CHAPTER 13

Quantitative analysis of translations in Spanish-language periodical publications (1900–1945)

A methodological proposal

Ventsislav Ikoff and Pablo Martínez
Universitat Oberta de Catalunya / 300.000 Km/s / Institute of Advanced Studies of Catalonia

This paper addresses the quantification of literary production in Spanish between 1900 and 1945. To obtain quantitative data, we must consider catalogue records both of main libraries and of newspaper and periodicals libraries. The latter preserve literary magazines that feature articles and translations from peripheral authors who were overshadowed or forgotten by historical narrative. Given the absence of proper cataloguing in newspaper and periodicals libraries, it is necessary to develop a workflow that enables massive and accurate searches. The aim of the proposed methodology is to build a meta catalogue that allows to search for authors and works together with translators or illustrators. Only this way will we be able to obtain an overall image of that literary period.

Keywords: periodical publications, literary translation history, Ibero-American literary field, research methodology, digital archives, digital humanities

Introduction

Libraries around the world are digitizing a myriad of archives with historical, periodical publications. This process is of paramount importance to the preservation of these historical documents for posteriority and, at the same time, for their accessibility, making them available for researchers' consultation anywhere on the planet. Nevertheless, extracting knowledge from digitalized archives remains difficult, and these resources remain under-exploited, despite their great potential for researchers. One of the difficulties of extracting knowledge from digitized periodicals stems from their unstructured archives. Said archives' catalogues provide

https://doi.org/10.1075/btl.155.13iko

information limited to journal name and number, but lack information on the included texts, such as articles, published in a given issue. Consequently, when trying to adopt a large-scale approach and perspective, it is impossible to fully grasp what the periodical presses of the past have published.

This does not mean, however, that digitized textual corpora are inexplorable or unsuitable for cultural studies. Computational linguistics and digital tools have enabled research on cultural trends in corpora based on digitized books (Michel et al. 2011; Gulordava and Baroni 2011; Juola 2013) and historical newspapers (Lansdall-Welfare et al. 2017; Cristianini, Lansdall-Welfare, and Dato 2018). These studies, based on applying statistical methods to the entire corpus (Tahmasebi et al. 2015), quantitatively depict the development of linguistic, cultural, and historical phenomena over time. However, as Koplenig (2015) demonstrates, metadata itself is an important source of information and is needed to contextualize and qualify results.

The quantitative analysis of book translations is already an established approach in Translation Studies, and, although it overlooks book content, it has most notably given rise to a Sociology of Translation, as proposed by Heilbron (1999). Bringing large-scale quantitative analysis to the study of translations in periodical publications, however, is novel. Specifically, in Spanish and Latin-American cultural periodicals, how many translations have been published remains unknown, and there is no conception of which works by which authors have been translated and by whom. Digital approaches like the one proposed in this paper aim to put us in the position not only to answer these questions in general, but also to conduct an in-depth study of the circulation and reception of different authors, movements, and ideas in the cultural periodical presses from particular spaces and periods of time. We believe that this method can contribute to advance book history, literary history, and the history of literary translation in the Hispanic world. On a more general note, the analysis of Spanish-language modernist literary journals from the first half of the twentieth century will provide new vantage points for understanding literary modernity in the Hispanic field and will provide an interesting basis for comparisons with book translations, allowing us to draw parallels, or otherwise, between these two platforms for literary circulation.[1]

As a publication platform, periodicals show very specific features as well as their own logic and dynamics. As an object of study, however, they also pose many methodological difficulties. One of the main challenges is the inexistence of a catalogue on their contents and contributors. Even when journals did publish a

1. This paper draws on the experience gained throughout the Trad-Data research project. The project studies book translations in Spain and Latin America between 1900 and 1945 on a large scale, using data from library and online catalogues.

summary of their contents on one of their pages, the data has not been registered in a searchable catalogue. This leads to the above-mentioned challenge when it comes to structuring the contents of a large corpus of digitized periodicals, consequently hindering the large-scale analysis of the literary works and authors they published and helped circulate, while obscuring the role that translators and critics played in mediating new forms of expression in their pages. As a solution, these details could be found in the periodicals' pages, but the automation of this process faces the challenges of inconsistent format and content distribution – journals vary in paper size and page count, and sometimes these features change even within a single publication over time. Often, articles can be distributed in several columns or across the whole width of the page and vary in length and font. The level of detail in terms of the information on the journals, such as authors and translators, also fluctuates, with references to original texts and mediating languages occasionally omitted.

