hosted by Hamburg University Press

apropos
[Perspektiven auf die Romania]

Sprache/Literatur/Kultur/Geschichte/Ideen/Politik/Gesellschaft

Where are Romance Studies Heading?
A Bibliographic Data Science Analysis Using Regression

José Calvo Tello

apropos [Perspektiven auf die Romania]
hosted by Hamburg University Press
2022, 9
pp. 119-152
ISSN: 2627-3446

Online
https://journals.sub.uni-hamburg.de/apropos/article/view/1894

Zitierweise

Calvo Tello, José. 2022. „Where are Romance Studies Heading? A Bibliographic Data Science Analysis Using Regression.“ *apropos [Perspektiven auf die Romania]* 9/2022, 119-152.
doi: https://doi.org/10.15460/apropos.9.1894

2022, n°9
pp. 119-152
doi: https://doi.org/10.15460/apropos.9.1894

José Calvo Tello

Where are Romance Studies Heading?

A Bibliographic Data Science Analysis Using Regression

José Calvo Tello
is subject librarian and researcher at
Göttingen State and University Library.
calvotello@sub.uni-goettingen.de

Keywords

Romance Studies – Bibliographic Data Science Analysis – Library Records – Linear Regression

1. Introduction

Researchers of any discipline share certain opinions and intuitions about how their discipline has developed in recent years. This is normally influenced both by their own experience and by other senior researchers' opinions about the previous decades. Of course, shared intuitions do not necessarily need to be based on actual facts, they are often misled by particular experiences, the specifics of a research institution or the development of a specific sub-discipline.

In this study, I tackle the question about how the Romance Studies have developed in the past decades. For this purpose, I use library records as research objects and apply statistical methods. In addition to the description of the past decades, I use statistical models to make predictions of the impact of current trends in future years.

Romance Studies are a challenging discipline. In contrast to other philologies such as German or English Studies, a plurality of languages are at the core of this discipline (Kramer 2002, 13). Besides, in the countries where these languages are spoken, the research mainly focuses on one language (French Studies, Spanish Studies, etc.). A further challenge for the Romance Studies is the multilingualism of their research production. Any study trying to be representative for this discipline needs to cover at least publications written in French, Spanish, Italian, Portuguese, German and English, with a long tail of further important languages such as Romanian, Catalan, Occitan, Sardinian, etc.

Previous research about the history of the Romance Studies has mainly focused on the periods before 1950 (Richert 1913; Kalkhoff 2010; Wolf 2012; Lieb and Strosetzki 2013; Kremnitz 2016; Kramer 2020). As Kremnitz points out, it is more difficult to assess current developments than to describe the historical processes (Kremnitz 2016, 287). Many of these historical studies worked on a narrow selection of scholars with great impact in the fields of Romance Linguistics or Literary Studies. Such an approach is possible when the historical distance to the research object is enough to identify the most influential researchers. Since this is not possible for the last decades, I decided to take a quantitative perspective by using data curated by professionals, i.e. library records. In the past years, Digital Humanities have shown a new interest in working with data from bibliographies and library records (Henny-Krahmer 2017; Jannidis, Konle, and Leinen 2019; González 2021; Ehrlicher and Lehmann 2021; Herrmann et al. 2021; Gittel 2021). This work can be framed in the new paradigm of the *bibliographic data science*, an emerging sub-discipline closely related to the Digital Humanities that analyzes library records and bibliographies to study the historical development of several types of publications, including literature, research, journals, etc. (Tolonen et al. 2020; 2019; Vaara et al. 2019; Maryl and Wciślik 2016).

In this study, I focus on the development of the past 40 years, i.e. between 1980 and 2019. The interest for these years is based on several historical changes, such as the inclusion of many European Romance countries into a shared political structure such as the European Union, the Bologna process, the development of new technologies such as the Internet or e-books, or the rapid development of the scholarly publishing sector (Kramer 2002; Becker et al. 2020; Krefeld 2020; Monjour 2020). These four decades lend itself to analysis because of the availability and quality of the data in the catalogs, which is higher for publications from the last decades than for previous ones.

This analysis focuses on the Romance Studies from the perspective of the German-speaking area. For this, I use as data the library records from German university libraries, as I will explain in detail in the next section.

One could ask whether it is acceptable to use data from German libraries to analyze a discipline about foreign languages. Although a study based on using the data from several linguistic regions and nations could be a valuable option, it could be asked whether such an approach would cover the Romance Studies and the Romance languages accurately. As many researchers acknowledge, in the Romance-speaking countries the disciplines tend to focus only on the national language (Kramer 2002, 17; Holtus and Sánchez Miret 2008; Kremnitz 2016). The German-speaking area is seen as the place where the Romance Studies started and where a comparative approach finds more support (Wandruszka 1988; Holtus and Sánchez Miret 2008; Kremnitz 2016), or as Gumbrecht states “Romance philology arose in Prussia (not in France, Spain, or Italy)” (Gumbrecht 2002, 2). For this reason, I consider the German libraries as one representative source of data for this analysis, although not the only one.

2. Dataset

2.1 Library Records from the Hebis and GVK - GBV Union Catalog

For this analysis, I use records from two large library catalogs. Research and university libraries in Germany are organized in consortia or networks within which infrastructure and data are shared (for example for the catalogs). The consortia tend to cover the research libraries of one or more federal states and there are currently six consortia (Gantert 2016, 44):

1. *Gemeinsamer Bibliotheksverbund* (GBV, translated into English as the GBV Common Library Network): responsible for the libraries of seven federal states plus the Prussian Cultural Heritage Foundation;
2. *Südwestdeutscher Bibliotheksverbund* (SWB): responsible for the libraries of three federal states;
3. *Hessisches BibliotheksInformationsSystem* (hebis): responsible for the libraries of Hesse and a region of Rhineland-Palatinate;
4. *Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen* (hbz): responsible for libraries in the federal states North Rhine-Westphalia and Rhineland-Palatinate;
5. *Kooperativer Bibliotheksverbund Berlin-Brandenburg* (KOBV): responsible for the libraries of two federal states;
6. *Bibliotheksverbund Bayern* (BVB): responsible for the libraries in Bavaria.

Since 2019, the GBV and the SWB have joined their efforts and have created the K10plus network. Within this network of consortia, the libraries share their library records in a database called the K10plus, which can be consulted as a catalog.[1]

Three of these consortia use the same format and software based on the PICA format, which stands for Project for Integrated Catalogue Automation (Voß 2020). The rest of the consortia work with different formats and software solutions such as Aleph and SISIS. Although it would have been of great interest to have included all catalogs in this study, there is no integrated pool of bibliographic data from these resources. This translates to several challenges when such an analysis is performed. First, different frameworks and technologies need to be applied. Second and most importantly, the record fields from the several formats need to be mapped into a single model in order to compare the categories across databases. This task cannot be done by a single researcher and the consortia currently do not offer solutions for this problem.

How representative are the records of these three consortia for the library landscape for Romance Studies? This dataset does cover the records from university libraries of the majority of the German states, since records from 11 out of 16 German states are being covered, with a diverse geographical distribution.

[1] <https://opac.k10plus.de/DB=2.299/START_WELCOME>.

Besides, the records cover at least a part of two *Specialised Information Services* (in German *Fachinformationsdienste für die Wissenschaft* or simply FIDs), which constitute projects carried out by libraries focused on a specific discipline (Gantert 2016, 155–58). For the Romance Studies, two FIDs are especially relevant:

1. The Specialised Information Service (FID) Romance Studies, at the University Library of Bonn and at the State and University Library of Hamburg.
2. The Specialised Information Service (FID) Latin America, Caribbean and Latino Studies, at the Library of the Ibero-American Institute (part of the Prussian Cultural Heritage Foundation) in Berlin.

Both the Prussian Cultural Heritage Foundation and the Hamburg State and University Library are part of the GBV and therefore their records can be found in the K10plus database. However, the records of the University Library of Bonn are integrated in the hbz consortium and therefore are not part of this analysis. In addition, the records of the FID for Russian, East and Southeast European Studies at the Bavarian State Library also had to be excluded from this analysis. This will be mentioned again when looking at the results and trends for Romanian.

To summarize the answer to the question about the representativeness of the dataset, it represents the greatest dataset of library records about Romance Studies and it does represent the majority of the libraries in Germany. However, important sections of the German librarian landscape are not being considered, in particular the resources from the hbz and the BVB. This opens the possibility that similar analyses could be performed in the future using datasets created in a distributed manner, with each consortium contributing its respective records.

The data from both networks can be accessed via Application Programming Interfaces (APIs). From these sources, the data can be downloaded in Pica+ format, expressed in an XML serialization. Although unknown in other areas, Pica+ is the standard format for cataloging in libraries from several countries, among others in Germany.

The retrieved databases are also the sources for the standard Online Public Access Catalog (OPACs) of the libraries. These catalogs contain data of independent works (in German *selbständige Werke*) such as monographs, collective works and journals (Gantert 2016, 228–29). This means that chapters of collective works or articles in journals are not part of the analyzed dataset.

2.2 Classification Systems for the Identification of Romance Studies Publications

After downloading a dataset for the last decades from the APIs of both consortia (K10plus[2] and hebis),[3] the next step is to extract from the original dataset only those publications related to Romance Studies. For that, I use several library classification systems. These classification systems are hierarchical structures of classes that represent subjects, such as Chemistry, Theology or Romance Studies (Gantert 2016, 203). In theory, any class can be divided into more specific classes, creating a tree-like structure or taxonomy. One or more classes of these classification systems are then assigned to any publication. This annotation can be used by users to retrieve from the catalog the publications of any specific area. In my case, these classification systems allow me to identify the publications relating to Romance Studies and thus define the dataset for the analysis.

For filtering of the original data, I apply three classification systems:

- Dewey Decimal Classification and Subject Categories:[4] The German National Library's catalog is using a set of around 100 Subject Categories (in German *Sachgruppen*). These groups are a simplification of one of the most widely accepted classification systems, especially in English-speaking countries: the Dewey Decimal Classification (DDC; see further details in Chowdhury et al. 2008, 96–99).[5] Specifically, for this analysis I consider publications assigned with the DDC classes starting with 44, 45, 46, 84, 85, or 86.
- Regensburger Verbundklassifikation (RVK):[6] This classification system is probably the most widespread in the German-speaking area. In its current version, it contains more than 800.000 classes and these do not only represent subjects, but also publications about or by specific people.[7] For this analysis, I consider publications assigned with the RVK classes starting with the letter I.
- Basisklassifikation (BK):[8] This classification system was originally developed in the Netherlands, and it has a wide acceptance in the libraries of the GBV consortium. It contains around 2.000 classes and, as its name describes, it is seen as a basic classification system in contrast to other more complex systems such as the complete DDC or the RVK. I consider publications assigned with BK classes starting with 18.2 or 18.3.

[2] <https://wiki.k10plus.de/display/K10PLUS/SRU>.

[3] <http://sru.hebis.de/sru/DB=2.1>.

[4] <https://www.dnb.de/DE/Professionell/DDC-Deutsch/DDCinDNB/ddcindnb_node.html

[5] The German National Library decided to add three classes which were not present in the original Dewey Decimal Classification: Literary fiction (class B, in German *Belletristik*), Children and Youth literature (class K, in German *Kinder- und Jugendliteratur*) and Textbooks (class S, in German *Schulbücher*).

[6] <https://rvk.uni-regensburg.de/>.

[7] For example, the RVK class IR 8005 represents secondary literature about Fernando Pessoa.

[8] <https://wiki.k10plus.de/pages/viewpage.action?pageId=437452809>.

After a first analysis, it became clear that primary literature (i.e. published novels, poetry, theater plays, etc.) and secondary literature (research publications) differ in many of their characteristics. For example, in many cases publishers do not offer any e-book license for primary literature to university libraries. Besides, the prices of primary literature are much lower than those of secondary literature, since the first ones are meant for a general public, while the latter are meant for a specialized readership and have therefore lower print runs. Because of these and other differences in many of the analyzed categories, I decided to exclude all primary literature from the dataset. For this step, I use again the three classification systems, since all of them have one or more classes that specify that the publication is primary literature:

- *Sachgruppen*: class B;[9]
- BK: classes 17.97 and 17.98;
- RVK: classes containing in their labels the phrases *Gesammelte Werke* (‘Collected works’) or *Einzelwerke* (‘individual works’). This is the case for almost 8.000 RVK classes. For example, the class IE 5101 represents the individual works of primary literature by the author Adem de la Halle, while for the next author in the RVK, Audefroi (le Bastard), the class IE 5107 represents both his collected and individual works. These cases exemplify the challenges of working with the RVK: what could be done for the *Sachgruppen* and BK with one or two steps, the researcher is forced to repeat them 8,000 times for the RVK, which is only feasible when scripts can be programmed.

