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, 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 in several studies. In this paper we present an approach for the augmentation of structured knowledge about Transnational writers by automatically extracting biographical information from Wikipedia. The approach is based on 4 distinct modules: Coreference Resolution, Event Detection, Named Entity Recognition, and Entity Linking. Modules are combined through Lexico-Semantic Patterns, which represent a tool for mapping extracted knowledge into Wikidata semantic model. Results show that our approach dramatically increases the number of biographical triples on Wikidata for Transnational writers. Such enhanced knowledge fosters the discovery of these writers both by a general public, which can discover less known works through fairer Recommendation Systems, and researchers, who have access to more complete sources of structured information about them.

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

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

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 Transnational people [12]. Such underrepresentation reduces the possibility of discovering, identifying, and suggesting Transnational 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 World Literature Knowledge Graph (WL-KG)[13], a knowledge base of writers and their works gathered from Wikidata and aligned with three external archives: OpenLibrary, Goodreads, and Google Books. Such an alignment does not re-balance the ratio between the absolute number of Western and Transnational writers on Wikidata, but increases the information about them in structured form, since it provides insightful knowledge about their literary works.

To push forward such data augmentation, in this paper we present a pipeline for biographical event extraction aimed at increasing the number of biographical facts about Transnational writers. Our approach combines Coreference Resolution, Event Detection [14], Named Entity Recognition (NER), and Entity Linking (EL), and organizes the extracted knowledge with Lexico-Semantic Patterns (LSPs) [15, 16], namely rules for the detection of biographical expressions in texts through patterns composed of semantic and syntactic elements.

Although our pipeline is agnostic with respect to the kind of biographical information, the method was tested on four properties encoded in Wikidata semantic model: 'educated at' (P69), 'employer' (P108), 'award received' (P166), and 'nominated for' (P1411). These properties were chosen because they are the most frequent among the ones that link people to relevant career events[4]. The results of our experiment show that the number of biographical triples about Transnational writers drastically increases by applying our pipeline for biographical event extraction, passing from $17,273$ to $315,878$ ($+280\%$). This approach also led to an increase of subjects associated with at least one of the properties listed above: $+1,126$ with P69 property, $+4,392$ with P108, $+183$ with P166 and P1411.

The augmentation of biographical information and literary works about Transnational writers represents two crucial improvements. Even if they do not re-balance the absolute number of Transnational writers against Western ones, they provide additional knowledge that can foster literary research on less known authors, but can also be used for reducing this form of underrepresentation in other practical uses, like the development of fairer recommendation systems based on embeddings. [17]. Moreover, it constitutes the prerequisite for the creation of empirical researches tailored to Transnational Writers' biographies and informed on the literature on Post-Colonial and World Literature: specific patterns of biographical events, such as migrations or political activism to mention the most obvious, need to be verified on the largest possible data sets in order to become effective, specialized tools for studying the lives of Transnational writers from written sources.

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 pipeline, while in Section 5 an in-depth evaluation of results is presented. In Section 6 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 [18] and Wikipedia is a well-known issue related to racial[5] and gender [19] 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

[4]Together they are associated $637,243$ to entities of the type Person on Wikidata, according to the last data dump available on Academic Torrent (https://academictorrents.com/download/229cfeb2331ad43d4706efd435f6d78f40a3c438.torrent). More information about this statistic is provided in Section 4.4

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

to groups vulnerable to discrimination. [20] 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. [21] 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 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. [18] 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 [22] 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 [23] from unstructured data like news articles [24] and social media protests [25]. Several resources are available for this task: OntoNotes [26] is multi-layer and multi-genre corpus with a semantic annotation based on the PropBank framework [27] and an annotation of coreference. TimeBank [28] and NewsReader [24] are corpora of news annotated according to the TimeML standard for the annotation of events and temporal expression [29], while LitBank [30] is a collection of 100 literary excerpts annotated according to ACE event annotation framework [23]. 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 [31, 32] and developing methods for identifying the most important events which occur in people's lives [33, 34]. The interest in such a task is shared by different communities, since it has many applications like prosopography [35], digital archives [36], and analysis of biases [20].

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 [37], the Stanford CoreNLP toolkit [38], and Semafor [39]. Relying on such tools, Russo et al. [40] detected relevant events, dates, and places about 782 people deported to Nazi concentration camps. [41] extracted linguistic motion frames [42] and places from biographies of notable people on Wikipedia. [43] identified latent personas of movie characters, namely sets of stereotypical events which define them and their role within a plot. [44] 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 [20], 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 derives from our previous work on the World Literature Knowledge Graph [13, 22] and it is preceded by a brief discussion of the criteria we adopted for defining the concept of 'Transnational'.

