Topic modelling of titles across different languages in the 19th Century Digitised Books Collection of the British Library

Petar Soldo
r1076709
petar.soldo@student.kuleuven.be

January 5th, 2026

1 Introduction

Around 30% of titles in the 19th Century Digitised Books Collection of the British Library are not in English. This report investigates the differences and similarities between topics of English, French and German titles. We start by inspecting all languages present in book titles and describing them using simple statistical measures. Next, we briefly inspect how are genres distributed over our three target languages. Finally, we describe the preprocessing the titles, performing topic modelling and interpret the results.

The code and the dataset are available on the following GitHub repository: https://github.com/petar-soldo/IDH25-individual.

2 Titles and languages

The main data we need to perform topic modelling across languages are the titles and the language in which the titles are written. To access the titles we simply use the `'Title_clean'` column, which contains 49 395 titles[1]. There are three columns in the dataset related to languages: `'Language code (008)'`, `'Language code (041)'` and `'Title_clean_language_detected'`. The two "language code" columns are part of the original dataset and actually contain information on the languages of the work itself, not the title. `'Language code (008)'` is used to describe the language of a monolingual book, while `'Language code (041)'` is used to describe books written in several languages. There are reasons why we can't use these two columns for our research.: (i) it informs us about the language of the *content* of the book and (ii) it lacks a considerable amount of data. While the first argument could be ovelooked, keeping in mind a

[1] It seems 60 titles are missing or were lost in the cleaning process, because that's how many empty cells there are.

Source	Total length	Distinct languages
`Google`	49455	105
`langdetect`	49455	38
`'Language code (008)'`	49455	28

Table 1: Total number of values and distinct languages per source.

certain margine of error, the lack of data makes it unusable. There are only 596 non-zero values in the `'Language code (041)'` column, which is about 1.21% of all rows, while 38% of languages in `'Language code (008)'` are 'undetermined' as can be seen in Figure 1. This makes the amount of available data very small.

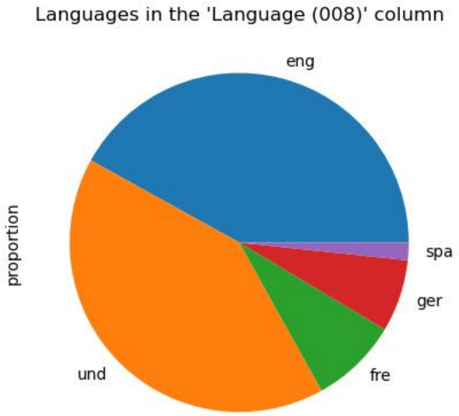

Figure 1: Languages in the 'Language 008' column

`'Title_clean_language_detected'` was created during the data cleaning and enriching process, using the Google Cloud service for language detection. While it tells us the language of every single title, it differentiates between 105 different languages. In contrast, the '(008)' contains 28 different languages. This could indicate that some of the languages were wrongly labelled, which a short inspection of the titles of some the languages confirmed.

Because of this, we add a new column of languages, using the `langdetect` module in Python. We have 38 languages, which seems much more feasible, so we will use this data. We can see all this data compared in Table 1.

According to our newly added column, six topmost languages among the titles are English, French, German, Italian, Dutch and Spanish. Of these, we will work with English, French and German, because Italian already has less

Language	Frequency	Relative frequency
English	35442	71.67%
French	4274	8.64%
German	3744	7.57%
Italian	988	2%
Dutch	850	1.72%
Spanish	799	1.62%

Table 2: Language distribution of the titles by `langdetect`.

than 1000 titles, which means the results of the topic modelling will be very hard to interpret and of dubious quality. The language distribution is illustrated in Table 2 and Figure 2. This more or less aligns with the language distribution given by the Google Cloud.

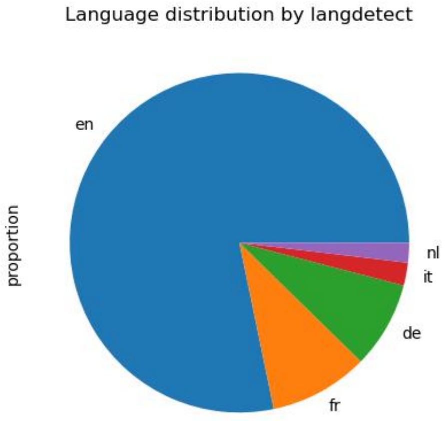

Figure 2: Language distribution of the titles by `langdetect`.

We can see from the figures that the large majority of titles are in English, which comes as no surprise, followed by a much lower, but similar number of French and German titles. This also means that topic modelling will yield less reliable results for French and German.

3 Languages and genres

Because genre is closely related to a topic of the work, we also inspect the language genre cross-tables. We should first mention that the large majority

Language	Drama	Music	Poetry	Prose
English	9.24%	5.08%	30.35%	55.32%
French	3.45%	1.06%	52.52%	42.97%
German	1.39%	4.18%	63.23%	31.19%

Table 3: Genres by languages.

of the books do not have a genre assigned. There are 30 146 (i.e. 60.96%) of titles without a genre. The relative frequencies for each genre by language can be seen in Table 3. We can see that Drama and Music have the least amount of title in all languages. In English, Prose is the dominant genre, while Poetry is attributed to majority of titles in both French and German.

One last thing worth exploring is the the length of the titles, because it could be important for the interpretation of topic modelling results. The average length of the titles is more or less the same across both languages and genre. For languages it is in the span of 15.5-16 words per title[2]. There is more variance when it comes to genres (and non-genre), with Drama having around 14.5, Poetry 15.3, Prose 12.8 and Music 19 words per title. This should not pose problems for topic modelling.

