University of Turin

Doctoral School of University of Turin

PhD Program in Computer Science

XXXVI cycle

PhD Dissertation

Marco Antonio Stranisci

Semantic-Aware Methods for the Analysis of Bias and Underrepresentation in Language Resources

Advisors
Rossana Damiano
Viviana Patti
Dipartimento di Informatica, Università di Torino, Italy

PhD Coordinator
Viviana Patti

Academic Year 2023-2024

Abstract Designing fair Natural Language Processing (NLP) technologies is a crucial issue in a moment of such a great interest for their many downstream applications. It has been demonstrated that these models are affected by social bias and this can results in the generation of harmful contents against minorities or in their systematic exclusion. Recent works on bias in NLP attempt to detect and mitigate these behaviours but their focus is mostly confined to the study of models' embeddings while less research has been done on bias in the documents used to train them.

My thesis contributes to research on bias by presenting a novel framework that integrates NLP and Semantic Web (SW) methods to analyze bias in datasets and annotated corpora. The framework is designed to identify three types of bias that can be found in documents: *i. representational* bias, which is the association between a given social group and a set of stereotypical features; *ii. allocative* bias, which relates to the underrepresentation of minorities in datasets; *iii. research design* bias, which comprehends all harms that may be provoked during the creation of NLP corpora.

The framework relies on two resources. The first is a network of ontologies that has been developed to model knowledge about how people are represented in social media and digital archives and to align different corpora annotated for the same topic under a common taxonomy. The second is a model for biographical event detection, which enables a thorough analysis of how people belonging to minorities and their lives are represented in datasets. The framework has been tested on three case studies: the detection of *representational* bias in English Wikipedia biographies; the detection and mitigation of *allocative* bias in Wikidata; the identification of research design bias in corpora annotated for Hate Speech (HS) and abusive language.

Results show that the framework supports the discovery of complex forms of bias in datasets and corpora. More specifically, it supports an intersectional approach to bias detection in biographies and enables the adoption of transparent and replicable approaches for the explanation and mitigation of bias in public archives and annotated corpora.

Abstract Progettare tecnologie di Natural Language Processing (NLP) fair è una questione cruciale in un momento di grande interesse per le loro numerose applicazioni pratiche. È stato dimostrato che questi modelli sono influenzati da pregiudizi sociali e ciò può portare alla generazione di contenuti discriminatori contro le minoranze o alla loro sistematica esclusione. I recenti lavori sui bias in NLP tentano di rilevare e mitigare questo fenomenon, ma si focalizzano principalmente sullo studio degli embeddings dei modelli, mentre un numero minore di ricerche si occupa dei bias nei documenti utilizzati per addestrare i modelli.

La tesi contribuisce alla ricerca sul bias presentando un nuovo framework che integra metodi di NLP e del Web Semantico per analizzare i bias nei dataset e nei corpora annotati. Il framework è stato concepito per identificare tre tipi di bias che possono essere trovati nei documenti: *i. representational bias* che consiste nell'associazione tra un determinato gruppo sociale e un insieme di caratteristiche stereotipiche; *ii. allocative bias* che si riferisce alla sottorappresentazione delle minoranze nei set di dati; *iii. research design bias*, che studia i potenziali pericoli che possono emergere durante la creazione di corpora per NLP. Il framework si basa su due risorse: *i.* una rete di ontologie sviluppata per modellare la conoscenza su come le persone sono rappresentate nei social media e negli archivi digitali e per allineare diverse risorse annotate per lo stesso argomento all'interno di un'unica tassonomia; *ii.* un modello per la rilevazione di eventi biografici, che consente un'analisi approfondita di come le persone appartenenti a minoranze e i loro percorsi biografici vengono rappresentati nei dataset. Il framework è stato testato su tre casi di studio: la rilevazione dei *representational bias* nelle biografie di Wikipedia in lingua inglese; la rilevazione e mitigazione dei *allocative bias* in Wikidata; l'identificazione dei bias di ricerca in corpora annotati per l'identificazione dello Hate Speech e del linguaggio abusivo.

I risultati mostrano che il framework supporta la scoperta di forme complesse di bias in dataset e corpora. Più nello specifico, il framework supporta un approccio intersezionale alla rilevazione dei bias nelle biografie e consente l'adozione di approcci trasparenti e replicabili per la spiegazione e la mitigazione dei bias negli archivi pubblici e nei corpora.

List of Figures

List of Tables

Contents

List Of Abbreviations

AAE: African American English

ACE: Automatic Content Extraction

API: Application Programming Interface

CRM: Conceptual Reference Model

CWC LOD: Canadian Writing Research Collaboratory Linked Open Data

DH: Digital Humanities

DS: Dangerous Speech

DOLCE: Descriptive Ontology for Linguistic and Cognitive Engineering

EL: Entity Linking

ERE: Entities, Relations, and Events Annotation

FAIR: Findable Accessible Interoperable and Reusable

FN: FrameNet

FRBR: Functional Requirements for Bibliographic Records

GPE: Geo-Political Entity

HDI: Human Development Index

HS: Hate Speech

IAA: Inter-Annotator Agreement

IE: Information Extraction

KG: Knowledge Graph

LLOD: Linguistic Linked Open Data

LM: Language Model

LSP: Lexico-Semantic Pattern

MFT: Moral Foundations Theory

ML: Machine Learning

NA: Network Analysis

NE: Named Entity

NER: Named Entity Recognition

NIF: NLP Interchange Format

NIST: National Institute of Standards and Technology

NLP: Natural Language Processing

OA: Open Annotation

ODP: Ontology Design Pattern

OLia: Ontologies of Linguistic Annotation

ORG: Organization

ORSD: Ontology Requirement Specification Document

PiM-O: People in the Media Ontology

PROV-O: Prov Ontology

O-Dang: Ontology of Dangerous Speech

RDF: Resource Description Framework

RE: Relation Extraction

RED: Richer Event Description

SemEval: International Workshop on Semantic Evaluation

SRL: Semantic Role Labeling

SW: Semantic Web

T-IPS: Time-Indexed Person Status

UR-ON: Under-Represented Ontology Network

URB-O: Under-Represented Books Ontology

URW-O: Under-Represented Writers Ontology

UVI: Unified Verb Index

VIAF: Virtual International Authority File Name

WEAT: Word-Embedding Association Test

WL-KG: World Literature Knowledge Graph

Chapter 1

Introduction

Studies on inequalities embedded in the Western-centric research tradition had a strong impact during the second half of the twentieth century in different disciplines: Kamin [1974] challenged the I.Q. test, defining it "as an instrument of oppression against the poor — dressed in the trappings of science" [Kamin, 1974]; Spivak [2015] claimed for the inclusion of the voices of traditionally silenced people in the retelling of history, arts, and literature; Crenshaw [1989] called for a rethinking of social studies aimed at accounting for intersectional forms of discrimination. Despite their heterogeneity, these approaches share the idea that research is affected by hidden discrimination that are the reflection of my societies.

These forms of discrimination may be defined as 'bias', a term that is broadly adopted to identify a deformation of research that affects certain groups. According to Hammersley and Gomm [1997], bias are procedural errors with a systematic effect on a given population and that can be made unconsciously or on purpose. The above mentioned nature of I.Q. tests falls within this definition, since it produces a systematic discrimination of the poorest through a methodological setting that claims to be objective.

Cultural Bias. Social disparities has also a strong impact on the cultural underrepresentation and misrepresentation of traditionally marginalized groups (e.g.: women, ethnic minorities). Such an issue is deeply rooted in the history of colonization [Said,

1977, Spivak, 2015] and still today affects societies in many forms: from the portrayal of minorities in news [Sui and Paul, 2017] to the Western-centric setting of school textbooks Wolf [1992]. This cultural bias is also present in allegedly-neutral sources of knowledge like Wikipedia, where a stereotypical representation of women [Sun and Peng, 2021] and a systematic underrepresentation of gender [Weathington and Brubaker, 2023] and ethnic [Adams et al., 2019] minorities represent an open issue that derives from community of contributors mostly composed of white Western men[1].

Bias in Natural Language Processing. Studies of bias in Natural Language Processing (NLP) represent a growing field of research aimed at uncovering harms that NLP technologies can produce against vulnerable minorities. However, a stable definition of this phenomenon has not been yet developed due to a wide range of related applications and conceptualizations. Hovy and Prabhumoye [2021] identify several sources of bias affecting different aspects of the NLP pipeline. Bias can be related to data collection and annotation, underlying models, or the cultural perspectives of scholars when they design their research [Santy et al., 2023]. Barocas et al. [2017] provide a taxonomy of bias based on two types of harm: *allocative* and *representational*. The former refers to the unequal distribution of opportunities among social groups. For instance, the systematic underrepresentation of minorities in datasets [De-Arteaga et al., 2019] can lead to models that systematically exclude them. Representational harms, then, are about the association of categories of people to stereotypical features [Bolukbasi et al., 2016].

An example of unconscious representational bias can be retrieved in BERT technical report [Kenton and Toutanova, 2019]. The following excerpt is the only part of the paper where pre-training datasets are mentioned. Here, Wikipedia is uncritically used as a source of training although it has been demonstrated that it is affected by gender and racial bias[2]

For the pre-training corpus I use the BooksCorpus (800M words) [Rajpurkar

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

[2] https://en.wikipedia.org/wiki/Gender_bias_on_Wikipedia https://en.wikipedia.org/wiki/Racial_bias_on_Wikipedia

et al., 2015] and English Wikipedia (2,500M words). For Wikipedia I extract only the text passages and ignore lists, tables, and headers.

Despite these proposed taxonomies of bias, a comprehensive framework for their understanding is still lacking. This leads to a number of drawbacks, as pointed out by Blodgett et al. [2020]: several works on this topic are built upon a vague definition of bias. In addition, most of them are exclusively focused on *representational* harms [Blodgett et al., 2020]. Finally, only few works study *allocative* bias in datasets [De-Arteaga et al., 2019, Zhao et al., 2019] used for training NLP technologies, while documentation practices that could support a more transparent and reliable analysis of datasets [Jo and Gebru, 2020, Bommasani et al., 2023] are still rarely adopted to report on the creation of new NLP resources.

Bias in Hate Speech Detection. The high number of resources and approaches developed for Hate Speech (HS) detection by the NLP communities gave birth to a new group of studies aimed to assess and challenge common practices behind the creation of corpora for this task. Despite sharing a common goal, it has been demonstrated that these resources are affected by different forms of bias. Sap et al. [2019] demonstrated that in these corpora messages of African-American users are more likely to be annotated as expressing HS because people belonging to this minority are not involved in the annotation. Such an issue is amplified by issues deriving from vague definitions of HS [Fortuna et al., 2020] and from the absence of established procedures for corpus annotation [Vidgen and Derczynski, 2020].

1.1 Motivation and Objectives

In this thesis I present a framework for bias detection in datasets that enhances NLP methodologies with Semantic Web (SW) technologies. The aim of this approach is to implement a methodology that jointly considers *allocative* and *representational* harms as sources of bias in datasets and annotated corpora. The adoption of SW technologies is

crucial to perform such an integration, since they enable the exploration of several forms of underrepresentation in datasets and a reliable and replicable comparison between different sources of knowledge. NLP methods are adopted to automatically integrate data from many sources of knowledge, enabling comparative analyses about the presence of biases in different public archives.

My framework represents a novelty in the field of bias detection for two reasons. First, it supports extrinsic analyses of bias in datasets that emphasize the alignment and reuse of existing linguistic resources [Cimiano et al., 2020]. Existing works have already studied the presence of bias in specific datasets [Luccioni and Viviano, 2021] while few attempts have been made to compare the presence of bias across different resources. My approach provides the development of a common semantic model under which aligning different sources of knowledge that in such a way can be comparatively assessed. The framework is also innovative because it is built upon a model for the automatic detection of biographical events that allows the systematic analysis of stereotypical representations of people belonging to vulnerable minorities and minoritized groups. Existing works on this topic adopt off-the-shelf tools for this task [Lucy et al., 2022] and are limited to gender bias [Sun and Peng, 2021]; my approach relies on a system that was natively designed for biographical event detection and adopts an intersectional approach that jointly considers different socio-demographic features to investigate the presence of bias.

1.1.1 Research Questions

This work specifically focuses on three research questions that are developed throughout the thesis and that respectively explore *representational*, *allocative*, and research *design* biases.

- **RQ 1. How can *representational* biases be detected and measured?** Existing approaches to bias detection in digital public archives rely on coarse grained analysis of textual features [Field et al., 2022] and are limited to the analysis of gender disparity [Sun and Peng, 2021]. In this thesis I try to overcome these limitations by implementing an Information Extraction (IE) pipeline for biographical

event detection in support of an intersectional analysis of English biographies. I focus on Wikipedia as a case study and use structured information from Wikidata to perform the intersectional analysis according to four axes: gender, ethnicity, age, and occupation. The approach shows that observing *representational* biases from an intersectional perspective allows for a more thorough understanding of how they affect datasets.

- **RQ 2. Which strategies can be adopted to detect and mitigate *allocative* biases?** Digital archives give access to an unprecedented amount of knowledge, which is often used to train NLP technologies. However they still include forms of underrepresentation of minorities and minoritized groups that are propagated to NLP models and real-life applications that rely on them. In this thesis I develop an approach to quantify and mitigate such underrepresentation and I test it on the case study of writers in Wikidata. The mitigation strategy is twofold: *i.* I adopt SW technologies to align Wikidata with less biased archives and *ii.* I automatically extract biographical triples from raw text. Such an approach allows for a significant increase of works associated to non-Western people and of biographical facts about them

- **RQ 3. Which measures can be implemented to discover research design biases in annotated corpora for abusive language detection?** A growing number of works show that corpora for abusive language detection are affected by issues related to annotation bias that span from the lack of rigorous annotation guidelines [Fortuna et al., 2020] to the systematic exclusion of minorities from the annotation process [Sap et al., 2019]. In this thesis I align and analyse nine English corpora for abusive language within a common semantic model that enables three types of analysis: (i) the interplay between different forms of abusiveness; (ii) the consistency of annotated corpora with legal definitions of HS; (iii) the potential impact of moral and emotional knowledge on HS detection. The analysis shows that corpora have different degree of interoperability between them and with legal

definitions of HS, paving the way for more informed data collection strategies.

1.1.2 Thesis Contribution

My thesis contributes to the field of bias detection with a novel framework based on a network of ontologies designed to represent relevant aspects for the representation of people, a system specifically developed for biographical event detection, and a class of experiments that are built upon my framework. My work resulted in the following research outputs:

1. **Semantic Modelling.** I developed an ontology network designed to represent knowledge about how people are represented in the media.
   (a) The People in the Media Ontology (PiM-O) is a mid-level ontology aligned with DOLCE semantic model [Gangemi et al., 2002] that encodes relevant concepts for the multi-faceted representation of people in social media and public archives.
   (b) The Underrepresented Writer Network (UR-ON) is a domain ontology that enables the collection and representation of non-Western writers and their works.
   (c) The Ontology of Dangerous Speech (O-Dang), a domain ontology suited for the comparison of existing corpora annotated for HS and abusive language.
2. **Corpora.** I created four corpora for the detection of biographical events, appraisal, and moral values.
   (a) Two corpora for biographical annotated for biographical event detection according to existing guidelines for event detection and co-reference resolution.
   (b) A corpus annotated for the detection of emotional appraisal.
   (c) A corpus annotated for the detection of moral values.
3. **NLP Systems.** I developed the following systems.

(a) A model for the automatic detection of biographical events.

(b) A pipeline for the extraction of biographical triples from raw text.

4. **Methods for Bias Detection.** I designed a set of bias analyses that covers *representational* and *allocative* bias in datasets and bias in annotated corpora for HS and abusive language detection.

   (a) I designed a methodology for the analysis of *representational* bias in Wikipedia English biographies that relies on biographical event detection and accounts for an intersectional perspective on this issue.

   (b) I conducted an investigation of *allocative* bias in Wikidata and implemented a methodology to reduce them through its alignment with other public archives and through the extraction of biographical triples.

   (c) I performed a comparative analysis of corpora annotated for HS and abusive language. The analysis relies on O-Dang [Stranisci et al., 2022b], which has proven to be useful for the discovery of research design bias that may affect annotated corpora.

1.1.3 Structure of the Thesis

The thesis is organized in two main parts.

1.1.3.1 Part 1: Building a Framework for Bias Detection

The first part focuses on the creation of my framework for the detection of bias. It includes an overview of actual trends in bias detection in NLP (Chapter 2), a focus on the ontology network that I built to enable the extrinsic analysis of bias in datasets (Chapter 3), and a chapter that describes the development of a system for biographical event detection (4).

Chapter 2. In this chapter I present an overview of research on bias. The review exploits Network Analysis (NA) techniques to identify the most relevant work on the

topic of bias in NLP in last years. Then, I group and describe the existing literature in sub-fields of research, empirically showing that works on *allocative* bias have a limited influence in the actual research on bias.

Chapter 3. In this chapter I present the semantic model that has been designed to systematically study *allocative* harms and annotation biases. First, I review existing work on modelling biographical knowledge and linguistic annotations. Then, I describe the ontology network that I developed.

Chapter 4. In this chapter I describe a number of approaches to biographical event detection that I developed throughout the PhD. All these approaches relies on existing methods and principles elaborated in the field of event detection, whose main lines of research are reviewed at the beginning of the chapter. Then, I present three methods: one based on Lexico-Semantic Patterns (LSP), one on Semantic Role Labelling (SRL), and one on the integration of different annotation schemes for event detection and co-reference resolution. Finally, I report on a series of classification experiments that I performed to automatically detect biographical events.

1.1.3.2 Part 2: Experimental Analysis of Bias

The second part of the thesis describes to application of my framework to the analysis of three forms of bias: *representational* bias in English Wikipedia (Chapter 5), *allocative* bias in Wikidata (Chapter 6), and research design bias in HS corpora (Chapter 7).

Chapter 5. In this chapter I present an analysis of *representational* bias that relies on biographical event detection and adopts an intersectional perspective. The analysis investigates how the interaction of four demographic axes may contribute to the spreading of bias: gender, ethnicity, age, and occupation.

Chapter 6. In this chapter I analyze the impact of *allocative* bias against non-Western writers in Wikidata and present two strategies to reduce it: *i.* a data augmentation

strategy based on the alignment of Wikidata with Goodreads and Open Library; *ii.* a data augmentation strategies based on the automatic extraction of biographical triples.

Chapter 7. In this chapter I explore bias in annotated corpora adopting three types of analyses performed through the integration of these resources within O-Dang. I first analyze how different abusive phenomena can be generalized to each other. Then I explore the compatibility of HS corpora with existing HS definitions. Finally, I analyze the impact of morality and emotional appraisal in the detection of discriminatory contents.

Chapter 8. In the last chapter, I report the obtained results and the observations emerged from our analyses. We briefly discuss the ethical issues and the limitations of the work. We individuate also the remaining challenges that we want to address in future works, and summarize the contributions to the research community in terms of findings, methodologies, resources, and publications.

Appendix A. All resources developed and discussed in this PhD thesis are publicly available. They are listed in Appendix A in Table 8.1. The table is intended to follow the Data Summary Section included in the Open Research Data Pilot (ORDP) template. Each resource is presented together with the link to its **Github repository**, information about its licensing, and its backup on **Zenodo**[3].

Appendices B and C. Annotation guidelines adopted for the creation of Bio-SRL corpus (Section 4.3) are available in Appendix B. Annotation guidelines adopted for WikiBio (Section 4.4) are available in Appendix C

Appendix D. The interview created for testing the usability of the World Literature Knowledge Graph (6) is reported in the Appendix D.

[3] https://zenodo.org/records/11195984

Part I

Building a Framework for Bias Detection

Chapter 2

Social Bias and Underrepresentation: an Overview

Despite the growth of research on bias in NLP, a stable definition of this phenomenon has not been yet developed due to its wide range of applications and conceptualizations. Hovy and Prabhumoye [2021] identify several sources of bias that affect different aspects of the NLP pipeline. Bias may be related to data collection and annotation, to models, or to the cultural perspectives of researcher when they design their research [Santy et al., 2023]. Barocas et al. [2017] provide a taxonomy of bias based on two types of harm: *allocative* and *representational*. The former refers to the unequal distribution of opportunities among social groups. For instance, the systematic underrepresentation of minorities in datasets [Dodge et al., 2021] can lead to models that systematically exclude them. Representational harms are about the association of categories of people to stereotypical features [Bolukbasi et al., 2016]. Such a polysemy of bias results in a high number of works that diverge for their objectives, approaches and focus. Additionally, it seems that *allocative* harms are often overlooked in this field of research [Blodgett et al., 2020].

In this chapter I present an overview of NLP research on social bias aimed at providing the background of this thesis. The overview has three objectives:

1. Providing a quantitative analysis of works on social bias in the NLP community

since the introduction of Language Models (LM).

2. Performing a qualitative study of major research trends on this issue.

3. Highlighting current gaps that I attempt to address with the present thesis with a specific focus on the documentation debt that affects many works in NLP.

The Chapter is organized as follows. In Section 2.2 I quantitatively analyze works on bias adopting NA techniques to identify the most relevant papers on this topic. In Section 2.2 I cluster the most influential works on bias in NLP in six research trends, in order to present a qualitative overview of existing directions in bias detection. In Section 2.3 I present an analysis of documentation debts that affect datasets adopted for training LLMs.

2.1 Bias in NLP studies: a Quantitative Analysis

As a first step of my quantitative analysis I focused on papers published at the four main conferences of the Association of Computational Linguistics (ACL): ACL, Empirical Methods in Natural Language Processing (EMNLP), the European Chapter of the Association of Computatational Linguistics (EACL), the North American Chapter of the Association of Computational Liguistics (NAACL)[1]. Since bias detection has become particularly relevant with the rise of LMs, I narrowed my investigation to a period comprised between 1 January 2017, which is when the paper *Attention is All You Need* [Vaswani et al., 2017] was published, and 31 December 2023. I identified six topics to investigate: bias, gender, race, stereotype, underrepresentation, and fairness. Topics were chosen to perform an analysis that encompasses cultural (gender, race, underrepresentation), cognitive (stereotype, bias) and technical aspects (fairness) related to the issue. For each of them I selected one or more seed terms that I used to gather all papers published in the proceedings of the main conferences or in their findings. I kept only works mentioning at least one of the seed terms listed in Table 2.1 in their title.

[1]I did not consider the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT) to restrict my focus on NLP conferences

Topic	seeds
Bias	bias
Gender	gender, sex, sexism
Race	racis*, race, intersectional
Stereotype	stereotype
Underrepresentation	under-*, imbalance, unbalance
Fair	fair, fairness

Table 2.1: The list of topics of my quantitative analysis and of the seed terms that I used for my filtering

Once a first set of papers was obtained, I manually checked them to remove all articles that mentioned the seed terms without referring to the topic of biases. For instance, all researches on positional bias in text summarization and works on grammatical gender. As a result, I obtained a list of 460 papers. Figure 2.1 depicts the distribution of the papers broken down by categories and years. As it can be observed, papers containing the term 'bias' in their title are the majority (59.7%), while the second most influential topic regards 'gender' (19.3%). 'Fairness' represents the 10.6% of the total. Works on 'underrepresentation' and 'race' combined do not reach 5%. If observed diachronically, the distribution of papers about bias shows a high increase from 2017, when only four papers were published on these topics, to 2023, when its amount raised to 178. This growth is confirmed even if it is compared against the total number of publications in ACL conferences year-by-year. Papers on 'bias' are 0.5% of all works published in 2017 against the 3.2% in 2023.

Such overall results suggest three tendencies. *i.* The first and the most intuitive is that there is a growth of interest in the topic in the last years. *ii.* The finding of Blodgett et al. [2020] about the disproportion of works about *representational* harms against *allocative* ones seems to hold true: papers containing seed terms related to underrepresentation are only nine in six years. *iii.* research works on bias are almost always focused on gender rather than race.

To further investigate the bias issue in NLP literature, I tried to quantify the influence of NLP papers on this topic with a NA approach based on a snowball sampling technique. For all the 460 collected papers I gathered all papers that they mention and all the papers

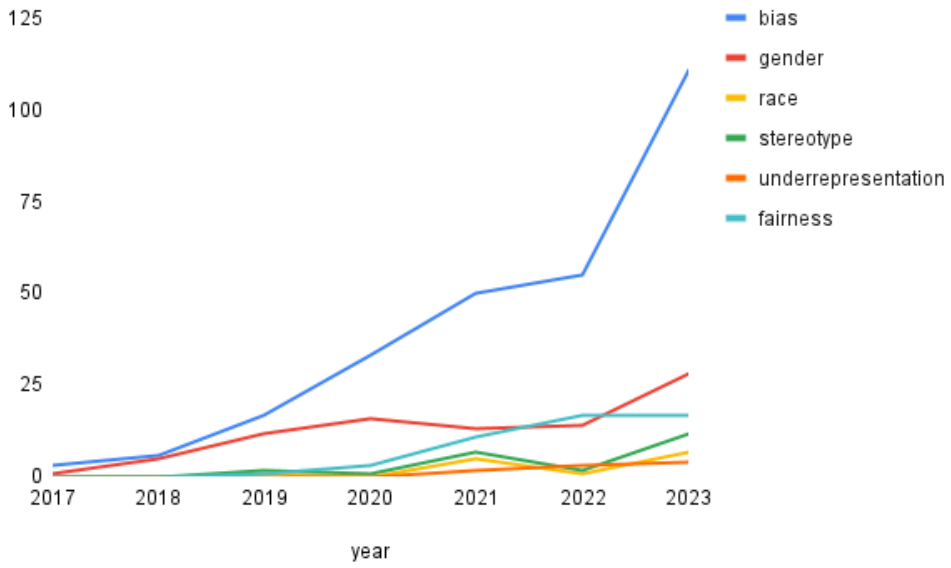

Figure 2.1: The number of papers about *bias*, *gender*, *race*, *fairness*, *stereotype*, and *underrepresentation* published in the four main ACL conferences between 2017 and 2023.

mentioning them and their references through Semantic Scholar APIs[2]. As a result I obtained a list of 119.018 articles and 350.652 links between them. Finally, I computed the eigenvector centrality of each paper [Bonacich, 2007]. Eigenvector centrality is a measure of the influence of a node within a network on the basis of its direct and indirect connections with other nodes. This enables the identification of the most influential papers on the six topics related to bias. Table 2.2 shows the 16 papers that scored an eigenvector centrality equal or higher than 0.3

A review of the most influential papers on bias confirms and expands my previous hypothesis and findings. NLP studies on this issue seems to be deeply entrenched with the spreading of LMs. BERT technical report [Kenton and Toutanova, 2019] is the most central paper in the network. *Attention is All You Need* [Vaswani et al., 2017] is also in this list together with 4 other LMs' technical reports and with Pennington et al. [2014] paper on GloVe. This ranking also confirms the skewness of bias studies toward gender representation: six papers in this list are about this topic. Among them, it is

[2] https://www.semanticscholar.org/product/api

Paper	Centrality
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Kenton and Toutanova [2019]	1
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings Bolukbasi et al. [2016]	0.79
Semantics derived automatically from language corpora contain human-like biases Caliskan et al. [2017]	0.61
Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints Zhao et al. [2017]	0.5
Attention is All you Need Vaswani et al. [2017]	0.48
Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods Zhao et al. [2018]	0.48
Language Models are Unsupervised Multitask Learners Radford et al. [2019]	0.47
Gender Bias in Coreference Resolution Rudinger et al. [2018]	0.47
RoBERTa: A Robustly Optimized BERT Pretraining Approach Liu et al. [2019]	0.44
Language Models are Few-Shot Learners Brown et al. [2020]	0.41
GloVe: Global Vectors for Word Representation Pennington et al. [2014]	0.41
Language (Technology) is Power: A Critical Survey of "Bias" in NLP Blodgett et al. [2020]	0.36
Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them Gonen and Goldberg [2019]	0.34
Deep Contextualized Word Representations Peters et al. [2018]	0.31
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer Raffel et al. [2020]	0.31
The Risk of Racial Bias in Hate Speech Detection Sap et al. [2019]	0.30

Table 2.2: The 16 papers that scored an eigenvector centrality equal or greater than 0.30

possible to find the paper that originally presented the Word-Embedding Association Test (WEAT) [Caliskan et al., 2017], which represents one of the first attempts to measure stereotypical associations between gender and professions. Remaining articles are divided in two groups: one is aimed at exploring gender bias in word embeddings, the other explores the issue in co-reference resolution. The only influential paper about race is Sap et al. [2019] seminal work on racial biases in HS detection. Finally, it is worth mentioning the presence of Blodgett et al. [2016] paper, which is the only survey in this list and the only work mentioning *allocative* harms.

2.2 Research Trends in Bias Detection: a Qualitative Analysis

In this section I provide a qualitative analysis of bias in NLP, by clustering them in 6 research trends. The analysis is based on the method that I adopted for identifying highly influential papers about bias detection in Section 2.1. In order to expand the analysis to a higher number of works, I selected all the 50 articles that scored an eigenvector centrality equal or higher than 0.2 and filtered all those that are not relevant for my topic or that focuses on bias in other fields like computer vision. In general, it is possible to identify a consistent number of papers that are focused on very specific bias issues and that jointly present a part devoted to bias detection and one to its mitigation. A smaller number of works is instead devoted to survey the literature review. As a last step of my analysis, I read all abstract and organized them in six groups.

Measuring Bias. A first group of works is aimed at the definition of metrics and methodologies for measuring Bias. Through the lens of these work it is possible to observe the recent technological evolution of NLP technologies. Caliskan et al. [2017] proposed metrics for word embeddings, May et al. [2019] for sentence encoders, Kurita et al. [2019] for contextualized word embeddings, and Sheng et al. [2019] for generative models.

Debiasing Embeddings. This group of work focuses on approaches for reducing bias at the level of embeddings. Most of them are on gender and related to traditional word embeddings (e.g.: [Bolukbasi et al., 2016, Zhao et al., 2017]). Among them, there is space for debiasing racial stereotypes in word embeddings [Manzini et al., 2019] and for bias mitigation in contextualized word embeddings [Zhao et al., 2019]. Gonen and Goldberg [2019] provide an experimentally-grounded critique to debiasing techniques demonstrating that they only have a superficial impact on embeddings.

Debiasing Classifiers. Under this research trend it is possible to group all works that operate at any point in the text classification pipeline. Their focus is to identify and reduce the emergence of latent features that may be learned by a classifier for abusive language, which may result in a systematic classification of words related to minorities as triggers of abusive messages. Dixon et al. [2018] proposed an approach based on adversarial learning to mitigate unintended biases in toxic message classification. Park et al. [2018] tested two data augmentation approaches to produce fairer classification of sexist tweets: one based on including additional datasets in the training set, the other based on a gender-swap augmentation techniques. Sap et al. [2019] discovered that annotated corpora for HS detection systematically associate African American English (AAE) to hatred contents.

Developing Benchmark Datasets. This research trend includes all works that released a new benchmark for bias detection. Among them there are WinoBias [Zhao et al., 2018], a dataset for co-reference resolution aimed at testing stereotypical associations between women and certain types of profession; StereoSet [Nadeem et al., 2021], a dataset for testing the presence of stereotypical knowledge in LMs; RealToxicityPrompt [Gehman et al., 2020], a list of annotated prompts that is intended to measure the toxicity of text generated by LMs; The Equity Evaluation Corpus [Kiritchenko and Mohammad, 2018], specifically designed to measure gender and racial biases in models trained for SA.

Surveying Bias Literature. Three surveys and secondary research are among my list of most influential papers. Sun et al. [2019] proposed a neutral survey aimed at identifying research direction for gender bias detection and recognition. On the contrary, the work of Blodgett et al. [2020] strongly criticized the state of the art in this field, while Bender et al. [2021] problematize the processes behind the creation of LLMs. These works are particularly relevant for the topic of bias in NLP, since they emphasize the issue of research design bias that are often overlooked and more difficult to be analyzed and measured.

Documenting Datasets. Only two influential papers share a focus on documentation issues. Both proposed documentation templates for the documentation of datasets [Bender and Friedman, 2018] and LMs [Mitchell et al., 2019], which are now adopted by a growing number of papers that release resources for NLP. However, this line of research is mostly limited to annotated corpora. Research aimed at assessing datasets adopted for the pretraining is confined to a few number of works on the topic [Luccioni and Viviano, 2021, Dodge et al., 2021]. It is however worth mentioning the release of a growing number of pre-training corpora that account for this issue, such as the ROOTS corpus [Laurençon et al., 2022] and the Dolma corpus [Soldaini et al., 2024], and recent attempts to assess the pipeline behind the development of LLMs. [Bommasani et al., 2023]

2.3 The Documentation Debt of Pretraining Datasets

The last part of my overview provides a pilot analysis of documentation debts that encompasses the creation and description of pretraining datasets. This issue has been raised by Jo and Gebru [2020], who advocated for the adoption of best practices from archival sciences, in order to provide a more transparent documentation of data used in Machine Learning (ML) and to mitigate harms that may be perpetrated by LLMs. However, few works focused on this issue: Luccioni and Viviano [2021] performed an analysis of the Common Crawl dataset[3], discovering the presence of significant levels

[3] https://commoncrawl.org/

of hatred contents. Field et al. [2022] provided a *post-hoc* analysis of the C4 corpus [Raffel et al., 2020], discovering the underrepresentation of messages written in non-standard English. Gururangan et al. [2022] demonstrated that LLMs trained on existing pretraining datasets tend to attribute high quality scores only to documents written by white and well educated people. Despite these preliminary efforts, systematic analyses of the documentation of pretrained datasets have not been yet performed. My analysis focuses on datasets adopted to train LMs that are ranked in the baseline of SuperGLUE [Wang et al., 2019]. In Section 2.3.1 I will report on LMs selected for the analysis. Section 2.3.2 describes how pretrained datasets are documented in LLMs reports.

2.3.1 Language Models

The first step of my analysis consisted in reviewing all LMs which obtained a score above the baseline of SuperGLUE [Wang et al., 2019]. On 30 November 2023, there were 25 submissions outperforming the threshold, 16 of which were accompanied by a technical report or a paper. I only retained reports with a reference to the dataset creation process, thus obtaining a total number of 10 LMs: ST-MoE-32B [Zoph, 2022], Turing NLR v5 [Bajaj et al., 2022], ERNIE 3.0 [Sun et al., 2021], PaLM 540B [Chowdhery et al., 2022], DeBERTa [He et al., 2020], T5 [Raffel et al., 2020], NEZHA-Plus [Wei et al., 2019], RoBERTa [Vu et al., 2022], ADAPET (ALBERT) [Lan et al., 2019], GPT-3 [Brown et al., 2020]. For each report I identified a series of features to analyze in my overview (Table 2.3):

- Which are the datasets used for the pre-training process.
- How they are referenced within the report. The reference may be 'direct', if the provenance of the dataset is specified in the report Bajaj et al. [2022]; 'indirect', if authors refer to previous works where data creation is further specified [Sun et al., 2021]; 'internal', if data creation is part of the paper itself [Raffel et al., 2020].
- Which space is given to dataset description: 'mentioned', 'paragraph', 'section', 'appendix'.

Model	Reference	Collocation	Filtering	Assessment	Released by
ST-MoE-32B [Zoph, 2022]	indirect	mentioned + appendix	None	No	Google
METRO-LM [Bajaj et al., 2022]	direct	section	None	No	Microsoft
ERNIE 3.0 [Sun et al., 2021]	indirect	paragraph	None	No	Baidu
PaLM 540B [Chowdhery et al., 2022]	internal	section + appendix	None	Yes	Google
DeBERTa [He et al., 2020]	direct	mentioned + appendix	None	None	Microsoft
T5 [Raffel et al., 2020]	internal	sections	stopwords language	Yes	Google
NEZHA-Plus [Wei et al., 2019]	direct	paragraph	None	No	Huawei
RoBERTa [Vu et al., 2022]	direct	paragraph	None	No	Facebook
ADAPET (ALBERT) [Lan et al., 2019]	indirect	mentioned	None	No	Google Toyota
GPT-3 [Brown et al., 2020]	indirect	paragraph	deduplication low quality removal	Yes	OpenAI

Table 2.3: All LMs submitted to SuperGLUE that obtained a score above the baseline and are accompanied with a description of datasets adopted for pre-training

- Which filtering strategies have been applied to datasets.
- Which type of assessment has been performed over the dataset.
- Which institution released the model.

Datasets. During the first part of my data collection I was able to identify 11 datasets used for pretraining. I first retrieved the bibliographical information about each of them, if it exists, and then checked its online availability. 7 datasets out of 11 are publicly available and can be accessed through HuggingFace [Lhoest et al., 2021], where they are accompanied with a data card [Pushkarna et al., 2022]. The procedure for retrieving the Gutenberg Corpus [Gerlach and Font-Clos, 2020] is available on its project folder on

Github[4] and can be used to recreate it from scratch; STORIES corpus [Trinh and Le, 2018] can be accessed upon request by contacting its authors. GLaM [Du et al., 2022] and Infiniset Thoppilan et al. [2022] are not publicly available. Within this research I did not analyze Common Crawl[5], even if it is mentioned in some LMs papers. This is because it is not a proper dataset, but rather a live service that crawls content from the web.

After collecting all datasets, I obtained some statistics about their size, the type of documents they are composed of, and the available metadata about documents. Table 2.4 contains a general overview of the collected information about datasets.

2.3.2 Datasets Overview

In this section, I present an overview of pretraining datasets focused on how they are described within technical reports of LMs that have been submitted to SuperGLUE (Section 2.3). Such an overview is inspired by the work of Jo and Gebru [2020] who extensively argued in favour of applying archives curation strategies to ML datasets. Therefore, the focus of this section is to identify the presence and the extent of the documentation of pretraining datasets. The analysis is organized in three subsections: the first is devoted to describing a series of general attitudes towards dataset documentation that emerge from all technical reports (Section 2.3.2.1); I then describe the adoption of datasets which were not specifically created for training LMs (Section 2.3.2.2); finally I describe how datasets specifically created for pretraining are built and documented (Section 2.3.2.3).

2.3.2.1 Dataset Documentation

While significant differences emerge between technical reports, there are some attitudes shared by all of them. Even if all technical reports provide a breakdown of data sets used for LMs training process, none of them make available the samples that were actually

[4] https://github.com/pgcorpus
[5] https://commoncrawl.org/

Dataset	N. of records	Type	Availability	Metadata	Models
Book [Zhu et al., 2015]	$740K$	sentence	yes	no	ERNIE 3.0 METRO-LM RoBERTa DeBERTa GPT-3
C4 [Raffel et al., 2020]	$365M$	document	yes	url, timestamp	PaLM 540B ST-MOE T5
CC-News [Hamborg et al., 2017]	$708K$	document	yes	domain,title, description, url, image_url.	ERNIE 3.0 METRO-LM RoBERTa
Discovery [Sileo et al., 2019]	$3.4M$	sentence	yes	no	ERNIE 3.0
GLaM [Du et al., 2022]	–	–	no	–	PaLM 540B
Gutenberg Corpus [Gerlach and Font-Clos, 2020]	$50K$	document	yes	title	ALBERT
Infiniset [Thoppilan et al., 2022]	–	–	no	–	PaLM 540B
OpenWebText [Radford et al., 2019]	$8M$	document	yes	url	METRO-LM RoBERTa DeBERTa GPT-3
Reddit [Völske et al., 2017]	$3.8M$	document	yes	title, id, author summary subreddit	ERNIE 3.0
STORIES [Trinh and Le, 2018]	document	–	upon request	–	METRO-LM DeBERTa RoBERTa
Wikipedia	$6,4M$	document	yes	id, url, title	ERNIE 3.0 METRO-LM RoBERTa DeBERTa ALBERT GPT-3

Table 2.4: The 11 datasets used for training the best 10 LMs according to SuperGLUE.

used during pretraining. This affects reproducibility, since it is not possible to replicate the pretraining without this information. The issue is amplified by the lack of specific reference to datasets used within this task. Several examples of such an attitude can be made: in the BERT report Kenton and Toutanova [2019] mentioned Wikipedia without referencing the actual dump that has been used; in the GPT-3 paper [Brown et al., 2020] two corpora of books were described as part of the curated datasets used to filter documents, but none of them is referenced.

Dataset descriptions vary among reports, showing that some common practices emerged through the years. In Table 2.3 it is possible to observe a growing attention on this aspects as papers with an extensive analysis of datasets in their appendix only appears from 2020. The Table also shows how companies internally manage data documentation: once a dataset is documented for a LM, the resource is reused for other models without additional analysis (see Zoph [2022], Sun et al. [2021]) and datasets are referenced indirectly. This seems to show a static idea of data curation practices, which are not presented as something that can be updated and improved.

Finally, it is worth mentioning that explicit descriptions of data filtering strategies aimed at guaranteeing data quality are only present in the technical reports of T5 [Raffel et al., 2020] and GPT [Brown et al., 2020].

2.3.2.2 Datasets that were not created for training LMs

Wikipedia. Wikipedia is the first corpus that has been used for pre-training a LM [Kenton and Toutanova, 2019] together with the Book Corpus [Zhu et al., 2015]. It is considered an authoritative source of quality texts and used as a benchmark for filtering texts for pretraining. Chowdhery et al. [2022] trained a hash-based linear classifier on Wikipedia and other selected sources to only retain good quality documents. However, this dataset is never precisely referenced in LMs papers. Despite being subject to a constant change due to its crowd sourcing nature, authors never mention which snapshot has been used for pretraining. Additionally, long-standing studies on the presence of cultural [Callahan and Herring, 2011] and gender bias [Wagner et al., 2015] in Wikipedia

pages are never considered when the risk mitigation strategy is assessed. The dataset is available on HuggingFace[6]

Book Corpus [Zhu et al., 2015]. This corpus is widely adopted for pretraining, but it suffers several reference issues. Its original version, proposed by Zhu et al. [2015] as a resource to support alignment between books and their movie releases, is not available anymore. Some versions of the corpus are still accessible, but poorly referenced. For instance, it is not clear which are the differences between bookcorpus[7] and bookcorpusopen[8], both hosted on HuggingFace and referenced to the same paper. In some LMs reports [Brown et al., 2020] a 'book corpus 2' is mentioned, but not referenced. Bandy and Vincent [2021], who provided a *post-hoc* datasheet [Gebru et al., 2021] for this dataset, described Book Corpus 2 as a portion of the Pile dataset [Gao et al., 2021], which is never mentioned as a source for these documents, though. Additionally, from the analysis several vulnerabilities emerge: copyright infringement, presence of duplicates, and genre imbalance.

CC-News [Hamborg et al., 2017]. CC-News is a snapshot of news texts derived from the CommonCrawl initiative[9]) that has been exploited for training RoBERTa, ERNIE 3.0 and METRO-LM. On HuggingFace, the corpus is attributed to Hamborg et al. [2017][10] who presented a scraper rather than a collected dataset. As a matter of fact, all LMs technical reports refer to Nagel [2016] which is actually a dead link. Additionally, Facebook's LMs [Vu et al., 2022, Bajaj et al., 2022] have been trained on a version of CC-News that has never been made public.

Discovery [Sileo et al., 2019]. The dataset was created for a discourse marker prediction task and is composed of $1.74M$ pairs of sentences linked by a discourse marker.

[6] https://huggingface.co/datasets/wikipedia
[7] https://huggingface.co/datasets/bookcorpus
[8] https://huggingface.co/datasets/bookcorpusopen
[9] https://commoncrawl.org/
[10] https://huggingface.co/datasets/cc_news

The dataset is derived from an existing resource [Panchenko et al., 2019] by retaining only pairs in which the second sentence starts with a word followed by a comma (e.g.: subsequently). The corpus has been used to train ERNIE 2.0 [Sun et al., 2020] and for its subsequent version [Sun et al., 2021]. The corpus is available on HuggingFace[11]

Standardized Project Gutenberg Corpus [Gerlach and Font-Clos, 2020]. The corpus has been conceived as the first systematic attempt to provide a standardized version of the Project Gutenberg initiative, aimed at avoiding potential biases caused by its sampling and exploring several phenomena like language variation and distant reading. Rather than releasing a corpus, authors published a procedure to recreate it from scratch[12].

Reddit [Völske et al., 2017]. Despite being used to train ERNIE 3.0 [Sun et al., 2021], a clear reference to this dataset is not present in the paper. On HuggingFace it is possible to access a public resource created as a benchmark for the abstractive summarization task[13].

2.3.2.3 Datasets created for training LMs

C4 [Raffel et al., 2020]. The Colossal Clean Crawled corpus has been created during the development of T5 [Raffel et al., 2020]. It is composed of $365M$ documents gathered from Common Crawl and filtered with a list of stop words to remove dangerous content[14]. Additionally, poor quality documents have been removed through a series of heuristics (e.g.: *lorem ipsum* pages, documents with less than 5 sentences). The corpus is available on HuggingFace[15] and has been used to train subsequent LMs developed by Google.

