقد	M_2
وهكذا	M_2
وهو	M_2
فضلا	M_3
كآلية	M_3

كما	M_1
كما	M_1
كما	M_1
كما	M_1
وإنجاز	M_1
وتتواصل	M_1
وهي	M_1
نفس	M_10
وهكذا	M_10
64	M_11
وستنظم	M_11
وسيمكن	M_11
وهكذا	M_11
حيث	M_12
كما	M_12
ومن	M_12
ويتعلق	M_12
وتتفرع	M_13
حيث	M_13
فإذا	M_13
فإعداد	M_13
فالتمويل	M_13
وستشرف	M_13
ومن	M_13
حيث	M_14
حيث	M_14
كما	M_14
وفي	M_14
وقد	M_14
وقد	M_14
وقد	M_14
وقد	M_14
وقد	M_14
ونسبة	M_14
ويتم	M_14
حيث	M_15
كما	M_15

لكن	L_24
وهو	L_24
كذلك	L_25
كما	L_25
هذا	L_27
وبالتالي	L_27
وأصيب	L_28
وانخفض	L_28
والعمل	L_3
وخضعت	L_3
ويعمل	L_3
يشكل	L_3
كذلك	L_4
إلا	L_5
على	L_5
عملوا	L_5
وقد	L_5
وقد	L_5
ولكنه	L_5
ويوزع	L_5
إلا	L_6
بهدف	L_6
وإلتزمت	L_6
وسيتم	L_6
ومن	L_6
ويبقى	L_6
وأظهر	L_7
والدولة	L_7
وهي	L_7
فالنظرة	L_8
كذلك	L_8
وعلى	L_8
وهذا	L_8
ويحصل	L_8
حيث	L_9
كأن	L_9
وهي	L_9

وكذلك	Y_12
ومنها	Y_12
ويسهم	Y_12
حيث	Y_13
وتقوم	Y_13
وتمثلت	Y_13
وسيتم	Y_13
وهذه	Y_13
احتوت	Y_2
حيث	Y_2
فإذا	Y_2
كما	Y_2
وإذا	Y_2
وإعطاء	Y_2
والمعروض	Y_2
وقد	Y_2
وقد	Y_2
وكذا	Y_2
ومن	Y_2
وهي	Y_2
ويحكم	Y_2
يتضمن	Y_2
استهدفت	Y_3
تم	Y_3
تم	Y_3
وذلك	Y_3
بأنه	Y_4
وقد	Y_4
وقدمت	Y_4
وكذلك	Y_4
وكذلك	Y_4
حيث	Y_5
وتقوم	Y_5
وسن	Y_5
وعند	Y_5
وقد	Y_5
ولا	Y_5

ويتوقف	M_9
ويستفيد	M_9
ويمكن	M_9
العديد	Y_1
حيث	Y_1
حيث	Y_1
لكن	Y_1
والمادة	Y_1
وعلى	Y_1
ونصت	Y_1
ويجب	Y_1
إلا	Y_10
كما	Y_10
وعلى	Y_10
وفي	Y_10
وقد	Y_10
وقد	Y_10
وقد	Y_10
وقد	Y_10
وهذا	Y_10
وهذا	Y_10
ويتولى	Y_10
ويجب	Y_10
ويجب	Y_10
ويحدد	Y_10
حيث	Y_11
حيث	Y_11
شمل	Y_11
والخروج	Y_11
وتم	Y_11
ويتم	Y_11
حيث	Y_12
كما	Y_12
كما	Y_12
وأيضاً	Y_12
وتحقيقاً	Y_12
وتعمل	Y_12

كما	M_3
وتصل	M_3
وتواصل	M_3
ودعم	M_3
كما	M_5
وفي	M_5
ويندرج	M_5
في	M_6
كما	M_6
واعتباراً	M_6
وخلال	M_6
وفتحت	M_6
وقد	M_6
ويعتبر	M_6
كما	M_7
كما	M_7
كما	M_7
هذا	M_7
وبالمقابل	M_7
ورغم	M_7
أما	M_8
كما	M_8
وقد	M_8
وقد	M_8
ولكنه	M_8
واتخذ	M_9
أما	M_9
فإن	M_9
فحسب	M_9
في	M_9
في	M_9
كما	M_9
وقد	M_9
وقد	M_9
وقد	M_9
وقد	M_9
ويتمتع	M_9

ولا	Y_5
ويجوز	Y_5
أحد	Y_6
متضمنة	Y_6
وذلك	Y_6
وسيتم	Y_6
وعرضه	Y_6
وفي	Y_6
وكذا	Y_6
حيث	Y_7
فإذا	Y_7
كما	Y_7

كما	Y_7
وإذا	Y_7
وإذا	Y_7
ونورد	Y_7
وهي	Y_7
ويجوز	Y_7
فإذا	Y_8
وتكون	Y_8
وقد	Y_8
وهذه	Y_8
ويعاقب	Y_8
ويعتبر	Y_8

ويعد	Y_8
يتضمن	Y_8
أنشأت	Y_9
رغم	Y_9
فيكون	Y_9
كما	Y_9
وإجمالي	Y_9
وإذا	Y_9
وتسري	Y_9
ويحكم	Y_9

Anexo IV

Relación de las 983 *stopwords* localizadas en la traducción del árabe al inglés, con datos desglosados por documentos y segmentos (documento_segmento).

