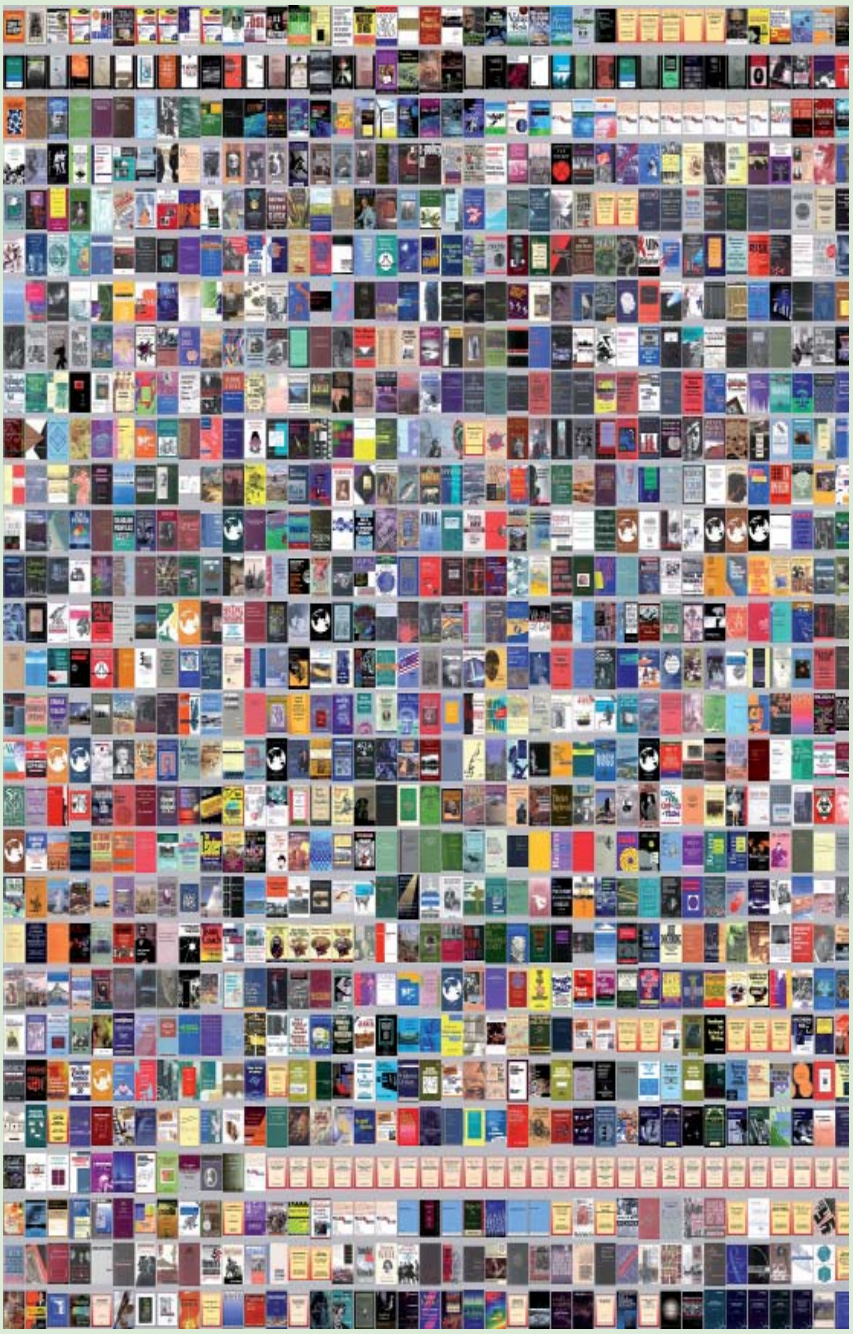
Universales de la traducción: ¿Existen?

La *convergencia* y la *simplificación* son dos de los llamados universales en los estudios de traducción. El primero postula que los textos traducidos tienden a ser más similares que los textos no traducidos. El segundo sostiene que los textos traducidos son más simples y fáciles de entender que los no traducidos.

Este trabajo analiza los resultados de un proyecto que aplica técnicas de programación neurolingüística en córpora comparables de textos traducidos y no traducidos en español, con el fin de determinar la validez de estos dos universales.