For the exclusion of the primary literature, I also use the field content type. This field[10] “contains factual terms to describe the content of this publication”.[11] This is a mandatory field for many types of publications, such as comics, biographies or exhibition catalogs, but not for primary literature. Even when it is not mandatory, many records are marked as fictional representation (*Fiktionale Darstellung*), anthology (*Anthologie*) and collection of letters (*Briefsammlung*). The records from both cataloges (hebis and K10plus) with these values were also excluded from the analysis.

The specific implementation of the steps can be followed in the companion Jupyter Notebooks of this publication, which will be described in detail in Section 2.5.

As mentioned before, I only analyze records published between 1980 and 2019 (including both years). These four decades represent a compromise between a historical overview of the field and the quality and homogeneity of the data. Cataloging rules and workflows in libraries evolve constantly. The data of the catalog for publications in 1900 is very different to the data for publications of the

[9] Only present in the German *Sachgruppen* and not in the original DDC.

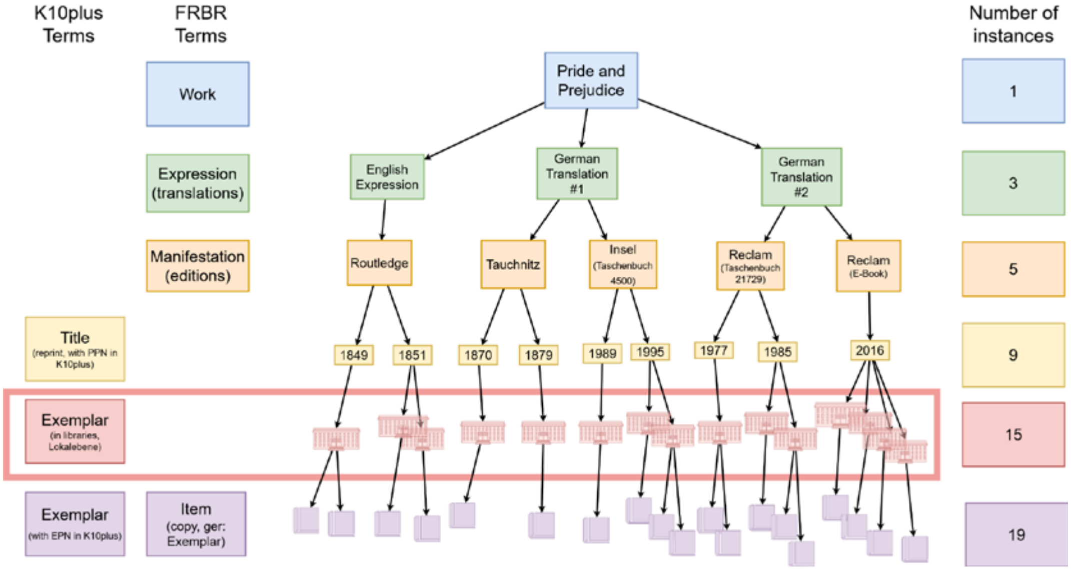
[10] With the Pica3 code 1131, Pica+ code 013D.

[11] Following the documentation of the K10plus, my translation: <https://swbtools.bsz-bw.de/cgi-bin/k10plushelp.pl?cmd=kat&val=1131&katalog=Standard&adm=1>.

year 2000. For instance, the classification systems applied to identify Romance Studies publications (*Sachgruppen*, BK and RVK) were developed only in the second half of the 20th century. In any case, exploring ways of expanding the analyzed period while ensuring data quality is worth further studying.

2.3 Analyzed Instance: FRBR and Pica Model

A further filter needs to be applied to the dataset: what kind of instance is analyzed exactly. I have already stated that the hebis and K10plus databases contain monographs, journals and series and tend to exclude chapters and journal articles (among other scholarly publications). However, the databases consider different levels of abstractions relating to the publications. In this section, I explain these instances using the example of *Pride and Prejudice* visualized in Figure 1, even though the analysis does not cover primary literature as explained before. The novel *Pride and Prejudice* was written originally in English by Jane Austen (used as example in Wiesenmüller and Horny 2017). The model of the Functional Requirements for Bibliographic Records (FRBR, Wiesenmüller and Horny 2017) uses the term *work* to relate to the abstract unit. This unit is then expressed in several linguistic *expressions*. These expressions include the original text (in this case, the English text written by Austen) and the different translations. If the text is translated more than once, every translation is counted as a further expression of the work. When a publisher publishes one of these expressions, then it creates a *manifestation*. If several publishers publish the same expression (the original text or any translation), they are considered several manifestations (Wiesenmüller and Horny 2017, 18). However, if the same publisher launches several reprints of the text, they are still considered part of the same manifestation (Wiesenmüller and Horny 2017, 19).



1 | Comparison of FRBR and Pica+ models with the example of *Pride and Prejudice*

Although the fields from both databases are organized following this FRBR model, the databases in these consortia reflect an alternative model. The reason for that is that, following the FRBR model, reprints of several years would be part of the same manifestation. However, the users of the library could be interested in knowing whether the copy of *Pride and Prejudice* in the library was published in 1849 or in 1851. For this reason, the presented model in the analyzed databases considers more specific instances. On the top of this model, we find the concept of title (*Titel* in German). This could be seen as any manifestation but distinguishing the reprints of different years. Each different title has a unique identifier (called *Pica-Produktionsnummer* or PPN) in the database.

When a library acquires a publication (in the case of printed publications and e-books licenses) or decides to consider a publication in their catalog (in the case of Open Access publications), it creates an *exemplar* in the library's catalog. From each exemplar, the library can purchase one or more copies (*items* in FRBR terms). Each item receives a different call number that enables library users to find it in the library. In the database, each item is identified with a unique identifier (called *Exemplar-Produktionsnummer* or EPN).

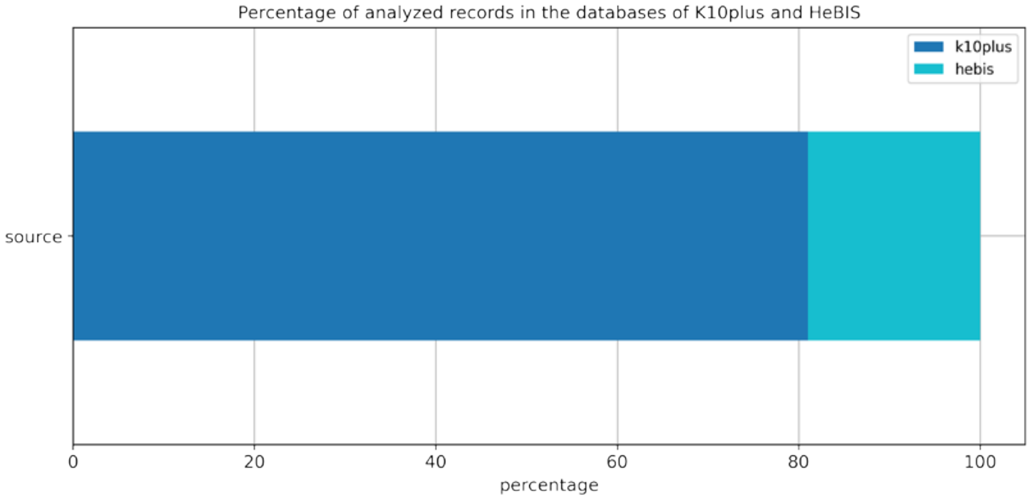
Theoretically, I could have analyzed any instance of the FRBR model or the databases, from the most abstract one (work) to the most specific (item). This decision has strong implications on at least two aspects. First, working with FRBR instances would have forced us to reconcile many of the data of the databases since the FRBR instances are actually not explicitly identified for the majority of the cases. That would have meant to modify and edit many of the analyzed data, resulting in perhaps errors and noise. Second, working on more abstract instances (such as work or expression) would have made the analysis blind to many aspects.

It is valuable information about a work knowing whether several translations can be found in the catalog, whether several publishers were interested in the text, whether several reprints were made. Besides, it is informative whether a publication is present in only one German library (perhaps purchased by the FIDs) or if many German libraries have it in their catalog. Since part of the goals of some FIDs is to purchase relevant publications of a discipline, working at the title level would have disproportionately magnified the impact of the FIDs which would have skewed the results of the analysis. In order to avoid this, I decide to work at the exemplar level. In this way, the important library stocks of the FIDs are a part of the analysis, but all libraries in both consortia are considered equally.

I reject the options of working at the item level, and therefore ignoring the number of copies of each publication in each library. Although this could have some interest, the number of copies is stronger related to very specific aspects of each department, such as the number of students, the budget funds, or whether the publication was used in class.

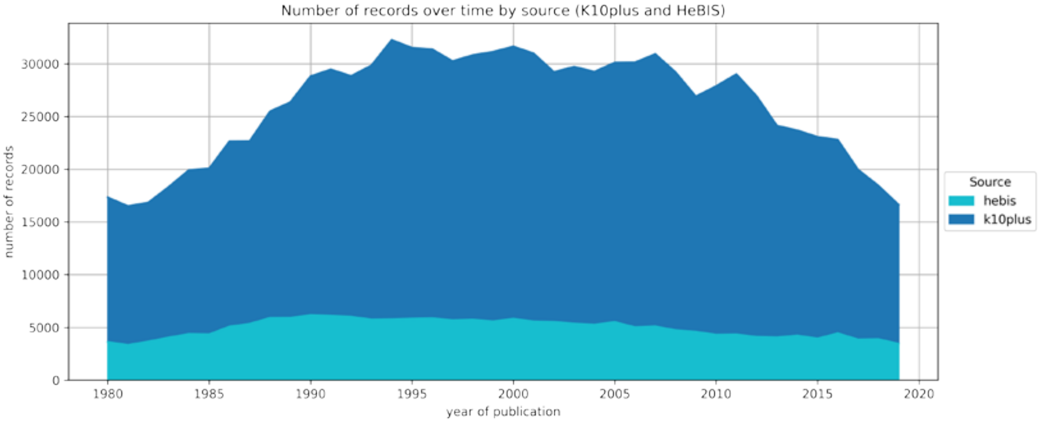
2.4 Description of the Data Set

After these steps, the available dataset contains 1,041,157 exemplars of publications in the libraries of both consortia. These exemplars are based on 334,221 different titles (identified with different PPNs in both consortia). That means that on average, each title tends to be present in three libraries of both consortia.



2 | Percentage of analyzed records in the databases of K10plus and hebis

Figure 2 shows the distribution of records in both consortia. The dataset contains 843,558 records from the K10plus database, which constitutes 81% of the analyzed exemplars. Correspondingly, hebis contributes with 197,599 records, around 19%. The difference of the number of records of both databases is not surprising since the K10plus covers a much wider area and therefore more libraries than hebis. Figure 3 shows the chronological distribution of the four decades with the number of records from both consortia.



3 | Number of records over time by source (K10plus and hebis)

The chronological distribution of Figure 3 shows an increase of records in both databases until the 1990s. After that, the number of annual records in the hebis decreases, while it remains stable in the K10plus, at least until the end of the 2000s. The decrease of annual publications is especially noticeable in the last 10 years in both databases. While there is a certain variation, the distribution between both databases remains rather stable over the entire period. The number of records obtained from the hebis database oscillates between 15% (the lowest point in 2011) and 24% (the highest point in 1987).

Has the number of publications in the field of Romance Studies been decreasing since 2010 but especially since 2015, as Figure 3 shows? Before jumping to conclusions, the characteristics of the analyzed dataset need to be remembered: records from libraries filtered based on classification. Libraries do not only purchase or obtain publications of the current or last year, but also from former years and decades. These publications then need to be cataloged and assigned to classification systems in order to be identified as publications for Romance Studies. In other words, it might take some years for libraries to complete the process of identifying, obtaining, cataloging, and classifying publications properly. For this reason, I consider all results from the last five years as preliminary.

Finally, I would like to add a remark about joining the records from both consortia. Although I have argued for the combination of the records from the K10plus and the hebis databases, this step also means an increase of the heterogeneity of the dataset. Although both consortia share many characteristics, some specifics about cataloging rules, classification systems or formats differ. For example, some Pica+ codes are different in both databases, and therefore the extraction of the data needs to be done separately. Another case relates to the classification systems: while BK is one of the most widely used classification systems in the K10plus, it is not used in hebis, where the RVK is more widespread. Although these subtle differences have little impact in this analysis, it needs to be considered the trade-off between the greater interest of combining data from different sources and the resulting decline of the quality of the data.

2.5 Extraction of Information and Normalization of the Data

The data of the catalog was saved as the Pica+-XML files. In order to analyze the data, a selection of the sub-fields of the catalog were extracted from the Pica+-XML fields and saved as columns in a tabular format. These tables then were saved as parquets files, a format for large or sparse tables. However, the content present in each field needs certain amount of pre-processing in order to be analyzed. For example, the Pica+ field 34D (sub-field a) contains the number of pages of each publication. Table 1 shows a few examples of the content that can be found there.

