COMPUTATIONAL TEXT ANALYSIS IN SPANISH/ARABIC TRANSLATION: FOUR EMPIRICAL STUDIES

Mohamed M. Moustafa

Facultad de Filosofía y Letras

Programa de Doctorado en Lingüística, Literatura y Traducción

Universidad de Málaga

TESIS DOCTORAL
2022

Supervisor: Dr. Nicolás Roser

AUTOR: Mohamed Mahmoud Moustafa

https://orcid.org/0000-0001-5456-7967

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

D. Nicolás Roser Nebot, Profesor del Departamento de Traducción e Interpretación de la Universidad de Málaga y Director de la tesis doctoral del alumno D. MOHAMED MAHMOUD MOUSTAFA titulada *COMPUTATIONAL TEXT ANALYSIS IN SPANISH/ARABIC TRANSLATION: FOUR EMPIRICAL STUDIES* inscrita en el programa de doctorado "Lingüística, literatura y traducción" de la Facultad de Filosofía y Letras de la Universidad de Málaga I N F OR M A que da su Visto Bueno para que la tesis sea registrada como paso previo a su lectura y defensa.

Y para que conste y surta los efectos oportunos ante quien proceda, firma el presente Informe en Málaga a diecisiete de febrero de dos mil veintidós.

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña MOHAMED M. MOUSTAFA
Estudiante del programa de doctorado TRADUCCIÓN de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: COMPUTATIONAL TEXT ANALYSIS IN SPANISH/ARABIC TRANSLATION: FOUR EMPIRICAL STUDIES

Realizada bajo la tutorización de NICOLÁS ROSER NEBOT y dirección de NICOLÁS ROSER NEBOT
(si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 17 de julio de 2022

Fdo.: Mohamed M. Moustafa Doctorando/a	Fdo.: NICOLÁS ROSER NEBOT Tutor/a

Informe del director sobre la idoneidad de presentar la presente Tesis por compendio de publicaciones

La tesis doctoral de D. MOHAMED MAHMOUD MOUSTAFA titulada *COMPUTATIONAL TEXT ANALYSIS IN SPANISH/ARABIC TRANSLATION: FOUR EMPIRICAL STUDIES* inscrita en el programa de doctorado "Lingüística, literatura y traducción" de la Facultad de Filosofía y Letras de la Universidad de Málaga, es una tesis elaborada por compendio de publicaciones. Estas publicaciones son cuatro trabajos de investigación publicados en las siguientes revistas científicas:

1) *Journal of Information Technology Case and Application Research*
2) *Journal of Language Teaching and Research*
3) *Journal of Intercultural Communication Research*

El tema subyacente en los artículos se encuentra en la interfaz que proporcionan la traducción y la minería o exploración de textos. En este sentido, la presentación de la tesis por compendio de publicaciones permite ir viendo, de manera nítida y escalonada, el desarrollo de la investigación llevada a cabo; y cómo en cada nuevo artículo ha habido una asimilación de los resultados de la fase previa del estudio, representada por el artículo anterior. En cada nuevo artículo esta asimilación va acompañada por una superación y/o una ampliación o nueva aplicación de los resultados obtenidos hasta ese momento, junto con la aparición de nuevas ideas, proporcionadas por la misma investigación, que ensanchan, cada vez más, los horizontes del análisis y, por tanto, de sus conclusiones.

Por otro lado, la tesis por compendio de publicaciones ofrece no sólo resultados validados por revistas científicas de prestigio y de impacto, sino también planteamientos igualmente validados por los comités científicos de esas revistas al exigirse una revisión por pares con anterioridad a la publicación del artículo. Ello supone asegurar el nivel no únicamente académico de los contenidos de la tesis sino, también y por añadidura, su validez de aplicación en el campo científico al que se adscriben. De esta manera, la publicación de los artículos que van conformando el cuerpo de la tesis van enriqueciendo el ámbito de las ideas y de los datos del tema que trata aquélla y, al mismo tiempo, recibe e incorpora las opiniones y críticas que cada una de sus publicaciones suscita, con lo que la perspectiva y los argumentos de la tesis se refuerzan y adquieren más consistencia y precisión.

Las cuatro publicaciones que conforman la tesis tienen como objeto analizar los sentimientos que se yerguen detrás del uso de ciertos vocablos españoles de origen árabe, algunos términos árabes referidos al islam y la propia palabra "árabe" en su uso en los escritos de tres autores de la Generación

del 98, movimiento literario del último tercio del siglo XIX y primer tercio del siglo XX.

Por otro lado y para realizar esta investigación, se han utilizado técnicas sofisticadas de minería de datos que son una completa novedad en este tipo de estudios, con lo que la tesis aúna la novedad de tema, de metodología y de presentación.

Por todas estas razones, el Director de la tesis considera su presentación por compendio de publicaciones como el formato más idóneo para llevarla a término.

En Málaga, a veinticinco de junio de 2022

Fdo.: Nicolás Roser Nebot.

Authorisation Letter

I, Nicolás Roser Nebot, consent as co-author Mr. Mohamed M. Moustafa to use all the scientific articles we published together in his PhD thesis in order to present it to discussion at UMA with the aim to obtain the PhD degree.

I pledge not to use those articles which compound Mohamed M. Moustafa' s PhD thesis in another PhD thesis, not in the University of Malaga and not in another university else.

For this purpose, I sign this Authorisation Letter in Malaga on June 24th 2022.

Signed: Nicolás Roser Nebot

Response to External Reviewers' Comments Report

I would like to thank the external reviewers for their constructive comments, which led to a great improvement in the substance and the presentation of the thesis. In response to the reviewers' comments, I have revised the thesis as followed.

External reviewer 1 (Dr. Akeel Almarai):

No comments

External reviewer 2 (Dr. Mourad Zarrouk):

Cambios obligatorios:

"Reescribir los textos en árabe que aparecen en las páginas 113-117."

Response:

Based on the reviewer's comments, all Arabic texts on the thesis (pp. 113-117) were corrected. Examples:

" ومن سـخرية القدر أن النجاح الضـئيل الذي حققه صـديقي ريبيرو مع السـيدات قد اعتمد على كونه أشـقر ، وهو ما يشيع بين السويسريين أو الفرنسين ، على حين كانت السيدات والفتيات ينتظرن فتى إسباني يتصرف بتلقائية وأسمر اللون مع ملامح عربية . وعلى الرغم من هذا الانطباع الأول ، فقد واصـل ريبيرو زيارة المنزل وتكوين صـداقات مع الجميع "

" وهي الضروريات التي تتيح لرجل أن يعيش في خيمة كعربي في الصحراء "

" وهي الضروريات التي تتيح لرجل أن يعيش في خيمة كعربي في الصحراء "

Index

CHAPTER 1

Introduction

This Doctoral Thesis is based on four research papers published in three journals: *Journal of Information Technology Case and Application Research, Journal of Language Teaching and Research and the Journal of Intercultural Communication Research.* The underlying theme of the papers lies at the interface of translation and text mining. For example, this first study is motivated by the important role played by mass media in shaping public opinion. The study aims to analyze sentiments expressed in traditional and social media towards terms and expressions coined in Spanish to refer to Islam. Given a large number of mass media in Spain, the study focuses on both the leading Spanish newspaper *El País* and on Twitter microblogs. For example, in the study examining the Sentiment analysis of Arabic language influence on Spanish vocabulary: An El País newspaper and Twitter case study, we investigated a corpus of Spanish words of Arabic terms appearing in the Spanish newspaper El País from January 2011 to December 2016. We selected a random sample of 20 terms and built our corpus around them. The vast amount of electronic texts available on the newspapers' web sites facilitated the creation of our corpora in a relatively short time. The El Pais corpus included 570,312 words. The Corpus satisfied several rigorous criteria such as authorship, topic, genre, and register (Biber, 1993). Our corpora size is comparable to several published studies, including Grabowski's (2013) study (705,460 words), Ferrero's (2011) study (692,751 words), and Merakchi and Rogers' (2013) study (288,306 words). We used a newspaper based corpus since it is argued that, compared to other modalities, the journalistic style has changed the most by the advent of the Internet (Bielsa 8c Bassnett, 2009). In fact, such corpora have been used in several studies, including Lobanova, Kleij, and Spenader (2010) who investigated the definition of antonyms using three Dutch daily newspapers, and Jones (2002) who used the British newspaper The Independent to build a corpus of 55,411 sentences to investigate the functions of synonyms and antonyms in context. Tse (2003) used a corpus including 514,691 words based on three British newspapers, namely The Independent, the Guardian, and The Daily

Telegraphy to investigate theoretical factors determining the use or omission of the definite article preceding multi-word organization names. Similarly, Baroni and Bernardini (2006) used a two-million-word Italian newspapers corpus to study the language properties of the "translationese". In this study, we also analyzed data representing a random set of Twitter posts from November 15, 2016 to December 15, 2016. The random sampling technique was used to eliminate potential bias arising from the influence of a specific event. The data comprised 4586 out of 45860 tweets generated. All retweets and duplicated tweets were eliminated. Sample selection has been varied by day of the week and hours in the day in order to guarantee representativeness. The sample is comparable in size to similar research studies. For example, Qiu et al. (2010) used a sample of 3783 opinion sentences.