[11] https://huggingface.co/datasets/discovery
[12] https://github.com/pgcorpus/gutenberg
[13] https://huggingface.co/datasets/webis/tldr-17
[14] https://github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words
[15] https://huggingface.co/datasets/c4

OpenWebText [Radford et al., 2019] . OpenWebText is the dataset used as part of the training of GPT-2 Radford et al. [2019] and GPT-3 [Brown et al., 2020]. The resource has been created by selecting all outbound links from Reddit that achieved at least three likes. Together with Wikipedia and Book corpus, the dataset represents a baseline to ensure the quality of other documents gathered from Common Crawl. OpenWebText is available for download on its Github page[16].

STORIES Trinh and Le [2018] . This corpus has been created to train models aimed at solving the Winograd Schema Challenge [Levesque et al., 2012]. Documentation of the data gathering process is not present in the paper and the dataset is not publicly available. Nevertheless, STORIES may be obtained contacting its authors.

GLaM [Du et al., 2022] and Infiniset Thoppilan et al. [2022]. These corpora have been both developed by Google. GLaM is the best documented dataset, since it comes with a datasheet and a thorough analysis of its structure and potential biases. Infiniset is presented as its derivation. Unfortunately, none of them is publicly available.

From this overview of datasets, it is possible to observe the growth of a new attitude on dataset documentation, which may be the result of foundational work by Gebru et al. [2021] and Bender et al. [2021]. More recent works are more likely to include detailed descriptions of datasets and their assessment. However, this attitude does not result in a disclosure of the actual documents on which LMs are trained. In no case it is possible to replicate training experiments since the actual snapshots of documents used during pretraining are never available and always generically referenced. This seems to show a formal rather than a substantial commitment of AI companies in sharing transparent information about datasets.

[16] https://skylion007.github.io/OpenWebTextCorpus/

2.4 Conclusion of the Chapter

In this chapter I presented an overview of works on bias with the aim of identifying current trends and pitfalls about this field of research. The overview shows that *allocational* bias are still little explored by the NLP community and that the lack of documentation embraces almost all datasets used for LLMs pretraining. In general, the analysis confirms my research motivation that focuses on the design of framework that jointly considers *allocative* and *representational* bias.

Chapter 3

Ontologies for the Exploration of Bias

In the previous chapter I presented the high heterogeneity of studies and approaches for the analysis of social bias in NLP and underrepresentation in digital archives. In the context of this fragmentation, Semantic Web Technologies may represent a crucial tool for the alignment and systematization of this works, as it has been recently demonstrated by Franklin et al. [2022]. However, there is a lack of resources aimed at providing comprehensive semantic foundations for a systematic reorganization of this new field of research.

In this chapter I present the People in the Media Ontology (PiM-O), a mid-level ontology aligned with DOLCE [Gangemi et al., 2002] and designed to encode knowledge about how people communicate, are represented, and often discriminated in traditional and digital media. The ontology relies on four pillars, which together form a comprehensive representation of the interaction between people and the media ecosystem: biographical knowledge, annotation, reception, and communicative situation. PiM-O is integrated by two domain ontologies: The Under-Represented Writers Ontology Network (UR-ON) and the Ontology of Dangerous Speech (O-Dang). UR-ON models the lives of Transnational writers, namely authors who were born in non-Western countries that

may suffer a lack of representation in digital archives. O-Dang is conceived as a tool for the alignment of existing corpora for HS detection and correlated phenomena. UR-ON will be adopted as the semantic background for my analysis of underrepresentation of non-Western people in datasets (Chapter 6), which is tied to my second research question:

RQ 2. Which strategies can be adopted to detect and mitigate *allocative* biases?

O-Dang will support the analysis related to my third research question in Chapter 7:

RQ 3. Which measures can be implemented to discover research design biases in annotated corpora for abusive language detection?

The chapter is organized as follows. In Section 3.1 I provide a review of the existing work on the semantic modelling of biographies and on models for the annotation of linguistic resources. Section 3.2 describes the PiM-O ontology, while Sections 3.3 and 3.4 respectively present the UR-Network and O-Dang.

3.1 Related Work

3.1.1 Biographical modelling

The task of defining a robust semantic encoding of biographical knowledge is characterized by philosophical implications about the nature of events and by the intrinsic limitations of the RDF/OWL syntax.

From a descriptive-logic perspective [McCarthy and Hayes, 1981], a biographical event is a fluent: a function that holds true within a given situation. The assertion in Example 1 is true only under the state of Obama's presidential term, which is time-bounded.

1. Barak Obama is the President of the USA

Representing such an assertion in RDF is problematic since this language stores the knowledge in triples of the type 'subject, predicate, object', thus hindering the representation of additional information about the same event that is crucial to verify its

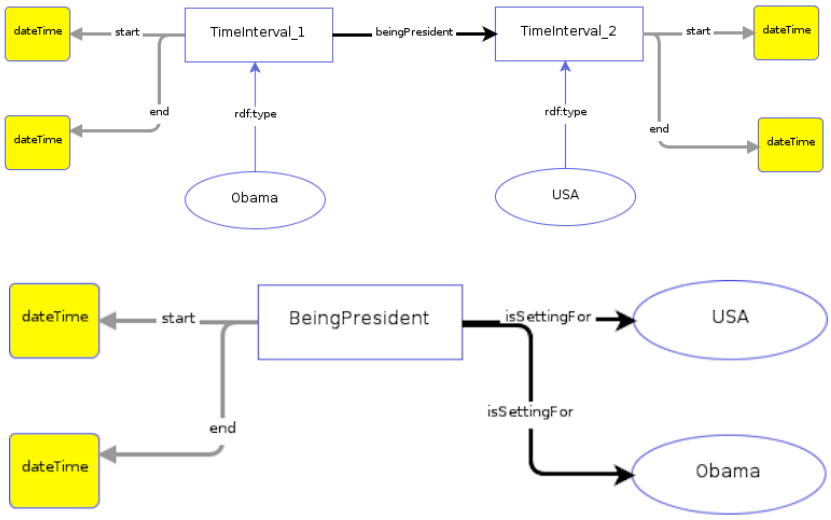

Figure 3.1: Representations of a biographical events based on two alternative reification strategies.

truthfulness. The example below shows that it is not possible to add temporal information to a triple.

```
:Barak_Obama :presidentOf :UnitedStates .
```

RDF reification is a strategy to overcome this issue by adding a semantically void triple (blank node) that functions as a collector for *n-ary* relations between a subject and a series of objects. Different strategies for event representation through reification have been proposed through years [Krieger, 2014, Krieger and Declerck, 2015]. Since a thorough overview of them falls outside the scope of this work, in this Section I only present two alternative encodings of the biographical event expressed in Example 1.

In the upper part of Figure 3.1 the biographical event is encoded through the reification of time intervals [Welty et al., 2006, Krieger, 2008]. At the core of the representation there is a triple where subject and object are two time slices. The former is associated with Obama, the latter to USA. Both have the same start and end date, which are connected separately to each time slice. This model emphasizes all the states of an entity

through its association to n time slices that represent all its biographical events. The representation in the lower part of 3.1 is event-centric [Gangemi, 2011, Hoffart et al., 2011]: the property *:beingPresident* is reified and linked to both entities that participate in the event and time boundaries. Here the focus is on chain of events that can be associated with one or more participant entities.

Almost all domain ontologies that are designed to represent events involving people are event-centric. StoryTeller [Yeh, 2017] is built upon the Time[1] and the StoryLine[2] ontologies. Token-reified biographical events are arranged in StoryLine slots, and further decomposed to express more detailed spatial and temporal information about them. Events are also the core units of the Narrative Ontology [Meghini et al., 2021], where they are organized in narrative units that can be explored through an interactive visualization tool[3]. The Bio-CRM ontology [Tuominen et al., 2018], which has been used to organize Finnish biographies[4], is built over the event class from the CIDOC Conceptual Reference Model (CRM) [Bruseker et al., 2017]. The Simple Event Ontology [Shi and Lin, 2019] and the Ontology Design Patterns [Gangemi and Presutti, 2009] are not designed to specifically represent biographical representation, but they are suited to encode them as occurrences where entities participate holding a specific role. 'Being president' is considered an event that includes Barak Obama with the role 'President'.

To conclude the section, it is worth mentioning two domain ontologies aimed at representing the lives of people who historically suffered several forms of under-representation and discrimination. The CWRC Ontology[5] [Brown et al., 2007] has been developed to support the Orlando project[6]: a data set of $1,300$ women British writers aimed at widening the study of feminist literary research. The ontology has an extensive taxonomy of classes describing the biography of a woman writer with the set of characteristics that determines her condition, such as ethnicity, political affiliation, reproductive history, and

[1] https://www.w3.org/TR/owl-time/
[2] https://www.bbc.co.uk/ontologies/storyline
[3] https://tool.dlnarratives.eu/
[4] https://www.timemachine.eu/ltm-projects/biographysampo-finnish-biographies-on-the-semantic-web/
[5] http://sparql.cwrc.ca/ontologies/cwrc.html
[6] http://www.artsrn.ualberta.ca/orlando/

sexuality. The Enslaved Ontology[7] [Shimizu et al., 2020] is a modular ontology aimed at mapping several databases about African slavery in a single Knowledge Graph (KG)[8] aligned with Wikidata.

3.1.2 Linguistic Linked Data

The application of SW Technologies for the annotation of linguistic resources resulted in a number of models and approaches that can be grouped under the definition of Linguistic Linked Open Data (LLOD) [Cimiano et al., 2020], a paradigm aimed at making linguistic resources Findable Accessible Interoperable and Reusable (FAIR). Corpora and dictionaries published according to FAIR principles must rely on a set of standards that foster their interoperability with other resources and standardise the output of texts processed by existing NLP tools [Cunningham, 2002]. Such standards vary depending on the granularity of the linguistic phenomenon analyzed. In this section I review a set of ontologies designed for LLOD describing their logic on the Example 2 from HateXplain [Mathew et al., 2021], a corpus of social media posts annotated for HS.

2. id: 17398393_gab; text: 'Let be compassionate to the legal citizens of this country stop funding illegal immigrants stop encouraging lawlessness and build that fucking wall'

The Open Annotation Model (OA) [Sanderson et al., 2013] provides a general framework for encoding annotations of a web document. The ontology provides a class of the type OA:ANNOTATION that is the collector of two elements: the body of the annotation, which is linked through the property **oa:hasBody** and expresses the value of the annotation; the target of the annotation, which represents the piece of media that is annotated and is linked to the OA:ANNOTATION through the **oa:hasTarget** property. Below it is possible to observe how the message in Example 2 is encoded according to this ontology. The annotation class is associated to a unique id and linked to an annotation

[7] https://docs.enslaved.org/ontology/
[8] https://enslaved.org/

value, 'Hate Speech', and a chunk of the document that expresses this phenomenon, 'stop funding illegal immigrants'.

```
<http://example.org/anno42> a oa:Annotation;
:hasBody 'Hate Speech';
:hasTarget: 'stop funding illegal immigrants' .
```

The Ontologies of Linguistic Annotation (OLiA) iniative [Chiarcos, 2012][9] provides a common format for the integration of more than 100 grammatical annotation schemes for different languages. The ontology distinguishes between a Reference Model (OLIA:LINGUISTICCONCEPT) that defines a common terminology and the Annotation Model (OLIA:LINGUISTICANNOTATION) that is specific to each annotation tagset aligned within this model. For instance, the adjective 'illegal' in Example 2 can be represented with Penn [Marcus et al., 1993] and Alpino [Van der Beek et al., 2002] tagsets in order to explore its occurrences in their respective treebanks.

```
<http://example.org/anno42> a oa:Annotation;
:hasBody 'Hate Speech';
penn:hasBody 'jj';
alpino:hasBody adj;
:hasTarget 'illegal' .
```

Since OLiA is mainly designed to align grammatical categories and tagsets, its expressiveness for the representation of semantic phenomena is limited.

OntoLex-Lemon [McCrae et al., 2017] is a lexicography module that overcomes this limitation with a scheme for the representation of a lexical item and its meaning. According to this model, a token may be represented as a ONTOLEX:LEXICALENTRY that can be directly linked to an encyclopedic concept through the property **ontolex:denotes** or reified through its lexicalized sense (ONTOLEX:SENSE). For instance, the term 'illegal' can be linked to a lexical sense that references (**ontolex:reference**) the definition of

[9] https://github.com/acoli-repo/olia

'illegal' in WordNet [Miller, 1995] and its semantic role in FrameNet [Baker et al., 1998]. Such an encoding represents the word as 'prohibited by law or by official or accepted rules', according to WordNet, and 'The Action is the behavior which complies with or violates the Code', according to FrameNet.

```
'illegal' a :LexicalEntry;
:sense :illegal_sense .
:illegal_sense :reference wn:illegal.a.01, fn:Action .
```

Finally, it is worth mentioning the Linguistic Linked Open Data cloud (LLOD) [Cimiano et al., 2020][10], Meta-Share [Federmann et al., 2012][11], and Framester [Gangemi et al., 2016][12] platforms. These initiatives aim at gathering within the same infrastructure a number of linguistic resources aligned through a common semantic model. LLOD gathers within the same platforms data that are published according to SW principles; Meta-Share is an initiative aimed at representing metadata about linguistic resources under a common model, Framester encoded FrameNet [Baker et al., 1998] under the SW paradigms and linked it with other semantic resources such as DBpedia [Auer et al., 2007], and WordNet [Miller, 1995].

3.2 People in the Media Ontology

Research on bias often focuses on how people belonging to minorities or minoritized groups are associated to certain features. These representations are fragmented, since different sources of knowledge may be affected by different types of bias about a target group. People in the Media (PiM-O)[13] is an overarching ontology designed to represent the multifaceted way in which people interact and are represented in the media environment, with a particular focus on social media platforms. Such a model is conceived as a foundational semantic background that organizes and aligns knowledge about how

[10] https://linguistic-lod.org/
[11] http://www.meta-share.org/
[12] https://framester.github.io/
[13] https://purl.archive.org/purl/people-in-the-media

implicit discrimination (e.g., social bias, underrepresentation) has an impact on people and groups. More specifically, PiM-O encodes and aligns four types of information in a unifying model:

- Biographies. Providing a formal representation of biographical knowledge is a key aspect to better understand how people are represented in media and social media. Since most of the information about people is unstructured, the definition of a semantic model for the storage, alignment, and comparison of biographical knowledge obtained through Information Extraction (IE) methods.

- Annotations. Corpora are not neutral resources, but artifacts shaped by the cultural contexts where they are developed. The positionality of research groups [Santy et al., 2023] and annotators' cultural background [Cabitza et al., 2023] have a strong influence on the quality and representativity of these resources. A semantic encoding of annotations may support a comparative analysis of corpora and other linguistic resources created for the analysis of people-related contents.

- Reception. Theories of reception, which span from political communication [Zaller, 1991] to literary studies [Jauss and Benzinger, 1970], emphasize the role of receivers in building consensus over ideas and cultural works. Different communities of users may react to the same message in opposite ways; in the cultural field, communities are active participants in the definition of canons. Reception's encoding enables for the modeling and comparison of how events and cultural artifacts are shaped by communities.

- Communication. Most of the corpora annotated for the analysis of blatant and implicit discrimination include texts gathered from social media. These type of messages are conversational and discrimination often happens within the interaction between users. Therefore, a semantic model of online conversations may be useful to provide a richer representation of these messages.

3.2.1 Design Principles

The first step of my semantic modelling has been the identification of a foundational model that represents a reference for my ontology among the several alternatives that have been conceived. The Basic Formal Ontology [Smith et al., 2005], which is widely adopted in bioinformatics, differs from other models for two distinctions: universals versus particular and dependent versus independent entities. The former provides a systematization of relations between types (e.g., hemoglobin is_a protein) and individuals (e.g., an arm is part_of a person). The latter enables the modelling of entities that depends on others, such as a virus that depends on the organism it infects. The Unified Formal Ontology [Guizzardi, 2005], which has been first applied in business modelling, relies on a taxonomy that represents the interaction between particulars (e.g., a boy) and universals that enables their identification (e.g., redhead). Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) [Gangemi et al., 2002] differs from the previous models in two aspects: it does not model the relation between universals and particulars; it supports a representation of reality as it emerges from human language and cognitive perception. A crucial feature of DOLCE is that any instance may be represented as it is perceived by a given observer. This characteristic fits the need of developing a semantic model for the identification of bias, since it supports the representation of different perspectives over the same phenomenon. For such a reason I adopted DOLCE as a reference model for PiM-O.

DOLCE provides a general distinction between endurants and perdurants . The term 'endurant' identifies all concepts that are always present through time. For instance, being a person is something that holds throughout the whole life of an individual. Perdurants identify concepts that are present only within a specific time interval, such as 'being a student'. Perdurants in DOLCE are defined as events (DUL:EVENT) and situations (DUL:SITUATION). The event class encodes anything that happens or occurs; the situation class might be considered as a view of a set of entities mediated by a given observer. Situations and Events are not mutually exclusive: they can contribute together to the representation of reality. In Figure 3.2 it is possible to observe a snapshot of DOLCE

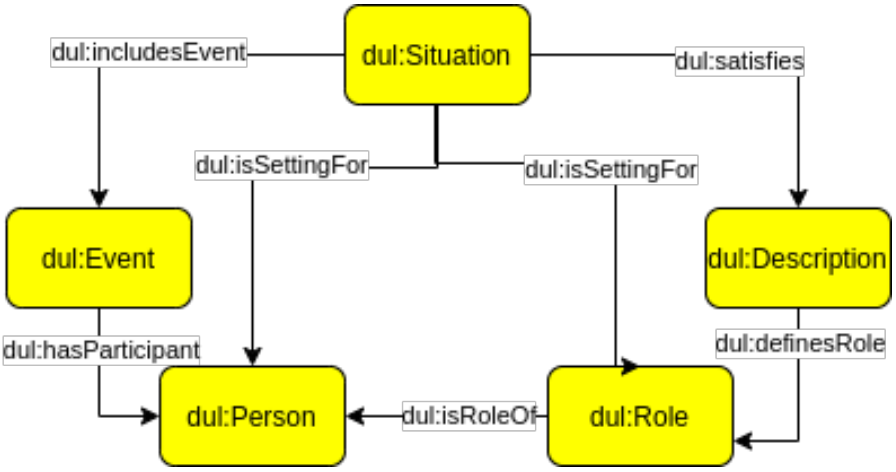

Figure 3.2: A graph representation of the interaction between DUL:EVENT and DUL:SITUATION classes in DOLCE.

where a DUL:SITUATION functions as a unique perspective that integrates three entities: an DUL:EVENT, a DUL:PERSON, and a DUL:ROLE. Within this perspective, a entity of the type person DUL:PERSON participates to an event. Finally, the situation satisfies a DUL:DESCRIPTION, which in turn defines the role of the person within the situation itself.

For instance, the sentence "The social media manager published a message" can be represented in DOLCE as a situation where a person participates in the event 'publishing a message' with the role of 'social media manager'. The role is defined by a description that represents the view on the event. Multiple descriptions may be satisfied within the same situation. If the message posted by the social media manager mentions a specific person, it can be added a second description that defines the new role 'addressee of the message' that is associated to a second entity of the type person who participates to the same event 'publishing a message'.

For the representation of people, I relied on the Ontology Design Pattern (ODP) framework introduced by Gangemi and Presutti [2009]. ODPs are small ontologies for specific use cases that are aligned with DOLCE and implementable in different contexts. Examples of ODPs[14] span from patterns like 'time indexed person role', which provides

[14]Actually there are 240 ODPs in the project repository: http://ontologydesignpatterns.org/

a high level encoding of human activities, to more specific uses like 'aquatic resources'. For this work I reused the 'time indexed person role' to encode the status of a person within a given interval of time, and the 'basic execution plan' to represent changes of status related to specific phenomena like migration.

A second relevant source that is reused in PiM-O is the Prov Ontology (PROV-O)[15] [Lebo et al., 2013] that models the provenance of knowledge encoded in a semantic model. The integration of PROV-O in my resource aims at emphasizing the presence of different and potentially conflicting perspectives over the same observation, thus allowing for the encoding of biases over events and categories of people. Three properties are mapped in my resource: (i)'was derived from' (**prov:wasDerivedFrom**) expresses all the sources from which knowledge is gathered. An usage example of the 'was derived property' is its adoption to compare two descriptions of the same event. 'Macron hosts Hungary's Orban in bid to unlock EU support for Ukraine'[16] was derived from France 24 and focuses on France's president, while 'Hungary's Orban says EU should first sign strategic partnership accord with Ukraine'[17] was derived from Reuters and emphasizes the role of Hungary's president. (ii) 'was attributed to' (**prov:wasAttributedTo**) identifies the responsible entity of the creation of a given resource. Davidson's corpus [Davidson et al., 2017], which is the first and one of the most famous annotated resources for Hate Speech (HS) detection, is attributed to two institutions: Cornell and Hamad Bin Khalifa universities. (iii) 'was associated with' (**prov:wasAssociatedWith**) associates an annotated entity to the agent that provided the annotation. For instance, the social media post "12 h ago ching chong accepted your friend request", which is included in the HateXplain corpus [Mathew et al., 2021], is associated with three annotators. Two of them marked it as offensive, while the third labeled it as expressing HS.

The design of the ontology followed the Ontology Requirement Specification Doc-

wiki/Community:ListPatterns

[15] https://www.w3.org/TR/prov-o/

[16] https://www.france24.com/en/live-news/20231207-macron-hosts-hungary-s-orban-in-bid-to-break-ukraine-deadlock

[17] https://www.reuters.com/world/europe/hungarys-orban-says-eu-should-first-sign-strategic-partnership-accord-with-2023-12-01/

ument (ORSD) [Suárez-Figueroa et al., 2009], a best practice for the definition and documentation of ontologies adopted within the NeOn methodology [Suárez-Figueroa et al., 2015], an approach to ontology design that identifies nine different scenarios to adopt for the development of a new semantic resource (e.g., reusing and merging existing ontologies). The documentation is divided in three main sections: the former highlights the general objective of the semantic model and its scope; the second focuses on potential users and usages of the resource; the third specifies Competency Questions (CQ), namely a set of queries that can be performed over a knowledge base encoded according to the ontology and that must return results.

The general purpose of PiM-O is to provide a comprehensive model for the representation of people in different media that enables the analysis of potential bias against them, while each pattern is tied to its own objective, intended usages, and set of CQs.

PiM-O is composed of four patterns: 'biographical situation' (PIM:BIOGRAPHICALSITUATION), 'annotation' (PIM:ANNOTATION), 'reception' (PIM:RECEPTION), and 'communicative situation' (PIM:COMMUNICATIVESITUATION). These patterns can be used as stand alone ontologies or they can be combined to form complex relations, such as an annotated communicative situation or the reception of a prominent biographical event.

Each pattern has been evaluated throughout its whole life cycle along three dimensions: consistency, completeness, and conciseness [Gómez-Pérez, 2004]. Consistency refers to the assessment of the ontology against CQs, in order to identify potential fallacies in the semantic model. This dimension was assessed at two levels. Together with my supervisor, I recursively evaluated the alignment between my research objectives, pattern main purposes, and derived CQs, identifying and fixing inconsistencies. Once I outlined the definitive CQs and designed the pattern, I tested them by writing a SPARQL query for each CQ and querying it over a small A-Box to identify whether the CQ is aligned with the ontology. The evaluation of completeness has been performed by analyzing the effectiveness of patterns for the re-encoding of data from different knowledge bases that may include information for which the pattern was not suited. For instance, to assess the Reception pattern along the 'completeness' dimension I analyzed all types of reception

1	**Purpose** Providing a semantic representation of any event involving at least one person.
5	**Intended uses** Usage 1: identification of gaps and bias across resources
6*b*	**Group of Competency Questions** CQ_1.1: Which biographical events of a person person are present in a given source of knowledge? CQ_1.2: Which are the most common events to specific categories of people?

Table 3.1: A snapshot of PiM-O's ORSD that highlights purpose, intended uses and competency questions for the 'Biographical Situation' ODP.

that are included in online communities of readers end enriched the taxonomy with specific indicators of receptions (e.g., the number of readers). Finally, I evaluated pattern conciseness by gathering feedback on pattern from expert and non-expert users. Such a step has been crucial to identify the main issues that people who are not familiar with DOLCE encounter when accessing to patterns and guided the development of property chains that improved patterns' intelligibility for external users.

In the following sections each ODP will be described separately together with a snapshot of the ORSD and examples of case uses and the corresponding semantic models.

3.2.2 Biographical Situation

The aim of the biographical situation ODP is to provide a rich semantic representation for a fine-grained analysis of bias and underrepresentation of specific categories of people in existing sources of knowledge (Table 3.1). From such a purpose and intended usage, two CQs are elicited. The first (CQ_1.1) focuses on the identification of all biographical events related to a single person in a given source, enabling a quantitative and qualitative analysis of their representation in specific knowledge bases. The second (CQ_1.2) is related to the analysis of groups, since it is conceived to return all biographical information related to socio-demographic features like gender, ethnicity, or citizenship.

The biographical situation (PIM:BIOGRAPHICALSITUATION) is a subclass of a DUL:SITUATION that requires at least the presence of a person (PROV:PERSON) and sat-

isfies at least one condition (PIM:CONDITION). A PIM:CONDITION is a subclass of a DUL:DESCRIPTION that defines the role of the person involved in the biographical situation. Within the same ODP it is possible to specify additional information, such as spatio-temporal information or the presence of other entities. Below the pattern is represented in the Description Logic (DL) notation

$$BiographicalSituation \sqsubseteq Situation \sqcap \exists isSettingFor.Person \sqcap \exists isSettingFor.Condition \sqcap \exists isSettingFor.Role \sqcap Condition \sqsubseteq \exists defines.Role \sqcap Role \sqsubseteq \exists isRoleOf.Person$$

The same biographical situation can include multiple conditions framing different observations of the described person. The following sentence, extracted from the English Wikipedia page about Chimamanda Ngozi Adichie[18], can be encoded in the ontology in several ways.

1. At the age of 19, Adichie left Nigeria for the United States to **study** communications and political science at Drexel University in Philadelphia, Pennsylvania.

Figure 3.3 depicts an example of how the part of the sentence framed by the event 'study' may result in a situation of the type 'study abroad' where Chimamanda Ngozi Adichie is described by the condition of being a migrant, a student, or both, depending on the type of observation that is adopted. The encoding shows the participation of additional entities concurring to provide a richer representation of the biographical situation: the time interval (1996), the location (United States), and the university attended by the person (Draxel University). It is worth mentioning that an encoding focused on the event 'left' leads to a different representation, where Chimamanda Ngozi Adichie participates in an event of the type 'migration' that includes as participants a place of departure (Nigeria) and a place of arrival (United States).

The PIM:BIOGRAPHICALSITUATION ODP is also designed to support Relation Extraction (RE) [Bassignana and Plank, 2022b] with a taxonomy of nine biographical situations based on the entities that co-occur with a person. The taxonomy has been

[18] https://en.wikipedia.org/wiki/Chimamanda_Ngozi_Adichie

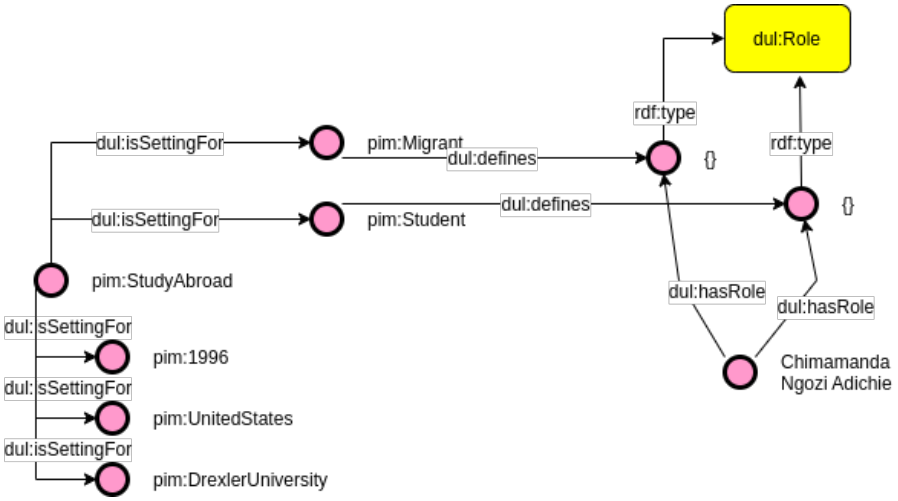

Figure 3.3: Example of a possible encoding of a sentence from the English Wikipedia biography of Chimamanda Ngozi Adichie in the Biographical Situation ODP

empirically defined by observing the relation types present in two existing RE datasets: REBEL [Cabot and Navigli, 2021], a silver corpus of semantic relations extracted from Wikipedia and Wikidata, and CrossRE [Bassignana and Plank, 2022a], a manually annotated dataset for RE of documents belonging to six domains. In Table 3.2 the nine relation types are listed, each of which characterized by the presence of a specific entity. According to such a taxonomy, the sentence 'In 2018, Zhao *directed* her third feature film, *Nomadland*, starring Frances McDormand' which in Wikidata is expressed by the triple 'Nomadland (Q61740820) director (P57) Chloé Zhao (Q21078321)', can be encoded in my ontology as follows:

```
[ a pim:Contributes;
isSettingFor pim:ChloéZhao;
isSettingFor pim:Nomadland;
isSettingFor pim:Director.
pim:ChloéZhao dul:hasRole pim:Director .

]
```

REL type	co-occurring entity	example
contributes	work + role	In 2018, Zhao *directed* her third feature film, *Nomadland*, starring Frances McDormand
topic	work	Napoleon appears briefly in the first section of Victor Hugo's *Les Misérables*, and is extensively referenced in later sections
field	occupation or discipline	Stephen William Hawking was an English *theoretical physicist*, *cosmologist*
geographical relation	place	Born in *Ogidi*, Colonial Nigeria, Achebe's childhood was influenced by both Igbo traditional culture and postcolonial Christianity
language	language	Seedorf speaks six languages fluently: *Dutch*, *English*, *Italian*, *Portuguese*, *Spanish* and *Sranan Tongo*
member	organization	Ahead of the 2009–10 season, Ronaldo joined *Real Madrid* for a world record transfer fee at the time of £80 million (€94 million)
position-held	organization + role	Meredith Whittaker is the *president* of the *Signal Foundation* and serves on their board of directors
relationship	person	[Billy Porter] married *Adam Smith* on January 14, 2017, after meeting him in 2009
participated	event	On November 2, 1970, Fonda was *arrested* by authorities at Cleveland Hopkins International Airport on suspicion of drug trafficking

Table 3.2: A list of biographical situations designed for RE. Situations are distinguished on the basis of the co-occurring entity of a triple. All examples are derived from the English Wikipedia.

Reducing the number of relation-types from hundreds to nine and providing strict constraints about the entity types that can be part of the relation is a strategic aspect of my ontology design process, since it enables for the reuse and evaluation of existing datasets for this task and for the integration of additional sources of knowledge within the same semantic model.

3.2.3 Annotation

The Annotation pattern is aimed at providing a transparent representation of annotated corpora (Table 3.3). As it has been observed in Section 3.1, existing ontologies for linguistic annotation and NLP enable the interoperability between existing resources, but they are not suited for the most recent trends in corpora annotation: the positionality of

research teams developing datasets [Santy et al., 2023], the presence of cultural bias in existing corpora [Sap et al., 2020], the rise of alternatives approaches to methods based on annotators' agreement [Abercrombie et al., 2023]. The pattern is intended to have two main usages. The first is comparing resources and messages that have been annotated within different annotation processes or that received annotations for multiple phenomena (Usage 2). From this usage two complementary CQs have been derived. CQ_2.1 enables gathering all the annotations of a given message for the same phenomenon with different annotation schemes. It is the case of many examples annotated for HS detection in the corpus developed by Davidson et al. [2017] and re-annotated in subsequent corpora. CQ_2.2 focuses on the retrieval of all messages annotated for more than one phenomena (e.g., HS and irony), which enables how orthogonal dimensions interact within the same message. The second intended usage is about facilitating the analysis of bias in annotated corpora by providing a semantic background for their identification (Usage 3). Through CQ_3.1 it is possible to obtain the provenance of each corpus, namely the research institutions that developed these resources. CQ_3.2 enables the exploration of disaggregated annotations in order to identify if some annotation patterns correlate with certain features of annotators.

The Annotation ODP introduces an explicit modeling of annotations' provenance and integrates them with existing authoritative ontologies for linguistic annotation: NLP Interchange Format (NIF) [Hellmann et al., 2013], OLiA [Chiarcos, 2012] and OA [Sanderson et al., 2013]. The backbone of the ODP is the relation between an annotation (PIM:ANNOTATION), which is a subclass of DUL:SITUATION, that satisfies one or more annotation scheme (PIM:ANNOTATIONSCHEME). The annotation is attributed to the organizations that managed the annotation task, while the annotation schemes satisfied by the annotation are associated with agents who actually labeled the message. Below, the pattern is described in DL.

$$Annotation \sqsubseteq Situation \sqcap \exists satisfies.AnnotationScheme \sqcap \exists isSettingFor.Person \sqcap \exists isSettingFor.Organization \sqcap (AnnotationScheme \sqsubseteq \exists wasAttributedTo.Organization \sqcap wasAssociatedWith.Person)$$

Following OA [Sanderson et al., 2013], each annotation has a target (**oa:target**), which is the piece of an information object (DUL:INFORMATIONOBJECT) that is annotated, and a body (**oa:body**), which represents the annotation label. For the sake of interoperability with existing ontologies, the annotation scheme is a super class of a NIF string (NIF:STRING), of a OLiA Linguistic Annotation (OLIA:LINGUISTICANNOTATION), and of an Annotation class encoded in OA (OA:ANNOTATION).

Figure 3.4 shows the semantic representation of the same message "🙌 RT @iDO_me2: this vodka be having all the hoes bent over 😩💃" that has been part of two different annotation processes. The former is attributed to Cornell and Hamad Bin Khalifa Universities [Davidson et al., 2017]; the latter to the University of Washington and Stanford [Sap et al., 2020]. Both annotation schemes have been designed to evaluate the offensiveness expressed by the message, but diverge about annotations releases. The ontology represents such a difference by linking the annotator role to a group, whenever annotations are provided in an aggregated form, or to a person, if they are not. Such a model is also suited for comparing annotation strategies. Davidson et al. [2017] scheme is binary and targets all the message; Sap et al. [2020] asked annotator to identify offensive text spans and to provide a rationale of their annotation. Such a complex task can be encoded as follows:

```
[ a pim:AnnotationScheme;
oa:hasTarget 'hoes';
oa:hasBody 'offensive';
oa:hasBody 'refers to someone that's sexually promiscuous'.
]
```

In this case the annotation scheme refers to the word 'hoes' as a target of the annotation that specifies two types of annotation: the offensiveness of the word and the rationale about its offensiveness.

The pattern can also be adopted for representing the annotation of different phenomena over the same message. For instance, the text in Figure 3.4 has been also anno-

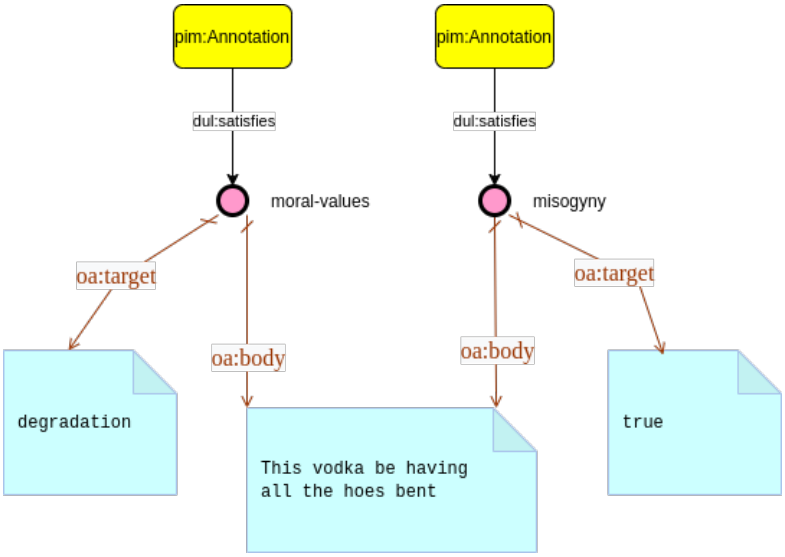

Figure 3.4: The representation of a message that has been annotated in two different corpora.

tated as expressing degradation in the context of an annotation of moral values [Hoover et al., 2020]. As it can be observed in the example below, it is possible to associate a DUL:INFORMATIONOBJECT to different annotation schemes in order to express multiple descriptions of the same message.

```
[ a dul:InformationObject;
dul:isDescribedBy pim:ann_01, pim:ann_02 .
pim:ann_01 a pim:AnnotationScheme;
oa:hasBody ‘offensiveness’ .
pim:ann_02 a pim:AnnotationScheme;
oa:hasBody ‘degradation’ .
]
```

Such a representation can be adopted for explore the interaction of different phenomena that jointly contribute to the perception and diffusion of discriminatory contents, which may support transfer learning [Mozafari et al., 2020] and multi-task learning [Plaza-Del-Arco et al., 2021] approaches for the identification of HS and correlated

1	**Purpose** Encoding annotated corpora in a transparent way.
5	**Intended uses** Usage 2: aligning different corpora annotated for the same phenomenon Usage 3: facilitating the identification of biases in corpus creation
6*b*	**Group of Competency Questions** CQ_2.1: which are the messages annotated for the same phenomenon with different annotation schemes? CQ_2.2: which are the messages that receive annotations for more than one phenomenon? CQ_3.1: which is the provenance of all corpora annotate for a specific phenomenon? CQ_3.2: which are the messages that were published with disaggregated labels?

Table 3.3: A snapshot of PiM-O's Ontology Requirement Specification Document that highlights purpose, intended uses and competency questions for the 'Annotation Situation' ODP.

phenomena.

3.2.4 Reception

The Reception ODP is distinguished from the Annotation pattern because it encodes forms of evaluation that are not structured in an annotation scheme and that are not specifically generated for the training of NLP systems. In the context of my ontology, reception might be considered as an abstract concept that may be specified according to different theories: the reception and acceptation of political messages [Zaller, 1991] or the reception of cultural works from specific audiences [Jauss and Benzinger, 1970]. The main purpose of this ODP is to trace patterns of discrimination related to events or works in specific contexts by encoding their reception (Table 3.4). The usage of this pattern in the PiM-O ontology is to assess the impact of human-related events or works across different communities (Usage 4). From this use case, three CQs are elicited: CQ_4.1 is designed to identify the number of reactions to a given item from a specific online community. CQ_4.2 and CQ_4.3 enables the retrieval of all reactions to a work from an online community or from a specific country.

The Reception pattern (PIM:RECEPTION) is the setting between an information object and its realization (DUL:INFORMATIONREALIZATION in different contexts[19]. Each realization receives a reception (DUL:RECEPTION) that functions as a collector of a number of data values aimed at measuring the intensity or quality of reception: number of likes (**pim:likes**), comments (**pim:comments**), reposts (**pim:reposts**), etc. A description of the pattern in DL is provided below.

$$Reception \sqsubseteq Situation \sqcap \exists isSettingFor.Manifestation \sqcap \exists satisfies.Reception \sqcap$$
$$(Manifestation \sqsubseteq \exists receives.Reception)$$

Figure 3.5 depicts an example of the ODP. The information object is the text of a message written by Barack Obama and published on Facebook and X the $17th$ of November 2023. As it can be observed, the same message received different reactions from the two audiences in terms of magnitude: $2,434$ likes on Facebook *versus* $39,611$ on X, 759 comments *versus* $6,143$, and 404 *versus* $8,540$. Receptions may also be intended qualitatively. The first post showed on Facebook under Obama's post is a enthusiastic appreciation ("I just finished watching the movie. Thank you for a fantastic film about a leader many knew nothing about."), while the first under X is an ironic and negative comment ("What grade would you give Obama and his agenda?"). The encoding of qualitative and quantitative reception indicators can support systematic analysis of the harmfulness of specific communication contexts against certain categories of people, as well as the identification of common reception patterns across different communities.

3.2.5 Communicative Situation

The Communicative Situation ODP models conversations between users at an abstract level. The aim of this pattern is to represent any form of interaction between users that may express a discrimination or a counter-speech (Table 3.5), being complementary to existing ontologies that encode communication from the perspective of the interaction

[19]A thorough description of description of the interaction between forms of expressions and their manifestations based on the FRBR model will be provided in Section 3.3

1	**Purpose** Representing the reception of messages by different communities of users
5	**Intended uses** Usage 4: Assessing the impact of human-related events or works over different audiences
6*b*	**Group of Competency Questions** CQ_4.1: How many users reacted to the same post in a given social network? CQ_4.2: How a given work is received by a given online community? CQ_4.3: How a given work is received in a given country?

Table 3.4: A snapshot of PiM-O's Ontology Requirement Specification Document that highlights purpose, intended uses and competency questions for the 'Reception' ODP.

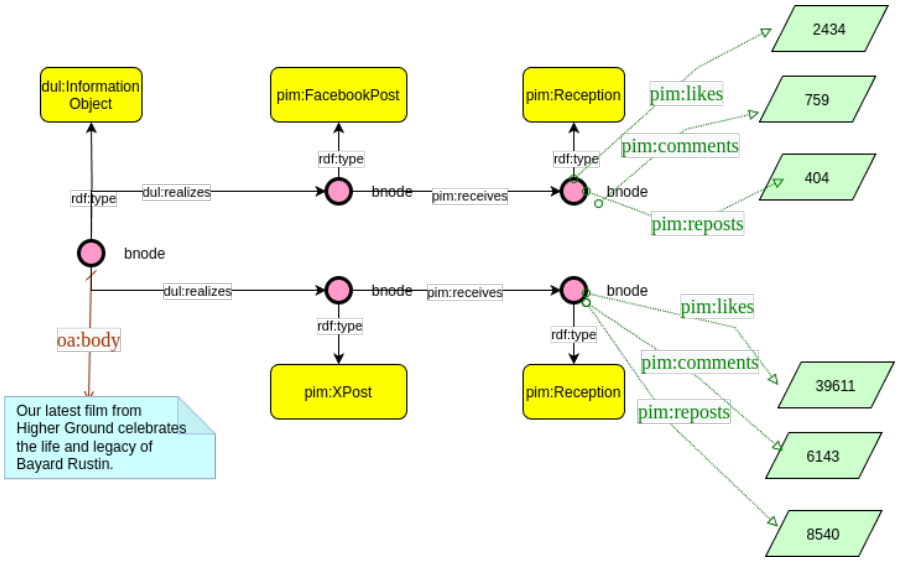

Figure 3.5: Two reception of the same post published by Barack Obama the 17*th* of November 2023

between groups and topics [Breslin et al., 2006]. Such aim is crucial since it allows encoding a message not only for its textual representation but for its intended pragmatic effect [Austin, 1975]. Two CQs have been derived from this aim. CQ_5.1 focuses on the retrieval of all messages that might have a negative impact on a specific target group while CQ_5.2 allows for obtaining all messages that challenge discriminatory contents, such as counter-speech [Chung et al., 2019] or counter-narratives [Lueg and Lundholt, 2020].

As for the other patterns, it is based on the interaction between a situation and a description. Elements from any type of communication theory may be represented by the description itself in order to encode simple and complex conversation dynamics. More specifically. A communicative situation (PIM:CONVERSATION) satisfies a communication theory (PIM:COMMUNICATIONTHEORY), which defines n communication roles (COMMUNICATIONROLE). In such a way it is possible to organize interactions by theories, role types, or by intention types. Below it is possible to observe the pattern in DL format.

$$Conversation \sqsubseteq Situation \sqcap \exists satisfies.CommunicationTheory \sqcap \exists isSettingFor.$$
$$CommunicationRole \sqcap CommunicationTheory \sqsubseteq \exists defines.CommunicationRole$$

Figure 3.6 shows how the pattern may be adopted to represent an example from the CONAN dataset [Chung et al., 2019], a corpus of hate speech messages and counter-speech. In the example, elements from two communication theories are jointly present: Adjacency Pairs [Schegloff and Sacks, 1973], namely units of conversation consisting of sequences of two adjacent utterance lengths, produced by different speakers, and Speech Act Theory [Austin, 1975], according to which utterances are used to perform specific actions. The interaction between the first message and its reply is encoded as an adjacency pair, where the former holds the role of 'first pair part', while the latter the role of 'second pair part'. The reply is associated to a second role of the type PIM:COUNTERSPEECH, that is derived from the Speech Act Theory. This role emphasizes the speech act of contrasting a certain message with another message.

