Roxana Patras, Carolin Odebrecht,
Ioana Galleron, Rosario Arias, J. Berenike Herrmann, Cvetana Krstev, Katja Mihurko Poniž, Dmytro Yesypenko

Thresholds to the "Great Unread": Titling Practices in Eleven ELTeC Collections

Résumé

Le but principal de cet article est de décrire et, jusqu'à un certain point, de comprendre les pratiques d'intitulation dans la littérature, à travers une exploration d'un corpus multilingue de romans européen publiés entre 1840 et 1920. L'étude est basée sur l'analyse de 11 des 16 sous-collections de romans en cours de préparation dans l'action COST 16204 « Distant reading for European Literary History ». Ces collections couvrent les domaines allemand, anglais, espagnol, français, italien, polonais, portugais, roumain, serbe, slovène et ukrainien. Nous proposons une analyse de la présence de personnes, de lieux et d'indicateurs génériques dans le titre, et faisons quelques observations à propos de la « syntaxe » de ces entités.

Abstract

The main aim of the paper is to describe and, to a certain extent, to understand, *titling practices* in literary discourse through the exploration of a multilingual literary corpus comprising European novels published between 1840 and 1920. The study is based on the analysis of 11 out of the 16 sub-collections of novels in preparation within the COST Action 16204 "Distant reading for European Literary History", namely the English, French, German, Italian, Polish, Portuguese, Romanian, Serbian, Slovenian, Spanish, and Ukrainian sub-collections. We focus on an analysis of persons, places

and genre entities in titles, and observe some regularities involving the "syntax" of these various entities in titles.[1]

[1] The research described in this paper was conducted in the context of the COST Action "Distant Reading for European Literary History" (CA16204 – "Distant-Reading"). Find more about the Action at: http://www.distant-reading.net. COST is funded by the Horizon 2020 Framework Programme of the EU. The collection encoding and documentation are done by Working Group 1 (https://www.distant-reading.net/wg-1/). The creation, and correction, of ELTeC is an extensive team effort. Therefore, we would like to thank all contributors to ELTeC.

Introduction

In the last fifty years, literary history has undergone important changes, based on the rediscovery of the so-called "second" and even "third hand" authors and works. In the meantime, scholars worked towards a better understanding of the mutual influences between different cultural areas. Most recently, the rise of the digital humanities has pushed further forward both movements, opening up the way for direct interrogation of the "great unread".[2]

Much remains, however, a desideratum, both for acquiring works in machine-readable formats and for building methods of analysis and interpretation in a digital paradigm. This paper aims at contributing to this second goal, through focusing on a small, yet notoriously decisive, element of the novel, namely its title and subtitle. We will try and show what an annotation-based exploratory study offers to the literary historian, while suggesting some directions to be taken so as to uncover new ideas.

Interacting with book titles and subtitles is an intrinsic part of cultural and literary-historical practice. However, for a long time, title functions such as the designation of contents, author and form, the elicitation of connotations, the stirring of readers' attention, have not been studied as such. It is Gérard Genette's seminal work that addressed these tacit practices,[3] followed by a rich literature putting into the light that: a. titles are artifacts that depend on the place, moment, users, scope and manner of use; b. titles (and paratext in general) tend to be exposed to erosion, thus to become shorter, but their complexity is not directly related to length; c. there are no works without titles - even the untitled ones put forward a sort of titling practice -, but there are a lot of titles without works and usually the lost works' titles awake a feeling of nostalgia;[4] d. the functions of titles are designation/ identification, description (through thematic, rhematic or mixed indicators), connotation, and seduction (or - if Genette's "seduction" is a too strong word - "attraction");[5] e. our presuppositions on known or canonical titles might be misleading, hence they might generate the so-called "fake titles"; the authenticity of a "true title" being always a question of checking the historical dynamic of 3 or even 4 components: "the title [as such]", "the subtitle", "the generic indicator" and sometimes "the super-title" (e.g. "La Comédie humaine").[6]

---

[2] Margaret COHEN, *The Sentimental Education of the Novel*, Princeton, Princeton University Press, 1999.

[3] Gerard GENETTE, *Paratexts*, 59-106.

[4] See also Harry LEVIN, "The Title", XXIII-XXV.

[5] Madeline HAGGAN, "Research Paper Titles in Literature, Linguistics and Science: Dimensions of Attraction", in: *Journal of Pragmatics*, 2004, 36, 2, 293-317.

[6] Jerrold LEVINSON. "Titles", in: *The Journal of Aesthetics and Art Criticism*, 1985, 44, 1, 29-39.

Genette's studies triggered an unparalleled amount of interest in this device, to the point that research focusing on "paratexts" extended to a "titology" (or "titrologie" in French),[7] before spreading into adjacent fields such as aesthetics and philosophy,[8] book history,[9] or visual studies.[10] Over time, "paratext" has even become a self-marketed merchandise, a "show sold separately".[11] Also, due to the diversification of media and to the dominance of visual over textual culture, scholars have more and more paid their attention to outlandish forms of *peritext* and *epitext* such as the errata, the dedication, the publisher's series, the interview, the preface, the table of contents and so on, laying a special emphasis on the visual (the illustration, the author's portrait pictures), or the performative (the bookshop's window) components of paratext. Digital humanities made no exception to this constant "extension",[12] with (largely unacknowledged) contributions going back as far as 1995.[13] More recently, Moretti's essay on 7,000 titles of British Novels raised a greater amount of interest,[14] and similar studies have started to develop.[15]

To sum up, there appears to be three traditions of titology studies, with surprisingly little overlap. In the field of *poetics,* studies address important dimensions of form and function, but largely revolve around selected (canonical) examples, often carefully picked to illustrate a point in hand; *literary history* focuses on the titling practices specific to authors or periods, and tends to aggregate diachronic

---

[7] Gérard GENETTE, *Palimpsests. Literature in the Second Degree* (*Palimpsestes. La Littérature au second degré*), Nebraska, University of Nebraska Press, 1982. Gérard GENETTE, *Paratexts: Thresholds to Interpretation* (*Seuils*), Cambridge, Cambridge University Press, 1987/1997. Harry LEVIN, "The Title as a Literary Genre", in: *The Modern Languages Review*, 1977, 72, 4, XXIII. Claude DUCHET, "La Fille abandonnée et La Bête humaine, éléments de titrologie romanesque", in: *Littérature*, 1973, 12, 49-73. Serge Bokobza, *Contribution à la titrologie romanesque: variations sur le titre 'le rouge et le noir'*, Geneva, Librairie Droz, 1986.

[8] See, for instance, Colin SYMES, "You Can't Judge a Book by Its Cover: The Aesthetics of Titles and Other Epitextual Devices", in: *The Journal of Aesthetic Education*, 1992, 26, 3, 17-26. Greg PETERSEN, "Titles, Labels, and Names: A House of Mirrors", in: *The Journal of Aesthetic Education,* 2006, 40, 2, 29-44. Nycole PAQUIN (ed.), *Le titre des œuvres : accessoire, complément ou supplément*, special issue of *Protée*, 2008, 36, 3), https://doi.org/10.7202/019629ar.

[9] Eleanor F. SHEVLIN, "'To Reconcile Book and Title, and Make 'em Kin to One Another': The Evolution of the Title's Contractual Functions", in: *Book History,* 1999, 2, 42-77.

[10] Giwoong BAE & Hye-jin KIM, "The impact of movie titles on box office success", in: *Journal of Business Research*, 2019, 103, 100-109.

[11] Jonathan GRAY, *Show Sold Separately. Promos, Spoilers, and other Media Paratexts*, New York, New York University Press, 2010, 6-8.

[12] Guido Mattia GALLERANI, Maria Chiara GNOCCHI, Donata MENEGHELLI & Paolo TINTI, "Introduction", in: *Seuils/Paratexts*, special issue of *Interférences littéraires/Literaire interferenties*, 2019, 23, 1-15. Andrea DEL LUNGO, "*Seuils*, vingt ans après. Quelques pistes pour l'étude du paratexte après Genette", in: *Littérature*, 2009, 3, 155, 102-103.

[13] See Michel BERNARD, "À juste titre: a lexicometric Approach to the Study of Titles", in: *Literary and Linguistic Computing*, 1995, 10, 2, 135-141.

[14] Franco MORETTI, "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)", in: *Critical Inquiry*, 2009, 36, 1, 134-158.

[15] Paul NULTY, "Titles in Digital Book Collections", 2016, [online], http://paulnulty.net/wp-content/uploads/2016/11/booktitles1.html.

studies,[16] but a chronological analysis of titles approached formally - as nominal phrases and predicative devices, across various national or cross-national settings - currently remains a desideratum. Finally, *computational literary studies* (CLS) gather large amounts of data and focus on the formal dimensions of titling practices and their modelling, but do not always integrate the findings and theories of the previous approaches.

Our study tries to bring together these different strands, aiming to contribute to the understanding of titling practices in the European novel of the 19th century and early 20th century, as far as the multilingual literary corpus ELTeC can afford. We will start by presenting our sub-collections and our methodology, then, in a second part, we will describe the most frequent entities one can find in our titles. In a third part, we will analyse the syntax of titles, not in the grammatical sense of understanding relationships between parts of speech or phrases, but through looking at most frequent combinations of entities. Thus, by contrast to Bernard, Moretti or Nulty, our approach is not a lexicometric one. Also, while the number of our titles is considerably smaller than the amount of data taken into consideration in these studies, they concern a much larger and carefully sampled *cultural area* (11 European languages). In addition, our study is characterised by the fact that we can manually categorise this data, then aggregate and visualise it, employing various approaches. Finally, it is important to stress our focus on working on novel titles, as contrasted to Nulty, who works on fiction and non-fiction titles, and to Bernard, who took into account titles of works pertaining to various literary genres.

