Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 99-110 recibido 15-12-2022 revisado 07-02-2023 aceptado 15-02-2023

Construcción del RomCro, un corpus paralelo multilingüe

Construction of RomCro, a multilingual parallel corpus

Gorana Bikić-Carić[1], Bojana Mikelenić[1], Metka Bezlaj[2]
[1] Facultad de Humanidades y Ciencias Sociales de la Universidad de Zagreb
{gbcaric, bmikelen}@ffzg.hr
[2] Departamento de Lingüística de la Universidad de Zadar
mbezlaj21@unizd.hr

Resumen: En este trabajo se presentan las fases de construcción de un corpus paralelo multilingüe de cinco lenguas romances y croata. El corpus contiene oraciones originales provenientes de textos literarios de los siglos XX y XXI, alineadas con sus traducciones al resto de los idiomas. El orden original de las oraciones ha sido cambiado. El corpus cuenta con 15,9 millones de palabras y está disponible en las plataformas *Sketch Engine* y ELRC.

Palabras clave: corpus multilingüe, corpus paralelo, lenguas romances, croata.

Abstract: In this article we present the phases of construction of a parallel multilingual corpus of five Romance languages and Croatian. The corpus contains original sentences from literary texts from the 20th and 21st centuries, aligned with their translational equivalents in remaining languages. The original order of sentences is scrambled. The corpus counts with 15.9 million words and is available on platforms *Sketch Engine* and ELRC.

Keywords: multilingual corpus, parallel corpus, Romance languages, Croatian.

1 Introducción

A diferencia de los corpus monolingües, los corpus multilingües contienen textos correspondientes a varias lenguas. Estos corpus pueden ser paralelos (formados por textos escritos en una lengua y sus traducciones a otras)[1] o comparables (construidos por textos en varias lenguas que pertenecen al mismo tipo, o se usan para su construcción las mismas técnicas de muestreo) (McEnery, Xiao y Tono, 2006: 47). En los corpus paralelos, es preciso distinguir entre corpus unidireccionales (lengua fuente → lengua meta), bidireccionales (lengua fuente ↔ lengua meta) y multidireccionales (lengua(s) fuente(s) ↔ varias lenguas meta) (McEnery, Xiao y Tono, 2006: 48; Mikelenić y Tadić, 2020: 3932).

En cuanto a la utilidad de los corpus paralelos, es sabido que hoy en día la conveniencia de utilizar esos recursos en las investigaciones lingüísticas es enorme. Se pueden usar para todo tipo de investigaciones lingüísticas (contrastivas, traductológicas, fraseológicas, lexicográficas, glotodidácticas, etc.) (p. ej. Granger, Lerot y Petch-Tyson, 2003; Teubert, 2007), pero también pueden ser muy valiosos tanto en la enseñanza de la traducción (López Rodríguez, 2016) como en el entrenamiento de sistemas de traducción automática (Koehn et al., 2007) y en la extracción terminológica (Lefever, Macken y Hoste, 2009).

[1] Para los diferentes usos del término *paralelo* (ing. *parallel*) en la literatura, v. McEnery, Xiao y Tono (2006: 47).

ISSN 1135-5948 DOI 10.26342/2023-70-8

En este trabajo se presenta el proceso de construcción del RomCro, un corpus paralelo multilingüe que cuenta con seis lenguas. El RomCro está alineado, lematizado, anotado morfosintácticamente y compuesto de textos literarios escritos en cinco lenguas romances (en español, francés, italiano, portugués, rumano) y en croata. Este es el primer corpus paralelo multilingüe y casi completamente multidireccional que alinea textos literarios en diversas lenguas romances y una eslava. En otros corpus similares la lengua eslava era la lengua pivote (Terzić et al., 2020; Grabar et al., 2018; Akimova et al., 2020). Por eso, es importante incidir en que se trata de un corpus multidireccional, ya que incluye traducciones de cada texto a todas las demás lenguas del corpus.

La creación de este corpus empezó en 2019 como parte de un proyecto[2] de la Cátedra de Lingüística Románica del Departamento de Estudios Románicos de la Facultad de Humanidades y Ciencias Sociales de la Universidad de Zagreb bajo la dirección de Gorana Bikić-Carić. Además de la directora, en el proyecto participan los colaboradores Dražen Varga, Bojana Mikelenić y Metka Bezlaj. Las determinadas fases de desarrollo y los resultados de varias investigaciones lingüísticas hechas a base de unos subcorpus del RomCro ya se han presentado en algunas conferencias[3] y publicaciones (Bikić-Carić y Bezlaj, 2022; Bikić-Carić, 2020).

Este artículo se organiza en cinco partes: después de la introducción, en la segunda parte se mencionan algunas bases de textos existentes que contienen las mismas lenguas; en la tercera parte se presenta la composición actual del corpus, mientras que en la cuarta parte describimos todas las fases de desarrollo del corpus con enfoque en varios desafíos y problemas que surgieron a lo largo de su construcción. La quinta parte está dedicada a la conclusión y propone unas pautas para futuras investigaciones.

2 Trabajo previo

Antes de pasar a la descripción de las varias fases de desarrollo del RomCro, cabe destacar las razones por las cuales optamos por construir un nuevo corpus, si bien hay muchos corpus multilingües que ya están disponibles a los investigadores. A pesar de la gran utilidad de los recursos existentes, según nuestro conocimiento, el RomCro es el primer corpus que abarca las cinco lenguas romances ya mencionadas y el croata, tanto en la versión original como en las traducciones.

