Reducing the Underrepresentation of Transnational Writers through Biographical Event Extraction

Marco Antonio Stranisci [a,*], Viviana Patti [a] and Rossana Damiano [a]
[a] *Department of Computer Science, University of Turin, Italy*
E-mails: marcoantonio.stranisci@unito.it, viviana.patti@unito.it, rossana.damiano@unito.it

Abstract. Wikidata represents an important source of literary knowledge, which is collaboratively created and curated by a large community of users. In this archive, it is possible to find hundreds of thousands pages about writers and their works. However, Wikidata is affected by the underrepresentation of Transnational authors, as recently demonstrated. Such an issue is present at different levels, since not only Transnational writers are less in number, but there are also fewer biographical information about them in their pages. In this paper we present an approach for reducing such form of underrepresentation by automatically extracting biographical information from Wikipedia through transformers and lexico-semantic patterns, and encoding it into Wikidata semantic model. Results show that our approach allows increasing the number of biographical triples on Wikidata for all writers, rebalancing at the same time the knowledge base in favour of Transnational writers.

Keywords: Underrepresentation, Biographical Event Extraction, Wikidata

1. Introduction

In the last few years, digital media have impacted on the fruition of literary works, unfolding new opportunities for readers and scholars. New practices of reading were born through social platforms like Goodreads[1] [1], and the relationship between writers and their audience was reshaped, driving cultural and economic transformations [2]. The diffusion of large open sources of knowledge [3, 4] has engaged literary scholars in new research issues and practices [5] that rely on Semantic Web technologies [6, 7]. Such a transformation has influenced also computer scientists, who have exploited data on literary works stored in self-publishing platforms [8] and digital archives [9] to train models for Natural Language Processing [10].

The impact of digital media on the literary ecosystem is not free from flaws, though. The ecosystem of digital archives of literature is vast, but fragmented, and not all resources acknowledge the Linked Data paradigm. For instance, there is no systematic mapping of writers' pages on Wikidata onto other sources such as OpenLibrary[2] and Worldcat[3]. Since these knowledge bases have proven to be flawed by the lack of neutrality, such a limitation is even more critical as it hinders their comparative analysis. For instance, as highlighted by some recent studies, Wikidata and Wikipedia include biases [11] as well as a lack of information about non-Western people [12]. Such

*Corresponding author. E-mail: marcoantonio.stranisci@unito.it.
[1] https://www.goodreads.com/
[2] https://openlibrary.org
[3] https://www.worldcat.org

1570-0844/$35.00 © 0 – IOS Press. All rights reserved.

underrepresentation reduces the possibility of discovering, identifying, and suggesting non-Western writers and their works both to the general public and to domain experts like scholars in Digital Humanities (DH).

In order to provide a data-driven, thorough analysis of such underrepresentation, we developed the Under-Represented Writers Knowledge Graph (URW-KG), a knowledge base of writers and their works gathered from Wikidata and aligned with three external archives: OpenLibrary, Goodreads, and Google Books. The KG includes information about $194,346$ writers (including country of birth, year of birth, ethnic group and gender) and their $1,3M$ works with information about their publication history (number of editions, publishers, subjects, year and place of publication). Alongside the creation of such resource, we developed a Natural Language Processing approach for extracting biographical events from unstructured texts with the goal of reducing the underrepresentation of Transnational writers by augmenting the amount of structured biographical information about them.

To collect biographic information, we first tested the use of Lexico-Semantic Patterns (LSPs) [13], namely rules for detecting biographical expressions in texts from patterns of semantic and syntactic elements. However, such an approach showed several limitations both in terms of precision and recall, since it was not effective to deal with verb polysemy and entity disambiguation and did not account for non-verbal events. Therefore, in this paper we present a novel approach for biographical event extraction that integrates LSPs with Language Models (LMs). Language Models (LMs) serve the function of identifying the target entity of a biography with all the events related to it, while LSPs provide a knowledge-driven method for extracting and encoding results according to a semantic model.

Our method was tested on four properties encoded in Wikidata semantic model ('educated at' (P69), 'employer' (P108), 'award received' (P166), and 'nominated for' (P1411)), which are relevant for appraising the professional achievements in writers' lives. The results of our experiment show that the number of biographical events about writers on Wikidata drastically increases through biographical event extraction, passing from $126,989$ to $315,878$. The extraction also contributes to rebalancing the number of biographical information about Transnational writers compared to non-Transnational ones. In particular, the percentage of Transnational writers with at least one 'employer' triple grew from 17% to 68.7%, while non-Transnational grew from 19% to 65.8%.

This paper is structured as follows. After reviewing the related work in Section 2, in Section 3 we quantify the underrepresentation of Transnational Writers. Section 4 describes our biographical event extraction approach, while Section 5 and Section 6 respectively provide the evaluation of our methodology and the analysis of the reduction of underrepresentation after the integration of the extracted biographical events. In Section 7 we summarize our findings and describe future work.

2. Related Work

In this section, we present an overview of the studies on biases and underrepresentation on Wikidata and Wikipedia and a review of the research on biographical event detection.