Our study will mainly focus on the *identification* and *description* functions of novel titles, following Genette. However, we will also offer some insights into *connotation* and *seduction* (*attraction*) functions, especially when describing the entity categories and titles' differentiating strategies. Finally, we will draw some tentative conclusions about titling practices in Europe, destined to stimulate future analysis that can be developed on the basis of our data, or using the same methodology on a larger corpus.

1. Data collection and methodology

The starting point of this study is the European Literary Text Collection (ELTeC). ELTeC aims to provide a corpus of 2500 digitized European novels in at least 10 languages, covering a period spanning from 1840 to 1920. Currently, the members of the COST Action 16204, which drives the

[16] Claude LACHET (ed.), *À plus d'un titre. Les titres des œuvres dans la littérature française du Moyen Âge au XXe siècle*, Lyon, Université Jean Moulin – CEDIC, 2000.

creation of ELTeC, focus on the creation of novel sub-collections for each of the represented languages, carefully selecting texts according to chronological, sociological and literary market criteria. Moreover, texts are supposed to be fully annotated at the linguistic level.[17]

ELTeC contains metadata that enable researchers to build sub-collections, e.g. by period (T1=1840-1859, T2=1860-1879, T3=1880-1899, T4=1900-1920), by author's gender, by the length of the novels (short, i.e. between 10.000 and 50.000 words; medium: between 50.000 and 100.000 words; long: more than 100.000 words), and by the degree of popularity, measured by the number of reprints (between 1970 and 2000).

Out of the sixteen sub-collections present in ELTeC to date, we have annotated title data from eleven language sub-collections, namely English, French, German, Italian, Polish, Portuguese, Romanian, Serbian, Slovenian, Spanish, and Ukrainian. While we are aware that our dataset does not fully cover the European linguistic and cultural diversity (other ELTeC sub-collections still wait to be explored), we consider that our list of title data remains quite valuable since it covers canonical and non-canonical texts, while representing a wide variety of European languages and cultural backgrounds.

The title data we are using reflects the working version of ELTeC dated from the 10th of June, 2020;[18] as the sub-collections continue to grow and to be reorganized so as to better respond to the ELTeC selection criteria, it was necessary to draw a line beyond which changes were not to be considered any more in our study. We extracted the titles and other information from the ELTeC summary page,[19] except for the Polish and Ukrainian language sub-collections, whose titling data has been extracted directly from the XML files uploaded in the GitHub repository. Changes in the language sub-collections that have been included after the above-mentioned date have not been considered for our title study.

The 798 titles we are dealing with are unevenly spread over the ELTeC languages and time periods (designated as "slots" in the following figures), as it can be seen in Figure 1.[20] While some sub-

[17] For more details and updates on the advancement of our work, see https://distantreading.github.io/ELTeC/.

[18] Sub-collections are to be found at https://distantreading.github.io/ELTeC/index.html. For the whole dataset and visualizations, see Patras, Roxana, Odebrecht, Carolin, Galleron, Ioana, Arias, Rosario, Herrmann, J. Berenike, Krstev, Cvetana, Mihurko Poniž, Katja, Yesypenko, Dmytro. (2020). Dataset for ELTEC titles [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4268669. In subsequent notes pointing to this repository, we will only use its short title (Data Set), followed by the DOI.

[19] See https://distantreading.github.io/ELTeC/index.html.

[20] For the script used to produce this visualization, see Data Set, http://doi.org/10.5281/zenodo.4268669

collections are both complete (i. e. gathering 100 novels or almost) and quite well balanced (see ENG, FRA, DEU), others are still work in progress (UKR, SRP or SPA, for instance). In several sub-collections, T1 and T2 are less well represented, partly because 1840-1859 is still an incipient period for the novel genre in certain cultures, partly because the books published during these periods are not always easy to locate, they are in bad conservation shape, or they have proved difficult to convert to machine-readable formats because of the paper quality, typographic specificity or orthographic variation.

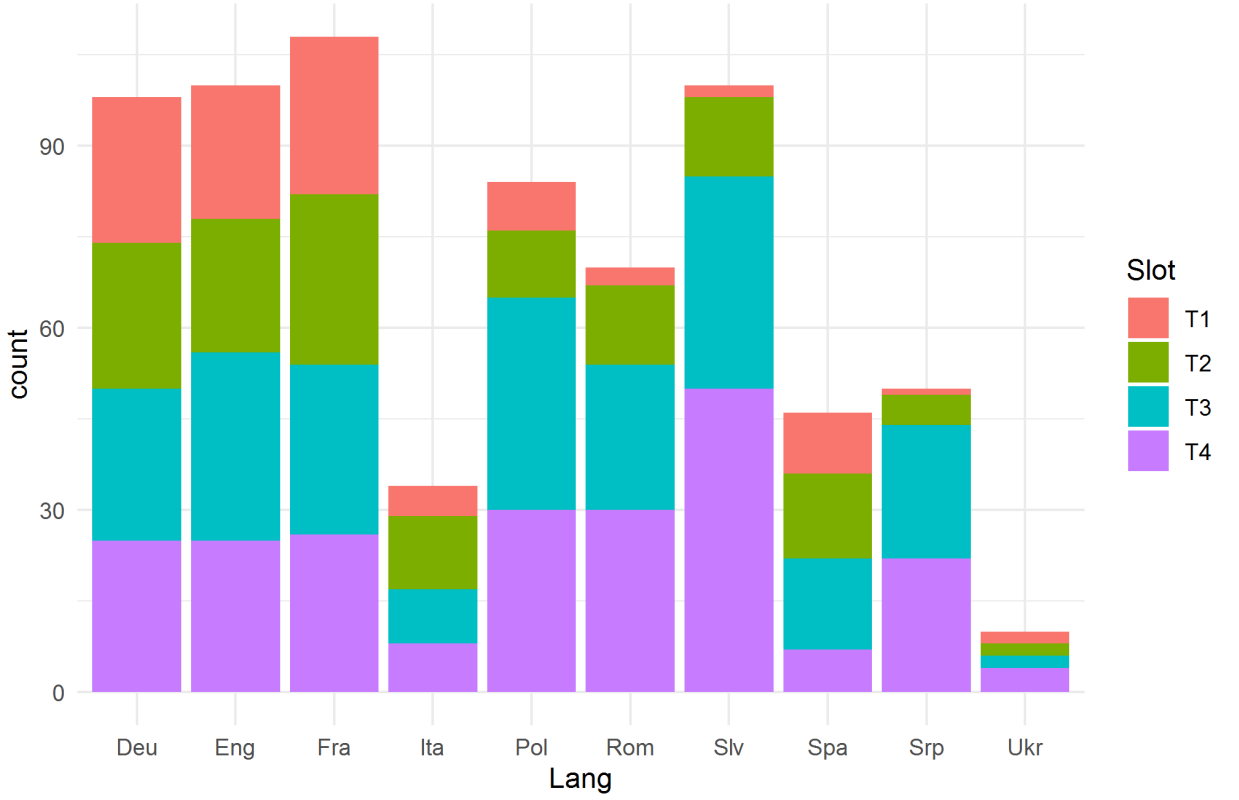

Figure 1. Distribution of novels by language sub-collection and time period slots

An important preliminary step to our study was the verification of titles, since, in some cases, subtitles or alternative titles, as well as genre indicators, were missing from the initial ELTeC XML title declaration. For instance, *I Promessi Sposi. Storia milanese del sec. XVII scoperta e rifatta di Alessandro Manzoni* was reduced to *I Promessi Sposi*. Also, because any titling subsequent to the original title can

only have the status of interpretation, we decided to observe "title authenticity."[21] We thus completed the ELTeC titles and subtitles by consulting information about *first editions*. As a consequence, some of our titles turned out to be much longer than first envisaged. However, it is to be borne in mind that in a few cases we could not locate image documents of the books on the web, while libraries were still inaccessible because of the COVID 2019 pandemics.

Rather than trying to look at the complete linguistic components of the titles, we decided to focus on the nouns in the titles. There are two main reasons for this choice. First, it is well known that titles are composed chiefly of nominal phrases; indeed, other studies have shown that other parts of speech are poorly represented in these chunks.[22] Second, this allows us to focus on person and place names in titles, both being complex semantic categories that need to be treated with methodological care.

Because of the large number of different languages covered in our study, we considered that a manual annotation was the most suitable approach. The annotation was conducted by native speakers of the 11 languages covered, with five exceptions for which the annotators had near native-speaker command (English, Portuguese, Polish, French and Italian titles).