3.1. Definying Transnational Writers

Before measuring the underrepresentation of Transnational writers, it is necessary to provide a set of criteria to distinguish them from the writers who are not characterized by such a condition[6]. Since this classification must be implemented in an automatic data gathering process from Wikidata, its definition must be theoretically sound and easily implemented for the task. Our classification rationales derives from post-colonial studies [45–47], according to which writers from former colonies are more prone to underrepresentation since they have been historically silenced. Relying on this body of theories, however, is not fully suitable for automatically gathering data. Gayatri Chakravorty Spivak [45] introduces the idea that writers born in former colony countries who belong to local elites must not be considered post-colonial because they were raised as Western children, and belonging to a local elite is not a property available on Wikidata. We therefore chose two criteria for implementing such a classification: (i) the country of birth. We consider Transnational writers only people who were born in a country that has been a former colony and has a Human Development Index below 0.8 [22]. (ii) The ethnicity. We manually reviewed all the ethnicities from Wikidata, keeping only the ones that represent minorities on Western countries (eg: African Americans). All writers associated with one of these ethnicities are classified as Transnational, even if they were not directly born in former colonies.[7]

A further aspect we needed to consider in designing our classification was the terminology for referring to Transnational writers, which is not a trivial issue. A few group of 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 do not rely on clearly bounded definitions like post-colonial, but to adopt the broader term Transnational, which refers to people who "operated outside their own nation's boundaries, or negotiated with them" [48] (p.9). This terminological choice prevents the exposure of our classification to a small set of famous false positive like the aforementioned Coetzee and Smith. Additionally, it is a way to signal to the user of our resource that the classification must be validated by a finer-grained validation that can be only manually performed, even if these false positives are rarely present in the Knowledge Graph.

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), and 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 (derived from P19 property), citizenship (P27), ethnic group (P172), gender (P21), date of birth (P569), Wikipedia page, and all the works associated with them (P50). Among the total number of writers on Wikidata, 17,368 (9%) are labeled as Transnational according to the definition provided in the previous section, 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 generations, but Transnational women writers suffer an additional lack of representation on Wikidata: there are

[6] All authors of this paper were born and live in Italy from Italian parents. Therefore, their foremost objective in defining Transnational writers is to avoid a colonial view on this classification

[7] The list of countries and minorities is available in a dedicated Zenodo repository: https://doi.org/10.5281/zenodo.8399935

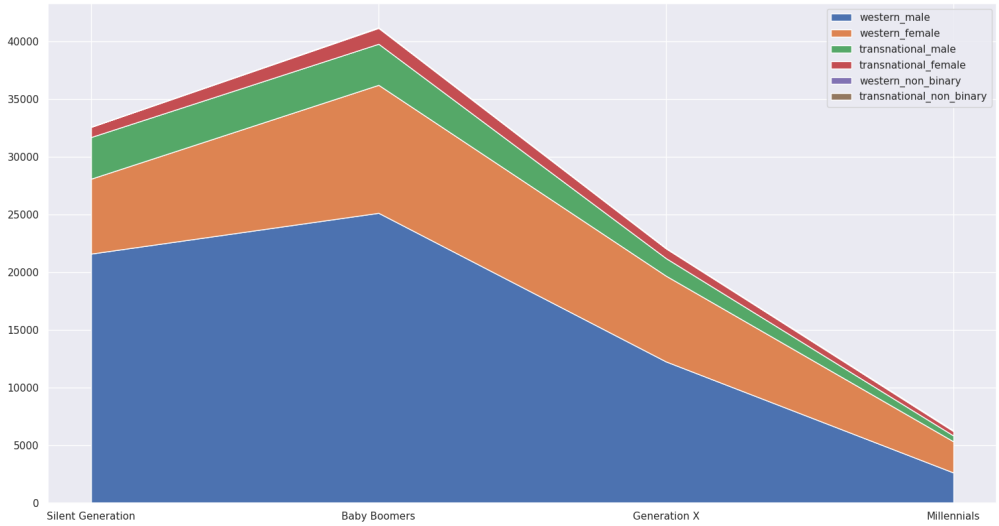

Fig. 1. The area chart depicts four generations of Western and Transnational writers grouped by gender. The blue area represents Western men, the orange Western women, the green Transnational men, the red Transnational women. Non-binary authors (Western and Transnational) are too few that do not appear in the figure despite they are computed in the dataset.

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 aims at the augmentation of the pages of Transnational writers on Wikidata with a higher amount of biographical information, with the ultimate goal of leading to a more accurate representation (and deeper understanding) of their lives.

4. Biographical Triples Extraction

In this section we present our pipeline for the extraction of biographical triples about Transnational writers, a task organized in four interrelated sub-tasks: coreference resolution of the entity target of the biography and event detection (both described in Section 4.1), Named Entity Recognition (NER) (Section 4.2), and Entity Linking (EL) (Section 4.3). While an approach for addressing the first two tasks has been addressed in a previous work [14, 49], the introduction of NER and EL within the pipeline is specific to the present work.