PPN	Title	Pages
1132286107	Antología poética	128 S
1604360445	Everest diccionario práctico de americanismos	238 S.
1604428902	Crônica da casa assassinada	XXXVII, 810 S.
1612842208	La @mort d'Agrippine	XXIV, 90 S.
228112982	Gewohnheit - heilen	476 S.
278864376	Cultures of the Aztecs, Mayas and Incas	216 S.
34184957X	Regards sur la littérature québécoise	312 p
489256120	French studies	Online-Ressource
745009786	Revista de Cancioneros Impresos y Manuscritos	Online-Ressource
798179643	Der @französische Wortschatz der Vorklassik	377 Seiten

Table 1 | Random sample of titles, with the original information of pages

Except for the online resources, the rest contain information about the number of pages. However, they are encoded slightly differently: with the German word *Seiten*, its first letter, or the letter "p". Some cases have a full stop at the end, others do not. Besides, two examples mark the length of an introductory section with Roman numerals.

If a researcher wants to calculate the mean of pages in publications, it is needed to extract the numerical values from the column "Pages" in Table 1. For this field, a regular expression was enough to extract the numerical values of these strings and assign them to the new column "pages extracted" in Table 2.

PPN	Title	Pages	Pages extracted
1132286107	Antología poética	128 S	128.0
1604360445	Everest diccionario práctico de americanismos	238 S.	238.0
1604428902	Crônica da casa assassinada	XXXVII, 810 S.	810.0
1612842208	La @mort d'Agrippine	XXIV, 90 S.	90.0
228112982	Gewohnheit - heilen	476 S.	476.0
278864376	Cultures of the Aztecs, Mayas and Incas	216 S.	216.0
34184957X	Regards sur la littérature québécoise	312 p	312.0
489256120	French studies	Online-Ressource	NaN
745009786	Revista de Cancioneros Impresos y Manuscritos	Online-Ressource	NaN
798179643	Der @französische Wortschatz der Vorklassik	377 Seiten	377.0

Table 2 | Random sample of titles, with the original and extracted information of pages

As can be observed, the simpler cases are correctly extracted. The pages in Roman numerals are ignored and therefore the information relating to the number of pages of these publications is simplified. Besides, the electronic resources are set as *NaN* ('Not a number') and these cases can easily be ignored in later calculations.

In other categories with the option of multiple values (for example, publications in several languages), I extract the data using ad-hoc tokenizers for the encoded information in Pica+ which normalize the data to a certain point. More details can be found in the companion Jupyter Notebooks, described in the following section. However, I decide to not adding many normalizing steps to the data and trust the work of the librarians who edit the catalog. Quantitative normalization of the data for the analysis can introduce new types of errors and noise, for example, joining together entities that should be treated separately.

When dealing with catalog data it is important to remember that only a few fields contain information relevant to all records. While the catalog contains information about the length in pages of 95% of the records, information about the price can be only found for 27% of the cases. This missing information is due to many causes: it was not feasible to obtain the information (for example, many publications do not contain any references to the year or publisher), some information may not be applicable to all records (for example, name of publisher in manuscripts) or the cataloging practice considers the information optional (such as price, see Section 4.5). This means that the analysis of some categories with a coverage close to 100%

of the records (such as the medium or the language of the publication) is much more representative than others (such as the price).

2.6 Publication of Code and Data

The data and code used for this publication are available online for anyone interested. Both components have been saved in two repositories: DARIAH Repository[12] and Zenodo.[13] The folder "data" contains the tables with the bibliographic records of the Romance Studies publications. The folder "code" contains scripts written in the programming language Python. These are in two formats: functions written in a Python script (with the ending ".py") and Jupyter Notebooks (with the ending ".ipynb"). Jupyter Notebooks are documents that can combine documentation, programming code and its output into a single file (VanderPlas 2016; Dombrowski, Gniady, and Kloster 2019). This way, any reader can have access to all steps, parameters and outputs that I considered during the analysis, many of which cannot be addressed in this article.

3. Linear Regression as Method and Example of Pages

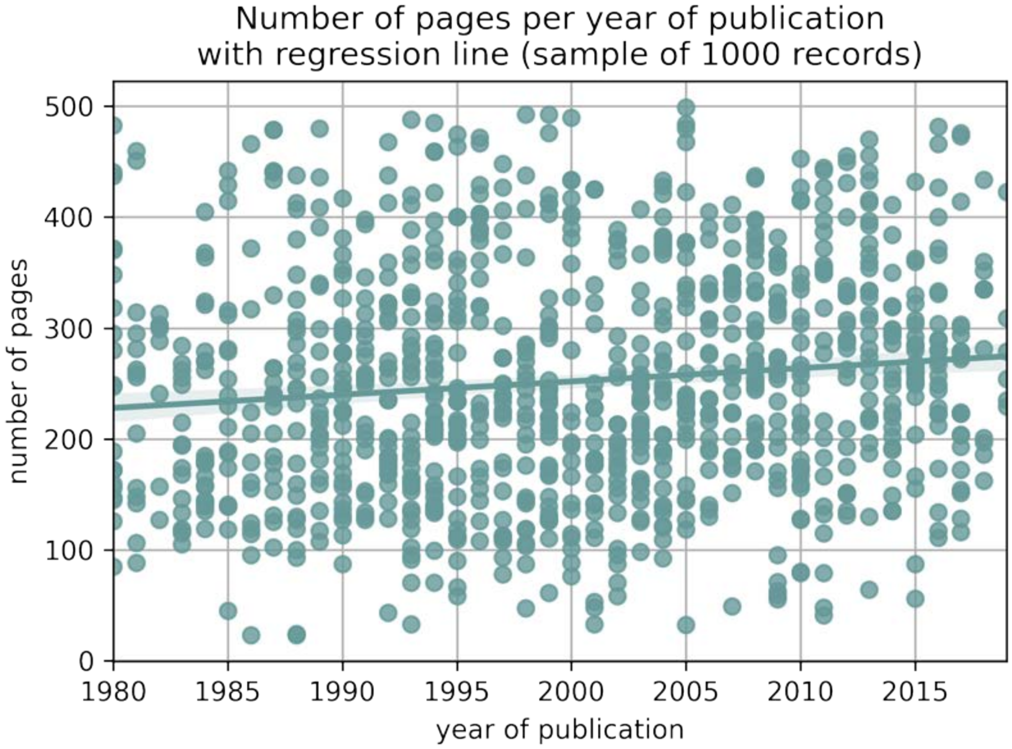
As mentioned before, in this article I mainly use linear regression for the analysis of the categories. Linear regression is part of the family of techniques from statistics known as regression, whose goal is to create a model from some observed data in order to predict numerical values for new cases (Evans 2014, 160–69). Therefore, regression can be seen as a method of Machine Learning (Müller and Guido 2016). Like classification, it is also a supervised method since the input data contains cases with the correct output-labels (in contrast to unsupervised methods such as clustering or dimensionality reduction). However, while the the task of classification in Machine Learning predicts categorical values (such as the genre of publication or the name of the author), regression predicts a numerical value, such as a price or a probability that something will happen. In this article, I use linear regression to predict what can be expected in future years for Romance Studies publications.

For an intuitive idea of this technique, in this section I take the information of the number of pages of each publication. Before looking at the real data, let us imagine an unreal scenario in which all publications in the field of the Romance Studies published in 1980 would be exactly 80 pages long. In 1981, all authors and publishers went a step further and every single publication would be 81 pages long. This would have repeated in 1982 (82 pages), 1983 (83), and kept on going during the entire period, so that all publications from 2019 would be 119 pages long. It is easy to predict in this unlikely scenario that for the year 2030, publications should be 130 pages long. Of course, this prediction is based on the premise that authors and publishers would follow the previous trend. In other words, this technique assumes a conservative perspective that the development will remain similar and therefore ignores the possibility of sudden changes.

[12] <https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000F-73EA-8>.

[13] <https://zenodo.org/record/7098830#.YyqblbTP1aS>.

Let us now move to the real data of the length of publications.[14] In Figure 4, each publication is a data point and these are sorted following two axes: the horizontal axis sets the year of the publication and the vertical axis shows the number of pages of each publication, with values between ten and 500 pages. As can be expected, each year shows a variety of publication length. However, looking only at the lowest and highest values on the vertical axis, it can be seen that publications become slightly lengthier over time: While in the first years there is no publication close to 500 pages, this milestone is surpassed by several publications after the year 1990.



4 | Number of pages per year of publication with regression line

This tendency is not only observable in these long outliers, but in the central tendencies by year. While the median number of pages for publications in 1980 is 219 pages, it is 256 for publications published in the year 2019. The linear regression model formalizes this tendency as a slope, a function that is visualized as a line in Figure 4. This slope is positive, meaning that an increase in the values of the horizontal axis increases the predicted values in the vertical axis. In other words, we can expect lengthier publications in the future. The exact slope of this category in this data is 0.9, which is the increase of number of pages that can be expected for each year. Using different components of the models, it can be predicted that publications will be 277 pages long on average in the year 2030, and 286 in 2040. As mentioned before, this will only happen in the (unlikely) premise that things will develop exactly as they have done until now. Although I do not expect these exact

[14] In order to obtain a clearer visualization, Figure 4 does not contain the entire dataset, but a randomly selected data sample of 1000 publications from Romance Studies. This selection was used with the function sample of the Python library Pandas, which allows the user to get random samples of a given size. The reported slope and p-value are based on the entire dataset and very similar to the results based on the sample.

values to be perfect predictions for the future, they show the expected tendency for future years.

This uncertainty can be seen as a weakness of this study, but actually this is a problem inherent to any Machine Learning application. Regardless of the complexity of the data or the algorithms, current Machine Learning is based on the premise that new or future cases will follow the same principles as the ones that were observed in the past. This weakness of Machine Learning is perhaps clearer in this study because of the simplicity of the algorithm used and because the predicted values lay in the future, which of course cannot be predicted with certainty.

The exact function in Python that I use for the regression model can be seen in the Jupyter Notebooks. This function (from the library Scipy) gives also other values, such as an intercept, an r-value, a p-value and the standard deviation of the r-value. In this article, I report the p-value to observe whether the calculated tendency of the linear regression is statistically significant. Since the p-values correlate negatively with the size of the data and the dataset is relatively large, I assume a p-value lower than 0.001 for statistical significance. This is the case for the analyzed data of the pages in this section.

4. Analysis

In the following sections, I describe and analyze several categories extracted from the fields of the catalog, mentioning the Pica+ codes that were extracted from the original sources, the percent of records with this information in the catalog (coverage), the historical development of the past decades and then the prediction for the next decades.

4.1 Language of Publication

The language of publication is a frequent topic of reflection and discussion in Romance Studies. In general, it is accepted that three groups of languages are the main options (Lieber and Wentzlaff-Eggebert 2002):

1. The Romance languages, with a traditional predominance of French, followed by Spanish and Italian (Schrott 2003)
2. German, as the native language of many scholars in Romance Studies and the language of scientific communication in the Humanities in the German-speaking countries
3. English, as the international language of communication that the different communities can at least read

Each of these three options bring advantages and disadvantages, and these have been discussed in previous publications. In general, many researchers in Romance Studies argue that Romance Studies need to be a multilingual field, with many of them supporting German as language of publication; although it is accepted that the influence of English is increasing, it is seen as an undesired process which can bring harm both to German-speaking researchers and to the field of Romance

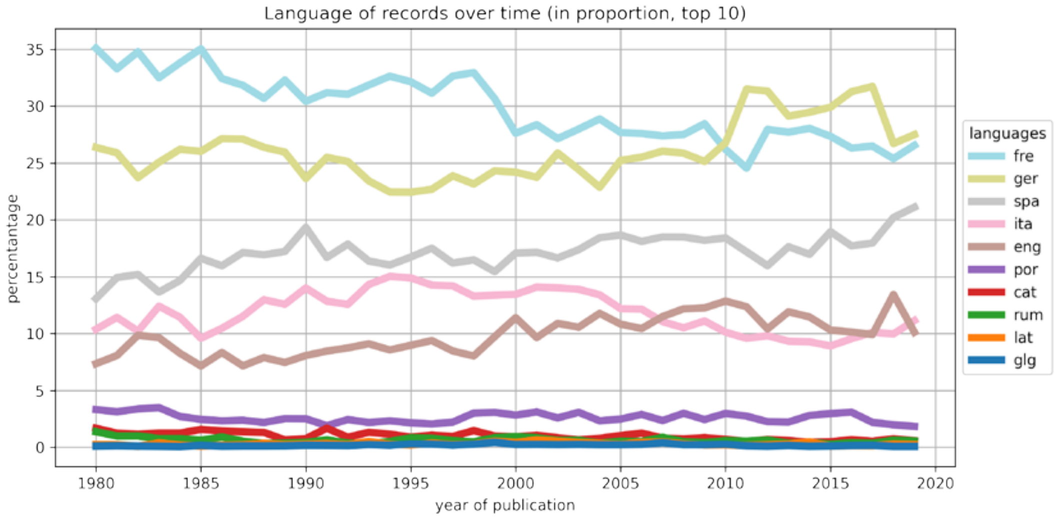
Studies (Lieber and Wentzlaff-Eggebert 2002; Kramer 2002; Constantinescu 2002; Nitschack 2002). Previous research has been centered on the role of the Romance Studies as publication language in different academic areas (Burr 2008; Kramer 2008; Haarmann 2008).

