PUBLICACIONES DE LA
ASOCIACIÓN ARGENTINA
DE HUMANIDADES DIGITALES

Volumen 4 2023
e-ISSN 2718-7470
https://doi.org/10.24215/27187470e051
https://revistas.unlp.edu.ar/publicaahd/

Análisis de sentimientos de traducciones: un experimento para contextos de Humanidades Digitales multilingües

Sentiment Analysis in Translation: A Pilot Experiment for Multilingual DH Contexts

Jennifer ISASI VELASCO
Pennsylvania State University
https://orcid.org/0000-0002-4295-895X

Cita recomendada:

Isasi Velasco, J. (2023). Análisis de sentimientos de traducciones: un experimento para contextos de humanidades digitales multilingües. *Publicaciones de la Asociación Argentina de Humanidades Digitales*, *4*, e051. https://doi.org/10.24215/27187470e051

RECIBIDO: 17/10/2023 **ACEPTADO:** 25/11/2023

RESUMEN

Como plenarista del V Congreso de la AAHD presenté varias metodologías que aplico a cuestiones literarias de forma experimental. Utilizando de ejemplo un experimento sobre análisis de sentimientos, justifico la necesidad de dar cuenta de los entornos HD multilingües desde mi experiencia bilingüe en los Estados Unidos, dentro de las Digital Humanities en general y de los talleres en particular, para proponer experimentos piloto con metodologías existentes como prototipos para puntos de entrada al análisis de texto en diferentes idiomas. Este artículo no afirma que el método aquí usado sea el mejor enfoque lingüístico o cultural para el análisis de sentimientos de los productos culturales hispanoamericanos. Sin embargo, defiende el potencial de explorar técnicas disponibles para crear espacios más inclusivos en escenarios de aprendizaje rápido y en los que, por norma general, los textos literarios utilizados pertenecen al canon anglosajón.

PALABRAS CLAVE: análisis de sentimientos, traducción, Humanidades Digitales, análisis textual, literatura hispanoamericana.

ABSTRACT

As a plenary speaker of the Fifth AAHD Congress, I presented several methodologies that I apply to literary studies experimentally. Using an experiment on sentiment analysis as an example, I justify the need to account for multilingual DH environments from my bilingual experience in the United States, within digital humanities in general and workshops to propose pilot experiments with existing methodologies, as prototypes to start working on text analysis in different languages. This article does not claim that the method used here is the best linguistic or cultural approach for sentiment analysis of Hispanic American cultural products. However, it defends the potential of exploring available techniques to create more inclusive spaces in rapid learning scenarios and in which, generally, the literary texts used belong to the Anglo-Saxon canon.

KEYWORDS: Sentiment Analysis, Translation, Digital Humanities, Textual Analysis, Ibero-American Literature.

1. INTRODUCCIÓN: UN BREVE PANORAMA DEL MULTILINGÜISMO EN EL MUNDO DIGITAL

El dominio del hemisferio norte y del inglés sobre la tecnología y el ámbito digital es innegable y tanto la evidencia como la lucha contra dicha situación están bien documentadas (Priani Saisó et al., 2014; Beigel, 2014; Sivertsen, 2018; Risam, 2018; Beigel, 2019; Fiormonte, 2021; del Rio Riande y Fiormonte, 2022; entre muchos otros). La idea de que el acceso al conocimiento en todo el mundo iba a ser más democrático con las tecnologías de la información y el acceso a Internet resultó ser una fantasía (o un engaño). Sin ocupar mucho espacio para el particular, podemos señalar que Galperín y Mariscal ya demostraron que el impacto

positivo de la implementación de servicios de ancho de banda fue sobreestimado en el contexto de América Latina: a pesar del efecto positivo general del acceso al ancho de banda, su impacto en el capital humano fue bastante perjudicial, ya que no tuvo impacto en la cantidad de trabajos disponibles, benefició más a los hombres que a las mujeres y no tuvo un impacto significativo en los sistemas educativos (Galperín y Mariscal 2016, pp. 279-283). La organización Whose Knowledge?[1], igualmente, nos dice que Internet en realidad está profundizando los sesgos del conocimiento en el mundo: el contenido disponible es demasiado blanco, masculino y del Norte Global, pero el 75% de los usuarios en línea son del Sur Global (Siko, et al. 2018).

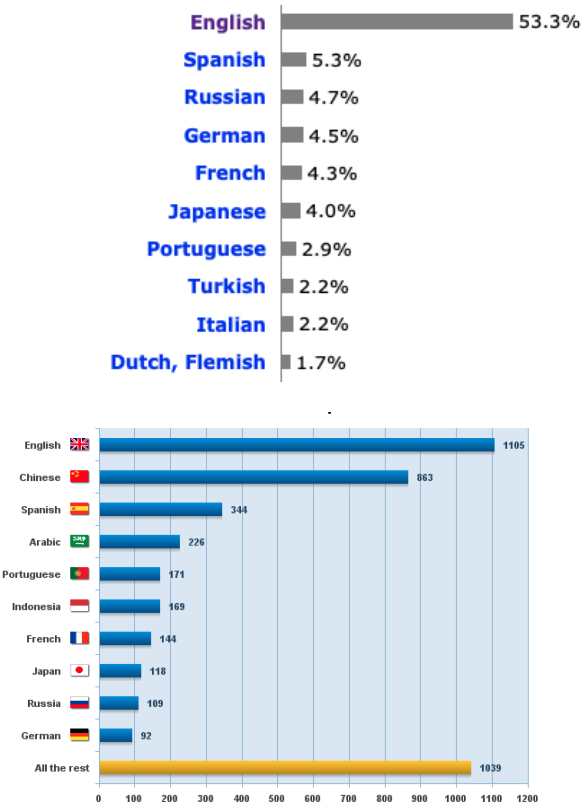

Figura 1. Idioma de acceso de las páginas web frente a los diez idiomas principales en Internet en millones de usuarios. Fuente: Web Technologies Surveys (2023) e Internet World Stats (2019).