SEG.	STOPW.
B_1	وتم
B_1	حيث
B_1	حيث
B_1	وقاموا
B_1	وقاموا
B_1	وتمكنوا
B_1	وقد
B_1	على
B_1	واستمرت
B_1	وقد
B_1	وعلى
B_10	حيث
B_10	حيث
B_10	ويكون
B_10	وألا
B_10	ووجود
B_10	إذ
B_10	ويتم
B_10	ونشير
B_10	ويتم
B_10	كما
B_10	وعلى
B_10	ويتم
B_10	ونشير

B_10	حيث
B_10	وذلك
B_11	وترسلها
B_11	واتخاذ
B_11	وعليه
B_11	حيث
B_11	ونص
B_11	وتضع
B_11	وللنزيل
B_11	وعلى
B_11	وتلقي
B_11	ويخطر
B_11	ووضعها
B_12	وهي
B_12	وجرى
B_12	ويتضمن
B_12	وقد
B_12	وتعمل
B_12	فضلاً
B_12	وطعنت
B_12	إضافة
B_12	ويظهر
B_12	وعلى
B_13	إضافة
B_13	وتقر

B_13	فصدر
B_13	على
B_13	وهي
B_13	تم
B_14	ومن
B_14	بما
B_14	ومن
B_15	وأنه
B_15	ولا
B_15	ويتم
B_15	بالإضافة
B_15	وقد
B_15	وفي
B_15	حيث
B_15	وهناك
B_15	كما
B_15	حيث
B_15	على
B_15	ففي
B_15	حيث
B_15	وأنه
B_15	فضلاً
B_15	وعليه
B_15	وفي
B_15	وفي

فشكل	B_3
وذلك	B_3
ويجب	B_3
وإحالة	B_3
وقد	B_3
إضافة	B_3
وهو	B_3
كما	B_3
ويصدر	B_3
ولا	B_3
وهي	B_4
وهما	B_4
وصرفت	B_4
أولاهما	B_4
حيث	B_4
إضافة	B_5
يرجى	B_5
مما	B_5
وذلك	B_5
حيث	B_5
وتتمثل	B_5
وانتهت	B_5
حيث	B_5
إضافة	B_5
وباشرت	B_5
وتقوم	B_5
وكذلك	B_6
ووفدوا	B_6
كما	B_6
والأكاديمية	B_6
حيث	B_6
وامتدت	B_6
وهما	B_7
وهذا	B_7
ولقد	B_7
وذلك	B_7
ولقد	B_7

نظراً	B_19
وبالنسبة	B_19
وقد	B_19
وأسند	B_19
وستمحى	B_19
اءً	B_19
وقد	B_19
وجاء	B_19
إذ	B_19
وذلك	B_2
وهما	B_2
تعمل	B_2
بفحص	B_2
وهي	B_2
وهي	B_2
وتتكون	B_2
وقد	B_2
وذلك	B_2
ومن	B_2
وقد	B_2
كما	B_2
وللأمانة	B_2
وفي	B_2
إلا	B_20
وصرفت	B_20
وحولت	B_20
وبالإضافة	B_20
وقد	B_20
حيث	B_20
وقد	B_20
حيث	B_20
وقد	B_20
حيث	B_20
وتم	B_21
وورد	B_21
وهي	B_21
وعلاوة	B_21

كما	B_15
ولكن	B_16
وتختص	B_16
وعليه	B_16
وتضع	B_16
وللنزيل	B_16
وعلى	B_16
ونص	B_16
ويرجى	B_16
حيث	B_16
وتلقي	B_16
كما	B_16
كما	B_16
وكذلك	B_17
وعليه	B_17
ومن	B_17
وقد	B_17
كما	B_17
حيث	B_17
وقد	B_17
حيث	B_17
ويرجى	B_17
حيث	B_17
بينما	B_18
نظراً	B_18
حيث	B_18
وتبين	B_18
ومنها	B_18
علما	B_18
هل	B_18
وعلاوة	B_18
ولقد	B_18
رغم	B_18
وكذلك	B_19
فيما	B_19
وهي	B_19
واستفاد	B_19

هذه	E_12
وتضطلع	E_12
أثمر	E_12
وهي	E_12
ووفرت	E_12
وفى	E_12
ويشمل	E_13
حيث	E_13
وتدر	E_13
وذلك	E_13
وتقوم	E_13
كما	E_13
وجاري	E_13
وتم	E_13
وجاري	E_13
كما	E_13
ناهيك	E_13
تتعدد	E_14
24	E_14
بمعني	E_14
حيث	E_14
وقد	E_14
وأن	E_14
أما	E_14
وتقوم	E_14
وتجري	E_14
وكنواة	E_14
وتشكيل	E_14
والمختلف	E_14
ويهدف	E_14
وللجنة	E_14
كما	E_14
وقد	E_15
ويركز	E_15
فتعمل	E_15
كما	E_15
كما	E_15

وقد	E_1
وفى	E_1
وقد	E_1
وهو	E_1
كما	E_1
والقروض	E_10
وذلك	E_10
وجاري	E_10
فقد	E_10
وتشير	E_10
حيث	E_10
وتواجه	E_10
ومن	E_10
وجاري	E_10
وتشجيع	E_10
وتتواصل	E_10
كما	E_10
كما	E_10
وجرم	E_10
كما	E_10
كما	E_10
وفي	E_11
وقد	E_11
وقد	E_11
كما	E_11
ويتم	E_11
وفيما	E_11
وفي	E_11
وتسعي	E_11
وتنتشر	E_11
وسوف	E_11
وتسير	E_11
وذلك	E_11
وإنشاء	E_12
في	E_12
والمأوي	E_12
ولدعم	E_12