The following paragraphs will describe technological solutions and methodologies as approaches to the analysis of translations in Spanish-language magazines from 1900 to 1945, detailing some of the technologies that facilitate the archive's transformation as well as the quantitative analysis of texts via search strategies. To this end, this paper will examine the procuring of digitized archives from several identified online libraries, said archives' treatment and visual recognition, and their subsequent indexing for consultation.

Measuring literary production

If we attempt to estimate figures that could allow us to compare the quantity of literary production over time, we will find ourselves before a lack of publication registries that could point to editorial activity, such as the International Standard Book Number (ISBN), which was adopted internationally as of 1970 and did not gather information prior to this period. The ISBN includes between 9 and 13 figures that identify all publications in the editorial market. This number provides information such country, editor, and article (the publication's issue), with which one may find the publication date and other information contained in centralized registries.

As an alternative, using a large corpus of digitized texts can provide a quantitative description of cultural trends (Michel et al. 2011). To this end, we may use the Google Ngram Viewer, which allows one to find the frequency of terms over time within the corpus created with the digitized books on the Google Books platform, which spans from 1500 to 2008.[2] This tool has been employed in other studies in

2. Google Ngram Viewer can be accessed at: https://books.google.com/ngrams.

the cultural field (Juola 2013) and has been used to find usage efficacy in different cases (Gulordava and Baroni 2011).

As such, we may carry out three searches as an experiment: first, using the Spanish-language word "*novela*", which includes all texts that mention said term, thus containing books within the genre as well; second, searching for the term "*quijote*", which references one of Spanish literature's most cited texts and would include reeditions as well as any texts referencing the novel; and third, the Spanish-language term "*traductor*", thus referring to possible citations of the persons in charge of the translation or referring to the role of the translator in the world of essay-writing and literature. Through this experiment, we may observe the high frequency and increasing appearance of these terms (see Figure 1).

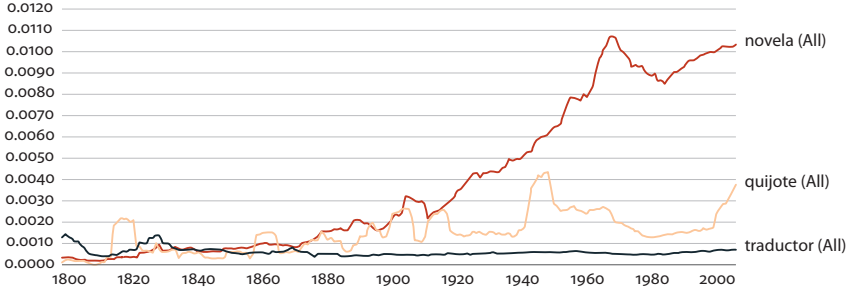

Figure 1. Distribution frequencies for "novela", "quijote" and "traductor" in the Google Ngram Viewer from 1800 and 2008

Though this information offers an estimate, it also contains many of the biases stemming from the lack of partiality among the library archives registered on Google as well as from the lack of precision around the search terms and the frequent errors of text digitalization (Pechenick, Danforth, and Dodds 2015; Koplenig 2015) when it comes to either textual content or the registry date of books.

Because of this, conducting the experiment with other sources in order to compare results would prove relevant. This process requires the meta-catalogue WorldCat (Chen 2012), which aggregates information from various library archives and permits the simultaneous consulting of content from each archive. In this case, the search takes place within the contents of the catalogue card, which includes descriptors and metadata rather than the book's content. The ultimate goal of this process is to count the number of books per year. Since the search tool precludes the possibility of consultation without any descriptive characters in the title field, we conducted the experiment using the letter "a".[3]

3. WorldCat is available at https://www.worldcat.org. The result of this experimental search can be reproduced via the following url: https://www.worldcat.org/search?q=kw:a&fq=yr:1900+>+>+x0:book+>+mt:fic+>+ln:spa&qt=advanced&dblist=638