Hay muchas iniciativas para crear contenido nuevo o traducir el que ya está disponible en varios servicios, siendo la Fundación Wikimedia

[1] https://whoseknowledge.org/

quizás el mayor ejemplo. Sin embargo, mientras que la Wikipedia en inglés cuenta ya con 6.715.000+ de artículos, su versión en español cuenta con un número bastante más reducido de 1.892.000+ de artículos (según datos de octubre de 2023). En la siguiente tabla podemos ver la diferencia de visitas a Wikipedia entre los idiomas con más hablantes, en un rango de tres meses, lo cual muestra claramente que la existencia de artículos en el idioma propio llama a los usuarios hablantes del mismo, así como a la escritura de más artículos:

Idioma y nº de hablantes	Páginas	Usuarios	Visitas de páginas
English (1.456b)	59.193.920	46.321.993	22.904.296.366
Chinese (1.138b)	7.524.332	3.415.514	1.161.832.960
Hindi/Urdu (609m)	1.282.426	784.037	199.054.797
Spanish (559m)	8.013.551	6.969.992	2.595.105.782
French (310m)	12.778.741	4.770.827	2.166.199.187
Arabic (274m)	8.255.953	2.491.731	679.914.351
Bengali (273m)	1.193.991	434.722	83.600.227
Portuguese (264m)	5.591.174	2.983.382	695.799.611

Tabla 1. Tres meses de visitas en Wikipedia por los lenguajes más hablados en el mundo a 2023. Fuente: Elaboración propia.[2]

Los números bajan cuando observamos las estadísticas de idiomas minoritarios. Como curiosidad podemos ver la situación de la Wikipedia en euskera, proyecto que se inició en 2001. Este cuenta con 421.914 artículos, 884.043 páginas y 153.401 usuarios (en 3 meses). En 2019 estos números posicionaron al euskera como el idioma número 29 sobre 336, según el perfil de Euskarazko Wikipedia.[3] No obstante, existen sólidas iniciativas respaldadas por el gobierno, una Academia de la Lengua Vasca (*Euskaltzaindia*) y grupos de sociales y académicos que

[2] Datos obtenidos en octubre de 2023 de la página *Pageviews Analysis* (análisis de visionado de páginas) de la Wikipedia, filtrando por idiomas.
[3] Estadísticas tomadas de la página de Wikipedia en lengua euskera: https://eu.wikipedia.org/wiki/Wikipedia:Estatistikak.

trabajan para un uso más amplio del euskera en la vida cotidiana, así como procedimientos gubernamentales y de investigación.

La situación en la práctica de las Humanidades Digitales (HD) a escala global no es muy diferente en términos de los idiomas que utilizamos. Consideremos estos tres ejemplos: las conferencias más grandes suelen ser en lugares anglosajones y, en la mayoría de las ocasiones, con el inglés como idioma principal o incluso único; las revistas más importantes solo aceptan artículos en inglés (salvo números especiales); y mientras que los que practicamos las humanidades y trabajamos en otros idiomas o desde las periferias citamos ampliamente trabajos en inglés, el número de referencias a artículos escritos en otros idiomas disminuye al mínimo en trabajos escritos desde regiones anglófonas y hegemónicas (Isasi y del Rio Riande, 2022). Esta situación no es específica de las Humanidades Digitales (HD), pero nos parece de gran impacto si tenemos en cuenta algunas de sus premisas: uso de recursos de licencia y código abiertos, prácticas compartidas y la ampliación del archivo digital cultural.

Algunos humanistas digitales anglófonos y, en particular, los que realizan análisis computacionales de la literatura, se benefician de diferentes iniciativas que, ya sea intencionalmente o no, están creando copias digitales de documentos que parecen lo suficientemente importantes para digitalizar. Así, basta con chequear las obras con formato de libro disponibles en HathiTrust, una iniciativa estadounidense que "ofrece una colección de millones de títulos digitalizados de bibliotecas alrededor del mundo"[4]. Una búsqueda avanzada de formato *book* en inglés muestra un total de 4.453.328 items, 1.861.451 de los cuales están disponibles con vista completa. Si buscamos el mismo formato en español, nos quedamos con 577.005 elementos con tan solo

[4] Página con la información sobre HathiTrust https://www.hathitrust.org/about/. Un vistazo a la lista de las bibliotecas contribuidoras basta para ver que, a pesar de ser ciertamente un importante grupo de bibliotecas, el "alrededor del mundo" no es tan amplio como pudiera parecer: https://www.hathitrust.org/member-libraries/contribute-content/contributors/. La traducción es mía.

118.883 disponibles en vista completa. Todas estas obras varían grandemente entre temas, fecha y lugar de publicación, y es verdad que algunas obras en la lista de inglés están en otros idiomas, o viceversa.

Con un acceso totalmente desigual a textos, además mínimamente anotados, es difícil entrenar modelos computacionales que requieren grandes cantidades de datos; y esto al mismo tiempo resulta en menos trabajos de análisis de texto computacional en otros idiomas, menos acercamientos al análisis de texto en otras lenguas y, finalmente, trabajos que no se aceptan para presentar en conferencias porque se cree que el corpus no es lo suficientemente grande, etc. En suma, el acceso desigual a recursos textuales y métodos digitales causa un desbalance tremendo en el campo de las Humanidades Digitales y la forma en que se organizan cursos, talleres, conferencias, revistas, etc. Hay humanistas digitales trabajando en sus propios idiomas por todo el mundo y es fácil encontrar trabajos de España o Alemania –dos países en los que hablar y escribir en inglés es obligatorio para todo estudiante universitario– o Latinoamérica. Pero, ¿qué ocurre en el contexto específico de los Estados Unidos de Norteamérica, donde no hay un idioma oficial y, además, el español sigue siendo visto con menosprecio a pesar de ser el segundo idioma más hablado?