ولقد	B_7
وجاري	B_7
كما	B_7
وفي	B_7
وإجراء	B_8
وهي	B_8
وتباشر	B_8
هذا	B_8
كما	B_8
وتختص	B_8
وتم	B_8
كما	B_8
بالإضافة	B_8
وأي	B_8
وتمارس	B_8
وتختص	B_8
وشدد	B_8
بادرت	B_8
فقد	B_8
ونؤكد	B_8
وقدم	B_9
ويجوز	B_9
ويبلغ	B_9
ويتم	B_9
وإضافة	B_9
وهي	B_9
وتم	B_9
تم	B_9
واستغرق	B_9
حيث	B_9
تأسيسا	E_1
لما	E_1
كما	E_1
وبالنسبة	E_1
وقد	E_1
وبالتالي	E_1
ويضمن	E_1

كما	E_20
بالتالي	E_20
من	E_20
وأعقب	E_3
وذلك	E_3
ومن	E_3
باعتبار	E_3
وفي	E_3
بالنظر	E_3
تركت	E_3
ونتيجة	E_3
من	E_3
وتم	E_3
بالإضافة	E_3
وقد	E_3
إذ	E_3
وحق	E_3
ولا	E_3
وللمواطنين	E_3
وحظر	E_3
ولا	E_3
ولا	E_3
وبالتالي	E_4
فإذا	E_4
ولا	E_4
ويصدر	E_4
فان	E_4
وقد	E_4
ومن	E_4
وتم	E_4
وتعد	E_4
ويعد	E_4
وسنوضح	E_4
ومن	E_4
وسوف	E_4
الأمر	E_5
في	E_5

وتوالي	E_18
وقد	E_18
بالإضافة	E_18
وذلك	E_18
وفي	E_18
وتواصل	E_18
ومن	E_18
معظمهم	E_18
وهي	E_19
في	E_19
بحكم	E_19
حيث	E_19
وتقوم	E_19
هذا	E_19
وقد	E_19
لهذا	E_19
وكفلت	E_19
فضلا	E_19
فضلا	E_19
وقد	E_19
إذ	E_19
كما	E_19
وبسقوط	E_2
وذلك	E_2
وأن	E_2
وهو	E_2
وبسقوط	E_2
وهو	E_2
فأن	E_20
ومواكبة	E_20
ويمتع	E_20
ومن	E_20
وتعد	E_20
وتتشكل	E_20
ورحلة	E_20
وذلك	E_20
وبالنسبة	E_20

من	E_15
وجارى	E_15
وقد	E_15
وجارى	E_15
وتعمل	E_15
وتشير	E_15
وقد	E_15
مع	E_15
ويتم	E_15
وتشير	E_15
بنسبة	E_15
وتفعيل	E_16
وبالرغم	E_16
وتشير	E_16
وتشير	E_16
وتشير	E_16
وذلك	E_16
وذلك	E_16
ويتم	E_16
وبالنسبة	E_16
وقد	E_16
وإن	E_17
ويتولى	E_17
وتقدم	E_17
ويجوز	E_17
فأجاز	E_17
ويثار	E_17
ولا شك	E_17
كما	E_17
وتلزم	E_17
وقد	E_17
ومسابقات	E_18
وتزايد	E_18
كما	E_18
وباعتبار	E_18
وهو	E_18
والوزارة	E_18

وبعد	I_1
وقد	I_1
واهتماماً	I_1
ويمكن	I_1
وعلى	I_1
إن	I_1
ولم	I_1
فهي	I_10
وتستخدم	I_10
وعلى	I_10
وهم	I_10
ولذلك	I_10
ولكن	I_10
إلا	I_10
ولكن	I_10
ولهذا	I_10
وبما	I_10
باستثناء	I_11
وبسبب	I_11
أما	I_11
وتقوم	I_11
ولذلك	I_11
وتركز	I_11
كما	I_11
ويقوم	I_11
وبسبب	I_11
وبسبب	I_11
كما	I_11
ومع	I_11
وبسبب	I_12
ولكن	I_12
وتتمثل	I_12
فيما	I_12
حيث	I_12
حيث	I_12
مضافاً	I_12
كما	I_12

وقد	E_7
وتعد	E_7
واشتمل	E_7
وصدر	E_7
وتشديد	E_7
كما	E_7
وذلك	E_8
والتي	E_8
واستحدث	E_8
واستحدث	E_8
وإنشاء	E_8
كما	E_8
كما	E_8
ومنح	E_8
وإلزام	E_8
والفئات	E_8
كما	E_8
وفي	E_8
وبدعم	E_9
من	E_9
وقد	E_9
وقد	E_9
تقوم	E_9
اتسعت	E_9
وقد	E_9
كما	E_9
ويتضح	E_9
وكذلك	E_9
وخاصة	E_9
وفي	I_1
وأتم	I_1
إلا	I_1
ومن	I_1
حيث	I_1
والتي	I_1
كل	I_1
وتكلل	I_1