En cuanto a los corpus existentes, se trata más bien de unas colecciones más extensas en varias lenguas que permiten la extracción de textos paralelos en los idiomas incluidos en el RomCro,

[2] http://www.ffzg.unizg.hr/roman/odsjek/projekti/romcro/

[3] Mikelenić, B. y M. Bezlaj. Desafíos en la construcción de un corpus paralelo multilingüe. *XIII International CORPUS Linguistics Conference – CILC2022*, Università degli studi di Bergamo, mayo 26-28, 2022;

Bikić-Carić, G. y M. Bezlaj. Neke specifičnosti upotrebe određenog člana u romanskim jezicima (s posebnim naglaskom na francuski i španjolski). *70 godina izučavanja romanskih kultura, jezika i književnosti na Filozofskom fakultetu Univerziteta u Sarajevu*, Filozofski fakultet Univerziteta u Sarajevu, diciembre 3-4, 2021;

Bezlaj, M. y G. Bikić-Carić. Le choix entre l'infinitif et une forme conjuguée après les verbes d'opinion dans cinq langues romanes. *Considérations philologiques en contexte français et francophone*, Filološki fakultet Blaže Koneski Sveučilišta Sv. Ćiril i Metod u Skoplju, Skopie, noviembre 19-20, 2021;

Mikelenić, B. y M. Bezlaj. Construcción del RomCro: un corpus paralelo de lenguas romances y croata". *III Encuentro de Jóvenes Hispanistas*, Eötvös Loránd Tudományegyetem, Budapest, marzo 3-5, 2021;

Bikić-Carić, G. y M. Bezlaj. Construcción de un corpus multilingüe y su aplicación en el análisis contrastivo de los artículos. *XLIX Simposio de la Sociedad Española de Lingüística*, Universitat Rovira i Virgili, Tarragona, enero 21-24, 2020.

pero su uso implica numerosos problemas metodológicos. Por ejemplo, existen varios recursos lingüísticos multilingües formados por documentos legales y otros textos de la Unión Europea[4]. El más conocido de estos recursos es el *JRC-Acquis*, un corpus paralelo con más de mil millones de palabras (Steinberger et al., 2006). Estos recursos, aunque traducidos por traductores profesionales, suelen caracterizarse por un estilo jurídico y un vocabulario legal muy específico (Mikelenić y Tadić, 2020: 3932-3933; Mikelenić, 2020: 186).

Por otra parte, recursos como el *OpenSubtitles*[5] (Tiedemann, 2012; Lison y Tiedemann, 2016) están compuestos por traducciones de subtítulos y de esta manera ofrecen una gama más extensa de registros. Sin embargo, hay que tener en cuenta no solo la estructura y el formato determinado de los subtítulos, sino también el hecho de que los traductores de estos textos en la mayoría de los casos se desconocen. Además, la lengua fuente suele ser desconocida también, es decir, podemos suponer que es una sola, normalmente el inglés (Mikelenić, 2020: 186-187). Ambos tipos de recursos, es decir, tanto los subtítulos como los corpus de la UE, están recopilados y procesados automáticamente, lo que aumenta la posibilidad de errores ortográficos, de alineación, etc. (Mikelenić y Tadić, 2020: 3932).

Teniendo en cuenta todas las desventajas de los recursos existentes, decidimos crear un nuevo corpus multilingüe con las características que se irán presentando a lo largo de este trabajo. En breve, quisiéramos destacar que el RomCro es un corpus diferente de los ya disponibles, dado que está compuesto de textos literarios traducidos por traductores profesionales, lo que implica una calidad de traducción del más alto nivel. Como se verá en el capítulo siguiente, consideramos que el lenguaje de las novelas se acerca más al lenguaje general (si lo comparamos con el lenguaje legal, por ejemplo).

Asimismo, los textos que forman parte del RomCro se seleccionaron manualmente. Tanto el procesamiento inicial de los textos como la alineación se hicieron automáticamente, pero los resultados luego fueron revisados y corregidos a mano (v. subcapítulos 3.3 y 3.4), disminuyendo así la contaminación del corpus por errores de diversa índole y creando un recurso más fiable.

Finalmente, se debe subrayar que el RomCro es un corpus multidireccional, lo que significa que está compuesto de textos originales en cada lengua y de sus traducciones a todas las demás lenguas del corpus. En otras palabras, la lengua fuente es siempre conocida y sabemos exactamente, entre otras cosas, qué texto se tradujo a partir de qué original. Lo dicho supone que el RomCro es especialmente valioso a los traductores y traductólogos. Los beneficios del uso de corpus lingüísticos en los estudios de traducción están ya muy bien destacados y argumentados (v. Baker, 1993; Laviosa, 1998; Johansson y Oksefjell, 1998). Un corpus multilingüe como este ofrece varias posibilidades en este sentido también, p. ej. en investigaciones sobre las características de la lengua de traducción.

Conviene decir también que, al incluir el croata, una lengua eslava de bajos recursos digitales, el RomCro abre un nuevo camino para investigaciones contrastivas entre la lengua croata y las lenguas romances. Hasta ahora, el croata se ha incluido en los corpus paralelos que incluyen el inglés u otras lenguas eslavas (como el checo, el búlgaro o el macedonio) (Mikelenić y Tadić, 2020: 3933). Según nuestro conocimiento, el único corpus paralelo que incluye el croata junto con una lengua romance es el corpus bilingüe español-croata compuesto por Mikelenić (Mikelenić y Tadić, 2020; Mikelenić, 2020).