En este contexto, algunos proyectos como *Torn Apart/Separados*[5], *Cartas a la familia*[6], *Borderlands Archives Cartography*[7], o los varios proyectos creados en el US Latino Digital Humanities Program en la Universidad de Houston, entre otros, llegan a una amplia audiencia que demanda contenidos que visibilicen archivos de comunidades minoritarias y oprimidas, en diferentes idiomas. Por este motivo, Ortega pide "zonas de contacto" en las *Digital Humanities* (DH), con esfuerzos de traducción como una de las estrategias, ya que la traducción es una herramienta política:

[5] Disponible en: http://xpmethod.columbia.edu/torn-apart/volume/2/index
[6] Disponible en: https://familyletters.unl.edu/
[7] Disponible en: https://www.bacartography.org

> (...) capaz de revelar la complementariedad entre prácticas y (des)entendidos comunes, así como, crucialmente, facilitando el reconocimiento y validación de diferentes modelos de conocimiento cuando sea necesario (Ortega, 2019)[8].

En cierto sentido, la revista Programming Historian[9] y el grupo online Multilingual DH[10] están tratando de seguir esa idea con el fin de remediar el acceso a metodologías digitales para estas diferentes comunidades. El desarrollo de habilidades de métodos digitales en entornos (físicos) académicos, como laboratorios o centros, y las bibliotecas en las que se brindan este tipo capacitaciones, debería seguir estos ejemplos.

En el contexto de mi investigación de análisis de textos en español en los Estados Unidos, la traducción suele ser la mejor opción. Mi trabajo se centra en la literatura y estoy particularmente interesada, primero, en proyectos que pongan a disposición obras literarias en español como colección de conjuntos de datos (tanto el texto como los registros bibliográficos) y, segundo, en los métodos que podemos aplicar a esos textos, ya que pueden informar nuestra comprensión del registro cultural digital y el análisis digital a medida que los construimos. Además, ahora que mi trabajo se centra en proveer apoyo para el desarrollo de conocimientos en estos métodos digitales en una facultad de artes liberales, junto con las bibliotecas y siendo participante de proyectos digitales multilingües, he visto más claramente la necesidad de adaptar las estrategias de traducción en el contexto de la difusión y enseñanza de las HD. Pero la traducción de las instrucciones no es suficiente y por ello, para llevar a cabo este trabajo, busco metodologías que puedan ser utilizadas también en múltiples idiomas.

El objetivo de este artículo es, por lo tanto, mostrar la necesidad de utilizar diferentes idiomas al diseñar talleres, con la esperanza de ayudar a una comunidad de humanistas digitales cada vez más diversa, con el

[8] La traducción es mía.
[9] Sitio web del Programming Historian: https://programminghistorian.org/.
[10] Página oficial: https://multilingualdh.org/en/.

estudio de un análisis de sentimientos multilingüe de traducciones como ejemplo. Este estudio, además, sirve para mostrar no sólo como aplicar un método a varios idiomas sino como puerta de entrada al uso crítico de dichos métodos en diferentes contextos culturales.

2. BASE PARA EL ANÁLISIS

2.1. El análisis de sentimientos y consideraciones metodológicas

Años de análisis del lenguaje en los campos de la Lingüística Computacional y el Procesamiento del Lenguaje Natural (PLN) han producido una amplia gama de herramientas para realizar análisis de texto en los idiomas más hablados. La atribución de autoría y la estilometría fueron, por ejemplo, dos de los principales métodos adoptados en la literatura (Mosteller y Wallace, 1964; Irizarry, 1997). Un método tradicionalmente aplicado a la revisión y el marketing de productos, a saber, el análisis de sentimientos o la extracción de opiniones, se ha adoptado más recientemente para realizar análisis computacionales de textos literarios. En principio, esta metodología consiste en asignar un valor de sentimiento positivo o negativo a las oraciones en el análisis de datos no estructurados con el objetivo de estudiar la polaridad de los textos y el progreso de los sentimientos a lo largo del tiempo.

Al igual que con la mayoría de las metodologías de análisis digital y los experimentos llevados a cabo en los últimos años, estos diccionarios de análisis de sentimientos, flujos de trabajo y corpus, que tienen el fin de comprobar resultados, se han desarrollado y realizado principalmente en inglés. Pero otros idiomas están siguiendo su ejemplo. Henríquez Miranda y Guzmán (2017) encontraron que, en español, la técnica léxica sobre análisis de sentimientos se utiliza en un 42% de los casos, seguida del aprendizaje automático con un 35% de uso y un modelo híbrido en un 15% de ocasiones.

También se han realizado muchos estudios de minería de opinión sobre tuits y reseñas en español, pero en su mayoría son experimentos el

contexto de opiniones breves sobre productos y utilizando técnicas de PLN, Machine Learning (ML) e Inteligencia Artificial (IA) (Rodríguez Aldape, 2013). Fradejas Rueda ha adaptado un diccionario para usar en R junto con el paquete *tidytext* (Fradejas Rueda, 2019), y el Grupo Tecnolengua de la Universidad de Málaga, en España, está desarrollando Lingmotif (Moreno-Ortiz, 2017), una herramienta para analítica textual. Jockers desarrolló el paquete *syuzhet* en R pensando específicamente en textos literarios en inglés (Jockers, 2017), pero también agregó la opción de ejecutar el análisis en otros idiomas aprovechando el léxico multilingüe del NRC (National Research Council Canada) desarrollado por Mohammad (s.f.) y traducido automáticamente[11]. El propio Mohammad, al mismo tiempo, ha realizado experimentos con análisis de sentimientos en novelas y cuentos de hadas (Mohammad, 2011).