2.1. Underrepresentation

The presence of racial and gender inequality on Wikidata [14] and Wikipedia is a well-known issue related to racial[4] and gender [15] gaps among the contributors. This issue has been analyzed from two main perspectives. A first line of research is mainly devoted to analyzing the presence of biases in Wikipedia pages about people belonging to groups vulnerable to discrimination. [16] performed an event extraction task on $10,412$ biographies, showing that women's Wikipedia pages contain more personal events than men's, while the latter biographies are more focused on event related to their career. [11] provided a method for systematically extracting and comparing biographies of different groups of people (e.g., by gender, ethnicity, sexual orientation) using Wikipedia categories to generate samples. They further provided a multi-dimensional index for performing this comparison, finding that general differences exits between biographies based on these groups. A second line of research is focused on quantitatively analysing under-representation on Wikipedia. [17] provided a thorough study of famous people on Wikipedia, showing that women and non-Western people are less likely to appear in a relevant number (25) of language editions of

[4]https://en.wikipedia.org/wiki/Racial_bias_on_Wikipedia

Wikipedia. Among the 100 most popular biographies, then, only 3 are about women, and 8 about non-Western people. [12] studied sociologists' Wikipedia pages, finding that non-white male and female sociologists are more prone to under-representation. [14] performed a comparison between the number of software developers, engineers, and scientists on Wikidata and the real-world population, showing that people from Europe and North America are over-represented. Our recent work on underrepresentation of non-Western writers [18] led to similar results: The ratio between African writers on Wikidata and African population is the lowest (1 *per* 374, 260 people) while Europe reaches the highest ratio in favor of writers (1 *per* 9, 136 people).

2.2. Biographical Event Extraction

Event detection and modeling have a long tradition in NLP and DH. Such interested is motivated by the aim of automatically extracting contents [19] from unstructured data like news articles [20] and social media protests [21]. Several resources are available for this task: OntoNotes [22] is multi-layer and multi-genre corpus with a semantic annotation based on the PropBank framework [23] and an annotation of coreference. TimeBank [24] and NewsReader [20] are corpora of news annotated according to the TimeML standard for the annotation of events and temporal expression [25], while LitBank [26] is a collection of 100 literary excerpts annotated according to ACE event annotation framework [19]. Even though these corpora rely on different annotation schemes and tackle different genres, they all share the same granularity of annotation, which is situated at token level. This allows an extensive reuse of these resources over different tasks.

Biographical event detection might be considered as a specialization of this task focused on modelling biographies [27, 28] and developing methods for identifying the most important events which occur in people's lives [29, 30]. The interest in such a task is shared by different communities, since it has many applications like prosopography [31], digital archives [32], and analysis of biases [16].

Despite such efforts from different communities, the number of resources and works for biographical event detection and extraction is limited, and mainly based on the adaptation of existing tools such as the NewsReader pipeline [33], the Stanford CoreNLP toolkit [34], and Semafor [35]. Relying on such tools, Russo et al. [36] detected relevant events, dates, and places about 782 people deported to Nazi concentration camps. [37] extracted linguistic motion frames [38] and places from biographies of notable people on Wikipedia. [39] identified latent personas of movie characters, namely sets of stereotypical events which define them and their role within a plot. [40] developed a system in which event detection and Semantic Web Technologies are integrated in a single pipeline, which takes as input news articles and outputs event-centric knowledge graphs.

The lack of annotated corpora for this task is an additional issue. With the exception of the work of [16], who released a set of tuples of the type event-target entity rather than a corpus of annotated documents, no resources were specifically developed for extracting biographical events. Such absence hinders the creation of benchmarking tools, preventing scholars from evaluating their approaches to biographical event extraction.

3. Measuring Underrepresentation on Wikidata

In this section we describe a quantitative analysis of Transnational writers on Wikidata. The analysis is preceded by a brief discussion of the criteria we adopted for defining the concept of 'Transnational'.

3.1. Definying Transnational Writers

For measuring the underrepresentation of Transnational writers, it is necessary to define a clear classification of people based on their ethnicity. To do so, we chose two complementary criteria for drawing such a classification: (i) the country of birth; (ii) the fact of belonging to an ethnic minority in a Western country. The former is derived from post-colonial studies [41], according to which writers from former colonies are more prone to underrepresentation since they have been historically silenced; the latter is derived from the tradition of black studies [42].

A final aspect we needed to consider in designing our classification was the terminology for referring to Transnational writers, which is not a trivial issue. Many writers, in fact, should not be considered post-colonial despite

being born in former colonies, because they belong to white minorities (e.g., J. M. Coetzee) or are the children of European or American parents (e.g., Wilbur Smith). Since Wikidata lacks coverage of people's family origins and ethnicity, we decided to adopt the broader term Transnational, which refers to people who "operated outside their own nation's boundaries, or negotiated with them" [43]. Although this representational and terminological issue is expected to affect a relatively small set of Transnational writers, such as the aforementioned Coetzee and Smith, we are aware that it is likely to attract the attention of the public, since these writers usually have a high visibility exactly because of their Western origins. However, from a quantitative perspective, we consider it acceptable for the purposes of this work, leaving to future work the task identifying reliable information sources for filling this gap.

3.2. Data Gathering and Analysis of Underrepresentation

The data gathering process was performed on Wikidata in October 2021. We first obtained from Wikidata all the $393,441$ entities of type Person (wd:Q5) with occupation (wdt:P106) writer (wd:Q36180), novelist (wd:Q6625963), or poet (wd:Q49757), with their year of birth. We then filtered out all the people born before 1808, which is a crucial year because it marks the beginning of the Spanish-American war, which can be considered as the first decolonization process. The number of writers in our collection was thus reduced to $194,346$. Finally, for each author, we collected the country of birth, ethnic group, gender, date of death, Wikipedia page, and all the works associated with them. Among the total number of writers on Wikidata $17,368$ (9%) are labeled as Transnational, while non-Transnational authors are $176,697$ (91%). The distribution is even more skewed when works are considered: of the $145,375$ works gathered from Wikidata, only $8,380$ (5.8%) are associated with Transnational Writers.