حيث	E_5
والتي	E_5
ما	E_5
الأمر	E_5
وعقب	E_5
وبطبيعة	E_5
وبعد	E_5
وذلك	E_5
وقد	E_5
فضلا	E_5
ومن	E_5
وقد	E_5
وستتضمن	E_5
ذلك	E_6
وعدم	E_6
والتوجيه	E_6
كما	E_6
وقد	E_6
ويعد	E_6
وبالنسبة	E_6
وقد	E_6
وتلتزم	E_6
وتوالي	E_6
وقد	E_6
يتولاه	E_6
وقد	E_6
وسبل	E_6
حيث	E_6
بتخصيص	E_7
إذ	E_7
لتصل	E_7
مع	E_7
وعقب	E_7
حيث	E_7
وعقب	E_7
وقد	E_7
وقد	E_7

ويشترط	I_6
ويتبين	I_6
وهناك	I_6
على	I_6
بالإضافة	I_7
علماً	I_7
وهذا	I_7
كان	I_7
وأن	I_7
وتعتبر	I_7
وتابعت	I_7
وتشير	I_7
وكذلك	I_7
مع	I_7
ويتراوح	I_7
وتبعاً	I_7
ويعيش	I_7
ولسوء	I_7
وأن	I_7
أدى	I_7
وبيّن	I_7
وتراوحت	I_8
وتهدف	I_8
منهم	I_8
بينما	I_8
وكان	I_8
وعلى	I_8
وهذه	I_8
استمر	I_8
وفي	I_8
ولم	I_9
ويتم	I_9
ويبلغ	I_9
أبرز	I_9
تعمل	I_9
أما	I_9
وقد	L_10

تعمل	I_3
مع	I_3
فيبقى	I_3
حيث	I_3
والمادة	I_4
وهذا	I_4
ونظم	I_4
وقد	I_4
فإذا	I_4
ويشترط	I_4
والاختلاف	I_4
وأورد	I_4
والعمل	I_4
وعلى	I_4
ويشتمل	I_4
ويضطلع	I_4
وانسجاماً	I_5
كما	I_5
فقد	I_5
وكان	I_5
وقد	I_5
وكان	I_5
واستناداً	I_5
وقد	I_5
كانت	I_5
لذا	I_5
وهي	I_5
وهناك	I_5
وذلك	I_5
وتنوعت	I_5
ثم	I_6
لاحظنا	I_6
وتكون	I_6
ورغم	I_6
وندرج	I_6
ونجد	I_6
حيث	I_6

حيث	I_12
كما	I_12
كما	I_12
كما	I_12
والجدول	I_12
وتبذل	I_13
كما	I_13
ومع	I_13
وتم	I_14
ويعتبر	I_14
وقد	I_14
وقد	I_14
وقد	I_14
وقد	I_14
ويكون	I_14
والذين	I_14
وقد	I_14
وتم	I_2
وكانت	I_2
وقد	I_2
أما	I_2
أهم	I_2
وشرح	I_2
ويعني	I_2
كان	I_2
لذلك	I_2
كما	I_2
بالإضافة	I_2
وقد	I_2
حيث	I_2
ولكنها	I_3
مهمتها	I_3
ترتبط	I_3
ويتم	I_3
كما	I_3
وبسبب	I_3
كما	I_3

مما	M_15
حيث	M_15
كما	M_15
وقد	M_15
كما	M_15
كما	M_15
كما	M_15
وقد	M_15
وفي	M_16
بفتح	M_16
بحيث	M_16
مقابل	M_16
إذ	M_16
وتغيير	M_16
لا	M_16
وقد	M_16
وهذه	M_16
كما	M_16
أما	M_16
ليمثل	M_17
وهو	M_17
وهم	M_17
الذي	M_17
كل	M_17
الذي	M_17
فعلى	M_17
وهي	M_17
وهو	M_17
وتتمثل	M_17
على	M_17
وقد	M_17
كما	M_17
ولهذه	M_17
وبهذا	M_17
غير	M_17
وفي	M_18
في	M_18

والتي	M_12
الأمر	M_12
إلى	M_12
وتعويض	M_12
إضافة	M_12
ومن	M_12
كما	M_12
ويتعلق	M_12
حيث	M_12
فالتغطية	M_13
وسوف	M_13
ويهم	M_13
ووضع	M_13
فالتمويل	M_13
ومن	M_13
حيث	M_13
فإعداد	M_13
فإذا	M_13
أهمها	M_13
وما	M_14
حيث	M_14
وتأسيس	M_14
والذي	M_14
وقد	M_14
حيث	M_14
ويتم	M_14
حيث	M_14
وقد	M_14
وقد	M_14
وفي	M_14
كما	M_14
وقد	M_14
مع	M_15
تتوزع	M_15
فإنه	M_15
فإن	M_15
وفتحنا	M_15

هدفت	L_10
ويستفيد	L_18
وتتولى	L_21
وسنداً	L_25
وصولاً	L_3
وخضعت	L_3
كما	M_1
مع	M_1
أن	M_1
بإعطاء	M_1
حيث	M_1
تستثمر	M_1
لما	M_1
تمثلت	M_1
وتتواصل	M_1
وهي	M_1
كما	M_1
كما	M_1
كما	M_1
كما	M_10
كما	M_10
وتضاعف	M_10
حيث	M_10
نفس	M_10
ومتابعة	M_11
واعتماد	M_11
حيث	M_11
مما	M_11
وقال	M_11
بالإضافة	M_11
حيث	M_11
وهكذا	M_11
وستنظم	M_11
وسيمكن	M_11
مما	M_12
حيث	M_12
ستعمل	M_12