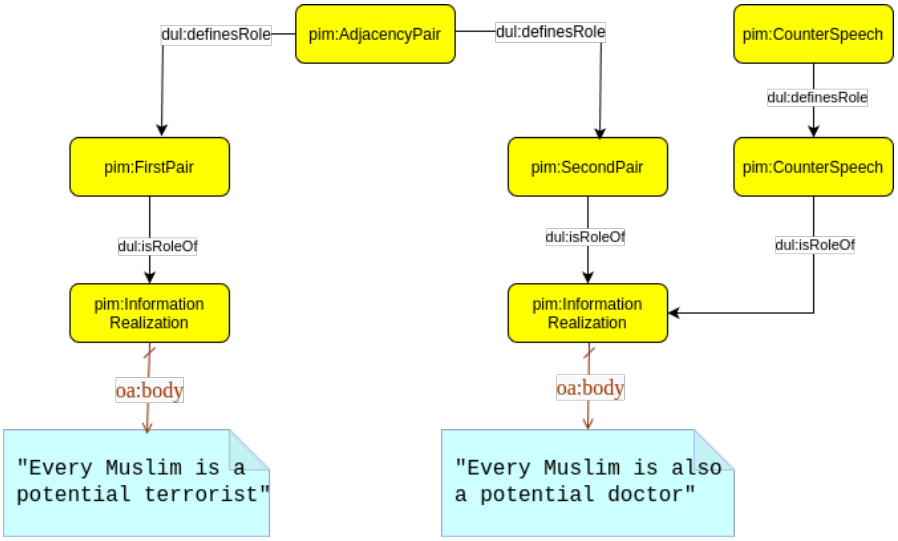

Figure 3.6: The example of a communicative situation pattern, where two tweets are encoded according the adjacency pair and the speech act theory.

This pattern also enables the representation of people who participate in a conversation by expressing their conversational roles that may be encoded in the pattern together with roles hold by messages. The example below represents the message 'What a hypocrite. You are part of the very antithesis of everything good.' as having the role SecondPairPart and is attributed to an user that has a role of the type pim:Addresser.

```
[ a pim:AdjacencyPair;
dul:definesRole pim:Addresser, pim:Addressee .
pim:Addressee dul:isRoleOf pim:user\_01 .
pim:Addresser dul:isRoleOf pim:user\_02 .
msg a :InformationRealization;
dul:hasRole pim:SecondPairPart;
dc:hasCreator pim:user\_02;
dc:comment 'What a hypocrite. You are part of the very
antithesis of everything good.'
]
```

1	**Purpose** Identifying conversation dynamics that may have an harmful impact on people.
5	**Intended uses** Usage 5: Encoding different types of interaction between users
6*b*	**Group of Competency Questions** CQ_5.1: Which are all the interactions between users that are harmful for a certain category of people? CQ_5.2: Which are all the counter-speech to messages marked as harmful?

Table 3.5: A snapshot of PiM-O's Ontology Requirement Specification Document that highlights purpose, intended uses and competency questions for the 'Biographical Situation' ODP.

3.3 The Under-Represented Ontology Network

Social media, and other User Generated Content platforms have given voice to an unprecedented number of people, while the Semantic Web offers encyclopedic knowledge about the world in an open, machine readable format. However, such technological transformation has not completely resulted in a more pluralist communicative environment, because the voices of people from non-Western countries are often unheard in crucial contexts. For instance, the involvement of minority journalists in mainstream newspapers is an open issue [Nishikawa et al., 2009], as long as the integration of post-colonial perspectives within school textbooks [Mikander et al., 2016]. This under-representation could be problematic since it precludes a full understanding of diversity in my society.

The Under-Represented Ontology Network (UR-ON)[20] is a network of two domain ontologies aimed at encoding life events of potentially under-represented writers and their works. UR-ON is the first step toward the creation of a semantic resource for the discovery and mitigation of the underrepresentation of non-Western writers in digital archives. The network reuses and specializes some of the biographical patterns modeled in the People in the Media (PiM-O) in order to provide more specific representation of two phenomena: (i) the interplay between writers and places where they lived; (ii) the description of their works as events where several actors contribute to the publication

[20] https://purl.archive.org/urwriters, https://purl.archive.org/urbooks

with different roles.

The work is structured as follows. In Section 3.3.1 I describe criteria that I adopted for modelling underrepresentation, while Section 3.3.2 and 3.3.3 describe the two ontologies that compose the UR-ON.

3.3.1 Modelling Underrepresentation

A prior step for the modelling of underrepresentation is to provide a set of criteria to categorize people that are potentially affected by this issue. My classification rationales derives from post-colonial studies [Spivak, 2015], 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, especially if the ontology is conceived to be populated by data gathered automatically. Spivak [2015] 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. Such type of knowledge is not present is public archives like Wikidata and DBPedia. I therefore chose two criteria for implementing such a classification: (i) the country of birth. I consider Transnational writers only people who were born in a country that has been a former colony and has a Human Development Index (HDI) below 0.8 [Stranisci et al., 2021c]. (ii) The ethnicity. I manually reviewed all the ethnicity types from Wikidata, keeping only the ones that represent minorities in Western countries (e.g., African Americans). All writers associated with one of them are classified as Transnational, even if they were not directly born in former colonies.

A further aspect I considered in designing my 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). Again, digital archives lack coverage of people's family origins and ethnicity, therefore I decided to do not rely on clearly bounded definitions like post-colonial, but to adopt the broader term Transnational, which refers to

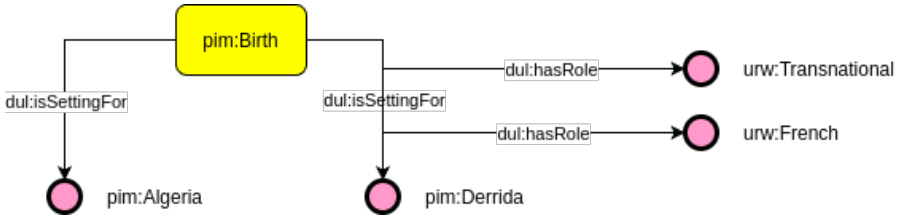

Figure 3.7: The representation of a Transnational writer of European origins.

people who 'operated outside their own nation's boundaries, or negotiated with them' [Boter et al., 2020] (p.9). When possible, the condition of being 'Transnational' is accompanied by additional properties like the country of citizen, to offer a transparent representation of false positives. Figure 3.7 represents the birth of Jacques Derrida as a situation in which the writer is associated with two roles: Transnational, for its birth in Algeria, a former colony with a HDI of 0.745; French, since his family obtained the French citizenship in the $19th$ Century.

3.3.2 The Under-Represented Writers Ontology

The URW-O provides the implementation of two additional biographical patterns to the PiM-O, in order to represent two situations: the process of migrating, and the status of a foreign person in a given country. Both are embodied in a specific time interval, and this relation of time-dependency need to be formally expressed for two reasons: on one side, it is essential to order life events in a chronological fashion; on the other side, it allows drawing a link between a writer's life, and their cultural production. As for the designing of PiM-O, my solution relies on the Ontology Design Pattern (ODP) framework [Gangemi and Presutti, 2009]. More specifically, I adopted the DUL:BASICPLANEXECUTION ODP to describe a migration, since a migration represents the execution of a intentionally devised line of action, and the DUL:TIMEINDEXEDPERSONROLE for modeling the legal status of a person, because the legal status of a person with respect to a country is typically non-rigid and can be modelled as a role.

The URW:MIGRATION class (see Figure 3.8) is subclass of a DUL:PLANEXECUTION and, as such, **dul:isSettingFor** six elements: the action of URW:MIGRATING, which is an

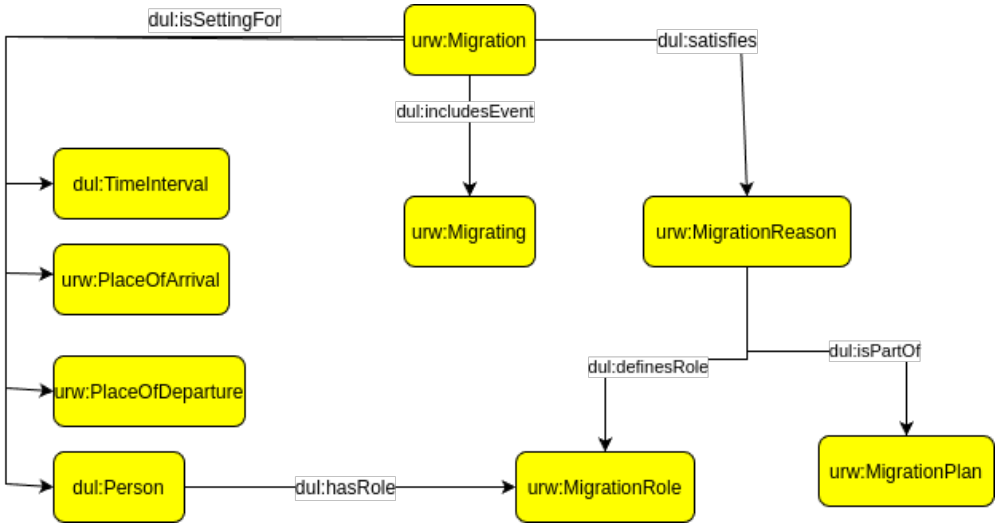

Figure 3.8: The representation of the Migration biographical Pattern.

event, a DUL:PERSON, namely the agent who is migrating, and her/his URW:MIGRATION ROLE in the migration process. The URW:MIGRATION class is also a setting for the spatio-temporal coordinates of the migration: the DUL:TIMEINTERVAL along which it occurs, the URW:PLACEOFARRIVAL of the migration, and its URW:PLACEOFDEPARTURE. We modeled the reasons for a person to leave her/his country for another as a URW:MIGRATION REASON that is subclass of a DUL:DESCRIPTION and part of a URW:MIGRATIONPLAN satisfied by the URW:MIGRATION situation. The URW:MIGRATIONREASON **dul:defines** the URW:MIGRATIONROLE of a person.

The URW:TIMEINDEXEDPERSONSTATUS class (Figure 3.9) **dul:isSettingFor** a DUL:PERSON and their role of the type URW:TRANSNATIONAL in a given URW:COUNTRY within a specific DUL:TIMEINTERVAL. The biographical pattern satisfies a PIM:CONDITION that is primarily used to express the status of a person (e.g., citizen, economic migrant, refugee), but it also can be used to describe other features that determines her/his condition. For instance, religion, sexual orientation, or social class. These additional aspects currently fall outside the scope of the ontology, so they are purposely left open to the integration with other semantic resources (e.g., Brown et al. [2007], Shimizu et al. [2020]).

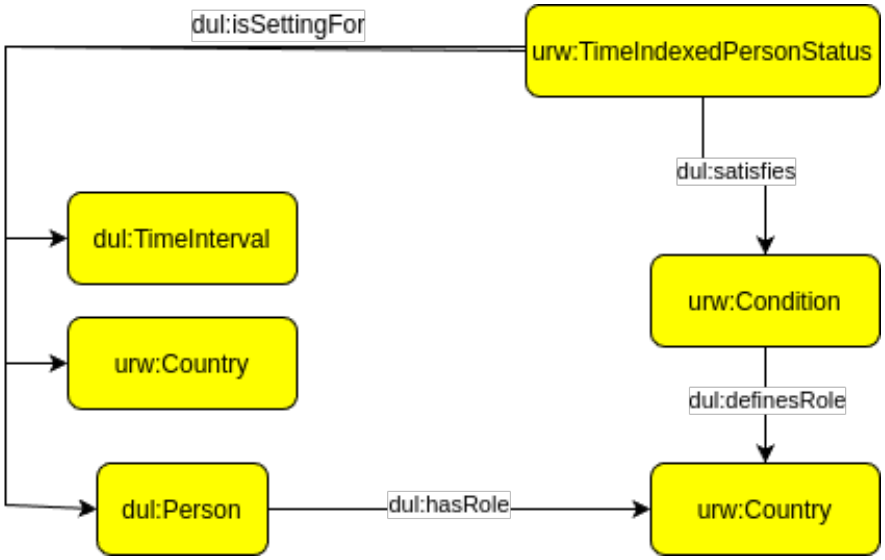

Figure 3.9: The representation of a Time-Indexed Person Status biographical Pattern.

3.3.3 The Under-Represented Books Ontolgy

The Under-Represented Books Ontology (URB-O) introduces the publication event, a concept that encodes a number of information about a work and its production process.

The ontology is a specialization of PiM-O and is mapped onto the Functional Requirements for Bibliographic Records (FRBR) [Tillett, 2005], a standard for modeling the relationship between a work (FRBR:WORK), its expressions (FRBR:EXPRESSION), and manifestations (FRBR:MANIFESTATION). Following the FRBR ontology, I defined a work as an instance of type FRBR:EXPRESSION, which is described as the 'intellectual or artistic realization of a work in the form of alpha-numeric, musical, or choreographic notation'. We then defined the concept of URB:EDITION as a subclass of FRBR:MANIFESTATION, namely 'the physical embodiment of an expression of a work'. These two concepts are linked through the property **frbr:embodiment**.

Each semantic relation between an expression and its edition is wrapped in a URB:PUBLICATION pattern, which is a subclass of a DUL:EVENT, an event in DOLCE can be used as a reification to provide rich descriptions of something that happens or occurs. In my case this type of pattern is adopted for two reasons: (i) expressing a large number of

facts about an edition (place, date, language of publishing and publisher) in a compact way; (ii) encoding roles of people who contributed to a publication without being the author of a work.

Finally, the model integrates the PIM:RECEPTION pattern with a number of attributes that are specific to the reception of literary works. Depending on the source of knowledge from which a work is derived, it may have an average rating (**urb:rated**), a number of ratings (**urb:numberOfRatings**), or a number of readers (**urb:numberOfReaders**). Figure 3.10 shows an example of my representation of works. 'Harry Potter e il Prigioniero di Azkaban', namely the Italian version (FRBR:EXPRESSION) of the 3rd Harry Potter book, **prov:wasAttributedTo** to J. K. Rowling and it has as **frbr:embodiment** the '1999 edition'. The latter in turn participates (**dul:isParticipantIn**) to a URB:PUBLICATION, a blank node entity that can be used for expressing several information: country of publication, year of publication, publisher, and translator. In the Figure the publisher, 'Salani Editore', and the translator, 'Beatrice Masini' are shown as entities linked to this event through the property **prov:wasAssociatedWith**. Finally, the URB:EDITION is linked to an entity of the type PIM:RECEPTION that functions as a collector for a number of information such as the source of the reception, the average review score, and the number of reviews are associated.

3.4 The Ontology of Dangerous Speech

Inside the NLP community there is a considerable amount of language resources created, annotated and released every day with the aim of studying HS and HS-related phenomena. This interest produced a proliferation of annotation schemes and approaches to this topic that resulted in a wide, but fragmented ecosystem of resources, as it was demonstrated by the surveys of Vidgen and Derczynski [2020] and Yin and Zubiaga [2021]. Since HS corpora differ in their annotation scheme, data gathering process, and annotation aggregation strategies, its alignment is challenging especially for HS, that is not

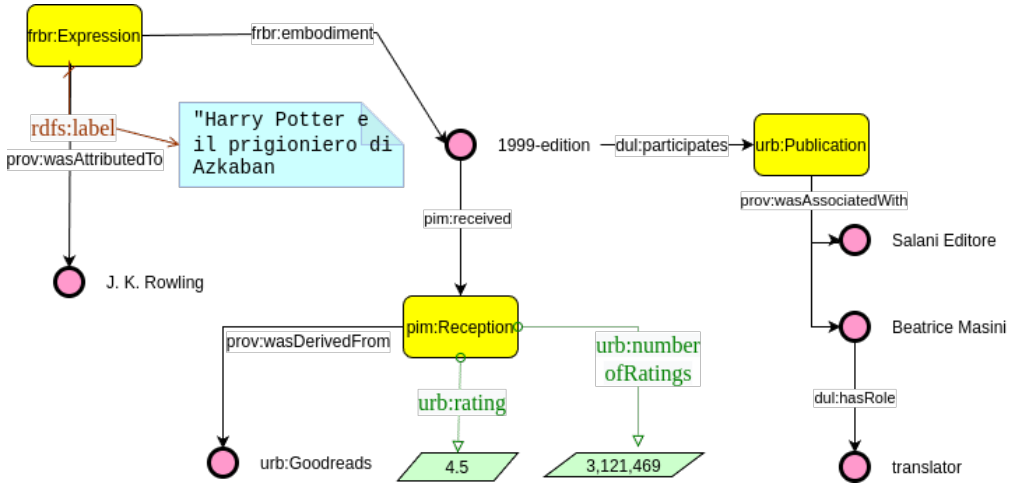

Figure 3.10: The representation of a work and its publication and reception process in URB-O.

only a linguistic phenomenon but a crime as prescribed by international laws[21]. In addition, an increasing number of works are exploring how many concurring phenomena contribute to the spreading of online discrimination [Plaza-Del-Arco et al., 2021, Frenda et al., 2022], emphasizing the need to implement a multi-tasking setting to extensively address the HS detection issue.

In this work I present the Ontology of Dangerous Speech (O-Dang)[22], a resource aimed at aligning different corpora for HS detection within a unique semantic model. O-Dang is a domain ontology derived from PiM-O Annotation module, which: (i) includes a lexicographic model based on OntoLex-Lemon [McCrae et al., 2017]; (ii) defines a relation between annotation schemes and laws that define and regulate HS; (iii) specializes the DUL:CONCEPT class with a number of concepts that may co-determine the spreading of discrimination. All these implementations are thought to provide a common ground for the ecosystem of resources annotated for HS.

The work is organized as follows. In Section 3.4.1 I define the concept of Dangerous Speech and its implication in the semantic modelling of O-Dang. Section 3.4.2 describes

[21]See for instance the definition of the Council of Europe: https://search.coe.int/cm/Pages/result_details.aspx?ObjectID=0900001680505d5b

[22]https://purl.archive.org/odang

the ontology and its potential applications to encode existing resources

3.4.1 Dangerous Speech

Dangerous Speech (DS) has been defined by Benesch [2012] as a speech that "has a reasonable *chance* of catalyzing or amplifying violence by one group against another, given the circumstances in which it was made or disseminated". Dangerous speech, therefore, is a type of speech that aims at contributing to create a climate of violence and intolerance against protected groups of people, such as women, immigrants, religious minorities, and others.

As some scholars highlighted, there are various rhetorical and pragmatic devices that play a part in the expression of dangerous utterances. For instance, Grimminger and Klinger [2021] and Frenda et al. [2019] reflected on the use of offensive and toxic communication in tweets expressing a stance towards specific political candidates (such as Biden and Trump) or sensible social issues involving a particular target such as women (like feminist movements or abortion). Others focused more on the use of the ironic language to lessen the negative tones of the hateful messages, making their automatic recognition challenging [Nobata et al., 2016, Frenda et al., 2022]. The employment of these kinds of devices actually lets speakers or users to be less explicit in their claims, limiting, thus, their exposure.

3.4.2 The Semantic Model

In this section I describe the characteristics of O-Dang and apply to the Twitter post in Example 1, which is included in the HS corpus developed by Waseem and Hovy [2016].

1. 'id:215, text:@ianbremmer @nrllhkose Fuck these people that want to steal my freedom of speech for the sake of a fascist religious ideology.'

As mentioned in the introduction, O-Dang reuses and extends the PIM:ANNOTATION pattern, from which it inherits the conceptualization of an annotation as a situation satisfied by a given annotation scheme (PIM:ANNOTATIONSCHEME). O-Dang introduces

two additional description classes: ODANG:LAW that is used for encoding the definitions of HS that appear in International laws about HS, and ODANG:GUIDELINES for the representation of Community Guidelines of social media. Annotation schemes can be linked to these descriptions through the property **dul:isRelatedToDescription**. The rationale of such an implementation is to express the connection between annotation schemes that are designed by research groups and norms that regulate this phenomenon. For instance, Example 1 has been annotated as HS according to the following annotation guidelines. The annotation can be tested over a number of existing norms to check if there are some discrepancies between them and the scheme.

> A tweet is offensive if it uses a sexist or racial slur; attacks a minority; seeks to silence a minority; criticizes a minority (without a well founded argument); promotes, but does not directly use hate speech or violent crime; criticizes a minority and uses a straw man argument; blatantly misrepresents truth or seeks to distort views on a minority with unfounded claims; shows support of problematic hash tags, E.g, "#BanIslam", "#whoriental", "#whitegenocide", negatively stereotypes a minority; defends xenophobia or sexism; contains a screen name that is offensive. [Waseem and Hovy, 2016]

A second aspect that is implemented in the ontology is the interaction between messages (DUL:INFORMATIONOBJECT), annotation schemes (PIM:ANNOTATIONSCHEME), and concepts (DUL:CONCEPT) that are evoked by them. The interaction is encoded both at the level of token and message. The token-level interaction is supported by the implementation of OntoLex-Lemon with existing lexica for HS and correlated phenomena. In Figure 3.11 it is possible to observe how the word 'fucking' is linked to a lexical sense that may reference at the same time to a concept defined in HurtLex, a multilingual lexicon of offensive, aggressive, and hateful words [Bassignana et al., 2018], and to an emotion and SenticNet affective lexicon [Cambria et al., 2010].

The same interaction between concepts can be expressed at a message level. In the example depicted in Figure 3.4 the same message as annotated both as containing HS

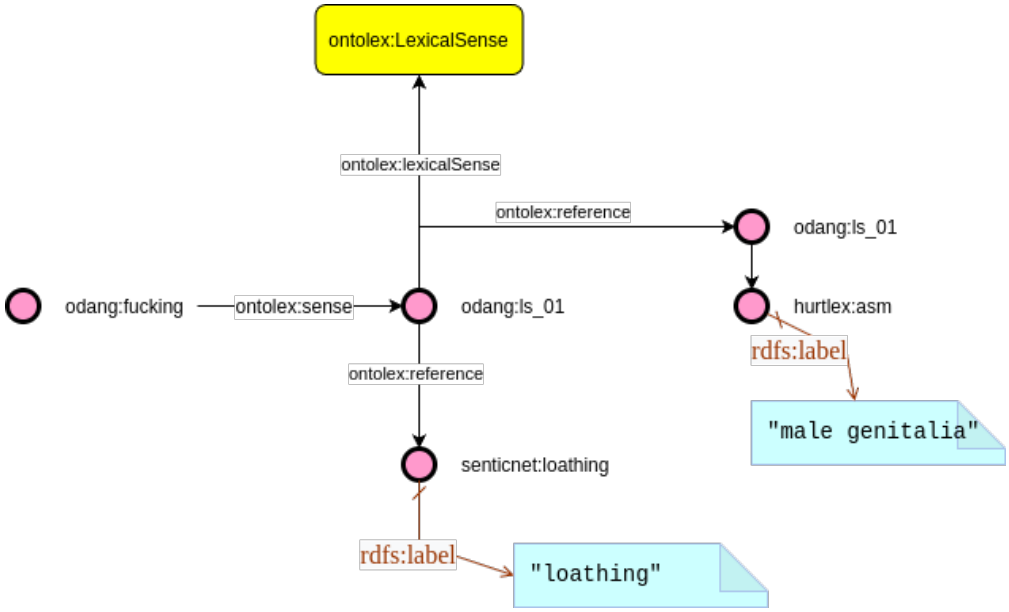

Figure 3.11: Two lexicalized senses of the word 'fucking' as they are classified in HurtLex and SenticNet

and expressing moral degradation.

3.5 Conclusion of the Chapter

In this chapter I presented a set of ontologies aimed at encoding relevant knowledge about how people are represented in media. PiM-O ontology encodes four relevant patterns at a general level: Biographical Situation, Annotation, Communicative Situation, and Reception. UR-ON network and O-Dang are derived from this ontology and they are specifically designed for the analysis of bias in Chapters 6 and 7.

Chapter 4

Bias Detection through Biographical Event Extraction

Biographies represent a privileged observation point to better understand how people are represented or misrepresented. They can be exploited to discover interactions between people, historical events and other entities and to explore *representational* bias affecting vulnerable groups to discrimination [Sun and Peng, 2021]. The interest in biographies is shared by many research communities that may benefit from methods for the automatic detection of biographical events.

However, few attempts to develop such methods have been made so far, the majority of which are based on off-the-shelf tools [Bamman and Smith, 2014, Russo et al., 2015, Menini et al., 2017]. This lack of approaches is also a crucial limitation for the analysis of *representational* bias in datasets, which mostly rely on surface-level features [Field et al., 2022] and reuse of tools that were not developed for this objective [Lucy et al., 2022].

In this chapter I fill this gap by presenting a series of resources for biographical event detection. I developed two corpora and tested three approaches for this task that resulted in the release of an effective model for the detection of biographical events.

Results presented in this chapter contributed to my first research question:

RQ 1. How can *representational* biases be detected and measured?

The classifier that I developed for biographical event detection enables a fine-grained investigation of stereotypical representation of people with certain socio-demographic features in widely-adopted public archives (Chapter 5). The classifiers is also tied to my second research question:

RQ 2. Which strategies can be adopted to detect and mitigate *allocative* biases?

Its adoption for the automatic extraction of biographical events from unstructured data contributes to the augmentation of knowledge about Transnational people, as I will demonstrate in Chapter 6 with a case study of writers in Wikidata.

The chapter is organizes as follows. In Section 4.1 I describe related work, surveying existing approaches for the annotation of events (Section 4.1.1) and existing work on Biographical Event Detection (Section 4.1.2). In Section 4.2 I present a Biographical Event Extraction experiment based on VerbNet and LSP [Stranisci et al., 2021b], while in Section 4.3 I describe a novel annotation scheme that relies on SRL [Stranisci et al., 2022c]. Finally, in Section 4.4 I describe a new semantic resource and a model for the detection of biographical events [Stranisci et al., 2023b].

4.1 Related Work

4.1.1 Resources for Event Detection

Event descriptions are complex linguistic entities whose study is placed at the intersection of multiple disciplines such as Philosophy, Linguistics, and Artificial Intelligence. The automatic identification of events has a longstanding tradition in Natural Language Processing (NLP) as shown by the availability of corpora covering multiple languages, text types, and representation formats [Baker et al., 1998, Pustejovsky et al., 2003b, Hovy et al., 2006, Bos et al., 2017], the organization of shared tasks [Doddington et al., 2004, Verhagen et al., 2007, Mitamura et al., 2015a], and dedicated workshops [Hovy et al., 2013, Caselli et al., 2015, Bamman et al., 2019].

This widespread interest in the topic is mainly due to the fact that *events* remain a

contested category and that *event descriptions* encapsulate lot of semantic information (i.e., participants and roles, time and location) that can be used for reasoning and inference. Such an interest produced a high fragmentation of approaches to the creation of resources for event detection, though. In this section I review some of the most adopted framework for the annotation of events focusing on a series of aspects that characterize them.

The Automatic Content Extraction (ACE) and Entities, Relations, and Events (ERE) annotation frameworks [Doddington et al., 2004, Song et al., 2015, Mitamura et al., 2015a] are part of two US government programs ran by by the National Institute of Standards and Technology (NIST) and aimed at exploiting NLP technologies for tracking entities, events, and their relations in documents. All these annotation schemes are interrelated (see Aguilar et al. [2014]): they do not track any type of event, but only a close set of event types; events annotations are not syntactically bounded. The original taxonomy of events include eight categories: LIFE, MOVEMENT, TRANSACTION, BUSINESS, CONFLICT, CONTACT, PERSONELL, JUSTICE[1]. According to these guidelines, events may be triggered by verbs, nouns, adjectives, pronouns or by groups of words that have been later defined as Event Nugget [Mitamura et al., 2015b, Song et al., 2016]. Each event may include a fixed number of arguments, namely entities that participate to the event with roles that can be general or specific to each event type. In Example 1 only three tokens trigger an event: 'arrested','charged', and 'deaths'. The event triggered by the noun 'deaths' belongs to the category LIFE. The chunk 'two Vancouver-area men' is annotated as Arg-Agent, while '329 passengers and crew members of an Air-India Boeing 747' is marked as Arg-Victim. '1985' and 'the Irish Sea' are general arguments that respectively define the time and place of the event.

1. Canadian authorities **arrested two Vancouver-area men** on Friday and **charged** them in the **deaths** of **329 passengers and crew members of an Air-India Boeing 747** that blew up over **the Irish Sea** in **1985**, en route from Canada to

[1] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-events-guidelines-v5.4.3.pdf

London.

A second family of approaches derives from the Frame Semantics theory [Fillmore, 1976] and from Beth Levin's research on verb classes [Levin, 1993]. These frameworks are specifically focused on verbs as 'categorizations of experience' [Fillmore et al., 2006] and that can be grouped in classes that share the same set of semantic roles and behaviors. Annotation frameworks sharing this approach are heterogeneous. FrameNet (FN) [Baker et al., 1998] is a collection of $1,224$ frames that can be evoked by a closed number of lexical units and that have a core structure of semantic roles. The word 'arrested' in Example 2 is one of the possible linguistic manifestation of the 'Arrest' frame that can be also evoked by words like 'apprehend', 'bust', and 'collar'. The frame has a set of three arguments that complete the categorization of the experience: Authorities ('Canadian authorities'), Charge ('the deaths of 329 passengers and crew members'), and Suspect ('two Vancouver-area men').

2. Canadian authorities **arrested** two Vancouver-area men on Friday and **charged** them in the deaths of 329 passengers and crew members of an Air-India Boeing 747 that **blew** up over the Irish Sea in 1985, en route from Canada to London.

PropBank [Kingsbury and Palmer, 2002] inherit from FN the frame-based approach, but narrows its scope to the annotation of verbs and abstracts semantic roles to a numbered list of role-types. The same verb in Example 2 is annotated as a frame of the type 'arrest.01', that represents one of the meaning that the verb may express. 'Canadian Authorities' is marked as Argument 0, while 'two Vancouver-area men' as Argument 1. Such simplification does not imply a reduction of attributes that may vary in the context of each verb semantics: when linked to 'arrest.01', Argument 0 represents the role 'police', while Argument 1 the role 'criminal'. It is worth mentioning that PB is a syntactically-bounded approach. Even if the annotation of a third argument that expresses the crime is provided in the 'arrest.01' frame and the information is expressed in the sentence ('the deaths of 329 passengers and crew members'), such semantic role is not annotated because it is syntactically linked to the another verb. The NomBank

project [Meyers et al., 2004] applies the same PropBank framework to nominal events. The noun 'deaths' in Example 2 is annotated as an instance of 'death.01' with 'of 329 passengers and crew members' as Argument 1. In the context of this nominal event, the argument represent 'the deceased'. More recently, the Abstract Meaning Representation initiative [Banarescu et al., 2013] proposes an integration of PropBank rolesetes within a Penn-based tree style [Marcus et al., 1993], in order to provide a more flexible and general semantic representation of events.

LIRICS [Bunt and Romary, 2002, Petukhova et al., 2008] and VerbNet [Schuler, 2005, Kipper et al., 2006] provides general representations of verb classes and of relational notions that link participants to events and situations expressed by verbs. The verb 'arrested' in Example 2 is part the 'prosecute-33.2-1' VerbNet class, which includes 'arrest', 'charge', 'indict', 'persecute', 'prosecute', 'punish', and 'try'. The verb class involves the roles of Agent and Patient and can be the expression of two semantics: DECLARE (an Agent describes a Patient criminal); CHARGE (An Agent apprehend a patient for a crime).

TimeML [Pustejovsky et al., 2003a, Saurí et al., 2006] represents a third approach to the annotation of events, which is the reference for a number of derivative guidelines specialized on news analysis [Minard et al., 2016], event co-reference [O'Gorman et al., 2016], and causal relations [Caselli and Vossen, 2017]. Like ACE and ERE guidelines, TimeML event annotation is not bounded to verbs and generally follows the principle of the "minimum span annotation", according to which an event must possibly be associated to a single token that represents its head. TimeML is characterized by its taxonomy that is not thematic, but highlights pragmatic functions of events. Events may be classified as I_ACTION whenever they trigger an intention (e.g.: try), PERCEPTION for all triggers related to physical perception (e.g.: hear), REPORTING for all declarative triggers (e.g.: say), ASPECTUAL that defines all triggers with an aspectual function (e.g.: continue), I_STATE that is used for all triggers that introduce the subjective stance of an individual about something that may occur (e.g.: believe, want, hate), STATE and OCCURRENCE cover any state or event that does not fall in the previous categories. TimeML provides

Corpus	Availablity	Licence Type
FrameNet [Baker et al., 1998]	free	MIT
TimeML [Pustejovsky et al., 2003a]	free	LDC
ACE [Doddington et al., 2004]	paid	LDC
PropBank [Hovy et al., 2006]	free	LDC
AMR [Banarescu et al., 2013]	paid	LDC
ERE [Song et al., 2015]	paid	LDC
Event [Song et al., 2016] Nugget	paid	LDC
RED [O'Gorman et al., 2016]	paid	LDC

Table 4.1: A list of corpora for event detection.

a set of time links that are suited to temporally order events. In Example 3 all event triggers are selected regardless their POS or the themes they belong to. Additionally, the scheme is suited to order them. For instance, the trigger 'arrested' may be classified as an OCCURRENCE, while 'charged' as an I_ACTION, since it expresses the intention of prosecutors. The two events are linked by a time link of the type <'arrested' BEFORE 'charged'>.

3. Canadian authorities **arrested** two Vancouver-area men on Friday and **charged** them in the **deaths** of 329 passengers and crew members of an Air-India Boeing 747 that **blew** up over the Irish Sea in 1985, en **route** from Canada to London.

Several attempts to unify such resources have been made throughout years. Pustejovsky et al. [2005] tried to merge TimeML, PropBank, and NomBank in a unique resource, Bonial et al. [2011] provided a mapping between LIRICS and VerbNet semantic roles, and the effort of Bunt and Palmer [2013] resulted the Semantic Annotation Framework International Standard (ISO 24617-4[2]. Palmer [2009] aligned PropBank, FrameNet, Wordnet, and VerbNet in a unique resource within the SemLink project[3]. Gangemi et al. [2016] performed a similar alignment of FrameNet with other semantic resources encoding themn in a KG[4]. Other alignment attempts resulted in the design of ontologies aimed at linking ACE and ERE annotation schemes to FrameNet [Parekh et al., 2023] and

[2] https://www.iso.org/standard/56866.html
[3] https://verbs.colorado.edu/semlink/
[4] https://framester.github.io/

VerbNet [Brown et al., 2017]. Finally, the work of Van Son et al. [2018] and Caselli and Bos [2023] is aimed at analyze the interoperability of annotation schemes and corpora for event detection, while Huang et al. [2023] released a benchmark that clean and harmonize existing benchmark for Event Detection in a unique resource.

A final analysis of annotation approaches must consider the availability of resources annotated according to their guidelines. Table 4.1 shows that a high number of corpora are not freely available. In particular, all resources derived from the ACE initiative are not freely available. This hinders a comprehensive reuse of corpora annotated for event detection. Only resources that are not issued under paid licence will be considered for my biographical event detection experiments.

4.1.2 Biographical Event Detection

From a methodological perspective, Biographical Event Detection might be considered as a specialization of Entity-Centric Event Detection Task (see Chambers and Jurafsky [2008]), namely the detection of all events in which the same entity is involved. Such a task has been relevant for a series of real-life application: the monitoring of news [Vossen et al., 2016], that in more recent times met the need of providing reliable methods for the identification of misinformation [Piskorski et al., 2020]; the extraction of biographical knowledge for the augmentation and curation of digital archives [Menini et al., 2017, Fokkens et al., 2017]; the use of biographical information for fine-grained analysis of social biases in datasets [Sun and Peng, 2021, Devinney et al., 2023].

Despite the relevance of these fields of application, resources and approaches for biographical event detection are fragmented and limited and almost none of the existing corpora has been annotated according to semantic standards described in Section 4.1.1. This hinders the possibility of extracting fine-grained biographical information from texts and of evaluating models trained for such a task.

Most of the approaches of biographical event detection are unsupervised. Narrative Chains [Chambers and Jurafsky, 2008] are clusters of events associated to the same entity that are automatically extracted by combining co-reference resolution and dependency

parsing. Bamman et al. [2013] adopt a similar approach to infer abstract descriptions of movie characters (*personas*) from Wikipedia pages.

Digital Humanities (DH) projects strongly rely on the usage of off-the-shelf NLP tools. Russo et al. [2015] and Menini et al. [2017] utilize existing tools for Semantic Role Labelling SRL and Frame detection to respectively extract biographical events of Nazi Camp refugees and historical figures. The Biography Portal of the Netherlands [Fokkens et al., 2017] has been enriched with an NLP pipeline based on existing tools, while the German Biography Portal "Deutsche Biographie" Reinert et al. [2015] has been augmented through the implementation of a set of regular expressions.

Few resources have been specifically designed for Biographical Event Detection or Entity-Centric Event Detection. MEANTIME [Minard et al., 2016] is a multilingual datasets of 120 English news articles and their translations in Spanish, Italian, and Dutch. For each document, 8 sentences are annotated for event and entity detection, following TimeML-inspired guidelines [Tonelli et al., 2024]. The corpus has been adopted as a benchmark for a SemEval shared task [Minard et al., 2015] where participants were asked to automatically identify and order entity-centric events across documents. Ding and Riloff [2018] developed a corpus of affective events: frame-alike tuples of the type <Agent, Predicate, Theme, PP> annotated for their polarity and categorized on the basis of the human needs they refer to (e.g.: health, finance). Wiegand et al. [2022] recently released a corpus of tweets annotated for the identification of biographically relevant utterances. Sun and Peng [2021] developed a semi-supervised dataset of events for the analysis of social biases on Wikipedia, while Devinney et al. [2023] proposed a procedure for the automatic generation of biographical datasets from Wikipedia in different languages.

In this Section I provided a review of resources annotated for Event Detection and described works aimed at the identification of entity-centric or biographical knowledge from text. Unfortunately, manually annotated corpora designed with existing semantic annotation schemes (Section 4.1.1) still do not exist. In next sections I present a series of works focused on the creation of such resources.

4.2 Extracting Biographical Events Through Lexico-Semantic Patterns

LSP matching is a rule-based approach [Jacobs et al., 1991, IJntema et al., 2012] that has been developed for supporting IE and Ontology Population tasks. Similarly to Lexico-Syntactic Patterns [Hearst, 1992], a methodology for the extraction of grammatical relations from large corpora, LSP is a framework aimed at extracting meaningful relations between entities and events from raw text. The language provides the creation of rules composed of semantic and syntactic elements that are related to classes and properties of an ontology. When a rule matches a string of text, the ontology is automatically populated with one or more RDF triples. LSPs have been implemented in a series of IE pipelines (see Frasincar et al. [2009]) and adopted in several scenarios, such as financial events discovery [Borsje et al., 2010], extraction of scientific knowledge [Ovchinnikova et al., 2021], and urban planning [Saeeda et al., 2020]. An example of a Lexico-Semantic Pattern, designed to extract geographical information from text [Saeeda et al., 2020] is shown in Example 4.2:

1. *$subject : Concept COMP RB? IN? $object : Concept* matches the phrase *Administrative territory of Prague is divided into localities* retrieving a mereological relation between *Prague*, and *localities* to be stored in an ontology.

In this work I adopted the LSP methodology in combination with VerbNet [Schuler, 2005] to populate the URW-O (Section 3.3) with automatically extracted biographical event with a focus on the relation between people and geographical entities. Such type of knowledge is important since it allows to track people movements during their lives and to investigate their connections through their mutual relationships with Organizations (ORG) and Geo-Political Entities (GPE).

Such a pipeline has been tested over a corpus of $7,979$ English Wikipedia biographies of Transnational writers, namely people born in former colony or belonging to ethnic minorities in Western countries.

4.2.1 LSP creation

The first step of my pipeline has been the definition of a set of LSP rules. In order to do so I first preprocessed biographies: I split them in sentences with the SpaCy library[5], and only kept the ones containing at least one entity of the type ORG or GPE. I selected a sample from the 87,091 extracted sentences and identified the presence of three types of pattern:

- a sequence of a verb, a preposition, and an entity of a type ORG or GPE, as in the sentence: "She **worked for Jive Records** in New York City during the summer of 2007";
- a verb and an ORG or GPE, as in "Maya Angelou often **left New York**, but she always came back";
- a preposition and an ORG or a GPE preceded or followed by a verb, as in "In 1988 he proceeded the post-graduate studies **at the Moscow Mining Institute** (Moscow State Mining University), defended his dissertation and **obtained** Ph.D degree.";

The subsequent step in the definition of the LSPs has been the clustering of verbs through a mapping with general verb types, aimed at reducing the number of patterns and increasing their recall. To do so, I employed the Unified Verb Index[6], a repository resulting from the mapping of several lexical resources that provides syntactic and semantic frames of English verbs. In particular, I linked the verbs in my data to the VerbNet classes in the Unified Verb Index (UVI) [Schuler, 2005]. An example of relevant class for mapping movement verbs onto the Migration ontology patters is *escape-51.1-1-1*, under which the following lemmas are included: depart, disembark, escape, exit, flee, leave, vacate. Such a refinement led to the creation of 48 LSPs for the conversion of texts in two biographical patterns: Time-Indexed Person Status (T-IPS) and Migration. The former

[5] https://spacy.io/
[6] https://uvi.colorado.edu/

LSP	Biographical Pattern
supervision-95.2.2-1 at|in GPE|ORG	T-IPS
create-26.4 in|by|on GPE|ORG	T-IPS
auxpass admit-64.3-1 in GPE|ORG	T-IPS
send11.1 to|at GPE|ORG	Migration
escape-51.1-1-1 to|at|in GPE|ORG	Migratiion

Table 4.2: Examples of LSPs together with the biographical patterns they target within the URW Ontology.

encodes any biographical events happened in a foreign country, the latter a movement from a country to another. Table 4.2 shows some examples of LSPs that have also been design to account for passive forms of verbs.

The sentence in the Example 2 may be encoded as a biographical pattern (see Section 3.2.2) through the LSP *escape-51.1-1-1 to|at|in GPE|ORG*.

2. On July 16, 1996, at the age of 43, **Malika Oufkir emigrated** to **Paris** accompanied by her brother Raouf and her sister Soukaina.

The expected result of this match is to encode the sentence in a biographical pattern of the type Migration (see Section 3.3), that includes a person, a GPE, and an event:

```
[ a :Migration;
:isSettingFor :MalikaOufikir, :Paris.
:inlcludesEvent :emigrate.
]
:MalikaOufikir :participatesIn :emigrate.
```

.

4.2.2 Entity Linking

An Entity Linking EL step has been implemented within the pipeline for two reasons: (i) reducing the number of false positives detected as GPE and ORG; (ii) optimizing the knowledge about locations in the extracted triples by inferring the sovereign country of the extracted named entities.

The EL approach is based on Wikidata that has been used as a knowledge base against which recognized entities have been searched and linked. All the 62,912 strings identified as geopolitical entities or organizations by the SpaCy Named Entity Recognition module have been used as an input for search through the Wikipedia API[7] that returned a list of 187,680 candidates with their Wikidata ids. For each of them, the property 'country' (P17) has been searched in Wikidata, thus obtaining a cleaned list of 45,894 entities and 65,607 candidates.

As a final step of the EL process, the similarity between named entities and candidates has been computed by averaging two metrics: Gestalt Pattern matching [Ratcliff et al., 1988], which is a measure of the number of characters shared by two strings, and the Cosine Similarity [Rahutomo et al., 2012] between the BERT vector representations [Kenton and Toutanova, 2019] of retrieved entities and their candidates. Table 4.3 shows an example of the EL pipeline. 5 candidates were selected for the string 'the cavendish laboratory'. The candidate, 'Cavendish Laboratory', is the candidate with the highest similarity score (0.9) and the only one with a link to a sovereign state: 'United Kingdom' (Q145).

Recognized Entity	**Candidate**	**Score**	**P17**
the cavendish laboratory	Cavendish Laboratory	0.9	UK (Q145)
the cavendish laboratory	Mark Oliphant	0.275	x
the cavendish laboratory	Henry Cavendish	0.65	x
the cavendish laboratory	Brian Josephson	0.3	x
the cavendish laboratory	James Chadwick	0.45	x

Table 4.3: An example of the EL pipeline's results. 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. Third and fourth columns respectively report the similarity score between the named entity and the candidate, and the sovereign country of the entity.