4.1. Biographical Event Detection

Despite the meaning of Biographical Event Detection is clear, a stable formalization of this concept aimed at computationally approaching this task has not yet been formulated. In a previous work, we defined Biographical

Event Detection as an entity-based detection task [14]: given a biography, a model must be able to identify only the events directly related to the entity target of the document. After a review of existing resources, we found that some annotated corpora may be reused for this task: TimeBank [28], OntoNotes [26], LitBank [43], NewsReader [24], and GUM [50]. Since none of them was designed for Biographical Event Detection, we developed Wikibio, a resource aimed at assessing the performance of models trained on these resources. WikiBio is a corpus 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. The annotation of events in WikiBio is mainly based on TimeML [29] and RED [51] guidelines, according to which events can be expressed by different parts of speech and must be annotated as single tokens. Example 1 shows 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 [50] 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: 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.

Once annotated, the corpus is used within a series of experiments of entity coreference resolution and event detection, based on finetuning a DistilBert-based Language Model (LM) [52] on different combinations of documents from datasets that are reusable for this task. For the entity detection we adapted the coreference annotation layers of GUM [50] and OntoNotes [26]. For event detection we reused the OntoNotes [26] semantic layer, TimeBank [28], NewsReader [24], and LitBank [30].

4.1.1. Entity Detection

For this task we finetuned a LM over different training sets composed of documents from WikiBio, and from adapted versions of OntoNotes [26], and GUM [50] and tested on a sample of documents extracted from WikiBio. Examples 3 and 4 show the adaptation of OntoNotes and GUM for the task. Original annotations contain a set of mentions clusters determined by the number of named entities in each document. Therefore, in Example 3 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. .

For the experiment we defined six combinations of 100 documents that were split in sequences with a maximum length of 128 wordpieces. Each set of sequences was grouped in a single batch, an approach that has been already adopted for performing coreference resolution with LMs [53]. 5 documents from WikiBio corpus have been used in some training set combinations, 5 in the development set, and 10 in the test set. We finetuned the public checkpoint of DistilBert[8] for 30 epochs on each combination of documents. Table 1 shows results of the experiments. As it can be observed, the only two models that obtained an F-score of at least 0.8 on the test set were the ones fully trained on OntoNotes, and composed of 95 documents from OntoNotes and five from WikiBio. We kept the model fully

[8]https://huggingface.co/distilbert-base-uncased

trained on OntoNotes since it performed slightly better thant the one trained over OntoNotes and WikiBio (+0.008 F-Score) and used it for the detection of target entities in the 48,486 biographies gathered from Wikipedia, filtering out all the sentences that do not contain a mention of the target entity. Such heuristic is exemplified in Figure **??** where the first sentence is removed from the data set since it does not include a mention of the writer.

Training | Dev | Test (30 EPOCHS)	F-Score_train	F-Score_dev	F-Score_test
GUM | WikiBio| WikiBio	0.820	0.728	0.752
GUM +WikiBio| WikiBio| WikiBio	0.819	0.728	0.753
Onto | WikiBio| WikiBio	0.896	0.782	**0.808**
Onto+WikiBio| WikiBio| WikiBio	0.846	0.774	0.800
Misc | WikiBio| WikiBio	0.824	0.766	0.792
Misc+WikiBio| WikiBio| WikiBio	0.828	0.764	0.789

Table 1

Results of entity detection experiments. In the table 3 combinations of existing resources have been used for training: GUM, OntoNotes, and a miscellaneous of both. Alongside this combinations, 3 further training sets containing a portion of WikiBio were created.

4.1.2. Event Detection

We composed 16 different training sets by selecting sentences from existing corpora annotated for event detection and tested their performance with or without adding texts from WikiBio. The training was not performed over the whole document, but on samples of sentences. The 1,691 sentences containing events annotated in our corpus were split into three sets of equal size that were used for development (563), testing (564), and training (564). The upper part of Table 2 shows the results of the model trained on each of the 16 training sets after five epochs, while in the lower part it is possible to observe the results of the models trained on the four training sets that led to a better performance. As it can be noticed, all the latter achieved an F-score on the test set above 0.85. The training set composed of sentences from Timebank and WikiBio achieved the best result (0.859), but it is worth mentioning that the model trained on a miscellaneous of all corpora and WikiBio shows the lowest drop of performance between F-score obtained on training and test set. This may signal a higher generalization of model's prediction to other type of texts.

4.2. Named Entity Recognition

Since the objects of the triples based on properties 'educated at' and 'employer' are in most cases entities of the type 'Organization', while 'award received' and 'nominated' are related to entities of the type 'Prize', we adopted two different NER strategies for their recognition.