For better exploring such underrepresentation, we analyzed the distribution of Transnational writers against non-Transnational writers, grouped by gender across four generations: Silent Generation (1928-1945), Baby Boomers (1946-1964), Generation X (1965-1980), and Millennials (1981-1996). As it can be observed in Figure 1, non-Transnational male writers are predominant within the Silent Generation (66.2%) and Baby Boomers (60.9%). Surprisingly, the number of non-Transnational women progressively increases until it overcomes the number of men within the Millennials: $2,699$ (43.1%) vs. $2,605$ (41.6%). Transnational writers are significantly less across all the generations, but Transnational women writers suffer an additional lack of representation on Wikidata: there are only 508 (8.1%) male and 379 (6%) female Transnational writers on Wikidata among Millennials. Non-binary writers are the most underrepresented, regardless of their condition. In the KG there are only 23 (0.01%) non-binary Transnational authors and 146 (0.009%) non-binary among the remaining writers.

The underrepresentation of Transnational Writers is also present in Wikipedia. Out of $194,346$ writers, only $48,486$ of them have an English Wikipedia page and Transnational writers represent only the 16.4%.

Finally, such a disproportion is also reflected in the average number of biographical properties describing writers on Wikidata. There are 0.93 properties of the type 'educated at' (P69) about Transnational writers *versus* 0.987 about the others; 0.292 *versus* 0.332 properties of the type 'employer' (P108), and 0.929 *versus* 1.22 properties of the type 'award received' (P166).

Our biographical event extraction pipeline is not aimed at balancing the number of authors on Wikidata; instead, it focuses on the augmentation of their Wikidata pages with a higher amount of biographical information about Transnational writers. This could lead to a more accurate semantic representation (and deeper understanding) of their lives.

4. Biographical Event Extraction

As described in the review of the related work (Section 4.3), there is a lack of specific resources for the biographical event extraction task. In this section, we present WikiBio (Section 4.1), a corpus that we developed for training and testing biographical event classifiers, and describe a pipeline for automatically detecting events from biographies (Section 4.2), together with a strategy for automatically extracting and encoding triples from text (Section 4.3).

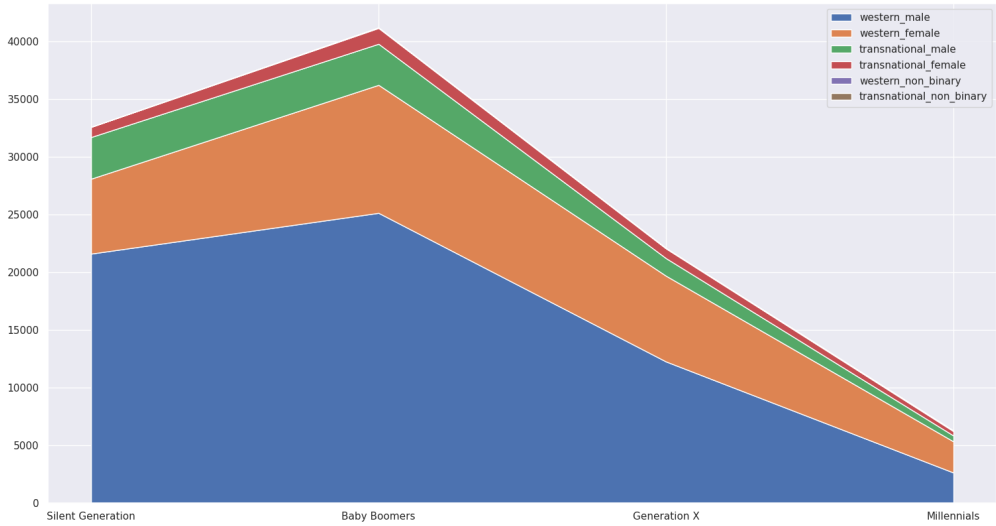


Fig. 1. Four generations of Western and Transnational writers grouped by gender

4.1. The WikiBio Corpus

Developing a corpus for biographical event detection is a crucial step for evaluating systems. The availability of training data from the bulk of corpora annotated for events, in fact, makes the creation of a large resource superfluous, since these data can be reused for the detection of biographical events; rather, a benchmark data set of biographical events that is compatible with the existing event annotation schemes is crucial for evaluating the performance of classifiers on biographic events.

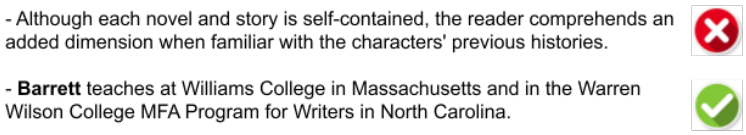
WikiBio is a corpus annotated for biographical events, composed of 20 Wikipedia biographies of African and African-American writers. The annotation was made at a token level, an approach which is widely adopted in existing resources and is well-suited for extracting fine-grained knowledge about biographical events. Event annotation is mainly based on TimeML [25] and RED [44] guidelines, according to which events may be expressed by different parts of speech and must be annotated as single tokens. Example 1 shows an examples of annotated events, of which 2 are verbs (‘dreaming’ and ‘resigned’), 1 is an adjective (‘tired’), and 1 is a name (‘routines’).