Por: **Gloria Corpas Pastor**, Departamento de Traducción e Interpretación, Universidad de Málaga, España.

Ruslan Mitkov, Instituto de Investigación en Procesamiento de Información y Lenguaje, Universidad de Wolverhampton, Reino Unido.

Naveed Afzal, Instituto de Investigación en Procesamiento de Información y Lenguaje, Universidad de Wolverhampton, Reino Unido.

Viktor Pekar, Oxford University Press, Reino Unido.

Traducción: Roddie Mazzuchi MacSwain.

Convergencia y simplificación: Estudio de programación neurolingüística sobre la base de córpora.

1. Introducción

Durante mucho tiempo, los estudios de traducción han enfocado su interés en las características del texto traducido o, más específicamente, en las características distintivas que exhiben, por lo general, los textos traducidos y el modo en que se diferencian de textos originales no traducidos escritos por hablantes nativos. La investigación inicial data de 1995, cuando Toury presentó las leyes de *estandarización* creciente y la ley de interferencia, si bien fue Baker en 1993 y 1996 quien formuló muchos de los llamados universales y sugirió

el uso de córpora para su estudio. Los universales suscitaron mucha atención entre los expertos en traducción, si bien su formulación y explicación inicial se ha basado en la intuición y la introspección con investigación de corpus de seguimiento restringida a córpora de poca extensión comparativamente, textos literarios o periodísticos y análisis semimanual. Por otro lado, la investigación previa no orientó lo suficiente respecto de cuáles eran las características que hacen que estos universales sean considerados válidos, Corpas Pastor (2008).

Este trabajo aplica un enfoque completamente diferente e innovador al utilizar robustas técnicas de programación neurolingüística sobre córpora de textos traducidos al español y sobre córpora comparables de español no traducido con el fin de investigar la validez de dos universales de la traducción, la *simplificación* y la *convergencia*. El universal de simplificación se manifiesta en el hecho de que los textos traducidos tienden a ser más simples y fáciles de entender que los textos no traducidos. Según el universal de convergencia, los textos traducidos suelen ser más similares entre sí que los textos no traducidos. El propósito de este estudio es el de determinar si estos dos universales tienen validez cuando el español es el texto destino. A este fin, analizamos córpora de textos traducidos al español y córpora comparables de textos en español no traducidos. Mediante el uso de herramientas de procesamiento del lenguaje, analizamos los córpora respecto de una variedad de características léxicas, gramaticales y estilísticas.

2. Córpora utilizados

En nuestro estudio, examinamos pares de córpora comparables de dos campos de especialización, la médica y la técnica. Dentro del área de la medicina, trabajamos con dos tipos de córpora: traducciones hechas por traductores profesionales y traducciones realizadas por alumnos. He aquí abajo el listado de los córpora que fueron específicamente compilados para este experimento:

- Corpus de traducciones en el área de la medicina realizadas por profesionales (TMP).
- Corpus de traducciones en el área de la medicina realizadas por alumnos (TMA).
- Corpus de traducciones técnicas (TT).
- Corpus de textos originales en el área de la medicina, comparables con traducciones realizadas por profesionales (CTMP).
- Corpus de originales en el área de la medicina comparables con traducciones realizadas por alumnos (CTMA).
- Corpus de textos originales en el área técnica comparables con traducciones técnicas (CTT)

