The Onomastic Repertoire of the *Roman d'Alexandre* (ORNARE). Designing an Integrated Digital Onomastic Tool for Medieval French Romance

Marta Milazzo[1], Giorgio Maria Di Nunzio[2]
[1]Department of Linguistics and Literary Studies, University of Padua
Via Vendramini, 13, 35137 Padova, Italy
[2]Department of Information Engineering, University of Padaua
Via Gradenigo 6b, 35131 Padova, Italy
marta.milazzo@phd.unipd.it, giorgiomaria.dinunzio@unipd.it

Abstract

The paper reports on the first results of the design and implementation of a new digital tool for romance philology: the digital Onomastic Repertoire for the medieval French romance (12th-15th centuries). This tool, projected with a modular and integrable architecture, was implemented from a selection of romances, the corpus of the Medieval French Roman d'Alexandre. After introducing the peculiarities of the onomastic system in the Middle Ages (and, more generally, the peculiarities of medieval literary texts), the paper describes 1) the methodological challenges faced in the preparatory work, illustrates and comments on the first results achieved and 2) the design and implementation of the first integrated system for the interactive creation of the Onomastic Repertoire of the romaN d'AlexandRE (ORNARE), and 3) the current research output in terms of both a digital edition and the digital onomastic index of the corpus.

Keywords: onomastics, digital philology, annotation tools

1. Introduction

According to the Latin *adagio nomen omen*, the proper name is a sign, a destiny: to call oneself is to be, in name and in fact. Even more than in antiquity, it is the centuries conventionally included in the definition of the Middle Ages that make the Latin expression a true commandment. If medieval aesthetic and intellectual praxis is first and foremost unveiling – "la veritade ascosa sotto bella menzogna"[1] (Dante) – one understands then how the name is the depository of the deepest truth. This is even more evident when we look at literary production, where names, far from being accidental, possess symbolic meanings and rich allusions. Add to this the fact that the Middle Ages are affected by drastic onomastic phenomena on a historical level: leaving aside the frequent changes of proper name, entirely regular in a system of 'onomastic plurality', we will at least recall that "the sophisticated Roman practice of double/triple naming had been replaced by German single naming practice which then in turn, after four to five centuries of dominance, gave way to the two-element system (personal plus family name) as it has evolved into our own day" (Geary, 2002). For all these reasons, onomastics – the study of the forms, functions and meaning of proper names –, in particular medieval literary onomastics, constitutes an important interdisciplinary field of study which deeply involves romance philology and digital humanities (DH).

Among the most important tools for Romance philology - the study of medieval Neo-Latin languages and literature - are the onomastic repertories, reasoned dictionaries of proper names taken from literary corpora: for example, in the Old French domain, the *Table des noms propres avec toutes leurs variantes figurant dans les romans du Moyen Age* by Louis-Fernand Flutre (Flutre 1962) and the *Répertoire des noms propres de personnes et de lieux cités dans les chansons de geste françaises et les œuvres étrangères dérivées* by André Moisan (Moisan, 1986). These are very useful tools, albeit in need of updating: Flutre published his table in 1962, Moisan in 1986. Since then, no updating work has been produced that considers the innovations offered by digital humanities. Moreover, a substantial increase in new texts makes it necessary to deal with more extensive corpora than those considered so far. While two digital medieval onomastic repertories have been created (*Repertorium van Eigennamen in Middelnederlandse Literaire teksten*[2], dedicated to Middle Dutch and *Diccionario de nombres del ciclo amadisiano*[3], dedicated to a corpus of medieval Spanish romances), they allow online consultation, but overall constitute traditional tools.

These are the premises that justify the present research, the result of which aims to 1) design and implement the first integrated system for the analysis of Medieval French Romance and 2) create a new digital Onomastic Repertoire for the Old-French romance. The objectives we set ourselves are instead broader and more complex. Some further clarifications are necessary: in the following (Sect. 2), we will specify the methodological choices that have been made and

[1] "The truth hidden under a beautiful lie".
[2] https://bouwstoffen.kantl.be/remlt/
[3] https://dinam.unizar.es

LREC-COLING 2024, pages 15982–15987

the specific case study, while in Section 3, we will present the design and implementation of ORNARE with the current research output. Finally (Sect. 4), aware of the modular nature of our ongoing project, we will give our final remarks and present future considerations.