1. In mid-1958, **tired** of her daily **routines** and **dreaming** of bigger things, Head **resigned** her job.

The annotation of entities in the WikiBio corpus follows a simplified version of the GUM guidelines [45] for coreference resolution, a task whose aim is to identifies clusters of mentions of each named entity in a document. According to our guidelines, annotators are asked to annotate only the mentions of the target entity of the biography which are associated to the events in which they are involved. Example 2 exemplifies our guidelines: while a traditional coreference annotation would consider all the named entities in a document as candidates for coreference resolution (e.g., locations like ‘Bori’ and ‘Ogoniland’), our approach only considers only the entity of type person which is the subject of the biography. Moreover, not all the mentions of this entity are considered. Rather, only the mentions in which the entity is involved as a participant in an event are considered. For instance, the pronoun ‘His’ is not annotated even if it is a mention of the target entity, because it is not related to an event that involves it.

2. **Kenule Saro-Wiwa** was born in Bori [...] *His* father’s hometown was the village of Bane, Ogoniland.

Fig. 2. A graphic representation of how the entity detection task works.



4.2. Biographical Event Detection

We defined biographical event detection as a two-step task where we first finetune a pretrained Language Model (LM) to identify all the mentions of the target entity of a biography. Then, we finetune the same pretrained model for event detection. Training sets were developed by combining a portion of WikiBio and documents from existing resources that were adapted to the task. For the entity detection we adapted the coreference annotation layers of GUM [45] and OntoNotes [22]. For event detection we reused the OntoNotes [22] semantic layer, TimeBank [24], NewsReader [20], and LitBank [26].

Entity Detection. For this task we created a training set composed of a sample of 5 documents from WikiBio augmented with 100 documents OntoNotes [22] and GUM [45] coreference annotation layers. Example 3 shows the adaptation of such corpora for the entity detection task. Original annotations contain a set of mentions clusters determined by the number of named entities in each document. Therefore, in Table **??** 4 named entities can be identified: 'Fidel Castro', 'Cuba', 'Hugo Chavez', 'Venezuela'. Our adaptation consists in keeping only the mentions of the entity of the type 'person' which occurs most times in the document (Example 4), 'Fidel Castro' in this case.

3. President **Fidel Castro (1)** from **Cuba (2)** and **Hugo Chavez (3)** of **Venezuela (4)** made beautiful music together. .
4. President **Fidel Castro (1)** from Cuba and Hugo Chavez of Venezuela made beautiful music together. .

We performed a series of experiments on a distilbert-based model [46] testing different training sets on a sample of documents from WikiBio, obtaining a best F-score of 0.792. We then used this model for the detection of target entities in the $48,486$ biographies gathered from Wikipedia and filtered out all the sentences that do not contain a mention of the target entity. Such heuristic is exemplified in Figure 2 where the first sentence is removed from the data set since it does not include a mention of the writer. As a result of this process, we reduced the number of sentences to be classified for event detection from $1,486,320$ to $1,163,475$ (-21.8%).

Event Detection. Similarly to entity detection, we selected a set of existing corpora for event detection to train a DistilBert-based model in combination with our corpus. Annotated sentences from WikiBio were splitted in three subsets, used respectively for the training, for the validation, and the testing. We then augmented the training set with a combination of sentences from the following four corpora: the semantic annotation layer of OntoNotes [22], TimeBank [24], LitBank [26], and NewsReader [20]. We ran some experiments with different samples until we found a model that obtains the best performance with an F-Score of 0.859. We ran our model on all sentences containing at least one mention of a writer, thus identifying $2,533,990$ events.

4.3. Triples Extraction

For the extraction of triples from text we adopted a rule-based language based on Lexico-Semantic Patterns (LSP) [47]. LSPs are rules in which syntactic and semantic elements are combined into patterns for extracting information from text. In previous work [13], we exploited such method in combination with VerbNet classes [48] for creating a set of rules of the following form:

VerbNet class $preposition organization|location

The following is an example of how a LSP matches sentences with different verbs and preposition, and encodes them as a biographical triple:

LSP: obtain-13.5.2 from|for|at|by|in|as GPE|ORG

5. Ajunwa[5] **received** her BA **at University of California**, Davis in 2003.
6. He held a master's degree in Theatrical Directing which **he obtained from the University of Sofia**

As it can be observed, both sentences 5 and 6 match verbs which are part of the cluster *obtain-13.5.2*, namely 'receive' and 'obtain' in combination with different prepositions. This allows extracting triples with property 'educated at' (P69).

However, such an approach resulted in a high number of false positives due to three factors that show the limitation of existing VerbNet clusters [48] and of LSPs: (i) the high polysemy of verbs; (ii) the absence of a rationale for disambiguating mentions of the writers from other people mentioned in the biography; (iii) errors in the detection of entities of the type 'organization' or 'location'. Additionally, such a verb-centric approach overlooks all nominal events, which are present in high number within this type of documents.

To overcome these limitations, we strengthened the LSP approach by combining it with entity and event detection based on LMs. In order to be compliant with the Wikidata semantic model, we also limited the extraction of triples to a 2 specific properties that express relevant biographical information, namely 'educated at' (P69) and 'employer' (P108), and 2 properties related to prizes that were treated together, namely 'award received' (P166) and 'nominated for' (P1411).