Methodologically, the thesis is based on the use of stylometrics, Twitter analytics, and corpus analysis to extract and translate short texts from Spanish into Arabic. The focus throughout the thesis is on the use of sophisticated data mining techniques to conduct the analysis. For example, chapter 4 is based on the analysis of large corpora of *El País* newspaper along with a corpus representing a Twitter corpus. Hu and Kearney (2020) argued that since Twitter data "can be collected non-intrusively, it allows researchers to examine human behaviors in more ecologically valid or 'natural' settings, at least compared to survey-based methods" (pp. 489-490). Chapter 7 also uses large corpora representing the Generation of '98- a group of Spanish novelists, philosophers, essayist,s and poets active during the 1898 Spanish-American war. Outstanding figures of this group include the novelists P Baroja (1872-1956), Vicente Blasco Ibañez (1867-1928), and Ramón Mar del Valle-Inclan (1866-1936), the philosophers Miguel de Unamuno (1864-1936) and Jose Ortega y Gasset (1883-1955), and the poets Antonio Machado (1875-1939) and Manuel Machado (1874-1947), among others. More specifically, we analyze and translate all the occurrences of the word "árabe" as used in the corpora. We find that the word "árabe" has been used eight times in Baroja's 384,957 words corpus, eleven times in Ibanez's 1,118,883 words corpus and just once in Unamuno's 198,403 words corpus.

Research questions

Throughout this thesis, we aim to find answers to several important research questions. For example:

- Can digital opinion mining techniques be used to detect Arabic language impact on Spanish vocabulary? More specifically, can digital corpora and social networks' opinion mining techniques be used successfully to detect hidden patterns in layman's sentiment towards Spanish words of Arabic etymology related to Islam?

- What is the general sentiment of Spanish Twitter users towards the Arab image?

- What are the major topics discussed by the Spanish Twitter users in relation to the Arab image?

- Is there a reinforcing spiral linking media content exposure to the Arab image sentiment formed on Twitter?

Workflow and software

Computational text analysis has been applied extensively in fields as diverse as linguistics, media communication and political science (Welbers et al., 2017). Computational text analysis approaches are also known as text mining (Netzer et al., 2012), computer-aided text analysis (Pollach, 2012), computer-aided content analysis (Dowling & Kabanoff, 1996) or automated text analysis (Humphreys & Wang, 2017). Computational text analysis methods include, among others, sentiment analysis designed to gauge the tone of a specific text (Liu, 2015), topic modeling developed to extract meaningful “topics” from large corpora (Blei et al., 2003) and text scaling,

which aims at analyzing word frequencies to position political ideologies along a continuum ranging between liberal/conservative or left/right (Laver et al., 2003). Computational text analysis allows researchers to "identify patterns …that traditional analysis would not" (Boumans &Trilling, 2016, p. 9). I argue that using these methods can achieve several benefits. First, the collection of data may be automated through sophisticated techniques such as web crawling/web scraping. Second, the access to the massive amount of digital data could encourage answering newtypes of research questions. Finally, through this method, research replicability can be enhanced as other researchers can use the same data and code to reproduce the same results.

In the published work for the bulk of this thesis, we follow the general methodology usually applied in prior research, which involves four steps (Fesler et al., 2019; Ferguson-Cradler, Forthcoming; Madrid-Morales, 2020) as detailed below.

- **Data collection and storage**

Two major methods were used to collect data: web crawling and web scraping. Web crawling involves downloading the content of specific web pages, whereas web scraping involves "the targeted extraction of specific content from those downloaded pages" (Madrid-Morales, 2020, p. 75). The data collection phase usually involves three main steps. First, a list of relevant websites is established. For example, if the study involves building a corpus of Spanish major newspapers over a decade, a crawler can be used to visit all the URLs of the Spanish newspaper and extract all the available links. Second, a web scraper can be deployed to compile a relevant list of all stories available on the homepage at a specific point in time. In the final step the content of each news story is downloaded along with the meta-information of interest such as the title, the date and the full text.

The collected data are typically stored in a table-like format. In such format, rows represent cases (texts), whereas columns represent variables (date, author, etc.). R software provides some dedicated packages that can facilitate the data storage process. Examples include the "data.table" library (Dowle & Srinivasan, 2019), which is capable of reading/writing large amounts of data.

- **Data processing**

Although the data collection and storage might take considerable amount of time, the data processing phase can even be more time-consuming. This is because both digital and social media data are noisy and need extensive cleaning. In the studies reported here, we used several R packages to do the cleaning, including "tidytext" (Silge & Robinson), "quanteda" (Benoit et al., 2018), and "tm" (Feinerer et al., 2008). These packages follow the "bag-of-words" approach by focusing on the word frequencies and discarding word order. In the initial stage, the text is tokenized (reduced to units such as words). Usually, numbers and punctuations are deleted. Words are also converted into lower case and white spaces are deleted in this phase.

- **Data analysis**

Computational text analysis can generally be conducted using lexicon-based methods, supervised or unsupervised machine learning methods (Boumans & Trilling, 2016; Grimmer & Stewart, 2013; Welbers et al., 2017). Figure 1 shows an outline of the text-as-data methods. The lexicon method (dictionary) is based on the use of a pre-defined set of words/adjectives to be compared with the words/adjectives in the corpus. In the case of sentiment analysis, a score will be generated based on the score obtained for each word. lexicon-based methods have recently been used to examine topics as diverse as measuring policy uncertainty (Baker et al., 2016) and determining the content of online posts (Bettinger et al., 2016). In supervised machine learning methods, researchers train a model based on hand-coded documents. The trained model is then used to predict un-coded documents. Supervised machine learning methods have also been widely used in the literature (Kelly et al., 2018; Gegenheimer et al., 2018). Unlike supervised machine learning methods, unsupervised machine learning methods do not rely on prior hand-coding in identifying groupings or clusters in data. This makes such methods "particularly appropriate for researchers in a more descriptive and hypothesis-generating posture" (Fesler et al., 2019).

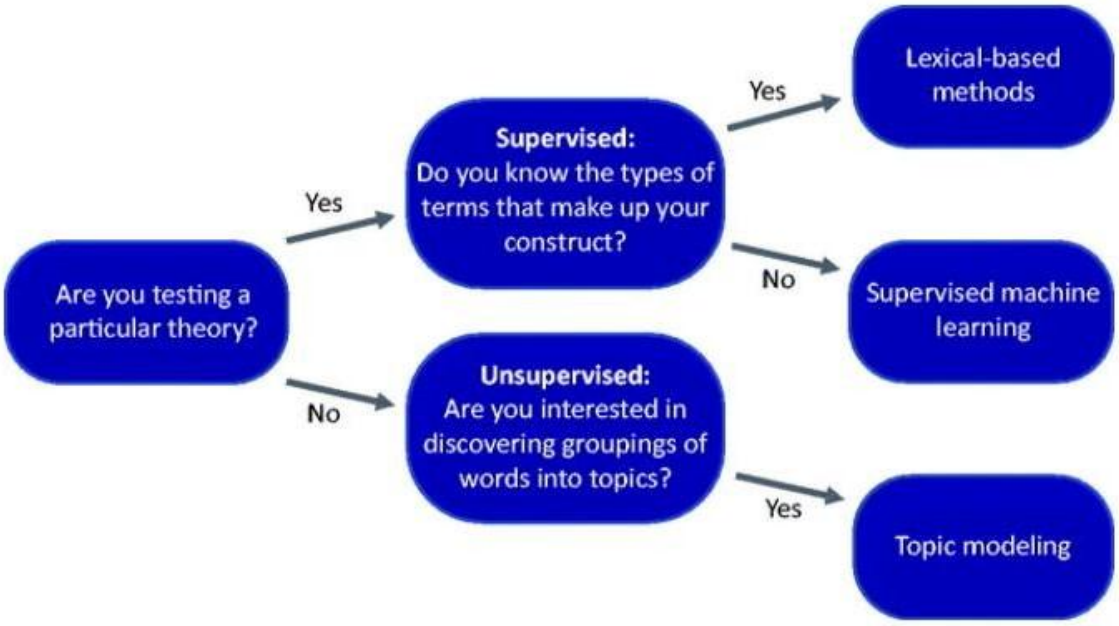

Figure 1. Text-as-data methods

Source: Fesler et al. (2019).

From the graph, we also see that computational text mining approaches can also use "topic modeling," which allows the researcher to (1) fathom the underlying latent thematic elements across a corpus; and (2) gauge the prevalence of each topic within each document. In Chapter 6 we use topic modeling to analyze the Arab image in Spanish social media. Results reveal that such images might be clustered into four main groupings. Based on topic modeling techniques conducted across the four clusters, the first cluster was labeled "cultural aficionados." This cluster comprises 7030 tweeters (70% of the sample). This is the largest cluster, and it is dominated by topics such as "Arab music" and "dance." This group of tweeters expresses an elevated level of interest in the Arab culture as reflected in music and dance. The second cluster included 510 tweeters (5% of the sample). This is the smallest cluster, and it is dominated by topics related to food. This cluster was termed "gastro aficionados." It seems that tweeters in this cluster represent individuals who express appreciation toward Arab cuisine. The third cluster had 1770 consumers (18% of the sample). Topics such as "panic" and "blood" dominate this cluster. This cluster was

termed “phobos” as it seems that this cluster is dominated by individuals who express fear or intense dislike towards the image of the Arab. Finally, the fourth cluster included 690 consumers (7% of the sample). Topics related to islamophobia dominate this cluster. This cluster was termed “islamophobos.” It seems that this group of tweeters express negative feelings towards Islam as a religion. Results also show that it seems that issues related to terrorism and islamophobia started to appear only recently. The Islamophobia topic shows a dramatic rise in 2017 with some tailing off later in 2018. It seems that this rise is closely related to recent Jihadist attacks in Europe, which started by the January 2015 Charlie Hebdo attacks, generally regarded as Frances’s 9/11 (Judaken, 2018), followed by the devastating July 2016 events in Nice, France, where an ISIS-affiliated rammed and killed eighty-six people, the Westminster’s attack in March 2017 and the Manchester’s Arena attack on May 22, 2017 killing 22 people. In the aftermath of the Manchester Arena attack, a British news website related to the Mail published an article detailing the massacre through 24 videos, including one by a far-right activist attacking journalists reluctant to link the event to “Islam” and “Muslims,” seven pictograms walking the reader through a “timeline of terror,” 115 pictures in which the reader is immersed in chaotic scenes (Gilks, Forthcoming). This result might be interpreted within the context of Slater’s (2007, 2015) Reinforcing Spirals Model (RSM). This model posits that communication and attitudes can often be mutually reinforcing. Thus, news coverage can support attitudes individuals already have, while such attitudes impact their communication choices. Similar results have recently been reported in the literature (Beam, Hutchens, & Hmielowski, 2018; Hutchens, Hmielowski, & Beam, 2019).