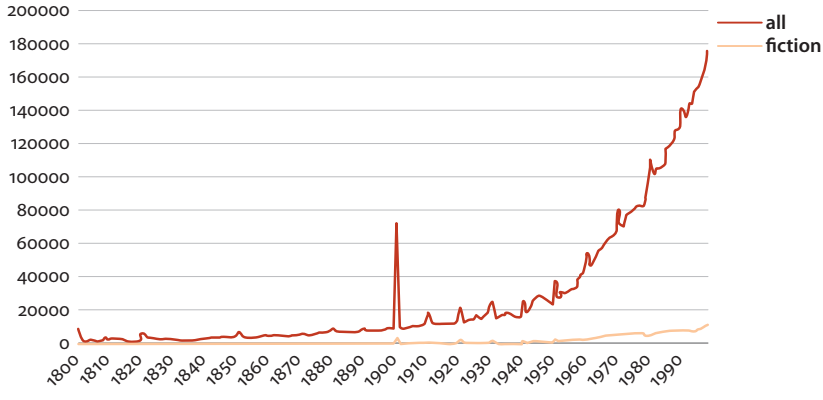

Figure 2. Distribution of books published in Spanish between 1800 and 2000 according to the WorldCat catalogue

According to the number of results for each search, Figure 2 clearly depicts an upward curve with different slopes from those observed in the Google Ngram Viewer searches. We can find regular peaks at the beginning of each decade, given the works' imprecise date registries.

To evaluate the quality of these results, we have consulted the National Library of Spain's original catalogue, in its MARC21 format – the protocol through which the library exchanges information with WorldCat.[4] Here, we can observe that some fields, such as the date, are filled in completely (an indicator of quality), while other fields, such as genre, are only occasionally filled out. As such, any estimate of the number of literary works produced during this period would not prove significant, given the errors in the original catalogues.

In sum, we cannot determine a stable figure when quantifying literary production throughout the period of study, due to the lack of unique identifiers as well as the inconsistency in meta-catalogues such as Google Books and WorldCat. Nonetheless, gathering indicators – assuming there are internal errors in the composition of information sources – can help describe production proportionality.

Limitations in the study of periodical publications

If we hope to quantify literary production in periodical publications, the task becomes far more complicated. While the ISBN is highly useful when quantifying the editorial production of recent decades, periodical publications have only used an International Standard Serial number since 1976, leaving the period of our analysis

4. http://www.loc.gov/marc/marcspa.html

uncovered. Further, library catalogues either register periodical publications by time period or year, or generate a unique register for the entire publication. Issues are not registered individually, nor are the articles published within them. In the best of cases, we can determine the existence of a publication and the amount of released issues, but the articles within, its authors, and its language remain unknown.

With the goal of approaching this void in cataloguing and in order to take advantage of new technologies, digital libraries have emerged over the last few years. In Spain, for instance, the Library of Catalonia launched its digitalization program (Lamarca Morell and Serra Aranda 2005) and, in 2007, the National Library of Spain published its digital library. These digital libraries, which allow for the online consultation of archives, digitalize contents in order to facilitate searches either within documents or within the catalogue card or issue descriptors.

In some cases, there are no standards for exchange akin to the MARC21. Rather, our tools can complement the more-than-generic cataloguing efforts described above. While these tools allow for the identification of magazines and their issues, one cannot access a descriptive index of the articles' contents and authors.

Thanks to the optical character recognition (OCR) of publications' content, users and researchers may find possibly relevant texts. Nonetheless, optical character recognition is patchy at best: the publications that were first included due to their relevance were treated with less sophisticated OCR, while more recent publications boast better processes. Furthermore, which OCR processes have been used for different texts remains unknown, implying the presence of potentially undiscovered errors or biases caused by these tools.

Meanwhile, searches only sometimes go through the full text, while other times only a few sections are included in the search. Sometimes, regular expressions (a sequence of characters defining a search pattern) can be used for search terms, while other times this is not the case.

The main tools used for the cataloguing and publication of serial publications in the main libraries we consulted are Pandora, in the case of the Digital Library of the National Library of Spain; and Content dm, in the case of the ARCA archive of the Library of Catalonia.[5]

As such, the main limitations to these systems' availability for data analysis include, on the one hand, the dispersion of libraries, which precludes the possibility of conducting a single, joint search, and on the other, the irregularity among optical character recognition processes and our lack of control over search tools.

5. More information about Pandora and Content dm is available at https://imastudia.unav.edu/studia/content/manual/manual.html and https://www.oclc.org/es/contentdm.html, respectively. The Digital Library of the National Library of Spain is accessible at http://hemerotecadigital.bne.es/index.vm. The ARCA archive of the Library of Catalonia is accessible at https://arca.bnc.cat.