For the extraction of triples we created the LSP rules from the events identified through LMs and the named entities of the type 'organization' and 'prize'. To do so, we first manually selected all the detected events with at least $1,000$ occurrences that were thematically coherent with the three types of biographical information that are the subject of our study. We then grouped them into thematic clusters to be processed by the LSP rules. For instance, events like 'studied', 'attending', 'degree', and 'graduated' were included in the same cluster associated with the property 'educated at'. We eventually created the following LSPs for the extraction of triples:

1. education_cluster + organization. The cluster is formed by the following events that were detected in biographies $57,456$ times: 'studied attended graduated degree graduating education studying learned attending attend graduate'.
2. work_cluster + organization. The cluster is formed by the following events that were detected in biographies $156,600$ times: 'worked served work working joined taught held founded professor career director teaching signed collaborated retired serving works job Professor research teacher hired founder resigned admitted position columnist scholar serve teach lectures join serves employed'.
3. prize_cluster + prize. The cluster is formed by the following events that were detected in biographies $61,269$ times: 'won awarded named appointed nominated recipient selected win Award awards winning inducted winner award'.

For the first two LSPs, we selected all sentences containing an event in the cluster and then used the spaCy[6] NER tool to identify all the entities of the type 'organization'. Since entities of the type 'prize' are not recognized in existing NER tools, we gathered a list of literary prizes from Goodreads[7] and built a regular expression pattern. At the end of this process we obtained $49,403$ triples with property 'educated at', $129,494$ with property 'employer', and $7,224$ with property 'award received' or 'nominated for'.

The example below shows an extracted triple, encoded in Wikibase ontology[8].

```
wd:Q10281199 a wikibase:Item ;
    rdfs:label ``Fernanda Young''@en ;
    wdt:P69 ``FAAP'' .
    wdt:P69 prov:wasDerivedFrom nodeID://b29989498 .
    nodeID://b29989498 rdfs:label ``Young later stated that she'd sworn
```

[5] https://en.wikipedia.org/wiki/Ifeoma_Ajunwa
[6] https://spacy.io/
[7] https://www.goodreads.com/award
[8] http://wikiba.se/ontology.

Table 1
Results of the evaluation of our biographical event extraction pipeline

Property	Precision
'educated at'	0.831
'employer'	0.801
'award received' and 'nominated'	0.983
Total	**0.872**

```
never to step on a university campus after the experiments, but
later attended Fine Arts at FAAP.''@en .
```

According to such representation, Fernanda Young (wd:Q10281199) was 'educated at' (wdt:P69) 'FAAP'. This claim was derived from (prov:wasDerivedFrom) the following sentence from her Wikipedia biography: "Young later stated that she'd sworn never to step on a university campus after the experiments, but later attended Fine Arts at FAAP".

All the extracted triples are stored on Zenodo under Creative Commons Attribution 4.0 International license (CC BY 4.0).[9]

5. Evaluation

For evaluating the precision of our pipeline for biographical event extraction, we checked a random sample of 200 triples extracted from Wikipedia biographies for each of the 3 biographical properties The results of this evaluation, listed in Table 1, show a high variety in performance between different types of triples. Triples with properties 'award received' and 'nominated' were extracted with a precision of 0.983, while the precision of triples with the 'educated at' property was 0.831. Finally, the extraction of triples with the 'employer' property was the less precise with a score of 0.801.

Notwithstanding the improvement brought about by the use of language models, the manual mapping of verbs onto event types described in Section 4.3 still introduces some ambiguity. For instance, the verb 'admit' is frequently associated to biographical event of the type 'educated at', since it highlights the acceptance at a school. In Example 7, however, it is a reporting verb. Similarly, in Example 8 the verb 'study' was found together with 'Holy Roman Empire', which is a study subject rather than an organization a student may attend.

7. Thompson **admitted** during a 1978 BBC interview that he sometimes felt pressured to live up to the fictional self that he had created, adding, "I 'm never sure which one people expect me to be."
8. While in service there, he **studied** German, and explored his growing interest in the military history of the old German states of the Holy Roman Empire.

The detection of named entities represents a second source of error. In Example 9, the NER tool identified 'Computer Science' as an organization leading to the wrong extraction of a triple with property 'employer'. Finally, a limited number of errors are due to the wrong association of the event to the target entity of the biography. In example 10, the information about employment is not related to the writer, but to her brothers.

9. He completed three years of *Computer Science* whilst performing as a solo musician at many of South Africa's most prestigious musical venues, **working** on average six nights a week between 1981 and 1985.
10. She completed her study in Chongqing University with her brothers' aid financially who were **working** for the KMT.

The high precision in extracting triples related to prizes is due to a combination of factors: (i) the low number of event types in this cluster; (ii) the lack of tools for automatically detecting entities of the type prize. Therefore, we created an LSP based on regular expressions, which led to few errors, but probably affected the recall.

[9] https://doi.org/10.5281/zenodo.7662258

Table 2
The comparison of results between a method based on LMs and LSPs and a pure LSPbased method on Transnational writers biographies.

Method	Precision	Sentences extracted
LSP	0.68	12,147
LMs + LSP	0.872	20,994

5.1. Comparison of Results with Previous Methods

A second type of evaluation is performed against our previous attempt to extract biographical events solely based on LSPs [13]. Despite the two works rely on the same rationale, namely the creation of rules based on clusters of events and certain types of named entities, they differ on the scope. The approach described here focuses only on three types of biographical events, while in [13] we adopted a broader approach. This is reflected in the number of LSPs, which are 3 in this pipeline and were 53 in [13]. In addition, the previous experiment only targeted biographies of Transnational writers. In Table 2 a comparison of the two approaches on this subset of authors is presented. As it can be observed, the current method shows a better performance in precision (0.88 *versus* 0.68) and in the number of extracted sentences (20,994 *versus* 12,147).