[4] https://joint-research-centre.ec.europa.eu/language-technology-resources_en

[5] https://www.opensubtitles.org/es

Asimismo, el croata está presente en algunos corpus literarios multilingües, como, por ejemplo, *TransLiTex* (Fraisse et al., 2018) o *InterCorp* (Čermák, 2019), del que forman parte también los textos literarios. Sin embargo, *TransLiTex* contiene traducciones de un solo libro a 23 lenguas e *InterCorp* se compone de 40 lenguas, entre ellas todas las que están incluidas en el RomCro, pero como lengua pivote tiene el checo, lo que significa que todos los textos están traducidos del o al checo, pero no necesariamente entre sí.

3 Composición actual del corpus

En la Tabla 1 se muestran todos los títulos originales que forman parte del RomCro, 27 en total. Aunque en la primera fase del proyecto elegimos dos textos originales en cada lengua, al final optamos por incluir otros títulos disponibles. Actualmente en portugués y en croata hay 3 títulos, en italiano y en rumano 4, en francés 6 y en español 7. De esta manera, el RomCro se convirtió en una especie de corpus oportunista (McEnery y Hardie, 2012: 11). Todos estos títulos están traducidos a las demás lenguas del corpus, con 5 excepciones[6]. Construido así, el corpus actualmente cuenta con 157 textos (lo que incluye 27 textos originales y 130 traducciones), en lugar de 162 textos. En total, el corpus compuesto de esta forma contiene 15,9 millones de palabras y 18,5 millones de *tokens*.

		Títulos:
1	**ES**	La sombra del viento (C.R. Zafón, 2001)
2		La catedral del mar (I. Falcones, 2006)
3		El juego del ángel (C.R. Zafón, 2008)
4		El asombroso viaje de Pomponio Flato (E. Mendoza, 2008)
5		Soldados de Salamina (J. Cercas, 2001)
6		El mapa del tiempo (F. J. Palma, 2008)
7		El tiempo entre costuras (M. Dueñas, 2009)
8	**FR**	Seras-tu là ? (G. Musso, 2006)
9		HHhH (L. Binet, 2010)
10		Un barrage contre le Pacifique (M. Duras, 1950)
11		La Fée Carabine (D. Pennac, 1987)
12		L'amant (M. Duras, 1984)
13		A l'ombre des jeunes filles en fleur (M. Proust, 1919)
14	**IT**	Imprimatur (Monaldi & Sorti, 2002)
15		Le otto montagne (P. Cognetti, 2017)
16		La forma dell'acqua (A. Camilleri, 1994)
17		L'amica geniale (E. Ferrante, 2011)
18	**RO**	Maitreyi (M. Eliade, 1933)
19		Întâmplări în irealitatea imediată (M. Blecher, 1936)
20		Nostalgia (M. Cărtărescu, 1993)
21		Cartea şoaptelor (V. Vosganian, 2009)
22	**PT**	A viagem do elefante (J. Saramago, 2008)
23		Nenhum olhar (J. L. Peixoto, 2000)
24		As intermitências da morte (J. Saramago, 2005)
25	**CR**	Muzej bezuvjetne predaje (D. Ugrešić, 1998)
26		Mediteranski brevijar (P. Matvejević, 1987)

[6] No se pudo conseguir la traducción al portugués de la novela *Maitreyi* y al italiano de *Muzej bezuvjetne predaje*, aunque estas existen. Por otra parte, *El asombroso viaje de Pomponio Flato* no está traducido al rumano, mientras que *Dora i Minotaur: Moj život s Picassom* no tiene la versión española ni portuguesa.

27		Dora i Minotaur: Moj život s Picassom (S. Drakulić, 2015)

Tabla 1: Títulos originales incluidos en el RomCro (según las lenguas).

La Tabla 2 presenta la distribución de todos los textos del corpus por lenguas. Cada lengua del corpus cuenta con alrededor de tres millones de *tokens*, lo que incluye tanto los textos originales como las traducciones. El subcorpus francés es el más grande con casi 3,5 millones de *tokens*, mientras que el croata es el más pequeño con 2,8 millones.

En cuanto a los textos originales, la mayoría de los subcorpus tiene menos de 1 millón de palabras: el subcorpus español es el único que cuenta con un número mayor de *tokens* y el croata es otra vez el más pequeño. Naturalmente, la distribución de *tokens* en las traducciones refleja una situación contraria: el español tiene el menor número de *tokens*, mientras que las demás lenguas cuentan con más de 2,5 millones.

	Número total de *tokens*	**Número de *tokens* (textos originales)**	**Número de *tokens* (traducciones)**
Español	3,146,711	1,170,843	1,975,868
Francés	3,419,815	676,806	2,743,009
Italiano	3,024,176	483,537	2,540,639
Portugués	2,997,771	212,016	2,785,755
Rumano	3,158,078	437,474	2,720,604
Croata	2,838,744	174,781	2,663,963

Tabla 2: Distribución de *tokens* por lenguas y en textos originales y traducidos.

La distribución de *tokens* en textos originales y traducidos se ha proyectado igualmente en el Gráfico 1, que muestra los mismos datos en forma porcentual. El porcentaje de los textos originales desciende desde el español (37,2 % de textos originales) hacia el portugués (7 % de textos originales) y el croata (6,2 % de textos originales), mientras que los subcorpus francés, italiano y rumano son bastante similares en cuanto a su composición: el francés tiene un 19,8 % de textos originales; el italiano, 16 %, y el rumano, 13,9 %.