2. Methodology

In this section, we discuss a step-by-step methodology for the design and development of a web application to facilitate the integration of the different elements that are required for the annotation and in-depth linguistic analysis of an onomastic index of the repertoire of the Medieval French Romance and, in particular, on the corpus of the Medieval French *Roman d'Alexandre* (see section 2.1). In the literature there are only a few recent examples of DH approaches dedicated to the study of onomastics. (Dannéls and Brodén, 2020) discuss some of the challenges and opportunities this field brings as well as the importance of a language technology infrastructure that supports interdisciplinary research. (Waldispühl et al, 2020) is the closest research project to our proposal. The authors present NordiCon, a database where formally interpreted and richly interlinked onomastic data is combined with information on material properties and digitized versions of the medieval manuscripts from which the data originate.

Our approach extends these ideas into one integrated environment that allows linguists and philologists to work in a collaborative environment from the scans of the original editions to the analysis of the text. The methodology proposed for a reproducible and replicable approach is the following:

1. OCR and Data Extraction: Acquire digital copies of the Old French *Roman d'Alexandre* (RdA) and use OCR tools (like Tesseract[4]) to convert scanned pages into machine-readable text.
2. Cleaning and Pre-processing: Correct OCR errors and clean the text by removing unnecessary characters, line breaks, and formatting issues.
3. Data Structuring and Normalization: Organize the text into a structured format suitable for analysis. Identify and standardize variants of names, including misspellings and abbreviations. Create relationships or linkages between names and their contextual information within the text.
4. Data Annotation: allow experts in Old French and onomastics to manually verify and annotate the onomastic data. Create annotation guidelines and standards for consistency. Validate the quality and accuracy of the annotations through iterative reviews.
5. Named Entity Recognition (NER): Implement Natural Language Processing (NLP) techniques to recognize and extract named entities (onomastic data) from the text.
6. Data Curation and Harmonization: Resolve any mismatch between manual annotation and automated extraction and manual annotation. Standardize the representation of names and associated metadata.
7. Collaboration and Data Sharing: Implement collaboration features, allowing multiple researchers to work on the onomastic repertoire simultaneously. Enable data sharing mechanisms and permission controls.

This comprehensive methodology covers (parts of) the entire process of creating a digital tool for an accurate onomastic repertoire, from OCR and data extraction to a systematic and harmonized approach to building the repertoire. For space reasons, in the following sections we will focus on some of the steps and leave some part of the implementation (for example the OCR process and the relational database to store names and annotations).

2.1 Case Study: Roman d'Alexandre

This digital Onomastic Repertoire deals with 1) all proper names of persons ('anthroponyms') 2) attested in romances 3) written in Old French 4) from 12th to 15th century. The main methodological problems relate to point 1) and 2). First of all, it is not clear what is to be understood by 'anthroponym', since in medieval production many characters may be called by combinations of common names (e.g., *Riche Chevalier*) that differ from anthroponyms, but which perform the same functions as the latter: it is therefore a question of deciding, for each case, what is considered a proper name and what is not. Moreover, the genre of medieval texts is by no means peaceful: defining what 'romances' are in the Middle Ages is an open problem. Using the most up-to-date tools (Frappier and Grimm, 1978-1984; Woledge 1954, 1975; Colombo et al., 2014; DeafBibl) we have collected our corpus, consisting of 235 critical editions of romances. All editions were then scanned and finally subjected to OCR. At the end of these complex steps, we obtained a corpus consisting of approximately 120,000 scans.

In the following sections, we present the current work on the design and implementation of the system for the steps 1-4 and partially step 7 for sharing the dataset and the current produced in this first phase of gold standard annotation of a subset of the whole corpus which consists of the

[4] https://tesseract-ocr.github.io

so-called RdA, a vast and ancient textual galaxy in verse where numerous materials (classical and oriental) relating to Alexander the Great are channeled. The breadth of cultural references in the romance has direct implications in the onomastic system, and vice versa. Medieval production is articulated on stereotypes, clichés particularly evident in the presentation of characters, cultural coordinates, geography, etc. Very often, a set of names is employed for a certain type of character: e.g., what are enemies or foreigners called? What name is chosen for a Greek or Indian knight? Onomastics thus offers evidence of cultural modalization in medieval West. Despite its extraordinary importance, there is only one edition of the RdA (Armstrong et al., 1937-1976), which lacks the *Index nominum* (the index of all the proper names). For these reasons, we have turned to the Alexandre as a highly significant sample to show the initial results and future potential of our work.