The final output of this EL step is a list of 18,579 recognised entities that are linked to a sovereign country and that obtained a score between the named entity and the candidate equal or greater than 0.75. Below is an example of how a linked entity is

[7] https://en.wikipedia.org/w/api.php

finally stored as a dictionary.

```
{
    entity: Q835960,
    label: 'University of São Paulo',
    named_entities: ['the university of são paulo',
    'universidade de são paulo'],
    P17: Q155
}
```

4.2.3 Analysis of Results

Our pipeline allowed identifying 93,946 triples from 46,334 unique sentences. The pattern with the highest number of matches is 'assessment-34.1 in|on|from|at ORG|GPE', through which 7,412 triples of the type T-IPS have been extracted. One of the most recurring match involves the verb 'study' as in the example 'He **studied** biomedical sciences **at Bristol University**', showing that from this pattern information about people's education are extracted. Other relevant LSPs are 'lecture-37.11-1 to|at|in GPE|ORG' and 'employment-95.3 at|in GPE|ORG'[8], which mainly focus on writers' and respectively matched 5,970 and 6,073 triples of the type T-IPS. An example of the former is 'she became the first woman to **speak at the ancient al-Azhar University** in Cairo', while the latter may be exemplified in 'In 1989 he **taught** a seminar on African culture **in Santo Domingo**'. The most relevant LSP for the extraction of biographical events of the type Migration is 'escape-51.1-1 to|at|in GPE|ORG' that matched 5,162 triples. An example of this pattern is presented in 'Gilmore **emigrated** from New Delhi **to London** as a teenager'.

Table 4.4 shows the number of extracted triples grouped by continent and the number of writers grouped by their continent of birth. The three more representative birthplaces in this sample are Asia (14.924), Latin America (1,940), and Africa (1,583), while the

[8]All the created LSPs were made available in the thesis' Github repository: https://github.com/marcostranisci/semantic-aware-bias/tree/main/resources/lexico_semantic_patterns

three continents linked to the highest number of biographical triples are North America (15,843), Asia (14,942), and Europe (11,347). These results are mainly in line with expectations: Transnational people are classified on the basis of being born on a former non-Western colony, while Europe and North America are the destination of some of the most important migratory flows. The high number of triples with an Asian ORG or a GPE might be explained by the high presence of biographies about writers born in India and Pakistan (60%) where English language is widespread. This could produce an overrepresentation of such authors, since my experiment is only based on English Wikipedia pages. Considering only triples set in a different continent than the one where authors were born reduces the number of triples with an Asian GPE or ORG. The sankey diagram in Figure 4.1 shows the two main destinations of movements between continents are Europe and North America, while the number of writers that from outside Asia that move in the continent are significantly less than ones that migrate within it.

Continent	**n. of triples**	**n. of writers**
North America	15,843	722
Asia	14,942	2,908
Europe	11,347	26
Latin America	7,407	1,940
Africa	6,469	1,583
Oceania	581	4

Table 4.4: The number of extracted triples through LSPs and writers' birthplaces, broken down by continent

Results obtained through LSPs can be useful for the exploration of people's migration patterns in different historical periods and provide insights about the interaction between writers, geopolitical entities, and organizations. Additionally, they enable the analysis of gaps that affect the English Wikipedia. The prevalence of biographical triples about people from India and Pakistan may be interpreted as a very specific knowledge about colonial history, which is especially bounded to a single macro-region. Finally, this pattern-based approach is well suited to identify the most recurring information of which Wikipedia biographies are composed both in terms of event types (e.g.: study,

work, eg.) and lexical richness. In this sense, only few verbs that are included in VerbNet classes [Schuler, 2005] are effective for the retrieval of meaningful knowledge about people.

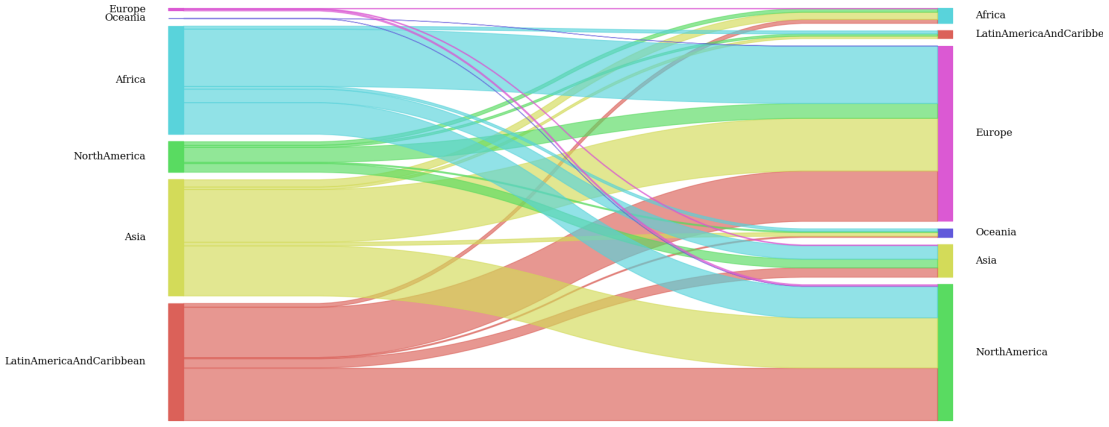

Figure 4.1: Sankey diagram of writers movements between continents. The left node represents writers' continent of births, the right represents the continent where triples are staged.

4.2.4 Evaluation Stage

The evaluation stage of LSPs focused on the assessment of their precision. Only sentences that matched with at least one of these patterns were considered as a benchmark for the evaluation. 2,555 sentences, corresponding to the 3% of the total, have been sampled and manually labelled by one annotator. Each sentence has been marked as expressing a 'T-IPS|Migration' or 'None', then compared with the predictions obtained with patterns. It is worth mentioning that such an evaluation does not assess the recall since a corpus of false negatives was not provided for this study.

Table 4.5 shows that LSPs performed with a precision of 0.68. It is important to mention an imbalance in the performance of the two biographical patterns: Migration

Pattern	Precision
TIPS	0.665
Migration	0.805
TIPS and Migration	0.68

Table 4.5: Results of the evaluation of biographical patterns

situations are retrieved with a precision of 0.805, in line with recent findings from the literature [Saeeda et al., 2020], while precision for TIPS is 0.665. This difference is probably due to the nature of the latter pattern, which is highly heterogeneous and needs a deeper analysis to specialize it into specific patterns for different status types. The manual analysis of prediction errors revealed they had several causes.

Co-reference. In some cases the error is determined by co-reference issues. Even if biographies are assumed to focus on a single individual, many sentences include events whose protagonists are other people. In the following case the triple of the type Migration is not relevant for the writer's life but for their father: '**His father** left India in early 1963 to study at Oxford University'.

Factuality. Another source of error is the factuality of biographical events extracted from biographies. That is the case of events that are extracted despite them not happening or their happening is not certain, as in the following example where the sentence does not provide enough context to determine whether the writer divorced or not: 'She eventually moved to Mexico City and unsuccessfully **tried** to file for divorce').

Entity Linking. The EL pipeline is also a cause of wrong predictions. Some entities are classified as ORG even if they do not belong to this type. It is the case of 'African American Studies0, which is a research field and that has been linked to the page 'African Americans' (Q49085), thus showing two interrelated issues: the Named Entity Recognition (NER) misclassification and the wrong identification of the corresponding entity on Wikidata. Other types of error are related to the role of the named entity within the

sentence. In the following example Twitter is not mentioned as a company with which the writer has a relation, but as a social media platform: 'in the phrase 'Shatrughan Sinha, has also spoken in Kumar's favour on **Twitter**'.

In addition to the above mentioned issues, the approach based on LSPs is vulnerable to low recall for several reasons. The fact that VerbNet is a non-exhaustive taxonomy of verbs may hinder the discovery of triples based on verbs that are not mapped. Similarly, nominal events [Meyers et al., 2004] are excluded from potential results due to my verb-based approach.

For these reasons LSPs alone are not completely reliable for the extraction of biographical triples and VerbNet is not sufficient as a knowledge base for event extraction. Finer grained corpora and more focused approaches are needed to implement a more effective biographical event extraction pipeline.

4.3 A Semantic Role-Based Approach to Biographical Event Detection

Semantic Role Labeling (SRL) [Kingsbury and Palmer, 2002, Màrquez et al., 2008] emerged as a suitable standard for the characterization of events and their participants in a frame-based fashion [Baker et al., 1998]. Given an event, typically expressed by a verb, this theoretical framework enables the identification of roles held by entities that are linked to it. Example 1, derived from PropBank guidelines [Bonial et al., 2010], shows an event expressed by 'left' that in the sentence means 'move away from'. 'Mary', marked as Argument 0 is the 'entity leaving', while 'the room' is marked as Argument 1, the 'place left'. Example 2 includes the same verb that expresses a different event, 'give', and a different roleset: the Argument 0 represents the 'giver', Argument 1 the 'thing given', and Argument 2 the 'beneficiary'.

1. Mary (ARG0) left (Pred) the room (ARG1)

2. Mary (ARG 0) left (Pred) her daughter-in-law (ARG 2) her pearls (ARG 1)

The SRL approach has been proposed in many variants (see Petukhova et al. [2008], Banarescu et al. [2013]) and applied to a number of domains like news analysis [Liu et al., 2010] or law text annotation [Bakker et al., 2022], showing its potential use cases and adaptations.

In this work I implement such an approach for biographical event detection, by releasing a novel set of annotation guidelines and BioSRL, a novel corpus specifically annotated for this task. Guidelines are mainly focused on the identification of arguments that include entities of the type ORG or GPE with the aim of providing a richer semantic representation of connections between people, events, and places. Guidelines have been designed to be consistent with two existing Semantic Annotation Frameworks, ISO 24617-1 [Pustejovsky et al., 2010], and ISO 24617-4 [Bunt and Palmer, 2013].

The guidelines have been adopted to annotate a corpus of 834 sentences extracted from Wikipedia pages of Transnational writers, namely writers born in non-Western countries, migrants or belonging to ethnic minorities [Stranisci et al., 2021c]. The annotation is designed to be interoperable with existing language resources (see Pustejovsky et al. [2003b], Hovy et al. [2006]), in order to potentially augment the corpus with additional data through a systematic mapping.

The work is structured as follow. Section 4.3.1 describes data collection and annotation guidelines design. In Section 4.3.2, results of the annotation are presented. Section 4.3.3 presents the mapping of the resource with existing corpora.

4.3.1 Data Collection and Annotation Scheme Design

In this section, the data gathering and preprocessing from Wikipedia is described; then, the annotation guidelines are presented.

4.3.1.1 Data Gathering

The corpus is a collection of sentences extracted from 7,979 Wikipedia English pages of Transnational writers (Chapter 3.3.2). The data gathering process was performed in four steps: (i) each biography has been split in sentences using Stanford Core NLP

[Manning et al., 2014]; (ii) for each sentence, all the named entities of the type Location or Organization have been identified using the same tool; (iii) an automatic semantic role labelling was performed on each sentence, using SRL Bert [Shi and Lin, 2019]. The resulting dataset of 218,198 tuples of predicates and semantic arguments contains at least one Location or one Organization. Below some examples are reported:

- **predicate**:move,**ARG2**:to New York City;
- **predicate**:study,**ARGM-LOC**:in the Convent of Jesus and Mary School in New Delhi;
- **predicate**:confer,**ARG0**:by the municipality of Kautokeino and the Kautokeino Sámi Association.

In the final step (iv), I identified the most frequently occurring combinations of 'predicate, ARG0', 'predicate, ARG1', and 'predicate, ARG2' in order to select a representative sample of the sentences in the data set for annotation.

4.3.1.2 Annotation Guidelines

Annotation guidelines[9] were developed in order to annotate all events in which the subject of the biography is a participant in the event. Annotations are inspired by ISO 24617-4 [Bunt and Palmer, 2013], according to which events and arguments are stored in sets of tuples defining their semantic roles and connections. Example 3 includes three events, each of which must be encoded as couples of markables and their semantic specifications. For instance, 'left' is encoded as <left, LEAVE>. The same annotation holds for other elements like places. Eg: 'South Africa' must be encoded as <South Africa, PLACE>. Markables can then be linked adopting the following semantic syntax: <South Africa, left, ARG1>.

3. In 1974 he **left** South Africa, **living** in North America, Europe and the Middle East, before **returning** in 1986

[9] Guidelines are available in Appendix B

The selection of the most significant semantic arguments in biographical events is guided by previous work [Stranisci et al., 2021b] in which a set of combinations of life events and named entities types were recognized as salient for biographies: locations for migrations; organizations for education and career events. Therefore, my guidelines mainly focus on events in which the subject of the biography is involved with such named entities. Moreover, since time is a crucial feature for biographical narratives, guidelines includes the identification of temporal expressions.

Identification of the entity and their semantic role. The prerequisite for an event to be annotated was that it had to involve the biography subject. This involvement was not always direct, though. For instance, an author could be mentioned through their works, as in "Her third novel, Missing in Machu Picchu (2013), was awarded" or through a group they were part of, as in "At the age of nine, her family moved to Ghana". According to the Richer Event Description (RED) guidelines[10], the former case was a BRIDGING relation, while the latter was a SET-MEMBER link. In my guidelines all these types of entity had to be annotated as if they were an instance of the writer, in order to consider important biographical events of the type 'his book win a prize', in which the writer is only indirectly mentioned.

Together with the identification of the writer, annotators had to specify her/his semantic role, in order to classify their participation in the event. Two labels were created for this purpose, both inspired by the Propbank framework: 'writer-ARG0', when the entity plays roles covered by this argument, such as 'Agent' or 'Perceiver', 'writer-ARGx', if they play roles covered by other argument types, like 'Patient'. Even though grouping such arguments slightly reduces the expressiveness of the PropBank framework, it has the advantage of helping the annotators to focus on a more general distinction between events in which writers have an active role and events in which they have not.

Identification of events, and their taxonomy. Events had to be annotated according to the TimeML scheme [Pustejovsky et al., 2010] and were categorized according to a

[10] https://github.com/timjogorman/RicherEventDescription

subset of TimeML event types tag: ASP-EVENT to mark all verbs conveying aspectual information (e.g.: start); REP-EVENT, for verbs that reports about other events other states and events (e.g.: tell); STATE for all the situations that hold true for a certain period (e.g.: live); EVENT for every change of state (e.g.: move). Each annotation could contain only one EVENT or STATE, while REP-EVENT and ASP-EVENT may be combined within the same annotation. In Example 4, 'traveled', an EVENT, and 'working', a STATE, need two separate annotations. The token 'started' can be marked as an ASP-EVENT and co-occur with 'working'.

4. Then, she traveled to Venezuela, where she started working in linguistics at the Department of Justice of Venezuela

Since some sentences contained nominal utterances and there were semantically void verbs like the copular *be* [Bonial and Palmer, 2016], guidelines allowed for the annotation of names as events or states in subordinate clauses like "After a brief time in Toronto", or in nominal predicates such "He was a professor". The annotation of nominal events follows the NomBank project [Meyers et al., 2004], that provides a list of names organized by their frames.

Identification of arguments containing a location or an organization. The third component of the guidelines was aimed at identifying the relation between the writer and some named entities that may signal their migration or their condition of being a migrant in a given place. Annotators were asked to select the entire semantic argument containing a location or an organization, and to mark the latter as 'ARGx-ORG', and the former 'ARGx-LOC'. The focus of this annotation stage was not to identify the specific semantic argument, but to label the cases in which a named entity is part of a semantic role. This allowed to refine clusters of arguments and map them onto existing taxonomies. For instance, in 'He works for $organization', the ARGx-ORG may be mapped onto the VerbNet 'Beneficiary' thematic role.

Identification of temporal arguments. Finally, the guidelines establish the annotation of temporal arguments. Rather than identifying only the token triggering a time expression, the entire argument had to be selected and labelled as 'ARGM-TIME'. For instance, in the example "In 1974 he left South Africa" the entire semantic argument "in 1974" had to be annotated.

A full annotation of Example 3 is the following:

ϵ_1 = <he, WRITER>

ϵ_2 = <left, LEAVE>

ϵ_3 = <South Africa, LOCATION>

ϵ_4 = <living, LIVE>

ϵ_5 = <in South Africa, LOCATION>

ϵ_6 = <in 1974, TIME>

L_1 = <ϵ_1, ϵ_2, writer-ARG0>

L_2 = <ϵ_3, ϵ_2, ARGx-LOC>

L_3 = <ϵ_1, ϵ_4, writer-ARG0>

L_4 = <ϵ_5, ϵ_4, ARGx-LOC>

L_5 = <ϵ_6, ϵ_2, ARGM-TIME>

4.3.2 Annotation Task and Results

The annotation task involved four annotators who evaluated 1,000 sentences sampled from 7,979 Wikipedia English pages of Transnational writers. One of them (ann_01 in Table 4.6) evaluated all 1000 sentences, while the others annotated respectively 200 (ann_02), 100 (ann_03), and 200 (ann_04) sentences. The annotation has been performed on Label Studio[11], an Open Source platform that easily allows the organization of chunk annotation tasks. Annotators were asked to provide one separate annotation for every EVENT or STATE identified in each sentence. As it is shown in Figure 4.2, the same sentence has received two separated annotations. The first is the chunk 'jailed' labelled as an EVENT, the second is the chunk 'detained', labelled as a STATE.

[11] https://labelstud.io/

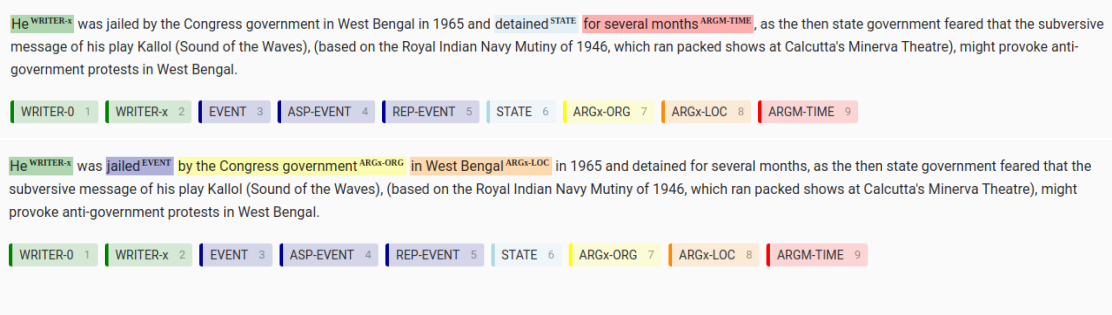

Figure 4.2: Two examples of annotation in Label Studio.

annotator	Event	State	Writer-ARG0	Writer-ARGx	ARGx-LOC	ARGx-ORG	ARGM-TIME
ann_01 (baseline ann_02)	0.83	0.72	0.90	0.87	0.78	0.75	0.91
ann_01 (baseline ann_03)	0.83	0.76	0.91	0.92	0.38	0.75	0.94
ann_01 (baseline ann_04)	0.84	0.66	0.91	0.90	0.65	0.83	0.85
ann_02 (baseline ann_01)	0.83	0.66	0.91	0.89	0.85	0.94	0.94
ann_03 (baseline ann_01)	0.82	0.64	0.93	0.95	0.91	0.92	0.94
ann_04 (baseline ann_01)	0.84	0.61	0.91	0.89	0.75	0.70	0.87
Average	**0.83**	**0.67**	**0.91**	**0.90**	**0.75**	**0.81**	**0.91**

Table 4.6: Inter-Annotator Agreement (F-measure).

Since the number of potential annotations for each sentence is not fixed, the Inter-Annotator Agreement (IAA) was computed through averaged pairwise F-measure: in this setting, the annotations of one annotator are used as the reference against which the annotations of the other annotator are compared. In order to maximize the agreement between annotators, I did not only consider the exact match between chunks, but also cases of partial agreement in which one chunk contained the other. Adopting such an approach has allowed to resolve some recurrent inconsistencies and to provide a coherent annotation throughout the corpus. Example 5 shows a partial agreement where one annotator only labeled the token expressing the verb and the other also annotated the auxiliary. In all these cases I only kept the one-token annotation, since they are annotated as such in existing resources. In Example 6 it is possible to observe a partial agreement between two annotations of ARGx-ORG. In these cases I kept the larger, in order to preserve the semantic role of the argument containing an entity of the type organization.

5. awarded / was awarded

6. the United Nations / to the United Nations

Table 4.6 shows the F-measure of the agreement between annotators for each class. Agreement is larger than 0.8 in almost all classes, with the exception of STATE and ARGx-LOC. From a qualitative analysis I observed that many annotators failed to recognize nominal events in propositions like one expressed in Example 7. Lower agreement in ARGx-ORG identification seems to be caused by the broadness of such a type of entity that results in a variety of irrelevant usages for the annotation task, as in Example 8, that has been wrongly labeled by one annotator.

7. after one year of **studies**

8. when Sri Lanka banned the burka on 2019, Nasrin took to **Twitter** to show her support for the decision

The resulting corpus contains 1,489 semantic annotations from 834 sentences. Table 4.7 summarizes the number of Semantic Types in the corpus, in which there are 894 events and 695 states. Furthermore, 215 aspectual or reported events were annotated; they occurred in 72 semantic annotations. In 143 cases, they jointly appear with an event or a state (e.g.: 'he $[started]^{ASP-EVENT}$ $[working]^{STATE}$'). Writers hold the semantic role of agent in 1,205 annotations, other roles in 445. Arguments containing an organization or a location are 1,203.

Semantic Type	Occurrences
EVENT	894
STATE	695
ASP-EVENT	114
REP-EVENT	101
writer-ARG0	1,090
writer-ARGx	388
ARGx-LOC	532
ARGx-ORG	671
TIME	525

Table 4.7: All the occurrences of Semantic Types in the corpus.

These statistics reveal some interesting aspects about limitations of approaches for biographical event detection that rely on existing SRL and NER tools (see Section 4.3.1). 164 of sentences sampled in such a way have been removed after the annotation since they did not express any biographical fact. 281 annotated sentences express a biographical event but not a semantic argument that includes a named entity of the type GPE or ORG. Together, non-relevant sentences and sentences that does not contains a relevant named entity represent the 44.5% of all the gathered ones, thus showing that existing SRL and NER tools are not fully reliable to obtain fine-grained biographical information. In Section 4.3.3 two mapping strategies will be performed to test the potential interoperability between my corpus and OntoNotes.

4.3.3 Mapping

The annotation guidelines and the corpus presented in this paper constitute a first, yet essential step towards the development of a system for the automatic extraction of biographical events. In this Section I test if and how the current corpus could be extended with existing resources to obtain an appropriate training dataset for such a task. We propose two types of experiments: (i) a comparison of semantic roles between the two corpora; (ii) the analysis of predictions made by a Language Model (LM) finetuned on my corpus over the OntoNotes dataset.

4.3.3.1 Mapping Through Comparison

The first mapping experiment aims at exploring the distribution of semantic roles in OntoNotes containing at least one ORG or GPE in order to understand if they align with my corpus. As I described in the annotation scheme (Section 4.3.1), argument types in my corpus were not specified during the annotation stage to maximize the agreement between the annotators and to provide flexible annotations to semantic links between verbs and locations. In this experiment I explore which are the most recurring semantic roles in OntoNotes that associated to the top-10 recurring events annotated in my corpus.

OntoNotes contains a multi-layer annotation of texts from several domains (e.g., newswires, magazine articles, broadcast news). For each such domain, a PropBank-based semantic annotation and the annotation of named entities is provided. The data set is composed of 99,974 sentences, 140,679 rolesets, and 554,307 semantic arguments. In order to align the two corpora, I extracted all verb occurrences and their arguments. Then, I computed the percentage of arguments containing a named entity of the type ORG, or GPE. As a result I obtained 19,763 sentences with at least a semantic link between a verb and an argument containing one of these named entities and a total number of 49,729 links of the type <VERB,ARGUMENT>.

The analysis focuses on the 10 most frequent verbs in my corpus: for each of them I computed its absolute frequency in semantic links derived from OntoNotes and the two most recurring semantic arguments associated with it.

Results of the mapping is shown in Table 4.3.3.1. From a first overview it emerges that semantic arguments of the type location that contain a GPE or an ORG (Example 11) are never the most frequent in OntoNotes. The only exception is 'teach', which is associated 18 times out of 34 to this type of argument.

11. Ko Ching [...] taught English [**at Taipei** First Girls' High School.]_ARGM-LOC

Argument 0, which at an abstract level represents the agent of the event, is the most frequent semantic link with 6 verbs. A recurrent personification of entities of the type ORG or GPE is identifiable in this pattern. In Example 12 'the EPA' is linked to 'work'. This shows that verbs that typically express biographical events often play a different function when they interact with ORG and GPE.

12. **[the EPA]**_ARG0 has also been working to draft.

Argument 1, which typically represents the patient of the event, is the most linked to two verbs and the second most linked to four ones. This argument type does not express biographical knowledge, though. In Example 13, 'Texas' is only indirectly related to a life event, since it is part of a broader argument associated to 'write' with the role 'things

Event	BioSrl	OntoNotes	ARGS
write	69	95	ARG1, ARG0
work	60	259	ARG0, ARGM-LOC
receive	56	187	ARG0, ARG1
study	41	74	ARG0, ARGM-LOC
publish	39	70	ARG0, ARG1
win	36	131	ARG0, ARG1
award	34	34	ARG0, ARG2
teach	28	34	ARGM-LOC, ARG0
attend	28	55	ARG1, ARG0
move	25	186	ARG2, ARG1

Table 4.8: The ten most frequent events and states within my corpus. Column *BioSRL* specifies the number of occurrences of each event in my corpus; *OntoNotes* its occurrences in OntoNotes, and *ARGS* the semantic arguments in OntoNotes related to the event that contain an entity of the type ORG or GPE.

written'. In this case, it is possible to observe that the argument itself has an inner structure where 'Texas' is related to the name 'attitude'.

13. In the past , writes Houston Chronicle columnist Jim Barlow , [outlanders were accepted only after passing a series of tests to prove they had the "right" **Texas** attitudes]_ARG1

The experiment shows that the most recurrent biographical events in my corpus do not have a corresponding semantic link in OntoNotes. Roles of GPE and ORG entities are different in the two corpora and many times they appear in arguments with a complex structure for which there is no available annotation.

4.3.3.2 Mapping Through Prediction

An alternative approach for mapping my corpus with OntoNotes is based on predictions obtained with a finetuned LM over my annotations for a Sequence To Sequence task. In order to do so, I first transformed all the annotations from my corpus in sequences where the lemmatized version of the verb precedes the sentence where it appears, labeling only the verb that trigger the event and squeezed arguments of the type ARGx-LOC and

ARGx-ORG in a unique label of the type 'location'. In Example 14, all tokens in the chunk 'In Norther Rhodesia' were labeled as 'location', while 'sitting' was labeled as 'event'.

14. sit [SEP] In Northern Rhodesia (now called Zambia), Achebe found himself sitting in a whites-only section of a bus to Victoria Falls.

Once I had processed the sentences, I finetuned a DistilBERT checkpoint [Sanh et al., 2019][12] using them as a training set. I randomly sampled the corpus 5 times, splitting it in 80% training, 10% validation, and 10% test set and for each sample I run a training for 40 epochs. I kept the model checkpoint that obtained the best $F-1$ score over the test set and used it to predict tuples of the type <event, location> in OntoNotes sentences encoded according to the Example 14. The total number of sentences is $140,679$, which is equivalent of all the possible roleset for each sentence in the corpus.

After the prediction step, I obtained $15,806$ sentences and $16,137$ where at least one verb and one location have been predicted. As can be observed in Table 4.9, the most recurring verbs obtained through predictions differ from ones obtained through comparison (Table 4.3.3.1). Only 'work' is present in both mappings. An overview of semantic arguments shows a more relevant presence of semantic links that seem to have a clearer correlation with events expressed by verbs. 'go' and 'come' are more likely to occur with the Argument 4, which in their rulesets represents the 'end point'. The argument of type location is the most frequent semantic link of four verbs. Argument 1 is relevant for the verb 'leave', since it represents the 'starting point' in leave's roleset. Light verbs like 'be' and 'take' [Chen et al., 2015, Bonial and Palmer, 2016] and 'say', which triggers reporting events (see Pustejovsky et al. [2010]) are frequently associated with more vaguer semantic links.

It is worth mentioning that in 33% of cases, predictions based on my corpus result in parts of the sentence that are not annotated with any semantic argument in OntoNotes. The chunk 'in Harlem' in Example 15 has been automatically recognized as an argument

[12] https://huggingface.co/docs/transformers/model_doc/distilbert

Event	Occ.	ARGS
go	585	ARG4, ARG1
be	277	ARG2, ARG1
take	247	ARG2, ARG1
live	187	ARGM-LOC, ARG0
work	186	ARGM-LOC, ARG1
come	181	ARG4, ARG2
leave	162	ARG1, ARG0
say	160	ARG1, ARG0
hold	155	ARGM-LOC, ARG0
do	152	ARGM-LOC, ARG0

Table 4.9: The list of most recurring biographical events predicted in OntoNotes. Column *ARGS* specifies the arguments to which they are linked to.

of the type location, but is not part of the roleset of 'go' in this sentence. This is coherent with PropBank guidelines, according to which arguments are linked to verbs for the joint presence of semantic and syntactic elements. However, such an approach unfolds the identification of semantic connections between events and all elements that are outside their rolesets.

15. go.02 [SEP] you know , he was going to a lot of sessions for a while in Harlem and stuff like that.

The creation of a corpus annotated for biographical events has been a crucial step to automatically extract biographical knowledge from texts. Existing authoritative resources cannot be easily aligned with my resource for such a purpose, though. Despite the fact that the corpus has been annotated with guidelines that rely on ISO for semantic annotation [Pustejovsky et al., 2010, Bunt and Palmer, 2013], mapping its annotations with OntoNotes resulted in a series of issues in the recognition of semantic links between verbs and their arguments. Additional resources and mapping experiments are needed to develop reliable systems for biographical event detection.

4.4 Building a Semantic Resource for Biographical Event Detection

In previous works (Section 4.2 and 4.4) I explored a series of strategies to detect biographical events by reusing existing resources. From these attempts emerged a series of limitations. First, resources based on verbs and their arguments like VerbNet [Schuler, 2005] and PropBank [Kingsbury and Palmer, 2002] cannot be easily reused for specific tasks, given their complex nature. Their focus on verb is a second core limitation: nominal events are not annotated in these resources, thus hindering the retrieval of a relevant number of biographical information. A final issue is the fact that I operated at the sentence level, while biographies are often long documents where the need of disambiguate between events referring to the subject entity of a biographies and events that do not is crucial.

In this work I present WikiBio[13], a new corpus annotated for biographical event detection at the document level, composed of 20 Wikipedia biographies. The corpus includes all the events which are associated with the entity target of the biography and is designed to be interoperable both with corpora based on TimeML annotation type [Pustejovsky et al., 2010] and with corpora that rely on SRL [Kingsbury and Palmer, 2002]. Additionally, the corpus is designed to perform coreference resolution, in order to only keep relevant events for biographies' subjects.

The work is organized as follows. In Section 4.4.1 I present my annotation scheme that covers both the annotation of events and entities within biographies. Section 4.4.2 describes the corpus and compares it to 5 resources annotated for event detection and coreference resolution. In Section 4.4.3 I report a series of experiments of biographical event detection that rely on the integration between WikiBio and other resources adapted for this task.

[13] https://github.com/marcostranisci/semantic-aware-bias/tree/main/resources/wikibio

4.4.1 Annotation Tasks

Since the biographical event detection task consists in annotating all tokens that trigger events related to the person who is the subject of a biography, annotation guidelines focus on two separate subtasks: (i) the identification of all the mentions of the target entity and the resolution of its co-reference chains; and (ii) the identification and linking of all the events that involve the target entity[14].

Entity annotation. The entity annotation subtask requires the identification of all mentions of a specific Named Entity (NE) [Grishman and Sundheim, 1996] of type Person, which is the target of the biography and all its coreferences [Deemter and Kibble, 2000] within the Wikipedia biography. For the modeling of this subtask, I used the GUM corpus [Zeldes, 2017], introducing different guidelines about the following aspects: *i)* only the mentions of the entity-target of the biography must be annotated; *ii)* mentions of the target entity must be selected only when they have a role in the event (Example 1, where the possessives "his" is not annotated); and *iii)* indirect mentions of the target entity must be annotated only if they are related to biographical events (Examples 2 and 3).

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

2. **He** married Wendy Bruce, whom **he** had known since **they** were teenagers.

3. In 1985, the Biafran Civil War novel **Sozaboy** was published.

Event Annotation. Although there is an intuitive understanding of how to identify event descriptions in natural language texts, there is quite a large variability in their realizations [Pustejovsky et al., 2003b]. Araki et al. [2018] point out that some linguistic categories, e.g., nouns, fits on an event *continuum*. This makes the identification of event mentions a non trivial task. Our event annotation task mainly relies on TimeML

[14]Guidelines are available in Appendix C

Annotation Layer	A0 & A1	A0 & A2
Event	0.72	0.86
Entity	0.65	0.86
LINK	0.76	0.64
CONT_MOD	0.71	0.64

Table 4.10: Inter-Annotator Agreement (Cohen's Kappa).

[Pustejovsky et al., 2003a] and RED [O'Gorman et al., 2016], where 'event' is "a cover term for situations that happen or occur." [Pustejovsky et al., 2003a]. Events are annotated at a single token level with no restrictions on the parts of speech that realize the event. Following Bonial and Palmer [2016], I introduced a special tag (`LINK`) for marking a limited set of light and copular verbs, as illustrated in Example 4. The adoption of `LINK` is aimed at increasing the compatibility of the annotated corpus with OntoNotes, the resource with the highest number of annotated events.

4. Ken Saro-Wiwa <LINK>**was**<LINK/> a Nigerian <EVENT>**writer**<EVENT/>
   <LINK source='be' target ='writer' />.

Lastly, to enable automatic reasoning on biographies, I annotate the contextual modality of events [O'Gorman et al., 2016]. In particular, to account for the uncertainty/hedged modality, i.e., any lexical item that expresses "some degree of uncertainty about the reality of the target event" [O'Gorman et al., 2016], I have defined three uncertainty values: `INTENTION`, for marking all the events expressing an intention (like 'try' or 'attempt'); `NOT_HAPPENED`, for marking all events that have not occured; `EPISTEMIC`, which covers all the other types of uncertainty (e.g., opinion, conditional). The uncertainty status of the events is annotated by linking the contextual modality marker and the target event, as illustrated in Example 5:

5. Feeling alienated, he **decided** to **quit** college, but was **stopped** [...]
   <CONT_MOD source ='decided' target = quit' value='INTENTION' / >
   <CONT_MOD source ='stopped' target = 'quit' value='NOT_HAPPENED' / >

Corpus Annotation and IAA. The annotation task was performed by three expert annotators (two men and one woman, who contributed to the research), near-native speakers of British English, having a long experience in annotating data for the specific task (event and entity detection). One annotator (A0) was in charge of preparing the data by discarding all non-relevant sentences to speed-up the annotation process. This resulted in a final set of 1,691 sentences containing at least one mention of a target entity. The entity and event annotations were conducted as follows: A0 annotated the entire relevant sentences, while a subset of 400 sentences was annotated by A1 and A2, who respectively labeled 200 sentences each. Since all the annotations have been provided at the token level and each token could be annotated with only one label, the total number of annotated items is fixed, unlike the approach adopted for BioSRL (Section 4.3) in which annotators could label a non-fixed number of biographical events. Therefore I rely on Cohen's kappa to compute the pairwise IAA (Table 4.10). In general, there is a fair agreement across all the annotation layers. At the same time, I observe a peculiar behavior across the annotators: there is a higher agreement between A0 and A2 for the event and entity layers when compared to A0 and A1, but the opposite occurs with the relations layers (LINK and CONT_MOD).

For the events, the higher disagreement is due to nominal events, often misinterpreted as not bearing an eventive meaning. For instance, the noun "trip" in example 6 was not annotated by A1.

6. When Ngũgĩ **returned** to America at the end of **his** month **trip** [...]

For the entities, I observed that disagreement is due to two reasons. The first is the consequence of a disagreement in the event annotations. Whenever annotators disagree on the identification of an event, they also disagree on the annotation of the related entity mention, as in the case of the pronoun 'his' in example 6. Another reason of disagreement regards indirect mentions. Annotators often disagree on annotation spans, as in "Biafran Civil War novel Sozaboy was published" where A1 selected 'SozaBoy', while A2 'novel Sozaboy'. When it comes to LINK, problems are mainly due to the identification of light

verbs. Despite the decision of considering only a close set of copular and light verbs to be marked as LINK (see Bonial and Palmer [2016]), annotators used this label for other verbs, such as 'begin' or 'hold'.

7. Walker **began** to take up reading and writing.

Corpus	**Size**	**Text types**	**Relevant task**
TimeBank	7,471 events	news	Event detection
OntoNotes	159,938 events, 22,234 entity mentions	frame-theory	Event & Entity detection
NewsReader	594 events	TimeML	Event detection
GUM	9,762 entity mentions	biographies	Entity detection
LitBank	7,383 events	literary works	Event Detection

Table 4.11: A list of five existing resources that have been employed in the biographical event detection task.

4.4.2 WikiBio: Overview and Comparison with Other Resources

The WikiBio corpus is composed of 20 biographies of African, and African-American writers extracted from Wikipedia for a total amount of 2,720 sentences. Among them, only 1,691 sentences include at least one event related to the entity target of the biography. More specifically, there are 3,290 annotated events, 2,985 mentions of a target entity, 343 LINK tags, and 75 CONT_MOD links.

Corpora size and genres We compare WikiBio against five relevant existing corpora that, in principle, could be used to train models for biographical event detection: GUM [Zeldes, 2017], Litbank [Bamman, 2020], Newsreader [Minard et al., 2016], OntoNotes [Hovy et al., 2006], and TimeBank [Pustejovsky et al., 2003b]. For each corpus, I took into account the number of relevant annotations and the types of texts. As it can be observed in Table 4.11, corpora vary in size and genres. OntoNotes is the biggest one and includes 159,938 events, and 22,234 entity mentions. The smaller is NewsReader,

with only 594 annotated events. TimeBank and LitBank are similar in scope, since they both include about $7.5K$ events, while GUM includes $9,762$ entity mentions.

Text types. With the exception of GUM, which includes 20 biographies out of 175 documents, all other corpora contain types of texts other than biographies such as news, literary works, and transcription of TV news. To get a high-level picture of the potential similarities and differences in terms of probability distributions, I calculated the Jensen-Shannon Divergence [Menéndez et al., 1997]. Such metric may be useful for identifying which corpora are most similar to WikiBio. The results show that WikiBio converges more with GUM (0.43), OntoNotes (0.48) and LitBank (0.49) rather than with TimeBank (0.51) and Newsreader (0.54). Such differences have driven the selection of data for the training set described in Section 4.4.5.

Annotations of entities, events, and coreference The distribution of the target entity within biographies in the WikiBio corpus has been compared with two annotated corpora for coreference resolution and named entity recognition: OntoNotes [Hovy et al., 2006] and GUM [Zeldes, 2017]. Since such corpora were developed for identifying the coreferences of all NEs in a document, I modified annotations to keep only the most frequent NEs of type 'person' in each document. The rationale was making these resources comparable with WikiBio, which includes only the coreferences to a single entity, namely the subject of each biography. After doing that, I computed the ratio between the number of tokens that mention the target entity and the total number of tokens, and the ratio between the number of sentences where the target entity is mentioned against the total number of sentences. While this operation did not impact on GUM, in which 174 out of 175 documents contain mentions of people, it had an important impact on OntoNotes, in which $1,094$ documents (40%) do not mention entities of the type Person. Tokens mentioning the target entity are 5% on OntoNotes, 8.7% on GUM and 4% on WikiBio. Such differences can be explained by the average length of documents in these corpora, which is of 388 tokens in OntoNotes, 978 in GUM, and $3,754$ in WikiBio. As a matter of fact, if the percentage of sentences mentioning the target-entity is considered

instead of the total number of tokens, WikiBio shows an higher ratio (61.7%) of sentences mentioning the target entity, than OntoNotes (20.8%) and GUM (42.6%).

	WikiBio	**GUM**	**Litbank**	**Newsreader**	**OntoNotes**	**Timebank**
WikiBio	0.00	0.43	0.49	0.54	0.48	0.51
GUM	0.43	0.0	0.49	0.54	0.39	0.49
Litbank	0.49	0.49	0.00	0.55	0.42	0.51
Newsreader	0.54	0.55	0.54	0.00	0.48	0.45
OntoNotes	0.48	0.39	0.42	0.48	0.00	0.40
TimeBank	0.51	0.49	0.51	0.45	0.40	0.00

Table 4.12: The similarity between corpora for event annotation computed with the Jensen-Shannon Divergence.

The three most frequently occurring lemmas in the WikiBio corpus seem to be strongly related to the considered domain: 'write' represents 3.2% of the total, 'publish' 2.9%, and 'work' 1.8%. 'Return' (1.3%) appears to have a more general scope, since it highlights a movement of the target entity from a place to another. The comparison with other corpora annotated for event detection shows differences concerning the most frequent events. The top three in OntoNotes [Bonial et al., 2010] are three light verbs: 'be', 'have', and 'do'. This may be intrinsically linked to its annotation scheme which considers all verbs as candidates for being events, including semantically empty ones. NewsReader [Minard et al., 2016] and TimeBank [Pustejovsky et al., 2003b] include two verbs expressing reporting actions among the top five, thus revealing that they are corpora of annotated news. Litbank [Bamman, 2020], which is a corpus of 100 annotated novels, includes in its top-ranked events two visual perception verbs and two verbs of movement, which may reveal the centrality of characters in this documents. The event 'say' is top-ranked in all the five corpora.

4.4.3 Detecting Biographical Events

In this section I describe a series of experiments for the detection of biographical events. Experiments involve the use of the existing annotated corpora for two tasks: entity mentions detection (Section 4.4.4) and event detection (Section 4.4.5). In both cases I used a 66 million parameters DistilBert model Sanh et al. [2019]. In this setting the WikiBio corpus is both used as part of the training set and as a benchmark for testing

how well existing annotated corpora may be used for the task. For such experiments a NVIDIA RTX 3030 ti was used. The average length of each fine-tuning session was 40 minutes.

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 4.13: Results of entity detection experiments.

4.4.4 Entity Detection

For this task I adapted the annotations in OntoNotes [Hovy et al., 2006] and GUM [Zeldes, 2017] keeping only mentions of the most frequent entities of type 'person'. As a result I obtained 870 documents from OntoNotes, 174 from GUM.

The WikiBio corpus was split into three subsets: five documents for the development, 10 for the test, and five for the training. Given the imbalance between the existing resources and WikiBio, I always trained the model with six different samples of 100 documents (Table 4.13), in order to reduce the overfitting of the model over the other datasets.

Experiments consist in training a DistilBert model for identifying all the tokens mentioning the target entity of a given model and were performed on six different training sets. Since the focus of my work is to develop a model for detecting biographical events, WikiBio was used as development set for better monitoring its degree of compatibility with existing corpora. Following the approach by Joshi et al. [2020], I split each document into sequences of 128 tokens, and for each document I created one batch of variable length containing all the sequences. Each experiment has been repeated three times with three different random samples (e.g., I randomly selected 100 documents from GUM three times and averaged the F_score obtained through all runs). Table 4.13 shows the results of these experiments. As it can be observed, including the WikiBio corpus in the

training set did not result in an increase of the performance of the model. This may be due to the low number of WikiBio documents in the training.The highest performance was obtained in two experiments: one using a training set only composed of documents from OntoNotes, which obtained a F-score of 0.808, and one with a miscellaneous of 50 OntoNotes and 50 GUM documents, that obtained 0.792. To understand if the difference between the two experiments is significant, I performed a One-Way ANOVA test over the train, development, and test F-scores obtained in both experiments. The test returned a p-value of 0.44, which confirms a non-significant difference between the two results

4.4.5 Event Detection

Event Detection experiments were guided by the comparison between WikiBio and the resources for event detection described in Section 4.4.2. Since OntoNotes was annotated according to the PropBank guidelines Bonial et al. [2010], which only consider verbs as candidates for such annotation, I partly modified its annotations before running the experiments. I first adapted the OntoNotes semantic annotation by replacing light and copular verbs [Bonial and Palmer, 2016] with nominal [Meyers et al., 2004] and adjectival events. Then I ran a battery of experiments by fine-tuning a DistilBert-based model using each dataset for training, and a series of miscellaneous of the most similar corpora to WikiBio according to the Jensen-Shannon Divergence metric (Table 4.12). Since I was concerned with both assessing the effectiveness of WikiBio for training purposes and testing how far biographic events can be extracted, I designed my training and testing data as follows. WikiBio was employed in different learning phases: in devising the training set (i.e., existing resources were employed either alone or mixed with WikiBio); additionally, the development set was always built by starting from WikiBio sentences. Finally, I always tested on WikiBio data.