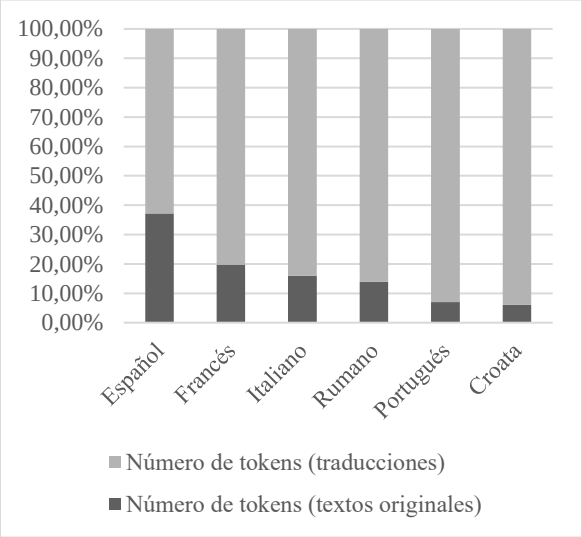

Gráfico 1: Distribución porcentual de *tokens* en textos originales y traducidos.

4 Fases de desarrollo

El desarrollo del RomCro se llevó a cabo de 2019 a 2022 y consistió en varias fases, comenzando por la selección de textos y terminando por hacer el recurso disponible en las plataformas *SketchEngine*[7] y ELRC[8].

4.1 Selección y recogida de textos

Teniendo en cuenta que la elaboración de cada corpus requiere la adopción de varios criterios en cuanto a su estructura y extensión, en un primer momento nos dedicamos al diseño del RomCro. Los criterios adoptados fueron los siguientes: 1) disponibilidad de textos traducidos a todas las lenguas del corpus, 2) textos literarios y 3) contemporáneos, 4) homogeneidad

[7] https://www.sketchengine.eu/

[8] https://www.lr-coordination.eu/

dialectal/variedades europeas, 5) formato digital. Como se verá a continuación, al compilar el corpus, algunos de estos criterios fueron modificados ligeramente por razones de naturaleza práctica.

El primer criterio que adoptamos fue la disponibilidad de los textos traducidos. A pesar de que todas las lenguas que forman parte del corpus sean idiomas oficiales de por lo menos un país, encontrar textos literarios originalmente escritos en cada una de las lenguas del corpus y traducidos a las demás lenguas resultó una tarea bastante difícil (sobre todo en cuanto a los textos en portugués, rumano y croata). Según nuestro conocimiento, la información sobre las traducciones a varias lenguas no está centralizada en ningún banco de datos, así que tuvimos que buscar las traducciones de cada título individualmente.

Se optó por las novelas para excluir del corpus los textos terminológicamente muy específicos, como p. ej. los de la legislación europea, es decir, para que el lenguaje del corpus fuera lo más neutral y general posible. Aunque somos conscientes de que la lengua literaria puede ser altamente idiosincrásica (Lawson, 2001: 294), argumentamos que estos rasgos distintivos no están tan presentes en las novelas como en otros géneros literarios (p. ej. cuentos u otras formas cortas, poesía o ensayos) (Mikelenić y Tadić, 2020: 3933).

Por otra parte, hay que tener en cuenta que la construcción de un corpus a partir de un solo tipo de texto puede ejercer una influencia en los resultados obtenidos. A pesar de esta limitación, vale destacar otra vez que el RomCro es el primer corpus paralelo que abarca esta combinación de lenguas y, por lo tanto, creemos que es una herramienta excelente para muchos tipos de investigaciones.

En cuanto al tercer criterio, es decir, a los límites temporales del corpus, queríamos confeccionar un corpus lo más sincrónico posible y, por consiguiente, intentamos encontrar textos del siglo XXI. No obstante, la versión final del corpus incluye algunos textos publicados en el siglo XX (una obra de la década de 1910, dos de la década de 1930, una de la década de 1950, tres de la década de 1980 y tres de la década de 1990) que decidimos incluir por el hecho de estar disponibles (v. Tabla 1).

Además, para evitar la diversidad dialectal y mantener el corpus lo más homogéneo posible, por el momento decidimos excluir los textos escritos en variedades no europeas del corpus. Esto es especialmente importante para el español y el portugués que tienen una producción literaria y traductora muy prolífica en América Latina. Las variedades americanas presentan unas características que las distinguen de las europeas y que muchas veces difieren según el país o la región. Dichos rasgos incluyen fenómenos a todos los niveles lingüísticos: desde el fonológico, morfológico y sintáctico hasta el léxico y pragmático.

Sin embargo, aunque todos los títulos originales incluidos en el corpus pertenecen a las variedades europeas, algunas de las traducciones se publicaron exclusivamente fuera de Europa, sobre todo en el caso del portugués. Por esta razón el RomCro en su versión actual cuenta con cuatro textos traducidos al portugués brasileño[9], pero el usuario puede elegir si quiere excluir estas traducciones de sus búsquedas a la hora de consultar el corpus, como se verá más adelante.

Finalmente, a fin de facilitar el procesamiento informático del corpus, nuestro objetivo fue encontrar el mayor número posible de textos seleccionados en formato digital. En varias ocasiones no fue posible cumplir con ese criterio,

[9] Se trata de las siguientes traducciones al portugués: *A fada carabina* de Daniel Pennac, *A forma da água* de Andrea Camilleri, *Acontecimentos na Irrealidade Imediata* de Max Blecher y *Nostalgia* de Mircea Cărtărescu.

así que muchos textos fueron escaneados y digitalizados automática y manualmente.