Thus, our existing tools will not suffice to quantify literary production in periodical publications from the years 1900 to 1945. Instead, for the purpose of this investigation, we would require new tools to efficiently consult the documents we have already digitalized.

Proposed workflow

Currently, an abundance of Spanish-language digital libraries is already widely available. Proceeding from various institutions and public administrations, all of these libraries use similar tools as the two libraries mentioned above: the Library of Catalonia and the National Library of Spain. In order to access content in a unified fashion, we would not only require access to each publication's metadata, which is available in different formats, but also access to the totality of documents in order to conduct optical character recognition via a documented methodology that would allow us to uncover the existing biases in searches and subsequently conduct complex searches.

A provisional list of existing digital libraries that could serve our study of translations in Hispanic magazines is presented in the Appendix. We would gather all of the documents in these libraries as well as each document's catalogue card. These catalogue cards, presented in the form of metadata or gathered through automated web data collection or "web scrapping" from the website, would be stored in high-capacity relational databases, such as PostgreSQL. The documents – often in PDF or PNG format – should be stored in servers, allowing them to be accessed for analysis. The information on each publication should include the source, download URL, and the downloaded content's location.

Subsequently, optical character recognition should be conducted for each of the acquired documents. The optical character recognition must use the same criteria for every document. This step is one of the most complex and critical, as it would significantly contribute to the quality of our existing systems. The goal is not only to homogenize the criteria for optical recognition, but also to take advantage of our most recent technologies. In this sense, one of the most widely used character-recognition programs today is Tesseract. It was first developed by Hewlett Packard in 1985 and has been supported by Google as of 2006, after which it was soon recognized as the best OCR software and has only improved over time. Tesseract can work with text patterns, identifying columns, and using dictionaries in various languages. Using these tools allows the text that has been digitized as an image to be converted into text, making it not only legible by humans, but also by computational tools that can take on subsequent analyses.

In the next phase, we would be interested in conducting searches throughout the entirety of digitized texts. To this end, we would require a search server capable not only of finding texts through exact coincidences but also through a myriad of similarity criteria. Tools such as Elastic Search allow queries according to methods including regular expressions, Levenshtein distance, phonetic similarity, and thesauri.

Searches using regular expressions can define patterns within the text and seek out partial coincidences and words that start in one way and continue in an indefinite form, without taking upper-case letters or accent marks into account, and can require the presence of certain characters.

Searches using the Levenshtein distance can establish similarities between words by considering the number of single-character edits needed to change one word into another. As such, one can measure the similarities between words which may have possible mistakes in the OCR transcription and thus overcome and ignore these errors, as well as plural forms and other declensions.

Searches using phonetic similarities compare factors such as pronunciation, rather than spelling. Meanwhile, searches via dictionaries assign the same value to terms written in different ways, whether they are similar terms, terms in different languages, or abbreviations. For instance, "Nueva York" and "New York" hold the same value, as do "Doctor" and "Dr.".

Our proposed workflow combines these tools with proper documentation on their use, giving way to the massive consultation of all the target digital libraries in this study. Nonetheless, identifying the structures of magazine contents is not yet possible in terms of locating text titles and authors, given that text extractions via OCR do not include said structures.

To address this problem, image analysis would be required in order to distinguish different parts of the articles' texts. Several experiences have been recorded (Andersen and Zhang 2003; Zhang and Andersen 2003; Paaß and Konya 2011) in which neural networks successfully extract articles' titles and subtitles. This is an important part of the process, as it distinguishes the article's different parts, its content, and its paratext in order to recognize information such as the headline, author, translator, notes, illustrations, and, of course, the body of the text.

Metadata on articles in periodical publications can have a twofold purpose. On the one hand, it allows us to study the general trends of literary publications in periodicals, the circulation of texts in journals across Ibero-America, and the networks of authors, translators, and other journal contributors. On the other hand, metadata also helps contextualize and qualify findings within textual corpora. In this sense, as (Koplenig 2015: 183) concludes in his study on the use of large textual data sets for research on cultural and linguistic change using Google's Ngram

corpora, "the importance of metadata cannot be underestimated: the availability of metadata is not just a nice add-on, but a powerful source of information".