As for the entity-detection experiments, the 1,691 sentences containing events annotated in the WikiBio corpus were split into three sets of equal size that were used for training (564), development (563), and testing (564). Given the disproportion between OntoNotes and other corpora, I sampled a number of sentences for training which did

not exceeded 5,073, namely three times the number of sentences annotated in my corpus. Such length was fixed also for miscellaneous training sets.

Experiments were organized in two sessions. In the first session I fine-tuned a DistilBert model for five epochs, using as training set the five corpora presented in Section 4.4.2 individually as well as three combinations of them: *i)* misc_01, a miscellaneous of sentences extracted on equal size from all corpora; *ii)* misc_02, in which sentences from NewsReader, the most different corpus with WikiBio (Table 4.12), were removed; *iii)* misc_03, a combination of sentences from OntoNotes and Litbank, namely the two most similar corpora with WikiBio. The model was fine-tuned on these training sets both with and without a subset of the WikiBio corpus for a total of 16 different training sets. In addition, I also fine-tuned and tested WikiBio alone. I then continued the fine-tuning only for the models which obtained the best F-scores. The batch size was set to 10 and each fine-tuning has been repeated three times of a different random sample. In Table 4.14 the average F_score of the three runs is reported for each training set composition.

Observing Table 4.14, it emerges that, differently from entity-detection experiments, including a subset of WikiBio in the training set, even if in a small percentage, always improves the results of the classifier. This especially happens for Litbank (+0.191 F-Score), and TimeBank (+0.031 F-Score).

When looking at results of finetuning for single corpora, it emerges that the model trained on the modified version of OntoNotes and TimeBank obtains the best scores. Such results are interesting for two reasons. They confirm the intuition that OntoNotes annotations may be easily modified to account for nominal and adjectival events. They also confirm the high compatibility of WikiBio and TimeBank guidelines (Sect. 4.3.1.2). Even if the latter is more divergent from WikiBio than other corpora, it seems to be compatible with it. As expected for its limited size and high divergence with WikiBio, the training set based on NewReader sentences obtains the worst results, with an F-Score below 0.5.

Results of miscellaneus training sets are interesting as well: they generally result in models with better performance, and they seem to work better on the basis of their

divergence with WikiBio. Trained on misc_01, a combination of all corpora, the model scores 0.827, which is below the result obtained with the modified version of OntoNotes. If Newsreader is removed, the model obtains 0.831, and 0.832 if also TimeBank is removed. It is also worth mentioning the delta between the F-score on the training and the test sets, which is −0.054 for misc_01, −0.029 for misc_02, and −0.013 for misc_03.

After the first fine-tuning step, I performed a One-Way ANOVA for testing the significance of differences between experiments. Analyzed in such a way, the four best-ranked models never showed a p-value of 0.5, which means that there are no significant differences between them. Thereby, I kept them for the second fine-tuning step that consists on training the model for 15 epochs on these datasets. Absolute results (Table 4.14) show that the model trained on Timebank obtained the best F-Score. However, as for the entity detection experiments, I considered the deltas between the training and test F-scores to select the best model for my analysis. All models acquired by employing a miscellaneous training set obtained a lower delta between training and test, and scored a similar F-Score.

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 4.14: Results of event detection experiments: complete table

4.5 Conclusion of the Chapter

In this chapter I presented three different approaches to biographical event detection, all focused on the reuse of existing corpora annotated for the detection of events. My attempts showed that state of the art resoruces [Schuler, 2005, Bonial et al., 2010] can support the task but must be integrated with specifically-designed resources that are actually missing. In order to fill this gap I developed two new corpora and exploited them to train a classifier capable of identifying only events directly related to the subject of a biography. The classifier will support my analysis of *representational* bias in Chapter 5 and my mitigation strategies of *allocative* bias in Chapter 6.

Part II

Experimental Analysis of Bias

Chapter 5

Representational Bias in Datasets

Wikipedia is one of the most important sources of collaborative knowledge, created and maintained by a community of almost 47 millions contributors[1]. Wikipedia is not only used by the general public, but is also adopted as a knowledge base in a number of heterogeneous researches. For instance, a Wikipedia dump has been used to train GloVE [Pennington et al., 2014] and few years later Kenton and Toutanova [2019] included it in the pretraining dataset of BERT. Wikipedia has also been considered a benchmark for the detection of high quality contents to be used for training LLMs. Brown et al. [2020] included it in a corpus of curated datasets that has been used to filter out poor-quality documents from the Common Crawl Dataset[2] during the composition of the dataset that has been used to train GPT-3.

Such wide adoption of Wikipedia in NLP is often acritical, though. Researchers tend to overlook existing findings about the presence of different forms of inequalities in this encyclopedia, which emerges and is actively addressed by the Wikipedia community itself[3] and that has been assessed by a high number of researches. Graells-Garrido et al. [2015] showed that Wikipedia biographies are not only skewed towards men, but they are also affected by stereotypical associations between gender and words that trigger social roles. In a similar analysis, Sun and Peng [2021] automatically extracted events from

[1] https://en.wikipedia.org/wiki/Special:Statistics, last accessed 2023-02-29.
[2] https://commoncrawl.org/
[3] https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Countering_systemic_bias

biographies, showing that in women's biographies it is more likely to find events related to the sphere of marriage and parenthood, while men are mostly associated to career-type events.

Despite the evidence about such disparities, NLP experiments aimed at debiasing Wikipedia as a datasets are few if compared to the efforts that are made to mitigate bias in word embeddings trained on Wikipedia (see Bolukbasi et al. [2016]). A second limitation regards the scope of NLP research on bias in Wikipedia. Most of them focus on gender disparities, while limited efforts have been made to identify and study the racial gap or their intersections. Intersectionality [Crenshaw, 2017] is a theoretical framework according to which discrimination may be fully understood only if different axes of inequality are jointly observed (e.g., race and sex). In her foundational work, Crenshaw [1989] showed that focusing only on gender or race was not sufficient to explain the unemployment discrimination that affected African American women workers. Some recent works in the NLP field adopt an intersectional approach, but mainly focus on its impact on annotated corpora [Lalor et al., 2022] or on socio-demographic features of labeled texts [Maronikolakis et al., 2022].

In this chapter I propose an approach to *representational* bias detection aimed at overcoming two existing limitations: *i.* the limited number of NLP researches on biases in Wikipedia; *ii.* the lack of an intersectional view on the presence of biases in this source of knowledge. In my approach I compare biographical events extracted from Wikipedia pages of people belonging to different groups on the basis of four socio-demographic features – gender, ethnicity, age, and occupation – in order to explore stereotypical associations between events and groups characterized by one or more features.

The approach combines the biographical event detection method described in Chapter 4 and structured knowledge extracted from Wikidata [Vrandečić and Krötzsch, 2014] to perform an analysis of bias in Wikipedia biographies along four interrelated axes: gender, race, age, and occupation. Results show the presence of hidden stereotypes that only unfold when different socio-demographic features are jointly considered. The chapter is organized as follows: in Section 5.1 I describe the methodological setup of my analysis

and the distribution of the dataset gathered from Wikidata. In Section 5.2 I present the results of my analysis, examining *representational* bias related to gender, ethnicity, age, and occupation.

5.1 Methodological Setup

The bias detection analysis is designed to simulate balanced distributions between social groups in order to provide an analysis of *representational* biases that is not affected by the skewness of Wikipedia towards certain categories of people (e.g., men and Westerners). Such an imbalance may hinder the identification of biographical pattern associated to specific social groups or bring out associations that are not statistically relevant. To address this issue, I implemented a procedure based on two steps.

- In the first step (Section 5.1.1) I gather structured information about people on Wikidata and encode them to facilitate the comparison between social groups. The encoding introduces the dichotomy Western/Transnational introduced in Section 3.3, according to which people are classified as Transnational if they were born in a former colony with a Human Development Index below 0.8 or if they belong to an ethnic minority in a Western country.

- In the second step I describe the Monte Carlo simulation that I implemented to perform comparisons between groups of biographies of the same size. I also introduce the list of groups that I selected for the intersectional analysis and present some descriptive statistics about the extracted biographical events.

5.1.1 Data Gathering

In order to perform my analysis, I first collected all 9,552,705 entities of the type Human (WDQ5) from a recent Wikidata dump[4] and their English Wikipedia pages through the Wikipedia API. Since my analysis focuses on the intersection between ethnicity, gender,

[4] https://academictorrents.com/download/229cfeb2331ad43d4706efd435f6d78f40a3c438.torrent

age, and field of work, I only kept entities that are associated with a place of birth (P19), gender or sexual orientation (P21), date of birth (P569), and occupation (P106). The total number of people linked to all these properties in Wikipedia is $1,567,865$, among which only $768,294$ (49%) have an English Wikipedia page.

All 'place of birth' properties were clustered in continents. In such a way it was possible to distinguish all people who are Transnational (Section 3.3.1) from those who are not. The 'sex or gender' property was treated similarly. I grouped people in three categories: 'men', 'women', and 'non-binary'. The latter was necessary because Wikidata pages about people who do not identify themselves as 'man' or 'woman' are almost not represented in the knowledge base. If combined in a single category, all people associated to one of the 23 genders that are not of the type 'male' or 'female' are only 933 in the whole dataset. Figure 5.1 shows the distribution of people clustered by their continent of birth and their gender. As it can be observed, the population is highly skewed toward Europe and North America that together reach 66.8% of continents of birth, while men represents 78.7% of the total.

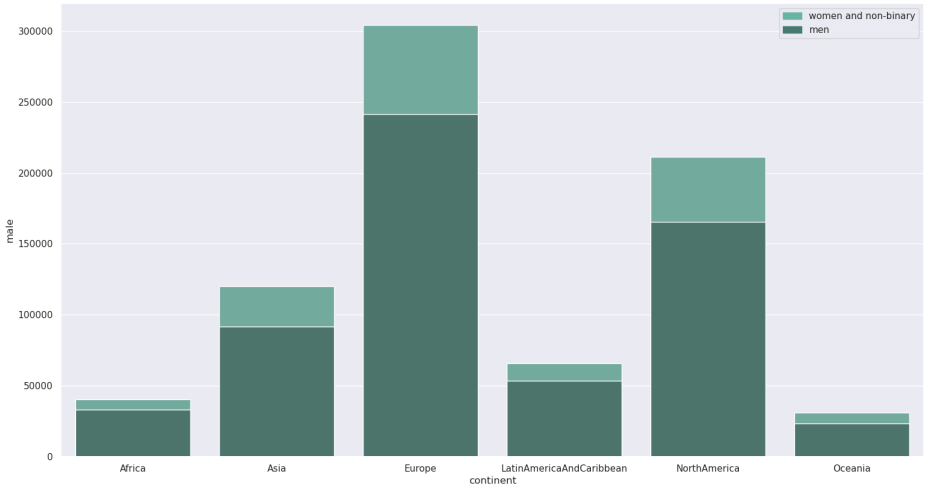

Figure 5.1: The distribution of people on Wikidata broken down by gender and ethnicity (continent of birth)

Dates of birth were clustered in 5 generations

- Silent Generation, which includes all people born between 1926 and 1945;
- Baby Boomers, between 1946 and 1964;
- Generation X, between 1965 and 1980;
- Millennials, between 1981 and 1996;
- Generation Z, between 1997 and 2012.

Figure 5.2 shows that Baby Boomers are the most represented generation in the knowledge base (28.5%) followed by X (24.7%) and Millennials (23.9%). The imbalance between men and women varies across generations: they are less represented among people from the Silent Generation 21.2% while they reach 38.9% among Millennials.

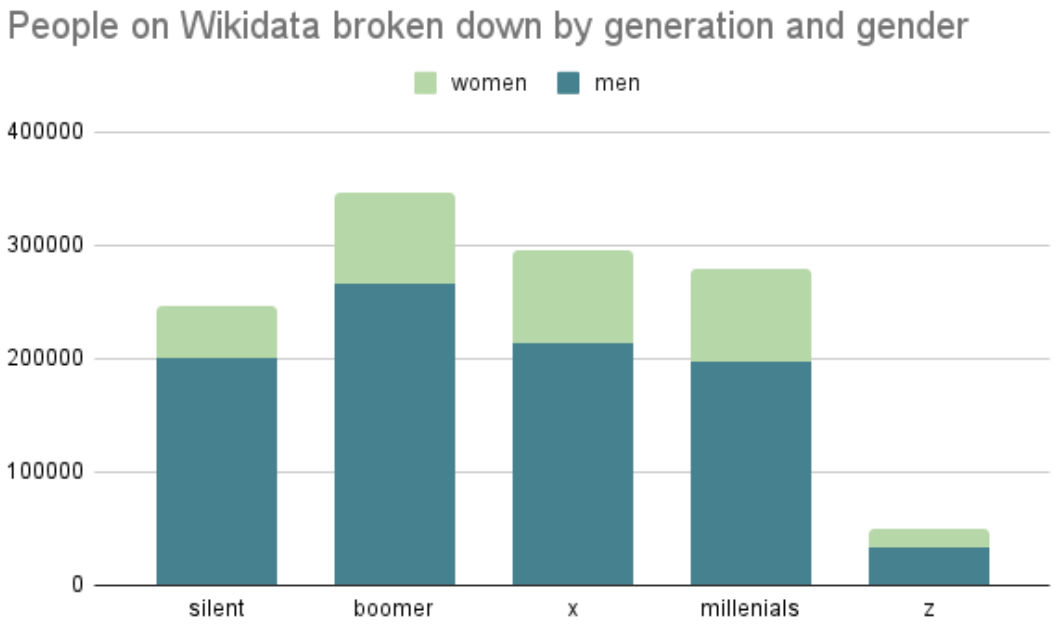

Figure 5.2: The distribution of people on Wikidata broken down by generations.

Occupation properties required a more significant manual clustering, since they are 4,735. I chose only the 31 occupations that are linked to at least 10% of people in the

dataset. I aggregated them in more general categories (e.g., association football player and basketball player were included in the same set) and kept the 10 larger categories of occupation, which are the following: athlete, actor, writer, politician, musician, director, coach, researcher, lawyer, model. In Figure 5.3 it is possible to see that 'athlete' is the most represented occupation type: 38.1% of people are linked to this property. The four other most recurring occupations are 'actor' (16.4%), 'writer' (12.6%), 'politician' (11%), and 'musician' (8.3%).

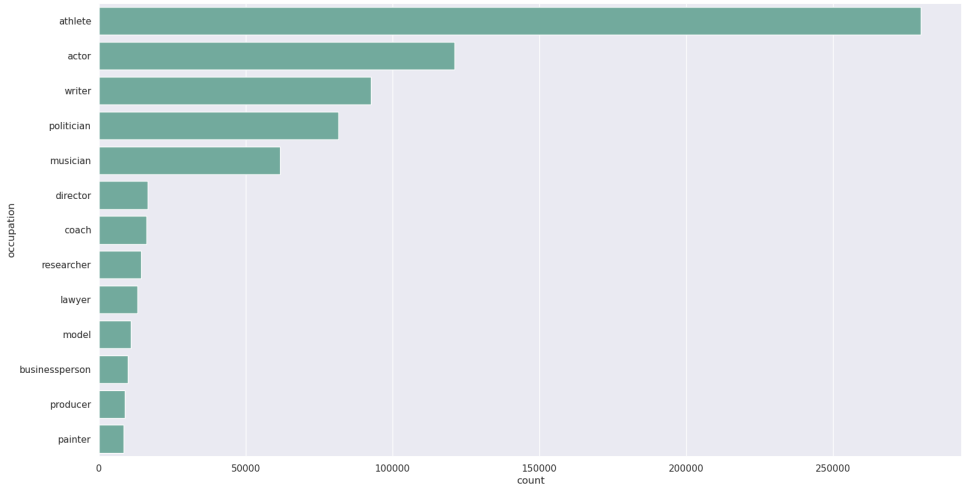

Figure 5.3: The distribution of people on Wikidata broken down by their occupation.

The correlation between generation and occupation properties in Table 5.1 shows that some professions are specific to certain generations. If five of them are mostly associated with Baby Boomers, 'athlete' and 'model' show a strong correlation with Millennials, while people belonging to Generation X are slightly more linked to 'actor' and 'coach' than Baby Boomers. Silent Generation is mostly associated with the 'researcher' occupation.

In Table 5.2 the correlation between occupation, gender and origin is reported. Except from the 'model' occupation, which is mostly composed of women, all occupation distributions are imbalanced towards men and towards Western people. The highest underrepresentation of women is among athletes, where they only are 7% and coaches

	Silent	Boomer	X	Millennial	Z
Athlete	8.5%	13.9%	21.7%	**45.5%**	10.1%
Actor	17.9%	25.0%	**29.5%**	24.2%	3.1%
Writer	28.4%	**38.1%**	24.6%	7.7%	0.9%
Politician	25.9%	**46.2%**	22.3%	5.2%	0.2%
Musician	17.3%	**29.3%**	28.8%	21.4%	0.3%
Director	21.4%	**34.7%**	31.4%	8.8%	0.1%
Coach	15.4%	35.6%	**37.0%**	11.3%	0.4%
Researcher	**46.4%**	40.7%	11.5%	7.1%	0.5%
Lawyer	30.8%	**44.2%**	21.0%	3.4%	0.03%
Model	4.6%	8.8%	27.2%	**55.0%**	0.4%

Table 5.1: The distribution of people on Wikidata according to their generation and occupation.

	Men-Women	Western-Transnational
Athlete	$0.93 - 0.07$	$0.70 - 0.30$
Actor	$0.56 - 0.44$	$0.72 - 0.28$
Writer	$0.69 - 0.31$	$0.75 - 0.25$
Politician	$0.78 - 0.22$	$0.64 - 0.36$
Musician	$0.68 - 0.32$	$0.73 - 0.27$
Director	$0.83 - 0.17$	$0.68 - 0.32$
Coach	$0.98 - 0.02$	$0.66 - 0.34$
Researcher	$0.81 - 0.19$	$0.85 - 0.15$
Lawyer	$0.77 - 0.23$	$0.75 - 0.25$
Model	$0.20 - 0.80$	$0.54 - 0.46$

Table 5.2: The distribution of people on Wikidata according to their occupation, gender and origin. Columns report the distribution of men and women, and of Western and Transnational people across all occupation types.

(2%). Transnational people suffers the highest underrepresentation among researchers (15%) and writers (0.25%).

5.1.2 Experimental Setting

My analysis was aimed at identifying potential biases that affect certain categories of people in English Wikipedia pages. I compared different social groups in pairs. For the analysis I adopted a Monte Carlo simulation [James, 1980]. Such statistical method is adopted for the analysis of large and normally-distributed datasets and provides the analysis of n samples of data on which a certain statistic is computed. The average of results obtained for all samples represents an approximation for the whole dataset. In my setting I used Monte Carlo to simulate normally distributed set of biographies across one or more socio-demographic axes (e.g., Transnational director women *versus* Western director men), in order to perform an analysis of *representational* bias that is less influenced by Wikipedia's distribution imbalance. More specifically, I considered each pair of social groups as two ballot urns and proceeded as follows: *i.* I randomly selected one of the two urns and picked a random biography; *ii.* I repeated the operation 100 times. *iii.* I created 20 samples of the same size by repeating the same procedure 20 times in order to reduce the variability induced by random sampling. As a result, I obtained a set of $2,000$ biographies for each analyzed pair of social groups.

In total, I performed 117 comparisons between social groups by combining the four socio-demographic features presented in Section 5.1.1: gender, ethnicity, age, and occupation. Table 5.3 shows all the combinations. I first compared biographies by gender and by ethnicity. The two features where then combined obtaining 15 pairs. An instance of this type is 'Western women *versus* Transnational men'. The integration of information about people occupation enabled the creation of 60 pairs of the type {gender + ethnicity + profession}. It is worth mentioning that non-binary people were excluded from all combinations that involve occupation and generation, since they are too few if their total number is broken down along these axis. Finally, I combined pairs according to generations in order to explore the role of age in *representational* bias. For instance,

Comparison	
{gender}	{gender}
{ethnicity}	{ethnicity}
{gender + ethnicity}	{gender + ethnicity}
{gender + ethnicity + profession}	{gender + ethnicity + profession}
{gender + ethnicity + generation}	{gender + ethnicity + generation}

Table 5.3: The list of socio-demographic axes chosen to perform my analysis of *representational* bias.

play, win, be, work, release, serve, bear, appear, receive, join,
say, write, make, begin, sign, move, start, take, return, study

Table 5.4: The 20 most frequent events extracted from biographies

Transnational Women belonging to Generation X were compared with Transnational Women belonging to Generation X. This way I obtained 40 pairs.

The overall number of biographies selected through the Monte Carlo simulation is 147, 754. For each of them I adopted the biographical event detection pipeline described in Section 4.4. I first obtained all the mentions of the target entity of a biography, then I performed event detection only on the sentences containing a mention of the entity. After this step I obtained 6, 019, 506 biographical events. Their distribution over the men/women and Western/Transnational axes denotes an imbalance of events towards men and Westerners. On average 42.2 events were detected in men biographies against 40.6 in women; 45.3 in Western against 36.1 in Transnational. This is due to the fact that even if the distributions were balanced at the level of biographies, their average length differs, penalizing minorities and minoritized groups. Table 5.4 shows the 20 most occurring events identified with my pipeline.

As a final part of my analysis, I compared the frequency of the extracted biographical events along pairs of social groups in order to identify events that are more associated to one group or another. Following the approach of Gallagher et al. [2021], I considered each pair of social groups a set of 20 distinct Monte Carlo simulations for which I computed the delta between the relative frequencies of extracted events between the two groups (e.g., men *versus* women)

Occupation	Women	Men
All	work, win, marry, appear, complete, bear, receive, perform, release, represent	play, sign, score, start, join, make, return, injury, lead, come
Acting	appear, actress, star, marry, receive, release, model, debut, performance, cast	actor, direct, director, write, defeat, be, produce, die, serve, lose
Athlete	win, compete, represent, name, finish, take, run, set, help, member	sign, score, release, make, return, join, start, spend, move, deal
Politician	work, serve, elect, bear, marry, study, degree, represent, move, win	be, chairman, die, arrest, become, businessman, sign, leader, death, play
Researcher	work, earn, focus, study, attend, teach, bear, research, graduate, elect	be, say, claim, argue, support, see, author, contribution, propose, die
Lawyer	work, appoint, award, bear, degree, graduate, receive, marry, hold, study	die, lose, run, defeat, return, chairman, win, attempt, call, arrest

Table 5.5: The comparison of the most relevant events for men and women, broken down by occupation.

$\rho_{\text{men}} - \rho_{\text{women}}$

In this case, events with a lower negative score are mostly associated with the left member of the subtraction (men), while events with the highest positive score with the right member (women). Since certain strong associations between social groups and events are caused by their rare presence in biographies, I only kept events that appeared in at least 10 round of the Monte Carlo simulations. Examples of the final output are 'play', which scored -0.018 resulting to be associated with men, and 'work' that showed a stronger association with women with a score of 0.007.

5.2 Analysis of *Representational* Biases

5.2.1 Gender

Men and women. The analysis of the events most associated to men and women (Table 5.5) reveals the presence of many generic events (e.g., start, join, receive, complete). Alongside these, some biographical knowledge seems to be most associated to each category. In Table 5.5 it is possible to observe that men are associated to a number of events that evoke sports (play, score, injury), while women are associated to a more faceted distribution: there are event related to the private domain like 'marry', events associate to culture and entertainment like 'perform', 'release' and 'appear', and to general forms of competition (win, compete).

Since such an analysis may be influenced by the unbalanced distribution of professions described in Section 5.1.1, I broke down the comparison by occupation. The analysis reveals that the association between women and marriage is not preserved along all types of occupation. When the comparison is made between athletes and researchers, such association disappears. A second interesting aspect emerges from the analysis of the most relevant events about actors and actress: the former are associated to roles of responsibility like 'direct' and 'produce', while the latter are more confined to the acting profession (e.g., star, performance) and to modelling. Finally, it is interesting to notice that the event 'die' is almost always associated to men, thus showing a potential influence of age in the analysis.

These findings suggest that the occupational status of a person has a role in reducing or amplifying gender bias. This happens especially in the case of athletes when references to marriage are not present among the events that are most associated to women. Additionally, the analysis suggests that men's careers embrace more than one occupation with prestigious roles, if compared to women.

Non-binary people. People who do not identify themselves as men or women are highly underrepresented in Wikidata. As shown in Section 5.1.1, less than $1,000$ people are associated with a different 'sex or gender' property than man or woman. This does not allow performing a joint analysis of gender and profession, which has been performed in the previous paragraph. A coarse grained comparison of non-binary and binary people shows interesting results, though. In Table 5.6 it is possible to observe the presence of two terms related to gender affirming surgery in association with non-binary people: 'transition' and 'surgery'. Other events highlight political activism (activist) and participation to the cultural and artistic sphere (release, perform, feature). Almost all events related to men and women are related to career. Similarly to what has been observed in Table 5.5 about men and women, a high number of events seem to be related to sport (play, score, defeat).

The analysis shows that Gender biases are amplified when Wikipedia pages about

Non-Binary	Binary
release, say, transition, perform, feature, state, appear, activist, surgery, use	play, win, score, sign, serve, finish, join, lose, defeat, return

Table 5.6: The most relevant events for non-binary people.

non-binary people are observed. Events related to their gender identity seems to show a stronger relevance than the association between women and marriage observed previously. However, the public dimension of events related to gender affirming surgery may characterize this association as a form of reclamation of these identities, which also echoes from the presence of the event 'activist' among the most relevant about this minority.

5.2.2 Ethnicity

The analysis based on ethnicity shows an even more salient intersection between this socio-demographic trait and occupation. If observed at a general level (Table 5.7), Western people are associated with sport events (play, coach, compete), Transnational ones with political jargon (elect, represent). The latter are more linked to the event 'marry'.

An additional pattern seems to emerge from the comparison at the level of single profession. While events about Transnational actors are all related to this domain (e.g., act, performance), 'write' appears as an event associated to Westerners. Another interesting aspect is observed in the comparison between politicians. Events pertaining to these domains (e.g., run, endorse) are significantly associated with Westerners, while Transnationals are more linked to events that do not belong to their profession (e.g., arrest, study, education). A similar observation emerges researchers. Terms like 'study', 'argue', and 'focus' are mostly associated with Western people, Transnationals with generic ones. It is worth mentioning the presence of 'politician' as an event linked to three different categories of Transnational professionals: politician, researcher, and lawyer. This may suggest that their professional careers are more intertwined with political careers than their Western counterparts.

Results of the analysis of *representational* bias based on ethnicity show mixed results. The association of Transnational people with terms related to political career

Occupation	Western	Transnational
All	sign, finish, play, record, attend, compete, retire, coach, say, spend	bear, become, elect, hold, represent, debut, marry, take, member, award
Acting	appear, say, role, appearance, attend, star, perform, guest, write, announce	act, debut, make, participate, career, start, actor, be, actress, performance
Athlete	attend, record, sign, have, name, earn, average, start, draft, lead	play, score, represent, win, debut, join, compete, part, move, take
Politician	say, run, announce, defeat, work, vote, support, serve, graduate, endorse	be, politician, bear, study, degree, hold, appoint, arrest, obtain, education
Researcher	professor, work, receive, study, argue, focus, research, play, graduate, perform	serve, obtain, return, bear, appoint, degree, run, resign, politician, be
Lawyer	say, vote, receive, announce, defeat, run, nominate, support, write, nomination	appoint, lawyer, study, elect, bear, politician, hold, obtain, member, become

Table 5.7: The comparison of the most relevant events for Western and Transnational people, broken down by occupation.

seems counter-intuitive. However, their high underrepresentation in Wikidata (Section 5.1.1) may suggest that the lower attention on lives of non-Western people leads to few biographies about highly prominent individuals. The analysis is prone to a significant interference of professions, though. As for the case of women, Transnational actors appear to be relegated to less prestigious role in the cultural environment while Westerners are associated to leading roles.

5.2.3 Intersection Between Gender and Ethnicity

In this section I analyze the presence of bias by looking at the intersection between gender and ethnicity. I first analyze how this interaction impacts on the comparison between women and men, then I perform the same analysis on non-binary people.

Transnational and Western Women. Table 5.8 shows four types of comparison: Transnational women *versus* Western and Transnational men; Western women *versus* Transnational and Western men. The analysis aims at understanding if and to which extent ethnicity introduces additional *representational* biases than ones emerged in Section 5.2.1. Some patterns are still clearly observable across all comparisons: the association between women and marriage and a stronger link between men and career events, with a specific focus on sports. The effect of ethnicity seems to have an impact when Transnational women are compared to Western men. The event 'model' is among the 10 most

Transnational Women work, bear, study, win, compete, participate, marry, hold, model, actress	**Western Men** play, sign, score, return, join, finish, spend, release, make, appearance
Transnational Women work, release, appear, star, participate, study, marry, act, bear, compete	**Transnational Men** play, score, sign, join, make, start, appearance, appoint, move, come
Western Women work, appear, win, star, compete, receive, say, marry, perform, place	**Western Men** play, sign, score, join, make, coach, spend, leave, return, start
Western Women win, compete, release, work, write, perform, marry, graduate, star	**Transnational Men** play, score, sign, join, return, appoint, serve, start, make, debut

Table 5.8: Results of an intersectional comparison on the axes of gender and ethnicity.

distinctive only in this comparison, while references to parenthood ('bear') are associated to Transnational women both against Transnational and Western men. The same does not happen for Western women, who never show a strong link to 'bear'. When Western women are compared to Transnational men, an unseen associations with a career-event like 'write' emerges together with a reference to academic success ('graduate'). This suggests that ethnicity has a twofold role in the interaction with gender. Being both a woman and a Transnational person seems to amplify gender bias, introducing an additional link between women and parenthood. The opposite happens for Western women who are associated to events related to the cultural field (e.g., write) when compared against Transnational men.

Transnational and Western Non-Binary People. The intersectional comparison between non-binary and binary people shows similar trends to the one between men and women. Transnational people biographies are characterized by a stronger association with events related to gender affirming surgery, as it emerges from the term 'surgery' that is linked to them but not to Westerners. Another significant distinction is about the term 'activist' that is among the 10 events that are most associated with Transnational people. Finally, the event 'write', which might denote an active participation in cultural production, is mostly linked to Western people. All these characteristics show three lines of intersection: women and non-binary people are mostly associated with private

Transnational Non-Binary release, say, activist, perform, surgery, be, change, transition, feature, come	**Western Binary** play, win, score, sign, finish, serve, join, lose, defeat, graduate
Transnational Non-Binary say, release, surgery, activist, change, perform, live, adopt, want, state	**Transnational Binary** play, win, score, sign, join, serve, finish, debut, lose, defeat
Western Non-Binary release, say, transition, state, feature, use, write, be, appear, describe	**Western Binary** play, win, finish, sign, score, serve, lose, join, return, defeat
Western Non-Binary release, say, write, appear, feature, perform, use, state, transition, describe	**Transnational Binary** play, win, score, sign, serve, join, return, appoint, represent, defeat

Table 5.9: Results of an intersectional comparison between non-binary and binary people.

events; Transnational lives are characterized by events related to political activism, and Westerners are more likely to be linked with events related to cultural production.

5.2.4 Intersection Between Ethnicity, Gender, and Age

For the analysis of biases about age I kept Baby Boomers as a fixed generation and compared biographies of people belonging to this generation with Silent, X, Millennial, and Z. I performed this comparison for Transnational women, Western women, and Transnational men, in order to identify the presence of specific patterns. Results are shown in Table 5.10. As can be observed, there is a stable presence of events related to prestigious public positions in any type of comparison when the population is composed of women ('serve', 'appoint', 'elect'). This is expected when Baby Boomers are compared with younger generations, but less intuitive when the comparison is performed with older ones that have had more chance to pursue a political career. If the focus is moved to Transnational men, events like 'elect' and 'appoint' are not among the most significant for Baby Boomers, suggesting that it is more likely for younger generations of Transnational men to experience a political career. A second relevant pattern regards the event 'write'. It appears linked to Transnational women belonging to Baby Boomers when are compared to all younger generation, to Western ones when are compared to Millennials and Z, to Transnational men only when compared to Z. This might suggest a lower number of young women who are associated to the cultural producation. A final consideration

	Transnational Women	**Western Women**	**Transnational Men**
vs Silent	serve, elect, appoint, win, be, degree, release, member, run, announce	serve, be, win, member, release, appoint, announce, sponsor, represent, defeat	win, coach, serve, say, be, announce, play, manager, lose, replace
vs X	serve, appoint, work, elect, die, write, publish, professor, bear, use	serve, work, elect, member, teach, appoint, hold, run, professor, die	serve, elect, bear, receive, appoint, hold, study, die, member, work
vs Millennial	serve, elect, appoint, work, publish, write, professor, receive, bear, degree	serve, work, write, elect, receive, publish, appoint, marry, award, study	serve, elect, bear, receive, appoint, hold, study, die, member, work
vs z	serve, work, appoint, publish, study, marry, elect, write, receive, professor	work, serve, elect, write, marry, publish, receive, appoint, study, graduate	serve, work, appoint, elect, study, publish, say, member, bear, write

Table 5.10: The analysis of the most relevant events for Baby Boomers against four other generations: Silent, X, Millennials, Z. The comparison is performed on Transnational women, Western women, and Transnational men

regards the event 'professor', which is always relevant for Baby Boomers when they are compared with younger generations. As for 'write', this could suggests that the youngest are associated with this prestigious profession with less frequency.

From these results it seems to emerge that age represents an additional bias to ones deriving from gender and ethnicity and their combination is crucial to better understand how *representational* biases actually affect people from minorities and minoritized groups. The association of Baby Boomers with events like 'write' and 'professor' shows that it is more difficult for people belonging to younger generations to become an established writer if they are also women and even more if they are women and Transnational.

5.3 Conclusion of the Chapter

In this chapter I provided a thorough analysis of *representational* bias in English Wikipedia pages, supported by a classifier specifically designed for biographical event detection (Section 4.4) and by the structured knowledge provided by Wikidata and encoded according to the UR network (Section 3.3. The analysis embraced an intersectional perspective, since it jointly considered four socio-demographic axes: ethnicity, gender, age, and occupation. Such an approach demonstrated that existing research on gender bias on Wikipedia failed to identify fine grained forms of bias deriving from the multiple sources

of oppression that affect certain categories of people. For instance, Transnational women are more prone to discrimination than their Western counterparts and being young is an additional issue. Results also showed that *representational* bias differ according to occupations. The more stereotypical representation of women are traceable in actress biographies, while in Wikipedia pages about sportspeople the gender bias is more mitigated. It is worth mentioning that the analysis has been performed over a highly unbalanced datasets that has been normalized to reduce the impact of *allocative* bias over *representational* ones. The majority of people in the English Wikipedia are Western men, and a high percentage of them are athletes. In Chapter 6 I will specifically focus on the underrepresentation of non-Western people in Wikidata, complementing this analysis with a study of the knowledge gap emerging from the Wikimedia ecosystem.

Chapter 6

Allocative Bias in Datasets

The underrepresentation of minorities is a long-lasting problem that either affects digital and traditional media. Silencing practices [Spivak, 2015] relegated ethnic minorities and Transnational people to a marginal role in textbooks [Wolf, 1992], movies [Erigha, 2015], and digital archives [Adams et al., 2019]. Such underrepresentation is the source of many forms of *allocative* harms that range from educational systems that systematically exclude minorities' perspectives from their curricula to NLP models that are not trained on documents that account for cultural diversity.

In this chapter I present an analysis of *allocative* bias in Wikidata aimed at measuring the magnitude of this phenomenon in this knowledge base. This work is built upon and continues the analysis performed in Chapter 5 and is specialized on the case study of writers. This not only enables the exploration of *allocative* bias about people, which was introduced in Section 5.1.1, but also extends the analysis to their works, opening to a more complete framework for the assessment of bias about cultural products as they emerge in the digital environment.

In addition to the detection of *allocative* bias I implemented two strategies for their mitigation. The first is based on the semantic alignment of Wikidata with two other public archives: Open Library[1] and Goodreads[2]. I framed these archives as three distinct

[1] https://openlibrary.org/
[2] https://www.goodreads.com/

communities of readers and analyzed their degree of inclusiveness given a fixed distribution of writers gathered from Wikidata. The second strategy is based on the automatic extraction of biographical triples from English Wikipedia pages, that led to the increase of available structured information about Transnational writers.

The resulting output of this analysis is the World Literature Knowledge Graph (WL-KG), a knowledge base composed of 194,346 writers and their works, gathered from Wikidata, Open Library, and Goodreads and augmented through biographical event extraction (Chapter 4. The WL-KG relies on the Underrepresented Network Ontology (UR-ON) that has been introduced in Chapter 3. The resource is available through a visualization platform specifically conceived for non-expert users that allows the discovery of potentially underrepresented writers and their work, thus representing not only a benchmark resource for *allocative* bias detection in NLP but also a tool for a more general mitigation of cultural bias in Digital Humanities research practices.

The chapter is organized as follows. In Section 6.1 I present my data gathering pipeline from Wikidata and some descriptive statistics about the underrepresentation of Transnational writers; Section 6.2 describes the mitigation strategy based on the semantic alignment of Wikidata with Goodreads and Open Library; Section 6.3 the strategy based on biographical triples extraction. Finally, in Section 6.3.1 I present the visualization platform designed for non-expert users.

6.1 Measuring Allocative Bias Against Transnational Writers

The first step of my analysis of *allocative* bias against Transnational writers has been the collection of information about them from Wikidata.

We gathered 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. Then I collected information about year and country of birth, in order to structure my KG accordingly to the UR-ON semantic model presented in Section 3.3:

- I filtered out all writers who were born before the outbreak of the Spanish-American war (1808) that I consider crucial event for the beginning of the decolonization process;
- for each writer I retrieved the 'place of birth' (P17) and then obtained the country of birth exploiting the 'country' (P19) property;
- if present, I collected information about minorities by first gathering the 'ethnic group' property (P172) and then retaining only ones representing minorities in Western countries (eg: African Americans);
- I obtained the 'gender' (P21) of each writers and grouped all people who do not identify themselves as men or women under the 'non-binary' label;
- I collected all works related to writers through the 'author' (P50) property.

After this process, I obtained 194,346 writers who were clustered in two groups: Western and Transnational. With Transnational I identified all people born in a former colony or belonging to a ethnic minority in a Western country. As a result, I formed a group of 176,697 (91%) Western writers, and 17,368 (9%) Transnational. The distribution is even more skewed when works are considered. Of 145,375 works gathered from Wikidata, 136,995 (94.2%) belong to Western writers, whereas 8,380 (5.8%) are associated with Transnational Writers.

For better exploring such underrepresentation, I analyzed the distribution of Transnational and Western 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 6.1 Western male writers are predominant within the Silent Generation (66.2%) and Baby Boomers (60.9%). Surprisingly, the number of Western women progressively increases until overcoming 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 and Transnational women writers suffer an additional lack of representation on Wikidata: there are only 508 (8.1%) male and 379 (6%) female

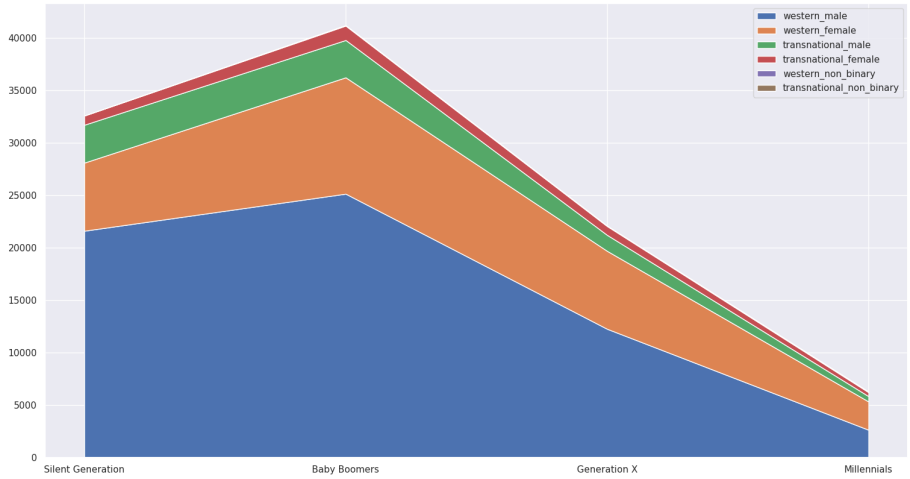

Figure 6.1: Four generations of Western and Transnational writers grouped by gender

Transnational writers on Wikidata among Millennials. Non-binary writers are the most underrepresented, regardless of their condition. In the KG there are only 146 (0.009%) non-binary Western authors and 23 (0.01%) non-binary Transnational authors.

6.2 Mitigating Bias Through Semantic Alignment

In this section I present my mitigation strategy of *allocative* bias based on the semantic alignment of Wikidata with other sources on knowledge.

Our first step was a quantitative analysis of writers' external identifiers. Wikidata pages are structured in two main sections: one includes all the knowledge that is internal to Wikidata; the other lists all external identifiers associated to an entity on the knowledge base. For instance, Chinua Achebe's Wikidata page is referenced to his profile on the Poetry Foundation database[3]. After a first recognition, I focused on three external identifiers: writers' Virtual International Authority File Name (VIAF) IDs, Open Library IDs and Goodreads IDs. A fourth platform, Library Things, was not included in the data

[3] https://www.poetryfoundation.org/poets/chinua-achebe

collection process given the low number of links from Wikidata and the impossibility of automatically obtaining authors' IDs from that website. In Table 6.1 it is possible to observe that the 84% of writers has a VIAF ID, the 18.5% an Open Library ID, and the 4.5% a Goodreads ID.

In order to increase the percentage of writers mapped to VIAF and Open Library identifiers, I adopted three heuristics:

- I retrieved all the names of the writers through the OpenLibrary APIs and kept only the entities fulfilling two conditions: (a) an exact string match between the author name in my KG and the one in OpenLibrary; (b) the same year of birth in my KG and in OpenLibrary. As a result, I obtained 19,737 additional ids (+54.6%).
- I scraped all writers' names from Goodreads sitemap[4] filtering out all homonyms. I then mapped all the names in my KG onto Goodreads author list, keeping only the string matches. I thus obtained 26,019 new ids (+280%).
- I searched all ISBNs related to each authors through VIAF and performed a search through ISBN on Open Library and Goodreads, that allowed retrieving 22,661 Open Library IDs (+40%) and 44,142 Goodreads IDs (+120%).

6.2.1 Quality Assessment of the Mapping

After the mapping, I performed a quality assessment of a sample of links between Wikidata and Goodreads, and between Wikidata and Open Library to remove incorrect links before gathering works. My evaluation strategy is composed of three steps. We computed the Gestalt pattern similarity [Ratcliff et al., 1988] between the names of the same writer in different platforms. Gestalt pattern similarity is a metric to compute the similarity between two strings that relies on the ration between the number of shared characters and the characters that are not shared. I chose such a metric over other similarity measures like cosine similarity because it emphasizes the explanation behind the mapping. Any association between names can be intuitively perceived and assessed by the user. An

[4] https://www.goodreads.com/siteindex.author.xml

example of mapping is about the author Esther Salaman[5] which is linked to a Goodreads page[6], where she is referred as 'Esther Polianowsky Salaman'. The two strings have a Gestalt pattern score [Ratcliff et al., 1988] of 0.7.

Once I setup the metric to assess the mapping, I checked its consistency on my data by manually checking random samples of name pairs with different degrees of similarity: $x < 0.1$, $0.1 \geq x < 0.2$, $0.2 \geq x < 0.3$, $0.3 \geq x < 0.4$, $0.4 \geq x < 0.5$, $0.5 \geq x < 0.6$, $0.6 \geq x < 0.7$. For each degree of similarity, I sampled 100 name pairs, which is a statistically significant sample since it always represents more than 1% of the population within each range of similarity. As it can be observed in Figure 6.2, the percentage of correct links is directly proportional to the similarity between the names by which the writer is referred to in different platforms. In particular, the precision dramatically increases with a similarity between 0.5 and 0.6 (77% of correct links) reaching a 89% of accuracy with a precision between 0.6 and 0.7. Considering the whole sample of 700 name pairs, the evaluation shows a recall of 0.31 for pairs with a similarity below 0.7

Such manual assessment allowed for setting a threshold similarity ≥ 0.7 that minimizes the number of wrong mappings and emphasizes the quality of data in the KG. As a result I removed 6,812 Open Library mappings and 3,754 Goodreads mappings. Projecting the recall of mappings with a similarity score below 0.7 is 0.31 over removed instances led to the removal of 2,111 potentially correct Open Library mappings (3.1%) and 1,126 potentially correct Goodreads mappings (1.42%). After this process I obtained 64,894 (33.4%) writers with an Open Library ID and 75,404 (38.7%) with a Goodreads ID (Table 6.1). The percentage of writers linked to at least one of the two platforms is 54%.