4.1.1 Aspectos legales en la construcción del corpus

Las cuestiones legales que afectan cada tipo de corpus son muchas. Dado que el RomCro está diseñado para que pueda usarse y citarse libremente con fines académicos y además está compuesto de textos contemporáneos, tuvimos que asegurarnos de que su composición no pudiera incurrir en una violación de los derechos de autor. Después de consultar los servicios jurídicos disponibles, resultó evidente que ni la legislación croata ni la legislación europea ofrecen pistas claras en cuanto a la construcción de los corpus digitales a partir de los textos literarios y sus traducciones.

No obstante, las fuentes consultadas nos corroboraron que los datos del corpus pueden usarse libremente para fines no-comerciales, siempre y cuando se citen adecuadamente, es decir, si figuran junto al título y nombre del autor. Además, la probabilidad de incumplir la normativa de los derechos de autor disminuye si las búsquedas del corpus se pueden hacer únicamente a nivel de la frase y si el corpus se diseñó con el único objetivo de ser utilizado en investigaciones científicas, es decir, con fines académicos y no-comerciales (p. ej. Wilkinson, 2006).

Por esta razón decidimos cambiar el orden de los segmentos del corpus para que nadie pudiera recuperar ninguno de los textos en su totalidad. Es más, la alineación se hizo a nivel de la oración (o dos oraciones, dependiendo de la traducción), así que nunca es posible recuperar más que esto. Por un lado, este procedimiento impide el análisis en los niveles mayores (p. ej. el análisis del discurso), pero, por otro lado, se protegen los derechos del autor.

4.2 Digitalización de los textos y preparación para la segmentación

Después de seleccionar y obtener los textos, tuvimos que escanear los que no pudimos encontrar en formato digital y hacer un reconocimiento óptico de caracteres automático (ing. OCR) con el programa *Abbyy FineReader*. Varios estudiantes de grado y de máster colaboraron en el proyecto, revisando y corrigiendo los resultados de esta digitalización, es decir, preparando los textos para la alineación automática. En esta fase se borraba todo lo que no era el cuerpo del texto (p. ej. imágenes, números de páginas) y se excluía todo lo que no se podía alinear con su traducción a otros idiomas (p. ej. dedicatorias, información sobre el autor). Este paso prolongó sustancialmente el trabajo, pero fue necesario para que el proceso de alineación fuera lo más fácil y rápido posible y el resultado final, el más correcto.

4.3 Segmentación, alineación y corrección manual

A continuación hemos segmentado y alineado los textos automáticamente con el programa *LF-Aligner*[10], un programa de acceso libre mediante el cual se puede utilizar la herramienta de alineación *Hunalign*[11] (Varga et al., 2005). La revisión del resultado de este proceso se hizo en el programa *Microsoft Excel*. Aunque trabajamos con 15 estudiantes que manejan varias lenguas romances, la mayoría de ellos no ha podido cumplir la tarea de revisar los segmentos alineados en todas las lenguas a la vez, lo que significaba que los textos se dividían entre dos o tres personas y se juntaban de nuevo. Al final, se hizo otra revisión por parte de uno de los tres profesores que participaron en el proyecto. En esta tarea se ha trabajado continuamente durante tres años académicos, combinándola con la revisión de resultados de la digitalización de nuevos títulos que se añadían al corpus.

Lo que también complicaba esta tarea fueron las diferencias en las traducciones de algunos textos, p. ej. libros con varias ediciones, aunque por la naturaleza del tipo de texto (novela), esto lo encontramos solo con títulos y autores ya conocidos en este sentido. Un ejemplo extremo es el del autor croata Matvejević, que conocía la mayoría de estas lenguas romances y traducía sus obras solo o en colaboración con el traductor, a menudo añadiendo o reformulando partes del texto.

[10] https://sourceforge.net/projects/aligner/

[11] https://github.com/danielvarga/hunalign

Hay que destacar que en ningún momento se añadió nada, tampoco en el caso de errores y omisiones no deliberadas (p. ej. la traducción de una oración no aparece), donde se optó por dejar el segmento vacío en el idioma en cuestión. Finalmente, como ya se destacó, en la última versión del corpus se cambió el orden de los segmentos para proteger los derechos del autor.

4.4 Lematización y anotación morfosintáctica

El paso siguiente fue la lematización y anotación morfosintáctica de los textos. Nos enfrentamos con la duda de seleccionar un etiquetador diferente para cada idioma u optar por uno que otorga resultados uniformes para todos, arriesgando tal vez la pérdida de algunos rasgos distintivos. Se presentaron dos opciones: los anotadores del proyecto *Universal Dependencies* – UD[12] (Nivre et al., 2016), que ofrecen etiquetas comparables para todos los idiomas o los anotadores del *Sketch Engine*[13] (Kilgarriff et al., 2004), que son *FreeLing*[14] (Padró, 2011) para español, francés, italiano y portugués y *MULTEXT-East*[15] (Erjavec et al., 2003; Erjavec 2017) para rumano y croata. Para poder tomar una decisión informada, hicimos pruebas con los dos grupos de anotadores, cuyos resultados se presentan en el apartado siguiente.

4.4.1 Selección de los etiquetadores

Se etiquetaron 50 oraciones en cada idioma, tanto de textos originales como de los traducidos, con ambos grupos de anotadores[16], marcando errores. La selección de los etiquetadores se basó en el análisis cualitativo de estos errores y a continuación comentamos algunos que se repetían.