In our annotation, the category "personEntities" comprises three types of reference to persons: via proper names ("*Милан Наранџић*", "Os filhos do padre *Anselmo*"), via statuses ("*Бабадевојка*", "Os *pobres*", "*Unŭ funcţionarŭ* sinucisŭ. *Fratele* şi *sora*"), or through the use of pronouns ("Il *mio* Carso", "Die Mappe *meines* Urgroßvaters", "Ciocoii vechi şi noi sau *ce* naşte din pisică şoareci mănâncă! Romanţ original"). "PlaceEntities" were annotated when mentioned in the title as proper names ("Les Trappeurs de l'*Arkansas*"), or as adjectives ("Der Sonnenwirt. Eine *schwäbische* Volksgeschichte"). All nouns pointing neither to persons nor places have been annotated as "otherEntities"; they could point to a temporal element ("*La velada* del helecho, o El donativo del diablo. Novela"), an object ("Feldblumen"), or an abstract concept ("*Conquista* del Perú: novela histórica original", "Fulga sau *ideal* şi *real*"), to name but a few. Table 1 bellow indicates the categories extracted for each place entity, and gives an overview of the values amongst which the annotators had to choose.

Place	place entity	place attribution	place determiner	place role	place syntax

[21] Hazard ADAMS, "Titles, Titling, and Entitlement To", in: *The Journal of Aesthetics and Art Criticism*, 1987, 46, 1, 7-9.

[22] Franco MORETTI, "Style".

yes/ no	[copy entity name]	yes/ no	def/ indef/ no/ na (for "non applicable")	existence patient agens location attribute possessor possessum	Head apposition pregen postgen prepmod adjective no

Table 1. Annotation table: abstract

Similar fields were to be filled in for "person" and "other" entities. In addition to the above-mentioned fields in relation to place entities, our annotation scheme made provision for four other types of information:

- for person entities, the gender was to be indicated as m (male), f (female), d (for collective characters gathering representatives of both sexes or for neutral nouns). If no gender category was applicable, annotators had to indicate "no".
- titles made of two components were to be signalled, either they were displaying an alternation (*Modern Flirtations, or a Month at Harrowgate*), a subtitle (*Le Petit-Chose. Histoire d'un enfant*) or a rhematic appendix (*Нове: роман*).
- the generic indicators in titles were to be listed, as well as their role and syntax. In contrast to Moretti, we decided to scoop all generic indicators, regardless of the amount of novelty they may have, or not, in certain cases.
- and finally, the focus of the title was to be understood as the single or multi-unit word that the annotator considered to be the most salient, the most semantically rich in the title. For this last column, we also decided to include a second round of annotation, conducted through a separate call for volunteers within the COST "Distant reading" network. The aim is to ascertain if different readers spot the same "main" word in titles, or to what extent their annotations may differ. This call being still in process, we will not discuss this aspect in what follows.

Annotators could, of course, duplicate or triplicate the columns in case titles displayed more than one entity of a kind.[23] In spite of being quite numerous, our categories proved to be quite robust and easy to handle, allowing us to observe several regularities and trends in the titling practices at the European level.

2. What's in a Title?

In this section, we will focus on the frequency of thematic entities (*persons*, *places*, *other entities*) and rhematic entities (*genre indicators*) in the sub-collections under study. From here, we hypothesise about "titling patterns" of "the great unread", and about particular aspects of titling in the eleven sub-collections of ELTeC.

The analysis of the evolution over time reveals that frequency of all entity types decreases in T3 and in T4. In other words, the total amount of entities decreases towards the end of the nineteenth century and early twentieth century. However, this trend might indeed reflect a sampling imbalance, as several sub-collections (POL, ROM, SLV, SRP) included more novels for T3 and T4.

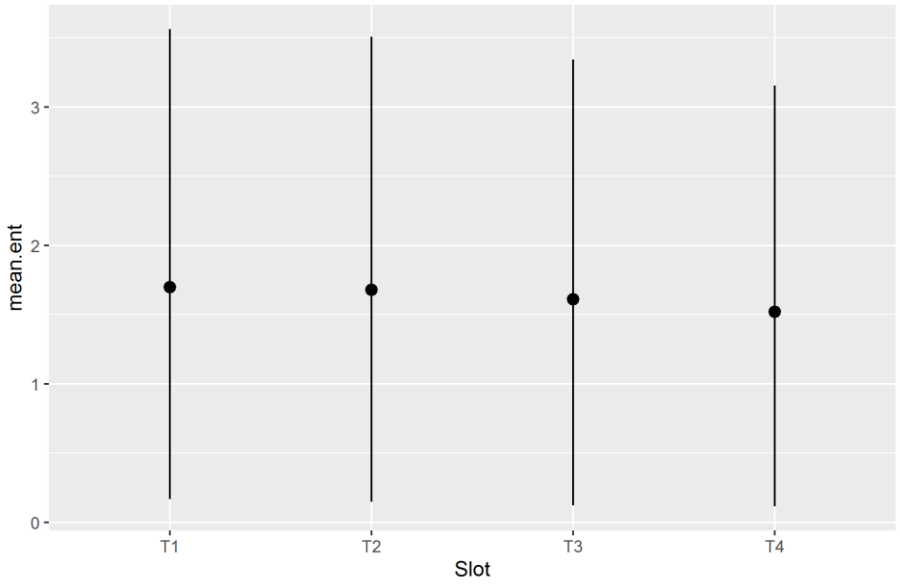

Figure 2. Plot with mean frequencies of entities (all types) in our 11 sub-collections[24]

[23] The annotation guidelines are available at https://github.com/distantreading/WG1/blob/master/titlePilotStudy/data/dataPreparation.md.

[24] We count the identified entities for person, place and other entity for each title in each sub-collection and group them by time slots and added standard error 0.05. We divided the entity frequencies per slot by the numbers of

As general trends per each type of entity, our data shows the following (see Figure 3): a. while the frequency of personEntities doubles from T1 (n=90) to T3 (n=180), in T4 it drops substantially (n=140); b. placeEntities show an overall steady increase from T1 (n=25) to T4 (n=50), with two leaps between T1 and T2, and between T3 to T4; c. otherEntities also show an increase trend, but with a leap between T2 (n=90) and T3 (n=145) and another substantial rise in T4 (n=159). Overall, and for almost all time slots, personEntities (represented as proper names, pronouns or statuses) have the highest frequency across ELTeC's titles and subtitles. T4, however, seems to indicate a relative shift of interest from person references to place and especially to 'other' references.

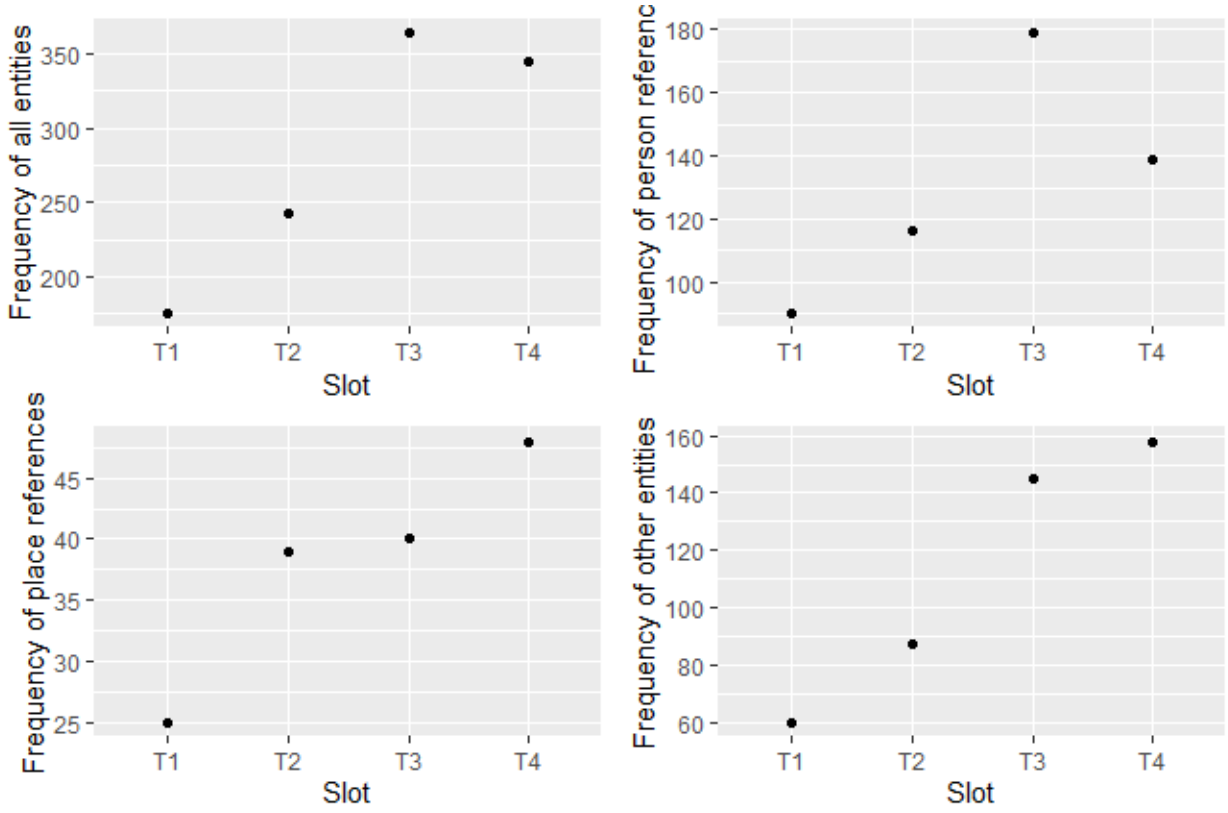

Figure 3. Frequencies of all entities, person, place and other entities per time slots across sub-collections (top left to bottom right).[25]

a. Persons

titles in a time slot. The plot is created with RStudio (RStudio Team 2020, Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

[25] Frequencies are aggregated over single titles. One title might contain more than one type of entity. The plot is created with R Studio (RStudio Team 2020).