- **Data visualization**

Large corpora might be summarized and explored using graphical representations. For example, the ggplot2 package (Wickham, 2016) in the R environment can be used to provide a versatile and robust environment for data visualization. This package can produce almost all types of graphs such as histograms, boxplots, pie charts, and bubble charts. Computational text analysis results can also be visualized using co-occurrence networks, wordclouds, and lexical dispersion plots. Spatial data can also be summarized using choropleth maps.

Thesis Organization:

This thesis is organized as follows. We start by providing a summary of the results of all our published papers. The conclusions are presented in Chapter three. In the next five chapters we present the published paper in their original formal and we cite the source for each paper at the beginning of each chapter.

CHAPTER 2

Summary and Results

The first paper (Chapter 4) argues that although digital online media and social networks have become a valuable resource for mining sentiments, there is no previous research investigating the layman's sentiment towards Spanish words of Arabic etymology related to Islam. Thus, Chapter 2 fills this gap through an analysis of a digital corpus of the Spanish leading daily newspaper *El País*, including 570,312 words along with a random sample of 4,586 out of 45,860 tweets collected using a specialized software package. In this study an expert-predefined Spanish lexicon of around 6,800 seed adjectives was used to conduct the analysis. Results indicate a generally positive sentiment towards several Spanish words of Arabic etymology related to Islam. By implementing both a qualitative and quantitative methodology to analyze sentiments towards Spanish words of Arabic etymology, this case research adds breadth and depth to the debate over Arabic language influence on Spanish vocabulary.

The second paper (Chapter 5) analyzes sentiment towards Spanish words of Arabic origin related to Islam using a social network approach. The study notes that with the arrival of Muslims in 711 till their expulsion in the 1600s, Arabic language was present in Spain for more than eight centuries. Thus, it aimed at analyzing Spanish words of Arabic origin related to Islam. A random sample of 4586 out of 45860 tweets was used to evaluate general sentiment towards some Spanish words of Arabic origin related to Islam. An expert-predefined Spanish lexicon of around 6800 seed adjectives was used to conduct the analysis. Like the first study, results also indicated a generally positive sentiment towards several Spanish words of Arabic etymology related to Islam.

The third paper (Chapter 6) examines the Arab image in Spanish social media using a Twitter sentiment analytics approach. The paper starts by noting that intercultural relations scholars have recently focused on investigating the impact of news coverage and the media on society in general. However, although social media have become a valuable resource for mining opinions, there are no previous studies investigating the Arab image sentiments expressed on social media. This study fills this gap through the analysis of a random sample of 26,905 tweets dealing with the Arab image in Spanish tweets. To conduct the analysis, an expert predefined lexicon of around 6,800 seed adjectives was used. The descriptive analysis detected a generally positive sentiment towardsthe Arab image while partitioning around medoids (PAM) clustering suggested that it is possible to cluster the tweets into four distinct clusters. Thus, it seems that the Arab image represents a complex construction. By analyzing opinions expressed on social media towards the Arab image, this research adds breadth and depth to the debate over such an under-represented area.

The fourth paper (Chapter 7) is entitled "a corpus-based computational stylometric analysis of the word **"**árabe**"** in three Spanish Generación Del 98 writers." This paper argues that although the Generation of '98 represents a group of renown Spanish novelists, philosophers, essayists and poets active during the 1898 Spanish-American war, no previous studies have attempted to analyze the diverse linguistic and stylistic features employed by such writers. This study aims to use computational stylometry to detect hidden stylistic and linguistic patterns employed by three Generation of '98 writers, namely Pío Baroja, Vicente Blasco Ibáñez and Miguel de Unamuno. We employ a large corpus comprising 1,702,243 words representing nineteen works by the three writers. Several rigorous criteria were satisfied in designing the corpora such as authorship, genre, topic and register. Concordance, wordclouds, consensus trees, multidimensional and cluster analyses were performed to reveal the different stylistic and linguistic patterns used by the three writers. Although the study has focused solely on the use of the word "árabe", it showed that computational stylometry techniques can be used to help detect hidden stylistic and linguistic patterns employed by different writers. This result is significant since it can help the reader navigate across various possibilities of expressions and terminologies employed by different writers.

CHAPTER 3

Conclusions

The major conclusions of this thesis can be summarized as follows:

- We have established that digital opinion mining can help in comprehending the impact of Arabic language on Spanish vocabulary. We believe that our study contributes to the existing sentiment analysis literature as the analysis of online digital corpora allows for "real time" continuous monitoring of public opinion. Compared to traditional polls, our sentiment analysis overcomes problems associated with offline polls such as lack of respondents' commitment, low response rates, social desirability artifacts, and restricted set of choices available to respondents. The use of Twitter and online digital media mitigate almost all such problems since the opinions expressed by users of social media are rich in content and spontaneous in nature. Within this context, such corpora can increase the possibility of hidden patterns of data by understanding real preferences of individuals.

- We have also analyzed the sentiment polarity of social media tweets expressing sentiments towards twenty Spanish words of Arabic etymology related to Islamic terminology. Although any tweet is limited to 140 characters, millions of tweets posted on Twitter almost daily. Such tweets might provide an unbiased representation of individuals' sentiment towards a specific topic. Moreover, such tweets may be used by policymakers to gauge opinions regarding a specific issue. Ignoring such sentiments might put policy-makers on the defensive and could also create significant image problems. The speed of social media might also render politicians' efforts based on traditional media useless.

- Furthermore, we analyzed the sentiments expressed towards the Arab image in Spanish tweets.

We argue that although face-to-face interviewing and focus groups can be used to gauge individuals' sentiments, such traditional marketing research methods are both time-consuming and suffer from high costs. In contrast, individuals who generated online sentiments are free and readily available. Microblogs do not suffer from social desirability or interviewer bias as is the case in traditional data collection methods. Furthermore, sentiments expressed in tweets can easilybe benchmarked against objective measures such as traditional news metrics. Our results also showthat most Twitter users in our random sample use iPhone and Android for their tweets. Based on this result, public campaigns aiming at enhancing intercultural understanding on social media should use smartphones to deliver positive portrayals of minorities. Finally, our results show thatthe Islamophobia topic showed a dramatic rise in 2017 with some tail off later in 2018. This rise was linked to the sharp increase in Jihadist attacks in several European cities, indicating some support for Slater's Reinforcing Spirals Model (RSM). This result is also in line with several studies conducted in the West. For example, Arendt and Northup (2015) demonstrated that biased international news coverage creates stereotypes and anxiety amongst communities, which in turn influences gut feelings/implicit attitudes towards certain minority groups. Similarly, Shaver, Sibley, Osborne, and Bulbulia (2017) found that exposure to "media-induced Islamophobia" in New Zealand is associated with both reduced warmth and increased anger towards the Muslim minority in the country.

- Our analysis has also demonstrated that stylometry plays a major role in the study of style and

linguistic patterns. Stylometry can also be used to validate linguistic and stylistic hypotheses. For example, in one of the published articles, we used corpora of nineteen works representing three Spanish Generation of '98 writers to investigate their stylistic and linguistic differences. Our corpora were designed to satisfy several rigorous criteria such as genre, register, authorship, and topic. We argue that our computational stylometric approach might help in obtaining context-specific information regarding syntactic and semantic usage of the term "árabe" by Spanish Generation of '98 writers. Translators can exploit the corpora used in this study in several ways.

For example, they can refer to concordances to find a suitable translation for a particular term. Miangah (2012) argued that "adjectives that collocate with nouns have been proven to be very useful in understanding the context." This is particularly true when traditional dictionaries do not suggest a suitable translation. Boulton (2012) shows the inherent limitations of traditional dictionaries as opposed to corpora using the following French example «Je suis paralysé entre le brfflot et la chanson d'amour.» A dictionary offers the following possible meanings for the term "le brulot", "fire ship", "pamphlet", or "gnat". However, the author used a large corpus to show that a good translation would be "rebel", "revolutionary", or "protest". In fact, this is what Renaud, the famous French singer, is famous for. Thus, Wright (1993, p. 70) noted that "documents must speak 'the language' of the target audience and should resemble other texts produced within that particular language community and subject domain. These considerations frequently require that translators move beyond merely correct strategies in terms of lexical and grammatical content in order to account for stylistically appropriate solutions." This is probably true since a translator is basically a text producer. Thus, in the first place, a translator should be able to envisage how words are used and how they relate to other words in a particular context.