TMP es comparable con CTMP, TMA es comparable con CTMA, y TT es comparable con CTT. La comparabilidad fue una consideración de importancia crucial para este estudio, ya que de otro modo se habría comprometido cualquier comparación de estilo o sintaxis. Los córpora se compilaron de modo tal que se garantizó la comparabilidad. Los criterios de diseño comprenden restricciones diatópicas, diacrónicas, diasistemáticas y de campo de especialización. Todos los textos traducidos tienen al inglés británico o estadounidense como idioma fuente y al español peninsular como idioma destino. Tanto los córpora de textos traducidos como los de textos no traducidos tienen aproximadamente la misma extensión. El TMP se compone de traducciones de biomedicina realizadas por traductores profesionales (en relación de dependencia o independientes trabajando para empresas registradas de traducción europeas). Se trata de un corpus de referencia especializado, ya que no contiene documentos enteros sino fragmentos compuestos por los segmentos de idioma destino de memorias de traducción (MT). Los tipos de texto van desde trabajos de investigación en publicaciones hasta ensayos clínicos, libros de texto, descripciones de productos y folletos informativos para pacientes, guías de usuarios e instrucciones para equipo quirúrgico. Su corpus comparable de textos de biomedicina en español no traducido incluye una selección similar de temas y tipos de textos. Es un corpus mixto, ya que contiene fragmentos y documentos enteros: segmentos de MT de lenguaje fuente diferentes de los empleados para compilar el TMP, un corpus pequeño de diabetes, y un corpus virtual *ad-hoc* compilado para coincidir con el TMP respecto de campos de especialización secundarios, temas, nivel de especialización comunicativa y tipos de texto. El otro corpus de biomedicina en español es un corpus textual especializado que contiene documentos enteros, esto es, traducciones realizadas por alumnos del último año de la carrera de Traducción e Interpretación durante los períodos académicos de 2004-2005, 2005-2006 y 2006-2007. Comprende casi los mismos tipos de texto y temas que el TMP, pero con una proporción más elevada de trabajos de investigación, descripciones de productos y folletos de información para pacientes. El CTMA es comparable con el TMP, ya que ambos comparten similares criterios de diseño.

Por último, el TT contiene segmentos de lenguaje fuente de MT de áreas técnicas y tecnológicas (telefonía, servicios de redes, telecomunicaciones, etc.) y el subcorpus en español CRATER. Comprende fragmentos de manuales de usuario, guías e instrucciones de operación, comunicados de prensa de empresas y, en menor grado, reglas y reglamentaciones, normas, proyectos y monografías. El CTT fue compilado *ad hoc* a partir de fuentes electrónicas evaluadas. Luego de analizar el TT en términos de tipos de texto, campos de especialización y temas, se generó un catálogo de palabras índice y ecuaciones de búsqueda. Como resultado de todo esto, terminamos compilando un corpus que es parcialmente comparable con el TT, ya que contiene documentos enteros (no sólo fragmentos). Debe remarcarse que ubicar este tipo de documentos técnicos en español peninsular fue más complicado que dar con documentos originales en español en el área de la medicina, ya que muchos textos de este tipo son traducciones encubiertas. Hemos garantizado la inclusión de sólo textos tecnológicos originales no traducidos mediante el filtrado y refinamiento de todas las búsquedas electrónicas.

Se indica a continuación el tamaño de los córpora arriba mencionados (número de tokens) :

- TMP: 1.058.122
- TMA: 780.006
- TT: 1.736.027
- CTMO: 1.402.172
- CTMA: 1.164.435
- CTT: 1.986.651

Por ende, los córpora de material en español traducido y no traducido son comparables con los siguientes argumentos:

a) Los pares de córpora traducidos y no traducidos incluyen aproximadamente el mismo rango de formas y tipos de texto

b) Pertenecen a los campos de especialización principales y secundarios

c) Muestran el mismo nivel de especialización y formalidad

d) Están circunscriptos diatópicamente al español peninsular

e) Fueron generados en el mismo lapso (2005-2008)

f) Su extensión es similar (número de tokens)

3. Características del corpus

Desafortunadamente, estudios previos sobre universales no han explicado lo que exactamente se clasifica como evidencia en términos de diferentes características de texto para su validación. Con miras a lograr una medida objetiva que cuantifique el grado en el que éste o aquel universal sea válido, resulta de importancia definir características o parámetros para que así puedan llevarse adelante estudios empíricos formales, con el fin de comparar textos en términos de simplificación o similitud, y más específicamente, para verificar nuestras hipótesis. En ausencia de tales pautas, el primer paso a seguir en este estudio es el de identificar características de textos. Proponemos evaluar estas características de los córpora sobre la base de los siguientes aspectos :

a) aspectos lexicográficos (riqueza lexicográfica y densidad lexicográfica)

b) aspectos estilísticos (longitud de la oración, uso de oraciones simples en lugar de oraciones complejas, uso de marcadores del discurso así como de conjunciones, legibilidad)

c) aspectos sintácticos (patrones de etiquetado de partes del habla)