Identifier	**Before Mapping**	**After Mapping**
VIAF	163,353 (84.0%)	
Open Library	36,097 (18.5%)	64,894 (33.4%)
Goodreads	8,997 (4.6%)	75,404 (38.7%)

Table 6.1: Number of authors with an external identifier

[5] http://www.wikidata.org/entity/Q4405658
[6] https://www.goodreads.com/author/show/618352

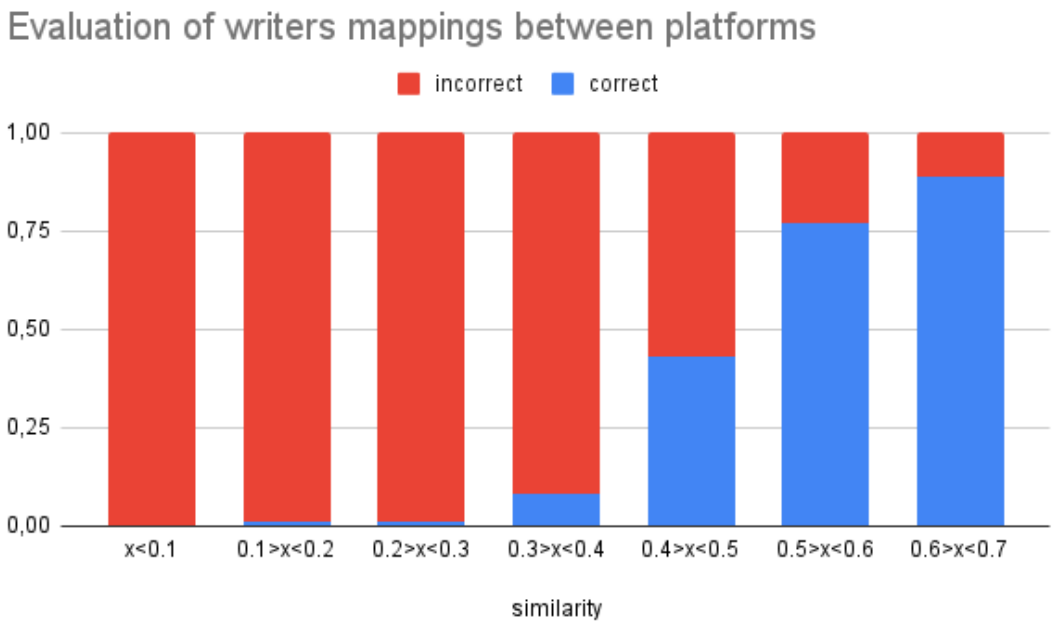

Figure 6.2: Results of the evaluation of writers mappings between Wikidata, Goodreads, and Open Library.

6.2.2 Data Collection and Statistics

After the augmentation of external identifiers of authors, I collected all their works in these platforms. OpenLibrary APIs allow retrieving all works, and for each work it is possible to obtain all editions. Results include a set of useful publishing information, readers count, ratings, and number of ratings. Goodreads does not provide APIs, but allows for web scraping. Hence, I first collected the list of all works from writers pages, their ratings and number of ratings, then I obtained publishing information through Google Books APIs.

In order to emphasize the role of readers communities, I only kept works that had received at least one reception or that were marked as read by at least one user. Table 6.2 shows the number of works collected from each platform and the number of writers associated with at least one work from them. As it can be observed, Goodreads includes a higher number of works and writers with at least one work. Furthermore, both

Open Library and Goodreads show a higher percentage of 'Transnational' writers than Wikidata: 12.6% and 11% against 8.6%.

Source	**N. of writers with ≥ 1 works (% transn.)**	**N. of works**
Wikidata	22,515 (8.6%)	117,798
Open Library	24,370 (12.4%)	226,108
Goodreads	60,201 (11.0%)	627,214
Total	71,443 (10.6%)	971,120

Table 6.2: Number of works for each platform

The analysis of readers communities may also be observed through the lens of the number of interactions between readers and works. While Wikidata does not include users evaluation of literary works, it is possible to obtain this information from Goodreads and Open Library. Both expose the number of ratings and the average rating, while the latter also exposes the number of readers. Table 6.3 shows the number of interactions between readers and literary works in the two platforms. As it can be observed, absolute numbers are incomparable: there are 112.708 ratings in Open Library against 1.7 billions in Goodreads. The percentage of ratings about Transnational works is higher on Open Library (6%) than in Goodreads (4.9%), while both platforms show a slightly higher average rating of Transnational writers.

Source	**Average rating**	**N. of works**	**N. of readers**
Open Library	3.91 (3.99)	112,708 (6.0%)	$1.2M$ (8.5%)
Goodreads	3.86 (3.77)	$1.7B$ (4.9%)	–

Table 6.3: Number of readers interactions in Goodreads and Open Library. Interactions about Transnational writers are reported in parenthesis.

Summarizing, aligning literary facts from different platforms in a unique semantic resource allows for a richer representation of World Literature, with a more balanced knowledge about Transnational writers (+2% of them are associated with at least one work). Furthermore, such data collections shows the impact of communities of readers on the diffusion of writers and their works.

6.3 Mitigating Bias Through Biographical Triple Extraction

In this section I present a strategy for the mitigation of Transnational writers underrepresentation based on biographical event extraction. The pipeline combines the methodology developed in Section 4.4 and the approach based on LSPs presented in Section 4.2 to automatically extract relations belonging to four career-relevant properties on Wikidata: 'educated at' (P69), 'employer' (P108), 'award received' (P166), and 'nominated for' (P1411).

Pipeline Implementation As mentioned above, I focused the implementation of my 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 my choice I 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 my dataset are 17,649, but only 7,979 of them (45.5%) have an English Wikipedia page[7]. In order to have a benchmark to evaluate the effects of my Entity Detection model (described in Section 4.4.4), I performed the sentence segmentation of raw biographies[8] 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, I detected all events as described in Section 4.4.5, thus identifying 11,876 event types that occur 216,666 times.

For the extraction of the triples from text I created three alternative LSPs by combining the output of the biographical detection step with this rule-based approach. I reviewed all the detected events and kept only the ones that (i) are quantitatively rel-

[7]The dataset was collected in October 2021

[8]We used the Natural Language Toolkit for this task: https://www.nltk.org/

evant, namely they occur at least 50 times and represent 72.9% of the total number of extracted events; (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 my focus on these types of events, I performed a NER step, 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 my pipeline I performed an EL step, which allowed us to extract 37,249 triples from 27,682 unique sentences.

The example below, encoded according to the Wikibase ontology[9] shows a set of extracted triples[10]:

```
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 .

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

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

[10] For readability, here I 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.

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)[11] and available through their SPARQL endpoint[12]

6.3.0.1 Evaluation of Results

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

Pipeline Evaluation In order to evaluate the quality of my pipeline, I 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, I first checked if it was correct and, if not, I specified the source of error among the four components of the pipeline: co-reference, event, NER, and EL. Table 6.4 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 6.4: Manual Evaluation of the extracted triples

When errors are broken down according to their types (Figure 6.3), an insight about my pipeline emerges. In fact, 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:

[11] https://doi.org/10.5281/zenodo.8399935
[12] https://kgccc.di.unito.it/sparql/biographical_triples

generic mentions identified as organizations, like 'Senate' in Example 8, which is wrongly labelled as the Senate of the United States of America, and misclassified 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.

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[13]. 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[14].

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] https://www.wikidata.org/wiki/Q4556206
[14] https://www.wikidata.org/wiki/Q269471

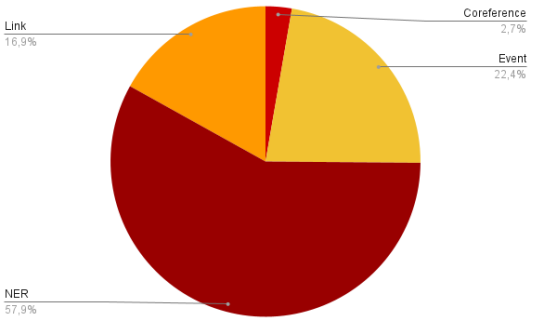

Figure 6.3: Breakdown of errors emerged during the manual assessment of the pipeline. I defined 4 types of errors: (i) coreference, if the event is not associated to the target of a biography; (ii) event, if the LSP does not apply to the Wikidata property; (iii) NER, if a wrong entity was recognized; (iv) Link, if the link was incorrect.

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 [Kingsbury and Palmer, 2002], NomBank [Meyers et al., 2004], and Unified Verb Index [Kipper et al., 2006] 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 (Auer et al. [2007]), (CaLiGraphHeist and Paulheim [2019]), and Google KG[15] in the EL step may result in a richer and more precise set of extracted triples.

Reduction of Underrepresentation The second part of the evaluation focuses on the impact of triple extraction in reducing Transnational writers' underrepresentation.

[15] https://www.google.kg/

To do so, I 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 my triple extraction task from their biographies. A first intuitive overview of this augmentation may be observed in Figure 6.4, 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 I implemented within my pipelines: given the absence of entities of the type 'prize' in NER corpora, I 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 6.3.0.1). The general impact of my approach seems to be significant, though. As it can be observed in Table 6.5, 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 6.5: The impact of my 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 this pipeline to other sources of knowledge and its implementation to other languages may dramatically increase the number of structured information that I have about Transnational writers and other categories

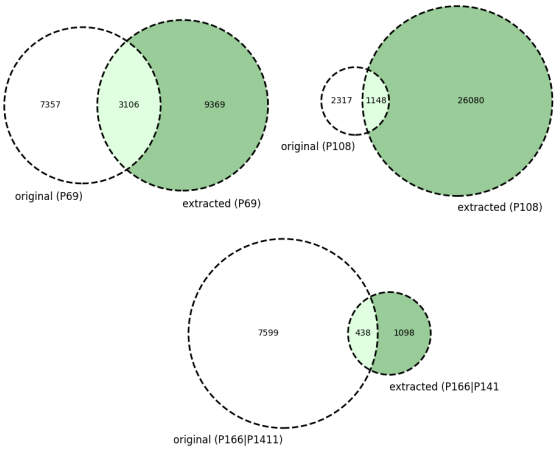

Figure 6.4: A visual representation of the effect of triples extraction on the total number of triples about Transnational writers. Extracted triples are in green in diagrams.

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.

6.3.1 Visualizing World Literatures

This final section presents a visualization tool that relies on the WL-KG (Section 6.2) and is designed to make it accessible to non-expert users. The visualization tool is the output of a collaboration with a research team from the University of Bari, which implemented and customized its previous work to the WL-KG [Bernasconi et al., 2022]. Section 6.3.1.1 describes the visualization platform, while in Section 6.3.1.2 I provide an evaluation of the interface performed through a series of interviews to professionals in the literary field.

6.3.1.1 The Visualization Platform

WL-KG is built to support advanced queries and is seamlessly integrated with SKATEBOARD, the Semantic Knowledge Advanced Tool for Extraction Browsing Organization Annotation Retrieval and Discovery, providing users with an intelligent and intuitive way to explore the vast world of literature. With WL-KG and SKATEBOARD interface, my

goal is to enable users to uncover deep insights and connections within literary works and enhance their understanding of the literary world. The SKATEBOARD platform presented in this research builds upon the work of Bernasconi et al. [2023] and represents an extension and updated version of their work to fit my specific context of use. The interface features two main views: 'Author' and 'Work'. The navigation flow that starts with an initial search for a topic of interest. Once a relevant topic is found, the user can drag the resource onto the central board and explore its relationships with other objects and predicates, creating a visual representation of the connections. This feature is illustrated in Figure 6.5.

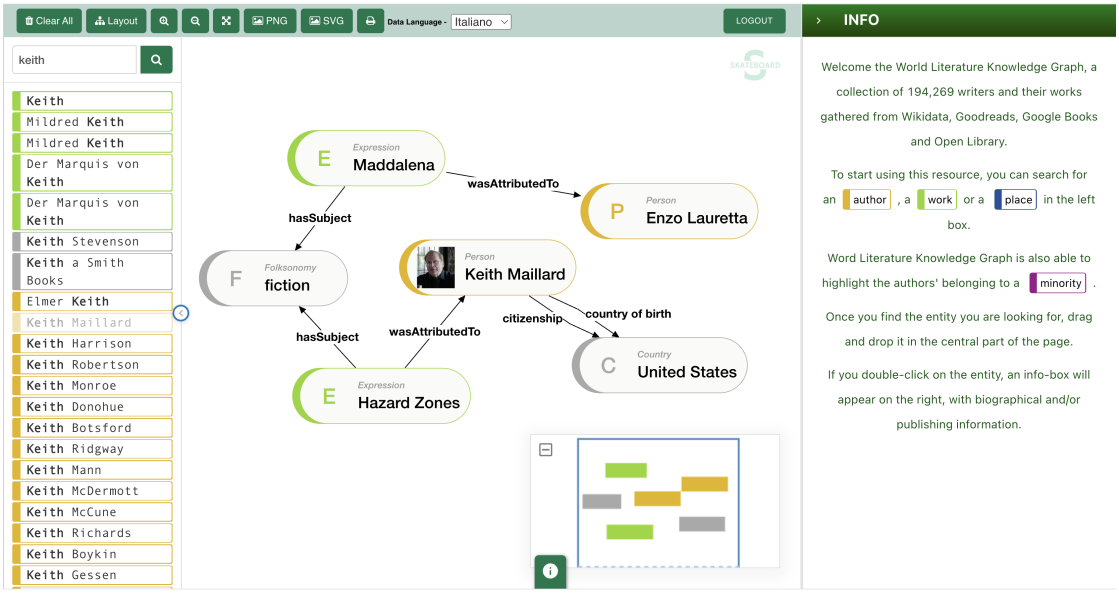

Figure 6.5: A snapshot of the visualization platform. On the left, the search box; in the middle, the whiteboard where entities can be dragged; on the right, info pane about the selected entity.

By clicking on resources of type 'Person' (as visible in Figure 6.6), the user can access information about an author, including both direct relationships such as published works and indirect relationships such as all the topics covered in their works, or a map of all the locations where their works were published. Clicking on resources of type 'Expression' (as visible in Figure 6.7) displays information specific to a particular work, such as editions, languages, and readers ratings.

Literary searches may also start from different types of entity in the KG. It is possible to retrieve all writers by their country of birth or by their citizenship, as well as perform searches based on specific minorities (eg.: African Americans). The platform also allows a navigation based on subjects: users can browse all works linked to a specific URB:FOLKSONOMY. The graph-based navigation encourages serendipitous discovery, allowing users to stumble upon unexpected connections and relationships.

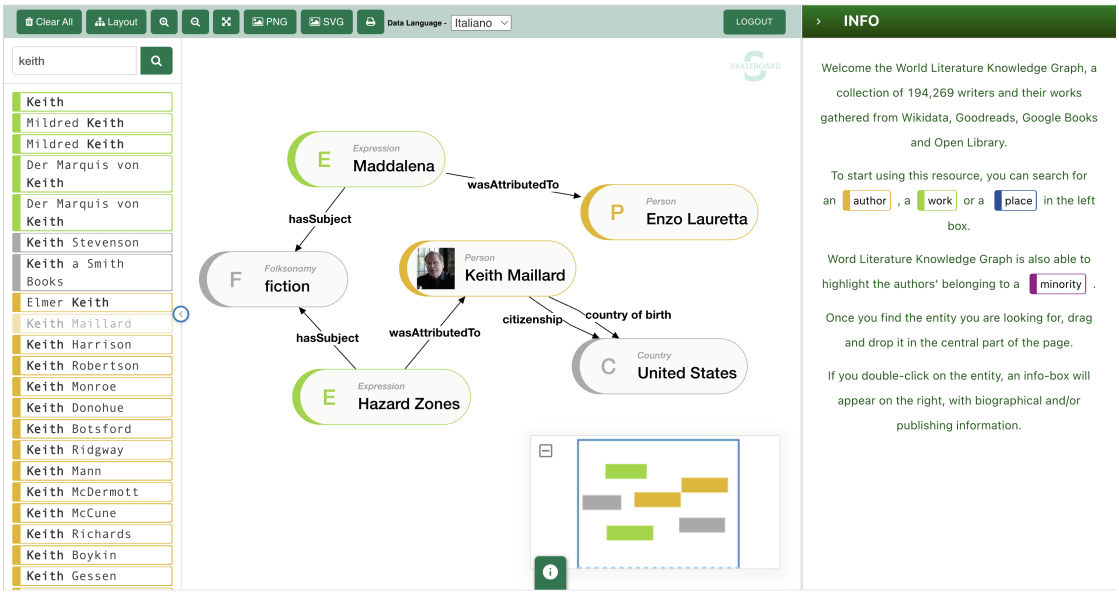

Figure 6.6: Person view: on the left, the central area of the interface, where selected entities can be dragged for visualising their provenance and associated media and their relations with other entities according to the node-link paradigm (here, Chinua Achebe); on the right, the Info pane displaying the information about the entity (e.g., biographical dates, citizenship).

In summary, the visualization platform presented in this research offers an updated and customizable interface for exploring and visualizing relationships between topics, authors, and works, with potential applications in various research fields.

6.3.1.2 Resource Evaluation

The current form of the WL-KG and its visualization platform are the result of a two-year interactive process of design and development carried out in constant interaction with

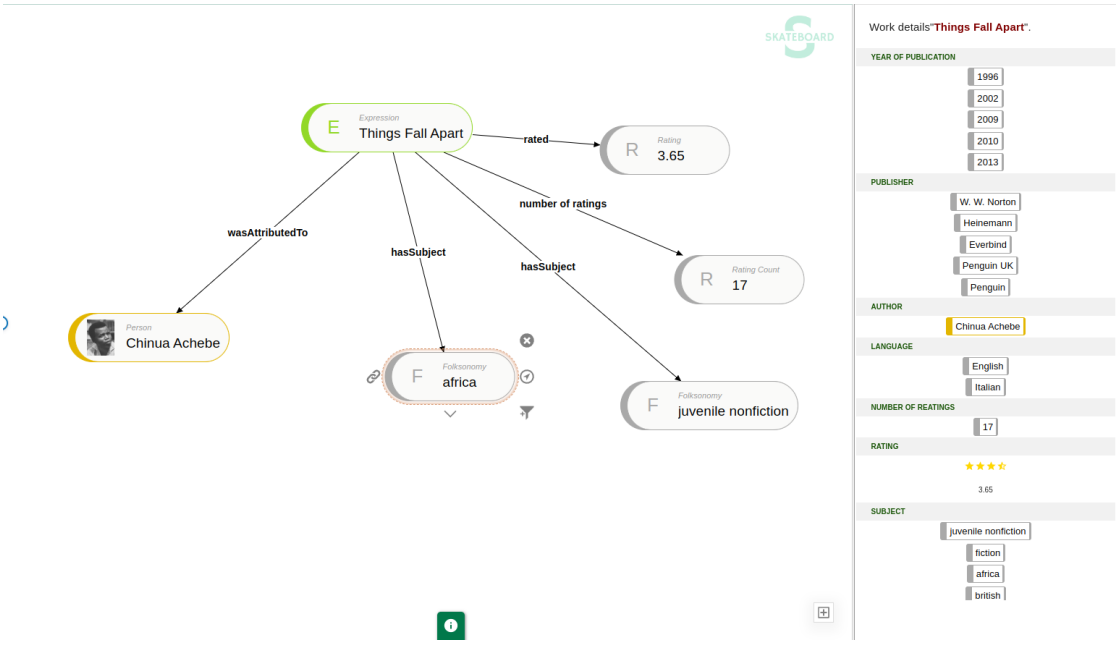

Figure 6.7: Expression view: on the left, the central area of the interface where a work (top left, “Things Fall Apart") is connected with its author (Chinua Achebe, see Fig.6.6). On the right, the Info pane displaying the information about the work in tabular form (an Expression in FRBR terms), such as publisher, language, rating, etc.

domain experts. The contribution of domain experts to the process has been twofold: on one side, they helped in defining the geopolitical and temporal boundaries of the resource, suggesting post-colonial studies as the conceptual reference framework for the study of underrepresentation; on the other side, they suggested that a graph-based visualization would be better suited to encourage exploration – and support their professional tasks – than the archival-based visual metaphors employed in the first prototype, for its capability to encourage the discovery of new authors through the connections displayed in the graphical interface.

For the evaluation of the WL-KG I organized a series of structured interviews[16] with a group of potential targets of my resource, in line with the paradigm of user-centered design [Wood, 1997]: 4 teachers, 6 researchers in the humanities, and 3 professionals in the publishing industry. Each interview was articulated in two parts: the first part,

[16]The structure of the interview is reported in Appendix D.

targeted on the search of Transnational writers and works, was focused on the use of the platform; the second focused on the potential uses of the resource in the users' field of work and research.

User Experience After asking the users to search for at least one Transnational author and work of their choice, the user experience was investigated along three dimensions: the usability of the platform, the completeness of the results, and the accuracy of the results.

Concerning the usability of the platform, most users experienced difficulties in navigating the WL-KG. First of all, they didn't realise that every element in the search area can be dragged into central whiteboard – according to the incremental paradigm that controls the interaction between the search area and the central whiteboard (as described in Section 6.3.1.2). Secondly, they failed to explore the information linked to the selected entity by expanding the relations between that entity and the other entities connected with it in the graph, which can be navigated in the whiteboard according the node-link paradigm. Conversely, a minority of respondents who had already experience of Knowledge Graphs found the platform easy to use and appreciated the possibility of selecting the entities of interest by dragging them into the whiteboard, a function that they saw as a way to overcome the limitations of the standard navigation tools for graph-based representations. Based on these observations, I hypothesize that the difficulties in the use of the visualization platform reported in the interviews can be mainly attributed to the users' lack of experience with graph-based resources. For these users, the drag and drop selection of entities and the link-based navigation were not intuitive and can be improved by providing more guidance in the exploration (e.g., through tooltips, demo-mode navigation, etc.). This is in line with the comment made by some respondents who suggested to initialize the platform with an already loaded example. Concerning the entry of the search parameters, some users expressed their difficulty in finding a suitable author or work, motivating it with their limited knowledge of the domain. To bypass this difficulty, a user suggested creating a list of writers' names, indexed by country of birth, in a

separated section of the site. I think that this suggestion is valuable, although it partly overlaps with the possibility of exploring the graph by starting from different types of entities (e.g., subjects, countries, topics), which is already available in the current version of the platform.

Concerning the completeness of the resource, a criticism derived from a misconception about its objectives shared by most respondents, who compared it with standard online archives, such as Wikipedia: the latter, being targeted at end users, include richer information about the entities in textual form, but are not suited for the development of applications that rely on the graph-based representations. This issue can be addressed by revising the description of the resource with a clearer definition of its intended usages. A more challenging request, then, emerged from the scholars in post-colonialism, who complained about some missing associations between works and subjects. It is the case of Andrea Levy's work 'The Long Song': although this book is about 'slavery', it is not linked to this subject in the KG, an issue derived from the lack of attribution of this subject within the digital sources from which data were gathered.

As for completeness, almost all respondents found the resource accurate, with a few errors that I could track from sources. For instance, 'Candide oder der Optimismus', namely the German translation of Voltaire's 'Candide', was attributed to Stephan Hermlin, its translator, due to an error propagated from Goodreads. To address this issue, a functionality for signaling missing and wrong information will be added in a future version of the platform.

Use Cases The discussion of use cases was structured in two main parts: the comparison of the resource with the existing known archives and the collection of feedback about use cases and missing functionalities. Participants tended to rate the resource as useful for the discovery of new writers, but not useful for exploring new works. Such feedback reflects my data collection strategy, that was limited to the existing entities of the type writer on Wikidata and to the works that had received at least one reaction on the platforms where they are archived, aiming at relevance rather than completeness of

works.

Interviews also showed that almost all respondents use general purpose archives like Google, Wikipedia, and Goodreads for the literary searches, showing a gap in the usage of knowledge bases designed for specific domains of application. The discovery of new literary facts has been pointed out as the major use case for all respondents. Interestingly, from the structured interviews with teachers, it emerged that the students themselves may be potential users of the platforms, since they could take advantage of subject-based search for supporting essay writing. Finally, the need of exposing emerged in the knowledge base all the places where authors lived during their lives, in order to discover deeper connections between them.

6.4 Conclusion of the Chapter

In this chapter I presented an analysis of *allocative* bias against Transnational writers. The analysis showed that the underrepresentation of this category of writers in Wikidata amplifies inequalities that are are present in the real world. Other public archives like Goodreads and Open Library are more balanced and aligning Wikidata with them reduces this type of underrepresentation. In addition, I proved that biographical triple extraction may contribute to the increase of structured information about Transnational people in this knowledge graph. Finally, I implemented a visualization tool that turned out to be useful to make my resource available to to a wider public composed of non-expert users.

Chapter 7

Bias in Annotated Corpora

The high number of resources and approaches developed for HS detection by the NLP community has given birth to a new group of studies aimed to assess and challenge common practices behind the creation of corpora for this task. Despite sharing a common goal, these studies go in different directions. A first group investigates how a high variety of annotations hinders the generalizability of models for HS detection. This problem clearly emerges from several surveys. Poletto et al. [2021] provide a clear distinction between HS and partially-overlapping phenomena like cyberbullying or offensiveness. Fortuna and Nunes [2018] propose a definition of HS that synthesize definitions emerging from international laws and social networks community guidelines. Yin and Zubiaga [2021]'s review shows that differences in annotation schemes and in the definition of the phenomenon hinder the training of generalizable classifiers. To overcome such an issue, Vidgen and Derczynski [2020] propose a set of best practices for dataset creation that involves (i) defining clear guidelines for the annotation task; (ii) documenting the whole process; (iii) accounting for the diversity of annotators; (iv) ensuring a representative and unbiased collection of data to annotate. A second line of research focuses on the presence of cultural biases in annotated datasets. The seminal work of Sap et al. [2019] demonstrated that in existing HS corpora messages written in AAE are more likely to be annotated as containing HS. Blodgett et al. [2016] release a classifier for the detection of AAE that has been used to further explore the correlation between authors' ethnicity

and presence of HS in their messages (see Kim et al. [2020]), corroborating previous findings. A third line of research analyzes the high subjectivity that characterizes the perception of discriminatory contents, proposing perspectivist annotation frameworks [Cabitza et al., 2023]. Leonardelli et al. [2021] release a dataset where each message is labelled by five annotators so that different levels of agreement could be explored. Poletto et al. [2019] compare the annotation of the same data performed with three different annotation schemes, showing that this choice has an impact on the annotators' propensity to label messages as hateful. Finally, some works investigate the effects of certain phenomena on HS recognition and perpetration. It is the case of Frenda et al. [2022] who provided evidence of the correlation between sarcasm and HS, Bassignana et al. [2018] who exploited SA to better classify HS, and Lai et al. [2021] who integrated resources designed for morality recognition to identify HS spreaders on Twitter.

In this chapter I present an attempt to integrate a number of existing corpora for HS and other abusive phenomena within a single semantic model in order to study their generalizability. More specifically, my aim is to test three hypotheses about existing resources on HS:

1. abusive phenomena are characterized by different degrees of generalizability [Pamungkas et al., 2020];

2. HS corpora differs in their compatibility with existing HS definitions [Fortuna et al., 2020]

3. cognitive and emotive phenomena have a role in the recognition of HS [Frenda et al., 2022, Lai et al., 2021]

We aligned corpora according to O-Dang [Stranisci et al., 2022b] (Chapter 3.4). Coherently with the research aim of the Linguistic Linked Open Data (LLOD) community, [Chiarcos, 2012, Hellmann et al., 2013], the ontology provides a semantic representation of annotations and corpora they are part of. At the core of O-Dang! there is the conceptualization of annotation labels as set of descriptions that can be examined in

terms of their appropriateness in a wide range of situations, such as automatic labeling a message as HS under a specific definition of this phenomenon. More specifically, I report on the alignment of 9 corpora for HS detection, 2 corpora annotated with moral values, and 2 corpora annotated for emotional appraisal. The alignment supports three analyses that I performed in order to test the hypothesis outlined above.

1. **Cross-domain learning of harmful language.** I grouped annotated messages in three manifestations of abusive language —misogyny, offensiveness, and HS— and studied how well a model trained on each manifestation generalizes on others.

2. **Definition compatibility and applicability on HS corpora.** I performed a zero-shot experiment aimed at measuring the effect of adding seven different legal definitions to a prompt-template for the classification of HS in all 11 corpora included in O-Dang!.

3. **Integration of moral values and emotional appraisal in HS detection.** I trained a model for the detection of morality [Graham and Haidt, 2012] and emotional appraisal [Roseman, 2013] and tested its contribution to predict HS.

Figure 7.1 provides a general overview the analysis that I performed. The box in the center of the image refers to transfer learning between different types of abusive language; the upper one refers to the compatibility between HS corpora and legal definitions; the two boxes in the left part of the image the analyses based on knowledge augmentation with moral values and emotional appraisal.

The chapter is organized as follows. In Section 7.1 I present all corpora that have been included in the first version of O-Dang! and how they are encoded in the KG. Section 7.2 I describe experiments of transfer learning between HS, misogyny, and offensiveness. Section 7.3 reports about how HS corpora are aligned with existing legal definition of HS. In Section 7.4 I present results of experiments aimed at transferring knowledge from corpora annotated for morality and appraisal to HS datasets.

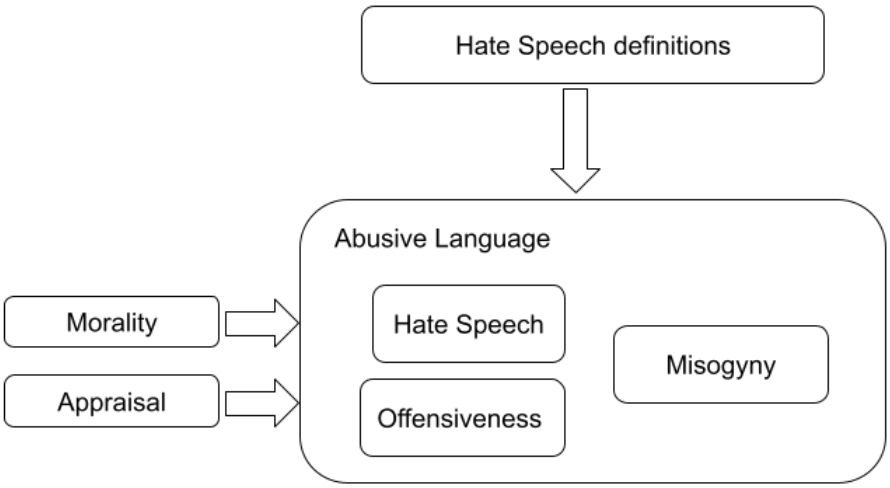

Figure 7.1: Three types of analyses of HS corpora. The first is between three types of abusive languages: HS, misogyny, and offensiveness. The second is between legal definitions of HS and corpora annotated for abusive language. The third between morality, appraisal, and HS.

7.1 Datasets Selection and Alignment

7.1.1 Hate Speech

Since the number of HS corpora is high, I only selected a sample of 9 resources. My choice has been guided by two criteria: comparing corpora released in different times and analyzing differences between between highly-related phenomena, such as HS and misogyny.

1. **Waseem [Waseem and Hovy, 2016].** A dataset of 16,000 tweets annotated along the axis of racism and sexism.

2. **Davidson [Davidson et al., 2017]**. A corpus of 24,802 tweets filtered through a keyword search based on HateBase[1] and annotated for offensiveness and HS.

3. **AMI [Fersini et al., 2018]** is a corpus of 2,000 messages for misogyny identification in a binary fashion and for misogyny behavior. The latter is a multi-label

[1] https://hatebase.org/

annotation where a behavior may be 'Stereotype & Objectification', 'Dominance', 'Derailing', 'Sexual Harassment & Threats of Violence', or 'Discredit'.

4. **OLID [Zampieri et al., 2019]** contains 14,000 tweets annotated for their offensiveness. Annotations provide also the type of offense, distinguishing between ones targeting a specific group and ones that not, and the target, where a distinction is made between individuals and groups.

5. **Ethos [Mollas et al., 2020]** comprises two corpora. The first is composed of 998 tweets annotated as HS or not. The second comprises only the 433 messages annotated as HS, which has been labeled according to eight categories: 'violence', 'directed_vs_generalised', 'gender', 'race', 'national_origin', 'disability', 'sexual_orientation', and 'religion'.

6. **HateXplain [Mathew et al., 2021]** includes 11,903 Twitter and Gab posts annotated as hateful, offensive or normal by three annotators who were also asked to identify the rationales of their annotations by selecting the span of text triggering offensiveness or hate.

7. **Implicit Hate Corpus [ElSherief et al., 2021]** is a corpus of 22,056 tweets annotated as containing HS or an implicit form of HS. 6,346 messages annotated as implicit were additionally labeled with a fine-grained taxonomy of implicit discrimination.

8. **EDOS [Kirk et al., 2023]** is a corpus of 20,000 Gab and Reddit posts annotated for sexism. The corpus includes a binary classification (sexist *versus* not sexist), a multi-label taxonomy of sexist messages (threats, derogation, animosity, prejudiced discussion), and a set of mutually exclusive labels for each of the previous categories.

9. **Incels [Gajo et al., 2023]** includes 5,203 messages gathered from Incels.is[2] and independently annotated for misogyny and racism.

[2] https://incels.is/

7.1.2 Appraisal

Recent work on HS detection found that users' emotional response has an impact on the proliferation of HS, thus paving the way for cross-domain experiments between corpora annotated for emotion and for HS [Plaza-Del-Arco et al., 2021]. In this context, appraisal theories of emotions [Roseman, 1991] are suited for this type of task.

Theories of appraisal Smith and Ellsworth [1985], Roseman and Smith [2001] emphasize the evaluation stage of an event or a situation, that leads to an emotional response and to a corresponding behavior aimed at coping with the situation and alleviating the response itself. Emotions, in this view, stem from cognitive evaluations of events and are followed by specific autonomic responses, behavioral configurations and action tendencies Smith and Ellsworth [1985]. Such evaluations work by assessing the current situation against a set of appraisal criteria, such as the congruence between an event and an agent's goal or the novelty of a specific situation Sander et al. [2018]. Different evaluations of the same situation elicit different emotions. For instance, a given event will cause joy if it helps the appraising agent fulfill their goal and will elicit surprise if unexpected. While appraisal theories have been largely employed in computational models of behavior Marsella and Gratch [2009], Dias et al. [2014], there is still a lack of linguistic resources drawing from this family of theories. Appraisal-based linguistic resources may be of great importance because they define, beyond emotions, evaluation processes for situation types and, even more importantly, a range of corresponding behaviors, of which linguistic behaviors are a subset. Together, all these features could provide more information and explanatory capacity to several tasks like stance detection, abusive language identification, and sentiment analysis.

NLP resources modeled on appraisal theories are still limited, though. Only in more recent times some resources have been developed for the automatic detection of appraisal variables. In the next paragraphs I present three existing resources that are annotated for this task and that have been integrated in O-Dang!.

ISEAR corpora [Hofmann et al., 2020]. Troiano et al. [2020] delivered two corpora

of German and English event descriptions for emotion recognition that generated through an experimental setup: deISEAR, and enISEAR. Corpora were crowdsourced on Figure-Eight in two rounds of annotations. A first group of annotators generated emotion-focused events of the form 'I feel ... when ..'. A second independent group annotated the emotion expressed by events, therefore validating the generated texts. Hofmann et al. [2020] completed this work by adding an annotation for the English corpus of 7 appraisal dimensions derived from the taxonomy proposed by Smith and Ellsworth [1985]: Attention, Certainty, Effort, Pleasantness, Responsibility, Control, and Circumstances.

x-enVENT [Troiano et al., 2022]. Developed by the same research team that created enISEAR, x-enVENT is a corpus of 929 texts gathered from existing corpora that have been annotated for appraisal with a scheme of 22. Messages are also annotated for the event that triggers the appraisal and for entities experiencing the event.

APPReddit [Stranisci et al., 2022a]. APPReddit is a corpus of social media posts annotated for appraisal. 1,091 events gathered from Reddit have been annotated with a scheme based on Roseman's model of appraisal [Roseman, 1991, 2013]. Annotators labeled each event along five dimensions: unexpectedness, certainty, consistency, responsibility, and control.

7.1.3 Moral Foundations

The impact of morality in the evaluation of social issues is a growing field of research. It has been demonstrated that annotators moral stance has an impact on annotated corpora [Forbes et al., 2020] and that morality may explain how and why HS is spread online [Hoover et al., 2019]. All these works are based on the assumption that morality is not universal [Shweder et al., 1997]: each individual has their own moral configuration that affects their view of the world.

In this work I rely on the Moral Foundations Theory (MFT) Graham et al. [2013], according to which morality is composed of five moral dyads.

1. Care/Harm. Prescriptive concerns related to caring for others and prohibitive concerns related to not harming others.

2. Fairness/Cheating. Prescriptive concerns related to fairness and equality and prohibitive concerns related to not cheating or exploiting others.

3. Ingroup Loyalty/Betrayal. Prescriptive concerns related to prioritizing one's ingroup and prohibitive concerns related to not betraying or abandoning one's ingroup.

4. Authority/Subversion. Prescriptive concerns related to submitting to authority and tradition and prohibitive concerns related to not subverting authority or tradition.

5. Purity/Degradation. Prescriptive concerns related to maintaining the purity of sacred entities, such as the body or a relic, and prohibitive concerns focused on the contamination of such entities.

The morality of each individual is built upon a specific configuration of these concerns that are considered within the theoretical framework as partly innate, partly developed through experience and social relationships. This allows MFT's dyads to describe morality as organized in advance of experience, highly dependent on environmental influences collected during development within a particular culture, and to see moral judgments as intuitions that happen before the subject starts to reason.

Several works within this framework have been devoted to investigate relations between moral foundations and political ideology, referring in particular to the moral differences between liberals and conservatives Graham et al. [2009], media studies Winterich et al. [2012]. In recent years, MFT foundations in the online environment have been studied by some scholars together with its correlation with other topics, such as hate speech Hoover et al. [2019], or political discourse Johnson and Goldwasser [2018]. Concurrently, many resources to investigate this phenomenon have been released: corpora of annotated tweets Hoover et al. [2020], dictionaries Graham and Haidt [2012], Hopp et al. [2020], and knowledge graphs Hulpus et al. [2020].

For my experimental setting, I used two existing resources: The Moral Foundations Twitter Corpus and the Moral Foundations Reddit Corpus.

Moral Foundations Twitter Corpus (MFTC) [Hoover et al., 2020]. This collection includes 35,108 tweets annotated for moral values according to the Moral Foundations Theory. The corpus is composed of 6 thematic subsets: 'Black Lives Matters', 'All Lives Matters', 'MeeToo Movement', a'Hurricane Sandy', '2016 US elections', and 'Baltimora Protests'. Additionally, a part of Davidson's corpus [Davidson et al., 2017] has been annotated for this task.

Moral Foundations Reddit Corpus (MRFC). [Trager et al., 2022]. The corpus includes 16,123 Reddit comments gathered from 12 different subreddits and clustered in three main categories: US politics, France politics, and everyday moral life. The corpus is annotated along a recently updated version of Moral Foundations [Atari et al., 2023] that replaces *fairness/cheating* with two moral dyads: equality/inequality that focuses on equal treatments and proportionality/disproportionately, which focuses on merit.

7.1.4 HS Normative Definitions

In addition to corpora for HS and harmful language, O-Dang is populated with 7 HS normative definitions from different domains, in order to support the compatibility of existing corpora with them. Table 7.1 shows such definitions, which can be grouped in three different domains. The first three derive from the legal domain: *UN* is provided by the United Nations, while *CoE_1* and *CoE_2* by the Council of Europe. A second group of definitions is collected from community guidelines of social media: Twitter (*TW*), Facebook (*FB*), and Google (*Ggl*). Finally, I stored in HAbLan the definition provided by Fortuna and Nunes [2018] in their survey (*FN18*).

Definition	Description	Source
UN	Shall declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin, and also the provision of any assistance to racist activities, including the financing thereof;	United Nations
CoE_1	For the purposes of the application of these principles, the term "hate speech" shall be understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin.	Council of Europe Recommendation 97/20
CoE_2	For the purposes of this recommendation, hate speech is understood as all types of expression that incite, promote, spread or justify violence, hatred or discrimination against a person or group of persons, or that denigrates them, by reason of their real or attributed personal characteristics or status such as "race", color, language, religion, nationality, national or ethnic origin, age, disability, sex, gender identity and sexual orientation.	Council of Europe Committee of Ministers
TW	You may not directly attack other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.	Twitter Community Guidelines
FB	I define hate speech as a direct attack against people — rather than concepts or institutions— on the basis of what I call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. I define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation.	Facebook Transparency Center
Ggl	Hate speech is not allowed on YouTube. I don't allow content that promotes violence or hatred against individuals or groups based on any of the following attributes, which indicate a protected group status under YouTube's policy: Age, Caste, Disability, Ethnicity, Gender Identity and Expression, Nationality, Race, Immigration Status, Religion, Sex/Gender, Sexual Orientation, Victims of a major violent event and their kin, Veteran Status.	Google Support Policy
FN18	Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used.	Fortuna and Nunes [2018]

Table 7.1: Hate Speech Definitions

7.1.5 Aligning Dataset

All datasets havr been encoded according to O-Dang (Section 3.4). Each message is represented as a DUL:INFORMATIONOBJECT, which is described by at least two PIM:ANNOTATIONSCHEME: one that signals the presence or absence of a given phenomenon (eg: misogyny), the other that specifies it. Below it is possible to observe an example of annotated text from AMI dataset [Fersini et al., 2018] in turtle syntax.

```
o-dang:ami_6 a dul:InformationObject;
    dul:isPartOf o-dang:AMI;
    dc:description 'ok babies i'll go to sleep ok bitch
    shut the fuck up';
    dul:isDescribedBy o-dang:ami_binary_desc_6;
    dul:isDescribedBy o-dang:ami_category_desc_6.
o-dang:ami_binary_desc_6 dul:defines o-dang:misogyny;
    oa:target o-dang:ami_6;
    oa:body 1 .
```

```
o-dang:ami_binary_desc_6 dul:defines o-dang:dominance;
    oa:target o-dang:ami_6;
    oa:body 1 .
```

In this representation, the same message (ami_06) is associated to a description that links it to the general concept of O-DANG:MISOGYNY and to a description that links it to one of its sub-category: O-DANG:DOMINANCE. Such an encoding enables the extraction of KG snapshots that may be used for more high-level or more fine-grained classification experiments. The association between texts and different descriptions can also be used to manage unaggregated annotations. Message '2 v2 challs kroc is scared of comp so dont be a dumb sweater retard' from HateXplain [Mathew et al., 2021] is associated with three descriptions from different annotators. Below it is possible to observe that the first description (hatexp_binary_desc_34) defines the message as expressing offensive contents, while the other two do not.

```
o-dang:hatexp_12 dul:isDescribedBy o-dang:hatexp_binary_desc_34,
o-dang:hatexp_binary_desc_35, o-dang:hatexp_binary_desc_35 .
o-dang:hatexp_binary_desc_34 dul:defines o-dang:offensiveness;
    oa_target o-dang:hatexp_12;
    oa_body 1 .
o-dang:hatexp_binary_desc_35 dul:defines o-dang:offensiveness;
    oa_target o-dang:hatexp_12;
    oa_body 0 .
o-dang:hatexp_binary_desc_36 dul:defines o-dang:offensiveness;
    oa_target o-dang:hatexp_12;
    oa_body 0 .
```

In summary, representing all datasets in a unique KG provides a flexible organization of knowledge that can be used to perform the analyses that are presented in next sections.

7.2 Transferring Abusiveness

Recent studies show how hard it is for models trained on one dataset to generalize well on others [Toraman et al., 2022, Yin and Zubiaga, 2021]. Although misogyny, offensiveness and HS have a common ground —they all involve harmful attitudes, expressions, or behaviors—, data differ. Despite their different definitions, misogyny and sexism both involve showing harmful attitudes against women, be it in a hostile or in a benevolent manner [Barreto and Doyle, 2023]. HS at textual level expresses hate on the basis of specific characteristics, such as race, religion, ethnicity, disability, sexual orientation, or gender identity [Fortuna and Nunes, 2018], while offensive language refers to any text containing abusive slurs or derogatory terms, regardless of the target [Zampieri et al., 2019]. Moreover, even data trained in-domain but across different datasets, with different annotation schemes, might not generalize well on other datasets. Our experiment aims to investigate this cross-domain and in-domain cross-dataset phenomenon.

Approach. In this experiment I exploited the KG to study how abusive knowledge can be transfer between its different forms. I obtained through SPARQL query three distinct datasets: (i) 'HS', composed of 58,430 entries from 6 corpora; (ii) 'misogyny', which includes 38,498 messages from 4 corpora; (iii) 'offensiveness' that contains 50,552 posts from 3 corpora. An example of SPARQL query is reported below.