- Finally, our analysis shows that computer technology plays a central role in the study of vocabulary and translation. By using computers in extracting and analyzing meaningful data from general and specialized corpora, corpus linguistics can contribute to the field of translation by performing important tasks such as determining collocations, word clusters, sub-categorizations, and validating linguistic hypotheses. In fact, corpora play a major role in every aspect of translation studies, including theoretical, descriptive, contrastive, and teaching. For example, in our final published paper (Chapter 8), we used French and Spanish corpora of *Le Monde* and *El País* daily newspapers to investigate English business terms translations. Our corpora satisfied several rigorous criteria such as authorship, topic, genre, and register. We argue that using these corpora might help translators understand the context in which English business terminology is used. Specifically, we claim that our corpus-based analysis might help in (1) building better dictionary entries regarding the English business terms used in *Le Monde* and *El País* daily newspapers, (2) compensate for the poor representation of such terms in a traditional dictionary, and (3) obtain

context-specific information about syntactic and semantic usage of the terms used in Le Monde and *El País* corpora.

Publicaciones

(1) Mohamed M. Mostafa & Nicolás Roser Nebot (2017). Sentiment analysis of Arabic language influence on Spanish vocabulary: An El País newspaper and Twitter case study, Journal of Information Technology Case and Application Research, 19, 145-157.

https://www.tandfonline.com/doi/full/10.1080/15228053.2017.1394143

Abstract

Although digital online media and social networks have become a valuable resource for mining sentiments, there is no previous research investigating the layman's sentiment towards Spanish words of Arabic etymology related to Islam. This study fills this gap through the analysis of a digital corpus of the Spanish leading daily newspaper *El País*, including 570,312 words along with a random sample of 4,586 out of 45,860 tweets collected using a specialized software. An expert-predefined Spanish lexicon of around 6,800 seed adjectives was used to conduct the analysis. Results indicate a generally positive sentiment towards several Spanish words of Arabic etymology related to Islam. By implementing both a qualitative and quantitative methodology to analyze sentiments towards Spanish words of Arabic etymology, this case research adds breadth and depth to the debate over Arabic language influence on Spanish vocabulary.

(2) Mohamed M. Mostafa & Nicolás Roser Nebot (2017). Sentiment Analysis of Spanish Words of Arabic Origin Related to Islam: A Social Network Analysis. Journal of Language Teaching and Research, Vol. 8, No. 6, 1041-1049.

https://www.academypublication.com/issues2/jltr/vol08/06/03.pdf

Abstract

With the arrival of Muslims in 711 till their expulsion in the 1600s, Arabic language was present in Spain for more than eight centuries. Although social networks have become a valuable resource for mining sentiments, there is no previous research investigating the layman's sentiment towards Spanish words of Arabic etymology related to Islamic terminology. This study aims at analyzing Spanish words of Arabic origin related to Islam. A random sample of 4586 out of 45860 tweets was used to evaluate general sentiment towards some Spanish words of Arabic origin related to Islam. An expert-predefined Spanish lexicon of around 6800 seed adjectives was used to conduct the analysis. Results indicate a generally positive sentiment towards several Spanish words of Arabic etymology related to Islam. By implementing both a qualitative and quantitative methodology to analyze tweets' sentiments towards Spanish words of Arabic etymology, this research adds breadth and depth to the debate over Arabic linguistic influence on Spanish vocabulary.

(3) Mohamed M. Mostafa & Nicolás Roser Nebot (2020) The Arab Image in Spanish Social Media: A Twitter Sentiment Analytics Approach, Journal of Intercultural Communication Research, 49, 133-155.

https://www.tandfonline.com/doi/full/10.1080/17475759.2020.1725592

Abstract

Intercultural relations scholars have recently focused on investigating the impact of the news coverage and the media on society in general. However, although social media have become a valuable resource for mining opinions, there are no previous studies investigating the Arab image sentiments expressed on social media. This study fills this gap through the analysis of a random sample of 26,905 tweets dealing with the Arab image in Spanish tweets. To conduct the analysis, an expert-predefined lexicon of around 6,800 seed adjectives was used. Descriptive analysis detected a generally positive sentiment towards the Arab image while partitioning around medoids (PAM) clustering suggested that it is possible to cluster the tweets into four distinct clusters. Thus, it seems that the Arab image represents a complex construction. By analysing opinions expressed on social media towards the Arab image, this research adds breadth and depth to the debate over such an under-represented area.

(4) Mohamed M. Mostafa & Nicolás Roser Nebot (2018). A Corpus-based Computational Stylometric Analysis of the Word "Árabe" in Three Spanish Generación Del 98 Writers Journal of Language Teaching and Research, Vol. 9, No. 5, pp. 928-938.

https://www.academypublication.com/issues2/jltr/vol09/05/05.pdf

Abstract

Although the Generation of '98 writers represents a group of renown Spanish novelists, philosophers, essayists and poets active during the 1898 Spanish-American war, no previous studies have attempted to analyze the diverse linguistic and stylistic features employed by such writers. This study aims to use computational stylometry to detect hidden stylistic and linguistic patterns employed by three Generation of '98 writers, namely Pío Baroja, Vicente Blasco Ibáñez and Miguel de Unamuno. We employ a large corpus comprising 1,702,243 words representing nineteen works by the three writers. Several rigorous criteria were satisfied in designing the corpora such as authorship, genre, topic and register. Concordance, wordclouds, consensus trees, multidimensional and cluster analyses were performed to reveal the different stylistic and linguistic patterns used by the three writers. Although we focus solely on the use of the word "árabe", we show that computational stylometry techniques can be used to help detect hidden stylistic and linguistic patterns employed by different writers. This result is significant since it can help the reader navigate across various possibilities of expressions and terminologies employed by different writers.

Resumen de la tesis

La tecnología informática juega un papel central en el estudio del vocabulario, la lexicografía y la traducción. De hecho, ya en la década de 1980, Sinclair señaló que la capacidad de las computadoras para almacenar y escanear textos muy largos "proporciona una base casi objetiva sobre la cual se pueden observar los patrones del lenguaje" (Sinclair, 1984, p. 3). También señaló que la evidencia de la lingüística computacional proporcionada por las concordancias es bastante superior a cualquier otro método tradicional. Por lo tanto, tan pronto como los traductores hagan pleno uso de esta evidencia, "será imposible volver a confiar en técnicas precomputacionales" (Sinclair, 1985, p. 87).

En una línea similar, Chapelle (2001, p. 38) afirma que la tecnología informática ha resultado en lo que se puede describir como una "revolución del corpus". Mediante el uso de computadoras para extraer y analizar datos significativos de corpus generales y especializados, la lingüística de corpus puede contribuir al campo de la traducción al realizar tareas importantes como determinar colocaciones, grupos de palabras, subcategorizaciones y validar hipótesis lingüísticas (Miangah, 2012, p. 321). El análisis del discurso basado en corpus también puede generar nuevas preguntas de investigación, eliminar sesgos e identificar normas lingüísticas y valores atípicos (Baker, 2006, p. 17).

Una tendencia importante en la lingüística de corpus es el uso de corpus altamente especializados para lograr varios propósitos diferentes, entre los que destacan la enseñanza de la escritura técnica (Noguchi, 2004), el aprendizaje de idiomas (Miangah, 2012, p. 321) y la enseñanza de técnicas de traducción (Frankenberg-Garcia, 2012, p. 473). De esta manera, mediante el uso de tecnología informática moderna en la traducción, se pueden corregir los sesgos semánticos y lingüísticos. Por ejemplo, si examinamos los dos verbos "quake" y "quiver", que pertenecen a la misma clase semántica "shake" en dos diccionarios ampliamente utilizados como el COBUILD y el Longman, encontramos que ambos diccionarios afirman que los dos los verbos son intransitivos. Sin embargo, la entrada para "quake" en el diccionario de Oxford califica el verbo de intransitivo, mientras que el mismo diccionario enumera "quiver" como un verbo transitivo. Para resolver este dilema, Atkins y Levin (1995, p. 132) utilizaron un corpus inglés muy grande que comprendía más de 50 millones de palabras y pudieron detectar posibles usos de ambos verbos en formas transitivas como "it quakeed her guts" y "…quivering its wings." Por lo tanto, al usar un corpus lo suficientemente grande, parece posible corregir los errores de los diccionarios tradicionales y proporcionar evidencia de entradas precisas (McEnery y Wilson, 1996, p. 532).

La tecnología asistida por computadora también se puede usar para detectar distintas acepciones o matices de significado. Por ejemplo, Nelson (2005, p. 112) utilizó un corpus especializado en inglés para investigar el uso de dos términos que a menudo se usan indistintamente: “global” e “internacional”. El autor descubrió que existen diferencias claras en el uso, ya que la palabra "global" a menudo se coloca junto con términos como "actividades comerciales", mientras que la palabra "internacional" generalmente se coloca junto con términos como "empresas e instituciones". Esto contrastaba con la palabra "local" que se ubicaba junto con una gran cantidad de palabras no comerciales. El autor concluyó que, si bien las tres palabras comparten la misma clase semántica, la palabra “local” se ubica significativamente en una actividad claramente no comercial.