A continuación, describiremos estos aspectos con mayor detalle:

3.1 Aspectos lexicográficos

Densidad lexicográfica: La densidad lexicográfica se computa como tipo/token dividiendo el número de tipos por el número total de tokens presentes en el corpus. Una densidad lexicográfica baja involucra una alta cantidad de repeticiones con las mismas palabras teniendo lugar una y otra vez. Por otro lado, una alta densidad lexicográfica significa que se utiliza una forma de lenguaje de mayor diversidad.

Riqueza lexicográfica: Sostenemos que la densidad lexicográfica no es indicativa de la variedad de vocabulario de un autor al contar variantes morfológicas de la misma palabra como tipos diferentes de palabra. Sin embargo, si bien *alumno* y *alumnos* técnicamente pueden ser palabras y tipos de palabras separados, representan la misma palabra desde un punto de vista lexicográfico. Para mitigar tal insuficiencia, proponemos una nueva medida de riqueza lexicográfica que se computa como el número de postulados dividido por el número de tokens presentes en el corpus y que explica la variedad de usos de una palabra por parte de un autor. El analizador Connexor retorna automáticamente el postulado de cada palabra (Tapanainen y Jarvinen, 1997).

> "Este trabajo aplica un enfoque completamente diferente e innovador al utilizar robustas técnicas de programación neurolingüística sobre córpora de textos traducidos al español y sobre córpora comparables de español no traducido con el fin de investigar la validez de dos universales de la traducción, *la simplificación* y *la convergencia*."

3.2 Aspectos estilísticos

Longitud de oración: La longitud de oración es un aspecto considerado típico de un estilo individual. Computamos la longitud de oración como el número de tokens en el corpus dividido por el número de oraciones en dicho corpus. En el presente trabajo, a diferencia del Estudio 1, hemos optado por no incluir la profundidad del árbol de análisis gramatical como aspecto estilístico debido a que: a) el árbol de análisis gramatical es más un concepto sintáctico; b) creemos que la profundidad del árbol de análisis gramatical y la longitud de oración no son aspectos completamente independientes.

Oraciones simples versus oraciones complejas: Sostenemos que el uso de oraciones predominantemente simples o complejas, o una combinación equilibrada de ambas, representa un aspecto relevante para el estilo de un autor. Con el fin de contar el número de oraciones simples o complejas, desarrollamos un algoritmo que identifique automáticamente el tipo de oración por medio del conteo del número de verbos finitos (y de sus correspondientes construcciones verbales) en una oración. Las oraciones con más de un verbo finito se clasifican como oraciones complejas. Contracciones tales como *haber*, *tener* o *ser* más participio pasado y *estar* más gerundio se contabilizan del mismo modo. Los verbos son detectados por

el analizador Connexor, al igual que los participios pasados y los gerundios. Hemos computado la proporción de casos en los que se emplean oraciones simples o complejas.

Marcadores del discurso: Según Biber (1988, 1995, 2003), el empleo de marcadores del discurso representa otra característica del estilo de un autor. A este fin, haciendo uso de un listado de marcadores del discurso en español, hemos extraído y calculado la proporción de marcadores del discurso del número de todas las palabras en un corpus.

Legibilidad: Experimentamos con tres medidas comunes de legibilidad en el texto: índice automatizado de legibilidad (ARI), índice Coleman-Liau (CLI), y examen de nivel de legibilidad de Flesch-Kincaid (FK).

El índice automatizado de legibilidad (Smith y Senter, 1967) fue creado originalmente para los manuales y documentación técnica de la Fuerza Aérea de los Estados Unidos de Norteamérica. Este examen de legibilidad está diseñado para medir la capacidad de comprensión de un texto. La siguiente es la fórmula para este examen:

$$ARI = 4{,}71 \frac{c}{w} + 0{,}5 \frac{w}{s2} - 21{,}43$$

donde *c* es el número de caracteres, *w* es el número de palabras y *s* es el número de oraciones en el texto. La fórmula calcula el mínimo nivel requerido para comprender un texto.