My hypotheses (based on articles or opinions spread throughout the community) for this period are the following:

1. The number of publications in French is being reduced over time (cfr. Wandruszka 1988; Haarmann 2008; Kramer 2008)
2. The number of publications in Spanish is increasing
3. The number of publications in English is also increasing

For the rest of the languages, including German, I do not have specific expectations but they are also considered for an exploratory analysis. It is especially interesting to observe the development of Romance languages such as Romanian, Catalan or Galician.

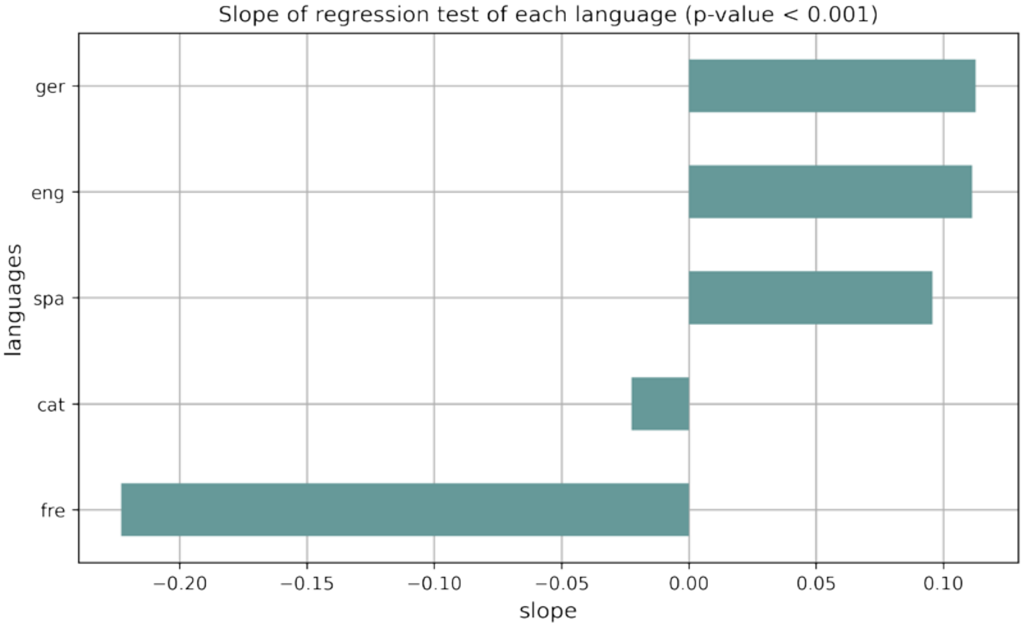
The data for the language of publication is extracted from the Pica+ field 010@ (sub-field a). The coverage of this category is surprisingly high, with values in 98% of the records. Multiple values are possible, the combination of German and French being the most frequent (29,302 publications). In total, the analyzed dataset contains publications in more than 170 languages (such as Russian, Dutch, Polish, etc.).



5 |Development of percentage of records by language (top ten)

Figure 5 shows the development of percentages of the records in the ten most frequent languages in the dataset. As can be seen, French is the language with the largest number of publications, with a total of 33.48%. A decreasing tendency of publications in this language can be observed for the entire period. French is closely followed by German with 29.17%, with a clear increase after 1995 and surpassing French in 2010 and therefore becoming the predominant language. Spanish is the third language with 19.44% of the records and a solid increasing tendency of publications for the entire period. 13.66% of the publications are in Italian, with an increase until the 2000s and since then first certain decrease and finally stability around 10% for the last years. English is the fifth language with 11.21% of the records, with an increasing tendency during the entire period and surpassing Italian after 2007. The difference between these five languages and the next ones is notable: only 2.92% of the records are in Portuguese, followed by Catalan (1.08%), Romanian (0.67%), Latin (0.35%), and Galician (0.22%). As mentioned before, if the records from the BVB library consortium with the FID for Eastern Europe were part of the dataset, the number of publication in Romanian would be likely higher.

The next step is to apply linear regression in order to quantify and evaluate the observed tendencies. For this, I run separate regression models for each language. Figure 6 shows the positive and negative slopes for those languages whose results have p-values smaller than 0.001. Only five languages show trends with statistical significance: German, English, Spanish, Catalan, and French. That means that the models are not able to make statistically significant predictions for languages such as Italian, Portuguese, Latin, or Galician, for example, because their situation is rather stable during the analyzed period.



6 | Positive and negative statistically significant slopes of linear regression models analyzing language of publication

As can be seen in Figure 6, German, English and Spanish have positive slopes, all three with values close to 0.1. This means that an increase of 0.1% of publications in these languages can be expected every year. For the year 2030, 29.46% of the publications can be expected in German, 20.04% in Spanish and 13.22% in English. On the lower part of Figure 6 the languages with negative slopes can be observed. While Catalan has a very small slope of only -0.02, French has the strongest absolute value with -0.22. With this tendency, the model predicts 23.04% of publication in French in 2030. By the year 2040, French is expected to be surpassed by Spanish and become the third most widely used language of communication in the Romance Studies, with still a large distance to English.

4.2 Place of publication

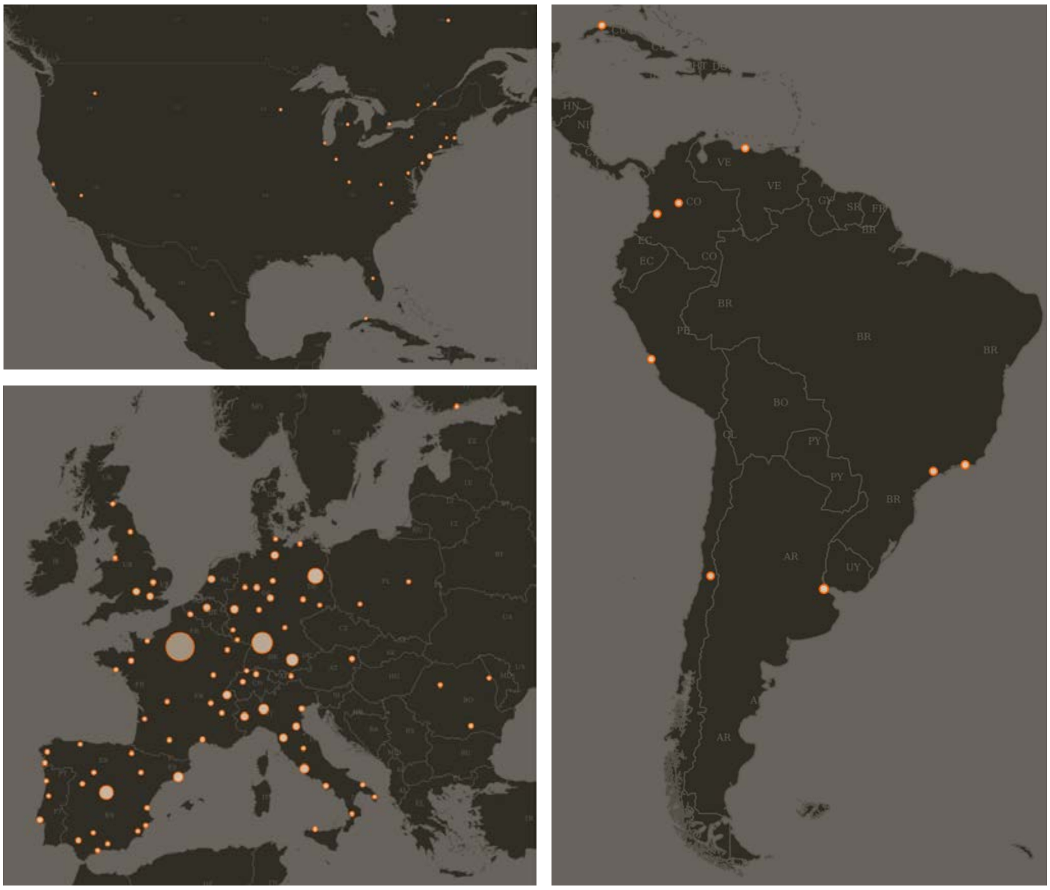
A further category strongly associated with the language of publication is the place of publication, which is the next analyzed category. Since the European countries have moved in the last decades towards a political and economic integration, it could be expected to see an increase of publications from the Romance-speaking countries. Besides, I want to explore the tendency of the number of publications from Romance-speaking countries from other continents.

This information can be found in the Pica+ field 33A, sub-field p. As a typical bibliographic information, its coverage in the catalogs is very high, with values in this field in 99% of the analyzed records. For those publications with several places of publications (for example, Berlin-Boston), the cataloging rules foresee several separated fields.[15] The field then is tokenized following this and trying to correctly extract places with names composed of several tokens (such as Frankfurt am Main, Buenos Aires, New York, etc.). Figure 7 shows an overview based on a random sample of 5,000 records,[16] visualized through the DARIAH Geo-Browser. The data loaded in the Geo-Browser is available online for further exploration.[17] The maps in Figure 7 show the dominance of places of publication in the German-speaking area, France, Spain and Italy. Besides, several publications come from the south-east of England (London, Oxford) and the East Coast of the United States and Canada. In Latin America, only exclusively the capitals of some countries are covered in this sample of 5,000 publications, which does not contain any publication from Africa, Asia or Oceania. Of course, this does not mean that the sample does not contain research from authors coming from or based in these areas, since many of them publish through printing houses based in other countries.

[15] However, the cataloging practice might have developed over the years, introducing in the catalog only the first place of publication for a period of time. This might influence the results, for example if a city tends to appear only as secondary place of publication.

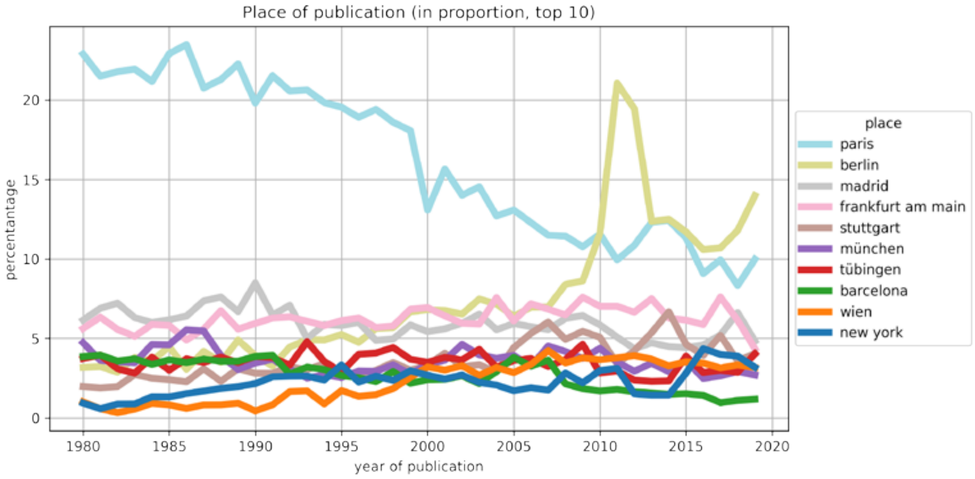
[16] As for the sample of Figure 4, this random sample is created with the function *sample* of the Python library Pandas.

[17] <https://geobrowser.de.dariah.eu/?csv1=https://cdstar.de.dariah.eu/dariah/EAEA0-6966-72CA-0E10-0>.



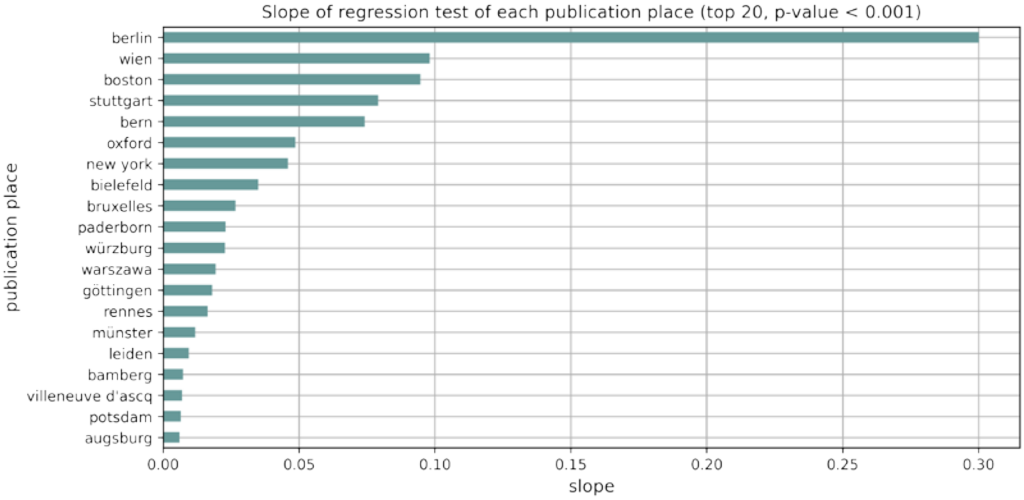
7 | Maps with the place of publications of 5,000 publications, centered in Europe, North and South America

Figure 8 shows the development of the top ten publication places in the analyzed period. The decrease of Paris as a place of publication for the Romance Studies is especially remarkable, with more than 20% of the records in the 1980s, but under 10% in some of the most recent years. The other remarkable development concerns Berlin, with only 3% in 1880 and peaking in 2011 of 21%. This peak will be further explained in the following sections about publishers and medium. Beyond these two places, the rest of the top ten places of publications vary between 8% and 1%. Five of them are in Germany, one is in Austria (Vienna) and only three are in Romance-speaking countries: Paris, Madrid and Barcelona.