```
SELECT DISTINCT ?id ?text ?label ?phenomenon WHERE {
?id a dul:InformationText;
dc:description ?text;
dul:isDescribedBy ?desc.
?desc ?defines ?phenomenon;
oa:body ?label .

FILTER(?phenomenon=o-dang:hate-speech) .
}
```

Training set	T1-HS	T2-Mis	T3-Off
HS	0.79	0.57	0.58
Mis	0.47	0.81	0.68
Off	0.66	0.61	0.82

Table 7.2: Macro F-1 score of predictions based on a classifier trained on one of the three phenomena and tested over the others (T1, T2 and T3).

After holding out one external test set for each dataset group (misogyny, HS and offensiveness), composed of a stratified sample (10%) of all the individual datasets, I trained a classifier based on a DistilBert checkpoint [Sanh et al., 2019] for 4 epochs and then predicted over all the external datasets. I repeated the training 5 times for each dataset with a different sample and averaged the obtained F-1 scores. Results highlight two aspects: (i) the consistency of annotations of the same phenomenon across different corpora and (ii) the compatibility of annotations for HS, misogyny, and offensiveness.

Results. Table 7.2 shows the results. The annotations about the same phenomenon seem to be compatible in corpora that have been released in different periods and with different annotation guidelines. F-1 scores are around 0.80 in all the three settings, as observable from the main diagonal, and in line with scores obtained in development sets. Transfer between phenomena always results in a drop, but some differences can be highlighted. Models trained on offensiveness generalize better than the others, reaching an F-1 score of 0.66 and 0.61 on hate speech and misogyny respectively. This finding is in line with Pamungkas et al. [2020]: The OffensEval training set always achieves the best performance when tested on three other hate-related datasets. If trained on HS, classifiers provide similar performance when predicting misogynous messages (0.57 F-1 score) and offensive ones (0.58 F-1 score). Classifiers trained on misogyny perform the worst when predicting HS (0.47) and averagely when predicting offensiveness (0.68). The 'offensive' class seems to be the most transferable one to an external domain, and also when training on HS and misogyny, testing on offensiveness produces the best results. The reason might be that both HS and misogynous instances contain offensive language, as, except for the Implicit Hate corpus [ElSherief et al., 2021], such datasets contain

mostly explicit hate. We carried out a manual inspection in order to verify it. Indeed, when training on offensiveness and testing on misogyny, correctly-predicted instances contain slurs against women.[3] Moreover, a large number of false positives (729) occur when instances contain slurs, which triggered the positive class to be predicted, but not against women.[4] Although in less quantity, there are also false negatives (409), with a tendency in showing hate against women implicitly or ironically.[5] The same tendency is observed when tested on HS, although I have more false negatives (2048) than false positives (1411).

On the contrary, models trained on misogyny and tested on HS do not generalize as well as models trained on HS and tested on misogyny. The reason might be that HS is a broader form of harmful language, containing cases of misogyny as well, while misogyny does not contain other forms of hate speech, therefore it might only classify correctly instances targeting women. Moreover, usually misogyny classifiers tend to over-rely on gender-based identity terms (e.g., women, girls) [Nozza et al., 2019], and that might cause a lot of false positives in the negative instances of hate speech datasets.

Indeed, when training on misogyny and testing on HS, I noticed that most religion- and race-based HS gets misclassified, producing large numbers of FNs (3710), compared to FPs (285). As expected, almost every FP instance contain gender-based identity terms (women, feminist, girls, females, bitch).

7.3 From HS Corpora to HS Legal Definitions

Defining HS has been a challenge not only for the NLP community but also for other disciplines, such as psychology, sociology, and law [Flick, 2020, Fortuna et al., 2020, Saha et al., 2019]. The multifaceted nature of HS it a very subjective phenomenon that can be influenced by personal biases [Goyal et al., 2022] and different legal contexts [Bleich,

[3]For instance, "This cunt likes every single post on that bitch's account and she doesn't even follow him".

[4]For instance, "When you know you're a piece of shit and own it like a boss <URL>".

[5]Examples: "But how can they be raped if no men are there to rape them?"; "What's the difference between your wife and your job? After 5 years, your job still sucks"

Prompt for normative definitions	[definition] According to the definition provided, is the following hate speech? Reply with 1 for YES or 0 for NO ?[text]
Prompt for moral foundations	[message_1] expresses [moral value] and hate speech. [message_2] expresses [moral value] but not hate speech. Does the following message express hate speech?

Table 7.3: Prompt Templates. The first is used to analyze the compatibility between HS normative definitions and corpora (Section 7.3). The second to assess the impact of moral values in HS recognition (Section 7.4).

	UN	CoE_1	CoE_2	TW	FB	Ggl	FN18	NODEF	Dataset AVG
AMI	0.42	0.53	0.51	0.52	0.54	0.60	0.68	**0.70**	0.56
DAVIDSON	0.49	0.71	**0.79**	0.42	0.35	0.51	0.43	0.65	0.55
HATEXPLAIN	0.50	0.38	0.36	0.52	0.64	0.24	**0.72**	0.65	0.51
EDOS	0.50	0.67	0.64	0.67	0.67	0.65	**0.70**	0.64	0.65
IMPLICIT HS CORPUS	0.72	0.48	0.57	0.72	**0.73**	0.54	0.72	0.65	0.65
INCELS	0.70	0.71	0.72	**0.74**	0.72	0.55	0.72	0.65	0.69
MRC	0.34	0.30	0.33	0.32	0.42	**0.45**	0.31	0.31	0.35
MTC	**0.40**	0.22	0.27	0.28	0.28	0.22	0.24	0.20	0.26
OLID	0.61	0.56	0.61	0.58	0.54	0.53	**0.63**	0.57	0.59
SEXISM	0.85	0.84	0.87	0.85	**0.88**	0.80	0.83	0.86	0.86
WASEEM	0.72	0.60	0.74	0.55	**0.76**	0.57	0.68	0.58	0.66
DEF AVG	0.59	0.56	0.60	0.59	**0.63**	0.51	0.59	0.59	

Table 7.4: Prompting with definitions scores (F1) against original gold standard labels.

2017]. Automatic HS detection in NLP is a task that is developed on the basis of hand-picked definitions and empirical evaluations. Although this tactic can prove effective in some cases, in others it can be proved detrimental as it is hardly reproducible [Korre et al., 2023]. This is most evident at the annotation level, where annotators are often asked to annotate HS by following vague guidelines [Fortuna et al., 2020] that hinder the reusability of corpora and their adoption for generalisable HS understanding [Yin and Zubiaga, 2021].

Approach. The experiment aims to evaluate to which extent annotated corpora are compatible with the normative definitions of HS presented in Section 7.1.4. For the purposes of this study, I analyze the ability of an LLM to classify a text for HS when is prompted with different normative definitions. The model that I used to generate the results was Stable Beluga 2; a Llama2 70B model finetuned on an Orca style Dataset [Ma-

han et al., 2023]. I provide an example of the prompts that I used in Table 7.3.

We queried from O-Dang a stratified sample of 1,000 messages for each of the 11 corpora described in Section 7.1 and performed 8 prompting classification experiment: one for each HS normative definition in Table 7.1 and one without any definition. The following is an example of sparql query that can be used to collect annotated texts and HS laws from O-Dang.