Organizations Detection. Following the same approach adopted for Biografical Event Detection, we trained a NER model on a training set composed of sentences from OntoNotes [26] and MultiNerd [54], a semi-supervised corpus for NER bootstrapped from Wikipedia. We adapted them to transform the task from a multi-label token classification to a binary token classification, since we were only interested in recognizing entities of the type 'Organization'. The model was trained over three different combination of sentences and always achieved an F-score above 0.98.

Prizes Detection. Since prizes are not among the set of entities present in NER corpora, we created a gazetteer of prize names based on a list present on Goodreads[9] and implemented a recognition task based on regular expressions.

[9] https://www.goodreads.com/award/

Training | Dev | Test (5 EPOCHS)	F-Score_train	F-Score_dev	F-Score_test
WikiBio| WikiBio| WikiBio	0.479	0.479	0.479
Litbank | WikiBio| WikiBio	0.847	0.640	0.622
Litbank + WikiBio| WikiBio| WikiBio	0.835	0.814	0.813
Misc_01 | WikiBio| WikiBio	0.885	0.863	0.801
Misc_01 + WikiBio| WikiBio| WikiBio	0.871	0.831	0.827
Misc_02 | WikiBio| WikiBio	0.866	0.816	0.819
Misc_02 + WikiBio| WikiBio| WikiBio	0.861	0.837	**0.832**
Misc_03 | WikiBio| WikiBio	0.850	0.811	0.817
Misc_03 + WikiBio| WikiBio| WikiBio	0.844	0.839	0.831
Onto | WikiBio| WikiBio	0.950	0.800	0.790
Onto + WikiBio| WikiBio| WikiBio	0.936	0.873	0.809
Onto_mod | WikiBio| WikiBio	0.997	0.823	0.814
Onto_mod + WikiBio| WikiBio| WikiBio	0.888	0.869	0.829
Timebank | WikiBio| WikiBio	0.89	0.801	0.790
Timebank + WikiBio|WikiBio| WikiBio	0.865	0.856	0.821
NewsReader | WikiBio| WikiBio	0.453	0.479	0.479
NewsReader + WikiBio| WikiBio| WikiBio	0.467	0.479	0.479
Training | Dev | Test (15 EPOCHS)	**F-Score_train**	**F-Score_dev**	**F-Score_test**
Misc_01 + WikiBio| WikiBio| WikiBio	0.890	0.852	0.853
Misc_02 + WikiBio| WikiBio| WikiBio	0.900	0.855	0.856
Misc_03 + WikiBio| WikiBio| WikiBio	0.896	0.859	0.855
Timebank + WikiBio| WikiBio| WikiBio	0.919	0.850	**0.859**

Table 2

Results of event detection experiments. 10 training sets are based on a sample of documents selected for each of 4 reused corpora, with and without the addition of data from WikiBio. Additionally, we composed 3 miscellaneous corpora: Misc_01 (all 4 resources); Misc_02 (OntoNotes, Timebank, and Litbank); Misc_03 (OntoNotes, and Timebank). The bottom part of the Table focuses on the 4 models that achieved the best F-score after five epochs.

4.3. Entity Linking

We intend EL as a Wikification task [55] where the entities recognized in the previous part of the pipeline are linked to Wikidata. Our approach encompassed four steps:

- for each recognized entity we obtained five candidates by performing a search through Wikipedia Search APIs[10]
- for each candidate we computed a score based on three dimensions: (i) its ranking in the order in which candidates are returned from Wikipedia, normalized to a 0-1 range; (ii) the string similarity between the candidate and the recognized entity [56]; (iii) the cosine similarity between the candidate entity and the result;
- we filtered out all candidates with a score below 0.85. If after this process more than one candidate were associated to a recognized entity, we kept the one with the highest score;
- we linked the selected candidate to Wikidata.

Table 3 shows an example of how the EL pipline is structured. Given the recognized entity 'the cavendish laboratory', we obtained a set of candidates from Wikidata APIs Cavendish Laboratory, Mark Oliphant, Henry Cavendish,

[10]https://www.mediawiki.org/wiki/API:Search

Brian Josephson, James Chadwick. For each of them a score based on the order in which they are returned (column “Wikipedia"), string (column “String"), and cosine similarity is computed (column “Cosine"), and the one with the highest score (column “Score"), namely ’Cavendish Laboratory‘, is linked to its Wikidata page (with id Q181892, last column).

Recognized Entity	Candidate	Wikipedia	String	Cosine	Score	Link
the cavendish laboratory	Cavendish Laboratory	1	0.9	0.9	0.93	Q181892
the cavendish laboratory	Mark Oliphant	0.9	0.05	0.5	0.48	-
the cavendish laboratory	Henry Cavendish	0.8	0.6	0.7	0.7	-
the cavendish laboratory	Brian Josephson	0.7	0.1	0.5	0.4	-
the cavendish laboratory	James Chadwick	0.6	0.3	0.6	0.5	-

Table 3

An example of the EL pipeline. The recognized entity is present in the first column (‘the cavendish laboratory’ , while in the second the first 5 candidates obtained through the Wikipedia search API are shown. In the subsequesnt 4 columns it is possible to observe the multidimensional score associated to each candidate: the ranking it has in Wikipedia search results (column 3); the string similarity between extracted entity and candidate (column 4); the cosine similarity between their embeddings (5), and the average of the 3 score. Last column shows the WikiBase Item id of the selected candidate.