تساهم	M_7
وترسيخاً	M_7
واحتراماً	M_7
فقد	M_7
فما	M_7
حيث	M_7
حيث	M_7
كما	M_7
وبالمقابل	M_7
كما	M_7
كما	M_7
علماً	M_8
كان	M_8
فهو	M_8
إلا	M_8
وقد	M_8
الشيء	M_8
كما	M_8
ولكنه	M_8
أما	M_8
وقد	M_8
وهي	M_9
وذلك	M_9
إضافة	M_9
وطبقاً	M_9
وأن	M_9
وبناء	M_9
الذي	M_9
كما	M_9
ويتوقف	M_9
ويستفيد	M_9
أما	M_9
ويتمتع	M_9
وقد	M_9
وقد	M_9
ويمكن	M_9
وقد	M_9

تضمن	M_4
وذلك	M_4
بتحويل	M_4
كما	M_4
بالإضافة	M_4
حيث	M_4
حيث	M_5
والذي	M_5
حيث	M_5
أو	M_5
اعتباراً	M_5
حيث	M_5
وإنما	M_5
إذ	M_5
لكونه	M_5
الذي	M_5
كما	M_5
وفي	M_5
ويندرج	M_5
ونصت	M_6
وخولت	M_6
ووسعت	M_6
وذلك	M_6
إذ	M_6
وارتفعت	M_6
حيث	M_6
وقد	M_6
واعتباراً	M_6
كما	M_6
وقد	M_6
وخلال	M_6
في	M_6
ويعتبر	M_6
وفتحت	M_6
كان	M_7
ويرجع	M_7
الهدف	M_7

غطت	M_18
وتغطية	M_18
والذي	M_18
كما	M_18
وبالموازاة	M_18
وفي	M_18
وبلغة	M_18
صدر	M_18
باعتباره	M_2
مع	M_2
على	M_2
وهي	M_2
الذين	M_2
وعاش	M_2
وتمتعت	M_2
وستأتي	M_2
كما	M_2
وقد	M_2
كما	M_2
وهكذا	M_2
وقد	M_2
وهو	M_2
والمغرب	M_2
وبالمعنى	M_2
فالمقتضيات	M_2
بحيث	M_3
ودعم	M_3
كما	M_3
ظهرت	M_3
حيث	M_3
وتقديم	M_3
وتصل	M_3
وتواصل	M_3
حيث	M_4
وكذا	M_4
لتشمل	M_4
الذين	M_4

وقد	Y_2
احتوت	Y_2
وكذا	Y_2
تم	Y_3
وأوراق	Y_3
وكذا	Y_4
وتحديد	Y_4
وكذلك	Y_4
وقدمت	Y_4
ويحظر	Y_5
وكذا	Y_5
وسن	Y_5
وتقوم	Y_5
ويجوز	Y_5
حيث	Y_5
وقد	Y_5
ولا	Y_5
ولا	Y_5
وعند	Y_5
ورفع	Y_6
وإدخال	Y_6
ومن	Y_6
وزيارة	Y_6
لتلمس	Y_6
والخروج	Y_6
وسيتم	Y_6
أحد	Y_6
ولكن	Y_7
ويحظر	Y_7
وذلك	Y_7
حيث	Y_7
حيث	Y_7
كما	Y_7
فإذا	Y_7
وإذا	Y_7
وإذا	Y_7
كما	Y_7

وإحالة	Y_11
ومناقشة	Y_11
بهدف	Y_11
والخروج	Y_11
والخروج	Y_11
ويتم	Y_11
حيث	Y_11
حيث	Y_11
وتم	Y_11
شمل	Y_11
إذا	Y_12
وتأمين	Y_12
وكان	Y_12
كما	Y_12
وتعمل	Y_12
حيث	Y_12
كما	Y_12
ويسهم	Y_12
وتحقيقاً	Y_12
وأيضاً	Y_12
وانخفضت	Y_13
بعض	Y_13
وانتهت	Y_13
وتقوم	Y_13
وتمثلت	Y_13
وسيتم	Y_13
حيث	Y_13
وهذه	Y_13
ويحكم	Y_2
ويحظر	Y_2
والمعروض	Y_2
ومن	Y_2
فإذا	Y_2
وإذا	Y_2
وقد	Y_2
حيث	Y_2
وإعطاء	Y_2

في	M_9
فإن	M_9
واتخذ	M_9
وقد	M_9
ولم	Y_1
ولا	Y_1
حيث	Y_1
ويجب	Y_1
والمادة	Y_1
حيث	Y_1
لكن	Y_1
وعلى	Y_1
ونصت	Y_1
وبالتالي	Y_10
وكذا	Y_10
والتي	Y_10
ويثبت	Y_10
فإذا	Y_10
ويتم	Y_10
وأيضاً	Y_10
وعلى	Y_10
ويجب	Y_10
وفي	Y_10
ويحدد	Y_10
وقد	Y_10
وهذا	Y_10
وقد	Y_10
ويجب	Y_10
ويتولى	Y_10
إلا	Y_10
وهذا	Y_10
وقد	Y_10
وقد	Y_10
ومنها	Y_11
بناء	Y_11
يتضمن	Y_11
وإعطاء	Y_11

فيكون	Y_9
كما	Y_9
وتسري	Y_9
رغم	Y_9
وإذا	Y_9
وإجمالي	Y_9
ويحكم	Y_9

وهذه	Y_8
الشخص	Y_9
ويبين	Y_9
وحرصاً	Y_9
ولا	Y_9
تشمل	Y_9
لكنها	Y_9

ويجوز	Y_7
وتكون	Y_8
والتي	Y_8
وتكون	Y_8
ويعد	Y_8
فإذا	Y_8
ويعتبر	Y_8

Anexo V

Relación de las 647 *stopwords* comunes localizadas en las traducciones del árabe al español y al inglés con datos desglosados por documentos y segmentos (documento_segmento).