Finally, after article recognition and indexing, the next step would be to detect translations. The simplest and most effective approach would be to detect the authors, match them against an existing authorities database such as The Virtual International Authority File (viaf.org), and select articles based on the author's nationality and usual language of expression. This step may also possibly identify articles in foreign languages, in which case further verification of the text language might be required. A text-searching strategy using the most popular words in several languages could resolve this issue.

Lastly, it is worth mentioning that the tools used throughout this process need to use open code and have documented and published functionalities, so that we may understand the tools' inner workings and reproduce the same analyses with different tools, casting light on any biases in analysis caused by functionality errors or configurations that may skew our findings.

Final considerations

With the workflow described above, we can now launch a new framework that would allow us to fully assess digitalized periodical publications by accessing their contents via search tools that can also recognize content structures. Traditionally, library archives do not register periodical publications' contents in detail, making the reutilization of these archives complicated when it comes to analyzing the literary production of a given time. However, our process would allow us to take a step forward when it comes to exploiting archives and extracting knowledge from them.

On a more general level, combining a quantitative approach with traditional literary studies can produce new insights for literary and translation history and improve our knowledge of the complex processes that have shaped the literary landscape in Spanish speaking countries.

Funding

This paper is part of the project "Trad-Data: Translation and Market. The Circulation of Books (Spain-Latin America, 1900–1945) and Big Data" funded by a BBVA Foundation Leonardo 2017 Grant for Young Scholars and Cultural Creatives, and the ERC Starting Grant "Social Networks of the Past. Mapping Hispanic and Lusophone Literary Modernity (1898–1959)", both directed by Diana Roig-Sanz at the Universitat Oberta de Catalunya.

References

Andersen, Timothy L., and Wei Zhang. 2003. "Features for Neural Net Based Region Identification of Newspaper Documents". In *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings*, 1:403–7. IEEE Comput. Soc.. https://doi.org/10.1109/ICDAR.2003.1227698

Chen, Xiaotian. 2012. "Google Books and WorldCat: A Comparison of Their Content". *Online Information Review* 36 (4): 507–516. https://doi.org/10.1108/14684521211254031.

Cristianini, Nello, Thomas Lansdall-Welfare, and Gaetano Dato. 2018. "Large-Scale Content Analysis of Historical Newspapers in the Town of Gorizia 1873–1914". *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 51 (3): 139–64. https://doi.org/10.1080/01615440.2018.1443862

Gulordava, Kristina, and Marco Baroni. 2011. "A Distributional Similarity Approach to the Detection of Semantic Change in the Google Books Ngram Corpus". In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, edited by Sebastian Pado and Yves Peirsman, 67–71. Edinburgh. https://www.aclweb.org/anthology/W11-2508

Heilbron, Johan. 1999. "Towards a Sociology of Translation: Book Translations as a Cultural World-System". *European Journal of Social Theory* 2 (4): 429–44. https://doi.org/10.1177/136843199002004002

Juola, P. 2013. "Using the Google N-Gram Corpus to Measure Cultural Complexity". *Literary and Linguistic Computing* 28 (4): 668–75. https://doi.org/10.1093/llc/fqt017

Koplenig, Alexander. 2015. "The Impact of Lacking Metadata for the Measurement of Cultural and Linguistic Change Using the Google Ngram Data Sets – Reconstructing the Composition of the German Corpus in Times of WWII". *Digital Scholarship in the Humanities* 32 (1): 169–88. https://doi.org/10.1093/llc/fqv037

Lamarca Morell, Dolors, and Eugènia Serra Aranda. 2005. "L'estratègia de La Biblioteca de Catalunya En Projectes Digitals". *Item: Revista de Biblioteconomia i Documentació*, no. 41: 41–53.https://www.raco.cat/index.php/Item/article/view/40866/68116.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. 2017. "Content Analysis of 150 Years of British Periodicals." *Proceedings of the National Academy of Sciences of the United States of America* 114 (4): E457–65. https://doi.org/10.1073/pnas.1606380114

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, et al. 2011. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science (New York, N.Y.)* 331 (6014): 176–82. https://doi.org/10.1126/science.1199644

Paaß, Gerhard, and Iuliu Konya. 2011. "Machine Learning for Document Structure Recognition". *Studies in Computational Intelligence* 370: 221–47. https://doi.org/10.1007/978-3-642-22613-7_12

Pechenick, Eitan Adam, Christopher M. Danforth, and Peter Sheridan Dodds. 2015. "Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution". Edited by Alain Barrat. *PLOS ONE* 10 (10): e0137041. https://doi.org/10.1371/journal.pone.0137041.