4.4. Pipeline Implementation

As mentioned in the Introduction, we focused the implementation of our pipeline to four career-relevant properties that are highly present on Wikidata: ‘educated at’ (P69), ‘employer’ (P108), ‘award received’ (P166), and ‘nominated for’ (P1411). To support our choice we gathered from the latest Wikidata dump on Academic Torrent all entities of the type Person and counted the most frequently occurring properties. Among the properties linking people with organization or other career-relevant entity types, P69 is the most frequent (323,017 triples), P108 is the second (185,826) and P166 the third (124614). P1411 has been selected for its semantic relatedness with P166.

The number of Transnational writers gathered from Wikidata and included in our dataset are 17,649, but only 7,979 of them (45) have an English Wikipedia page[11]. In order to have a benchmark to evaluate the effects of our Entity Detection model (described in Section 4.1.1), we performed the sentence segmentation of raw biographies[12] obtaining 234,606 sentences. Performing the Entity Detection step over biographies reduced the number of relevant sentences about Transnational writers to 187,082 (−20%).

For each of the filtered sentences, we detected all events as described in Section 4.1.2, thus identifying 11,876 event types that occur 216,666 times.

For the extraction of the triples from text we adopted a rule-based language based on Lexico-Semantic Patterns (LSP) [15]. LSPs are rules in which syntactic and semantic elements are combined into patterns for extracting information from text. In previous work [16], we exploited such method in combination with VerbNet clusters of verbs [57] 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[13] **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**

[11] The dataset was collected in October 2021
[12] We used the Natural Language Toolkit for this task: https://www.nltk.org/
[13] https://en.wikipedia.org/wiki/Ifeoma_Ajunwa

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 [57] 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.

Therefore we created three alternative LSPs by combining the output of the biographical detection step with this rule-based approach. We reviewed all the detected events and kept only the ones that (i) are quantitatively relevant, namely they occur at least 50 times; (ii) are thematically relevant for properties 'educated at' (P69), 'employer' (P108), 'award received' (P166), and 'nominated' (P1411). Resulting LSPs are the following:

1. 'Educated at'. The LSP is formed by the following event types, detected in 12,021 sentences from Transnational writers' biographies combined with an entity of the type 'Organization': 'studied attended degree graduated completed education studies obtained enrolled studying educated student schooling attend attending admitted PhD scholarship graduation training graduate learned trained degrees BA doctorate matriculated'.
2. 'Employer'. The LSP is formed by the following event types, detected in 34,534 sentences from Transnational writers' biographies combined with an entity of the type 'Organization': 'published worked wrote served joined founded taught work professor writing working editor writer earned career author director established translated Professor teaching invited founder lecturer write writes serving poet works job publishing resigned Director contributor serves position retirement teacher President hired columnist authored teaches research publish serve speaker teach scholar founders head researcher reporter advisor producing employed Editor Chair manager chair Chairman editor-in-chief tenure presenter translator commentator fired CEO co-founded resignation retiring recruited collaboration Lecturer directing acting'.
3. 'Award received' and 'Nominated'. The LSP is formed by the following event types, detected in 9,540 sentences from Transnational writers' biographies combined with an entity of the type 'Prize': 'won awarded appointed Award recipient nominated graduating award awards shortlisted winner winning Fellow conferred inducted nomination finalist recognised honors nominations'.

After restricting our focus on these types of events, we performed the NER step (Section 4.2), which allowed us to obtain 43,096 sentences that match at least one of the above LSPs. Example 7 shows how after this step it is possible to extract two types of relations through LSPs: 'educated at' and 'employer'.

7. After his post-graduated **studies (EVENT)**, he **joined (EVENT) Cotton College (ORG)**, Guwahati as a **lecturer (ORG)** in mathematics

As a final step of our pipeline we performed EL (Section 4.3, which allowed us to extract 37,249 triples from 27,682 unique sentences.

The example below, encoded in Wikibase ontology[14] shows a set of extracted triples[15]:

```
wd:Q10281199 a wikibase:Item ;
    rdfs:label ''Fernanda Young''@en ;
    wdtn:P214 <http://viaf.org/viaf/46422319>
    wdt:P69 wd:Q5424283.

wd:Q10281199 p:P69 s:...1 .
```

[14] http://wikiba.se/ontology.