M. Coleman y T. L. Liau (1975) presentaron su examen de legibilidad con el objeto de medir la capacidad de comprensión de un texto. De modo similar al ARI, este examen también se basa en caracteres en lugar de sílabas por palabra. La siguiente fórmula se emplea con el fin de calcular el índice Coleman – Liau:

$$CLI = 5{,}89 \frac{c}{w} - 0{,}3 \frac{s}{w} - 15{,}8$$

El examen Flesch – Kincaid (Flesch, 1948) fue diseñado con el objeto de indicar la dificultad de comprensión al leer un pasaje en inglés académico. Este examen se basa en sílabas por palabra en lugar de caracteres, y se calcula aplicando la siguiente fórmula:

$$FK = 0{,}39 \frac{w}{s} + 11{,}8 \frac{syl}{w} - 15{,}59$$

donde *syl* describe el número de sílabas en el texto.

3.3 Aspectos sintácticos

Aplicamos etiquetado de partes del habla / análisis sintáctico superficial por cada corpus y comparamos las secuencias de partes de etiquetas cuyo propósito es el de reflejar la estructura sintáctica de las oraciones. Con el objeto de determinar la similitud entre dos córpora, en términos de sus aspectos sintácticos, se comparan vectores de n-gramos empleando criterios de medición de coseno y recurrencia modelados como ensayos de permutación (Nerbonne y Wiersma, 2006).

En nuestros experimentos comparamos secuencias de etiquetas de partes del habla por cada par de córpora. Las secuencias de etiquetas de partes del habla explican la estructura sintáctica lineal de las oraciones. La idea detrás de nuestra metodología general consiste en comparar dos córpora cualesquiera considerando n-gramos. Con anterioridad, se utilizaron n-gramos de partes del habla para medir la distancia sintáctica, con los mejores resultados registrados para n=3 (Nerbonne y Wiersma, 2006). Las córpora que se comparan se representan como vectores de frecuencia de 3-gramos, y las medidas empleadas para la comparación son el coseno así como las medidas R y R_{sq} que fueron inspiradas por el criterio de medición de recurrencia ® (Kessler, 2001).

4. Hipótesis

La siguiente serie de hipótesis se formula tomando como punto de partida trabajos anteriores sobre estudios de texto traducido basados en córpora. De acuerdo con el postulado de simplificación, esperamos que los córpora traducidos:

- **a)** estén caracterizadas por vocabulario menos variado y más familiar;
- **b)** contengan un mayor número de oraciones simples que de oraciones complejas;
- **c)** contengan oraciones más cortas que oraciones de texto original;
- **d)** contengan menos marcadores del discurso que texto original;
- **e)** sean por lo general más legibles y fáciles de comprender según medidas establecidas de legibilidad.

De acuerdo con el universal de convergencia, esperamos que los aspectos lexicológicos, estilísticos y sintácticos descritos anteriormente (*véase sección 3*) revelen diferencias de menor grado dentro de una serie de córpora traducidos que entre una serie de originales. Específicamente, esperamos que una serie de textos traducidos muestren diferencias de menor grado respecto de: a) riqueza lexicográfica y densidad lexicográfica; b) longitud de oración y proporción de oraciones simples; c) uso de marcadores del discurso; d) tipos de construcciones sintácticas utilizadas en el texto.

"Los resultados de nuestros experimentos sugieren que la simplificación no afecta los textos traducidos, si bien ello no se aplica con referencia a la longitud de oración y al uso de oraciones simples versus complejas, y los textos realizados por traductores no profesionales no aparentan poseer tales características de simplificación."

5. Universal de simplificación

Con el fin de examinar la hipótesis de simplificación, computamos valores medios para aspectos lexicográficos y estilísticos para cada corpus. Con este propósito, cada corpus fue dividido en segmentos, y cada segmento contenía 6000 oraciones. Los valores medios se obtuvieron promediando los valores para segmentos individuales en cada corpus. Estos valores medios se compararon luego aplicando el examen t no pareado de dos colas. Debido a que las características sintácticas se comparan computando una medida de similitud entre los córpora, incluimos en este experimento todos los aspectos con excepción de los sintácticos. La tabla 1 muestra los resultados correspondientes: por cada uno de los pares de córpora, la tabla muestra la media para cada corpus y el nivel de importancia (α) determinado por medio del examen t (las diferencias de importancia estadística aparecen en negritas).