La dificultad de la diferenciación entre los adjetivos y los participios es bien conocida en la literatura, dado que comparten la forma. Por eso no es de sorprender que ambos grupos de anotadores etiquetaban algunos adjetivos como participios (verbos), aunque del contexto era evidente de qué se trataba. Por ejemplo, en *un alma bendita* el adjetivo fue etiquetado como participio por el anotador del *Sketch Engine*, al igual que algunos adjetivos en portugués (*uma noite lançada*), francés (*les guerres perdues*) e italiano (*disgustata com'è dall'idea che Rino…*) por los anotadores de UD.

En cuanto a la elisión, es decir, la supresión de la vocal en francés y en español, para el francés los anotadores del *Sketch Engine* cometen errores de clase de palabra (*qu'* etiquetado como verbo, *l', n'* como sustantivos, *d'* como adjetivo), que por otro lado se marca bien en los anotadores de UD, aunque están presentes los errores de lematización. Las contracciones *del* y *al* en español se etiquetan en el *Sketch Engine* solo como preposiciones, mientras que el anotador de UD les asigna dos etiquetas separadas (preposición + artículo).

Por último, el problema de la diferenciación entre algunas conjunciones y pronombres relativos que comparten la forma lo ilustramos en español e italiano. Aquí el anotador del *Sketch Engine* para español distingue bien entre las dos clases (p. ej. *que*), mientras que el anotador para italiano le designa su propia etiqueta a *che*, de esta manera evitando el problema. Los anotadores de UD para ambos idiomas en algunos casos funcionan bien, pero se encontraron varios errores (pronombres etiquetados como conjunciones).

En estos ejemplos se podía ver que ambos grupos de etiquetadores tenían dificultades, pero hay que destacar que en general funcionaron bien. Es más, conocidos los puntos débiles de cada etiquetador, en el futuro se podrá ajustar las búsquedas para compensar los errores. Al final, la elección se hizo en conexión con el siguiente asunto: el modo de acceder al corpus.

[12] https://universaldependencies.org/

[13] https://www.sketchengine.eu/

[14] http://nlp.lsi.upc.edu/freeling

[15] http://nl.ijs.si/ME/

[16] Como parte de UD, para algunos idiomas existen varios etiquetadores desarrollados en proyectos diferentes. En estos casos, primero averiguamos qué etiquetador para el mismo idioma funcionaba mejor para nuestros textos. Los etiquetadores elegidos fueron los siguientes: portuguese-bosque-ud-2.6-200830 (portugués), french-gsd-ud-2.6-200830 (francés), italian-isdt-ud-2.6-200830 (italiano), romanian-nonstandard-ud-2.6-200830 (rumano), croatian-set-ud-2.6-200830 (croata), spanish-gsd-ud-2.6-200830 (español).

4.5 Acceso al corpus

En cuanto al acceso al corpus, se presentaron dos posibilidades: crear una interfaz propia o usar una ya existente. La primera opción significa más libertad en diseño y funcionalidad, pero también más gastos de creación y mantenimiento. Por otro lado, utilizar una interfaz ya existente da acceso a asistencia técnica y más visibilidad. Se decidió elegir el *Sketch Engine* y también sus anotadores, por la funcionalidad y varias posibilidades de búsqueda y análisis que otorga este programa.

Así, los corpus paralelos se pueden consultar individualmente o en conjunto, con la opción de las concordancias paralelas (ing. *Parallel Concordance*), como se aprecia en la Imagen 1. El orden de idiomas es arbitrario y cambiable. Es más, según las necesidades del usuario, los subcorpus en cada idioma se pueden añadir u omitir (con el signo "X" en el rincón derecho de cada casilla), así que uno tiene la opción de, por ejemplo, usar solamente dos o tres subcorpus. Asimismo, es posible hacer la búsqueda solo en un idioma o proponer traducciones en uno o varios idiomas (en la Imagen 1: *Translated as (optional)*).

El programa ofrece varias posibilidades bajo *búsqueda avanzada* (en la Imagen 1: *Advanced*), donde los metadatos que contiene el RomCro se pueden usar como filtros para crear los subcorpus deseados. Como se observa en la Imagen 2, estos son el título del libro (*segment.title*), el autor del libro (*segment.author*) y el idioma original (*segment.original_lang*). De esta manera es posible hacer un subcorpus del portugués brasileño eligiendo los títulos mencionados antes (v. la nota a pie de página no. 9) o un subcorpus de textos originales en croata y sus traducciones a otros idiomas. Por supuesto, los filtros se pueden combinar. Cabe mencionar que el uso de los filtros es opcional y depende de las necesidades del usuario. Así, si uno desea consultar el corpus de portugués sin diferenciar entre las variantes, solo marcará *pt* como el idioma original.