In all the time slots, person reference features prominently in titling. This goes along with Moretti's observation that proper nouns appear frequently in nineteenth-century titles: "proper names are even more frequent, especially at the turn of the century".[26] However, our data shows that person entities are not always designated by a proper name, and some interesting differences in title practices appear between the sub-collections. In the English sub-collection, for example, names appear in 63% of all titles, whereas only 25% point towards a status, which is, more often than not, individual rather than collective (*The Tutor's Ward* rather than *A Tale for Mothers and Daughters*). This is also the case for the Spanish titles, with 67.3% out of the 46 titles including a name, and 45.6% a status. It is significant that collective person entities are scarce in both the English and the Spanish collections. To the contrary, this proportion is slightly equal or in favour of statuses in the Italian, French, Polish, Portuguese, and Romanian sub-collections. For instance, in the Portuguese collection 58% of the personEntities are designated by a status, while only 42% are named by proper names; in the Romanian collection, the predominance of statuses is even clearer (more than 60% of all personEntities). While the reasons of these preferences remain to be understood, they appear as a promising opening for further investigations.

Statuses in our collection can be grouped by *family* ("mothers", "wife", "son", "enfant", "sœur", "брат","сестрица"), *professions* ("capitan", "soldier", "professor", "leutnant", "docteur"), *social ranks* ("краља", "кнез", "krola", "marqués", "infanta", "senhora duquesa", "reis", "barao", "rainha", "principesă"), and *citizenship* ("italiano", "ardeleni"). The following figure 4 shows these person statuses as a word cloud.

[26] *Ibid* 143.

Figure 4. Person statuses in the entire collection[27]

Not very surprisingly, titles referring to women express their family affiliation rather than professional roles or citizenship: *mother* ("Mutter", "mere", "mati", "мајка"), *wife* ("Frau"), *daughter* ("filha", "hči"), *sister* ("Schwester", "soeur", "sora", "сестриця"), *mistress* ("courtisan"), *spinster* ("Бабадевојка"), ward ("pupille"). For male person entities we find quite symmetrical family statuses: *father, grandfather* ("dziadunio"), *great-grandfather* ("Urgroßvater"), *husbands* ("mężowie", "maris"), *cousin* ("cousin"), *son* ("filhos"), and *brother* ("frere", "fratele", "brat", "брат"). Nevertheless, the social statuses of women, unlike those of men, are the only ones to explicitly point towards violations of social norms: she may be a *sinner* ("peccatrice"), a *captive* ("branki"), a *divorcee* ("divorciada"), or a *witch* ("bruxa"). In this respect, the only notable exceptions about men are the Romanian outlaws *hajduk* ("haiduc") and *thief* ("tâlhar", "bandit"), and the anarchist/ extremist statuses such as the "radical", the "partisan", the "prophet(s)" (ENG), "errants", "usurpateur" (FRA), "agitador" (POR). Accordingly, a suggestion about the violation of the moral code can fulfil the seduction function.

[27] Voyant tools visualization, see https://voyant-tools.org/. See the full list of person statuses in Data Set, http://doi.org/10.5281/zenodo.4268669

On the contrary, male personEntities are often connected to diverse jobs such as *professor* ("professeur", "učitelj"), *physician* ("doktor", "docteur"), *weaver*, *mayor* ("burmistrz"), *lawyer*, *priest* ("padre", "поп"), *captain* ("capitan"). These denominations are almost non-existent for female entities, even if one finds mentions of a *peasant woman* ("Bäurin"), a *watchmaker* ("Uhrmacherin"), or an *ironing woman* ("engomadeira"). There is also a noticeable difference concerning first and last name: while female characters dominate by first names, male characters often have a surname. A title from the Polish collection seems particularly illustrative here: 'Krysia bezimenna...' [Chris who has no name]. Sometimes, one can find a last name without first name, but almost never just a first name for male persons. Sometimes there is a surname next to a woman's personal name too, but in some cases it expresses property or a kind of a dependency link (Radetić' Mara, Mikel's Zala).

In German, English, French, Portuguese, Polish, and Ukrainian sub-collections, female authors tend to point towards female person entities in titles[28]. This could be related to the longer tradition of female authorship in these literary traditions, and maybe to the intensity of the female movement in these literary traditions. An exception may be found in the Romanian collection, where female authors favour male and diverse personEntities ("*Haiducul*", "*Pandurul*", "*Martirii*") and otherEntities ("*Robia banului*", "*Spre desrobire*", "*Voința*"). Overall, in ELTeC there are almost no titles involving both female and male names – with a few exceptions such as *Louise et Barnavaux* or *Radetić' Mara*.[29]

Comparison of female gender across different time slots did not confirm the hypothesis that the progress of the Women's Rights Movement and the increasing number of women writers on the literary market, would increase the number of eponymous heroines. This can be seen in figure 5.

[28] For the figures supporting this affirmation and the following paragraph, see Data Set, http://doi.org/10.5281/zenodo.4268669

[29] Incidentally, one may also observe that this title does not follow the dominant pattern, where the masculine name is in the first place (see Devoney Looser, Preface, in: *Jane Austen. Sense and Sensibility*, New York: Penguin, 2018, xix).

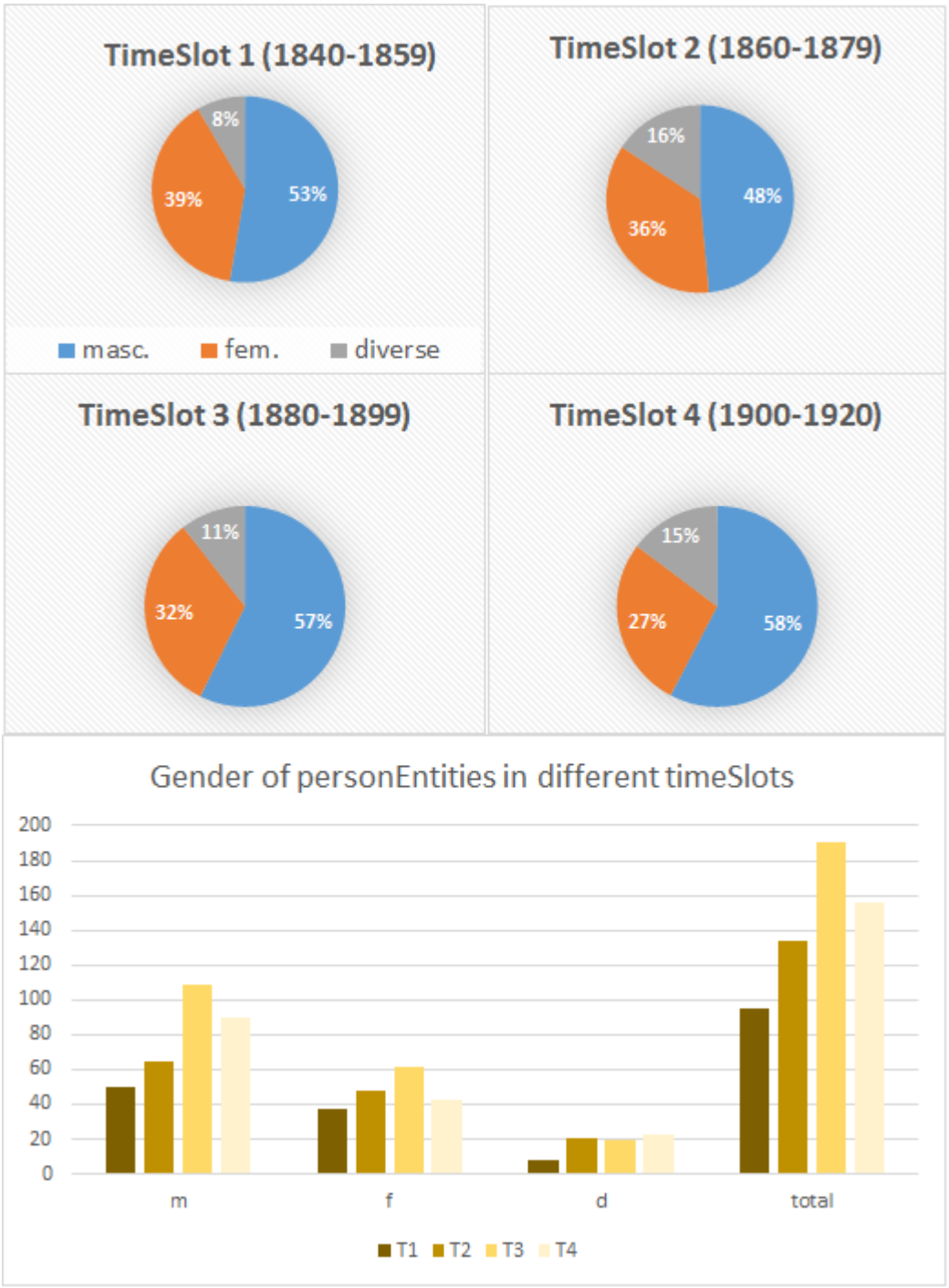

Figure 5. Proportion of person entities' gender by time slots[30]

[30] Microsoft Excel visualization. Complete figures for gender of personEntities can be seen in Data Set, http://doi.org/10.5281/zenodo.4268669

b. Places

According to a commonly used distinction between "absolute" and "relative" references,[31] place entities can be organised in two main categories: those that can be put on a map ("A conquista de *Lisboa*", "The Mysteries of *London*", "Colette Baudoche: histoire d'une jeune fille de *Metz*", "*Nürnberg*. Culturhistorischer Roman aus dem 15. Jahrhundert. Zweiter Band", "Nad *Niemnem*", "I Promessi Sposi. Storia *milanese* del sec. XVII scoperta e rifatta di Alessandro Manzoni") and those that cannot be located ("Mon *Village*", "La *gura sobel*", "V študentskih *ulicah*. Ljubezenska povest").[32] As shown in figure 6, this second category tends to increase over time, while "real" place names are more evenly distributed.