Al pensar en cuestiones de contextos multilingües en el aula, en la biblioteca y una audiencia global interesada en las HD, decidí realizar un experimento para comparar los resultados obtenidos al usar *syuzhet* en varios textos en su versión original y en traducción[12]. El método funciona en inglés y en español (Isasi, 2021) con mucha variabilidad debido a la inflexión en español y, por supuesto, el lenguaje figurado (que tampoco se tiene en cuenta en inglés). La pregunta que surge entonces es: en un análisis global, ¿obtenemos resultados iguales en el mismo texto en diferentes idiomas? Es decir, ¿un texto es siempre más positivo que negativo? También, si una novela tiene cierta forma narrativa, ¿coincide esa forma en otros idiomas a la que ha sido traducida? Y lo más importante a los efectos de una introducción al análisis de sentimientos

[11] "El Léxico de Emociones NRC es una lista de palabras en inglés y sus asociaciones con ocho emociones básicas (ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y disgusto) y dos sentimientos (negativos y positivos). Las anotaciones se realizaron manualmente mediante *crowdsourcing*" (Mohammad, s.f.). La traducción es mía.

[12] Los textos del experimento fueron: *El cisne de Vilamorta* de Emilia Pardo Bazán, *La gaviota* de Fernán Caballero (Cecilia Böhl de Faber), *Marianela* y *Trafalgar* de Benito Pérez Galdós*, Frankenstein* de Mary Shelley*, Pepita Jiménez* de Juan Valera*, David Copperfield* de Charles Dickens*, The Handmaid's Tale* de Margaret Atwood*, To the lighthouse* de Virginia Woolf y *Cien años de soledad* de Gabriel García Márquez.

en un contexto multilingüe, ¿qué podemos aprender de un método *no-perfecto*?

2.2. Análisis piloto con *syuzhet*

Los experimentos piloto se definen como pruebas a pequeña escala para evaluar mecanismos de investigación, calcular el costo y la duración de un proyecto, etc. antes de realizar una investigación a gran escala. Podríamos argumentar que los talleres sobre métodos de análisis computacional de textos están, de hecho, realizando un experimento piloto: con un pequeño texto de muestra, presentamos un método que los estudiantes, profesores o bibliotecarios pueden evaluar si es útil o no para sus propósitos de investigación.

El paquete *syuzhet* en R representa una buena oportunidad para realizar este tipo de experimentos en contextos HD multilingües precisamente porque sirve como una introducción al análisis de sentimientos, permite al usuario cambiar entre varios idiomas fácilmente y plantea muchas preguntas sobre el desempeño del método para analizar novelas (u otras formas textuales).

Brevemente, el método funciona de la siguiente manera. Primero, *syuzhet* toma un texto y lo divide en oraciones como:

> La realidad ha sido para él nueva vida, para ella ha sido dolor y asfixia, ha sido la humillación, la tristeza, el desaire, el dolor, los celos (...) (Galdós, 1878) (versión original en español).
>
> Reality to him meant a new life--to her, anguish, suffocation, humiliation, sorrow, contempt, an empty life, jealousy,--Death! (traducción propia).
>
> Reality has been for him new life, for her it has been pain and asphyxiation,it has been humiliation, sadness, contempt, pain, jealousy (...) (Google Translate).

Después asigna un valor negativo o positivo a cada palabra, así como un valor sobre las ocho emociones básicas incluidas en el diccionario, como se ve en la Tabla 2:

original	dolor	y	asfixia	ha sido	la humillación	la tristeza
Translation	anguish	,	suffocation		humiliation	sorrow
GTranslate	pain	and	asphyxiation	it has been	humiliation	sadness
ira	0|1|0	0	1|1|0	0	0	1|1|1
anticipación	0	0	0	0	0	0
asco	1|0|0	0	0	0	0|1|1	1|1|1
miedo	2|1|1	0	1|1|0	0	0	1|1|1
felicidad	0	0	0	0	0	0
tristeza	7|1|1	0	0	0	0|1|1	1|1|1
sorpresa	0	0	0	0	0	0
confianza	0	0	0	0	0	0
positivo	7|1|1	0	1|1|0	0	0|1|1	1|1|1
negativo	0	0	0	0	0	0

Tabla 2. Resumen de emociones y sentimientos en media oración de muestra. Fuente: Elaboración propia.

En este caso, la primera parte en las tres versiones de la oración tiene un valor de 0 en sentimiento y emociones (no se muestra en la Figura 2), pero luego comenzamos a ver cómo los sustantivos, todos referentes en este caso a estados de ser o sentimientos, tienen valores asignados a ellos. Notamos que el valor asignado a las palabras en la mayoría de los casos no coincide en las tres variables aquí estudiadas (separadas con una | en la Tabla 2). Como es de esperar al realizar un análisis de una *bolsa de palabras*[13], no se capturan todas las posibles traducciones de una palabra en el diccionario original, en inglés, y luego no se capturan en la lista todas las flexiones que puede tener una palabra en español. Esto da como resultado que la oración en español tenga una valencia negativa de 10, con un -7 en inglés y un -5 en la versión traducida automáticamente. Este resultado era más o menos el esperado y, por tanto, si uno está haciendo un análisis de sentimientos a pequeña escala en traducción, se podría argumentar que el sistema no funciona al notar oración por oración.

Sin embargo, la intención del método no es la pequeña escala. Las novelas deben estudiarse en su totalidad para comprender el desempeño

[13] Método utilizado en el PLN para representar y analizar las palabras en un texto, sin tener en cuenta su orden de aparición en el mismo.

de *syuzhet* en una escala general y como un experimento piloto en diferentes entornos. Tenemos que realizar lo que se conoce como lectura a distancia, con sus cualidades y defectos.

3. RESULTADOS

El grupo de novelas que he probado para este experimento son textos que conozco y que tenía disponibles en texto plano. Debido a que el vocabulario de NRC que usa *syuzhet* fue traducido automáticamente del inglés a otros idiomas, y la traducción inglés-español en Google Translate está logrando una precisión bastante buena, decidí agregar una versión traducida automáticamente sin supervisión para cada texto.