The higher precision may be explained by the narrower scope of the current work. The higher number of retrieved sentences is surprising, though, and can be explained by two main factors: on the one side, the higher performance of an approach based on LMs for the detection of events; on the other side, by the inclusion in our clusters of event types that are not expressed by verbs. This paves the way for the development of a semantic taxonomy of events specifically designed for this task.

6. Results and Discussion

In this section we analyze results of the extraction task in terms of how much it contributes to increase the structured biographical information about writers and reduce the underrepresentation of Transnational ones.

The absolute number of triples with 'employer' (P108) properties extracted from writers' biographies is more than 2.5 times larger than the ones recorded on Wikidata: 129,494 *versus* 48,463. The 47,397 extracted triples of the type *subject 'educated at' object* are slightly fewer than the 48,463 already present on Wikidata, while the impact of the 'nominated' and 'was awarded' combined is even smaller (7,224 *versus* 63,622).

Figures 3,4, and 5 depict the percentage of writers with at least one of the three biographical properties analyzed within our procedure. Results are aggregated by writers' condition, in order to assess the effectiveness of the augmentation in re-balancing the data set in favour of Transnational writers.

The augmentation produced by our approach has the most significant impact on the 'employer' (P108) property (Figure 4). In fact, it increases the percentage of Transnational writers with at least 1 triple with this property from 17% to 68.7%. Such improvement is higher than the one produced for non-Transnational writers, for which this property increased from 19.9% to 65.8%. This is the only case in which the augmentation results in a higher percentage of properties of Transnational respect to the other writers. The 'educated at' (P69) property is already frequent on Wikidata pages and thus obtains a smaller increase (Figure 3): The percentage of Transnational writers with such a property is the 55.3% before the extraction and 67.2% after, while non-Transnational writers go from 58% to 69%. Finally, the 'was awarded' and 'nominated' properties show the smallest increase (Figure 5). This is due to the absence of annotated corpora for Named Entity Recognition including entities of the type prize, which hinders their detection. Therefore, Transnational writers with at least 1 of these properties goes from the 35.8% to 39%, while the other writers go from 39.7% to 43.1%.

The augmentation of triples has also a significant impact on the average number of properties *per* writer, as shown in Table 3. The average amount of biographical information about employment increases to +2.89 for Transnational writers, and +2.66 for others, while education properties respectively go from 0.93 to 2 and from 0.98 to 2.07. The average amount of information about prizes shows a lower growth of +0.17 for Transnational and +0.15 for the other writers.

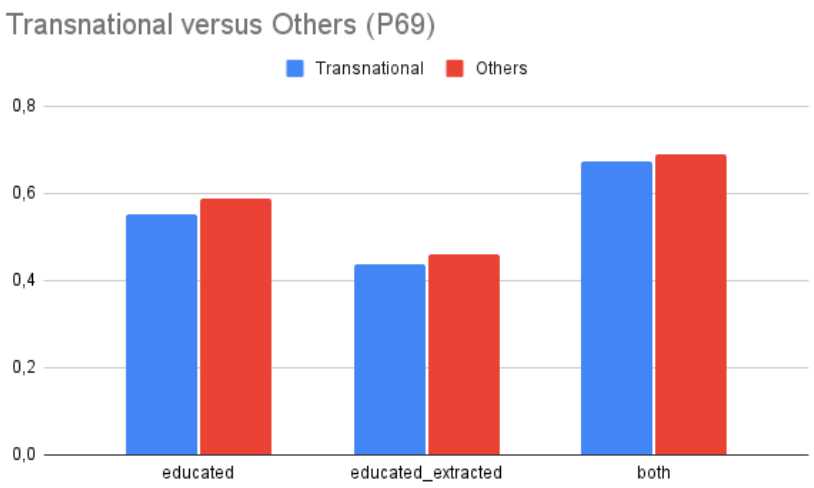


Fig. 3. The number of writers with at least one 'educated at' property on Wikidata before and after the extraction of triples.

Fig. 4. The number of writers with at least one 'employer' property on Wikidata before and after the extraction of triples.

7. Conclusion and Future Work

In this work we presented a method for extracting biographical information from Wikipedia and encoding it according to the Wikidata semantic model. This method, which is aimed at reducing the underrepresentation of Transnational writers on Wikidata, was tested on the triples based on 4 Wikidata properties: 'educated at' (P69), 'employer' (P108), 'award received' (P166), and 'nominated for' (P1411). Results show that our biographical event extraction pipeline not only increases the number of triples about writers by 3 times, but also rebalances the dataset proportionally providing a higher amount of information about Transnational writers.

Future work will focus on increasing the precision of the triple extraction step, through the implementation of a more effective Named Entity Recognition and Entity Linking approach. In addition, the biographical event extraction pipeline will be extended to a larger set of biographical properties and to different categories of people.