SEG.	STOPW.
B_1	على
B_1	واستمرت
B_1	وعلى
B_1	وقد
B_1	وقد
B_10	إذ
B_10	حيث
B_10	كما
B_10	وذلك
B_10	وعلى
B_10	ونشير
B_10	ونشير
B_10	ويتم
B_10	ويتم
B_10	ويتم
B_11	حيث
B_11	وتضع
B_11	وتلقي
B_11	وعلى
B_11	وللنزيل
B_11	ونص

B_11	ووضعها
B_11	ويخطر
B_12	إضافة
B_12	فضلاً
B_12	وتعمل
B_12	وطعنت
B_12	وعلى
B_12	وقد
B_12	ويظهر
B_13	تم
B_13	على
B_13	فصدر
B_13	وهي
B_14	ومن
B_15	حيث
B_15	حيث
B_15	حيث
B_15	على
B_15	فضلاً
B_15	ففي
B_15	كما
B_15	كما

B_15	وأنه
B_15	وعليه
B_15	وفي
B_15	وفي
B_15	وهناك
B_16	حيث
B_16	كما
B_16	كما
B_16	وتضع
B_16	وتلقي
B_16	وعلى
B_16	وللنزيل
B_16	ونص
B_16	ويرجى
B_17	حيث
B_17	حيث
B_17	حيث
B_17	كما
B_17	وعليه
B_17	وقد
B_17	وقد
B_17	ومن

وتم	B_9
وهي	B_9
ويبلغ	B_9
ويتم	B_9
ويجوز	B_9
كما	E_1
كما	E_1
وبالتالي	E_1
وبالنسبة	E_1
وفي	E_1
وقد	E_1
وقد	E_1
وقد	E_1
وهو	E_1
ويضمن	E_1
حيث	E_10
كما	E_10
كما	E_10
كما	E_10
كما	E_10
وتتواصل	E_10
وتشجيع	E_10
وتشير	E_10
وتواجه	E_10
وجاري	E_10
وجرم	E_10
ومن	E_10
كما	E_11
وتسعي	E_11
وتسير	E_11
وتنتشر	E_11
وذلك	E_11
وسوف	E_11
وفي	E_11
وفي	E_11
وفيما	E_11
وقد	E_11

أولاهما	B_4
حيث	B_4
إضافة	B_5
حيث	B_5
حيث	B_5
وانتهت	B_5
وباشرت	B_5
وتتمثل	B_5
وتقوم	B_5
وذلك	B_5
حيث	B_6
كما	B_6
والأكاديمية	B_6
وامتدت	B_6
كما	B_7
وجاري	B_7
وذلك	B_7
وفي	B_7
ولقد	B_7
ولقد	B_7
ولقد	B_7
بادرت	B_8
بالإضافة	B_8
فقد	B_8
كما	B_8
كما	B_8
وأي	B_8
وتختص	B_8
وتختص	B_8
وتم	B_8
وتمارس	B_8
وشدد	B_8
ونؤكد	B_8
تم	B_9
حيث	B_9
واستغرق	B_9
وإضافة	B_9

ويرجى	B_17
رغم	B_18
علما	B_18
هل	B_18
وعلاوة	B_18
ولقد	B_18
إذ	B_19
وأسند	B_19
اءً	B_19
وجاء	B_19
وستمحى	B_19
وقد	B_19
وقد	B_19
كما	B_2
وتتكون	B_2
وذلك	B_2
وفي	B_2
وقد	B_2
وقد	B_2
وللأمانة	B_2
ومن	B_2
وهي	B_2
حيث	B_20
حيث	B_20
حيث	B_20
وبالإضافة	B_20
وحولت	B_20
وقد	B_20
وقد	B_20
وقد	B_20
وتم	B_21
وعلاوة	B_21
وهي	B_21
وورد	B_21
كما	B_3
ولا	B_3
ويصدر	B_3

وقد	E_19
وقد	E_19
وكفلت	E_19
وأن	E_2
وبسقوط	E_2
وهو	E_2
وهو	E_2
بالتالي	E_20
كما	E_20
من	E_20
وبالنسبة	E_20
وتتشكل	E_20
وتعد	E_20
وذلك	E_20
ورحلة	E_20
إذ	E_3
بالإضافة	E_3
من	E_3
وتم	E_3
وحظر	E_3
وحق	E_3
وقد	E_3
ولا	E_3
ولا	E_3
ولا	E_3
وللمواطنين	E_3
ونتيجة	E_3
فإذا	E_4
فان	E_4
وبالتالي	E_4
وتعد	E_4
وتم	E_4
وسنوضح	E_4
وسوف	E_4
وقد	E_4
ولا	E_4
ومن	E_4

وقد	E_15
وقد	E_15
ويتم	E_15
وبالرغم	E_16
وبالنسبة	E_16
وتشير	E_16
وتشير	E_16
وتشير	E_16
وذلك	E_16
وذلك	E_16
وقد	E_16
ويتم	E_16
فأجاز	E_17
كما	E_17
وتقدم	E_17
وتلزم	E_17
وقد	E_17
ولاشك	E_17
ويثار	E_17
ويجوز	E_17
بالإضافة	E_18
معظمهم	E_18
والوزارة	E_18
وتواصل	E_18
وتوالي	E_18
وذلك	E_18
وفي	E_18
وقد	E_18
ومن	E_18
وهو	E_18
إذ	E_19
فضلا	E_19
فضلا	E_19
كما	E_19
لهذا	E_19
هذا	E_19
وتقوم	E_19