[15] For readability, here we simplify the instantiation of the Wikibase model, according to which the property-value pair in the statement *wd:Q10281199 wdt:P69 wd:Q5424283* should be explicitly related to its provenance (namely, Wikipedia) using *prov:wasDerivedFrom* and the value possibly ranked for correctness.

```
s:...1 ps:P69 wd:Q5424283 ;
    prov:wasDerivedFrom ref:b29989498 .

ref:b29989498 rdfs:label ''Young later stated that she'd sworn
    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' (wd:Q5424283). 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).[16]

5. Evaluation of Results

The evaluation of our experimental setup focuses on two aspects: assessing the quality of our Biographical Triple Extraction pipeline described in the previous section and understanding to which extent our approach contributes to reduce the underrepresentation of Transnational writers.

5.1. Pipeline Evaluation

In order to evaluate the quality of our pipeline, we selected a stratified random sample of 584 triples (1.5% of the extracted triples) and performed a manual check of its correctness. For each example, we first checked if it was correct and, if not, we specified the source of error among the four components of the pipeline: coreference, event, NER, and EL. Table 4 shows the result of the evaluation, which on average reach the 73.9% (432 correct and 152 wrong examples). There are high oscillations between the single properties, though: property P108 reaches the worst performance with 69.8% of correct triples, P108 scored 77.9%, while P166 and P1411 together reached 95.4%.

Property	n. of examples	% correct
P69	220	77.9%
P108	342	69.8%
P166|P1411	22	95.4%
Total	**584**	**73.9%**

Table 4
Manual Evaluation of the extracted triples

When errors are broken down according to their types (Figure 2), insight about our pipeline emerges. The most common source of error is the NER classifier, which determines the 57.9% of the errors. There are two main types of NER errors: generic mentions identified as organizations, like 'Senate' in Example 8 which is wrongly labelled as the Senate of the United States of America, and misclassfied entity types, like 'Huda Darwish' in Example 9 who is a person labeled as an organization.

8. During her term in **Senate**, Ramos-Shahani was the chair of various committees.
9. **Huda Darwish** continues her success by publishing sequels.

[16] https://doi.org/10.5281/zenodo.8399935

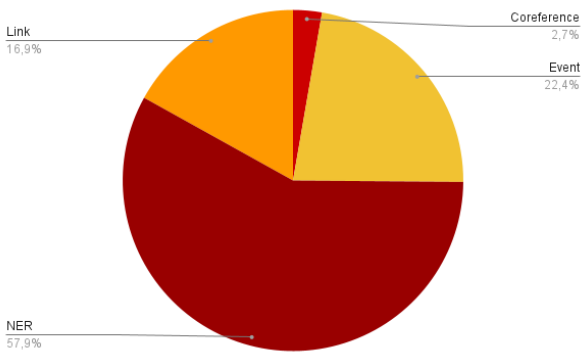

Fig. 2. Breakdown of errors emerged during the manual assessment of the pipeline. We defined 4 types of errors: (i) coreference, if the event is not associated to the target of a biography; (ii) event, if the LSP does not apply to the Wikidata property; (iii) NER, if a wrong entity was recognized; (iv) Link, if the link was incorrect.

The second major source of error is caused by event detection (22.4%), but such type of error is almost exclusively related to the 'employer' property with 32 errors out of 41 affecting these triples. A close observation of these errors shows that they are caused by their manual clustering in LSPs rather than classification errors. The polysemy of events like 'work' (Example 10) or events specific to the cultural industry like 'write' (Example 11) leads to a higher number of wrong predictions, since they can link a person to an organization (write for the Guardian) or to a cultural work.

10. since 1977, he has focused on **working** with his own band, with which he also appeared at the North Sea Jazz Festival.
11. In this year Wonder also **wrote** and produce the dance hit "Let's get serious"

EL-related errors represents 16.9% of the total. A first type of EL error derived from the lack of a specific entity on Wikidata. In example 11, 'St. Patrick's College, Asaba', missing in this knowledge source, is wrongly linked to an institution with the same name but located in Maynooth, Ireland[17]. A second type of error is the link to a disambiguation Wikidata page. The magazine Femina in Example 12 is correctly recognized by the NER, but linked to a list of potential candidates[18].

12. Okpewho attended **St Patrick's College** in Asaba, going on to university college, Ibadan, from where he earned a first-class honours degree in classics.
13. In 1918, Weber began publishing poems in the magazine **Femina** and soon began serving on the editorial board of the paper.

Summarizing, from the error analysis three main issues emerge. (i) While the Biographical Event classifier shows good performances, the mapping of events to Wikidata properties must be improved, especially for the 'employer' property. The manual organization of events in thematic clusters currently represents a bottleneck for a fully-automated Biographical Triple Extraction pipeline. Supporting this step with knowledge from resources like PropBank [27], NomBank [58], and Unified Verb Index [59] may lead to an automation of this mapping. (ii) The NER classifier must be improved in order to reduce the number of misclassified entities that are propagated to the linking step. (iii) The knowledge in Wikidata has some gaps that affect the EL step; including other sources of knowledge like DBpedia [60], CaLiGraph[61], and Google KG[19] in the EL step may result in a richer and more precise set of extracted triples.