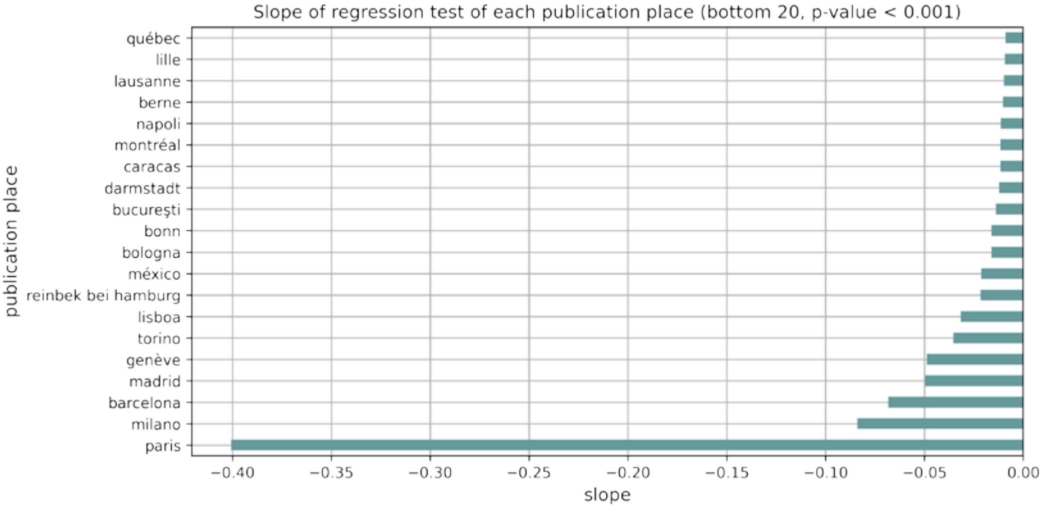
8 | Development of percentage of records of place of publication (top ten)

To observe the development of these and other places, I calculate the linear regression models for each place. Figure 9 shows the places with positive slope and a p-value under 0.001. Berlin shows the highest slope, with a predicted increase of 3% in the records of the catalog within 10 years. This is followed by several places with slopes over 0.03, all of them in the German or in the English-speaking area: Vienna, Boston, Stuttgart, Bern, Oxford, New York and Bielefeld. Besides Bruxelles, Rennes and Villeneuve d'Ascq, the rest of the places in this visualization are in Germany (with some exceptions for Poland and the Netherlands).



9 | Positive statistically significant slopes of linear regression models analyzing place of publication (top 20)

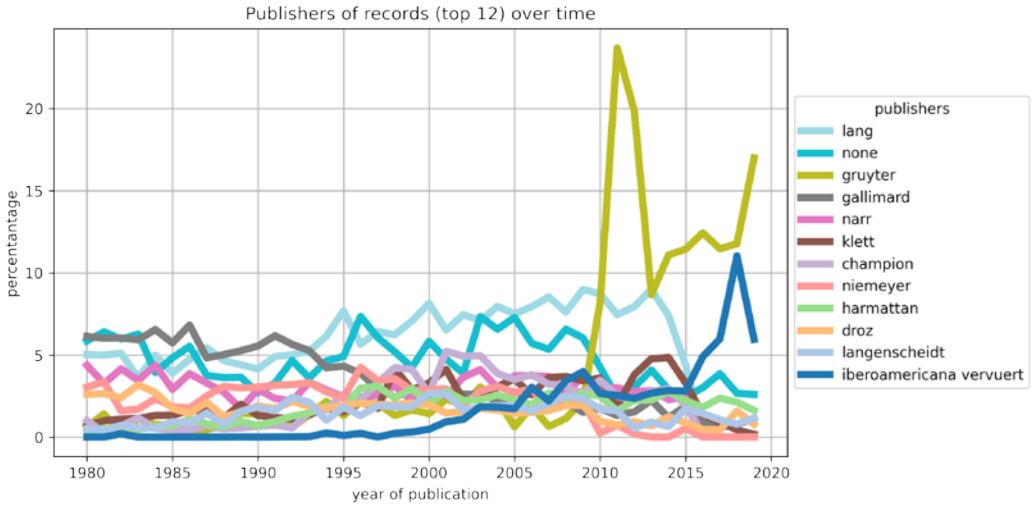
On the opposite side, Figure 10 shows the places with negative slopes. By far, the lowest slope is for Paris with almost a -0.4. The models predict 4.15% of the records being published in Paris for the year 2030. In general, the great majority of the places with negative slopes are from the different Romance-speaking countries: France, Italy, Spain, Portugal, the French-speaking areas of Switzerland and Canada, Mexico, Romania, Colombia. Besides, some places in Germany are in this figure, such as Bonn or Darmstadt.



10 | Negative statistically significant slopes of linear regression models analyzing place of publication (top 20)

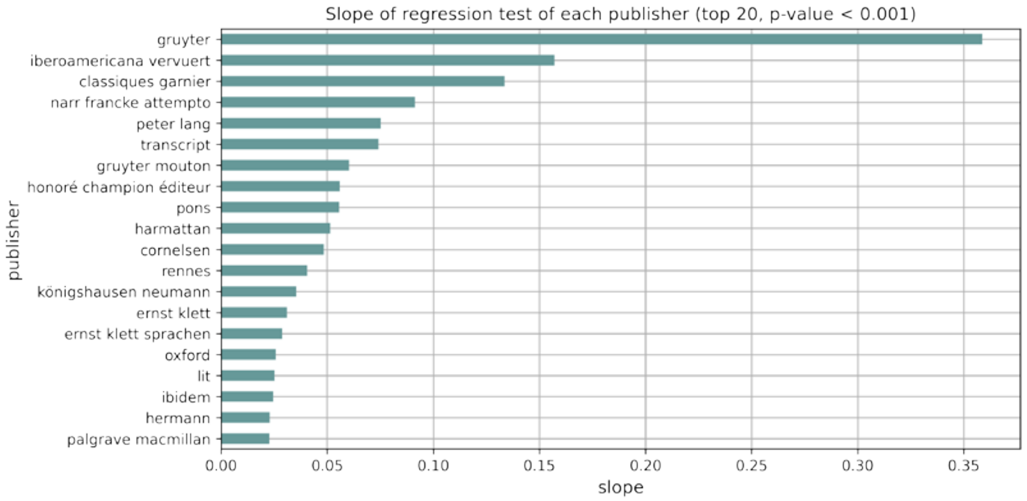
4.3 Publishers

Also strongly related to the language and place of publication is the development in the number of publications by publisher. As in the previous categories, the catalog has information for this field for a large majority of the records (98%) which is contained in the Pica+ field 033A (sub-field n). This field contains many abbreviations, such as *Verl., Univ., Ed., Éd., Pub., GmbH*, etc. I decided to delete all the references to the form of the institutions and a series of stop words in different languages (*&, of, de, von*, etc.). Besides, it needs to be considered that in many cases, the publishers change their name, sometimes because they are integrated into other publishing houses, sometimes because they create a specific imprint for some niches. For example, historical changes of the name can be observed for publishers such as Peter Lang, Narr, or Iberoamericana / Vervuert.



11 | Development of percentage of records of publisher (top twelve)

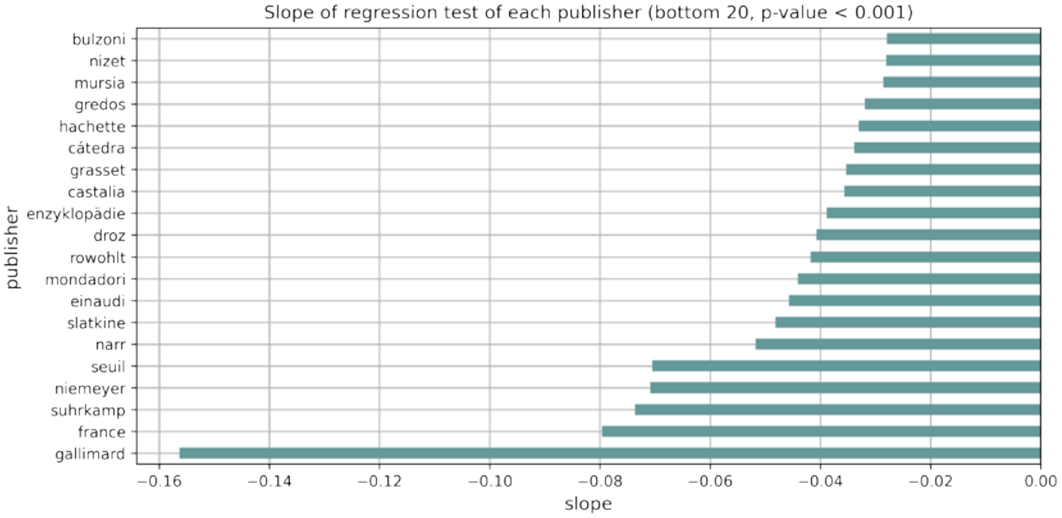
The historical development of the proportion of publications of the top twelve publishers is shown in Figure 11.[18] While in the 1980s the top publishers contributed between 6% and 1% to the catalog, after the 2000s publishers tended to increase or decrease their proportion of the publications notably. In general, a concentration of the publishing market can be observed in the sense that a few publishers (De Gruyter, Peter Lang, Iberoamericana / Vervuert) have been notably increasing their contribution to the catalog, while in general many smaller publishers reduce theirs. This will be tackled again in the next section about the development of the printed and digital publications.



12 | Positive statistically significant slopes of linear regression models analyzing publishers (top 20)

In Figure 12, the publishers with the highest slopes can be observed. De Gruyter obtains an exceptionally high slope over 0.35, followed by other publishers with slopes of 0.05 or greater, such as Iberoamericana / Vervuert, Classiques Garnier, Narr Francke Attempto, Peter Lang, transcript, Honoré Champion éditeur and Pons. Many of these publishers have their main location in the German-speaking area and therefore these results cuold be seen as an explanation of the results of Sections 4.2.

[18] The reasons for showing twelve cases and not ten is the development of Iberoamericana-Vervuert in the last five years. In the range between the top ten and 20 publishers, no other case has increased its presence similarly to this publisher.



13 | Negative statistically significant slopes of linear regression models analyzing publishers (top 20)

The opposite is shown in Figure 13, which lists the publishers with the lowest slopes. The publishers with the lowest slope are Gallimard (-0.15), followed by Presses Univ. de France (appearing only as *france* in the figure), Suhrkamp, Niemeyer, Seuil and Narr with slopes under -0.6. Some traditional publishers from France, Italy and Spain also appear in this figure, such as Droz, Mondadori, Castalia, Cátedra and Gredos.

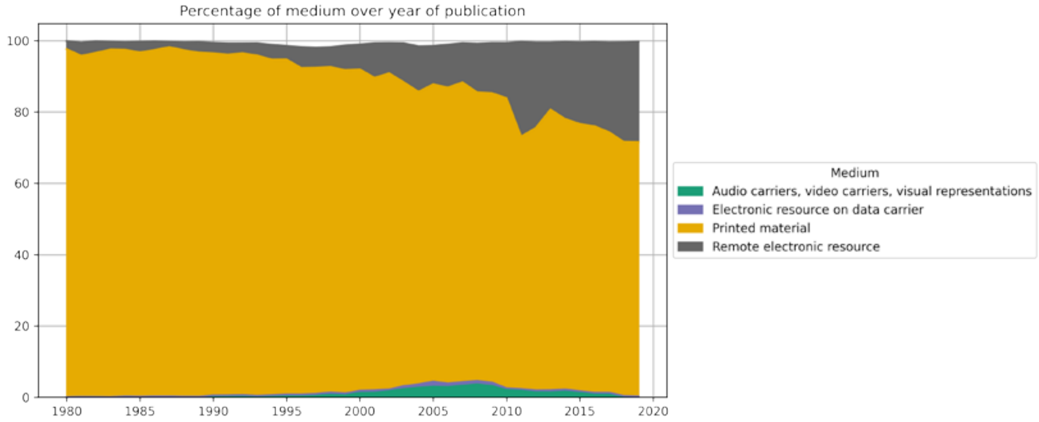
4.4 Medium

What is the current development of e-books in Romance Studies? This question is observed in this section, analyzing the data encoded in the Pica+ field 002@, subfield 0. Since this field is mandatory, it is included in 100% of the records of the catalog. Although the documentation of this field foresees many possible values, the majority of them are very rare in the catalog. In this analysis, I only focus on four categories:

- Printed material
- Remote electronic resource (e-books)
- Audio carriers, video carriers, visual representations (such as DVDs or Blu-rays)
- Electronic resource on data carrier (such as databases in CDs)

Of course, an increase in the number of e-books can be expected, and the main interest is how strong this is happening and if there are historical milestones in the past decades.