En promedio, en todas las novelas exploradas las versiones en inglés de los textos tienen una tasa más alta de valor de sentimiento asignado a las oraciones. Esto significa que la versión en inglés del método captura o identifica más palabras con un valor diferente a 0. Esto es, seguramente, debido a que estas palabras y sus valores están presentes en su diccionario.

Por ejemplo, en la novela *Cien años de soledad* de Gabriel García Márquez, la versión traducida por humanos al inglés tiene un 66,92% de oraciones con un valor asignado mientras que la versión original en español tiene un 64,34% y la versión traducida automáticamente tiene un 63,76%. Estos resultados no sorprenden dado que no se capturan muchas palabras en español que tendrían valor en inglés. Los resultados de *Frankenstein* son aún más altos, con un 71,07% en la versión original y un 69,85% de frases captadas como positivas o negativas en la versión traducida al español. No obstante, ha de resaltarse que solo un 25,97% de las frases estaban marcadas en la versión traducida automáticamente de la novela. Este es un resultado positivo por parte del método ya que la diferencia entre el porcentaje de frases con valor en el original y las versiones traducidas no es significativa, salvo un texto, *El cisne de Vilamorta* de Emilia Pardo Bazán, que tiene sólo un 23% de las frases contabilizadas con valor positivo o negativo en su versión original.

Por otro lado, todas las novelas que se identifican como más positivas que negativas en su versión original, en general, también son más positivas en las versiones traducidas. En ambos casos, hay valores atípicos incluso en este pequeño corpus, pero no son diferencias significativas. Como se muestra en la Figura 2, a modo de ejemplo, las tres versiones de *El cuento de la criada* y *Cien años de soledad* dan cuenta de poca diferencia entre sus valores positivos y negativos, mientras que *La gaviota* es significativamente más positiva que negativa.

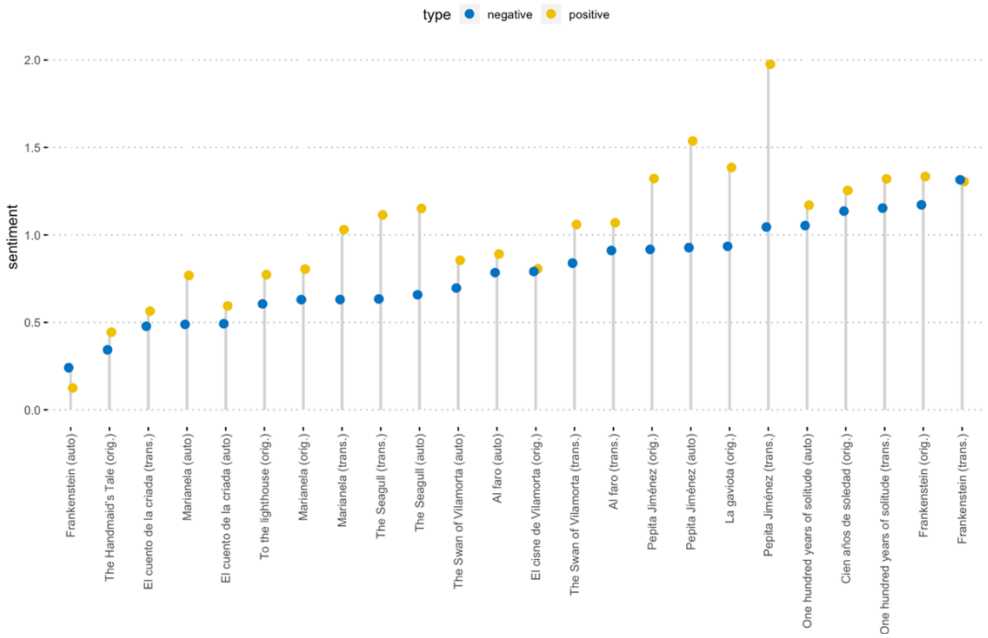

Figura 2. Grado medio de sentimiento en las novelas. Fuente: Elaboración propia.

Además de producir un valor de sentimiento general de un texto, *syuzhet* también nos permite estudiar el sentimiento en la trama narrativa a lo largo del tiempo para estudiar su forma. Según Vonnegut, tal como lo explica en su autobiografía *Palm Sunday* (Vonnegut, 1994) y en varias charlas, al extrapolar tramas de historias a una forma gráfica, es posible identificar ocho tipos de historias: *hombre en el hoyo*, *chico conoce a chica*, *de mal en peor*, *¿hacia arriba?*, *historia de la creación*, *Antiguo Testamento*, *Nuevo Testamento* y *Cenicienta* (Figura 3).

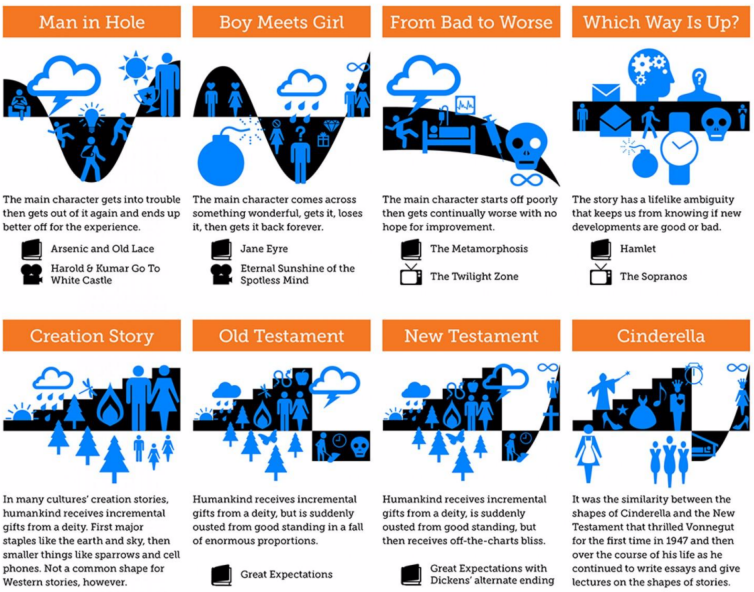

Figura 3. Las formas de las historias según Vonnegut, por Maya Eilam. Fuente: *La forma de las historia por Kurt Vonnegut*, Ersilias[14].