3. ORNARE System

In this section, we present the design and implementation of the first system and the creation of an Onomastic Repertoire of the RdA (ORNARE).

3.1 Dataset

The corpus of the RdA consists of about 1500 pages of editions and text variants, containing 33.119 verses plus 180 pages of variants still under redactions (the number will be updated as soon as the manual work will be completed).

3.2 Implementation

The application has been designed and implemented extending the R Shiny framework[5] for web interactivity proposed by (Milazzo and Di Nunzio, 2023), it includes the datatable (DT) package[6] to review the annotation, and the textplot package[7] for text visualization and analysis. ORNARE enables linguists to upload (or directly write) and search and annotate texts with onomastic information. The current interface of the ORNARE system is shown in Figure 1. The user can 1) select the edition to analyze, then 2) search for a character (for example, “the King of …”), and/or search for a “normalized” name (i.e., “Alixandre”), or a specific spelling (i.e., “Al’x”). The user can “update” 3) the onomastic information of all the selected verses that match the search in the main panel (functionality not visible in this screenshot). An additional filtering on the result can be used 4), and 5) the interaction. (edit and modify) directly the text of the verses when necessary.

This integrated Web environment is also designed to allow for on-the-fly adjustments, enabling users to fine-tune annotation results or modify display preferences. This web application would serve as a valuable tool for linguistic analysis and facilitate research on language contact and change, particularly in the context of the Old French RdA. The source code of the Web application is constantly updated and made available to the research community.[8]

3.3 Output

The ongoing work on step 4 (Data Annotation) of the methodology has produced the identification of 2,219 spelling of names. The exact number of characters and normalized names is still under evaluation given the fact that some of the names need careful philological research to be ascertained to discriminate between two characters with the same name.

Publishing linguistic data in accordance with the OntoLex-Lemon paradigm represents a crucial step in ensuring the interoperability and semantic richness of the data within the broader linguistic and semantic web community. Through this approach, the onomastic information can be seamlessly integrated into the LOD (Linked Open Data) cloud, enabling researchers and applications to access and utilize the data in a structured and consistent manner. In this sense, but also the TEI (Text Encoding Initiative) XML format to maintain the structure of the opera, as suggested by (Bohbot et al., 2018). The serialization in these two formats will be extremely useful from many viewpoints: the digital edition in TEI of the corpus would promote the collaborative research of DH researchers to create digital tools for the exploration and analysis of these text and ensure that texts are standardized and can be exchanged readily with other researchers. The Linked Data representation of these data (names of people, characters that appear in verses and toponymy in different branches) will provide a viable means of producing data that is machine-readable archival data and suited to automatic reasoning.

4. Preliminary Analysis

After OCRing the text of the critical edition of the *Alexandre*, all words (i.e., *Alixandres*, *Filotés*) and combinations of two or more words (i.e., *Dans Clins*, *Daires Persan*) beginning with a capital letter were automatically isolated, since by graphic convention proper names are rendered with a capital letter. While the extraction of such words and word combinations is automatic, the identification of the anthroponym remains manual. In fact, it must be taken into account that we are only interested in anthroponyms, so we have to eliminate proper names of another nature such as toponyms (*Gadres*, 'Gaza'), ethnonyms (*Grigois*, 'Greeks'), horse names (*Bucifal*), etc. Moreover, since these texts are in verse, every first word with

[5] https://shiny.posit.co/
[6] https://github.com/rstudio/DT
[7] https://github.com/bnosac/textplot
[8] https://github.com/gmdn/LREC-COLING-2024

which the verse begins is written with a capital letter: the work of selection is thus considerable.