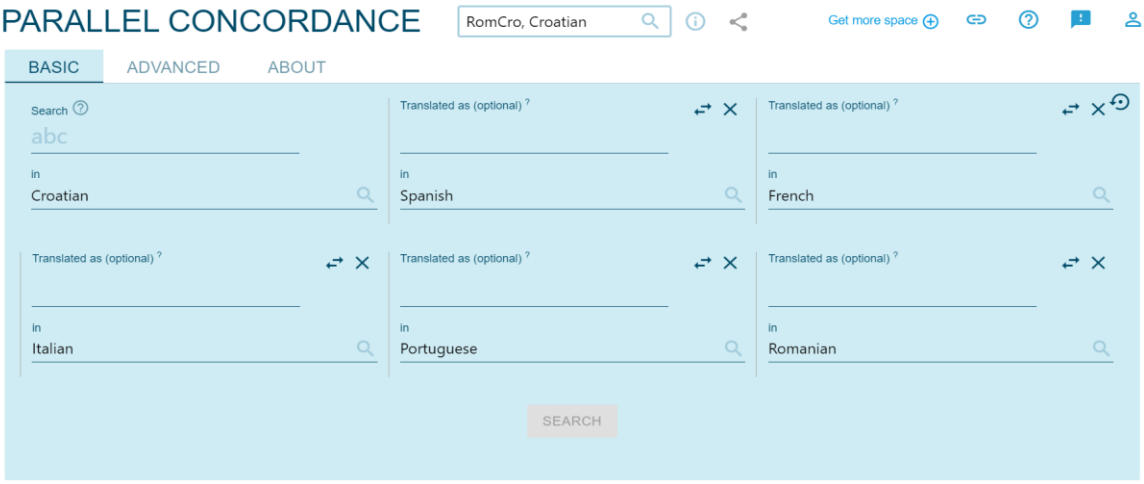

Imagen 1: La interfaz del *Sketch Engine* para obtener concordancias paralelas.

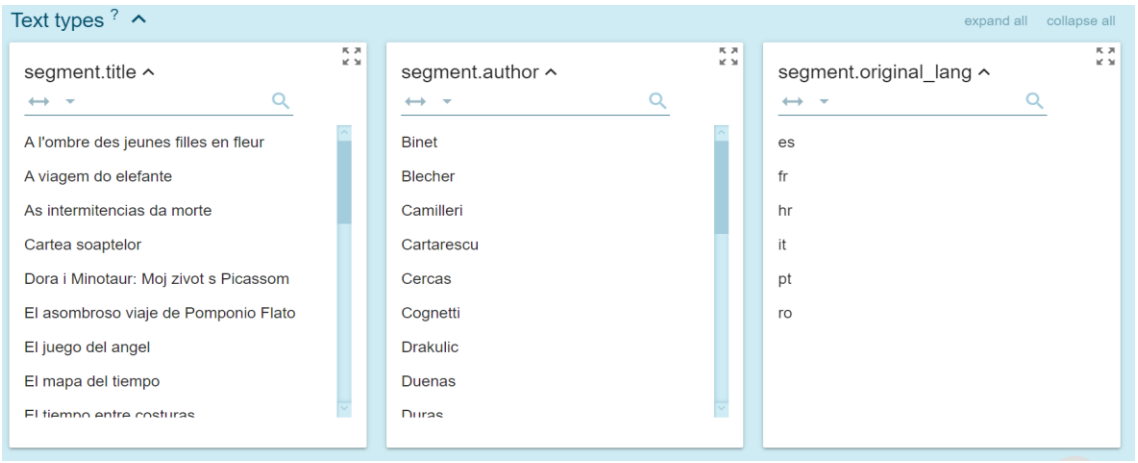

Imagen 2: Los metadatos del RomCro en el *Sketch Engine*.

El acceso al corpus es por ahora libre para los colegas de la Facultad de Humanidades y Ciencias Sociales de la Universidad de Zagreb. A otros usuarios del *Sketch Engine* les rogamos que nos contacten para otorgarles el acceso. Estamos haciendo lo posible para que el RomCro pronto sea accesible a todos los investigadores que deseen utilizarlo en esta forma.

Para los usuarios que prefieren trabajar directamente con el corpus, una versión en el formato TMX y otra en el formato TSV, también están disponibles en la plataforma ELRC (*European Language Resource Coordination*)[17], bajo la licencia CC-BY-NC-4.0. En ambos formatos, el orden de las lenguas es: español (es), francés (fr), italiano (it), portugués (pt), rumano (ro), croata (hr). Estas versiones no están lematizadas ni anotadas, pero ambos documentos contienen información sobre la lengua original, el escritor y el título de texto del que proviene cada segmento.

5 Conclusión y planes para el futuro

En este trabajo se ha presentado la construcción del RomCro, un corpus paralelo multilingüe de lenguas romances y croata que está disponible a través de las plataformas *Sketch Engine* y ELRC. Este corpus está alineado, lematizado y anotado morfosintácticamente. Es importante insistir aquí en que el RomCro es el único corpus multidireccional que cuenta con esta combinación de lenguas, lo que le hace muy valioso para lingüistas, traductores, profesores y otros profesionales.

En el futuro, esperamos ampliar el corpus con más textos literarios y/o de otros géneros e incluir más textos en variedades no europeas. Además, nos parece muy importante hacer el recurso disponible en su totalidad según los principios de la ciencia abierta, no solo a los investigadores croatas de nuestra facultad, sino también a la comunidad académica internacional.

También creemos que es necesario seguir con la realización de investigaciones de diversa índole. De hecho, este año hemos iniciado una continuación de nuestro proyecto que se dedica al análisis contrastivo de la oposición virtualidad-realidad en las lenguas romances y croata. Finalmente, todos los colaboradores en el proyecto están trabajando en la popularización de este recurso entre los colegas y estudiantes. Creemos que es tan importante fomentar el uso del RomCro en la enseñanza de las lenguas y de la traducción como animar a los estudiantes a que lo usen en la redacción de sus trabajos académicos.

Agradecimientos

Las autoras agradecen la financiación recibida por la Facultad de Humanidades y Ciencias Sociales de la Universidad de Zagreb. Igualmente

[17] https://elrc-share.eu/repository/search/?q=romcro

agradecemos a los estudiantes que han participado en el proyecto por su inestimable colaboración.