4 Topic modelling of the titles

The last step was to perform the topic analysis. The whole procedure was heavily inspired by and based on the online tutorial on topic modelling *Topic Modeling with Python (Gensim & SpaCy)* part of the Exploring Themes with Topic Modeling workshop from a series of workshops on computational text analysis by The Sherman Centre for Digital Scholarship.

The biggest divergence from the tutorial is the fact that we used the NMF model instead of the LDA model, following an idea suggested by generative AI and described in the article Non-negative Matrix Factorization (NMF) for the Grouping of Articles' Titles. In short, according to both, the NMF model should perform better on short texts like article titles. After experimenting with both models, NMF indeed proved to yield better results and performed much faster. Both of these models were available in the `Gensim` module.

We first separate the titles by languages. We preprocess each of these groups using the `SpaCy` module in Python. We tokenize and lemmatize the titles. Function words are discarded during preprocessing (only nouns, adjectives, verbs and adverbs are kept) as well as some very common words which add little of value to the topic of the title, thus creating noise (words like *edit, edition, illustrate, reprint, etc.* etc.). There are slight differences in these tweaks between the languages.

[2] More precisely, English titles have an average ogf 16, French of 15.83 and German 15.47 words.

Next, the most frequent phrases were concatenated into bigrams and trigrams to keep them as a single unit during the analysis. After this, the lemmas and the n-grams were used to build a frequency dictionary. Instead of using absolute frequencies, like in the above mentioned tutorial, we transformed the frequencies to TF-IDF values.

Finally, we used the `Gensim` model to generate 7 topics for each of the languages. Specific parameters passed can be seen commented in the code.

Here are the topics for each of the languages:

English

1. "comedy", "sketch", "new", "travel", "year", "account", "life", "letter", "revise", "author"
2. "story", "tale", "romance", "life", "love", "day", "modern", "historical", "adventure", "australian"
3. "act", "map", "translate", "year", "note", "travel", "historical", "sketch", "letter", "play"
4. "history", "note", "early", "time", "present", "biographical", "year", "translate", "town", "antiquity"
5. "poem", "canto", "song", "several_occasion", "various_subject", "book", "collection", "chiefly", "miscellaneous", "part"
6. "novel", "wife", "daughter", "only", "child", "secret", "girl", "poet", "woman", "love"
7. "tragedy", "act", "song", "historical", "write", "report", "perform", "daughter", "fall", "remark"

French

1. "voyage", "souvenir", "travers", "impression", "lettre", "récit", "journal", "pittoresque", "an", "année"
2. "historique", "étude", "notice", "recherche", "statistique", "archéologique", "description", "pittoresque", "département", "deuxième"
3. "ouvrage", "carte", "illustrer", "gravure", "contenir", "accompagner", "orner", "plan", "bois", "note"
4. "histoire", "politique", "servir", "illustration", "siècle", "règne", "république", "temps_plus_reculé_jour", "populaire", "précis"
5. "ville", "ancien", "origine", "document", "archive", "inédit", "mémoire", "province", "relatif", "note"
6. "révolution", "français", "document", "deuxième", "département", "premier", "inédit", "belge", "troisième", "siècle"
7. "pays", "environ", "voyage", "souvenir", "jour", "dessin", "auteur", "notice", "nouveau", "planche"

German

1. "deutsch", "volk", "jahrhundert", "reich", "mittelalter", "zweiter", "ausgabe", "erinnerung", "krieg", "auflage"

2. "geschichte", "quelle", "alterthum", "preussisch", "allgemein", "bd", "deutsche", "land", "kreis", "lehrbuch"
3. "historisch", "alt", "beschreibung", "land", "französisch", "revolution", "geographisch", "studie", "topographisch", "statistisch"
4. "stadt", "chronik", "quelle", "verfassung", "festschrift", "urkundliche", "archiv", "burg", "ursprung", "urkunde"
5. "beitrag", "jahr", "bild", "krieg", "herausgeben", "vergangenheit", "urkunde", "denkwürdigkeit", "studie", "vorzeit"
6. "herausgegeben", "dr", "neu", "handbuch", "zeit", "bearbeiten", "sammeln", "herausgeben", "statistik", "geographie"
7. "reise", "jahr", "karte", "abbildung", "tagebuch", "erde", "bericht", "südlich", "holzschnitt", "illustration"

5 Conclusion

We can see that the topics themselves are a bit messy and unclear and somewhat hard to interpret. Changing the number of topics did not result in clearer results, so we will try to discern some topics from these on our own..

When we look at the **English** topics, we see some fiction and romance emerging (*story, tale, romance, love, novel, secret, wife, girl, child* etc.), there is a tendency towards songs and poems (*poem, canto, song, collection*, travelling and history (*historical, adventure, australian, biographical, history,*), and drama (*comedy, sketch, act, tragedy, perform*). When we look at the **French** titles we see a strong presence of travelling and history (*histoire, voyage, souvenir, impression, journal, carte, ancien, ville, origine*), this time with an added political note (*politique, république, révolution* and even maybe some science (*recherche, statistique*. In **German** titles we find a similar situation, with a stronger emphasis on history (*geschichte, mittelalter, alt, historisch, chronik*), but also on travelling (*reise, stadt, geographisch*) and it seems to mention more local history (*deutsch, volk, reich, preussisch* and again some science (*studie, topographisch, statistisch*.

While it might be difficult to pinpoint specific topics, we can see that French and German titles are quite similar and while having an overlap with English titles in history and travel, poetry and drama are non-existing. French and especially German titles also show a faint hint of science.

We should keep in mind that most of the works are in the non-genre category. Having looked at this we might make an assumption that these books are in fact mostly about history or science, given the presence of these topics in all the titles.

It also seems that French and German books are not literary works as much as handbooks, guides and histories of nations and places.