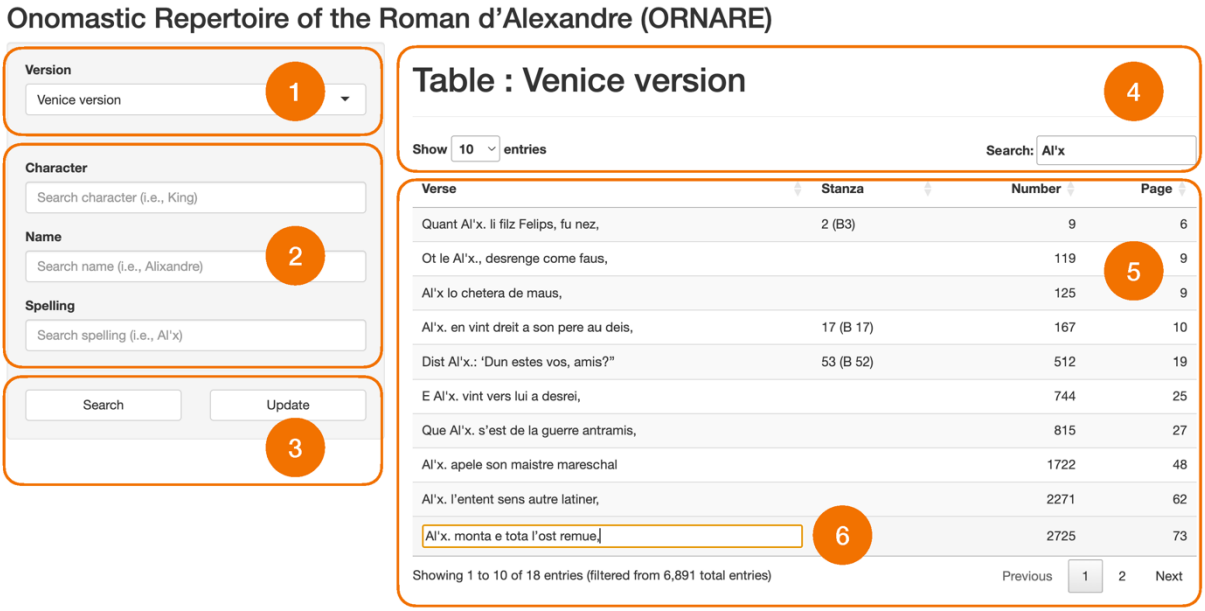

Figure 1 Screenshot of ORNARE system

To these significant problems one must add that the proper name in the Middle Ages is subject to great graphic variability: the same name can be declined, written with different spellings, abbreviations, forms. For instance, *Al'x.*, *Alixandres* and *Alis* are all variants of the protagonist's name, *Alixandre*, but *Alis* is also the hypothetical name of the son he would like to have by his queen. Again, *Aristés* and *Aridés* are two names of two different characters, but the name *Aristés* can also be spelled *Aridés*. Added to this is the difficulty, in OCR extraction, of languages for which there are very few attestations and where, indeed, irregular forms should not be normalized, but are often indicators of important linguistic information. As a result, the description of characters always requires careful reading and may necessitate, in special cases, even days of philological work.

Obviously, it is the philologist's task to group all the different spellings of a name into the same form, or conversely to distinguish the very frequent cases of homonymy and similar spellings. At the end of this work, one obtains a reasoned repertoire, where each character listed in the onomastic index corresponds to a nominal entry, including all the spellings and forms assumed by the name in the course of the romance.

4.1 Additional Requirements

The onomastic work is, however, only one part of the digital Repertoire. It will in fact allow two types of searches, depending on the perspective one chooses to favor. On the one hand, it will be possible to search for a specific spelling of a name: for example, by searching for the form *Al'.x* I will be directed to the nominal entry where this form is attested (in this case, Alexandre the Great). Each nominal entry is associated with a character, represented through instant descriptions and labels. It is then possible to search by character type, according to the medieval perspective. For instance, by querying the category 'pagan' or 'Greek', one is able to see which names Western *romanciers* used to describe pagan enemies or classical antiquity.

On the other hand, it will also be possible to search not only for names, but also for romances. Every onomastic entry is in fact provided with a reference to the text where it appears. They can be very many in the case of a literature that, like medieval literature, is highly intertextual. For instance, Alexander figures in dozens and dozens of medieval romances, as a character but also as an ambivalent *exemplum*. Through the Repertoire, it will be easy to check in which texts the figure of Alexander is evoked whether he is alluded to as a positive or negative example. The romances are described through synthetic filters (date/matter/form, etc.); a cross-search therefore makes it possible to investigate sections of the corpus (e.g.: searching for names used in Arthurian romances in prose written before the 13th century for 'pagan' characters; search for all names used for female characters in romances of ancient matter in verse).