[17] https://www.wikidata.org/wiki/Q4556206
[18] https://www.wikidata.org/wiki/Q269471
[19] https://www.google.kg/

5.2. Reduction of Underrepresentation

The second part of the evaluation focuses on the impact of triple extraction in reducing Transnational writers' underrepresentation. To do so, we created a benchmark by gathering from Wikidata all the properties of the type 'P69', 'P108', 'P166', and 'P1411' about the 7,979 Transnational writers with an English Wikipedia page and quantitatively compared them with the number of triples obtained through our triple extraction task from their biographies. A first intuitive overview of this augmentation may be observed in Figure 3, where the intersection between existing and extracted triples is represented through Venn's Diagrams. The increase of triples having P108 ('employer') as predicate is the most relevant, while P166 ('award received'), and P1411 ('nominated') grew less than the others. Such a disproportion reflects the strategies that we implemented within our pipelines (Section 4.2): given the absence of entities of the type 'prize' in NER corpora, we relied on a gazetteer and regular expressions to find them in sentences, thereby reducing the potential number of candidates for this type of triples. The opposite may be observed for P108 ('Employer'), which has been extracted more frequently but with a lower precision (Section 5.1). The general impact of our approach seems to be significant, though. As it can be observed in Table 5, the number of writers with at least one triple increases for each property: 'P69' properties grew from 4,382 to 5,508 writers (+1,126); 'P108' from 1,353 to 6,285 (+4,932); 'P166' and 'P1411' from 2,854 to 3,037 (+183).

Property	Wikidata triples	Extracted triples	Total triples
P69	7,357 (4,382)	9,369 (4,303)	13,614 (5,508)
P108	2,317 (1,353)	26,080 (6,182)	27,249 (6,285)
P166|P1411	7,599 (2,854)	1,098 (795)	8,3514 (3,037)

Table 5

The impact of our triple extraction approach on the total number of triples about Transnational writers. Numbers among parenthesis represent the amount of writers associated with at least one property of the type P69, P108, and P166|P1411. Second column shows the number of triples actually associated to the 7,979 people with a Wikipedia page; Third column the number of extracted triples; Fourth column the intersection of the two sets of data.

These results show that the even the knowledge injected from a small set of English Wikipedia pages is significant. The application of these pipeline to other sources of knowledge and its implementation to other languages may dramatically increase the number of structured information that we have about Transnational writers and other categories of people who suffer a lack of representation on Wikidata. This does not reduce their underrepresentation in absolute terms, since it can be applied also to Western writers who are significantly more in this knowledge base, but could mitigate it by providing a higher amount of biographical information about them.

6. 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 triples based on four Wikidata properties: 'educated at' (P69), 'employer' (P108), 'award received' (P166), and 'nominated for' (P1411). Results show that the pipeline significantly increases both the total number of triples with this properties and the number of Transnational Writers associated to at least one of them, thus reducing their underrepresentation. Additionally, knowledge extracted through this pipeline may be a prerequisite for the identification of biographical patterns that are specific to how Transnational people are represented or misrepresented (eg: migration, activism) in online archives. Future work will focus on increasing the quality of our biographical extraction pipeline, through a revision of NER and EL steps, and the implementation of a more effective strategy for LSPs creation. In addition, the biographical event extraction pipeline will be extended to other knowledge bases, in order to increase the information about Transnational writers.

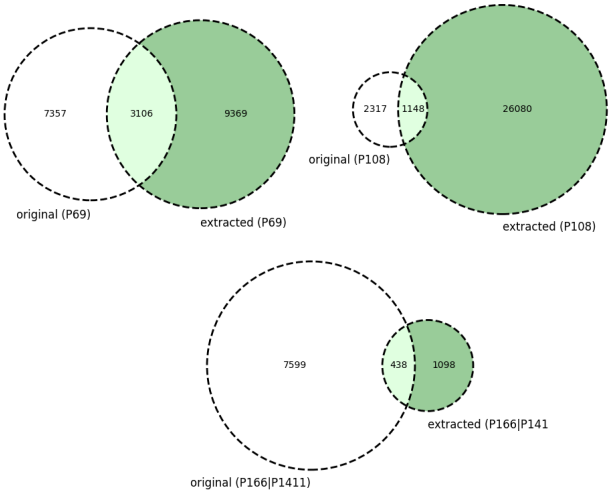

Fig. 3. A visual representation of the effect of triples extraction on the total number of triples about Transnational writers. Extracted triples are in green in diagrams.

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, E. Bernasconi, V. Patti, S. Ferilli, M. Ceriani and R. Damiano, The World Literature Knowledge Graph, in: *The Semantic Web–ISWC 2023: 22nd International Semantic Web Conference, Athens, Greece, November 6–10, 2023, Proceedings*, 2023.