```
SELECT DISTINCT ?id ?text ?label ?phenomenon ?law WHERE {
?id a dul:InformationText;
dc:description ?text;
dul:isDescribedBy ?desc.
?desc :defines ?phenomenon;
:isRelatedToDescription ?law;
oa:body ?label .
?law a :Law .
}
```

Results. A first result that emerges from my experiment is that the injection of HS normative definitions in prompts increases model performance in 10 cases out of 11. As it can be observed in Table 7.4, only AMI is negatively impacted by definitions, while other corpora show an increase of performance ranging from +0.02 to +0.2 F1 score. Interestingly, even corpora annotated for moral values shows some compatibility with HS normative definitions: MRC scored +0.14 F1, MTC +0.20.

A second relevant result regards the different degrees of compatibility between corpora and HS normative definitions. Corpora are poorly aligned with legal definitions. Best classifications on MTC are obtained in combination with UN, while on Davidson the best choice was CoE_2. Definitions gathered from social media community guidelines have the greater impact in five cases. More specifically: FB aligns better with predictions on Implicit HS Corpus, Sexism, and Waseem, TW with Incels, and Ggl with MRC. Finally, FN18 is the most impactful for the classification of three corpora (HateXplain, EDOS,

and OLID).

Summarizing, HS normative definitions not only seem to have an impact on classification, but they might be used to explore differences between corpora and they can be exploited to perform more accurate instruction-tuning tasks for HS detection.

7.4 Interaction between HS, Moral Values, and Appraisal

In this section I report a data augmentation experiment aimed at exploring the interaction between HS-related manifestations and phenomena that may have an impact on spreading and understanding HS. I chose morality and appraisal as case studies, since a growing number of research is investigating the impact of moral rhetoric both in datasets creation Forbes et al. [2020], van der Meer et al. [2023] and in text classification Liscio et al. [2023].

Approach. In the first approach, a model is trained to identify the five moral dyads included in the MTC corpus ——care, fairness, authority, purity, and loyalty—— and subsequently apply this model to HS-related datasets. The study aims to evaluate whether incorporating these moral values into HS detection models improves their overall performance. To achieve this objective, three key steps are undertaken: (a) training a model to identify the 5 moral values, (b) applying this model for the detection of moral values in hate speech datasets, and (c) performing an ablation study aimed at identifying the impact of moral foundations and appraisal in HS predictions. More specifically, we conducted a few-shot experiment where I designed a prompt template with examples of messages expressing a moral foundation or an appraisal dimension before asking the LLM to classify a given message (Table 7.3). We extracted a stratified sample of 1000 messages from the HS corpus and used it as a test set. Then, I randomly pick from the remaining dataset a message that expresses both the moral or appraisal dimension (e.g., purity) and HS, and a message that expresses the moral or appraisal dimension, but not HS, and I filled the prompt template with them. Finally, I used Stable Beluga [Mahan et al., 2023] to determine if the message was HS. In parallel, I performed a zero-shot classification and computed the delta between the F-1 scores obtained with and without

	Care	Fairness	Authority	Loyalty	Purity
Ethos	-0.06	-0.05	-0.11	**0**	-0.05
HateXplain	-0.02	-0.01	-0.02	**+0.02**	**+0.02**
Implicit	-0.04	-0.02	**-0.01**	-0.02	-0.05
Incels	**-0.03**	-0.07	-0.05	-0.07	-0.07
Waseem	+0.03	-0.04	-0-02	-0.16	**+0.07**

Table 7.5: Results of few-shot experiments on the interaction between HS and moral foundations. For each moral value, it is reported the delta between the F-1 scores with and without adding moral knowledge. A negative values means that adding moral knowledge results in a lower performance.

messages expressing moral values.

Results. Table 7.5 shows results of the few-shot experiment, where each row represents a dataset and each column a moral foundation injected in the prompt template. A negative score shows that injecting moral knowledge worsens the performance. A positive score shows an improvement. The impact of morality depends on the observed corpus and that each moral dyad has its own effect. HateXplain and Waseem benefit from the injection of moral knowledge in two dyads out of five: *loyalty* and *purity* for the former; *purity* and *care* for the latter. On the contrary, the effect of moral values on Implicit HS and Incels corpora is always negative. Dyads that lead to a minor drop of performance are *authority* for the former and *care* for the latter.

Table 7.6 shows results of the few-shot experiments with appraisal dimensions that seem to be more impactful than moral dyads. Except from Waseem, in every corpora there is at least one appraisal dimension that increases the performance against the zero-shot experiment. The impact of each dimension varies on the basis of the examined corpus. Incels dataset always scores a highest performance, with a maximum positive delta when the prompt template is filled with texts expressing 'control'. 'Certainty' is the only appraisal variable with a positive impact on Davidson (+0.03) and HateXplain (+0.01), 'consistency' on Implicit (+0.05). Ethos benefits from a positive impact from 'certainty' (+0.07) and 'responsibility' (+0.07).

	Certainty	Consistency	Control	Responsibility
Davidson	**+0.03**	-0.06	-0.02	-0.03
Ethos	**+0.08**	-0.16	-0.16	+0.07
HateXplain	**+0.01**	-0.02	-0.04	0
Implicit	-0.07	**+0.05**	0	-0.04
Incels	+0.05	+0.04	+0.11	**+0.08**
Waseem	**-0.01**	-0.08	-0.06	-0.07

Table 7.6: Results of few-shot experiments on the interaction between HS and appraisal dimensions. For each moral value, it is reported the delta between the F-1 scores with and without adding moral knowledge. A negative values means that adding moral knowledge results in a lower performance.

7.5 Conclusion of the Chapter

In this chapter I exploited O-Dang to align corpora annotated for HS and abusive language under a common semantic model. The alignment enabled a systematic study of bias in annotated corpora along three axes.

The analysis of transferrability of knowledge between different forms of abusive language showed that the knowledge encoded in messages annotated for offensiveness can be more easily transferred to HS and misogyny, confirming previous works on this topic [Pamungkas et al., 2023]. However, this higher generalizability may hide social biases: slurs typical of offensive language are not always harmful, as it is clear from the great number of false positives, as they are often used in informal language and spontaneous writing, and the degree of offensiveness perceived might vary across cultural background among other factors [Sap et al., 2019, Pamungkas et al., 2023].

The experiment aimed at assessing the compatibility of corpora with **HS normative definitions** demonstrates a significant role of definitions in HS detection and understanding. Results also show that annotated corpora differ in their alignment with existing norms, suggesting that the composition of training sets for HS detection must be guided by the context in which the model must be applied.

Results of the **integration of moral values and emotional appraisal** in HS detection demonstrated that HS classification is influenced by these phenomena, which can have a role in providing a deeper understanding of what causes the spreading of

online discrimination and through which means they are conveyed.

Chapter 8

Conclusion and Future Work

In this thesis I designed a framework for analyzing and addressing bias in datasets that integrates methodologies from NLP and SW to account for different forms of bias in a comprehensive manner. More specifically, I designed: *i.* an Ontology Network that enabled the documentation and comparison of existing archives and annotated corpora; *ii.* A set of resources for biographical event detection that supported a fine-grained analysis of how people are represented in biographies. Such a framework has been adopted to investigate the following research questions:

RQ 1. How representational bias can be detected and measured?

We investigated this research issue in Chapter 5, where I measured the impact of *representational* bias in English Wikipedia biographies in an intersectional perspective. The analysis showed that the comparison of single socio-demographic features is not effective for a thorough understanding of the phenomenon, since the combination of these features can amplify or reduce bias. For instance, women are more likely to be associated with events that denote stereotypical traits if they are Transnational and if they have certain occupations, such as modelling. Conversely, age is correlated with more prestigious events related to women careers. All these results confirmed the non-neutrality of Wikipedia contents, that are often uncritically used to train NLP technologies. My framework provides a semantic-aware approach to better understand the nature of bias in digital archives through the joint extraction of socio-demographic features from struc-

tured sources of knowledge and the automatic detection of biographical events.

RQ 2. Which strategies can be adopted to detect allocative biases?

In Chapter 6 I discovered that Wikidata is more prone to *allocative* bias against Transnational writers than other resources: the percentage of works written by them and recorded in Wikidata is considerably lower than ones in Goodreads and Open Library. This empirically confirms that the knowledge available in the Wikimedia ecosystem does not only reflect but amplifies inequalities in my societies. My framework also enabled the mitigation of this issue through the adoption of two strategies that led to the creation of the WL-KG: a knowledge base for the discovery of *allocative* bias against Transational writers. By aligning Wikidata with other digital archives I obtained a reduction of the underrepresentation of Transnational writers; by extracting biographical events from unstructured texts I increased the amount of structured information about them. Combined, these two approaches represent a valuable approach to create more inclusive archives. On top of this bias mitigation strategy, I developed the WL-KG visualization platform, a resource conceived for reducing bias also in other disciplines like Literature and Digital Humanities, by making my knowledge base accessible to non-specialist users.

RQ 3. Which measures can be implemented to discover research design biases in annotated corpora for abusive language detection?

In Chapter 7 I implemented a framework for comparing corpora annotated for HS and abusive language along different axes: the generalization over different types of abusive language; the compatibility of corpora with HS definitions; the impact of moral values and emotional appraisal on the detection of HS messages. My findings showed that each corpus is more likely to be compatible with certain HS definitions and with specific appraisal dimensions and moral values. On one side, this hinders the generalizability of HS corpora, showing that research bias prior to data annotation is a crucial issue that has not been resolved yet. On the other side, my framework allows for a more informed data selection strategy tailored to the task and research interest since it enables the alignment between corpora. For instance, training a NLP classifier for HS detection on a specific social media platform can be guided by the selection of the most compatible

datasets with the platform's guidelines. More generally, my approach, being based on the integration of dataset through a common ontology, may provide a contribution in challenging the documentation debt that affects many resources and tools in NLP.

8.1 Research Contribution

This thesis contributes to a better understanding of *allocative* and *representational* bias in datasets and annotated corpora by delivering a novel framework that emphasizes the role of documentation for a more transparent analysis of this phenomenon. SW technologies demonstrated to be particularly suitable to perform replicable and multifaceted investigations of bias, while biographical event detection paved the way for more complex analyses of the semantic expressed in biographies. During this work a number of resources have been developed to support the analysis of bias performed throughout the whole thesis.

Ontologies.

1. The People in the Media Ontology (PiM-O) is a mid-level ontology that encodes relevant concepts for the multi-faceted representation of people in social media and public archives.

2. The Under-Represented Writers' Ontology Network (UR-ON) is a domain ontology that enables the collection and representation of non-Western writers and their works.

3. The Ontology of Dangerous Speech (O-Dang), a domain ontology suited for the comparison of existing corpora annotated for HS and abusive language .

Corpora.

1. Two corpora for biographical annotated for biographical event detection according to existing guidelines for event detection and co-reference resolution.

2. A corpus annotated for the detection of emotional appraisal.

3. A corpus annotated for the detection of moral values.

NLP Systems.

1. A model for the automatic detection of biographical events.

2. A pipeline for the extraction of biographical triples from raw text.

Methods for Bias Detection.

1. A methodology for the analysis of *representational* bias in Wikipedia English biographies that relies on biographical event detection and accounts for an intersectional perspective on this issue.

2. An investigation of *allocative* bias in Wikidata and implemented a methodology to reduce them through its alignment with other public archives and through the extraction of biographical triples.

3. A comparative analysis of corpora annotated for HS and abusive language. The analysis relies on O-Dang, which has proven to be useful for the discovery of research design bias that may affect annotated corpora.

8.2 Relevant Publications

Conferences

- **Stranisci, M. A.**, Bernasconi, E., Patti, V., Ferilli, S., Ceriani, M., & Damiano, R. (2023, October). **The World Literature Knowledge Graph**. In International Semantic Web Conference (pp. 435-452). Cham: Springer Nature Switzerland.
  The World Literature Knowledge Graph that has been presented at this conference received a grant of 150,000 euro from the NGI-Consortium[1].

[1] https://spaces.fundingbox.com/spaces/the-next-generation-internet-ngi-community-ngi-search/652e67174b8f8d23cf985971

- **Stranisci, M. A.**, Damiano, R., Mensa, E., Patti, V., Radicioni, D., & and Caselli, T. (2023, July).

  WikiBio: a Semantic Resource for the Intersectional Analysis of Biographical Events. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (pp. 12370–12384). Association for Computational Linguistics.

  The paper earned an Outstanding Paper Award at the conference[2].

- **Stranisci, M. A.**, Patti, V., & Damiano, R. (2023, February). **User-Generated World Literatures: a Comparison between Two Social Networks of Readers**. In Proceedings of the Italian Research Conference on Digital Libraries (pp. 38-46).

- Lai, M., **Stranisci, M. A.**, Bosco, C., Damiano, R., & Patti, V. (2022, August). **Analysing Moral Beliefs for Detecting Hate Speech Spreaders on Twitter**. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 149-161). Cham: Springer International Publishing.

- **Stranisci, M. A.**, Frenda, S., Ceccaldi, E., Basile, V., Damiano, R., & Patti, V. (2022, July). **APPReddit: a Corpus of Reddit Posts Annotated for Appraisal**. In Proceedings of the Language Resources and Evaluation Conference (pp. 3809-3818). European Language Resources Association.

- **Stranisci, M. A.**, Patti, V., & Damiano, R. (2021). **Representing the Under-Represented: a Dataset of Post-Colonial, and Migrant Writers**. In 3rd Conference on Language, Data and Knowledge (LDK 2021). Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Journals

[2] https://2023.aclweb.org/program/best_papers/

- **Stranisci, M. A.**, De Leonardis, M., Bosco, C., & Patti, V. (2021). **The expression of moral values in the Twitter debate: a corpus of conversations**. IJCoL. Italian Journal of Computational Linguistics, 7(7-1, 2), 113-132.

- **Stranisci, M. A.**, Bosco, C., Cignarella, A. T., Frenda, S., & Patti, V. (2021). **Hate speech e dangerous speech in Twitter**. RASSEGNA ITALIANA DI LINGUISTICA APPLICATA, 3, 191-207.

Workshops

- **Stranisci, M. A.**, Mensa, E., Damiano, R., Radicioni, D., & Diakite, O. (2022, June). **Guidelines and a Corpus for Extracting Biographical Events**. In Proceedings of the 18th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (pp. 20-26).

- **Stranisci, M. A.**, Frenda, S., Lai, M., Araque, O., Cignarella, A. T., Basile, V., Patti, V. & Bosco, C. (2022, June). **O-Dang! The Ontology of Dangerous Speech Messages**. In Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data (pp. 2-8), co-located with LREC 2022, Marseille, France. European Language Resources Association.

- **Stranisci, M. A.**, Basile, V., Damiano, R., & Patti, V. **Mapping Biographical events to ODPs through Lexico-Semantic Patterns** (2021, October). In Proceedings of the 12th Workshop on Ontology Design and Patterns (WOP 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), volume 3011 of CEUR WORKSHOP PROCEEDINGS, pages 1–12. CEUR-WS, 2021.

Reports of Participation and Organization of Shared Tasks

- Di Bonaventura, C., Muti, A., & **Stranisci, M. A.** (2023, September). **O-Dang at HODI and HaSpeeDe3: A Knowledge-Enhanced Approach to Homotransphobia and Hate Speech Detection in Italian**. In Proceedings of the

Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023). *Our system was the best system at the third edition of HaSpeeDe.*

- Lai, M., Stranisci, M. A., Bosco, C., Damiano, R., & Patti, V. (2021, September). **HaMor at the Profiling Hate Speech Spreaders on Twitter**. In Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (pp. 2047-2055). *Our system has been awarded one of the Best of 2021 labs at CLEF*[3].

- Sanguinetti, M., Comandini, G., Di Nuovo, E., Frenda, S., **Stranisci, M. A.**, Bosco, C., Caselli, T., Patti, V. & Russo, I. (2020, December). **Haspeede 2@ evalita2020: Overview of the EVALITA 2020 hate speech detection task**. Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020).

Other Relevant Publications

- Saracco, A., Lillo, A., **Stranisci, M. A.**, & Gena, C. (2024, March). **Human Robot Interaction through an ontology-based dialogue engine**. In 19th Annual ACM/IEEE International Conference on Human Robot Interaction (HRI) (pp. 1-5). ACM.

- Arthur, T. E. C. L., Cignarella, A. T., Frenda, S., Lai, M., **Stranisci, M. A.**, & Urbinati, A. (2023, December). **Debunker Assistant: a support for detecting online misinformation.** In Proceedings of the Ninth Italian Conference on Computational Linguistics (CLiC-it 2023) (pp. 1-5). *Debunker Assistant, that has been presented at this conference received a grant of 150,000 euro from the NGI-Consortium*[4].

[3] https://clef2022.clef-initiative.eu/index.php?page=Pages/accepted_papers.html
[4] https://spaces.fundingbox.com/spaces/the-next-generation-internet-ngi-community-ngi-search/652e67174b8f8d23cf985971

- Tontodimamma, A., Anzani, S., **Stranisci, M. A.**, Basile, V., Ignazzi, E., & Fontanella, L. (2022, September). **An experimental annotation task to investigate annotators' subjectivity in a misogyny dataset**. In ASA 2022 Data-Driven Decision Making BOOK OF SHORT PAPERS (pp. 281-286).

- Frenda, S., Cignarella, A. T., **Stranisci, M. A.**, Lai, M., Bosco, C., & Patti, V. (2021, June). **Recognizing Hate with NLP: The Teaching Experience of the #DEACTIVHATE Lab in Italian High Schools**. In Proceedings of the Eighth Italian Conference on Computational Linguistics (pp. 1-7).

8.3 Ethics Section

This thesis covers the topic of bias in datasets, especially against people born outside Europe and North America. This raises a number of potential ethical issues that I tried to mitigate within my work. The first is the classification of people based on global taxonomies that have several limitations. The Global North *versus* Global South dichotomy has been introduced during the Cold War to represent a geopolitical system that does not exist anymore. In recent years it has been adopted by the UN's Development Agency[5] in one of the many initiatives aimed at reducing poverty in disadvantaged countries whose condition can be traced back to colonization. Additionally, in the list of Global South countries are still counted some of the richer economies in the world like China and South Korea. The opposition between Western and non-Western is affected by similar issues. Orientalism is a vague concept that compress different cultures and countries in a stereotypical representation of otherness against Europe and North America [Said, 1977]. My choice fell on the term *Transnational* to limit its scope to the engagement between a person and the border of their birth country. However, this categorization is prone to a high number of false positives, since many children of European officials were born in African and Asian countries. Moreover, it is not possible to distinguish between people

[5] https://www.undp.org/sites/g/files/zskgke326/files/migration/cn/UNDP-CH-PR-Publications-UNDay-for-South-South-Cooperation.pdf

belonging to local elites, which benefit from privileges, and the ones who belong to lower classes.

In addition, it is worth mentioning that Wikidata comes with many limitations in its taxonomy that hamper a fair collection of data. Squeezing two orthogonal features like 'gender' and 'sexual orientation' in a unique property is not fully respectful of non-binarism. A similar issue affects the 'nationality' class that is considered a sub-class of 'ethnic group' in Wikidata. Unfortunately, there are no comparable resources to Wikidata in size, thus I relied on this knowledge base by adopting citizenship as a classification strategy. I did not find alternatives to the issue related to the misrepresentation of non-binarism but in future work I will address this challenge.

A final ethics concern regards the background of the PhD student and his advisors, all of which are Europeans and do not belong to ethnic minorities. To mitigate such an issue I explored Post-colonial and African American studies in order to better understand the theoretical framework that supports these fields of research.

8.4 Limitations and Future Work

Our SW-based set of resources supports the performance of a more in-depth analysis of bias across corpora and datasets. SW standards are not uniformly adopted within the NLP community, though. Future work will be devoted to improving the accessibility of my resources with the implementation of more intuitive templates for accessing and querying my knowledge base. In this work I presented a novel framework for bias detection that enabled a thorough analysis of this phenomenon in datasets. However, I did not investigate the correlation between these findings and the models' behaviour. This represents a limitation of my work, since it prevents a full analysis of all sources of bias. In future work I will extend my framework to the analysis of bias in models' embeddings and in their classifications.

Bibliography

G. Abercrombie, V. Basile, D. Bernardi, S. Dudy, S. Frenda, L. Havens, E. Leonardelli, S. Tonelli, et al. 2nd workshop on perspectivist approaches to nlp (nlperspectives 2023). In *CEUR Workshop Proceedings*, volume 3494, pages 1–2. CEUR-WS, 2023.

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

J. Aguilar, C. Beller, P. McNamee, B. Van Durme, S. Strassel, Z. Song, and J. Ellis. A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards. In *Proceedings of the second workshop on EVENTS: Definition, detection, coreference, and representation*, pages 45–53, 2014.

J. Araki, L. Mulaffer, A. Pandian, Y. Yamakawa, K. Oflazer, and T. Mitamura. Interoperable annotation of events and event relations across domains. In *Proceedings 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 10–20, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-4702.

M. Atari, J. Haidt, J. Graham, S. Koleva, S. T. Stevens, and M. Dehghani. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*, 2023.

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*, pages 722–735. Springer, 2007.

J. L. Austin. *How to do things with words*, volume 88. Oxford university press, 1975.

P. Bajaj, C. Xiong, G. Ke, X. Liu, D. He, S. Tiwary, T.-Y. Liu, P. Bennett, X. Song, and J. Gao. Metro: Efficient denoising pretraining of large scale autoencoding language models with model generated signals. *arXiv preprint arXiv:2204.06644*, 2022.

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.

R. Bakker, R. A. van Drie, M. de Boer, R. van Doesburg, and T. van Engers. Semantic role labelling for dutch law texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 448–457, 2022.

D. Bamman. Litbank: Born-literary natural language processing. *Computational Humanites, Debates in Digital Humanities (2020, preprint)*, 2020.

D. Bamman and N. A. Smith. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376, 2014.

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)*, pages 352–361, 2013.

D. Bamman, S. Chaturvedi, E. Clark, M. Fiterau, and M. Iyyer, editors. *Proceedings of the First Workshop on Narrative Understanding*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-2400.

L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186, 2013.

J. Bandy and N. Vincent. Addressing "documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/54229abfcfa5649e7003b83dd4755294-Paper-round1.pdf.

S. Barocas, K. Crawford, A. Shapiro, and H. Wallach. The problem with bias: from allocative to representational harms in machine learning. special interest group for computing. *Information and Society (SIGCIS)*, 2, 2017.

M. Barreto and D. M. Doyle. Benevolent and hostile sexism in a shifting global context. *Nature reviews psychology*, 2(2):98–111, 2023.

E. Bassignana and B. Plank. Crossre: A cross-domain dataset for relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3592–3604. Association for Computational Linguistics, 2022a.

E. Bassignana and B. Plank. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, 2022b.

E. Bassignana, V. Basile, V. Patti, et al. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS, 2018.

E. M. Bender and B. Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

S. Benesch. Dangerous speech: A proposal to prevent group violence. *Voices That Poison: Dangerous Speech Project proposal paper*, 2012. URL http://worldpolicy.org/wp-content/uploads/2016/01/Dangerous-Speech-Guidelines-Benesch-January-2012.pdf.

E. Bernasconi, M. Ceriani, M. Mecella, and A. Morvillo. Automatic knowledge extraction from a digital library and collaborative validation. In G. Silvello, O. Corcho, P. Manghi, G. M. Di Nunzio, K. Golub, N. Ferro, and A. Poggi, editors, *Linking Theory and Practice of Digital Libraries*, Lecture Notes in Computer Science, pages 480–484. Springer, 2022. ISBN 978-3-031-16801-7.

E. Bernasconi, M. Ceriani, M. Mecella, and T. Catarci. Design, realization, and user evaluation of the ARCA system for exploring a digital library. *International Journal on Digital Libraries*, 24(1):1–22, 2023.

E. Bleich. Freedom of expression versus racist hate speech: Explaining differences between high court regulations in the usa and europe. In *Regulation of Speech in Multicultural Societies*, pages 110–127. Routledge, 2017.

S. L. Blodgett, L. Green, and B. O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, 2016.

S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

R. Bommasani, K. Klyman, S. Longpre, S. Kapoor, N. Maslej, B. Xiong, D. Zhang, and P. Liang. The foundation model transparency index. *arXiv e-prints*, pages arXiv–2310, 2023.

P. Bonacich. Some unique properties of eigenvector centrality. *Social networks*, 29(4): 555–564, 2007.

C. Bonial and M. Palmer. Comprehensive and consistent propbank light verb annotation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3980–3985, 2016.

C. Bonial, O. Babko-Malaya, J. D. Choi, J. Hwang, and M. Palmer. Propbank annotation guidelines. *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*, 2010.

C. Bonial, W. Corvey, M. Palmer, V. V. Petukhova, and H. Bunt. A hierarchical unification of lirics and verbnet semantic roles. In *2011 IEEE Fifth International Conference on Semantic Computing*, pages 483–489. IEEE, 2011.

J. Borsje, F. Hogenboom, and F. Frasincar. Semi-automatic financial events discovery based on lexico-semantic patterns. *International Journal of Web Engineering and Technology*, 6(2):115–140, 2010.

J. Bos, V. Basile, K. Evang, N. J. Venhuizen, and J. Bjerva. The groningen meaning bank. *Handbook of linguistic annotation*, pages 463–496, 2017.

B. Boter, M. Rensen, and G. Scott-Smith. *Unhinging the National Framework: Perspectives on Transnational Life Writing*. Sidestone Press, 2020.

J. G. Breslin, S. Decker, A. Harth, and U. Bojars. Sioc: an approach to connect web-based communities. *International Journal of Web Based Communities*, 2(2):133–142, 2006.

S. Brown, P. Clements, I. Grundy, S. Balazs, and J. Antoniuk. An introduction to the orlando project. *Tulsa Studies in Women's Literature*, 26(1):127–134, 2007.

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*, pages 87–97, 2017.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

G. Bruseker, N. Carboni, and A. Guillem. Cultural heritage data management: The role of formal ontology and cidoc crm. *Heritage and archaeology in the digital age: acquisition, curation, and dissemination of spatial cultural heritage data*, pages 93–131, 2017.

H. Bunt and M. Palmer. Conceptual and representational choices in defining an iso standard for semantic role annotation. In *Proceedings Ninth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9), Potsdam*, pages 41–50, 2013.

H. Bunt and L. Romary. Requirements on multimodal semantic representations. In *Proceedings of ISO TC37/SC4 Preliminary Meeting*, pages 59–68. KAIST, 2002.

F. Cabitza, A. Campagner, and V. Basile. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868, 2023.

P.-L. H. Cabot and R. Navigli. Rebel: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, 2021.

A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

E. S. Callahan and S. C. Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915, 2011.

E. Cambria, R. Speer, C. Havasi, and A. Hussain. Senticnet: A publicly available semantic resource for opinion mining. In *2010 AAAI fall symposium series*, 2010.

T. Caselli and J. Bos. Investigating interoperable event corpora: limitations of reusability of resources and portability of models. *Language Resources and Evaluation*, pages 1–31, 2023.

T. Caselli and P. Vossen. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, 2017.

T. Caselli, M. van Erp, A.-L. Minard, M. Finlayson, B. Miller, J. Atserias, A. Balahur, and P. Vossen, editors. *Proceedings of the First Workshop on Computing News Storylines*, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-45. URL https://aclanthology.org/W15-4500.

N. Chambers and D. Jurafsky. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, 2008.

W.-T. Chen, C. Bonial, and M. Palmer. English light verb construction identification using lexical knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

C. Chiarcos. Ontologies of linguistic annotation: Survey and perspectives. In *LREC*, pages 303–310. Citeseer, 2012.

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Y. Chung, E. Kuzmenko, S. Tekiroglu, M. Guerini, et al. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829. ASSOC COMPUTATIONAL LINGUISTICS-ACL, 2019.

P. Cimiano, C. Chiarcos, J. P. McCrae, and J. Gracia. *Linguistic linked data*. Springer, 2020.

K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, 1989(8), 1989.

K. W. Crenshaw. *On intersectionality: Essential writings.* The New Press, 2017.

H. Cunningham. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36:223–254, 2002.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the international AAAI conference on web and social media*, 11(1):512–515, 2017.

M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

K. v. Deemter and R. Kibble. On coreferring: Coreference in muc and related annotation schemes. *Computational linguistics*, 26(4):629–637, 2000.

H. Devinney, A. Eklund, I. Ryazanov, and J. Cai. Developing a multilingual corpus of wikipedia biographies. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 285–294, 2023.

J. Dias, S. Mascarenhas, and A. Paiva. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Emotion modeling*, pages 44–56. Springer, 2014.

H. Ding and E. Riloff. Human needs categorization of affective events using labeled and unlabeled data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1919–1929, 2018.

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, 2018.

G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, and R. Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf.

J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, 2021.

N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, O. Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.

M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, 2021.

M. Erigha. Race, gender, hollywood: Representation in cultural production and digital media's potential for change. *Sociology compass*, 9(1):78–89, 2015.

C. Federmann, I. Giannopoulou, C. Girardi, O. Hamon, D. Mavroeidis, S. Minutoli, and M. Schröder. Meta-share v2: An open network of repositories for language resources including data and tools. In *LREC*, pages 3300–3303, 2012.

E. Fersini, P. Rosso, M. Anzovino, et al. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228, 2018.

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*, pages 2624–2635, 2022.

C. J. Fillmore. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32, 1976. doi: https://doi.org/10.1111/j.1749-6632.1976.tb25467.x. URL https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1976.tb25467.x.

C. J. Fillmore et al. Frame semantics. *Cognitive linguistics: Basic readings*, 34:373–400, 2006.

C. Flick. The legal framework on hate speech and the internet good practices to prevent and counter the spread of illegal hate speech online. In *Language, Gender and Hate Speech A Multidisciplinary Approach*. Fondazione Università Ca' Foscari, dec 2020. doi: 10.30687/978-88-6969-478-3/011. URL https://doi.org/10.30687%2F978-88-6969-478-3%2F011.

A. Fokkens, S. Ter Braake, N. Ockeloen, P. Vossen, S. Legêne, G. Schreiber, and V. de Boer. Biographynet: Extracting relations between people and events. In *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik*, pages 193–227. new academic press, 2017.

M. Forbes, J. D. Hwang, V. Shwartz, M. Sap, and Y. Choi. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, 2020.

P. Fortuna and S. Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

P. Fortuna, J. Soler, and L. Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794, 2020.

J. S. Franklin, K. Bhanot, M. Ghalwash, K. P. Bennett, J. McCusker, and D. L. McGuinness. An ontology for fairness metrics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 265–275, 2022.

F. Frasincar, J. Borsje, and L. Levering. A semantic web-based approach for building personalized news services. *International Journal of E-Business Research (IJEBR)*, 5 (3):35–53, 2009.

S. Frenda, K. Noriko, V. Patti, P. Rosso, et al. Stance or insults? In *Proceedings of the Ninth International Workshop on Evaluating Information Access (EVIA 2019), a Satellite Workshop of the NTCIR-14 Conference*, pages 15–22. National Institute of Informatics, 2019. URL https://research.nii.ac.jp/ntcir/workshop/OnlineProceedings14/pdf/evia/03-EVIA2019-EVIA-FrendaS.pdf.

S. Frenda, A. T. Cignarella, V. Basile, C. Bosco, V. Patti, and P. Rosso. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193, 2022. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2021.116398. URL https://www.sciencedirect.com/science/article/pii/S0957417421016870.

P. Gajo, A. Muti, K. Korre, S. Bernardini, and A. Barrón-Cedeño. On the identification and forecasting of hate speech in inceldom. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 373–384, 2023.

R. J. Gallagher, M. R. Frank, L. Mitchell, A. J. Schwartz, A. J. Reagan, C. M. Danforth, and P. S. Dodds. Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(1):4, 2021.

A. Gangemi. Super-duper schema: an owl2+ rif dns pattern. In *Proceedings of DeepKR Challenge Workshop at KCAP11. Edited by Chaudry*, volume 2011, 2011.

A. Gangemi and V. Presutti. Ontology design patterns. In *Handbook on ontologies*, pages 221–243. Springer, 2009.

A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with dolce. In *International conference on knowledge engineering and knowledge management*, pages 166–181. Springer, 2002.

A. Gangemi, M. Alam, L. Asprino, V. Presutti, and D. R. Recupero. Framester: A wide coverage linguistic linked data hub. In *Knowledge Engineering and Knowledge Management: 20th International Conference, EKAW 2016, Bologna, Italy, November 19-23, 2016, Proceedings 20*, pages 239–254. Springer, 2016.

L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The pile: An 800gb dataset of diverse text for language modeling. *CoRR*, abs/2101.00027, 2021. URL https://arxiv.org/abs/2101.00027.

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.301. URL https://aclanthology.org/2020.findings-emnlp.301.

M. Gerlach and F. Font-Clos. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126, 2020.

A. Gómez-Pérez. Ontology evaluation. In *Handbook on ontologies*, pages 251–273. Springer, 2004.

H. Gonen and Y. Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 609–614, 2019.

N. Goyal, I. D. Kivlichan, R. Rosen, and L. Vasserman. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 28, 2022. URL https://api.semanticscholar.org/CorpusID:248496288.

E. Graells-Garrido, M. Lalmas, and F. Menczer. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 165–174, 2015.

J. Graham and J. Haidt. The moral foundations dictionary, 2012. URL https://moralfoundations.org/wp-content/uploads/files/downloads/moral%20foundations%20dictionary.dic.

J. Graham, J. Haidt, and B. A. Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029, 2009.

J. Graham, J. Haidt, M. Motyl, S. Koleva, R. Iyer, S. P. Wojcik, and P. H. Ditto. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47:55–130, 2013.

L. Grimminger and R. Klinger. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online, Apr. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wassa-1.18.

R. Grishman and B. M. Sundheim. Message understanding conference-6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996.

G. Guizzardi. *Ontological foundations for structural conceptual models.* Phd thesis - research ut, graduation ut, University of Twente, Oct. 2005.

S. Gururangan, D. Card, S. Dreier, E. Gade, L. Wang, Z. Wang, L. Zettlemoyer, and N. A. Smith. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, 2022.

F. Hamborg, N. Meuschke, C. Breitinger, and B. Gipp. news-please: A generic news crawler and extractor. In *15th International Symposium of Information Science, Berlin, Germany*. Zenodo, 2017.

M. Hammersley and R. Gomm. Bias in social research. *Sociological research online*, 2 (1):7–19, 1997.

P. He, X. Liu, J. Gao, and W. Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*, 1992.

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*, pages 219–236. Springer, 2019.

S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating nlp using linked data. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II 12*, pages 98–113. Springer, 2013.

J. Hoffart, F. M. Suchanek, K. Berberich, E. Lewis-Kelham, G. De Melo, and G. Weikum. Yago2: exploring and querying world knowledge in time, space, context, and many

languages. In *Proceedings of the 20th international conference companion on World wide web*, pages 229–232, 2011.

J. Hofmann, E. Troiano, K. Sassenberg, and R. Klinger. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, 2020.

J. Hoover, M. Atari, A. M. Davani, B. Kennedy, G. Portillo-Wightman, L. Yeh, D. Kogon, and M. Dehghani. Bound in hatred: The role of group-based morality in acts of hate, Jul 2019. 10.31234/osf.io/359me.

J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.

F. R. Hopp, J. T. Fisher, D. Cornell, R. Huskey, and R. Weber. The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, pages 1–15, 2020.

D. Hovy and S. Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021.

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*, pages 57–60, 2006.

E. Hovy, T. Mitamura, and M. Palmer. Workshop on events: Definition, detection, coreference, and representation. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, 2013.

K.-H. Huang, I. Hsu, T. Parekh, Z. Xie, Z. Zhang, P. Natarajan, K.-W. Chang, N. Peng,

H. Ji, et al. A reevaluation of event extraction: Past, present, and future challenges. *arXiv preprint arXiv:2311.09562*, 2023.

I. Hulpus, J. Kobbe, H. Stuckenschmidt, and G. Hirst. Knowledge graphs meet moral values. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 71–80, Barcelona, Spain (Online), Sept. 2020. Association for Computational Linguistics (ACL).

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: 37–50, 2012.

P. S. Jacobs, G. Krupka, and L. Rau. Lexico-semantic pattern matching as a companion to parsing in text understanding. In *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.

F. James. Monte carlo theory and practice. *Reports on progress in Physics*, 43(9):1145, 1980.

H. R. Jauss and E. Benzinger. Literary history as a challenge to literary theory. *New literary history*, 2(1):7–37, 1970.

E. S. Jo and T. Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316, 2020.

K. Johnson and D. Goldwasser. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia, July 2018. Association for Computational Linguistics (ACL).

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*, 8:64–77, 2020.

L. J. Kamin. The science and politics of iq. *Social Research*, 41(3):387, 1974.

J. D. M.-W. C. Kenton and L. K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

J. Y. Kim, C. Ortiz, S. Nam, S. Santiago, and V. Datta. Intersectional bias in hate speech and abusive language datasets. In *Proceedings of the Fourteenth International Conference on Web and Social Media (ICWSM), Data Challenge Workshop*. AAAI Organization, 2020.

P. R. Kingsbury and M. Palmer. From treebank to propbank. In *LREC*, pages 1989–1993, 2002.

K. Kipper, A. Korhonen, N. Ryant, and M. Palmer. Extending verbnet with novel verb classes. In *LREC*, pages 1027–1032, 2006.

S. Kiritchenko and S. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, 2018.

H. Kirk, W. Yin, B. Vidgen, and P. Röttger. SemEval-2023 task 10: Explainable detection of online sexism. In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.305. URL https://aclanthology.org/2023.semeval-1.305.

K. Korre, J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, L. Dixon, and A. Barrón-cedeño. Harmful language datasets: An assessment of robustness. In Y.-l. Chung, P. R\"ottger, D. Nozza, Z. Talat, and A. Mostafazadeh Davani, editors, *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 221–230, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.woah-1.24. URL https://aclanthology.org/2023.woah-1.24.

H.-U. Krieger. Where temporal description logics fail: Representing temporally-changing relationships. In *KI 2008: Advances in Artificial Intelligence: 31st Annual German Conference on AI, KI 2008, Kaiserslautern, Germany, September 23-26, 2008. Proceedings 31*, pages 249–257. Springer, 2008.

H.-U. Krieger. A detailed comparison of seven approaches for the annotation of time-dependent factual knowledge in rdf and owl. In *Proceedings 10th joint ISO-ACL SIGSEM workshop on interoperable semantic annotation*, page 1, 2014.

H.-U. Krieger and T. Declerck. An owl ontology for biographical knowledge. representing time-dependent factual knowledge. In *BD*, pages 101–110, 2015.

K. Kurita, N. Vyas, A. Pareek, A. W. Black, and Y. Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, 2019.

M. Lai, M. A. Stranisci, C. Bosco, R. Damiano, V. Patti, et al. HaMor at the profiling hate speech spreaders on twitter. In G. Faggioli, N. Ferro, A. Joly, M. Maistro, and F. Piroi, editors, *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 2047–2055. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-2936/paper-178.pdf.

J. P. Lalor, Y. Yang, K. Smith, N. Forsgren, and A. Abbasi. Benchmarking intersectional biases in nlp. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609, 2022.

Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

H. Laurençon, L. Saulnier, T. Wang, C. Akiki, A. Villanova del Moral, T. Le Scao, L. Von Werra, C. Mou, E. González Ponferrada, H. Nguyen, et al. The bigscience

roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826, 2022.

T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. Prov-o: The prov ontology. *W3C recommendation*, 30, 2013.

E. Leonardelli, S. Menini, A. P. Aprosio, M. Guerini, and S. Tonelli. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, 2021.

H. Levesque, E. Davis, and L. Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

B. Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

Q. Lhoest, A. V. del Moral, Y. Jernite, A. Thakur, P. von Platen, S. Patil, J. Chaumond, M. Drame, J. Plu, L. Tunstall, et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, 2021.

E. Liscio, O. Araque, L. Gatti, I. Constantinescu, C. Jonker, K. Kalimeri, and P. K. Murukannaiah. What does a text classifier learn about morality? an explainable method for cross-domain comparison of moral rhetoric. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14113–14132, 2023.

X. Liu, K. Li, B. Han, M. Zhou, L. Jiang, Z. Xiong, and C. Huang. Semantic role labeling for news tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 698–706, 2010.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

A. Luccioni and J. Viviano. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.24. URL https://aclanthology.org/2021.acl-short.24.

L. Lucy, D. Tadimeti, and D. Bamman. Discovering differences in the representation of people using contextualized semantic axes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

K. Lueg and M. W. Lundholt. *Routledge handbook of counter-narratives*. Routledge, 2020.

D. Mahan, R. Carlow, L. Castricato, N. Cooper, and C. Laforte. Stable beluga models, 2023. URL https://huggingface.co/stabilityai/StableBeluga2.

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*, pages 55–60, 2014.

T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. In *Proceedings of NAACL-HLT*, pages 615–621, 2019.

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus

of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. URL https://aclanthology.org/J93-2004.

A. Maronikolakis, P. Baader, and H. Schütze. Analyzing hate speech data along racial, gender and intersectional axes. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 1–7, 2022.

L. Màrquez, X. Carreras, K. C. Litkowski, and S. Stevenson. Semantic role labeling: an introduction to the special issue, 2008.

S. C. Marsella and J. Gratch. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, 2009.

B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI conference on artificial intelligence*, 35(17):14867–14875, 2021.

C. May, A. Wang, S. Bordia, S. Bowman, and R. Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, 2019.

J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981.

J. P. McCrae, J. Bosque-Gil, J. Gracia, P. Buitelaar, and P. Cimiano. The ontolex-lemon model: development and applications. In *Proceedings of eLex 2017 conference*, pages 19–21, 2017.

C. Meghini, V. Bartalesi, and D. Metilli. Representing narratives in digital libraries: The narrative ontology. *Semantic Web*, 12(2):241–264, 2021.

M. Menéndez, J. Pardo, L. Pardo, and M. Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997.

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*, pages 77–80, 2017.

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*, pages 24–31, 2004.

P. Mikander et al. Westerners and others in finnish school textbooks. *University of Helsinki, Institute of Behavioural Sciences, Studies in Education*, 2016.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38 (11):39–41, 1995.

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)*, pages 4417–4422, 2016.

A.-L. M. Minard, M. Speranza, E. Agirre, I. Aldabe, M. Van Erp, B. Magnini, G. Rigau, and R. Urizar. Semeval-2015 task 4: Timeline: Cross-document event ordering. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 778–786, 2015.

T. Mitamura, Z. Liu, and E. H. Hovy. Overview of tac kbp 2015 event nugget track. In *TAC*, 2015a.

T. Mitamura, Y. Yamakawa, S. Holm, Z. Song, A. Bies, S. Kulick, and S. Strassel. Event nugget annotation: Processes and issues. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 66–76, 2015b.

M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer,

I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

I. Mollas, Z. Chrysopoulou, S. Karlos, and G. Tsoumakas. Ethos: an online hate speech detection dataset, 2020.

M. Mozafari, R. Farahbakhsh, and N. Crespi. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer, 2020.

M. Nadeem, A. Bethke, and S. Reddy. StereoSet: Measuring stereotypical bias in pretrained language models. In C. Zong, F. Xia, W. Li, and R. Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.416. URL https://aclanthology.org/2021.acl-long.416.

S. Nagel. Cc-news. *URL: http://web. archive. org/save/http://commoncrawl. org/2016/10/newsdatasetavailable*, 2016.

K. A. Nishikawa, T. L. Towner, R. A. Clawson, and E. N. Waltenburg. Interviewing the interviewers: Journalistic norms and racial diversity in the newsroom. *The Howard Journal of Communications*, 20(3):242–259, 2009.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, and B. Y. Zhao, editors, *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 145–153. ACM, 2016. URL https://doi.org/10.1145/2872427.2883062.

D. Nozza, C. Volpetti, and E. Fersini. Unintended bias in misogyny detection. In *Ieee/wic/acm international conference on web intelligence*, pages 149–155, 2019.

K. Ovchinnikova, I. Kononenko, and E. Sidorova. Development of lexico-syntactic ontology design patterns for information extraction of scientific data. In A. Pozanenko, S. A. Stupnikov, B. Thalheim, E. Méndez, and N. Kiselyova, editors, *Supplementary Proceedings of the XXIII International Conference on Data Analytics and Management in Data Intensive Domains, DAMDID/RCDL 2021 - Supplementary Proceedings, Moscow, Russia, October 26-29, 2021*, volume 3036 of *CEUR Workshop Proceedings*, pages 349–361. CEUR-WS.org, 2021. URL https://ceur-ws.org/Vol-3036/paper28.pdf.

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)*, pages 47–56, 2016.

M. Palmer. Semlink: Linking propbank, verbnet and framenet. In *Proceedings of the generative lexicon conference*, pages 9–15. GenLex-09, Pisa, Italy, 2009.

E. W. Pamungkas, V. Basile, and V. Patti. Misogyny detection in twitter: a multilingual and cross-domain study. *Inf. Process. Manag.*, 57(6):102360, 2020. doi: 10.1016/J.IPM.2020.102360. URL https://doi.org/10.1016/j.ipm.2020.102360.

E. W. Pamungkas, V. Basile, and V. Patti. Investigating the role of swear words in abusive language detection tasks. *Language Resources and Evaluation*, 57(1):155–188, 2023.

A. Panchenko, E. Ruppert, S. Faralli, S. Ponzetto, C. Biemann, et al. Building a web-scale dependency-parsed corpus from common crawl. In *LREC 2018-11th International Conference on Language Resources and Evaluation*, pages 1816–1823. European Language Resources Association (ELRA), 2019.

T. Parekh, I.-H. Hsu, K.-H. Huang, K.-W. Chang, and N. Peng. Geneva: Benchmarking generalizability for event argument extraction with hundreds of event types and

argument roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686, 2023.

J. H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, 2018.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In M. Walker, H. Ji, and A. Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202.

V. Petukhova, H. Bunt, et al. Lirics semantic role annotation: Design and evaluation of a set of data categories. In *LREC*. Citeseer, 2008.

J. Piskorski, V. Zavarella, M. Atkinson, and M. Verile. Timelines: Entity-centric event extraction from online news. In R. Campos, A. M. Jorge, A. Jatowt, and S. Bhatia, editors, *Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts co-located with 42nd European Conference on Information Retrieval, Text2Story@ECIR 2020, Lisbon, Portugal, April 14th, 2020 [online only]*, volume 2593 of *CEUR Workshop Proceedings*, pages 105–114. CEUR-WS.org, 2020. URL https://ceur-ws.org/Vol-2593/paper13.pdf.

F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489, 2021.

F. Poletto, V. Basile, C. Bosco, V. Patti, and M. Stranisci. Annotating hate speech: Three schemes at comparison. In R. Bernardi, R. Navigli, and G. Semeraro, editors, *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2019. URL https://ceur-ws.org/Vol-2481/paper56.pdf.

F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, 2021.

M. Pushkarna, A. Zaldivar, and O. Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.

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*, 3:28–34, 2003a.

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*, volume 2003, page 40. Lancaster, UK, 2003b.

J. Pustejovsky, A. Meyers, M. Palmer, and M. Poesio. Merging propbank, nombank, timebank, penn discourse treebank and coreference. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 5–12, 2005.

J. Pustejovsky, K. Lee, H. Bunt, and L. Romary. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397, 2010.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language Models are Unsupervised Multitask Learners. In *Semantic Scholar*, 2019. URL https://api.semanticscholar.org/CorpusID:160025533.

C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

F. Rahutomo, T. Kitasuka, and M. Aritsugi. Semantic cosine similarity. *The 7th international student conference on advanced science and technology ICAST*, 4(1):1, 2012.

S. Rajpurkar, D. Bhatt, P. Malhotra, M. Rajpurkar, and M. Bhatt. Book recommendation system. *International Journal for Innovative Research in Science & Technology*, 1(11):314–316, 2015.

J. W. Ratcliff, D. Metzener, et al. Pattern matching: The Gestalt approach. *Dr. Dobb's Journal*, 13(7):46, 1988.

M. Reinert, M. Schrott, B. Ebneth, and M. Rehbein. From biographies to data curation-the making of www. deutsche-biographie. de. In *BD*, pages 13–19, 2015.

I. J. Roseman. Appraisal determinants of discrete emotions. *Cognition & Emotion*, 5(3): 161–200, 1991.

I. J. Roseman. Appraisal in the emotion system: Coherence in strategies for coping. *Emotion Review*, 5(2):141–149, 2013.

I. J. Roseman and C. A. Smith. Appraisal theory. *Appraisal processes in emotion: Theory, methods, research*, pages 3–19, 2001.

R. Rudinger, J. Naradowsky, B. Leonard, and B. Van Durme. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, 2018.

I. Russo, T. Caselli, and M. Monachini. Extracting and visualising biographical events from wikipedia. In *BD*, pages 111–115, 2015.

L. Saeeda, M. Ledvinka, M. Blaško, and P. Křemen. Entity linking and lexico-semantic patterns for ontology learning. In *The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings 17*, pages 138–153. Springer, 2020.

K. Saha, E. Chandrasekharan, and M. De Choudhury. Prevalence and psychological effects of hateful speech in online college communities. In *Proc ACM Web Sci Conf*, pages 255–264, Jun 2019. doi: 10.1145/3292522.3326032.

E. W. Said. Orientalism. *The Georgia Review*, 31(1):162–206, 1977.

D. Sander, D. Grandjean, and K. R. Scherer. An appraisal-driven componential approach to the emotional brain. *Emotion Review*, 10(3):219–231, 2018.

R. Sanderson, P. Ciccarese, H. Van de Sompel, S. Bradshaw, D. Brickley, L. J. G. a Castro, T. Clark, T. Cole, P. Desenne, A. Gerber, et al. Open annotation data model. *W3C community draft*, 2013.

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.

S. Santy, J. Liang, R. Le Bras, K. Reinecke, and M. Sap. NLPositionality: Characterizing design biases of datasets and models. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.505. URL https://aclanthology.org/2023.acl-long.505.

M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, 2019.

M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 5477–5490, 2020.

R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky. Timeml annotation guidelines. *Version*, 1(1):31, 2006.

E. A. Schegloff and H. Sacks. Opening up closings. *Semiotica*, 8(4):289–327, 1973.

K. K. Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon.* University of Pennsylvania, 2005.

E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, 2019.

P. Shi and J. Lin. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*, 2019.

C. Shimizu, P. Hitzler, Q. Hirt, D. Rehberger, S. G. Estrecha, C. Foley, A. M. Sheill, W. Hawthorne, J. Mixter, E. Watrall, et al. The enslaved ontology: Peoples of the historic slave trade. *Journal of Web Semantics*, 63:100567, 2020.

R. A. Shweder, N. C. Much, M. Mahapatra, and L. Park. The "big three" of morality (autonomy, community and divinity), and the "big three" explanations of suffering. In A. Brandt and P. Rozin, editors, *Morality and health*, pages 119–169. Routledge, 1997.

D. Sileo, T. Van de Cruys, C. Pradel, and P. Muller. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3477–3486, 2019.

B. Smith, A. Kumar, and T. Bittner. *Basic formal ontology for bioinformatics.* IFOMIS Reports, 2005.

C. A. Smith and P. C. Ellsworth. Patterns of cognitive appraisal in emotion. *Journal of personality and social psychology*, 48(4):813, 1985.

L. Soldaini, R. Kinney, A. Bhagia, D. Schwenk, D. Atkinson, R. Authur, B. Bogin, K. Chandu, J. Dumas, Y. Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

Z. Song, A. Bies, S. Strassel, T. Riese, J. Mott, J. Ellis, J. Wright, S. Kulick, N. Ryant, and X. Ma. From light to rich ere: annotation of entities, relations, and events. In *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, 2015.

Z. Song, A. Bies, S. Strassel, J. Ellis, T. Mitamura, H. T. Dang, Y. Yamakawa, and S. Holm. Event nugget and event coreference annotation. In *Proceedings of the Fourth Workshop on Events*, pages 37–45, 2016.

G. C. Spivak. Can the subaltern speak? In *Colonial discourse and post-colonial theory*, pages 66–111. Routledge, 2015.

M. Stranisci, M. De Leonardis, C. Bosco, and V. Patti. The expression of moral values in the Twitter debate: a corpus of conversations. *IJCoL. Italian Journal of Computational Linguistics*, 7(7-1, 2):113–132, 2021a. URL http://journals.openedition.org/ijcol/880;DOI:https://doi.org/10.4000/ijcol.880.

M. A. Stranisci, V. Basile, R. Damiano, V. Patti, et al. Mapping biographical events to odps through lexico-semantic patterns. In *Proceedings of the 12th Workshop on Ontology Design and Patterns (WOP 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021)*, volume 3011 of *CEUR WORKSHOP PROCEEDINGS*, pages 1–12. CEUR-WS, 2021b. URL https://ceur-ws.org/Vol-3011/paper3.pdf.

M. A. Stranisci, V. Patti, R. Damiano, et al. Representing the under-represented: A dataset of post-colonial, and migrant writers. In D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, and B. Heinisch, editors, *3rd Conference on Language, Data and Knowledge, LDK 2021, September 1-3, 2021, Zaragoza,*

Spain, volume 93 of *OASIcs*, pages 7:1–7:14. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021c. URL https://doi.org/10.4230/OASIcs.LDK.2021.7.

M. A. Stranisci, S. Frenda, E. Ceccaldi, V. Basile, R. Damiano, and V. Patti. APPReddit: a corpus of Reddit posts annotated for appraisal. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3809–3818, Marseille, France, June 2022a. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.406.

M. A. Stranisci, S. Frenda, M. Lai, O. Araque, A. T. Cignarella, V. Basile, C. Bosco, and V. Patti. O-dang! the ontology of dangerous speech messages. In I. Kernerman, S. Carvalho, C. A. Iglesias, and R. Sprugnoli, editors, *Proceedings of the 2nd Workshop on Sentiment Analysis and Linguistic Linked Data*, pages 2–8, Marseille, France, June 2022b. European Language Resources Association. URL https://aclanthology.org/2022.salld-1.2.

M. A. Stranisci, E. Mensa, R. Damiano, D. Radicioni, and O. Diakite. Guidelines and a corpus for extracting biographical events. In H. Bunt, editor, *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 20–26, Marseille, France, June 2022c. European Language Resources Association. URL https://aclanthology.org/2022.isa-1.3.

M. A. Stranisci, E. Bernasconi, V. Patti, S. Ferilli, M. Ceriani, and R. Damiano. The world literature knowledge graph. In T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, and J. Li, editors, *The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part II*, volume 14266 of *Lecture Notes in Computer Science*, pages 435–452. Springer, 2023a. URL https://doi.org/10.1007/978-3-031-47243-5_24.

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 A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12370–12384, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.691. URL https://aclanthology.org/2023.acl-long.691.

M. A. Stranisci, V. Patti, and R. Damiano. User-generated world literatures: a comparison between two social networks of readers. In A. Falcon, S. Ferilli, A. Bardi, S. Marchesin, and D. Redavid, editors, *Proceedings of the 19th The Conference on Information and Research science Connecting to Digital and Library science, IRCDL 2023, Bari, Italy, February 23-24, 2023*, volume 3365 of *CEUR Workshop Proceedings*, pages 38–46. CEUR-WS.org, 2023c. URL https://ceur-ws.org/Vol-3365/short3.pdf.

M. C. Suárez-Figueroa, A. Gómez-Pérez, and B. Villazón-Terrazas. How to write and use the ontology requirements specification document. In *On the Move to Meaningful Internet Systems: OTM 2009: Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009, Vilamoura, Portugal, November 1-6, 2009, Proceedings, Part II*, pages 966–982. Springer, 2009.

M. C. Suárez-Figueroa, A. Gómez-Pérez, and M. Fernandez-Lopez. The neon methodology framework: A scenario-based methodology for ontology development. *Applied ontology*, 10(2):107–145, 2015.

M. Sui and N. Paul. Latino portrayals in local news media: Underrepresentation, negative stereotypes, and institutional predictors of coverage. *Journal of Intercultural Communication Research*, 46(3):273–294, 2017.

J. Sun and N. Peng. Men are elected, women are married: Events gender bias on wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Compu-*

tational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 350–360, 2021.

T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. Belding, K.-W. Chang, and W. Y. Wang. Mitigating gender bias in natural language processing: Literature review. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://aclanthology.org/P19-1159.

Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8968–8975. AAAI Press, 2020. URL https://doi.org/10.1609/aaai.v34i05.6428.

Y. Sun, S. Wang, S. Feng, S. Ding, C. Pang, J. Shang, J. Liu, X. Chen, Y. Zhao, Y. Lu, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*, 2021.

R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

B. B. Tillett. Frbr and cataloging for the future. *Cataloging & classification quarterly*, 39(3-4):197–205, 2005.

S. Tonelli, R. Sprugnoli, M. Speranza, and A.-L. Minard. Newsreader guidelines for annotation at document level nwr-2014-2-2, 2024. URL http://www.newsreader-project.eu/files/2014/12/NWR-2014-2-2.pdf.

C. Toraman, F. Şahinuç, and E. Yilmaz. Large-scale hate speech detection with cross-domain transfer. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2215–2225, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.238.

J. Trager, A. S. Ziabari, A. Mostafazadeh Davani, P. Golazazian, F. Karimi-Malekabadi, A. Omrani, Z. Li, B. Kennedy, N. K. Reimer, M. Reyes, et al. The moral foundations reddit corpus. *arXiv e-prints*, pages arXiv–2208, 2022.

T. H. Trinh and Q. V. Le. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*, 2018.

E. Troiano, S. Padó, and R. Klinger. Crowdsourcing and validating event-focused emotion corpora for german and english. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4005–4011, 2019.

E. Troiano, L. Oberländer, M. Wegge, and R. Klinger. x-envent: A corpus of event descriptions with experiencer-specific emotion and appraisal annotations. In *13th International Conference on Language Resources and Evaluation Conference, LREC 2022*, pages 1365–1375. European Language Resources Association (ELRA), 2022.

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.

L. Van der Beek, G. Bouma, R. Malouf, and G. Van Noord. The alpino dependency treebank. In *Computational linguistics in the Netherlands 2001*, pages 8–22. Brill, 2002.

M. van der Meer, P. Vossen, C. Jonker, and P. Murukannaiah. Do differences in values influence disagreements in online discussions? In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15986–16008, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.992. URL https://aclanthology.org/2023.emnlp-main.992.

C. Van Son, O. Inel, R. Morante, L. Aroyo, and P. Vossen. Resource interoperability

for sustainable benchmarking: The case of events. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

M. Verhagen, R. Gaizauskas, F. Schilder, M. Hepple, G. Katz, and J. Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 75–80, 2007.

B. Vidgen and L. Derczynski. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300, 2020.

M. Völske, M. Potthast, S. Syed, and B. Stein. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, 2017.

P. Vossen, R. Agerri, I. Aldabe, A. Cybulska, M. van Erp, A. Fokkens, E. Laparra, A.-L. Minard, A. P. Aprosio, G. Rigau, et al. Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*, 110:60–85, 2016.

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

T. Vu, B. Lester, N. Constant, R. Al-Rfou, and D. Cer. Spot: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, 2022.

C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Proceedings of the international AAAI*

conference on web and social media, pages 454–463. AAAI Press, 2015. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10585.

A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

K. Weathington and J. R. Brubaker. Queer identities, normative databases: Challenges to capturing queerness on wikidata. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–26, 2023.

J. Wei, X. Ren, X. Li, W. Huang, Y. Liao, Y. Wang, J. Lin, X. Jiang, X. Chen, and Q. Liu. Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*, 2019.

C. Welty, R. Fikes, and S. Makarios. A reusable ontology for fluents in owl. In *FOIS*, volume 150, pages 226–236, 2006.

M. Wiegand, R. Wilm, and K. Markert. Biographically relevant tweets – a new dataset, linguistic analysis and classification experiments. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3669–3679, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.323.

K. P. Winterich, Y. Zhang, and V. Mittal. How political identity and charity positioning increase donations: Insights from moral foundations theory. *International Journal of Research in Marketing*, 29(4):346–354, 2012.

A. Wolf. Minorities in us history textbooks, 1945–1985. *The Clearing House*, 65(5): 291–297, 1992.

L. E. Wood. Semi-structured interviewing for user-centered design. *interactions*, 4(2): 48–61, 1997.

J.-h. Yeh. Storyteller: An event-based story ontology composition system for biographical history. In *2017 International Conference on Applied System Innovation (ICASI)*, pages 1934–1937. IEEE, 2017.

W. Yin and A. Zubiaga. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598, 2021.

J. Zaller. Information, values, and opinion. *American Political Science Review*, 85(4): 1215–1237, 1991.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, 2019.

A. Zeldes. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612, 2017.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.

J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, 2018.

J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender bias in contextualized word embeddings. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1064. URL https://aclanthology.org/N19-1064.

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*, pages 19–27, 2015.

B. Zoph. Designing effective sparse expert models. In *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1044–1044. IEEE Computer Society, 2022.

Appendix

Appendix A: Data Documentation

In Table 8.1 all resources developed throughout the present PhD are listed. The table is intended to follow the Data Summary Section included in the Open Research Data Pilot (ORDP) template. Each resource is presented together with the link to its Github repository, information about its licensing, and about its backup on Zenodo. A brief summary of each resource is reported below.

Under-Represented Writers Ontology [Stranisci et al., 2021c]. This ontology has been developed to explore the underrepresentation of Transnational writers in public archives. Together with the Under-Represented Books Ontology it forms the Under-Represented Ontology Network.

Moral ConvITA [Stranisci et al., 2021a]. Moral ConvITA is a corpus of 861 adjacency pairs gathered from Twitter and annotated for moral values, according to the Moral Foundations Theory (MFT) [Graham and Haidt, 2012].

APPReddit [Stranisci et al., 2022a]. APPReddit is a corpus of 1,091 life events annotated according to Roseman [2013] theory of appraisal.

Bio-SRL [Stranisci et al., 2022c]. This is a corpus of 834 sentences annotated for biographical event detection with an annotation scheme that relies on two ISO standards for Semantic Annotation: SemAF [Bunt and Palmer, 2013] and ISO-TimeML [Pustejovsky et al., 2010].

Ontology of Dangerous Speech [Stranisci et al., 2022b]. The Ontology of Dangerous Speech has been developed to align and compare existing datasets annotated for Hate Speech and related phenomena.

Under-Represented Books Ontology [Stranisci et al., 2023c]. The ontology extends the encoding of the Under-Represented Writers Ontology to the representation of

Type	Resource	Release Date	Zenodo backup	License
Ontology	Under-Represented Writers	2021	x	CC0 1.0 Universal
Dataset	MoralConvITA	2021	x	CC0 1.0 Universal
Dataset	APPReddit	2022	x	CC0 1.0 Universal
Dataset	BioSRL	2022	x	CC0 1.0 Universal
Ontology	Ontology of Dangerous Speech	2022	x	CC0 1.0 Universal
Ontology	Under-Represented Books	2023	x	CC0 1.0 Universal
Dataset	WikiBio	2023	x	CC0 1.0 Universal
Knowledge Graph	World Literature KG	2023	x	CC0 1.0 Universal
Ontology	People in the Media	2024	x	CC0 1.0 Universal

Table 8.1: A summary of all resources created within the present PhD project.

works.

WikiBio [Stranisci et al., 2023b]. WikiBio is a corpus of 20 English Wikipedia biographies annotated for biographical event detection. The corpus has been adopted to train a model for the detection of biographical events.

World Literature Knowledge Graph [Stranisci et al., 2023a]. The World Literature Knowledge Graph is a knowledge base organized according to the Under-Represented Ontology Network and exposed through a visualization tool for the exploration of writers and their works.

People in the Media Ontology. People in the Media is an overarching ontology that has been developed to provide a comprehensive semantic encoding of the representation of people in different media.

Appendix B: Annotation Guidelines of Bio-SRL corpus

These guidelines are related to the annotation of a corpus of biographical events related to migrant writers. The annotation is organized on three levels:

- identification of 4 types of entities: the writer (or other indirect ways of referring to him), places, organizations and time expressions;
- identification of states, events and types of events (verbal and nominal)
- identification of semantic roles.

General Annotation Rules

The objective of this corpus is training a model able to recognize biographical events of migrant people and the semantic roles within these events. To meet this need the record must meet some general rules:

1. text chunks must always be consecutive;
2. the chunk itself cannot hold multiple roles within a record;
3. if there are several events in the same sentence, a separate annotation for each of them must be created

Annotation Procedure

Table 8.2 shows the six labels that must be used during the annotation task.

For every event in a sentence

- Make sure that the phrase mentions the writer in a direct (He won an award) or indirect way (His latest book won an award) and its semantic role. If it is mentioned, select the chunk that mentions it, otherwise discard the phrase.
- Identify the event or state expressed in the sentence

WRITER	the subject of the biography and its semantic role in the sentence (He wrote five novels)	[WRITER-AG, WRITER-PA]
EVENT	an event that involves a change of state (He moved to Paris)	[EVENT, ASP-EVENT, REP-EVENT]
STATE	a state relating to a person in a given period (he worked for Nike in 1999)	[STATE]
LOCATION	A semantic argument linked to the event and containing a LOCATION (He plays basketball in New York)	[LOC]
ORGANIZATION	A semantic argument linked to the event and containing an ORGANIZATION (He earned a PHD from the University of Turin)	[ORG]
TIME	A semantic argument linked to the event and containing a time expression (In 1995 he founded a school in Antananarivo)	[TIME]

Table 8.2: A list of the annotation labels for the Bio-SRL corpus

- if necessary, annotate the complementary part of the event or state, only when the verb needs it to express an event. For example, the verbs award, study, be born express a state or an event on their own. Instead, the verbs win, leave, obtain need a complementary part.
- if present, annotate the entities of type 'place' or 'organization'
- if present, annotate the time expression

Writers

The WRITER is the person mentioned in the biography, but not always the agent of the sentence, namely the person who performs the action. Two things must be considered:

1) make sure that the sentence concerns the writer and not other people. For example, the sentence below should not be annotated.

- Paramilitary groups such as the Red Shirts continued to suppress black voting in the Carolinas, especially in the upland counties.

2) Use the WRITER-AG label if you are talking about the author and the author is the agent of the event (it is the one who performs the action, for example)

- **He** [WRITER-AG] moved to Texas

In some cases the writer is AGENT of a passive sentence. In this case, if possible, annotate it together with the preposition

- The song "Mis Dos Mundos" was performed by Maiah Ocando, written **by Torrelles** [WRITER-AG],

3) Use the WRITER-PA label if the author is the patient of the event. In the first example the author is the subject of a passive sentence, in the second the object of an active sentence. In both cases it should be considered patient.

- **He** [WRITER-PA] was elected president by ONU
- Luka Took **Dirk** [WRITER-PA] to London

The WRITER is the person mentioned in the biography, but not always the agent of the sentence, namely the person who performs the action. Two things must be considered during the annotation:

4) The author may be mentioned directly, but also through his works or other references to him. These references should also be annotated, along with their possessive pronouns, if any.

- **his influence** [WRITER-AG] inspired a lot of young Nigerian novelists

If they are consecutive, select in a single chunk both the generic and specific reference to the work. Otherwise, give preference to generic references.

- In 2019, Simon Schuster published **her latest book , Silver, Sword, and Stone: Three Crucibles in the Latin American Story** [WRITER-PA] (Orion Publishers released it in the United Kingdom).

- **Steinbergs first two books** [WRITE-AG] Midlands (2002), about the murder of a white South African farmer, and The Number (2004), a biography of a prison gangster, won South Africa's premier non-fiction award, the Sunday Times Alan Paton Award.

If the work is mentioned with the pronoun you should not select it:

- In Italy, it won the Premio Gregor von Rezzori-Città di Firenze

5) If the writer is part of a collective group that is the protagonist of an event (e.g., his family, his group), annotate the mention of the group.

- At the age of nine, **her family** [WRITER-AG] moved to Ghana, where she continued her education at Akosombo International School.

Events and States

A first important distinction to make is that between events and states.

1) An EVENT is a verb or a nominal expression indicating a change of state:

- He **joined** [EVENT] Microsoft in 1999
- After his **Graduation** [EVENT] in 1999

2) A state is a condition relating to the writer in a given time period (which may or may not be specified):

- He **lives** [STATE] in Florida with his family
- He **lived** [STATE] in London from 1986 to 1989

In some cases, you can distinguish between events and states based on the type of time expression they are associated with. The first example should be considered an event, because it refers to a defined time period. The second, however, must be annotated as a state because it refers to a cyclical and lasting activity.

- He **participated** [EVENT] to a manifestation yesterday
- He **participated** [STATE] to the book club for 10 years

3) Not all verbs express an EVENT or a STATE. For example, many copulative verbs such as be and seem do not contain a meaning. In this case they should not be annotated and in their place you have to select the nominal part connected to them.

- He is **sad** [STATE] for his mom
- He seems **sad** [STATE] for his mom

5) If there are auxiliaries, annotate only the head of the verb

- He was **named** [EVENT] president
- He has never **finished** [EVENT] high school

6) Nominal events or states should be annotated only when they are not dependent on other verbs, as in the first example, where 'his childhood' should be annotated separately (agent+state).

- During his [WRITER-AG] **childhood** [STATE], he met William

Event Types

In addition to the general 'event' category there are two other types of events, which alone do not express a complete event, but give additional information.

1) ASP-EVENT. These events are informative about the temporal articulation of a given event. Examples of this are 'start' and 'stop'. When faced with these cases it is necessary to select the verb with aspectual value as ASP-EVENT and the related event as EVENT, if expressed by a verb

- He **started** [ASP-EVENT] working [EVENT] in Microsoft in 1999

In some cases an ASP-EVENT is combined with a STATE

- He **continues** [ASP-EVENT] living [STATE] in Florida with his family

2) REP-EVENT. These events indicate reported speeches. They should also be annotated together with an event or related state, if it is expressed by a verb.

- John **declared** [REP-EVENT] he will run [EVENT] for the election.
- Julia **told** [REP-EVENT] me that she hates [STATE] Harry Potter saga

Attention, the REP-EVENT can be combined with another event or state only if it has the same subject, as in the case below:

- Less than a week before her death, the University of Notre Dame **announced** [REP-EVENT] that it would award [EVENT] Bowman the 1990 Laetare Medal. (Thea Bowman)

When the subject is different, namely the author is WRITER-PA of the REP-EVENT and WRITER-AG of the event, the two events must be annotated in separate annotations.

- The Prime Minister of Sri Lanka **stated** [REP-EVENT] that Shahi disappeared in 2001 and sightings of him were thereafter reported around the world. (Riaz Ahmed Gohar Shahi)
- The Prime Minister of Sri Lanka stated that Shahi disappeared [EVENT] in 2001 and sightings of him were thereafter reported around the world. (Riaz Ahmed Gohar Shahi)

Locations and Organizations

Entities of type place and organization must be annotated with LOC and ORG labels. In this case, however, these must be annotated as if they were an entire semantic argument

of any kind that contains a LOC or an ORG. For each event, a maximum of one LOC and one ORG can be annotated.

1) Select the entity together with the proposition, if it is present

- Orion Publishers [ORG] released it **in the United Kingdom** [LOC]

2) When multiple organizations are associated with the same event or state, annotate them in a single chunk if they are contiguous.

- Neal started modeling for Shiseido and Pond's [ORG] skin-care ads.
- He spent his childhood traveling **from Portugal to France to Belgium and Morocco** [LOC].

3) If the organizations are not contiguous, annotate only the first

- He worked for Schindler [ORG], an elevators' factory, and for Adidas

4) If the name of an organization contains the name of a place, such as the University of Turin, consider the whole of chunk as 'organization'

- He went on to study political science **at Baghdad University** [ORG] and international law **at Vienna University** [ORG].

5) When an organization is mentioned, and within the mention there is a place, as in the case of the Saxon genitives, select all the Chunk as the ORG.

- Her album was released **by Brussels based Belgian record company Fonti Musicali**[ORG].

6) In a sentence with place adverbs or relative pronouns linking two events in the same place or organization, select the place expressed in the previous event and the preposition introducing it. In this case, both 'grew up' and 'childhood' are compatible with the 'in' preposition,

- She grew up **in Westphalia** [LOC], where she later said that her childhood [STATE] was unhappy.

7) If in a sentence with multiple events linked by an adverb of place or relative pronoun the preposition is not compatible to select place but not the preposition. In this case 'moved' and 'presents' hold different prepositions.

- Later, Neubarth moved to **Globonews** [ORG], the Cable TV channel from Rede Globo, where she currently presents [STATE] the news at 18:00

Temporal Expressions

Temporal expressions shall be annotated if they meet two conditions:

1) If they relate directly to the event

- **In 1999** [TEMP] she left Nigeria
- **At the age of 19** [TEMP] she left Nigeria

2) If they indicate a well-defined time or time frame. The first case indicates a point in time; the second, instead, indicates a time period with a beginning and an end: both must be noted. The word 'after' in the third example should not be annotated because it indicates a period of time from the generic boundaries.

- **Yesterday** [TIME] she left Nigeria
- **For two years** [TIME] she worked at Unipol
- After leaving Nigeria, she worked as a teacher.

Appendix C: Annotation Guidelines of WikiBio corpus

This document describes the annotation guidelines for the entity-based event extraction task. Given a text, which can be a sentence, a paragraph or a full document, annotators must first identify all the events related to a target entity and then: (i) Label the single word triggering the event; (ii) Label the word or group of words triggering the entities.

As it can be observed in the example (a), only events that are directly related to the entity must be annotated. Thereby, the pair <Woods, BORN> is a proper annotation, while the pair <his family, LIVE> is not, since the target-entity has not a direct role in the event LIVE.

(a) Woods [TARGET-ENTITY] was born [EVENT] at Hobeni , Transkei , where his family had lived for five generations.

The Corpus

The corpus of documents to be annotated is a set of 20 Wikipedia biographies of writers born in Africa or African American writers. The complete list of writers in the following:

Donald Woods, Ada Aharoni, Lewis Nkosi, Ngũgĩ wa Thiong'o, Angela Davis, Nasr Hamid Abu Zayd, Wole Soyinka, Bessie Amelia Emery Head, Ken Saro-Wiwa, Ali Al'amin Mazrui,Abdel-Tawab Youssef Ahmed Youssef,Alice Malsenior Tallulah - Kate Walker, Jayne Cortez, Chloe Anthony Wofford Morrison, Etheridge Knight, Maya Angelou, Amiri Baraka, John Edgar Wideman, Dwight D. York, Ishmael Scott Reed

Entities

Our guidelines for the annotation of target entities start from the Co-reference Guidelines for English Ontonotes, introducing some simplifications and variations.

Simplifications

- Instead of annotating the mentions of all the entities, annotators are asked to label only mentions about the target entity, namely the subject of the biography.

- Annotators must not annotate appositive coreference as it can be observed in example (b).

(b) Luke [TARGET-ENTITY], a writer from Somalia, was born in 1973

Variations

Mentions without a role. Mentions must be annotated only if they result in a direct role of the target entity in the event. SO events in which the target entity is mentioned for its relation with another entity who has a role must not be annotated, as it can be observed in (c)

(c) His father was the Chief Kadhi of Kenya,

Metonymical mentions in biographical events. There are some cases in which events refer to a target entity without directly mentioning it, as for instance in cases where the book of an author is awarded, translated or published. Such a relation may be considered as metonymic, since the entity that receives a prize is the author and not the book. Traditional coreference resolution guidelines ask annotators to consider such cases as mentions of the book, though. Instead in our guidelines annotators must consider these as metonymic mentions of the target-entity whenever the event is biographical, as in (d). If however the mention of the book is not related to a biographical event as in (e), annotators must not annotate it.

(d) In 1975 , Morrison's second novel Sula [TARGET-ENTITY] (1973) , about a friendship between two black women , was nominated [EVENT] for the National Book Award .

(e) The book talks about the relationship between a journalist and his dog.

"Part of" mentions. A last variation from the OntoNotes coreference guidelines refers to the "part of" relation between the target-entity and a group it is part of. Unlike traditional coreference guidelines, ours require annotators to label such mentions. As it can be observed in (f), the pronoun "they" is marked as a mention of the entity target,

since it is involved in the event. However, whenever it is possible to distinguish the target entity, annotators must annotate only it, as in (g), where "and his wife" was not marked. Such a type of mention is also applied to groups (h), but not to organizations.

(f) He [TARGET-ENTITY] exhibited with his wife and they [TARGET-ENTITY] received enthusiastic reaction

(g) He [TARGET-ENTITY] and his wife won the Nobel Prize.

(h) He [TARGET-ENTITY] founded Nirvana. The group [TARGET-ENTITY] toured Europe in 1992.

(j) He [TARGET-ENTITY] founded Apple. Apple launched IPhone in 2007.

Events

Our definition of events derives from TimeML. "We consider "events" a cover term for situations that happen or occur. Events can be punctual (1-2) or last for a period of time (3-4). We also consider as events those predicates describing states or circumstances in which something obtains or holds true (5)." Annotators must think of events as something that occurred in a certain moment or for a certain period of time within the life of the entity.

In order to correctly detect events there are four aspects you must pay attention:

One event == one token . When you are annotating an event you must try to always annotate only one token. This means that you must not annotate auxiliaries (a), prepositions for phrasal verbs (b), and other words that form a MWE (c).

(a) He has been awarded

(b) He grew up in Ogidi

(c) He is the US President since 1999

Events may be expressed by several part of speech. Even if they are more frequently expressed by verbs, EVENTS may be also expressed by other parts of speech, such as names, adjectives, and pronouns. In this task annotators must annotate events

regardless of their part of speech, as it can be observed in (d), (e), and (f).

(d) He won [EVENT] the Nobel Prize

(e) He has been professor [EVENT]at Berkeley for 5 years

(f) He was really sad [EVENT] yesterday

and copular verbs. Not all verbs trigger events. There are in fact several verbs that do not express events, such as copular verbs and lexical items participating in light verb constructions. These verbs are often semantically void, but may have a role in specializing the semantic of an event, for instance providing information about aspectuality. In our guidelines we ask annotators to pay attention to the following verbs that may be light or copular:

- be, become, seem, have, do, make, get, come, put

If it is so, annotators must label them as REL (Bonial, Palmer, 2016) and link them to the event they refer to, as it can be observed in (g) and (h). However, these verbs may also express an event alone, asi in (i) and (j).

(g) He make [REL] a speech [EVENT]

(h) He get [REL] a scholarship [EVENT]

(i) They were [EVENT] in Greece for 6 weeks

(j) He made [EVENT] a cake.

Annotating uncertainty. Uncertainty is a crucial aspect in annotating events, since it may affect time reasoning, and it is crucial to the domain we are investigating. If a person "tries to be elected in Parliament", it is important to label the event "elected", but at the same time to mark the uncertainty of such an event. Annotators are asked to:

1. identify in the text events or other linguistic items that express uncertainty;
2. label them as EVENT if they are events or EVENT_MOD if they are not;
3. link them to the event that they are related to.

There are three types of uncertainty links:

1. INTENTION: if the event represents the intention of an agent (k);

2. NOT_HAPPENED: the event did not happen (l);

3. EPISTEMIC: all the other cases. In particular events related to opinions and hypothetical events (m).

(k) the government was trying [EVENT] to have him killed [EVENT]. <trying, killed, INTENTION>

(l) was not [EVENT-MOD] allowed to speak [EVENT] publicly.
<not, speak, NOT_HAPPENED>

(m) Dr Mamphela Ramphele , berated [EVENT] him for writing [EVENT] misleading stories about the movement <berated, writing, EPISTEMIC>

APPENDIX D: Structure of the Interview adopted for the WL-KG Usability Testing

How to use the platform

- Login to the platform with the credentials that have been sent to you by mail (you can only browse it through a desktop computer): https://purl.archive.org/wl-kg
- Search for the entities you are looking for (e.g., a writer, a work, etc.) in the box on the left.
- When you find it, drag and drop the entity in the central area (empty at the beginning). In order to explore information about the entity, you can click twice on it or click on the arrow on its right.

Resource evaluation: writers

Search information about a writer belonging to an ethnic minority or born in a non-Western former colony. Explore their connection with other entities in the Knowledge Graph (e.g., authors, subjects, places)

1. Please, evaluate your search experience according to the following dimensions (select a value between 1 and 4):
    - Is the resource complete?
    - Are the information stored in the resource correct?
    - Is the resource easy to navigate?
2. Did you find any missing information? If so, please list them
3. Did you find errors? If so, please list them

Resource evaluation: works

Search a work written by a non-Western writer or by a writer belonging to an ethnic minority. Explore its connection with other entities in the Knowledge Graph (e.g., authors, subjects, places)

1. Please, evaluate your search experience according to the following dimensions (select a value between 1 and 4):
    - Is the resource complete?
    - Are the information stored in the resource correct?
    - Is the resource easy to navigate?
2. Did you find any missing information? If so, please list them
3. Did you find errors? If so, please list them

Comparison with other platforms

1. What platforms do you use for searching literary information?
    - Google Books
    - Goodreads
    - Open Library
    - VIAF
    - Wikipedia
    - Other (please specify)
2. Compared to other resources, how would you rate the World Literature Knowledge Graph for the task of discovering non-Western writers (select a value between 1 and 4)?
    - Compared to other resources, how would you rate the World Literature Knowledge Graph for the task of discovering non-Western writers?

- Compared to other resources, how would you rate the World Literature Knowledge Graph for the task of discovering new works?

3. According to your experience, what kind of tasks could be supported by this resource?

   - Discovering of new authors and works
   - Designing educational activities
   - Supporting research activities
   - Other (please specify)

4. To your knowledge, are there other sources of knowledge that can be integrated in this resource?