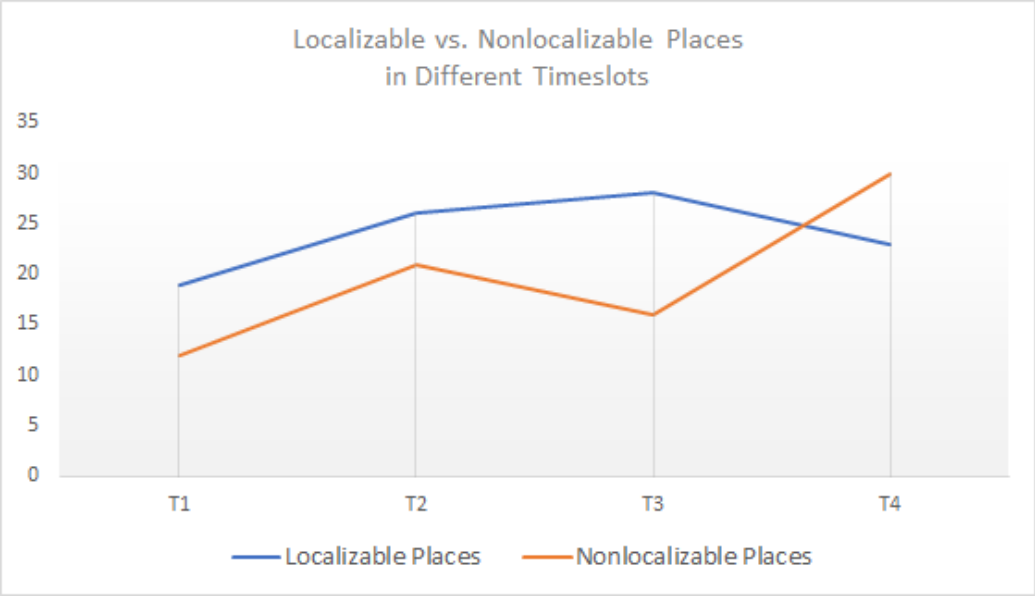

Figure 6. Absolute and relative place references in titles from different time slots[33]

This general trend hides, however, more specific national preferences, as it can be seen in the following figure 7. The Portuguese, French, German, and Italian sub-collections contain more absolute than relative spatial references, while the Spanish, Polish, Romanian and Slovenian titles

[31] Stephen C. LEVINSON, *Space in Language and Cognition. Explorations in Cognitive Diversity*, Cambridge, Cambridge University Press, 2014, 26.

[32] The full list of geographical references can be found in Data Set, http://doi.org/10.5281/zenodo.4268669.

[33] Microsoft Excel visualization. Complete figures can be seen in Data Set, http://doi.org/10.5281/zenodo.4268669.

appear to be quite balanced between the two types. The extreme cases are the English and the Serbian sub-collections, with the first one manifesting an obvious preference for placeEntities that cannot be put on a map, and the latter one systematically requiring the mobilization of "the reader's encyclopaedia".[34] Considering that several collections are still growing, it will be interesting to see if the final datasets confirm this preliminary observation and provide insights about the potential reasons of this contrasted distribution.

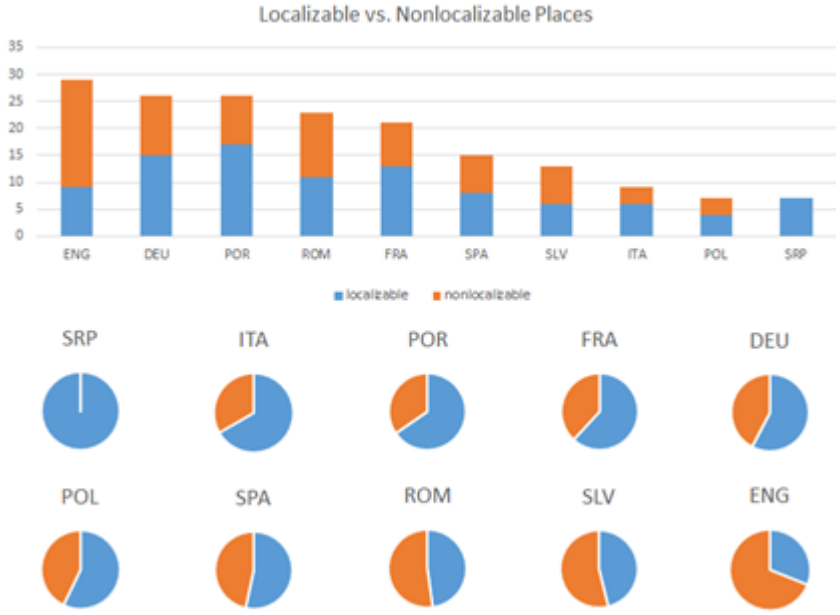

Figure 7. Localizable and non-localizable Places[35]

Drawing a map of the geographical references mentioned in our sub-collections appeared as extremely challenging, since the identified placeEntities present very different degrees of specificity/generality. As shown in the above-mentioned examples, in some cases the geographical reference is quite precise (a town, or even a specific street or place in a city), but quite often titles mention entire states or regions ("Les Trappeurs de l'*Arkansas*", "Conquista del *Perú*: novela histórica original", "Die Mandanenwaise. Erzählung aus den *Rheinlanden* und dem Stromgebiet des *Missouri*", "Arhanghelii. Roman din viaţa românilor *ardelenî*"), or even broader geographical contours ("Avram

[34] Bertrand WESTPHAL, *Geocriticism. Real and Fictional Spaces*, New York, Palgrave Macmillan, 2011, 98.

[35] For each sub-collection, placeEntities have been extracted and further annotated as "localizable" or "non-localizable". Proportions are obtained dividing the figures for each of these two categories by the number of placeEntities per sub-collection. The Ukrainian sub-collection has not been taken into account here as it contained no placeEntity so far. Microsoft Excel visualization. The absolute values can be seen in Data Set, http://doi.org/10.5281/zenodo.4268669

Iancu, regele *Carpaţilor*, continuatorul operii lui Horia", "Za svobodo in ljubezen. Roman z *Balkana*", "Прве жртве : приповетка из *српске* прошлости"). Further investigations are needed, both for scholars to find consensual ways of representing on a single map these different geographical scales, and to understand the reasons leading the authors, in various national and time settings, to prefer more or less specific geographical references.

In addition, practices appear quite heterogenous when looking at the relation between the nationality of the authors and the geographical locations they tend to mention in titles. Polish titles tend to refrain from mentioning places within the current Polish territory, preferring to refer rather to exogen locations such as Galicja, Niemen (previously in Poland), or Gehenna (Israel) and Calvados (Normandy). To the contrary, the Portuguese collection favours the mention of places within the current territory of Portugal (Lisbon, Coimbra, Adro, Mindelo). By far, the German and the French titles contain the most exotic and faraway places with respect to their current borders (Arkansas, Missouri, America, Ninive, Kabylie, Japan, Mexico). In the meantime, and interesting enough, when in Germany, places mentioned in German titles tend to be peripherical or regional (e.g. Stechlin).

c. Genre indicators

Beyond the thematic clues they contain, titles in our collection offer rhematic information in 44% of the cases. Over time, specification of the genre becomes more and more frequent, growing from 32.5% in T1 to 47% in T4.[36] Differences, however, remain important between the French, Portuguese, Spanish and English collections, where there is often no genre indication, and the Italian, Polish, Slovenian, Romanian and Serbian ones, whose authors (or maybe publishers) display it quite systematically.

Usually, genre indicators appear as an appendix to the main title, either as the head of a new sentence (*Alas! A novel*), or as an apposition (*El testamento de Don Juan I, novela histórica original*).[37] In a minority of cases, genre indicators are however the head of the title (*Diário de uma criança*) or of the subtitle (*Le Petit chose. Histoire d'un enfant*), with the Slovenian collection offering several examples of

[36] See the full list of genre indicators and their frequencies in Data Set, http://doi.org/10.5281/zenodo.4268669

[37] The annotation guidelines invited to annotate as "head" all genre indicators appearing after a strong punctuation (full stop, exclamation mark, question mark); genre indicators following a column or semi-column were considered as apposition. More delicate was the case of titles displaying no punctuation between the title and the genre indicator, for instance when the indicator appears on a new line under the title. In these cases, the capitalization has been considered as the start of a new sentence, and therefore genre indicator was to be annotated as head.

strictly rhematic titles (*Potresna povest*). However, we could not work out a clear pattern for these denominations in terms of time period, language or author's gender.