Tahmasebi, Nina, Lars Borin, Gabriele Capannini, Devdatt Dubhashi, Peter Exner, Markus Forsberg, Gerhard Gossen, et al. 2015. "Visions and Open Challenges for a Knowledge-Based Culturomics". *International Journal on Digital Libraries* 15 (2–4): 169–87. https://doi.org/10.1007/s00799-015-0139-1

Zhang, Wei, and Timothy L. Andersen. 2003. "Using Artificial Neural Networks to Identify Headings in Newspaper Documents". In *Proceedings of the International Joint Conference on Neural Networks, 2003*, 3:2283–87. IEEE. https://doi.org/10.1109/IJCNN.2003.1223767

Appendix. Digital archives of Spanish-language periodicals

Archivo Municipal de Cartagena – Fondo Hemerográfico Digitalizado
http://archivo.cartagena.es/pandora/index.html
Archivo Municipal de Huelva. Hemeroteca histórica
http://www.huelva.es/portal/es/archivo-municipal
Archivo Municipal de Murcia
http://www.archivodemurcia.es/
Library of Catalonia. Consorci de Biblioteques Universitàries de Catalunya. ARCA: Arxiu de Revistes catalanes antigües
http://www.bnc.es/digital/arca/index.html
Biblioteca Digital del Caribe
http://www.dloc.com
Biblioteca Municipal de Donostia-San Sebastián. Digital library
http://liburutegidigitala.donostiakultura.com/Liburutegiak/
Biblioteca Nacional de Colombia. Hemeroteca digital
http://catalogoenlinea.bibliotecanacional.gov.co
Biblioteca Virtual Joan Lluís Vives. Hemeroteca
http://www.lluisvives.com/psegundonivel.jsp?conten=hemeroteca
Biblioteca Virtual Miguel de Cervantes. Hemeroteca
http://www.cervantesvirtual.com/hemeroteca/
Catálogo de la Biblioteca Nacional de Argentina
https://catalogo.bn.gov.ar/
Centro de Estudios de Castilla-La Mancha. Biblioteca Virtual
http://ceclmdigital2.uclm.es
Dipòsit Digital de Documents de la UAB
http://ddd.uab.cat/
Directorio y Recolector de Recursos Digitales HISPANA
http://hispana.mcu.es/es/inicio/inicio.cmd
El Surrealismo y sus derivas
http://www.uam.es/proyectosinv/surreal/
Fondos digitalizados del Archivo Municipal de Castellón
http://archivo.castello.es/
Hemeroketa.com – Euskal Prentsaren Lanak
http://www.hemeroketa.com/
Hemeroteca Digital, Biblioteca Nacional de España
http://www.bne.es/es/Catalogos/HemerotecaDigital

Hemeroteca Nacional Digital de México
http://www.hndm.unam.mx
Hemeroteca Región de Murcia. Proyecto Carmesí
http://hemeroteca.regmurcia.com/
Jable. Archivo de prensa digital de Canarias
http://biblioteca.ulpgc.es/?q=jable
Junta de Andalucía Biblioteca Virtual de Andalucía. Hemeroteca
http://www.bibliotecavirtualdeandalucia.es/catalogo/publicaciones/listar_numeros.cmd
Koldo Mitxelena Kulturunea. Hemeroteca
http://www4.gipuzkoa.net/hemerotekakm/
Memoria de Madrid
http://www.memoriademadrid.es/
Ministerio de Cultura. Biblioteca Virtual de Prensa Histórica
http://prensahistorica.mcu.es/es/estaticos/contenido.cmd?pagina=estaticos/presentacion
Revistas de la Edad de Plata
http://www.edaddeplata.org/revistas_edaddeplata/index.html
Revistas y Periódicos Españoles: Colección Digital de Revistas Femeninas. Universidad de Connecticut.
https://guides.lib.uconn.edu/SPANMagazines https://lib.uconn.edu/location/asc/collections/spanish-periodicals/
SOMNI: Col.lecció digital de fons històric de la Universitat de València. Hemeroteca
http://trobes.uv.es/search*spi/X?(*)&searchscope=3&Da=&Db=&m=w&SORT=A

PART IV

Situating the agent and his network in the transnational space

Qualitative approaches of translation in periodicals