Probablemente una de las principales contribuciones de las tecnologías informáticas a los estudios de traducción ha sido la creación de lo que se denomina “corpus de monitor abierto”, que permite a los traductores realizar un seguimiento de las palabras nuevas que ingresan al idioma, de las palabras que cambian su significado habitual a lo largo del tiempo o de la adecuación en el uso de palabras basado en factores como la formalidad, el género, etc. Por lo tanto, la capacidad de usar *software* de computadora para detectar la colocación y los grupos/paquetes de palabras en lugar de la palabra en el plano individual significa que ahora es posible investigar sistemáticamente cómo las frases y las palabras distribuyen sus elementos en cada estructura que presentan y cómo tales colocaciones cambian con el tiempo (Castejon, 2012, p. 18).

Por otro lado, las redes sociales están cambiando la forma en que hablamos y nos comunicamos. No solo se están creando nuevas palabras, también estamos reutilizando palabras existentes. La mayoría de los sitios de redes sociales populares se originaron en los EE. UU., por lo que la mayoría de los componentes de la jerga y la terminología virtuales están en inglés; y a causa del uso particular del lenguaje en las redes sociales, la mayoría de estas palabras inglesas no necesariamente tienen equivalentes en otros idiomas. En otros idiomas europeos, en Facebook un botón muestra "like", en lugar de "me gusta". O piense en el verbo tuitear: algunos idiomas, como el árabe, al contrario que en español, tienen su equivalente acuñado, pero esto se relaciona con el gorgojeo de un pájaro, no con una publicación de Twitter. Una traducción literal de estas frases no tendrá el mismo significado o importancia que la versión en inglés.

La comunicación se ha convertido cada vez más en una actividad en línea, como lo indica la proliferación de dispositivos móviles y aplicaciones y plataformas en línea. Dado que la traducción es una forma de comunicación intercultural e interlingüística, se deduce que la actividad de traducción

también ha migrado al mundo virtual móvil. Si bien la investigación sobre la relación entre la web y la traducción se ha llevado a cabo desde las primeras iteraciones de la web en la década de 1990 (localización o *crowdsourcing* - externalización abierta de tareas -, por ejemplo), los aspectos específicos de esta vasta área de estudio aún deben investigarse y analizar, particularmente, las relaciones interdependientes y complejas entre la traducción y los medios sociales en línea (OSM, *Online Social Media* en inglés) o, en español, redes sociales.

La traducción de contenidos publicados en redes sociales (posts, comentarios, videos, etc.) plantea desafíos particulares para los traductores. En primer lugar, los traductores tienen que trabajar con longitudes de texto específicas, en particular con respecto a los tuits limitados a 280 caracteres en Twitter. Sin embargo, al traducir de un idioma a otro, la cantidad de palabras puede variar muy rápidamente. En consecuencia, el traductor no solo tendrá que transcribir la idea original al idioma de origen, sino también adaptar el texto a estas limitaciones técnicas.

Las redes sociales tienen otras peculiaridades como los *hashtags*, que suponen otro reto para los traductores. De hecho, no es recomendable tener solo una traducción directa de un *hashtag* original al traducir en las redes sociales. Se requiere de una investigación previa para encontrar los *hashtags* más relevantes en el país al que va destinada la traducción. A veces, puede surgir la duda de si se debe traducir o no el *hashtag*.

El tono utilizado en las redes sociales es muy específico en cada contexto virtual en el que aparece y varía de un medio a otro. También depende del público, de la empresa y de su sector de actividad que usa esa red social y de sus objetivos al utilizarla. Todos estos son factores que deben tenerse en cuenta en el proceso de traducción.

Traducir el contenido de los textos de las redes sociales es una tarea muy desafiante por muchas razones. En primer lugar, los tuits son textos muy breves y están escritos en formatos no estándar, ni formal. Asimismo, los usuarios de Twitter cometen muchos errores ortográficos y suelen expresar sus ideas en más de un idioma al mismo tiempo (Farzindar y Roche, 2013, p. 87).

La traducción de textos procedentes de redes sociales se convierte en una labor más compleja cuando se trata de un idioma morfológicamente rico como el árabe (Habash y Sadat, 2012, p. 138). Lo mismo sucede cuando los corpus paralelos para los tuits en la combinación árabe-inglés no están disponibles. En Salameh *et al*., 2015, p. 84; Mohammad *et al*., 2016, p. 564; Refaee y Rieser, 2015, p. 74, los autores recopilaron corpus paralelos para tuits en la combinación árabe-inglés pero estas recopilaciones de datos no fueron suficientes para construir un sistema de traducción automática eficiente para la traducción de los tuits entre el árabe y el inglés. Además, los datos recopilados en Jehl *et al*., 2012, p.

645, no son de código abierto.

La lengua árabe es conocida como una lengua morfológicamente rica y compleja. Las palabras árabes son palabras compuestas sobre una base de tres o cuatro fonemas consonantes a las que se añaden distintos fonemas vocales y otros fonemas consonantes en forma de prefijos, infijos y sufijos para crear las distintas flexiones, declinaciones y conjugaciones que la comunicación y la expresión en esa lengua precisan. Por lo tanto, una palabra en árabe puede corresponder a múltiples palabras en inglés lo que aumenta la tasa de términos que no se encuentran en el vocabulario usado habitualmente en una lengua natural (en inglés *OOV –Out Of Vocabulary- words* , es decir, palabras fuera del vocabulario -frecuente-) en un sistema de traducción automática estadística –SMT: Statistical Machine Translation en inglés- (Sadat y Habash, 2006, 83).

La compleja morfología de la lengua árabe junto con la subespecificación crea un alto grado de ambigüedad. Esta ambigüedad aumenta todavía más para el árabe utilizado en los textos de las redes sociales.

Recientemente aparecieron muchos problemas en la difusión de las plataformas de redes sociales. Los textos de los microblogs, por ejemplo, Twitter, son breves, intrusivos ("ruidosos" en inglés) y están escritos en un estilo ortográfico no estándar ni formal. De hecho, los usuarios de las redes sociales tienden a cometer errores ortográficos. Por ejemplo, suelen escribir palabras árabes utilizando alfabetos derivados del latín y números árabes, siendo que cada usuario translitera palabras árabes a su manera. Este fenómeno se denomina "Arabizi" (Bies et al., 2014, p. 219; Darwish, 2014, p. 34; Adouane et al., 2016, p. 372). Por ejemplo, la letra árabe 'ح' [ḥ] a menudo se translitera con el número '7', la letra 'ق' [q] con el número '9', etc. En los tuits, los usuarios traducen el nombre propio "احمد" [Aḥmad] de diferentes

maneras: "ahmed", "ahmad", "ahmd", "a7mad", "a7med", "a7mmd", "a7md" o "ahmmd" (Mubarak y Abdelali, 2016, p. 248).

Los tuits son cortos, intrusivos y tienen una estructura compleja. Contiene diferentes campos especiales como el nombre de usuario, *hashtags*, *retuitear*, etc. De hecho, los nombres de usuario han sido objeto de solicitudes de procesamiento en tanto que entidades nombradas en lengua árabe en (Mubarak y Abdelali, 2016, p. 249). También los *hashtags* han sido objeto de una aplicación de traducción automática mejorada en (Gotti et al., 2013, p. 63).

Esta Tesis doctoral se basa en cuatro trabajos de investigación publicados en tres revistas academicas: *Journal of Information Technology Case and Application Research*, *Journal of Language Teaching and Research* y *Journal of Intercultural Communication Research*. El tema subyacente de los

documentos se encuentra en la interfaz de la traducción y la minería de textos. Por ejemplo, este primer estudio está motivado por el importante papel que juegan los medios de comunicación en la formación de la opinión pública. El estudio tiene como objetivo analizar los sentimientos expresados en los medios tradicionales y sociales hacia los términos y expresiones acuñadas en español para referirse al islam. Dado el gran número de medios de comunicación en España, el estudio se centra tanto en el diario líder español *El País* como en los microblogs de Twitter.

En el estudio que examina el análisis de sentimiento junto a la influencia del idioma árabe en el vocabulario español en textos del periódico *El País* y en mensajes de Twitter (capítulo 4), investigamos un corpus de palabras en español de términos de origen árabe y/o relacionados con el islam que aparecieron en el periódico español *El País* desde enero de 2011 hasta diciembre de 2016. Seleccionamos una muestra aleatoria de 20 términos de etimología árabe y construimos nuestro corpus con ellos. La gran cantidad de textos electrónicos disponibles en los sitios web del periódico facilitó la creación de nuestros corpus en un tiempo relativamente breve. El corpus de El País incluyó 570.312 palabras. El corpus cumplió con varios criterios como autoría, tema, género y registro (Biber, 1993). El tamaño de nuestro corpus es comparable al de varios estudios publicados, incluido el estudio de Grabowski (2013) (705 460 palabras), el estudio de Ferrero (2011) (692 751 palabras) y el estudio de Merakchi y Rogers (2013) (288 306 palabras). Utilizamos un corpus basado en textos periodísticos ya que, en comparación con otras modalidades, el estilo periodístico ha cambiado en mayor medida con la llegada de Internet al adoptar más recursos expresivos procedentes de la lengua oral para introducirlos en la lengua escrita (Bielsa 8c Bassnett, 2009, p. 194). Y dado que los recursos expresivos de la lengua oral contienen un mayor grado de sentimiento, los textos periodísticos de ahora, aunque también –en menor medida- de antes, reflejan con más intensidad cualquier sentimiento o emoción y, por ende, usan términos de fácil uso en estudios de la índole del nuestro.