The most interesting aspect of the genre indicator is the wealth of denominations our titles display. As figure 8 shows, three different traditions are shared in the literary space, with most Latin countries and Germany favouring the "roman" label, Slavic speaking territories preferring "povest", while England and Spain talk about "novels".

Figure 8. Genre indicators in European literature as reflected in ELTeC collections[38]

Etymological studies have long showed the different meanings conveyed by these labels: the "roman" one reminding that the novel was, in the Middle Ages, considered a modern genre, written in the newly formed romance languages; "povest" ("narration") insisting on the narrative dimension of these texts; and "novel" pointing towards the demand for new contents and surprising adventures that such books satisfied, at least at the times of their rapid spread in the literary landscape. In the 19th century, however, these meanings are probably rarely perceived by the readers, and the three types of

[38] Manual adding labels with Inkscape; thanks to NASA for the map. Prominent words are the most frequent indicators per sub-collection, the other ones are the less frequent.

genre indicators seem to reflect linguistic variation, as well as mutual influences, rather than deliberate choices.

Beyond the three main genre indicators, one notices variation due to attributes ("Elena. Roman *original de datine politic filosofic*") or alternative lexemes (see figure 8 and figure 9). This may confirm Moretti's and Underwood's opinions about novel sub-genres.[39] Novels might be designated as "books", "memories", "lives", "histories", and many others. However, in this rhematic variety, we could trace some alternative titling practices: for instance, in the Romanian collection, the quasi-synonymous "roman" (26 occurrences), "nuvelă" (4 occurrences) and "romanţ" (1 ocurrence) are present. Almost all sub-collections have a genre indicator about the novel's unexpected short length or fragmented nature: "episodio(s)" (POR, SPA), "episode" (ENG), "episod" (ROM), "življenjepis" (SLV), "scene" (IT), "sketch" (ENG), "слика" (SRP). Similarly, almost all collections contain genre indicators indicating a longer length than expected: "history"/ "historia"/ "istorie"/ "narrativo"/ "kronica"/ "chronika".

Figure 9. Most frequent genre indicators in ELTeC collections

[39] Franco MORETTI, *Distant Reading*, London and New York, 2013, 71-86; Ted UNDERWOOD, "Distant Reading and the Blurry Edges of Genre", 22 October 2014, [online], <https://tedunderwood.com/2014/10/22/distant-reading-and-the-blurry-edges-of-genre/>.

Some sub-collections of ELTeC, such as the Italian and German ones, are particularly inventive with respect to alternative genre indicators ("storielle", "procesi verbali"; "Amouresken", "Volksgeschichte"). In some cases, the writing of the national history or the will to insist upon the human veracity of the story means finding more commendable labels than the "novel" one, tainted with a hint of frivolity: see, for instance, *Cantoni il volontario. Romanzo storico di Giuseppe Garibaldi*, *Јурмуса и Фатима или Турска сила сама себе једе: прича о ослобођењу шест округа 1832-1834*, or *Clemencia. Novela de costumbres*. In other cases, one may wonder if the alternative designation of the novel as "story", "narration", etc. does not translate an uneasiness to declare the textual product as pertaining to a known tradition – as in the case of Radu Rosetti's *Cu paloşul. Poveste vitejască din vremea descălecatului Moldovei*, or maybe of Zuzanna Morawska's *Wilcze gniazdo. Powieść z czasów krzyżackich dla młodzieży dorastającej*. It is difficult to say at this stage of our work if this relates to a form of precaution, especially from writers in newer cultures, or to a certain rejection of the genre, considered to be too codified for the expressive needs of the writer.

3. Title Patterns

Previous research has suggested that titles tend to become shorter over the centuries, possibly for pragmatic reasons, as Moretti has argued (2009). A short title is easier to handle, to display, to publicise; it fits nicely into a catalogue, be it of a library or a bookseller, and it is quickly memorised. Thus, the long and descriptive titles of the Middle Ages or the 16th, the 17th and even, sometimes, the 18th centuries, are abandoned with the rise of modern book industries.

To a large extent, our collections support this affirmation, whether one looks at the word counts (figure 10) or at the number of entities (figure 11). Titles can have up to 22 words and 6 entities (to which, in some cases, a genre indicator is to be added), but titles longer than 10 words are extremely scarce.[40] Most titles count 4 words and 3 entities, with a quite sharp decrease for 5 words and beyond. The lion's share remains that of titles displaying between 2 and 4 words (56.6%), and one entity (51.8%).

[40] The last length category appears bigger than 7-, 8- and 9-word titles, but this is only because it aggregates all titles displaying more than 10 words.

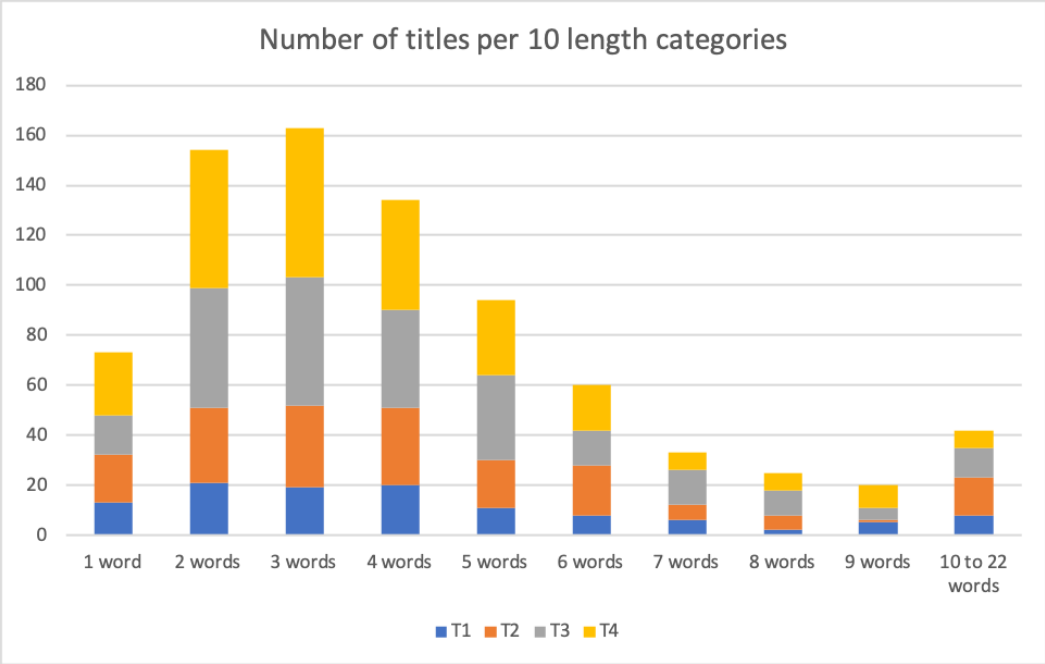

Figure 10. Length of titles in words, per time period

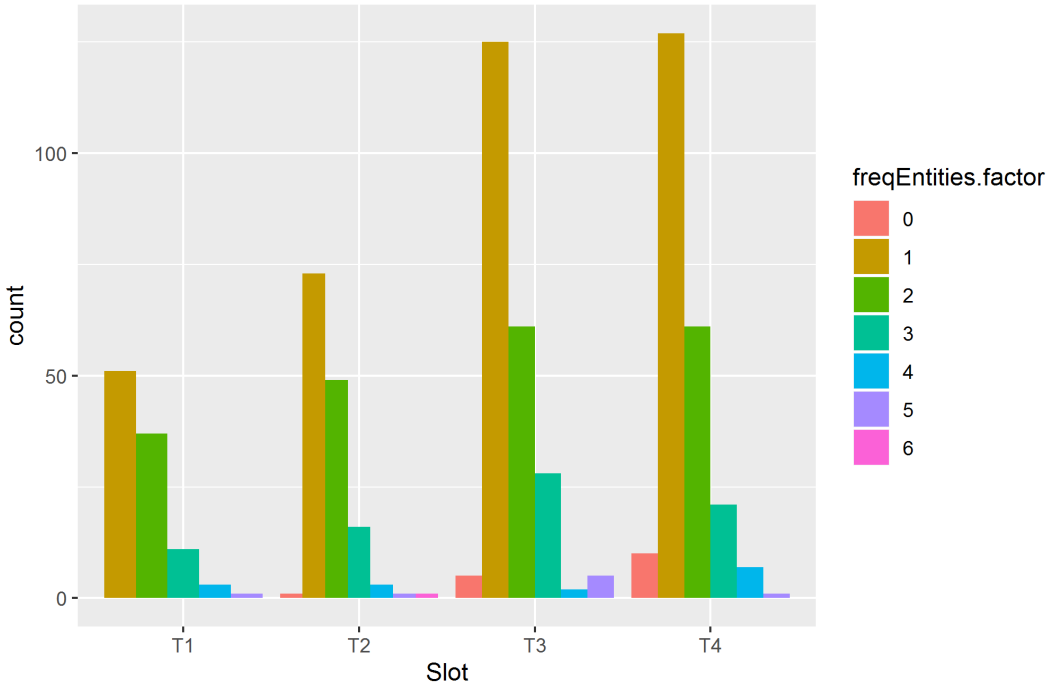

Figure 11. Proportion of titles per number of entities and time slot[41]

However, this observation needs to be somewhat qualified. First, the 'erosion phenomenon' does not go further during the period covered by our ELTeC sub-collections, and our mean value of 4.24 words/title is above those Moretti indicates to be the most frequent in his data (one, two or three words).