Tomando los resultados numéricos del sentimiento de cada palabra u oración en un texto, podemos trazar fácilmente los resultados en un gráfico como los sugeridos por Vonnegut. La función de trazado en *syuzhet* muestra dos gráficos:

> (...) el primero muestra los tres métodos de suavizado en el mismo gráfico. El segundo gráfico muestra solo la línea suavizada de la transformación discreta del coseno (DCT), pero lo hace en un eje de tiempo normalizado" (Jockers, 2017).[15]

El segundo gráfico es el que podemos usar para comparar el flujo del tiempo narrativo en las novelas, como en la siguiente figura, que agrega las tres versiones lingüísticas de *Trafalgar*:

[14] Véase: https://www.ersilias.com/la-forma-de-las-historias-por-kurt-vonnegut/.
[15] La traducción es mía.

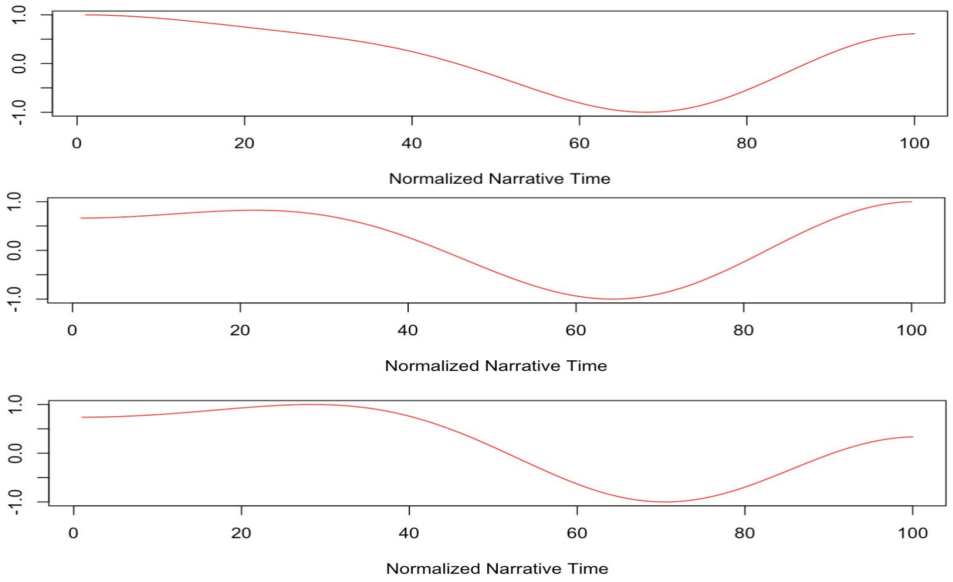

Figura 4. Forma macro simplificada de la novela *Trafalgar* en la versión original (español), la versión en inglés (traducida por humanos) y la versión automática (Google Translate). Fuente: Elaboración propia.

La Figura 4 muestra el argumento de *Trafalgar* (1873), una novela histórica ambientada en la España de 1805, de Benito Pérez Galdós. Estas tramas muestran que los sentimientos de la narración fluyen más o menos de la misma manera en las tres versiones de los textos utilizados aquí. Conocer la trama y ver el gráfico indica que estamos ante la historia de un hombre en un agujero; la narración comienza con el héroe compartiendo su vida y los acontecimientos antes de una batalla naval, la batalla en sí (en la que muere mucha gente y el propio héroe cree que va a morir) y su final feliz, pues sobrevive y avanza en la vida.

El resto de las tramas estudiadas para este artículo muestran alguna variación en términos de tiempo a lo largo de las narraciones, pero no mucha para un experimento piloto. Por ejemplo, las tres tramas de *Frankenstein*, así como su versión en francés (incluida aquí con fines experimentales), caracterizan este texto como una historia *de mal en peor*, siguiendo la teoría de Vonnegut. La trama comienza como positiva pero rápidamente se vuelve negativa en el segmento 20, solo para volverse neutral en los segmentos 40 a 60 y hundirse completamente al final de la historia (Figura 5).

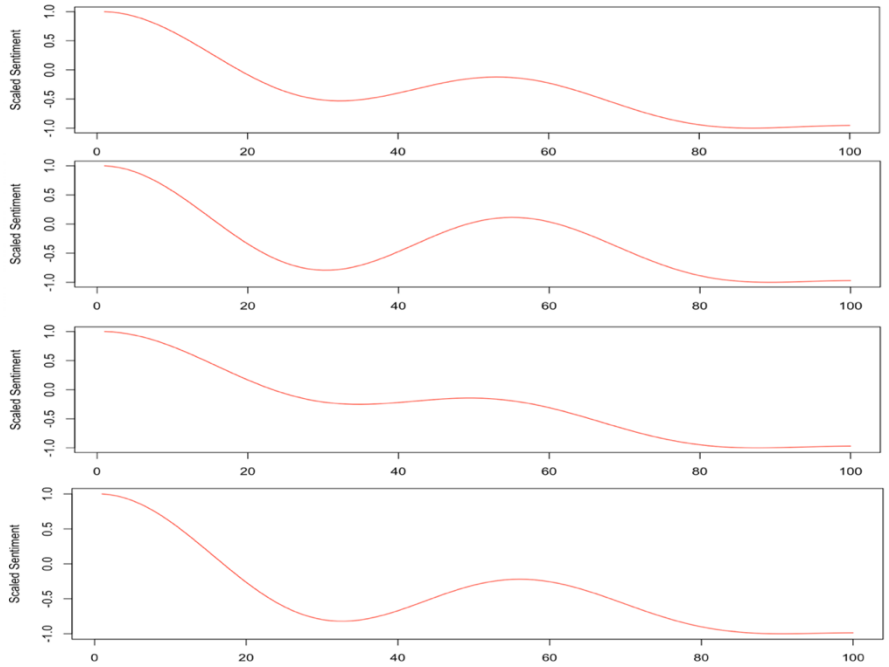

Figura 5. Cuatro tramas de *Frankenstein*: Texto original en inglés, traducción humana al español, traducción automática al español y traducción humana al francés. Fuente: Elaboración propia.