De hecho, los corpus que se han utilizado en varios estudios lingüísticos, incluidos Lobanova, Kleij y Spenader (2010, p. 549), que investigaron la definición de antónimos utilizando tres diarios holandeses, y Jones (2002), que utilizó el periódico británico *The Independent* para *recopilar* un corpus. de 55.411 oraciones para investigar las funciones de sinónimos y antónimos en contexto. Tse (2003) utilizó un corpus que incluía 514.691 palabras basado en tres periódicos británicos, a saber, *The Independent*, *The Guardian* y The *Daily Telegraph* para investigar los factores teóricos que determinan el uso o la omisión del artículo definido que precede a los nombres de organizaciones que constan de varios términos.

Otro aspecto de nuestro estudio tratado en el capítulo 5 (*Sentiment analysis of Spanish words of Arabic origin related to Islam: A social network analysis*), ha consistido en analizar los datos proporcionados por un conjunto aleatorio de publicaciones de Twitter del 15 de noviembre de 2016 al 15 de diciembre de 2016. La técnica de muestreo aleatorio se ha utilizado para eliminar el posible sesgo que surge de la influencia de un evento específico. Los datos recopilados han comprendido 4586 de 45860 tweets generados. Se eliminaron todos los retuits y tuits duplicados. La selección de la muestra ha variado según el día de la semana y las horas del día para garantizar la representatividad. La muestra es comparable en tamaño a estudios de investigación similares. Por ejemplo, Qiu et al. (2010) utilizó una muestra de 3783 oraciones de opinión.

Metodológicamente, la tesis se basa en el uso de estilometría, análisis de Twitter y análisis de corpus para extraer y traducir textos breves del español al árabe. El enfoque a lo largo de la tesis está en el uso de técnicas sofisticadas de minería de datos para realizar el análisis.

El capítulo 4 (*Sentiment analysis of Arabic language influence on Spanish vocabulary: An El País newspaper and Twitter case study*) se basa en el análisis un amplio corpus extraído del diario *El País* junto con otro corpus que representa un corpus de Twitter. Hu y Kearney (2020) afirman que, dado que los datos de Twitter "pueden recopilarse de manera no intrusiva, permite a los investigadores examinar los comportamientos humanos en entornos más ecológicamente válidos o 'naturales', al menos en comparación con los métodos basados en encuestas" (págs. 489- 490). El capítulo 7 (*A corpus-based computational stylometric analysis of the Word "Árabe" in three Spanish Generación Del 98 Writers*) también utiliza un gran corpus extraído de las obras de los novelistas Pío Baroja (1872-1956), Vicente Blasco-Ibáñez (1867-1928) y Ramón Mar del Valle-Inclán (1866-1936) y el novelista y filósofo Miguel de Unamuno (1864-1936). Más concretamente, analizamos y traducimos todas las apariciones de la palabra "árabe" en los distintos sentidos en los que se utiliza en el corpus seleccionado. Encontramos que la palabra "árabe" ha sido utilizada ocho veces en el corpus de 384.957 palabras de Baroja, once veces en el corpus de 1.118.883 palabras de Ibáñez y una sola vez en el corpus de 198.403 palabras de Unamuno.

A lo largo de esta tesis, nuestro objetivo ha sido encontrar respuestas a la siguientes preguntas:

a) ¿Se pueden utilizar técnicas de minería de opinión digital para detectar el impacto del idioma árabe en el vocabulario español?
b) ¿Pueden utilizarse con éxito las técnicas de minería de opinión de los corpus digitales y las redes sociales para detectar patrones ocultos en el sentimiento de los legos hacia las palabras españolas de etimología árabe relacionadas con el islam?

c) ¿Cuál es el sentimiento general de los tuiteros españoles hacia la imagen de lo árabe?
d) ¿Cuáles son los principales temas tratados por los tuiteros españoles en relación con la imagen de lo árabe?
e) ¿Existe una espiral de refuerzo que vincule la exposición del contenido de los medios con el sentimiento de imagen árabe formado en Twitter?

(El modelo de espirales de refuerzo (RSM) de Slater (2007, 2015) postula que la comunicación y las actitudes a menudo pueden reforzarse mutuamente).

Flujo de trabajo analítico y software

El análisis computacional de textos se ha aplicado ampliamente en campos tan diversos como la lingüística, los medios de comunicación y las ciencias políticas (Welbers et al., 2017). Los enfoques de análisis de texto computacional también se conocen como

a) minería de texto (Netzer et al., 2012, p. 651)
b) análisis de texto asistido por computadora (Pollach, 2012, p. 396)
c) análisis de contenido asistido por computadora (Dowling & Kabanoff, 1996, p. 72)
d) análisis de texto automatizado (Humphreys & Wang, 2017, p. 45)

Los métodos de análisis de texto computacional incluyen, entre otras aplicaciones, el análisis de sentimiento diseñado para medir el tono de un texto específico (Liu, 2015, p. 384), modelado de temas desarrollado para extraer "temas" significativos de grandes corpus (Blei et al., 2003, p. 169) y escalado de texto, cuyo objetivo es analizar frecuencias de palabras para posicionar ideologías políticas a lo largo de un continuo que oscila entre liberal/conservador o izquierda/derecha (Laver et al., 2003, p. 564).

El análisis de texto computacional permite a los investigadores "identificar patrones... que el análisis tradicional no podría" (Boumans & Trilling, 2016, p. 9). La recopilación de datos puede automatizarse a través de técnicas sofisticadas, como el rastreo web (*web tracking*) o el raspado web (*web scraping*). En segundo lugar, el acceso a una cantidad masiva de datos digitales permite fomentar la respuesta a nuevos tipos de preguntas de investigación. Por otra parte, a través de este método, se puede mejorar la replicabilidad de la investigación ya que otros investigadores pueden usar los mismos datos y código para reproducir los mismos resultados.

En los trabajos publicados para la elaborar mayor parte de esta tesis, hemos seguido la metodología general generalmente aplicada en investigaciones anteriores, que consta de cuatro pasos (Fesler et al.,

2019, p. 176; Ferguson-Cradler, Forthcoming; Madrid-Morales, 2020, p. 785) como se detalla a continuación.

Recopilación y almacenamiento de datos

Se han utilizado dos métodos principales para recopilar datos: el rastreo web y el raspado web. El rastreo web implica la descarga del contenido de páginas web específicas, mientras que el raspado web consiste en "la extracción dirigida de contenido específico de esas páginas descargadas" (Madrid-Morales, 2020, p. 75). La fase de recopilación de datos abarca tres pasos principales. Primero, se establece una lista de sitios web relevantes. Si el estudio necesita elaborar un corpus de los principales periódicos españoles durante una década, se puede utilizar un rastreador para visitar todas las URL del periódico español y extraer todos los enlaces disponibles. En segundo lugar, se realiza un *web scraping* (raspado de web) mediante programas software al uso para compilar una lista relevante de todos los contenidos (noticias, artículos, comentarios, …) disponibles en la página de inicio en un momento específico. En el paso final, se descarga el contenido de cada noticia junto con la metainformación de interés como el título, la fecha y el texto completo.

Los datos recopilados generalmente se almacenan en un formato similar a una tabla. En tal formato, las filas representan casos (textos), mientras que las columnas representan variables (fecha, autor, etc.). El software R proporciona algunos paquetes específicos que pueden facilitar el proceso de almacenamiento de datos. Los ejemplos incluyen la biblioteca "data.table" (Dowle & Srinivasan, 2019, p. 542), que es capaz de leer/escribir grandes cantidades de datos.

Procesamiento de datos

Aunque la recopilación y el almacenamiento de datos pueden llevar una cantidad de tiempo considerable, la fase de procesamiento de datos puede llevar incluso más tiempo. Esto se debe a que tanto los datos digitales o digitalizados existentes en la web como los de las redes sociales son intrusivos y necesitan una limpieza exhaustiva. En los estudios que hemos empleado para nuestro estudio, hemos usado varios paquetes R para hacer la limpieza, incluidos "tidytext" (Silge & Robinson), "quanteda" (Benoit et al., 2018) y "tm" (Feinerer et al., 2008). Estos paquetes siguen el enfoque de "bolsa de palabras" al centrarse en el índice de frecuencia de las palabras y descartar el orden de las palabras. En la etapa inicial, el texto es *tokenizado* (reducido a unidades como palabras). Por lo general, los números y los signos de puntuación se eliminan. Las palabras también se convierten a minúsculas y los espacios en blanco se eliminan en esta fase.

Análisis de los datos

El análisis de los textos computacionales se puede realizar utilizando tanto métodos basados en léxico como métodos de aprendizaje automático supervisados o no supervisados (Boumans & Trilling, 2016; Grimmer & Stewart, 2013; Welbers et al., 2017). Los métodos basados en el léxico (diccionarios) se basa en el uso de un conjunto predefinido de sustantivos/adjetivos para comparar con los sustantivos/adjetivos del corpus. En el caso del análisis de sentimiento, se genera una puntuación con base en la puntuación obtenida por cada palabra. Los métodos basados en el léxico se han utilizado recientemente para examinar temas tan diversos como la medición de la incertidumbre política (Baker et al., 2016) y la determinación del contenido de las publicaciones en línea (Bettinger et al., 2016). En los métodos de aprendizaje automático supervisado, los investigadores establecen, por medio de ensayos, un modelo basado en documentos codificados a mano. Luego, el modelo así establecido y ensayado se usa para predecir documentos no codificados. Los métodos de aprendizaje automático supervisado también se han utilizado ampliamente en la literatura específica sobre el tema (Kelly et al., 2018; Gegenheimer et al., 2018). A diferencia de los métodos de aprendizaje automático supervisados, los métodos de aprendizaje automático no supervisados no se basan en la codificación manual previa para identificar agrupaciones o grupos de palabras en los datos. Esto hace que dichos métodos sean "particularmente apropiados para investigadores en una postura más descriptiva y generadora de hipótesis" (Fesler et al., 2019).