Moreover, long titles are not numerous, but they do not disappear either; their share in our T4 (6.87%) is similar to the one they hold in T1 (7.08%), and one can even observe a spike during T2 (11.11%). Combined with the augmented presence of other entities to the detriment of persons, a tendency that has been documented in the previous section of this paper, this may be seen as an indication that authors are not satisfied with the established titling practices, and look for alternatives. Such experimentations are based on aesthetic as well as, probably, seduction reasons: in a market dominated by short titles, a long one may, paradoxically, better attract the eye. In most cases, long titles come with a first part syntactically separated to the rest, that can serve as a short identifier: see, for instance, *Le Docteur Omega (Aventures fantastiques de trois Français dans la Planète Mars* (Arnold Galopin, 1906), *Sonata de estío Memorias del Marqués de Bradomín,*[42] (Ramón María del Valle-Inclán, 1903) or *Die Abendburg. Chronika eines Goldsuchers in zwölf Abenteuern* (Bruno Wille, 1909). Writers seem thus to try and combine both the economically-driven need for providing a short identification, and their wish to follow their creative impulses, in spite of these two tendencies being somewhat at odds.

In the meantime, shorter or longer titles seem to be rather a matter of national culture. A sharp difference can be seen between the Portuguese and Spanish collections, with the first one displaying only 8.16% of titles longer than 6 words, while in the second one a title out of three is a long one (34.74% of the whole).

[41] For each title, we count if the title has one or more entities. We then group the titles containing 1, 2 or more entities in time slots. The plot shows that most of titles in all slots contain only one entity. The plot is created with RStudio.

[42] The lack of punctuation between the two parts of the title reflects the original disposition on two lines.

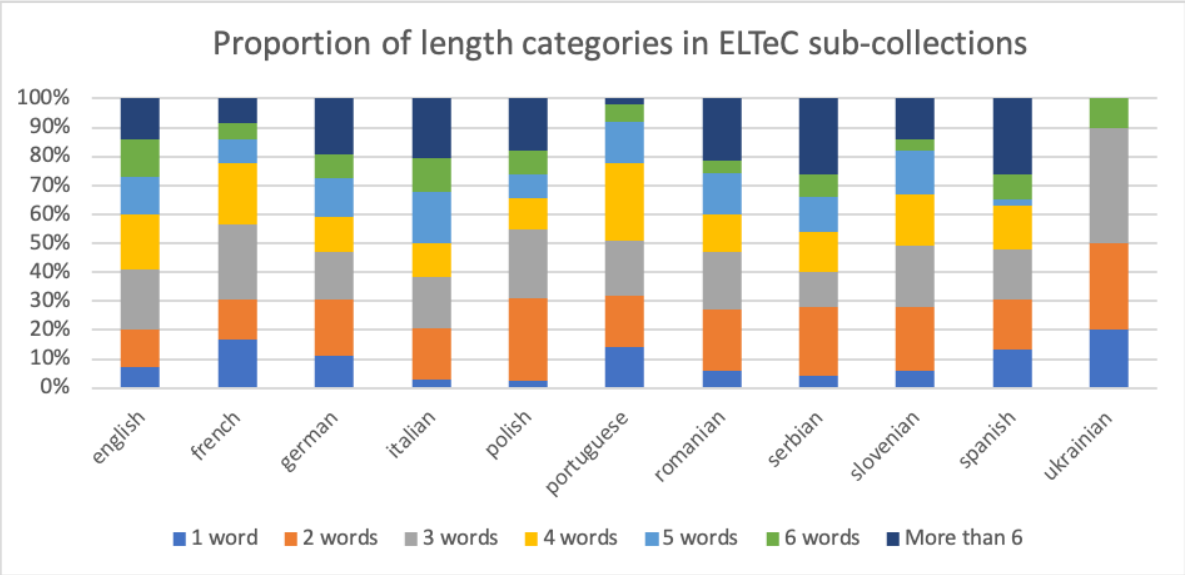

Figure 12. Proportion of length categories in ELTeC sub-collections

Quite surprisingly, Latin language collections do not group together when looked at from this point of view, with the Italian (32.35%) and the already mentioned Spanish novels titles proving to be more verbose than the French (13.89%) and the Portuguese ones. Polish, English and German collections are somewhere in between, with some 20% to 25% of long titles all periods included. Once again, these figures are to be taken with some caution, since the original title pages could not always be retrieved in the process of compiling data, and the initial wording may have been lost in the process or, to the contrary, supplemented with later added indications. Nonetheless, further investigation into these national differences is required, so as to understand book labelling practices in the various cultures, and to identify to what extent historical events or the "crowded market" phenomenon have left their imprint on them.

Whatever the title length, most writers seem to favour "double" titles, as it can be seen in figure 13.

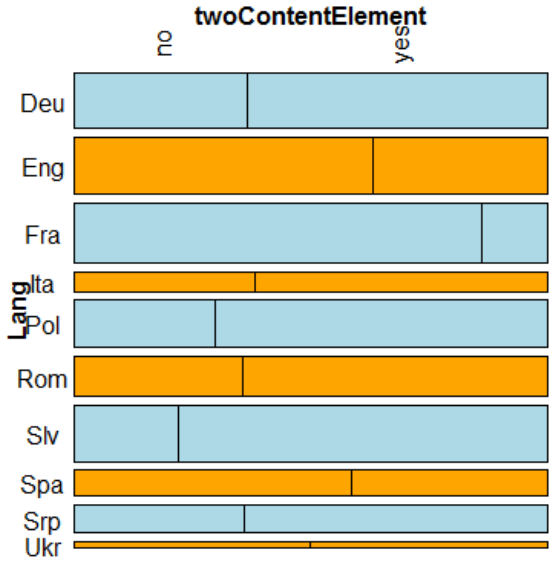

Figure 13. Share of double titles in ELTeC collections.

Single titles appear more frequently in English, French and Spanish novels, and several historical or aesthetic reasons can be imagined for explaining this: a more prescriptive environment, asking the author to provide a short, memorable and manageable label; greater individual and collective experience in titling novels, since the genre is quite old and well-established in these three cultures; authors' will to remain more suggestive than descriptive (and maybe to play with the reader through being more elusive), and so on. As observed above, these are also the cultures in which genre indicators appear less frequently, maybe because other elements of the context, such as the bookstore section where a book is sold, or the specialisation of the publisher, inform sufficiently the reader about the type of text s/he is buying. All these explanations, as well as the rough difference made here between younger and older novel markets, are to be furthered by larger overviews about the subsequent developments in cultures practising longer titles and genre indicators during the 19th century, as well as by consolidating the data within our time periods.

Beyond the "shrinking" tendency that our data supports, our annotation uncovers further stereotyping trends in the European novel. When involving more than one entity, titles favour certain combinations more than others, as it can be seen in figure 14. Persons tend to appear in connection with other entities (12% of the titles), rather than with places (4.6%) or with another person (4.4%).