Aspectos	TMP – CTMP			TMA – CTMA			TT – CTT		
	TMP	CTMP	´A	TMA	CTMA	α	TT	CTT	α
Densidad lexicográfica	.027	.042	0.005	.052	.041	0.4	.02	.025	0.001
Riqueza lexicográfica	.016	.029	0.005	.037	.028	0.4	.013	.015	0.001
Longitud de oración promedio	25.25	20.70	0.2	28.49	26.44	0.1	27.29	18.12	0.001
Oraciones simples (%)	.441	.638	0.01	.507	.521	0.7	.476	.592	0.002
Marcadores del discurso (coeficiente)	.0012	.002	0.05	.0018	.0021	0.2	.0007	.0016	0.002
ARI	16.85	15.08	0.4	19.14	19.01	0.75	17.85	12.85	0.001
CLI	16.27	16.9	0.3	17.16	18.28	0.05	16.28	15.5	0.1
FK	19.53	18.21	0.5	21.32	21.51	0.5	20.03	15.46	0.001

Tabla 1: Comparación de valores medios de los aspectos lexicográficos y estilísticos entre córpora comparables correspondientes.

Aspectos	Córpora traducidos			Córpora no traducidos		
	TMP-TMA	TMA-TT	TMP-TT	CTMP-CTMA	CTMA-CTT	CTMP-CTT
Densidad lexicográfica	0.002	0.001	0.079	0.14	0.201	0.001
Riqueza lexicográfica	0.001	0.001	0.14	0.14	0.015	0.001
Longitud de oración	0.011	0.522	0.202	0.145	0.002	0.368
Oraciones simples	0.057	0.673	0.202	0.096	0.462	0.212
Marcadores del discurso	0.001	0.005	0.351	0.063	0.001	0.072

Tabla 2: Valores P para diferencias entre córpora, computados aplicando el examen t.

6. Universal de convergencia

Con el objeto de examinar experimentalmente el universal de convergencia, comparamos similitudes dentro de una serie de textos traducidos (TMP, TMA, TT) y dentro de una serie de textos no traducidos comparables (CTMP, CTMA, CTT), véase también Corpas et al., 2008.

6.1 Comparación de aspectos lexicográficos y estilísticos

Al igual que en el experimento anterior, examinamos aspectos lexicográficos y estilísticos separadamente de los aspectos sintácticos, ya que estos últimos involucran puntaje respecto de similitudes en lugar de valores medios. Operacionalizamos la disimilitud dentro de cada grupo de córpora como promedios de probabilidades para las diferencias entre ellos, los que computamos con ayuda de dos exámenes: el examen t no pareado para cada aspecto individualmente y el examen Chi cuadrado para la serie completa de aspectos. Por ende, los valores *p* a partir de los exámenes Chi cuadrado dan como resultado un puntaje global de disimilitud dentro de una serie, mientras que los valores *p* que resultan de los exámenes *p* brindan una idea de disimilitud dentro de la serie únicamente respecto de aspectos particulares. Los valores medios para los aspectos lexicográficos y estilísticos se computan sobre los mismos segmentos de corpus indicados en la sección 5. La tabla 2 muestra los resultados de estos exámenes. La tabla 3 presenta medidas globales de similitudes entre córpora, tal como fueron computadas aplicando el examen Chi cuadrado.

Córpora	Valores p
Córpora traducidos	
TMP-TMA	0.01
TMP-TT	0.002
TMA-TT	0.023
Promedio	0.012
Córpora no traducidos	
CTMP – CTMA	0.059
CTMP – CTT	0.006
CTMA – CTT	0.071
Promedio	0.045

Tabla 3: Valores *p* para diferencias entre córpora computadas aplicando el examen Chi cuadrado.