Los enfoques basados en la minería de los textos computacionales también pueden usar propuestas de modelos por temas, lo que permite al investigador (1) comprender los elementos temáticos latentes subyacentes en un corpus y (2) medir la prevalencia de cada tema dentro de cada documento. En el capítulo 6 (*The Arab image in Spanish social media: A Twitter sentiment analysis approach*) utilizamos el modelado de temas para analizar la imagen de lo árabe en las redes sociales españolas. Los resultados revelan que tales imágenes pueden agruparse en cuatro grupos principales.

Según las técnicas de modelado de temas realizadas en los cuatro grupos, el primer grupo ha sido denominado "aficionados culturales". Este clúster está compuesto por 7030 tuiteros (70% de la muestra). Este es el grupo más grande y está dominado por temas como "música árabe" y "danza". Este grupo de tuiteros expresa un elevado nivel de interés por la cultura árabe reflejada en la música y la danza. El segundo grupo incluye 510 tuiteros (5% de la muestra). Este es el grupo más pequeño y está dominado por temas relacionados con la comida. Este grupo se ha denominado “aficionados gastronómicos”. Parece que los tuiteros de este grupo representan a personas que expresan su aprecio

por la cocina árabe. El tercer conglomerado tiene 1770 consumidores (18% de la muestra). Temas como “pánico” y “sangre” dominan este grupo. Este grupo se denominó "fóbicos", ya que parece que este grupo está dominado por individuos que expresan miedo o una intensa aversión hacia la imagen del árabe y de lo árabe. Finalmente, el cuarto conglomerado incluye a 690 usuarios de la plataforma (7% de la muestra). Los temas relacionados con la islamofobia dominan este grupo. Este grupo se denominó "islamófobos". Parece que este grupo de tuiteros expresan sentimientos negativos hacia el Islam como religión. Los resultados también muestran que parece que los problemas relacionados con el terrorismo y la islamofobia comenzaron a aparecer recientemente. El tema de la islamofobia muestra un aumento drástico en 2017, con cierta disminución a finales de 2018. Parece que este aumento está estrechamente relacionado con los, en ese momento, recientes ataques yihadistas en Europa, que comenzaron con los ataques a Charlie Hebdo de enero de 2015, generalmente considerados como el 11 de septiembre de Frances (Judaken , 2018), seguido del trágico suceso de julio de 2016 en Niza, Francia, donde un miembro de ISIS embistió y mató a ochenta y seis personas; el ataque de Westminster en marzo de 2017 y el ataque al Manchester's Arena el 22 de mayo de 2017, del que resultaron muertas 22 personas. Después del ataque al Manchester’s Arena, el sitio web de noticias del periódico británico *Daily Mail* publicó un artículo en el que detalla la masacre a través de 24 videos, incluido uno de un activista de extrema derecha que ataca a los periodistas reacios a vincular el evento con el "Islam" y los "musulmanes". Siete pictogramas que conducen al lector a través de una "línea de tiempo del terror", 115 imágenes en las que el lector se sumerge en escenas caóticas (Gilks, de próxima publicación). Este resultado podría interpretarse en el contexto del modelo de espirales de refuerzo (RSM) de Slater (2007, 2015). Este modelo postula que la comunicación y las actitudes a menudo pueden reforzarse mutuamente. Por lo tanto, la cobertura de noticias puede respaldar las actitudes que las personas ya tienen, mientras que tales actitudes afectan sus elecciones de comunicación. Recientemente se han constatado resultados similares en la literatura (Beam, Hutchens y Hmielowski, 2018; Hutchens, Hmielowski y Beam, 2019).

Visualización de datos

Los grandes corpus lexicográficos u de otro tipo pueden resumirse y explorarse mediante representaciones gráficas. Por ejemplo, el paquete ggplot2 (Wickham, 2016) en el entorno R se puede utilizar para proporcionar un entorno robusto y versátil para la visualización de datos. Este paquete puede producir casi todos los tipos de gráficos, como histogramas, diagramas de caja, gráficos circulares y gráficos de burbujas. Los resultados del análisis de un texto computacional también se

pueden visualizar utilizando redes de concurrencia o *co-ocurrencia*, nubes de palabras y diagramas de dispersión léxica. Los datos espaciales también se pueden resumir utilizando mapas de *coropletas*, que son cartografías de carácter cuantitativo que sirven para mostrar gráficamente fenómenos discretos mediante símbolos superficiales en relación directa y proporcional con su valor .

Así, para cumplir con los objetivos de la tesis, en la primera investigación (capítulo 4: *Sentiment analysis of Arabic language influence on Spanish vocabulary: An El País newspaper and Twitter case study*) se demuestra que si bien los medios digitales en línea y las redes sociales se han convertido en un recurso valioso para la minería de sentimientos, no existe una investigación previa que investigue el sentimiento del profano hacia las palabras españolas de etimología árabe relacionadas con el islam. Por ello, este capítulo 4 llena este vacío a través del análisis de un corpus digital extraído de *El País*, que es en la actualidad el principal diario español. Dicho corpus incluye 570.312 palabras junto a una muestra aleatoria de 4.586 de 45.860 tuits recopilados mediante un paquete de software especializado. En este estudio, se ha utilizado un léxico español predefinido por expertos de alrededor de 6.800 lexemas adjetivos para realizar el análisis. Los resultados indican un sentimiento generalmente positivo hacia varias palabras españolas de etimología árabe relacionadas con el islam. Al implementar una metodología tanto cualitativa como cuantitativa para analizar los sentimientos hacia las palabras españolas de etimología árabe relativas al islam, esta investigación de caso agrega amplitud y profundidad al debate sobre la influencia del idioma árabe en el vocabulario español.

El segundo artículo producto de la investigación realizada en esta tesis doctoral (capítulo 5: *Sentiment analysis of Spanish words of Arabic origin related to Islam: A social network analysis*) analiza de nuevo el sentimiento hacia las palabras españolas de origen árabe relacionadas con el Islam pero desde el enfoque con el que aparece en redes sociales. El estudio señala que desde la llegada de los musulmanes en el año 711 hasta su expulsión en el siglo XVII, la lengua árabe estuvo presente en España durante más de ocho siglos. Así, pues el objetivo del trabajo se propuso analizar palabras españolas de origen árabe relacionadas con el islam en las redes sociales. Se ha manejado una muestra aleatoria de 4586 de 45860 tuits para evaluar el sentimiento general hacia algunas palabras españolas de origen árabe relacionadas con el islam. Se utilizó un léxico español predefinido por expertos de alrededor de 6800 lexemas adjetivos para realizar el análisis. Al igual que el primer estudio, los resultados también indicaron un sentimiento generalmente positivo hacia varias palabras españolas de etimología árabe relacionadas con el Islam.

El tercer artículo fruto de la investigación doctoral (capítulo 6: *The Arab image in Spanish social media: A Twitter sentiment analysis approach*) examina la imagen de lo árabe en las redes sociales

españolas a partir del análisis del sentimiento en la plataforma Twitter. El documento comienza señalando que los estudiosos de las relaciones interculturales se han centrado recientemente en investigar el impacto de la cobertura de noticias y los medios de comunicación en la sociedad en general. Sin embargo, aunque las redes sociales se han convertido en un recurso valioso para extraer opiniones, no existen estudios previos que investiguen los sentimientos de la imagen de lo árabe expresados en las redes sociales. El estudio que nuestro articulo contiene llena este vacío a través del análisis de una muestra aleatoria de 26.905 tuits que tratan sobre la imagen árabe en los tuits españoles. Para realizar el análisis, se utilizó un léxico predefinido por expertos de alrededor de 6.800 lexemas adjetivos. El análisis descriptivo detectó un sentimiento generalmente positivo hacia la imagen de lo árabe, mientras que la partición en torno a la agrupación de *medoides* (PAM), que son los índices de desviación o diferenciacíon media de un objeto de estudio con respecto a los otros objetos de su mismo grupo, sugiere que es posible agrupar los tuits en cuatro grupos distintos. De este modo, la imagen de lo árabe en Twitter presenta una construcción compleja. Al analizar las opiniones expresadas en las redes sociales sobre la imagen de lo árabe, esta investigación agrega amplitud y profundidad al debate sobre un área tan subrepresentada.

El cuarto trabajo de investigación publicado como artículo (capítulo 7: *A corpus-based computational stylometric analysis of the word "Árabe" in three Spanish Generación Del 98 Writers*) evidencia que, si bien la Generación del 98 representa a un grupo de reconocidos novelistas, filósofos, ensayistas y poetas españoles activos durante la guerra hispanoamericana de 1898, ningún estudio previo ha intentado analizar los diversos rasgos lingüísticos y estilísticos empleados por tales escritores, al menos, en el plano cuantitativo y del sentimiento. Este estudio pretende utilizar la estilometría computacional para detectar patrones estilísticos y lingüísticos ocultos empleados por tres escritores de la Generación del 98, a saber, Pío Baroja, Vicente Blasco Ibáñez y Miguel de Unamuno. Empleamos un gran corpus compuesto por 1.702.243 palabras que abarcaron diecinueve obras de los tres escritores. En el diseño de los corpus se aplicaron varios criterios rigurosos como autoría, género, tema y registro. Se realizaron análisis de concordancia, nubes de palabras, árboles de consenso y multidimensionales y de conglomerados para revelar los diferentes patrones estilísticos y lingüísticos utilizados por los tres escritores. Aunque el estudio se ha centrado únicamente en el uso de la palabra "árabe", mostró que las técnicas de estilometría computacional pueden usarse para ayudar a detectar patrones estilísticos y lingüísticos ocultos empleados por diferentes escritores. Este resultado es significativo ya que puede ayudar al lector a navegar a través de varias posibilidades de expresiones y terminologías empleadas por diferentes escritores.