[14] M.A. Stranisci, R. Damiano, E. Mensa, V. Patti, D. Radicioni and T. Caselli, WikiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12370–12384. doi:10.18653/v1/2023.acl-long.691. https://aclanthology.org/2023.acl-long.691.

[15] 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.

[16] 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.

[17] B. Hui, L. Zhang, X. Zhou, X. Wen and Y. Nian, Personalized recommendation system based on knowledge embedding and historical behavior, *Applied Intelligence* (2022), 1–13.

[18] 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.

[19] 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.
[20] J. Sun and N. Peng, Men are elected, women are married: Events gender bias on wikipedia, *arXiv preprint arXiv:2106.01601* (2021).
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] P.R. Kingsbury and M. Palmer, From TreeBank to PropBank., in: *LREC*, 2002, pp. 1989–1993.
[28] 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.
[29] 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.
[30] 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.
[31] 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.
[32] H.-U. Krieger and T. Declerck, An OWL Ontology for Biographical Knowledge. Representing Time-Dependent Factual Knowledge., in: *BD*, 2015, pp. 101–110.
[33] F. Dib, S. Lindberg and P. Nugues, Extraction of Career Profiles from Wikipedia., in: *BD*, 2015, pp. 33–38.
[34] A. Plum, M. Zampieri, C. Orasan, E. Wandl-Vogt and R. Mitkov, Large-scale data harvesting for biographical data (2019).
[35] 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.
[36] 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).
[37] 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).
[38] 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.
[39] D. Das, D. Chen, A.F. Martins, N. Schneider and N.A. Smith, Frame-semantic parsing, *Computational linguistics* **40**(1) (2014), 9–56.
[40] I. Russo, T. Caselli and M. Monachini, Extracting and Visualising Biographical Events from Wikipedia., in: *BD*, 2015, pp. 111–115.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] G.C. Spivak, Can the Subaltern Speak?, in: *Colonial discourse and post-colonial theory*, Routledge, 2015, pp. 66–111.
[46] E.W. Said, Orientalism, in: *Social theory re-wired*, Routledge, 2023, pp. 362–374.
[47] K. Dotson, Tracking epistemic violence, tracking practices of silencing, *Hypatia* **26**(2) (2011), 236–257.
[48] B. Boter, M. Rensen and G. Scott-Smith, *Unhinging the National Framework: Perspectives on Transnational Life Writing*, Sidestone Press, 2020.
[49] M.A. Stranisci, E. Mensa, R. Damiano, D. Radicioni and O. Diakite, Guidelines and a Corpus for Extracting Biographical Events, in: *Proceedings of the 18th Joint ACL-ISO Workshop on Interoperable Semantic Annotation within LREC2022*, 2022, pp. 20–26.
[50] A. Zeldes, The GUM corpus: Creating multilayer resources in the classroom, *Language Resources and Evaluation* **51**(3) (2017), 581–612.

[51] 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.
[52] 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).
[53] M. Joshi, D. Chen, Y. Liu, D.S. Weld, L. Zettlemoyer and O. Levy, Spanbert: Improving pre-training by representing and predicting spans, *Transactions of the association for computational linguistics* **8** (2020), 64–77.
[54] S. Tedeschi and R. Navigli, Multinerd: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation), in: *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 801–812.
[55] R. Bunescu and M. Paşca, Using Encyclopedic Knowledge for Named entity Disambiguation, in: *11th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Trento, Italy, 2006, pp. 9–16. https://aclanthology.org/E06-1002.
[56] J.W. Ratcliff, D. Metzener et al., Pattern matching: The Gestalt approach, *Dr. Dobb's Journal* **13**(7) (1988), 46.
[57] K.K. Schuler, *VerbNet: A broad-coverage, comprehensive verb lexicon*, University of Pennsylvania, 2005.
[58] A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young and R. Grishman, The NomBank project: An interim report, in: *Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004*, 2004, pp. 24–31.
[59] K. Kipper, A. Korhonen, N. Ryant and M. Palmer, A large-scale classification of English verbs, *Language Resources and Evaluation* **42** (2008), 21–40.
[60] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak and Z. Ives, Dbpedia: A nucleus for a web of open data, in: *international semantic web conference*, Springer, 2007, pp. 722–735.
[61] N. Heist and H. Paulheim, Uncovering the semantics of Wikipedia categories, in: *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference, Auckland, New Zealand, October 26–30, 2019, Proceedings, Part I 18*, Springer, 2019, pp. 219–236.
[62] 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).
[63] 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.
[64] N. Hare, The battle for Black studies, *The Black Scholar* **3**(9) (1972), 32–47.