6.2 Comparación de sintaxis

Por otro lado, evaluamos la similitud sintáctica (disimilitud en nuestro caso) entre cada par de textos traducidos y no traducidos comparando secuencias de 3 gramos de etiquetas de par-

tes del habla por cada par de córpora. En primer lugar, corrimos el analizador Connexor para identificar todas las etiquetas de partes del habla y luego recopilamos vectores de frecuencia de 3 gramos cuya disimilitud se compara sobre la base de las medidas 1-C (C: coseno), R y R_{sq}.

Más específicamente, por cada corpus construimos un vector de frecuencia con todos los trigramas de etiquetas de partes del habla. Por ejemplo, la comparación de los vectores de frecuencia del corpus de todos los textos traducidos (TMP+TMA+TT) y el corpus de los textos no traducidos (CTMP+CTMA+CTT) involucra un total de 18.468 diferentes partes del habla. La tabla 4 más abajo representa los resultados obtenidos a partir de la comparación de los pares de córpora aplicando las medidas de disimilitud antes mencionadas. Los valores más elevados de las medidas aplicadas indican una mayor disimilitud (y menor similitud) entre dos córpora comparadas.

Córpora	1-C	R	R_{sq}
Textos traducidos			
TMP-TMA	0.206	252526	638848591
TMP-TT	0.337	388466	3146471863
TMA-TT	0.176	432725	2643068563
Textos no traducidos			
CTMP-CTMA	0.017	98448	82218137
CTMP-CTT	0.15	364322	851312764
CTMA-CTT	0.167	372940	100832299

Tabla 4: Resultados que miden las diferencias sintácticas

7. Debate y conclusiones

Con referencia a la hipótesis de simplificación, aparentemente se valida en algunos parámetros, pero no en todos. Por cierto, encontramos que los textos traducidos a menudo exhiben una densidad y riqueza lexicográfica significativamente más baja, y aparentan ser más legibles que los textos no traducidos (sin embargo, sólo en un par de corpus se pudo establecer la importancia estadística para las diferencias en cuanto a legibilidad). Inesperadamente, los textos traducidos mostraron una proporción significativamente menor de oraciones simples, y sus oraciones resultaron también en gran medida más cortas. Respecto de los marcadores del discurso, encontramos que en dos pares de tres los textos no traducidos emplean marcadores del discurso muchísimo más a menudo. Curiosamente, las características de simplificación resultan más visibles en córpora de traducción técnica y en menor grado en córpora de traducciones médicas realizadas por profesionales (en las cuales todos los aspectos con excepción de longitud de oración y coeficiente de oraciones simples revelan terminología y formulaciones más simples), mientras que no se logra encontrar simplificación en textos realizados por alumnos de traducción.

Con referencia al universal de convergencia, encontramos que los valores *p* para pares de córpora traducidos son de hecho por lo general más reducidos que aquéllos para pares de córpora no traducidos, ambos determinados con ayuda del examen *t* y el examen Chi cuadrado. Valores *p* más reducidos revelan una mayor probabilidad de que el par de córpora comparados sea diferente. Esto es verdadero para la mayoría de los aspectos individuales: riqueza lexicográfica, densidad lexicográfica, longitud de oración y coeficiente de oraciones simples. Respecto de los primeros dos aspectos, las diferencias entre los córpora traducidos son mayores en dos pares, si bien en uno (TMA – TT) el panorama es en verdad el opuesto. En términos de marcadores del discurso, sin embargo, los córpora traducidos son de hecho más similares entre sí, con excepción del par TMP – TMA.

Al considerar los valores *p* computados aplicando exámenes Chi cuadrados sobre todos los aspectos, notamos que los pares de córpora traducidos poseen siempre valores *p* menores que los no traducidos, lo que nuevamente entra en contradicción con la hipótesis de convergencia.

En lo que hace a las diferencias sintácticas entre córpora, nuestros resultados indican claramente que los textos traducidos difieren más en términos de sintaxis en el caso de todos los pares comparados, y desde el punto de vista de todas las medidas (1-C, R y R_{sq}). También resulta claro que la diferencia de sintaxis es mayor entre textos de diferentes áreas de especialización. Sobre la base de los resultados arriba indicados, podemos concluir que no existe evidencia que valide la convergencia en términos de sintaxis. En rigor de verdad, los resultados de la tabla 4 muestran, por otro lado, que los textos traducidos difieren más sintácticamente que los textos no traducidos en nuestros datos experimentales.