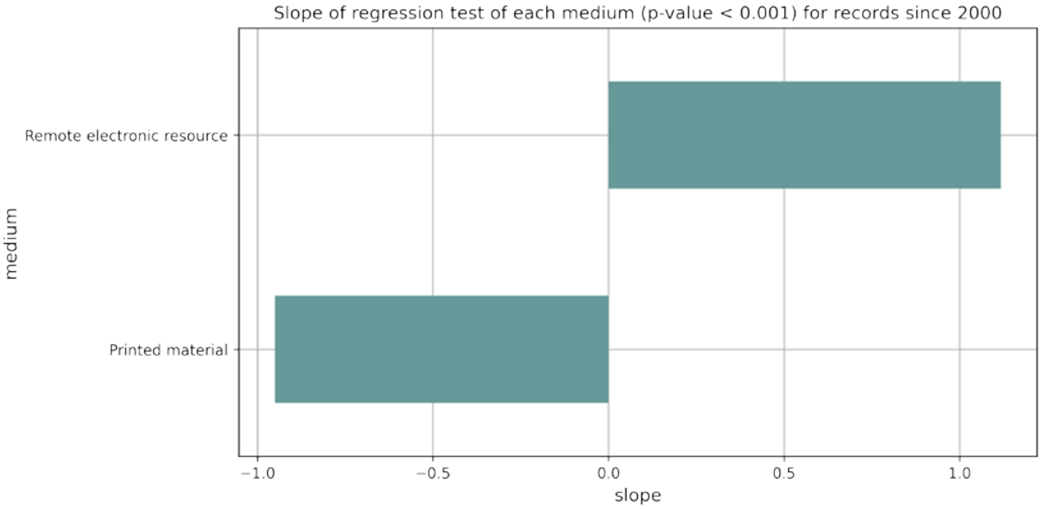
The chronological distribution of the medium can be observed in Figure 14. In general, printed material constitutes 88.16% of the total of records, while e-books represent 10.17%. The other two categories only represent around or less than 1% of the records. The contribution of the two other media is observable between the years 2000 and 2010, with a clear decline since then. Figure 14 shows that a section of the e-books are marked as published in the 1980s, when e-books were not purchased by libraries. That means that previous publications are being published in this format and therefore entering the catalog. However, after the year 2000 the share of e-books surpasses 10%. After 2010, e-books tend to represent over 20% of the publications.



14 | Development of percentage of media

In 2011 a peak of e-books can be observed. In previous categories, parallel peaks were visible also in this year, such as an increase of publications from Berlin and from the publisher De Gruyter. This publisher based in Berlin made many previous publications available as e-books and launching them as published in 2011. The treatment of the e-books by De Gruyter could be also one of the factors that explains the observed success of the publisher in Figure 11. In contrast to many other publishers from the Humanities, De Gruyter tends to offer e-book licenses to its publications. Besides, the prices of these are similar to the printed version and their e-books can be downloaded in a single file by any user of the library. In other publishing houses, the user of the library can only download sections of the publication (for example a certain number of chapters or pages) and the prices tend to double or triple the price of the printed version. In the following section, I will give more details about the prices, also distinguishing between printed material and e-books. In any case, these results arise the question whether the digital paradigm is actually reinforcing the national publishing markets.

For the prediction of this field in future years, I consider two models: one for the entire period and another one only with the data for the last 20 years. Since e-books were rather insignificant for libraries before the year 2000, it is questionable to take this data for predicting the future. Figure 15 shows that the expected tendency is an annual increase of e-books of 1%, with the consequent similar decrease of the printed material. The linear model of four decades gives similar but more conservative results, with negative and positive slopes for both media around 0.7% (further details in the Jupyter Notebooks).



15 | Positive and negative statistically significant slopes of linear regression models analyzing media

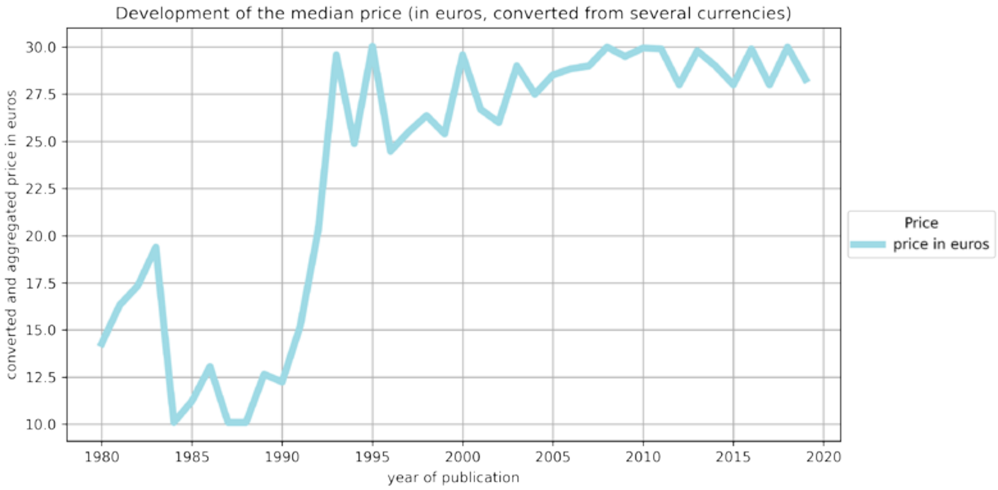
Although this shows a clear tendency, it does not describe a revolutionary change in the medium of the publications. Although it is unlikely that printed material will disappear completely, one could ask, with this current tendency, when a fully and entirely digital catalog for Romance Studies can be expected. The model predicts this taking place in the year 2080.

4.5 Price

In the previous section I have commented on the differences between the prices of printed material and e-books. In this section, I explore the development of the prices in the last decades and compare the price of both media. It is important to admit that the catalog is not the ideal source of data for prices since they are only kept as a comment to the ISBN in the Pica+ field 005A (sub-field f). In contrast to the previous categories, the coverage of this field in the catalog is much lower, with some information only for 27% of the records. The scarcity of the data is stronger during the 1980s with less than 10% of the records, while it increases up to 57% for the publications of the last years. Besides, this field brings further challenges:

- There is no homogeneous way to encode the price. Here are just some possibilities for the same price: 18.95 €; 18,95 €; 18.95 EUR; EUR 18.95 (DE); 18.95.
- The euro was introduced in several European countries, among others in Germany, where the analyzed libraries are located
- Some publications contain their prices in foreign currencies
- Some publications have information about the price in several currencies

All these problems required a special treatment of this field. The specific functions and regular expressions for each currency can be found in the Jupyter Notebooks and the Python code. For this analysis, I consider the prices assigned in euros, German marks, Swiss francs, British pounds, and US dollars. To compare the prices, I convert them into euros, following the average rate of the last years.[19] After these steps, I was able to obtain a price in euros for 19.86% of the records. I am aware that comparing absolute values from several currencies, countries and decades can be problematic. The goal is not to present an exhaustive analysis of this factor, but have a first glimpse of the development.



16 | Development of the median price

Figure 16 shows the historical development of the median price. As mentioned before, the dataset contains little data about the price of publications during the 1980s, therefore I argue that the apparent large increase of the price at the beginning of the 1990s is just an artifact of the scarcity of the data. Figure 16 shows a slow increase of the price since the 1990s, with a current median price being slightly under 30 euros.

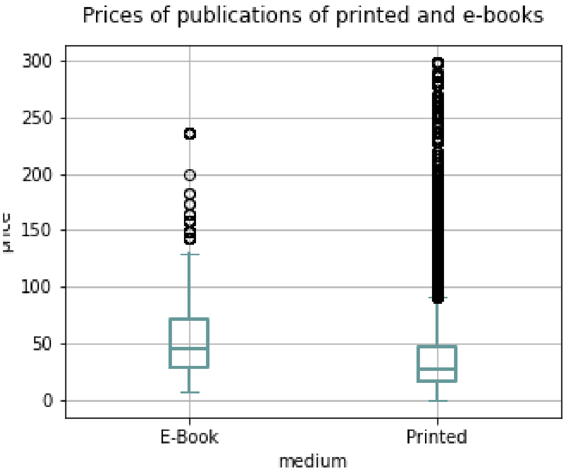
The linear regression model outputs a positive slope of 0.15 (p-value < 0.001), which corresponds to an annual increase of 15 cents of euro and a median price of 31.70 € for the year 2030.

When the current currencies are analyzed separately, it can be observed that while euros, Swiss franks and US dollars are stable, the prices in British pounds are increasing notably. While the median price in British pounds in 1995 was 39.60, it increased up to 76.56 for the year 2015.[20] A linear regression model gives a slope of 1.51 (p-value < 0.001), predicting for the year 2030 a price of 93.79 € if publications are originally priced in pounds.

[19] Which are 0.85 for US dollars, 1.32 for British pounds, 0.76 for Swiss franks and 0.51 for German marks. Further details in Jupyter Notebook and <https://www.ofx.com/en-au/forex-news/historical-exchange-rates/yearly-average-rates/>.

[20] Although this data represents the prices in pounds, the actual values here are expressed in euros.

Finally, I compare the prices of printed material and e-books. Figure 17 shows the distribution of the prices through box plots for both media in the dataset. While the median price for printed publications is 28 €, it is 46.32 € for e-books. A Welch's t-test throws a p-value lower than 0.001, meaning that, at least for this dataset, it can be said that e-books are statistically more expensive than printed versions.



17 | Comparison of prices printed and e-books

Of course, it needs to be considered that both media are highly imbalanced in the dataset. While I was able to obtain the price of 22.29% of the printed versions, this was only the case for 0.72% of the e-books. Although the results confirm the experience in the daily work of the library, further research and discussion about the price of e-books is needed.

Even though the development of the prices reflects only a subtle increase (except for publications from the United Kingdom), the higher prices of e-books need to be properly addressed. The budgets for the purchase of literature in German libraries have been frozen in the last decade. Digitization in the library and remote access to research publications have become central in the last decade and reinforced by the COVID-pandemic since 2019 (Bieselin 2005; M. Ernst 2021). An increase in the number of available e-books can only be possible with an increase in libraries' budgets, either for purchasing e-books or for the support of Open Access publications.

4.6 Keywords

The last category I explore in this analysis is related to the keywords (*Schlagwörter* in German). Keywords are controlled vocabularies composed by terms from one natural language, used to describe the content of a document in the catalog as accurately as possible (Gantert 2016, 198–99). Table 3 contains ten randomly selected examples from publications with their title and the keywords that can be found in the catalog.