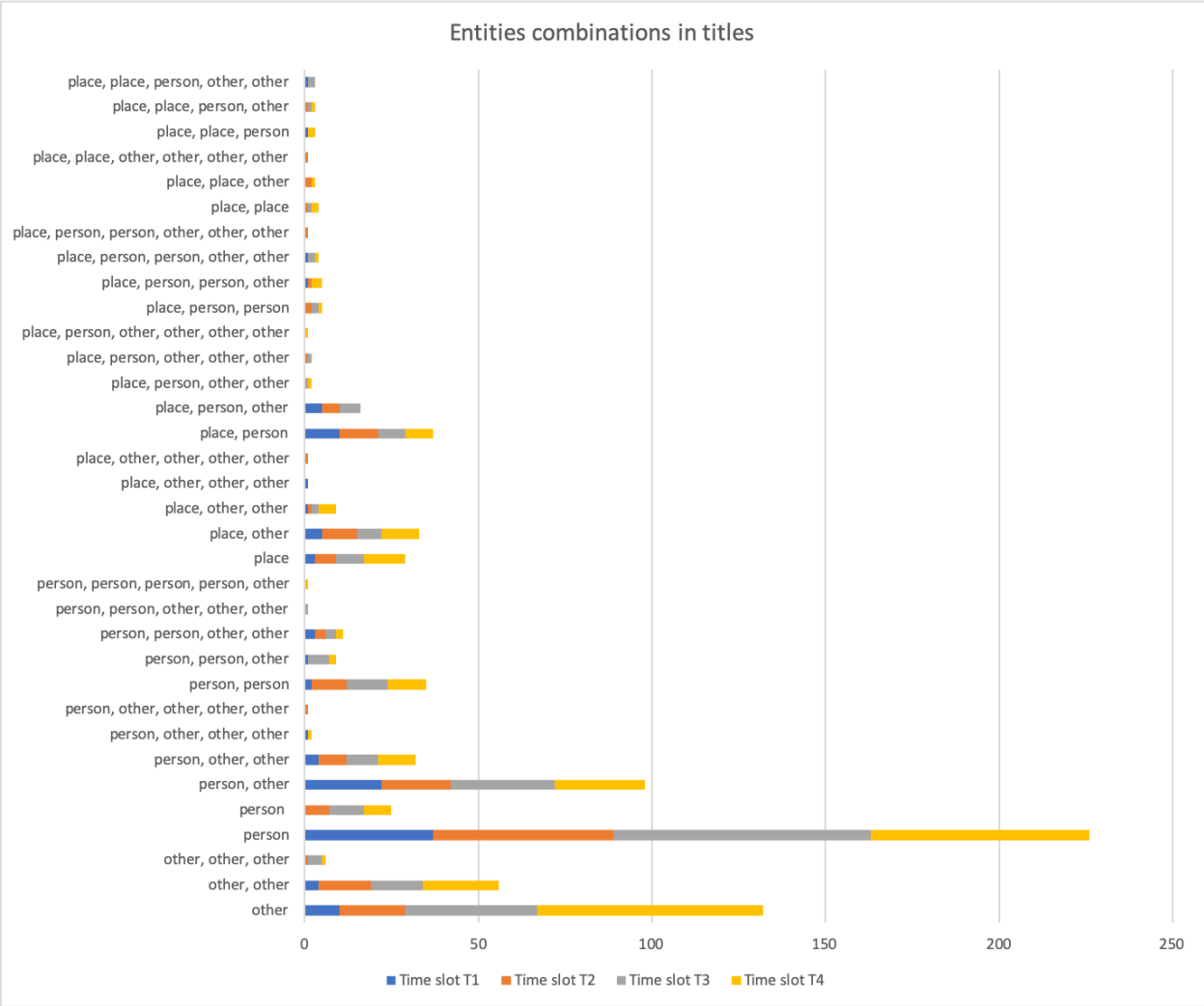

Figure 14. Sets of entities in titles[43]

This is somewhat different from the findings of Nulty in his 2016 blog,[44] where the most frequent POS combination is "PROPN_PROPN", suggesting a conjunction of two persons names, or of a person and a place name. While it remains to be confirmed on larger datasets, this tropism of "persons" towards other entities in European titles is further supported by the observation that the most frequent three entities combination is "person, other, other" (actually, the only one to score above 4%).

[43] Nouns in titles have been replaced by the type of entities they point to; the order of appearance of entities in titles has not been respected, so as to reduce entropy: "other, person, other" is equivalent, in this visualization built with Excel, to "person, other, other".

[44] Paul NULTY, "Titles", [online].

In most cases, the combination "person, other" dwells on a possessor/possessum relationship: *Păcatul rabinului*, *Die Mappe meines Urgrossvater*, *Beatin dnevnik*, *The Prophet's Mantle*, and so on. Events in which characters are involved or objects they possess are thus singled out and framed as able to attract readerly interest. This is to be connected to the fact that "other, other" is the second most frequent combination (7.02%), largely above "person, place" or "person, person", and raising sharply over the time periods – with 22 titles in T4 after starting with 4 titles in T1. On the contrary, "person, other" tends to diminish, going from 19% of the titles in T1 to 10% in the last period. ELTeC time frame is certainly not that of the "crisis of the character", but such observations make one wonder whether what we are seeing here is not an early uneasiness of the novel with characters, that will ultimately lead to their complete transformation in the 20th century.

Lastly, our dataset allows us to ponder on the way titles engage with the reader. To what extent do 19th-century writers build their titles as charades or, on the contrary, tend to be quite explicit in telling the reader (or pretending to tell) what the book is about? Obviously, longer titles carry more information than shorter ones, and in countries where authors are quite verbose one may form a more complete, if not always accurate, idea about the contents. But even when titles are short, there is a huge difference between *Fede e belezza* and *Memorie di Giuda*, to take but two examples of the same length from the Italian collection. While the conjunction of the two entities in the first case suggests a novel about a woman both faithful and beautiful, the second one allows to make a firmer guess that the book will be a first-person narration, either by the mythical (or historical) character of Judas, or a by a modern character committing some unforgivable treason. Beyond such personal interpretations, one may observe that, in the first case, the title states the existence of the two abstract concepts, while in the second case the possessor/possessum relationship, whose importance has already been underlined above, adds a supplementary layer of information. A closer scrutiny of our titles puts forward other more refined suggestions made by the various entities: some indicate locations that are not necessarily place names in the geographical sense of the term (the "afterlife", for instance), other have attributive roles, and in a minority of cases one finds indications about the patient or the agent of an action.

On the whole, two categories of titles can be identified (see figure 15). The first one groups titles that conjure up one, two, or, more rarely, three entities about which we only know that they *exist*. The second one intertwines persons, places and other types of nouns in subordination constructions, giving them more precise *roles*. While the evolution of these categories over time is more difficult to understand, the national practices have clearly their say, suggesting different communicative practices

between authors and readers. In five of our sub-collections (French, Serbian, Ukrainian, Slovenian and Italian), titles tend to be more impenetrable: they state the existence of a person, a place or another entity in more than 60% of cases, leaving the potential reader to speculate about their role in the book. On the contrary, German, Polish and Spanish writers tend to be more explicit in this respect. Last, English, Romanian and Portuguese subcollections are quite balanced, with a little more than half of the titles stating just an existence, while the other half explaining the role of the entities involved.

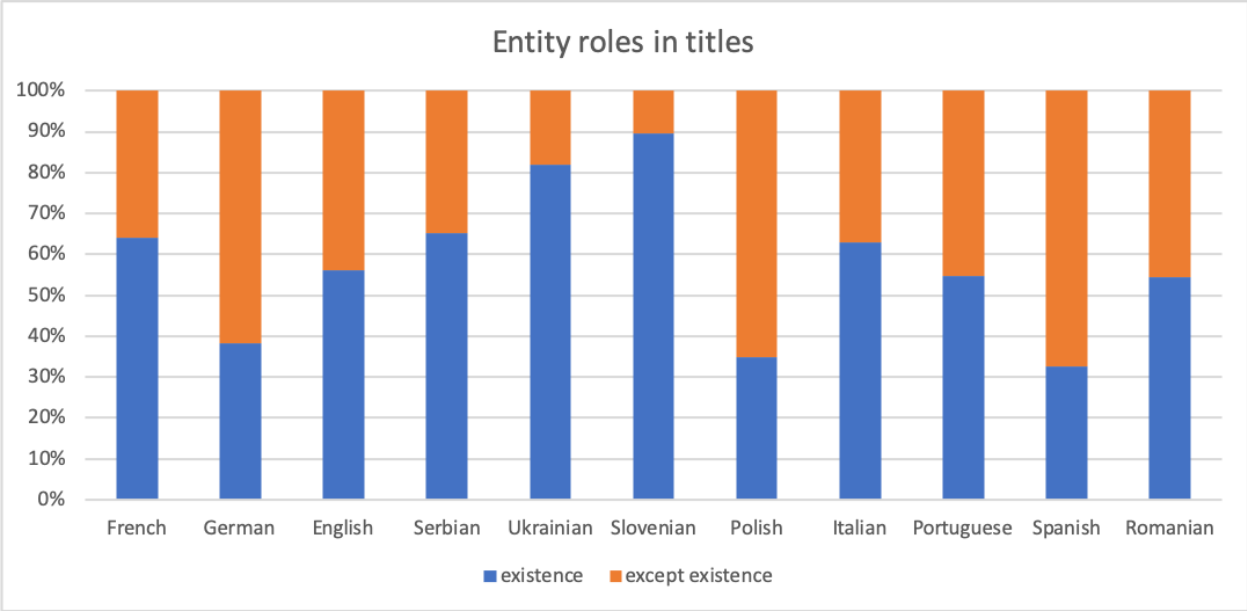

Figure 15. Entity roles in titles

Conclusion

A two to four-word title, involving one or two entities (one of which being a person), seems to be the most frequent formula for the European novel title during the long 19th century. However, as shown in our first part, over time, titles tend to become more and more about places and other entities, potentially pointing thus towards a shift in readership tastes. Also, titling practices suggest that authors progressively open to experiment with new types of stories and narrative formulas.

Interestingly, novel titles are either about men or about women, with the two appearing quite seldom in conjunction. This is maybe to avoid the immediate association of the novel with an erotic plot, and the negative representations (weakness, low taste, corruption of the youth) that are

sometimes associated with such themes. In the meantime, the number of places that cannot be located on a map suggests that novels in our sub-collections do not have difficulties in displaying and assuming their fictional nature. Denomination practices in the novels seem therefore to point both towards the triumph of the genre in the long 19th century, and towards a form of uneasiness, when compared to other fictional genres.

Maybe the most important finding of this study is the wealth of differences one may observe with regards to titling practices and traditions. Proportions and combinations of entities vary over time and in different national cultures, with no immediate explanation about how titling practices developed and influenced each other at the European level. Groups of countries appear and disappear according to the various criteria put under scrutiny on the previous pages. Also, it is quite unsettling that European novelistic traditions, as reflected in the ELTeC collections, do not seem to be determined by the traditional language groups, nor by the vicinity between certain countries. Maybe a closer look at the circulation of books and persons may explain the contrasts and the similarities observed so far, or rather the "panachage" of tendencies we observed above. In addition, the role of best-sellers is to be better understood in relation with the titling changes already delineated. Do major works adopt the same titling practices as more obscure texts? Or, to the contrary, do well-known 19th-century novels survive their times because they have adopted different titling practices from those usually employed by their contemporaries? While the aim of this exploratory study was not to answer such questions, we hope to have proven that annotation of digital materials can help to address them in the future.