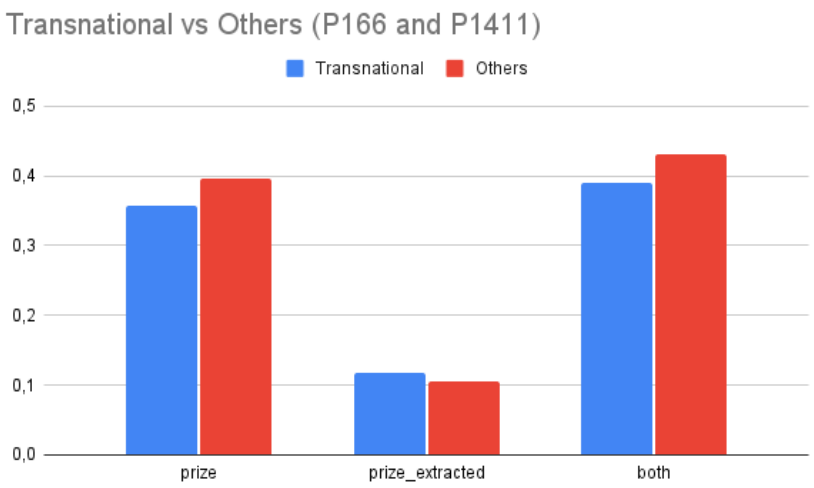


Fig. 5. The number of writers with at least one 'was awarded' or 'nominated' property on Wikidata before and after extraction of triples.

Table 3

The average number of biographical properties *per* writer

Property	**Transnational**			**Others**		
	Wikidata	Augmented	Increase	Wikidata	Augmented	Increase
'educated at' (P69)	0.93	2	+1.07	0.98	2.07	**+1.09**
'employer' (P108)	0.29	3.18	**+2.89**	0.33	2.95	+2.66
'award received' (P166) and 'nominated' (P1411)	1.02	1.19	**+0.17**	1.36	1.51	+0.15

References

[1] L. Nakamura, "Words with friends": socially networked reading on Goodreads, *Pmla* **128**(1) (2013), 238–243.

[2] D. Shaver and M.A. Shaver, Books and digital technology: A new industry model, in: *Special Issue on the Changing World of Publishing*, Routledge, 2020, pp. 71–86.

[3] B. Stroube, Literary freedom: Project gutenberg, *XRDS: Crossroads, The ACM Magazine for Students* **10**(1) (2003), 3–3.

[4] D. Vrandečić and M. Krötzsch, Wikidata: a free collaborative knowledgebase, *Communications of the ACM* **57**(10) (2014), 78–85.

[5] A. Gil and É. Ortega, Global outlooks in digital humanities: Multilingual practices and minimal computing, in: *Doing digital humanities*, Routledge, 2016, pp. 58–70.

[6] B. Tillett, What is FRBR? A conceptual model for the bibliographic universe, *The Australian Library Journal* **54**(1) (2005), 24–30.

[7] S. Brown, Scaling Up Collaboration Online: Toward a Collaboratory for Research on Canadian Writing, *International Journal of Canadian Studies* **48** (2014), 233–251.

[8] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba and S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[9] J. Brooke, A. Hammond and G. Hirst, GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus, in: *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, 2015, pp. 42–47.

[10] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).

[11] A. Field, C.Y. Park, K.Z. Lin and Y. Tsvetkov, Controlled analyses of social biases in wikipedia bios, in: *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2624–2635.

[12] J. Adams, H. Brückner and C. Naslund, Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test", *Socius* **5** (2019), 2378023118823946.

[13] M.A. Stranisci, V. Basile, R. Damiano, V. Patti et al., Mapping Biographical events to ODPs through Lexico-Semantic Patterns?, in: *CEUR WORKSHOP PROCEEDINGS*, Vol. 3011, CEUR-WS, 2021, pp. 1–12.

[14] Z. Shaik, F. Ilievski and F. Morstatter, Analyzing race and citizenship bias in Wikidata, in: *2021 IEEE 18th international conference on mobile Ad Hoc and smart systems (MASS)*, IEEE, 2021, pp. 665–666.