4. ¿QUÉ PODEMOS APRENDER DE ESTE EXPERIMENTO PARA UN CONTEXTO DE APRENDIZAJE RÁPIDO?

La creación de un corpus o la curación de datos para un proyecto de investigación puede llevar mucho tiempo y mucho trabajo antes de que se implemente una metodología de análisis. Elegir un método, aprenderlo (si no se conoce antes) e implementarlo puede ser algo más rápido. Comprender los resultados es más complicado. En un campo como las humanidades digitales, que cambia rápidamente, los estudiantes y profesores deben aprender métodos y herramientas nuevos o diferentes y, a menudo, buscan investigaciones previas que, por lo general, se han realizado en inglés, en contextos culturales hegemónicos y con autores canónicos. Adaptar una herramienta y/o método existente para realizar análisis de texto en diferentes idiomas requiere experiencia lingüística y, de nuevo, tiempo, ya que se crea un método digital desde cero para un corpus en particular. ¿Cómo sabemos si el método es el correcto? ¿Cómo

podemos imaginar preguntas de investigación? La diversidad lingüística ha sido ignorada en su mayor parte en este sentido.

Los experimentos piloto están destinados a proporcionar, precisamente, resultados preliminares para plantear preguntas y anticipar problemas en las metodologías de investigación. En el caso particular de contextos multilingües como en los Estados Unidos, un experimento con *syuzhet* puede proporcionar un sólido punto de entrada al tipo de preguntas que deben tenerse en cuenta en el análisis literario computacional. Al mismo tiempo, incluirá más voces en la conversación.

Primero, los estudiantes de un taller comienzan a entender el método de análisis de sentimientos y cómo funciona, que, en este caso, contrasta una lista de palabras de un texto con una bolsa de palabras con valores asignados. Esto les permite considerar los posibles beneficios e inconvenientes del método. Por ejemplo, el aspecto más destacado que debe discutirse es el hecho de que este método en particular solo funciona en una escala de lectura muy lejana. Además, los sentimientos y las emociones se capturan de forma individual, lo que significa que la negación de una emoción positiva no se captura como negativa.

Otro aspecto importante a considerar es que el diccionario multilingüe incluido en *syuzhet* se construyó preguntando a algunas personas sobre su "comprensión emocional" de las palabras en inglés en Norteamérica (Mohammad y Turney, 2013). Así, los resultados van a estar sesgados hacia la cultura de este grupo de personas y, por tanto, no puede ser una interpretación universal. Vale la pena pasar algún tiempo explorando la visualización interactiva del léxico de palabras y emociones antes de continuar con la prueba en el aula (Mohammad, s.f.). Y se puede preguntar si alguien en la sala no asocia ciertas palabras con las mismas emociones que se muestran en el sitio web.

Finalmente, el método está sesgado hacia un tipo de texto que podríamos denominar *estándar*. Por ejemplo, no vale la pena usar *syuzhet* para obtener el valor sentimental de un texto del siglo XV ni de un texto que dependa en gran medida de la representación del habla o de

dialectos; en ambos casos, se necesitará un diccionario específico. Y, por supuesto, siempre habrá que tener en cuenta la gran subjetividad de los textos literarios en general, la polisemia de las palabras, los contextos sociales, etc. de cada obra.

A pesar de estos problemas, que merecen ser discutidos, *syuhzet* permite a los usuarios cambiar de idioma fácilmente con solo una línea adicional de código, lo que lo hace ideal para contextos multilingües que usan caracteres latinos en la escritura (a la vez, un límite del paquete). Este experimento piloto iniciará una conversación. Ayudará a reconocer la conciencia de la diversidad lingüística y cultural entre los profesionales de las herramientas y métodos de HD al comienzo del proceso de aprendizaje en lugar de relegarlo a una última tarea en un proyecto.

5. CONCLUSIONES: VAMOS EN BUEN CAMINO

El análisis de sentimientos de textos es complicado. Más aún cuando tratamos de hacerlo en textos literarios. Fácilmente, necesitaríamos cientos, sino miles de bolsas de palabras dirigidas a conjuntos de corpus textuales particulares (por ejemplo, una para cada variedad regional del español). Las técnicas de aprendizaje automático o ML logran un mejor resultado que las bolsas de palabras, pero tiene una curva de aprendizaje mucho mayor. El paquete R que se utilizó aquí, *syuzhet*, pretende capturar la forma macro de una historia principalmente en inglés, pero puede ser un punto de partida para explorar el método en otros idiomas. El inglés-español es una pareja de idiomas que está altamente desarrollada para la traducción automática en la actualidad, y está claro que el método no funcionará tan bien en idiomas menos desarrollados en lo que hace a este método como, por ejemplo, el euskera. En este caso, y como ejemplo precisamente de la necesidad de desarrollar métodos específicos a diferentes idiomas, un equipo está desarrollando su propio método de minería de opinión (Alkorta et al., 2019).

Sin embargo, vamos en buen camino si reconocemos la presencia inevitable de la diversidad lingüística en el registro cultural digital, en las Humanidades Digitales en general y en particular en países plurilingües. Para derribar las barreras lingüísticas, como se discutió en el V Congreso de la Asociación Argentina de Humanidades Digitales de 2022, *Miradas desde el sur*[16], entre otras cosas, debemos hacer un mejor trabajo al pensar en el idioma que habla/lee/investiga la audiencia de un método, de colecciones digitales, o de una herramienta, tanto en el contexto más amplio de las audiencias globales, así como los contextos en los que estamos ubicados físicamente. En mi caso, ocupo una realidad académica con apariencia e imposición monolingüe ficticia, pues en realidad habito un mundo y un ámbito cultural multilingüe. La diversidad lingüística no existe sólo en el extranjero o en las HD no angloparlantes, por tanto, aunque con métodos imperfectos y a veces experimentos fallidos, debemos seguir mostrando la aplicación de métodos digitales en más idiomas.