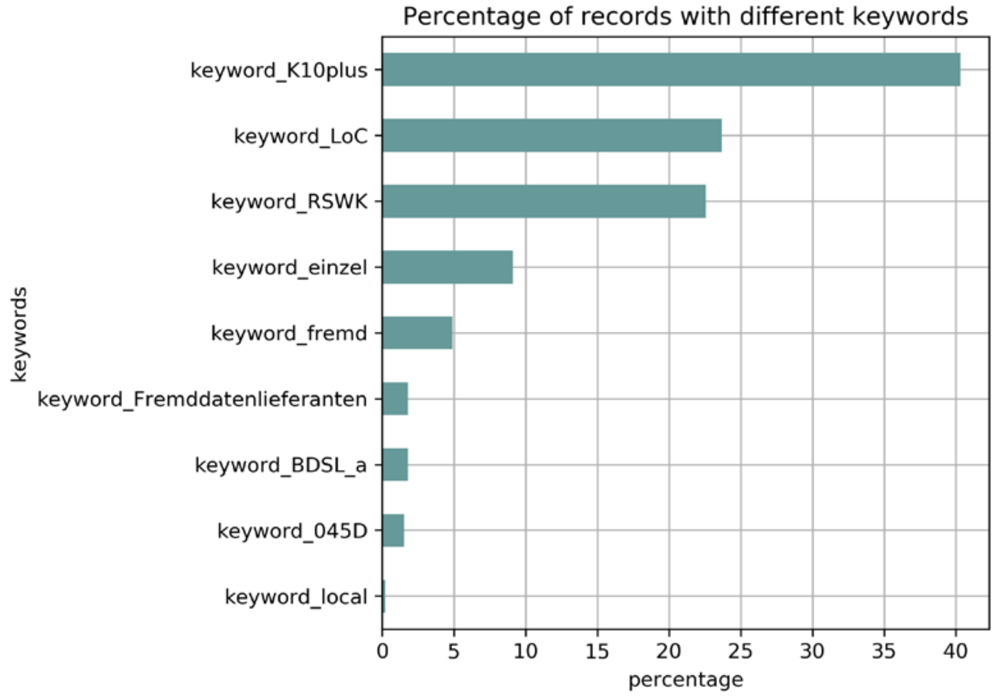
Title	Keywords
Edizione nazionale delle opere	Alberti
Méthodes de français à l'école secondaire en Suisse alémanique	Deutsche Schweiz|Personalpronomen|Fehleranalyse|Französischunterricht|Gymnasium|Pronominaladverb
Alexis de Tocqueville	Tocqueville
Lesage, écrivain	Le Sage
Von der "novela social" zur "nueva novela española"	Goytisolo|Roman|Goytisolo, Luis|1935-|Criticism and interpretation|Social problems in literature|Spanish fiction|20th century|History and criticism
Zwischen weißer und schwarzer Schrift	Jabès|Literaturtheorie|Schreiben
Diccionario fonético descriptivo de la lengua española	Spanish language|Spanisch|Aussprache|Deskriptive Phonetik|OBV
Elio Vittorini und die moderne europäische Erzählkunst (1926 - 1939)	Vittorini|Vittorini, Elio|1908-1966|Criticism and interpretation
Michel Tournier et le détournement de l'autobiographie. Suivi d'un entretien avec Michel Tournier	Tournier, Michel|Criticism and interpretation|Self (Philosophy) in literature|Tournier|Autobiografische Literatur
Der König im Kontext	Calderón de la Barca|Comedia|König|Herrschaft|Kontingenz|Geschichtsbild|Calderón de la Barca, Pedro|Kings and rulers in literature

Table 3 | Examples of ten publications and their assigned keywords in the catalog

In contrast to the previous categories, these keywords allow the researcher to gain insight into the actual content of the publication. For example, until now I have only explored which language was used for the text of the publication. Perhaps the decline in French can only be traced as publication language and researchers are still equally interested in French language, literature or culture, while publishing their results in other languages such as German or English.

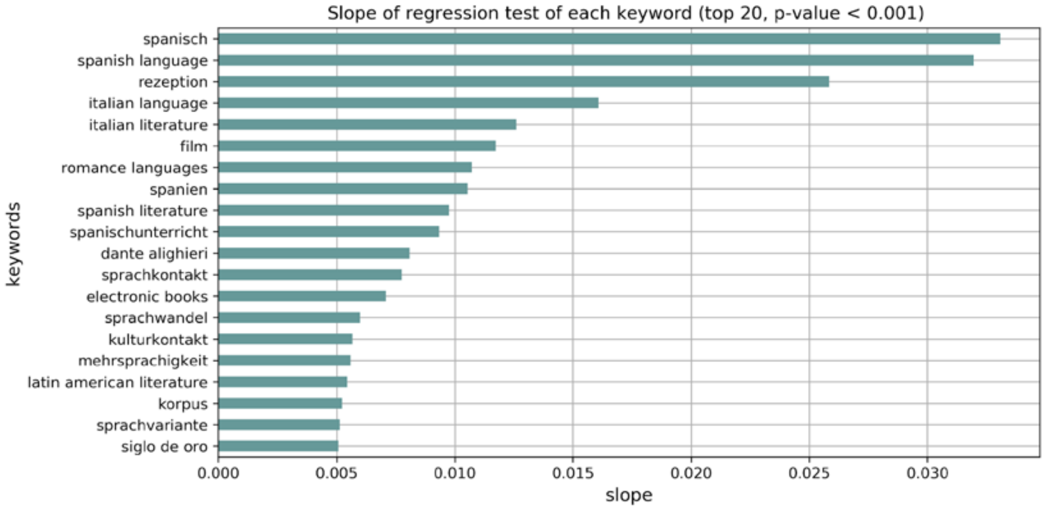
Although a similar analysis would have been possible by using the classes of the classification systems, I decided to reject this for two reasons. First, because the classes of classification systems were already applied to create the dataset. It could be argued that an analysis of the classes would be circular to a certain degree. Second and most importantly, keywords express more specific concepts than the classes of classification systems. Some examples of this can be seen in Table 3, such as “Kings and rulers in literature”, “Deskriptive Phonetik”, and “Fehleranalyse”.

The catalog contains different formalizations of the keywords depending on the source, the language used to express the keywords (English or German) or the controlled vocabulary applied. Figure 18 shows the coverage of several fields from the catalog containing this information. The fields with the highest coverage are the Pica+ fields 044K (in the Figure as keyword_k10plus), 44A (in the Figure as keyword_LoC) and 41A (keyword_RSWK). Each of these three fields contains data for more than 20%, while the rest covers fewer than 10% of the cases.



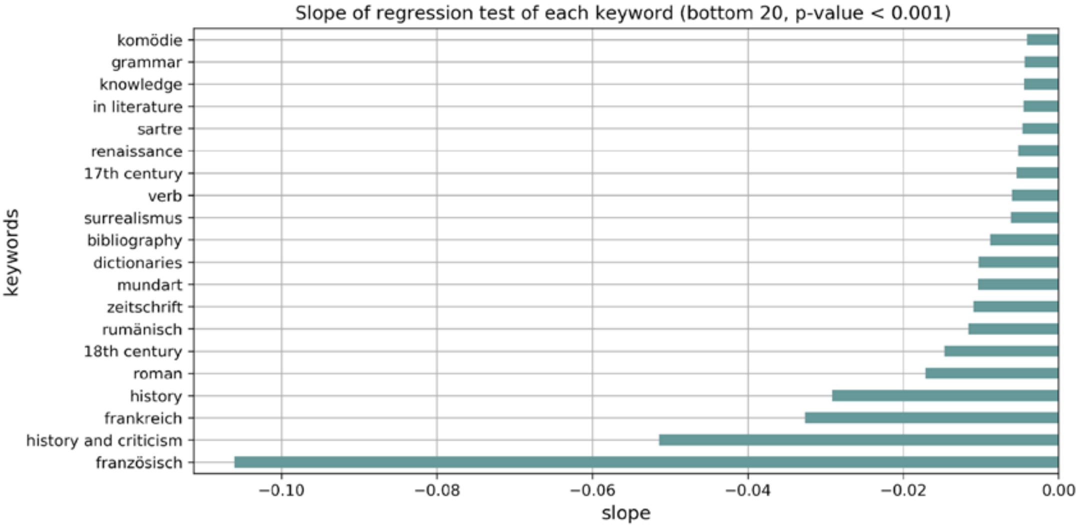
18 | Percentage of records with different keywords

For this reason, I decide to work only with these three categories. While the LoC keywords are expressed in English, the other two are in German. Together, the three fields cover 56.29% of the dataset for Romance Studies publications.



19 | Positive statistically significant slopes of linear regression models analyzing keywords (top 20)

The twenty keywords with the highest slopes with statistical significance can be observed in Figure 19. Many of these keywords relate to the Spanish language, literature and culture, with two keywords with the highest slopes followed by others (*Spanien, Spanish literature, Spanischunterricht, Latin American literature, siglo de oro*). In any case, the slope seen for Spanish in 4.1 was much higher than the slopes in Figure 19. This could be explained with the fact that the language is present in several keywords, keeping lower values. However, it could be pointed out that Spanish is increasing its role of publication language beyond the general interest of Spanish language and literature. In the case of Italian, in contrast to the stable situation seen in Section 4.1, the keywords show an increasing interest in Italian language and literature. This is reinforced by the presence of the keyword Dante Alighieri in these results. However, both keywords are expressed with the LoC vocabulary, which can point to an increasing interest in the English-speaking countries, while this could not be not the case in the German-speaking area. Numerous keywords associated with linguistics relate to contact and changes across and within languages (*Romance languages, Sprachkontakt, Sprachwandel, Kulturkontakt, Mehrsprachigkeit, Sprachvariante*). On its side, literary studies show positive trends relating reception and film. Finally, two further keywords can be associated with the new digital paradigm: *electronic books* and *korpus*.



20 | Negative statistically significant slopes of linear regression models analyzing keywords (top 20)

In contrast, Figure 20 shows the keywords with the lowest slopes. In coherence with the results seen in the previous sections, French appears in the keywords with the two lowest slopes, reinforced by the presence of a keyword for Sartre. In contrast to the case of Italian in Figure 19, the keywords relating to French are only expressed in German, which could reflect that this trend can be only observed in the German-speaking area. Romanian is the only other language on these bottom 20 keywords, which is the observed tendency in the analyzed catalogs, but it could have been notably different if the records from the BVB had been part of the dataset. Several keywords from Figure 20 are related to publications types, such as journals, dictionaries and bibliographies. Besides, two classical literary genres show strong decreasing tendencies: novel and comedy.[21] Literary Studies seem less

[21] However, this trend could be traced rather to a change of the librarian practice, assigning in the past years more specific keywords.

interested in previous periods (Renaissance, 17th and 18th century) and surrealism, while this is the case for verb and grammar in Linguistics.

5. Conclusions

In this article, I have used library records of publications in order to describe the development of the Romance Studies in the past decades and make predictions about the next decades. Although the predictions for the next years are interesting, all of them are based on the simplification that the tendencies of the past will remain, which might no be the case. This simplification is a general limitation of current Machine Learning approaches, which are based on the premise that new cases can be predicted following what was observed in the past. The analyzed dataset coming from the hebis and K10plus library consortia show different trends relating to the different analyzed categories.

The presence of German language publications and of publishers based in the German-speaking area is increasing in the Romance Studies. There is a decline in the importance of French which can be observed in the language of publication, place of publication, publishers and research topic. Spanish shows a clear increase both as publication language and topic of research, however this does not translate into an increase of publications coming from Spain. Italian shows a rather stable situation, with certain positive trends as a research subject.

E-books have become an important part of the publications of the Romance Studies, the change to the new digital paradigm will take decades in the current development. The results suggest that the new digital paradigm could be reinforcing the national publishing market, since German libraries could prefer acquiring e-books from German publishers. Besides, this study shows that e-books are notably more expensive, which needs to be addressed by an increase in the libraries' budgets.

In Linguistics, grammar and syntax seem to be in decline, while topics relating to language contact, variation and multilingualism are increasing. For Literature Studies, previous periods (1500-1800) and classical genres and publications have made room for new research subjects such as films and cultural contact.

Although these results are representative for a large section of the German territory, this analysis does not exhaust further possibilities. In future studies, the dataset could be expanded to more sources, countries, periods and other publication types (specially chapters and journal articles). However, the decline of data quality needs to be considered when combining several sources. Moreover, further possibilities of analysis are worth exploring, such as the generation of keywords (both from supervised or unsupervised tasks) or the annotation of the titles through lexical resources in several languages such as WordNet.

The results of this study can be used by researchers, Romance Studies departments and libraries to reflect their own decision. When researchers decide the language of the publication, the publisher of their next anthology for a section of a conference, or the topic of a new professorship, it can be decided to reinforce or not the current trend. The same is true for libraries, which influence the reception of a publication by their purchase decisions.

In this article, I have shown the potential of the bibliographic data science analysis applied to the last decades of a specific discipline. These research studies are only possible thanks to the valuable work and experience of the professionals in the libraries, who daily curate the data in the catalog. If we believe that data is especially valuable in the new digital paradigm, we also need to acknowledge and promote the role of the professionals curating the data on a day-to-day basis. In any case, this kind of analysis requires the combination of knowledge relating to libraries, the specifics of the discipline, and the use of digital and statistical methods. For this reason, I argue for closer collaboration between libraries and researchers.

References

BECKER, Lidia, Julia Kuhn, Christina Ossenkop, Anja Overbeck, Claudia Polzin-Haumann, and Elton Prifti, eds. 2020. *Fachbewusstsein der Romanistik: Romanistisches Kolloquium XXXII*. Tübinger Beiträge zur Linguistik 578. Tübingen: Narr Francke Attempto. <http://elibrary.narr.digital/book/99.125005/9783823394181>.

BIESELIN, Tanja-Barbara. 2005. 'Im Kampf gegen Etat-Kürzungen, Schließungen und morsches Image. Guerilla-Marketing für Bibliotheken' 29 (3): 361–75. <https://doi.org/10.1515/BFUP.2005.361>.

BURR, Isolde. 2008. 'Romanische Sprachen in internationalen Organisationen'. In *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen; Histoire linguistique de la Romania*, edited by Gerhard Ernst, Gerold Ungeheuer, and Armin Burkhardt, 3:3339–54. Handbücher zur Sprach- und Kommunikationswissenschaft; Handbooks of linguistics and communication science 23. Berlin: De Gruyter. <http://www.degruyter.com/doi/book/10.1515/9783110211412.3>.

CHOWDHURY, G. G., Paul F. Burton, David McMenemy, and Alan Poulter. 2008. *Librarianship: An Introduction*. London: Facet Publishing.