En resumen, los resultados de nuestros experimentos sugieren que la simplificación no afecta los textos traducidos, si bien ello no se aplica con referencia a la longitud de oración y al uso de oraciones simples versus complejas, y los textos realizados por traductores no profesionales no aparentan poseer tales características de simplificación. Otro hallazgo de nuestro estudio, de importancia y no esperado, es que ninguno de los aspectos lexicográficos, estilísticos o sintácticos que elegimos con el fin de evaluar la hipótesis de convergencia pudo revelar evidencia alguna que la valide. En los experimentos lle-

vados adelante a la fecha, la longitud y complejidad de la oración aparentemente no revela mucho sobre los universales de simplificación y convergencia, por lo que resultaría de interés identificar e investigar nuevos aspectos tales como el empleo de modismos y unidades multipalabra.

El presente trabajo de investigación podrá ampliarse hacia otros idiomas así como también hacia diferentes campos de especialización y la identificación de más aspectos en textos traducidos que podrán computarse utilizando programación neurolingüística y evaluando tales aspectos. Las implicancias para la Traducción Automática serían que el texto no traducido tiende a ser más simple que los textos traducidos (tabla 1). Por ende, con miras a potenciar los sistemas de traducción automática, los investigadores deberían apuntar hacia el análisis de córpora comparables de texto traducido versus texto no traducido, con el objeto de identificar los aspectos característicos de los textos no traducidos e intentar reproducirlos en lo generado a través de la traducción automática. También debe decirse que tales aspectos seguramente variarán de un campo de especialización a otro, por lo que resulta necesario aplicar un enfoque de registro limitado en cuanto a género.

Bibliografía

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

Baker, M., 1996. "*Corpus-based Translation Studies: The Challenges that Lie Ahead*", en H. Somers (ed.), 1996. Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager, Amsterdam & Philadelphia: John Benjamins, 175-186.

Biber, D, 1988, Variation across Speech and Writing. Cambridge, Cambridge University Press.

Biber, D., 1995, Dimensions of Register Variation: a Cross-Linguistic Comparison, Cambridge, Cambridge University Press.

Biber, D., 2003, "*Variation among University Spoken and Written Registers: A New Multi-dimensional Analysis*", en: P. Leistyna & C. F. Meyer (eds.), 2003. Corpus Analysis. Language Structure and Language Use, Amsterdam & New York, Rodopi, 47-70.

Coleman, M. and Liau, T. L., 1975, "A Computer readability formula designed for machine scoring", *Journal of Applied Psychology*, Vol. 60, pp. 283-284.

Corpas Pastor, G., 2008, *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Frankfort am Main, Berlin & New Cork, Peter Lang.

Corpas Pastor, G., Mitkov R., Afzal N., Garcia Moya L., 2008, "Translation Universals: Do they exist? A corpus-based and NLP approach to convergence", en *Proceedings of the LREC (2008) Workshop on "Comparable Corpora"*, LREC-08, Marrakesh, Marruecos.

Flesch, R., 1948, "A new readability yardstick", *Journal of Applied Psychology*, Vol. 32, pp. 221-233.

Kessler, B., 2001, *The Significance of Word Lists*, Stanford, CSLI Press.

Laviosa, S., 2002, *Corpus-based Translation Studies. Theory, Findings, Applications*, Amsterdam & Nueva York, Rodopi.

Nerbonne J. & Wiersma, X., 2006, "A Measure of Aggregate Syntactic Distance", en J. Nerbonne & E. Hinrichs (eds.), 2006. *Linguistic Distances. Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics*, Sydney, Australia, 82-90.

Smith, E. A. and R.J. Senter, 1967, *Automated Readability Index AMRL-TR*, 66-22, Wright-Patterson AFB, OH, Aerospace Medical Division.

Tapanainen, P., Jarvinen, T., 1997, "A non-projective dependency parser", en *Proceedings of the 5th Conference of Applied Natural Language Processing*, Washington D.C., USA, pp. 64–71.

Toury, G., 1995, *Descriptive Translation Studies and Beyond*, Amsterdam, John Benjamins.