REFERENCIAS BIBLIOGRÁFICAS

Alkorta, J., Gojenola, K., e Iruskieta, M. (2019). Towards discourse annotation and sentiment analysis of the Basque Opinion Corpus. En A. Zeldes, D. Das, E. Maziero Galani, J. D. Antonio y M. Iruskieta (Eds.), *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019* (pp. 144-152). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2718

Beigel, M. F. (2014). Publishing from the periphery: Structural heterogeneity and segmented circuits. The evaluation of scientific publications for tenure in Argentina's CONICET. *Current Sociology*, *62*(5), 743-765. https://doi.org/10.1177/0011392114533977

Beigel, M. F. (2019). Indicadores de circulación: una perspectiva multi-escalar para medir la producción científico-tecnológica latinoamericana. *Ciencia, Tecnología y Política*, *3*. https://doi.org/10.24215/26183188e028

[16] Sitio web del congreso: https://www.aacademica.org/aahd2022.

del Rio Riande, G., y Fiormonte, D. (2022). Una vez más sobre los sures de las digital humanities. *Acervo*, *35*(1). https://revista.an.gov.br/index.php/revistaacervo/article/view/1850/1711

Fiormonte, D. (2021). Taxation against overrepresentation?: The consequences of monolingualism for Digital Humanities. En D. Kim y A. Koh (Eds.), *Alternative historiographies of the Digital Humanities* (pp. 333-376). Punctum Books. https://www.jstor.org/stable/j.ctv1r7878x.13

Fradejas Rueda, J. M. (Ed.) (2022). *Cuentapalabras*. http://www.aic.uva.es/cuentapalabras/

Henríquez Miranda, C., y Guzman, J. (2017). A review of sentiment analysis in Spanish. *Tecciencia*, *12*(22), 39-49. http://dx.doi.org/10.18180/tecciencia.2017.22.5

Irizarry, E. (1997). *Informática y literatura*. Anthropos.

Isasi, J. (2021). Análisis de sentimientos en R con syuzhet. *Programming Historian en Español*, *5*. https://doi.org/10.46430/phes0051

Isasi, J., y del Rio Riande, G. (2022). ¿En qué lengua citamos cuando escribimos sobre Humanidades Digitales? *Revista de Humanidades Digitales*, *7*, 127-143. https://doi.org/10.5944/rhd.vol.7.2022.36280

Jockers, M. (2017). *syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text* (1.0.7) [Software]. CRAN. https://cran.r-project.org/web/packages/syuzhet/index.html

Mohammad, S. (s. f.). *NRC Emotion Lexicon*. Recuperado el 1 de octubre de 2023, de https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. En K. Zervanou, P. Lendvai (Eds.), *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 105-114). Association for Computational Linguistics. https://aclanthology.org/W11-1514

Mohammad, S. M., y Turney, P. D. (2013). Crowdsourcing a Word–Emotion Association Lexicon. *Computational Intelligence*, *29*(3), 436-465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Moreno-Ortiz, A. (2017). Lingmotif: Sentiment analysis for the Digital Humanities. En A. Moreno-Ortiz (Ed.), *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of*

the Association for Computational Linguistics (pp. 73-76). Association for Computational Linguistics. https://aclanthology.org/E17-3019

Mosteller, F., y Wallace, D. (1964). *Inference and disputed authorship: The Federalist*. Addison-Wesley.

Ortega, É. (2019). Zonas de contacto: A Digital Humanities ecology of knowledges. En M. K. Gold y L. F. Klein (Eds.), *Debates in the Digital Humanities 2019*. University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/4805e692-0823-4073-b431-5a684250a82d/section/aeee46e3-dddc-4668-a1b3-c8983ba4d70a

Priani Saisó, E., Spence, P., Russell, I. G., García, E. G. B., y Alves, D. (2014). Las humanidades digitales en español y portugués. Un estudio de caso: DíaHD/DiaHD. *Anuario Americanista Europeo*, *12*(2).

Risam, R. (2018). *New digital worlds: Postcolonial Digital Humanities in theory, praxis, and pedagogy*. Northwestern University Press.

Rodríguez Aldape, F. M. (2013). *Cuantificación del interés de un usuario en un tema mediante minería de texto y análisis de sentimiento.* [Tesis de Maestría, Facultad de Ingeniería Mecánica y Eléctricac, Universidad Autónoma de Nuevo León]. https://cd.dgb.uanl.mx/handle/201504211/5662

Siko, B., Sengupta, A., Allmann, K., y Pozo, C. (2018). *Decolonizing the internet*. Whose Knowledge? Recuperado el 1 de octubre de 2023, de https://whoseknowledge.org/wp-content/uploads/2018/11/DTI-2018-Summary-Report.pdf

Sivertsen, G. (2018). Balanced multilingualism in science. *BiD: Textos Universitaris de Biblioteconomia i Documentació*, *40*. https://doi.org/10.1344/BiD2018.40.25

Srinivasan, S., Comini, N., Koltsov, M., y Gelvanovska-Garcia, N. (2022). *Acceso y uso de Internet en América Latina y El Caribe. Resultados de las encuestas telefónicas de alta frecuencia ALC 2021.* Grupo Banco Mundial.

Vonnegut, K. (1994). *Welcome to the monkey house and palm sunday: An autobiographical collage*. Vintage.

Whose Knowledge?: Re-imagining and re-designing the internet to be for and by us all. (2021). Whose Knowledge? Recuperado el 1 de octubre de 2023, de https://whoseknowledge.org/wp-content/uploads/2021/04/WK-Prospectus-2021.pdf