In an initial experimental test, we tried to filter out semi-automatically common words that are capitalized when they are the first word of a verse. The user is initially prompted with all the capitalized words, while s/he annotates some names and skip some other words, the system automatically hides the skipped ones. For this work, however, we decided to resort to manual skipping (which is very fast when words are ordered alphabetically) since we wanted not to lose any particular name with some common

words as part of the name. The more data we collect about names, the more accurate the filtering system that we are currently testing becomes.

5. Conclusions and Future Work

At the end of lengthy preparatory work, which has involved significant efforts in terms of both methodological choices and digital coordination, our Onomastic Repertoire, with its dual search function (by names vs. by romances) makes it possible to fully appreciate the possibilities offered by onomastic work declined according to a digital perspective. The result is even more remarkable given the specificity of the task performed, for which, we would stress, there are no digital precedents. Indeed, digital onomastic tools made on contemporary corpora do not have to deal with the peculiarities of medieval texts, where the language is devoid of normalizations, the same name takes on a great plurality of spellings and forms, and the frequency of identical names (homonymy) or almost overlapping names is very high.

The Repertoire, by its very nature modular and integrable, will be able to greatly broaden the spectrum of its usefulness as new romances and names are added. Through it, it will be possible to investigate names, and thus representations, of space, time and geography. It will also be possible to cut out particular intersections using the appropriate filters. The technology thus makes it possible to recover data lost in the detail of many texts and to reach at overviews that would be, manually, impossible to obtain. The rigor of the lemmatization of names, which obviously remains philological, is strengthened by the management of a quantity of data that individual scholars cannot handle manually. The resulting integration can only benefit both philologists and computer engineers, each for their own reasons.

The integration of this peculiar linguistic data into the CLARIN (Common Language Resources and Technology Infrastructure) framework represents a strategic step for ensuring that the linguistic dataset, annotated in accordance with the OntoLex-Lemon paradigm, becomes an integral part of a broader research infrastructure. Furthermore, the integration into CLARIN ensures that the data is not only well preserved but also benefits from the sustainability and long-term archiving mechanisms that CLARIN provides, making it a valuable and enduring linguistic asset for the scientific community.

Due to space constraints, some parts of the implementation were intentionally left out and will be presented in an extended version of the paper. Namely, the OCR process and the semi-automatic correction of the produced OCRed text, and the design and implementation of the relational database to store the data and the annotation provided by the experts. In particular, since the application is based on the R programming language, the whole OCR pipeline was designed and implemented with R packages. In particular: i) pdftools to render each scanned page of the digitized text; ii) png and magick to work with the bitmap version on the scan in order to get the geometry of the page and the spaces of the different areas of interest; iii) tesseract to produce the OCR version of the page; iv) tidyverse to clean, manage, and extract the text of interest. There are some corpora and works that have been carried out in the context of Old French (see, for example, (Camps et al., 2021; Gabay et al., 2020; Garcia-Fernandez et al., LREC 2014). Nevertheless, we decided in this first part of the project to leave the raw OCR text without any modification (see (Del Fante and Di Nunzio, 2021) in order not to introduce any additional (unwanted) correction that may have led to wrong or biased philological reasoning. We intend to introduce the part of the (semi)automatic OCR correction in the second part of the project after the *index nominum* has been completed.

The source code of the Web application is constantly updated and made available to the research community as well as the data produced.[9]

6. Bibliographical References

Armstrong et al. (Eds.) (1937-1976). The Medieval French “Roman d’Alexandre”, 7 voll. Princeton UP, Princeton.

Balode, L. 2019, 'Latvian Given Names of Lithuanian Origin, International Conference “Personal Names and Cultural Reconstructions” Helsinki, 2019, 2-3.', Personal Names and Cultural Reconstructions, Helsinki , Finland , 21/08/2019 - 23/08/2019 pp. 2-3

Bohbot, H., Frontini, F., Luxardo, G., Khemakhem, M., and Romary, L. (2018, May). Presenting the Nénufar Project: A diachronic digital edition of the Petit Larousse Illustré. In GLOBALEX 2018-Globalex workshop at LREC2018 (pp. 1-6).