وقد	E_11
ويتم	E_11
أثمر	E_12
هذه	E_12
وتضطلع	E_12
وفى	E_12
ولدعم	E_12
وهي	E_12
ووفرت	E_12
كما	E_13
كما	E_13
ناهيك	E_13
وتقوم	E_13
وتم	E_13
وجاري	E_13
وجاري	E_13
وذلك	E_13
أما	E_14
كما	E_14
والمختلف	E_14
وأن	E_14
وتجري	E_14
وتشكيل	E_14
وتقوم	E_14
وكنواة	E_14
وللجنة	E_14
ويهدف	E_14
فتعمل	E_15
كما	E_15
كما	E_15
مع	E_15
من	E_15
وتشير	E_15
وتشير	E_15
وتعمل	E_15
وجارى	E_15
وجارى	E_15

كما	I_11
كما	I_11
وبسبب	I_11
وبسبب	I_11
وبسبب	I_11
وتركز	I_11
وتقوم	I_11
ولذلك	I_11
ومع	I_11
ويقوم	I_11
حيث	I_12
حيث	I_12
حيث	I_12
فيما	I_12
كما	I_12
كما	I_12
كما	I_12
كما	I_12
مضافاً	I_12
والجدول	I_12
وتتمثل	I_12
ولكن	I_12
كما	I_13
وتبذل	I_13
ومع	I_13
والذين	I_14
وقد	I_14
وقد	I_14
وقد	I_14
وقد	I_14
وقد	I_14
ويعتبر	I_14
ويكون	I_14
أما	I_2
أهم	I_2
بالإضافة	I_2
حيث	I_2

كما	E_8
واستحدث	E_8
واستحدث	E_8
وإلزام	E_8
والفئات	E_8
وإنشاء	E_8
وفي	E_8
ومنح	E_8
اتسعت	E_9
تقوم	E_9
كما	E_9
وقد	E_9
وقد	E_9
وقد	E_9
وكذلك	E_9
ويتضح	E_9
إن	I_1
حيث	I_1
كل	I_1
والتي	I_1
واهتماماً	I_1
وبعد	I_1
وتكلل	I_1
وعلى	I_1
وقد	I_1
ولم	I_1
ومن	I_1
ويمكن	I_1
إلا	I_10
وبما	I_10
ولذلك	I_10
ولكن	I_10
ولكن	I_10
ولهذا	I_10
وهم	I_10
أما	I_11
باستثناء	I_11

ومن	E_4
ويصدر	E_4
ويعد	E_4
فضلا	E_5
وبطبيعة	E_5
وبعد	E_5
وذلك	E_5
وستتضمن	E_5
وعقب	E_5
وقد	E_5
وقد	E_5
ومن	E_5
حيث	E_6
كما	E_6
وبالنسبة	E_6
وتلتزم	E_6
وتوالي	E_6
وسبل	E_6
وقد	E_6
وقد	E_6
وقد	E_6
وقد	E_6
ويعد	E_6
يتولاه	E_6
حيث	E_7
كما	E_7
واشتمل	E_7
وتشديد	E_7
وتعد	E_7
وصدر	E_7
وعقب	E_7
وعقب	E_7
وقد	E_7
وقد	E_7
وقد	E_7
كما	E_8
كما	E_8

وخضعت	L_3
كما	M_1
كما	M_1
كما	M_1
وتتواصل	M_1
وهي	M_1
نفس	M_10
وستنظم	M_11
وسيمكن	M_11
وهكذا	M_11
حيث	M_12
كما	M_12
ومن	M_12
ويتعلق	M_12
حيث	M_13
فإذا	M_13
فإعداد	M_13
فالتمويل	M_13
ومن	M_13
حيث	M_14
حيث	M_14
كما	M_14
وفي	M_14
وقد	M_14
وقد	M_14
وقد	M_14
وقد	M_14
ويتم	M_14
حيث	M_15
كما	M_15
كما	M_15
كما	M_15
كما	M_15
وقد	M_15
وقد	M_15
أما	M_16
كما	M_16

على	I_6
وتكون	I_6
ورغم	I_6
ونجد	I_6
وندرج	I_6
وهناك	I_6
ويتبين	I_6
ويشترط	I_6
أدى	I_7
كان	I_7
مع	I_7
وأن	I_7
وأن	I_7
وبيّن	I_7
وتابعت	I_7
وتبعاً	I_7
وتشير	I_7
وتعتبر	I_7
وكذلك	I_7
ولسوء	I_7
ويتراوح	I_7
ويعيش	I_7
استمر	I_8
بينما	I_8
منهم	I_8
وعلى	I_8
وفي	I_8
وكان	I_8
وهذه	I_8
أبرز	I_9
أما	I_9
تعمل	I_9
ويبلغ	I_9
ويتم	I_9
هدفت	L_10
ويستفيد	L_18
وتتولى	L_21