CONSTANTINESCU, Ioan. 2002. 'Deutschprachige Romanistik – Eine Wissenschaft Mit Zukunft'. In *Deutschsprachige Romanistik - Für Wen?*, edited by Maria Lieber and Harald Wentzlaff-Eggebert, 41–46. Heidelberg: Synchron, Wiss.-Verl. der Autoren.

DOMBROWSKI, Quinn, Tassie Gniady, and David Kloster. 2019. 'Introduction to Jupyter Notebooks'. *The Programming Historian* 8. <https://programminghistorian.org/en/lessons/jupyter-notebooks>.

EHRLICHER, Hanno, and Jörg Lehmann. 2021. 'La recolección de datos como laboratorio epistemológico. Algunas reflexiones acerca del entorno virtual de investigación Revistas Culturales 2.0'. *Signa: Revista de la Asociación Española de Semiótica* 30 (0): 59–81. <https://doi.org/10.5944/signa.vol30.2021.29298>.

ERNST, Michael. 2021. 'Ein Trend und seine Folgen'. *Verfassungsblog* (blog). 17 June 2021. <https://verfassungsblog.de/ein-trend-und-seine-folgen/>.

EVANS, Michael S. 2014. 'A Computational Approach to Qualitative Analysis in Large Textual Datasets'. *PLoS ONE* 9 (2). <https://doi.org/10.1371/journal.pone.0087908>.

GANTERT, Klaus. 2016. *Bibliothekarisches Grundwissen*. *Bibliothekarisches*

Grundwissen. Berlin, Boston: De Gruyter Saur. <https://www.degruyter.com/view/title/302969>.

GITTEL, Benjamin. 2021. 'An Institutional Perspective on Genres: Generic Subtitles in German Literature from 1500-2020'. *Journal of Cultural Analytics*, April. <https://doi.org/10.22148/001c.22086>.

GONZÁLEZ, Juana María. 2021. 'Análisis cuantitativo de la revista Índice Literario (1932-1936)'. *Artnodes* 27. <https://doi.org/10.7238/a.v0i27.374373>.

GUMBRECHT, Hans Ulrich. 2002. *Vom Leben Und Sterben Der Großen Romanisten: Karl Vossler, Ernst Robert Curtius, Leo Spitzer, Erich Auerbach, Werner Krauss*. Edition Akzente. München: Carl Hanser Verlag.

HAARMANN, Harald. 2008. 'Romanische Sprachen als Publikationssprachen der Wissenschaft: 19. und 20. Jahrhundert'. In *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen; Histoire linguistique de la Romania*, edited by Gerhard Ernst, Gerold Ungeheuer, and Armin Burkhardt, 3:3359–70. Handbücher zur Sprach- und Kommunikationswissenschaft; Handbooks of linguistics and communication science 23. Berlin: De Gruyter. <http://www.degruyter.com/doi/book/10.1515/9783110211412.3>.

HENNY-KRAHMER, Ulrike. 2017. 'Bib-ACMé: Bibliografía digital de novelas argentinas, cubanas y mexicanas (1810-1930).' In *Sociedades, políticas, saberes.*, edited by Nuria Rodríguez Ortega, 99–104. Málaga: Universidad de Málaga. <http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf>.

HERRMANN, J. Berenike, Giulia Grisot, Susanne Gubser, and Elias Kreyenbühl. 2021. 'Ein Großer Berg Daten? Zur Bibliothekswissenschaftlichen Dimension Des Korpusliteraturwissenschaftlichen Digital Humanities-Projekts „High Mountains – Deutschschweizer Erzählliteratur 1880–1930"'. *027.7 Zeitschrift Für Bibliothekskultur / Journal for Library Culture* 8 (1). <https://doi.org/10.21428/1bfadeb6.6e2feff6>.

HOLTUS, Günter, and Fernando Sánchez Miret. 2008. *'Romanitas', Filología Románica, Romanística*. Beihefte Zur Zeitschrift Für Romanische Philologie. - Berlin: De Gruyter, 1905- ; ZDB-ID: 200077-5 347. Tübingen: Niemeyer.

JANNIDIS, Fotis, Leonard Konle, & Peter Leinen. 2019. 'Makroanalytische Untersuchung von Heftromanen'. In *Digital Humanities: Multimedial & Multimodal*, 167–73. Mainz-Frankfurt: Dhd. <https://zenodo.org/record/2600812#.XLg1bUNS9hE>.

KALKHOFF, Alexander M. 2010. *Romanische Philologie Im 19. und Frühen 20. Jahrhundert: Institutionsgeschichtliche Perspektiven*. Romanica Monacensia. - Tübingen : Narr Francke Attempto, 2014- ; ZDB-ID: 3020712-5 78. Tübingen: Narr. <http://elibrary.narr.digital/book/99.125005/9783823375043>.

KRAMER, Johannes. 2002. 'Deutsch Als Publikationssprache Und Vielsprachige Romanistik — Ein Ärgernis in Der Internationalen Wissenschaftslandschaft?' In *Deutschsprachige Romanistik - Für Wen?*, edited by Maria Lieber and Harald Wentzlaff-Eggebert, 13–25. Heidelberg: Synchron, Wiss.-Verl. der Autoren.

KRAMER, Johannes. 2008. 'Romanische Sprachen als Publikationssprachen der Wissenschaft bis zum 18. Jahrhundert'. In *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen; Histoire linguistique de la Romania*, edited by Gerhard Ernst, Gerold Ungeheuer, and Armin Burkhardt, 3:3354–59. Handbücher zur Sprach- und Kommunikationswissenschaft; Handbooks of linguistics and communication science 23. Berlin: De Gruyter. <http://www.degruyter.com/doi/book/10.1515/9783110211412.3>.

KRAMER, Johannes. 200. 'Selbstdarstellungen der Romanistik während der Gründungsphase, um 1900 und nach 1988'. In *Fachbewusstsein der Romanistik: Romanistisches Kolloquium XXXII*, edited by Lidia Becker, Julia Kuhn, Christina Ossenkop, Anja Overbeck, Claudia Polzin-Haumann, and Elton Prifti, 1. Tübinger Beiträge zur Linguistik 578. Tübingen: Narr Francke Attempto. <http://elibrary.narr.digital/book/99.125005/9783823394181>.

KREFELD, Thomas. 2020. 'FAIRness weist den Weg – von der Romanischen Philologie in die Digital Romance Humanities'. In *Fachbewusstsein der Romanistik: Romanistisches Kolloquium XXXII*, edited by Lidia Becker, Julia Kuhn, Christina Ossenkop, Anja Overbeck, Claudia Polzin-Haumann, and Elton Prifti, 291–310. Tübinger Beiträge zur Linguistik 578. Tübingen: Narr Francke Attempto. <http://elibrary.narr.digital/book/99.125005/9783823394181>.

KREMNITZ, Georg. 2016. *Geschichte Der Romanischen Sprachwissenschaft: Unter Besonderer Berücksichtigung Der Entwicklung Der Zahl Der Romanischen Sprachen*. Bachelor Master Studies. - Wien : Praesens, 2014- ; ZDB-ID: 2806920-1 8. Wien: Praesens Verlag.

LIEB, Claudia, and Christoph Strosetzki, eds. 2013. *Philologie Als Literatur- Und Rechtswissenschaft: Germanistik Und Romanistik 1730 - 1870*. Euphorion. Beihefte Zum Euphorion. - Heidelberg : Winter, 1964- ; ZDB-ID: 503579-X 67. Heidelberg: Winter.

LIEBER, Maria, and Harald Wentzlaff-Eggebert, eds. 2002. *Deutschsprachige Romanistik - Für Wen?* Heidelberg: Synchron, Wiss.-Verl. der Autoren.

MARYL, Maciej, and Piotr Wciślik. 2016. 'Remediations of Polish Literary Bibliography: Towards a Lossless and Sustainable Retro-Conversion Model for Bibliographical Data'. In *Digital Identities: the Past and the Future*. <https://dh-abstracts.library.cmu.edu/works/2767>.

MONJOUR, Alf. 2020. 'Romanistik nach Bologna? Zum Nachdenken über zukünftige Positionen der romanistischen Sprach- und Kulturwissenschaften'. In *Fachbewusstsein der Romanistik: Romanistisches Kolloquium XXXII*, edited by Lidia Becker, Julia Kuhn, Christina Ossenkop, Anja Overbeck, Claudia Polzin-Haumann, and Elton Prifti, 195–203. Tübinger Beiträge zur Linguistik 578. Tübingen: Narr Francke Attempto. <http://elibrary.narr.digital/book/99.125005/9783823394181>.

MÜLLER, Andreas C., and Sarah Guido. 2016. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. Beijing, Boston: O'Reilly.

NITSCHACK, Horst. 2002. 'Deutschsprachige Romanistik — Für Wen? Dreifache Dialog'. In *Deutschsprachige Romanistik - Für Wen?*, edited by Maria Lieber and Harald Wentzlaff-Eggebert, 7–11. Heidelberg: Synchron, Wiss.-Verl. der Autoren.

RICHERT, Gertrud. 1913. *Die Anfänge Der Romanischen Philologie Und Die Deutsche Romantik*.

SCHROTT, Angela. 2003. 'Romanistische Sprachgeschichtsforschung: Zeitschriften'. In *Romanische Sprachgeschichte: ein internationales Handbuch zur Geschichte der romanischen Sprachen; Histoire linguistique de la Romania*, edited by Gerhard Ernst, Gerold Ungeheuer, and Armin Burkhardt, 1:421–27. Handbücher zur Sprach- und Kommunikationswissenschaft; Handbooks of linguistics and communication science 23. Berlin: De Gruyter. <http://www.degruyter.com/doi/book/10.1515/9783110146943.1>.

TOLONEN, Mikko, Mark J. Hill, Ali Ijaz, Ville Vaara, and Leo Lahti. 2020. 'Data-Driven Analysis of Canonical Works in Early Modern Britain.' In *15th*

Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2020, Ottawa, Canada, July 20-25, 2020, Conference Abstracts. <https://dh2020.adho.org/wp-content/uploads/2020/07/555_DatadrivenanalysisofcanonicalworksinearlymodernBritain.html>.

TOLONEN, Mikko, Jani Marjanen, Hege Roivainen, and Leo Lahti. 2019. 'Scaling Up Bibliographic Data Science.' In *Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8, 2019.*, 450–56. <http://ceur-ws.org/Vol-2364/41_paper.pdf>.

VAARA, Ville, Ali Ijaz, Iiro Tiihonen, Antti Kanner, Tanja Säily, and Leo Lahti. 2019. 'The Emerging Paradigm of Bibliographic Data Science'. In T*he Index of Digital Humanities Conferecnces*. <https://dh-abstracts.library.cmu.edu/works/9931>.

VANDERPLAS, Jake. 2016. *Python Data Science Handbook: Essential Tools for Working with Data*. First edition. Beijing, Boston: O'Reilly.

VOß, Jakob. 2020. *Einführung in die Verarbeitung von PICA-Daten*. Göttingen: GBV (VZG). <https://pro4bib.github.io/pica/>.

WANDRUSZKA, Mario. 1988. 'Deutsche Romanistik: Kritische Bilanz Und Perspektive'. In *Ein 'Unmögliches Fach': Bilanz Und Perspektiven Der Romanistik*, edited by Fritz Nies and Reinhold R. Grimm, 213. Tübingen: Narr.

WIESENMÜLLER, Heidrun, and Silke Horny. 2017. *Basiswissen RDA: Eine Einführung für deutschsprachige Anwender*. *Basiswissen RDA*. De Gruyter Saur.

Wolf, Johanna. 2012. *Kontinuität Und Wandel Der Philologien: Textarchäologische Studien Zur Entstehung Der Romanischen Philologie Im 19. Jahrhundert*. Romanica Monacensia. - Tübingen : Narr, 1968- ; ZDB-ID: 404830-1 80. Tübingen: Narr Francke Attempto.

Abstract

What have been the main trends in Romance Studies in the last decades? What can be expected for the next decade? These are the two main research questions of this article. To answer them, a large dataset of over one million publications of research in Romance Studies has been extracted from German library catalogs. This dataset is analyzed through descriptive statistics and linear regression in order to predict the development in future years. Several fields of the respective catalogs are analyzed, such as the language and place of publication, publishers, e-book vs. printed versions, price and subjects.

Zusammenfassung

Was waren die wichtigsten Trends in der Romanistik in den letzten Jahrzehnten? Was ist für das nächste Jahrzehnt zu erwarten? Dies sind die beiden Hauptforschungsfragen des vorliegenden Artikels. Zur Beantwortung dieser Fragen wurde ein großer Datensatz von über einer Million romanistischer Forschungspublikationen aus deutschen Bibliothekskatalogen extrahiert. Dieser Datensatz wird mittels deskriptiver Statistik und linearer Regression analysiert, um die Entwicklung in den kommenden Jahren vorherzusagen. Dabei werden verschiedene Felder der jeweiligen Kataloge analysiert, wie z.B. Sprache und Ort der Veröffentlichung, Verlage, E-Book und gedruckte Version, Preis und Themen.