Camps, Jean-Baptiste, Thibault Clérice, Frédéric Duval, Lucence Ing, Naomi Kanaoka, and Ariane Pinche. 2021. “Corpus and Models for Lemmatisation and POS-Tagging of Old French.” arXiv. https://doi.org/10.48550/arXiv.2109.11442.

Colombo Timelli, M., Ferrari, B., Schoysman, A., and Suard, F. (Eds.) (2014). Nouveau Répertoire des mises en proses (XIV-XVI siècles). Garnier, Paris.

Dannélls, D., and Brodén, D. (2020). Building a Language Technology Infrastructure for Digital Humanities: Challenges, Opportunities and Progress. Proceedings of the Twin Talks 2 and 3 Workshops at DHN 2020 and DH 2020

[9] https://github.com/gmdn/LREC-COLING-2024

Ottawa Canada and Riga Latvia, July 23 and October 20, 2020 / Edited by Steven Krauwer, Darja Fišer. CEUR-WS.org.
Dante, Convivio, II, 2.
Del Fante, D., and Di Nunzio, G.M.: Correzione dell'OCR per Corpus-assisted Discourse Studies: un caso di studio su vecchi quotidiani. Umanistica Digitale 5(11), 99–124 (2021). https://doi.org/10.6092/issn.2532-8816/13689
Dictionnaire Étimologique de l'Ancien Français, Bibliographie en ligne (DEAFBiblEl), Möhren F. (ed.). https://alma.hadw-bw.de/deafbibl/fr/
Diccionario de nombres del ciclo amadisiano (DINAM). https://dinam.unizar.es
Frappier, J., and Grimm, R.R. (Eds.) (1978-1984): Le roman jusqu'à la fin du XIIIe siècle, 2 voll., in J. Frappier et al. (Eds.), Grundriss der Romanischen Literaturen des Mittelalters, Heidelberg, C. Winter.
Flutre, L.-F. (1962). Table des noms propres avec toutes leurs variantes figurant dans les romans du Moyen Age, C.E.S.C.M, Poitiers.
Gabay, Simon, Alexandre Bartz, and Yohann Deguin. 2020. "CORPUS17: A Philological French Corpus for 17th Century." In Proceedings of the 2nd International Conference on Digital Tools & Uses Congress, 1–7. DTUC '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3423603.3424002.
Garcia-Fernandez, Anne, Anne-Laure Ligozat, and Anne Vilnat. 2014. "Construction and Annotation of a French Folkstale Corpus." In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, 2430–35. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/1070_Paper.pdf.
Geary, P. (2002). Foreword, in Beech, G.T. et al. (Eds.), Personal Names Studies of Medieval Europe: Social Identity and Familial Structures. Western Michigan University, Kalamazoo, p. ix.
Milazzo, M., and Di Nunzio, G.M. (2023). The First Tile for the Digital Onomastic Repertoire of the French Medieval Romance: Problems and Perspectives. In: Alonso, O., Cousijn, H., Silvello, G., Marrero, M., Teixeira Lopes, C., Marchesin, S. (eds) Linking Theory and Practice of Digital Libraries. TPDL 2023. Lecture Notes in Computer Science, vol 14241. Springer, Cham. https://doi.org/10.1007/978-3-031-43849-3_29
Moisan, A. (1986). Répertoire des noms propres de personnes et de lieux cités dans les chansons de geste françaises et les œuvres étrangères dérivées, Droz, Genève.
Repertorium van Eigennamen in Middelnederlandse Literaire teksten (REMLT). http://bou wstoffen.kantl.be/remlt/)
Waldispühl, M., Dannells, D., and Borin, L. (2020, May). Material Philology Meets Digital Onomastic Lexicography: The NordiCon Database of Medieval Nordic Personal Names in Continental Sources. Proceedings of the Twelfth Language Resources and Evaluation Conference, 860–867. Retrieved from https://aclanthology.org/2020.lrec- 1.108
Woledge, B. (1954, 1975). Bibliographie des romans et des nouvelles françaises antérieurs à 1500, Droz, Genève (1954); Supplement 1954–1973, Droz, Genève (1975)