Bibliografía

Akimova, M., A. Belousova, I. Pilshchikov, y V. Polilova. 2020. En *Actas de Parallel Corpora as Digital Resources and Their Applications*: https://parallelcorporadhn2020.github.io/talks/Akimova_Belousova_Pilshchikov_Polilova.html.

Baker, M., 1993. Corpus Linguistic and Translation Studies: Implications and Applications. En *Text and Technology: In Honour of John Sinclair*, páginas 233-250, John Benjamins Publishing Company, Philadelphia / Amsterdam.

Bikić-Carić, G. 2020. Quelques particularités dans l'expression de la détermination du nom. Comparaison entre cinq langues romanes. *Studia Universitatis Babes-Bolyai-Philologia*, 65(4):39-54.

Bikić-Carić, G. y M. Bezlaj. 2022. Neke specifičnosti upotrebe određenog člana u romanskim jezicima (s posebnim naglaskom na francuski i španjolski). Filozofski fakultet Univerziteta u Sarajevu. To appear.

Čermák, P. 2019. InterCorp. A parallel corpus of 40 languages. En *Parallel Corpora for Contrastive and Translation Studies: New resources and applications*. páginas: 93-101, John Benjamins Publishing Company, Philadelphia / Amsterdam.

Erjavec, T. 2017. MULTEXT-East. En *Handbook of Linguistic Annotation*. páginas 441-462, Springer, Dordrecht.

Erjavec, T., C. Krstev, V. Petkevič, K. Simov, M. Tadić, y D. Vitas. 2003. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. En *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, ACL (Budapest).

Fraisse, A., Q.-T. Tran, R. Jenn, P. Paroubek, y S. Fisher Fishkin. 2018. TransLiTex: A Parallel Corpus of Translated Literary Texts. En *Proceedings of the 11th Language Resources and Evaluation Conference*, European Language Resource Association.

Grabar N., O. Kanishcheva, y T. Hamon. 2018. Multilingual aligned corpus with Ukrainian as the target language. *SLAVICORP* (Prague).

Granger, S., J. Lerot, y S. Petch-Tyson (eds.). 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies.* Rodopi, New York.

Johansson, S. y S. Oksefjell (eds.). 1998. *Corpora and Cross-linguistic research: Theory, Method, and Case Stuidies.* Rodopi, Amsterdam / Atlanta.

Kilgarriff, A., P. Rychlý, P. Smrž, y D. Tugwell. 2004. The Sketch Engine. En *Proc Eleventh EURALEX International Congress*, páginas 105-116, Lorient, France.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, et al. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation.* En *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, páginas 177-180, Association for Computational Linguistics.

Laviosa, S. 1998. The Corpus-based Approach: A New Paradigm in Translation Studies. *Meta : journal des traducteurs / Meta: Translator's Journal*, 43(4):474-479.

Lawson, A. 2001. Collecting, aligning and analysing parallel corpora. En *Small Corpus Studies and ELT: theory and practice*, páginas 279-309, John Benjamins Publishing Company, Philadelphia / Amsterdam.

Lefever, E., L. Macken, y V. Hoste. 2009. Language-independent bilingual terminology extraction from a multilingual parallel corpus. En *Proceedings of the 12th Conference of the European Chapter of the ACL*, páginas 496-504 (Athens).

Lison, P. y J. Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. En *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, páginas 923–929,

European Language Resource Association (Portorož, Slovenia).

López Rodríguez, C. I. 2016. Using Corpora in Scientific and Technical Translation Training: Resources to Identify Conventionality and Promote Creativity. *Cadernos de traduçao*, 1:88-120.

Mikelenić, B. y M. Tadić. 2020. Building the Spanish-Croatian Parallel Corpus. En *Proceedings of the 12th Language Resources and Evaluation Conference*, páginas 3932-3936, European Language Resource Association.

Mikelenić, B. 2020. *Kontrastivna korpusna analiza prijedložne dopune u španjolskome i njezinih ekvivalenata u hrvatskome*, tesis doctoral. Filozofski fakultet, Zagreb.

McEnery, T. y A. Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

McEnery, T., R. Xiao, y Y. Tono. 2006. *Corpus-based language studies: An advanced resource book*. Routledge, London/New York:.

Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, . . . y D. Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. En *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, páginas 1659-1666. European Language Resource Association (Portorož, Slovenia).

Padró, L. 2011. Analizadores Multilingües en FreeLing. *Linguamatica*, *3*(2):13-20.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, y D. Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. En *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, páginas 2142-2147. European Language Resource Association (Genova, Italy).

Terzić, D., S. Marjanović, D. Stosic, y A. Miletic. 2020. Diversification of Serbian-French-English-Spanish Parallel Corpus ParCoLab with Spoken Language Data. En *Text, Speech, and Dialogue. TSD 2020*, páginas 61-70, Springer.

Teubert, W. (ed.). 2007. *Text Corpora and Multilingual Lexicography*. John Benjamins Publishing Company, Amsterdam / Philadelphia.

Tiedemann, J. 2012. *Parallel Data, Tools and Interfaces in OPUS.* En *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, páginas 2214-2218, European Language Resource Association (Istanbul, Turkey).

Varga, D., P. Halácsy, A. Kornai, V. Nagy, L. Nemeth, y V. Tron. 2005. Parallel corpora for medium density languages. En *Proceedings of the RANLP 2005*, páginas 590-596 (Borovets, Bulgaria).

Wilkinson, M. 2006. Legal aspects of compiling corpora to be used as translation resources: questions of copyright. *Translation Journal* 10(2): https://translationjournal.net/journal/36corpus.htm.