كان	I_2
كما	I_2
لذلك	I_2
وتم	I_2
وشرح	I_2
وقد	I_2
وقد	I_2
وكانت	I_2
ويعني	I_2
تعمل	I_3
حيث	I_3
فيبقى	I_3
كما	I_3
كما	I_3
مع	I_3
وبسبب	I_3
ويتم	I_3
فإذا	I_4
والاختلاف	I_4
والعمل	I_4
وأورد	I_4
وعلى	I_4
ويشترط	I_4
ويشتمل	I_4
ويضطلع	I_4
كانت	I_5
لذا	I_5
واستناداً	I_5
وتنوعت	I_5
وذلك	I_5
وقد	I_5
وقد	I_5
وكان	I_5
وكان	I_5
وهناك	I_5
وهي	I_5
حيث	I_6

وهذا	Y_10
ويتولى	Y_10
ويجب	Y_10
ويجب	Y_10
ويحدد	Y_10
حيث	Y_11
حيث	Y_11
شمل	Y_11
والخروج	Y_11
وتم	Y_11
ويتم	Y_11
حيث	Y_12
كما	Y_12
كما	Y_12
وأيضاً	Y_12
وتحقيقاً	Y_12
وتعمل	Y_12
ويسهم	Y_12
حيث	Y_13
وتقوم	Y_13
وتمثلت	Y_13
وسيتم	Y_13
وهذه	Y_13
احتوت	Y_2
حيث	Y_2
فإذا	Y_2
وإذا	Y_2
وإعطاء	Y_2
والمعروض	Y_2
وقد	Y_2
وقد	Y_2
وكذا	Y_2
ومن	Y_2
وقدمت	Y_4
وكذلك	Y_4
حيث	Y_5
وتقوم	Y_5

ويعتبر	M_6
كما	M_7
كما	M_7
كما	M_7
وبالمقابل	M_7
أما	M_8
كما	M_8
وقد	M_8
ولكنه	M_8
واتخذ	M_9
أما	M_9
فإن	M_9
في	M_9
كما	M_9
وقد	M_9
وقد	M_9
وقد	M_9
وقد	M_9
ويتمتع	M_9
ويتوقف	M_9
ويستفيد	M_9
ويمكن	M_9
حيث	Y_1
حيث	Y_1
لكن	Y_1
والمادة	Y_1
وعلى	Y_1
ونصت	Y_1
ويجب	Y_1
إلا	Y_10
وعلى	Y_10
وفي	Y_10
وقد	Y_10
وقد	Y_10
وقد	Y_10
وقد	Y_10
وهذا	Y_10

وقد	M_16
وهذه	M_16
على	M_17
غير	M_17
فعلى	M_17
كما	M_17
وبهذا	M_17
وتتمثل	M_17
وقد	M_17
ولهذه	M_17
وهو	M_17
وهي	M_17
كما	M_18
وبالموازاة	M_18
وبلغة	M_18
وفي	M_18
فالمقتضيات	M_2
كما	M_2
كما	M_2
والمغرب	M_2
وبالمعنى	M_2
وستأتي	M_2
وقد	M_2
وقد	M_2
وهكذا	M_2
وهو	M_2
وتصل	M_3
وتواصل	M_3
كما	M_5
وفي	M_5
ويندرج	M_5
في	M_6
كما	M_6
واعتباراً	M_6
وخلال	M_6
وفتحت	M_6
وقد	M_6

وسن	Y_5
وعند	Y_5
وقد	Y_5
ولا	Y_5
ولا	Y_5
ويجوز	Y_5
أحد	Y_6
وسيتم	Y_6
حيث	Y_7
فإذا	Y_7
كما	Y_7
كما	Y_7
وإذا	Y_7
وإذا	Y_7
ويجوز	Y_7
فإذا	Y_8
وتكون	Y_8
وهذه	Y_8
ويعتبر	Y_8
ويعد	Y_8
رغم	Y_9
فيكون	Y_9
كما	Y_9
وإجمالي	Y_9
وإذا	Y_9
وتسري	Y_9
ويحكم	Y_9

Anexo VI

Relación de las 100 *stopwords* más ocurrentes en la muestra segmentada.

POSICIÓN	STOPW.
1	في
2	من
3	على
4	أو
5	التي
6	إلى
7	مع
8	عن
9	الأطفال
10	أن
11	كما
12	خلال
13	هذه
14	ما
15	رقم
16	هذا
17	وزارة
18	تم
19	الطفل
20	القانون
21	حقوق
22	وقد
23	العمل
24	قانون
25	لا
26	عام
27	العامة
28	الى
29	المادة
30	الوطنية
31	الاجتماعية
32	عدد
33	العام
34	بين
35	علي
36	سنة
37	الذي
38	حيث
39	ذلك
40	حول
41	كل
42	لسنة
43	الدولية
44	اللجنة
45	المائة
46	الإنسان
47	الخاصة
48	غير
49	وفي
50	لبنان
51	وذلك
52	الوطني
53	المجتمع
54	عليها
55	الأحداث
56	الاطفال
57	فيها
58	قبل
59	التقرير
60	الدولة
61	بعض
62	بها
63	جميع
64	إلي
65	تلك
66	إطار
67	المجلس
68	ومن
69	التعليم
70	والتي
71	حماية

72	عمل
73	أي
74	قد
75	الحكومية
76	الخدمات
77	الإجراءات
78	الحكومة
79	الداخلية
80	الرعاية
81	فقد

82	لم
83	المؤسسات
84	كافة
85	فإن
86	نسبة
87	مجال
88	يلي
89	الشؤون
90	تنفيذ
91	منها

92	الجهات
93	الذين
94	العديد
95	الصحية
96	المدني
97	رعاية
98	عدم
99	يتم
100	التربية