Conclusiones

Las principales conclusiones de esta tesis se pueden resumir en los puntos que se detallan a continuación:

- Hemos establecido que la minería de opinión digital puede ayudar a comprender el impacto del idioma árabe en el vocabulario español y su uso actual. Nuestro estudio contribuye a la literatura de análisis del sentimiento existente, ya que el análisis de los corpus digitales en línea permite un seguimiento continuo de la opinión pública en "tiempo real". En comparación con las encuestas tradicionales, nuestro análisis de opinión supera los problemas asociados a las encuestas fuera de línea, que son la falta de compromiso de los encuestados, las bajas tasas de respuesta, el sesgo de deseabilidad social (deseo de ofrecer respuestas o comportamientos valiosos) del encuestado frente al encuestador y el conjunto restringido de opciones disponibles para los encuestados. El uso de Twitter y los medios digitales en línea mitiga la mayoría de estos problemas o reduce su impacto, ya que las opiniones expresadas por los usuarios de las redes sociales son ricas en contenido y de naturaleza espontánea. En este contexto, dichos corpus aumentan la posibilidad de descubrir secuencias y patrones de datos, en principio ocultos, al poder comprender de manera más minuciosa las preferencias reales de los individuos.

- También hemos analizado la polaridad del sentimiento en los tuits de las redes sociales que expresan sentimientos hacia veinte palabras españolas de etimología árabe relacionadas con la terminología islámica. Aunque cualquier tuit está limitado a 140 caracteres, se publican a diario millones de tuits en Twitter. Dichos tuits suministran una representación sumamente ajustada del sentimiento de las personas hacia un tema específico. En este sentido, los legisladores y gobernantes pueden utilizar dichos tuits para conocer y evaluar las opiniones sobre un tema específico. Ignorar tales sentimientos puede poner, y de hecho lo hace, a los responsables políticos en situaciones comprometidas y puede suscitar, también, un deterioro de su imagen pública. La velocidad de las redes sociales convierte en inútiles una parte nada despreciable de la actividad de los políticos cuya publicidad y resonancia se centre en los medios de comunicación tradicionales.

- Nuestra investigación ha analizado también los sentimientos expresados hacia la imagen lo

árabe en los tuits españoles. Aunque las entrevistas cara a cara y los grupos focales o de enfoque se pueden usar para medir los sentimientos de las personas, estos métodos tradicionales de investigación de mercados requieren mucho tiempo y tienen altos costos. Por el contrario, las intervenciones de personas que generaron sentimientos en línea son gratuitas, de fácil acceso y gran disponibilidad. Los microblogs no están sujetos a las convenciones sociales o al sesgo del entrevistador como es el caso de los métodos tradicionales de recopilación de datos. Además, los sentimientos expresados en tuits se pueden comparar fácilmente con medidas objetivas como las métricas de noticias tradicionales. Nuestros resultados también muestran que la mayoría de los usuarios de Twitter en nuestra muestra aleatoria usan iPhone y Android para sus tuits. Sobre la base de este resultado, las campañas públicas destinadas a mejorar el entendimiento intercultural en las redes sociales podrían utilizar los teléfonos inteligentes para ofrecer representaciones positivas sobre las minorías. Finalmente, nuestros resultados muestran que el tema de la islamofobia mostró un aumento dramático en 2017 con una cierta disminución, más tarde, en 2018. Este aumento estuvo relacionado con el fuerte aumento de los ataques yihadistas en varias ciudades europeas, lo que indica cierto apoyo para el modelo de espirales de refuerzo (RSM) de Slater. Este resultado también está en línea con varios estudios realizados en Occidente. Por ejemplo, Arendt y Northup (2015) demostraron que la cobertura de noticias internacionales sesgada crea estereotipos y ansiedad entre las comunidades, lo que a su vez influye en los sentimientos viscerales y actitudes implícitas hacia ciertos grupos minoritarios. De manera similar, Shaver, Sibley, Osborne y Bulbulia (2017) confirmaron que la exposición a la "islamofobia inducida por los medios" en Nueva Zelanda se asocia tanto con una menor calidez como con una mayor ira hacia la minoría musulmana en el país.

 En nuestro análisis se demuestra, por otro lado, que la estilometría juega un papel importante en el estudio del estilo y de los patrones lingüísticos de cualquier tipo de enunciado. La estilometría también se puede utilizar para validar hipótesis lingüísticas y estilísticas en los estudios literarios, políticos, sociales, históricos o antropológicos. Por ejemplo, en uno de los artículos publicados, hemos utilizado tres corpus terminológicos extraídos de diecinueve obras escritas por Pío Baroja, Vicente Blasco Ibáñez y Miguel de Unamuno, tres autores de la Generación del 98. Nuestro objetivo era investigar sus peculiaridades y diferencias estilísticas y lingüísticas. Nuestros corpus han sido diseñados para atender a los criterios de género, registro, autoría y tema. Nuestro enfoque estilométrico computacional nos ha proporcionado información específica del contexto sobre el uso sintáctico y semántico del término "árabe" por parte de los tres escritores seleccionados. Los traductores, los

traductólogos, los estudiosos de la literatura y los historiadores pueden aprovechar los corpus utilizados en este estudio de varias maneras con el fin de realizar estudios más precisos sobre el estilo, el pensamiento y los sentimientos de estos tres escritores y pueden, asimismo, aplicar el método a otros escritos, tanto de la Generación del 98 como de otras generaciones y movimientos literarios, ideológicos o históricos.

 La aplicación de nuestro método investigador en el proceso de traducción sirve, por ejemplo, para indagar acerca de las concordancias posibles con objeto de encontrar un equivalente apropiado para un término específico. Miangah (2012) establece que "se ha demostrado que los adjetivos que se colocan junto con los sustantivos son muy útiles para comprender el contexto". Esto es particularmente cierto cuando los diccionarios tradicionales no logran sugerir una traducción muy exacta de un término extranjero. Boulton (2012) muestra las limitaciones inherentes a los diccionarios tradicionales frente a los corpus utilizando el siguiente ejemplo francés «Je suis paralysé entre le brulot et la chanson d'amour». Un diccionario tradicional ofrece los siguientes significados posibles en español para el término brulot: "barco de fuego", "panfleto" o "mosquito". Sin embargo, Boulton utilizó un gran corpus para mostrar que una traducción correcta sería " (canción) rebelde", "(canto) revolucionario" o "(canción) protesta" . De hecho, el famoso cantante francés Renaud debe su fama a ser uno de los pioneros y mejores representantes de este tipo de canción en Francia .

De ahí que Wright (1993, p. 70) señalara que "los documentos deben hablar 'el idioma' de la audiencia a la que van destinados y deben parecerse a otros textos similares producidos dentro de esa comunidad lingüística y de ese dominio temático en particular. Estas consideraciones requieren, con frecuencia, que los traductores vayan más allá de las estrategias meramente correctas en términos de contenido léxico y gramatical para ofrecer soluciones estilísticamente apropiadas". Esta afirmación revela que un traductor es, básica y primordialmente, un productor de textos. Por lo tanto, las dos primeras y principales cualidades que ha de poseer un traductor son poder visualizar cómo se usan las palabras en el texto original y en su réplica traducida y entender cómo se relacionan con otras palabras en el contexto particular de cada texto producido, ya sea el original o su traducción.

 Finalmente, nuestro análisis muestra que la tecnología informática juega un papel central en el estudio del vocabulario y de la traducción. Mediante el uso de computadoras para extraer y analizar datos significativos de corpus generales y especializados, la lingüística de corpus puede contribuir a

desarrollar los métodos de traducción al permitir realizar tareas transcendentales en la misma como son el determinar las colocaciones, identificar los grupos de palabras, establecer las subcategorizaciones y validar hipótesis de equivalencias lingüísticas. De hecho, los corpus juegan un papel fundamental en todos los aspectos de los estudios de traducción tanto si son teóricos como descriptivos, contrastivos o didácticos. Por ejemplo, en nuestro último artículo publicado (Capítulo 8), utilizamos dos corpus, uno en francés y otro en español, extraídos de los diarios Le Monde y El País para investigar las traducciones de términos comerciales en inglés. En este caso, nuestros corpus respondieron a los criterios de autoría, tema, género y registro. El uso de estos corpus posibilita a los traductores comprender de forma más adecuada el contexto en el que se usa la terminología comercial en inglés y, por tanto, traducirla de manera más conveniente en español y francés, en la traducción directa, y en inglés, en la traducción inversa. En la actualidad, los análisis basados en corpus son una herramienta eficaz e indispensable para

a) Crear entradas de diccionario con explicaciones o equivalentes más contrastados

b) Proporcionar alternativas de traducción más dinámicas y rigurosas que las presentadas por un diccionario tradicional

c) obtener información específica del contexto sobre el uso sintáctico y semántico de los términos utilizados en los corpus analizados.