[15] B. Collier and J. Bear, Conflict, criticism, or confidence: An empirical examination of the gender gap in Wikipedia contributions, in: *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 383–392.
[16] J. Sun and N. Peng, Men are elected, women are married: Events gender bias on wikipedia, *arXiv preprint arXiv:2106.01601* (2021).
[17] A.Z. Yu, S. Ronen, K. Hu, T. Lu and C.A. Hidalgo, Pantheon 1.0, a manually verified dataset of globally famous biographies, *Scientific data* **3**(1) (2016), 1–16.
[18] M.A. Stranisci, V. Patti and R. Damiano, Representing the under-represented: A dataset of post-colonial, and migrant writers, in: *3rd Conference on Language, Data and Knowledge, LDK 2021*, Vol. 93, Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2021, pp. 1–14.
[19] G.R. Doddington, A. Mitchell, M.A. Przybocki, L.A. Ramshaw, S.M. Strassel and R.M. Weischedel, The automatic content extraction (ace) program-tasks, data, and evaluation., in: *Lrec*, Vol. 2, Lisbon, 2004, pp. 837–840.
[20] A.-L. Minard, M. Speranza, R. Urizar, B. Altuna, M. Van Erp, A. Schoen and C. Van Son, MEANTIME, the NewsReader multilingual event and time corpus, in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4417–4422.
[21] S. Giorgi, V. Zavarella, H. Tanev, N. Stefanovitch, S. Hwang, H. Hettiarachchi, T. Ranasinghe, V. Kalyan, P. Tan, S. Tan, M. Andrews, T. Hu, N. Stoehr, F.I. Re, D. Vegh, D. Atzenhofer, B. Curtis and A. Hürriyetoğlu, Discovering Black Lives Matter Events in the United States: Shared Task 3, CASE 2021, in: *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, Association for Computational Linguistics, Online, 2021, pp. 218–227. doi:10.18653/v1/2021.case-1.27. https://aclanthology.org/2021.case-1.27.
[22] E. Hovy, M. Marcus, M. Palmer, L. Ramshaw and R. Weischedel, OntoNotes: the 90% solution, in: *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, 2006, pp. 57–60.
[23] P.R. Kingsbury and M. Palmer, From TreeBank to PropBank., in: *LREC*, 2002, pp. 1989–1993.
[24] J. Pustejovsky, P. Hanks, R. Sauri, A. See, R. Gaizauskas, A. Setzer, D. Radev, B. Sundheim, D. Day, L. Ferro et al., The timebank corpus, in: *Corpus linguistics*, Vol. 2003, Lancaster, UK., 2003, p. 40.
[25] J. Pustejovsky, J.M. Castano, R. Ingria, R. Sauri, R.J. Gaizauskas, A. Setzer, G. Katz and D.R. Radev, TimeML: Robust specification of event and temporal expressions in text., *New directions in question answering* **3** (2003), 28–34.
[26] M. Sims, J.H. Park and D. Bamman, Literary Event Detection, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3623–3634. doi:10.18653/v1/P19-1353. https://aclanthology.org/P19-1353.
[27] J.A. Tuominen, E.A. Hyvönen and P. Leskinen, Bio CRM: A data model for representing biographical data for prosopographical research, in: *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*, CEUR Workshop Proceedings, 2018.
[28] H.-U. Krieger and T. Declerck, An OWL Ontology for Biographical Knowledge. Representing Time-Dependent Factual Knowledge., in: *BD*, 2015, pp. 101–110.
[29] F. Dib, S. Lindberg and P. Nugues, Extraction of Career Profiles from Wikipedia., in: *BD*, 2015, pp. 33–38.
[30] A. Plum, M. Zampieri, C. Orasan, E. Wandl-Vogt and R. Mitkov, Large-scale data harvesting for biographical data (2019).
[31] P. Leskinen, E.A. Hyvönen and J.A. Tuominen, Analyzing and visualizing prosopographical linked data based on biographies, in: *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*, CEUR Workshop Proceedings, 2018.
[32] A. Fokkens, S. Ter Braake, N. Ockeloen, P. Vossen, S. Legêne, G. Schreiber and V. de Boer, Biographynet: Extracting relations between people and events, *arXiv preprint arXiv:1801.07073* (2018).
[33] R. Agerri, I. Aldabe, Z. Beloki, E. Laparra, M.L. de Lacalle, G. Rigau, A. Soroa, A. Fokkens, R. Izquierdo, M. van Erp et al., Event detection, version 2 deliverable 4.2. 2, *Deliverable, NewsReader Project* (2014).
[34] C.D. Manning, M. Surdeanu, J. Bauer, J.R. Finkel, S. Bethard and D. McClosky, The Stanford CoreNLP natural language processing toolkit, in: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.
[35] D. Das, D. Chen, A.F. Martins, N. Schneider and N.A. Smith, Frame-semantic parsing, *Computational linguistics* **40**(1) (2014), 9–56.
[36] I. Russo, T. Caselli and M. Monachini, Extracting and Visualising Biographical Events from Wikipedia., in: *BD*, 2015, pp. 111–115.
[37] S. Menini, R. Sprugnoli, G. Moretti, E. Bignotti, S. Tonelli and B. Lepri, RAMBLE ON: Tracing movements of popular historical figures, in: *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 77–80.
[38] C.F. Baker, C.J. Fillmore and J.B. Lowe, The berkeley framenet project, in: *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998.
[39] D. Bamman, B. O'Connor and N.A. Smith, Learning latent personas of film characters, in: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 352–361.
[40] M. Rospocher, M. van Erp, P. Vossen, A. Fokkens, I. Aldabe, G. Rigau, A. Soroa, T. Ploeger and T. Bogaard, Building event-centric knowledge graphs from news, *Journal of Web Semantics* **37** (2016), 132–151.
[41] G.C. Spivak, Can the Subaltern Speak?, in: *Colonial discourse and post-colonial theory*, Routledge, 2015, pp. 66–111.
[42] N. Hare, The battle for Black studies, *The Black Scholar* **3**(9) (1972), 32–47.
[43] B. Boter, M. Rensen and G. Scott-Smith, *Unhinging the National Framework: Perspectives on Transnational Life Writing*, Sidestone Press, 2020.
[44] T. O'Gorman, K. Wright-Bettner and M. Palmer, Richer event description: Integrating event coreference with temporal, causal and bridging annotation, in: *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, 2016, pp. 47–56.
[45] A. Zeldes, The GUM corpus: Creating multilayer resources in the classroom, *Language Resources and Evaluation* **51**(3) (2017), 581–612.

[46] V. Sanh, L. Debut, J. Chaumond and T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108* (2019).
[47] W. IJntema, J. Sangers, F. Hogenboom and F. Frasincar, A lexico-semantic pattern language for learning ontology instances from text, *Journal of Web Semantics* **15** (2012), 37–50.
[48] K.K. Schuler, *VerbNet: A broad-coverage, comprehensive verb lexicon*, University of Pennsylvania, 2005.
[49] M.A. Stranisci, G. Spillo, C. Musto, V. Patti and R. Damiano, The URW-KG: a Resource for Tackling the Underrepresentation of non-Western Writers, *arXiv preprint arXiv:2212.13104* (2022).
[50] S.W. Brown, C. Bonial, L. Obrst and M. Palmer, The rich event ontology, in: *Proceedings of the Events and Stories in the News Workshop*, 2017, pp. 87–97.