Polifonia: a digital harmoniser for musical heritage knowledge, H2020

D4.2 : Interrogation and annotation of plurilingual corpora for discourse analysis (V1.0)

Deliverable information	
WP4	
Deliverable dissemination level	PU Public
Deliverable type	Other, software and report
Lead beneficiary	UNIBO, KNAW, OU
Document status	Final
Document version	V1.0
Date	June 30, 2022
Authors	Rocco Tripodi, UNIBO Eleonora Marzi, UNIBO Arianna Graciotti, UNIBO Valeria Zotti, UNIBO Antonella Luporini, UNIBO Monica Turci, UNIBO Ana Pano Alaman, UNIBO Peter Van Kranenburg, KNAW Andrea Scharnhorst, KNAW René Van Horik, KNAW Enrico Daga, OU Marilena Daquino - UNIBO
Peer review	Marilena Daquino - UNIBO Andrea Scharnhorst - KNAW

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004746

PAGE INTENTIONALLY BLANK

Project Information

Project Start Date: 1st January 2021
Project Duration: 40 months
Project Website: https://polifonia-project.eu

Project Contacts

Project Coordinator

Valentina Presutti

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA
Department of Language, Literature and Modern Cultures (LILEC)

E-mail: valentina.presutti@unibo.it

Project Manager

Marta Clementi

ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA
Research division

E-mail: marta.clementi3@unibo.it

POLIFONIA Consortium

No.	Short name	Institution name	Country
1	UNIBO	ALMA MATER STUDIORUM - UNIVERSITÀ DI BOLOGNA	Italy
2	OU	THE OPEN UNIVERSITY	United Kingdom
3	KCL	KING'S COLLEGE LONDON	United Kingdom
4	NUI GALWAY	NATIONAL UNIVERSITY OF IRELAND GALWAY	Ireland
5	MiC	MINISTERIO DELLA CULTURA	Italy
6	CNRS	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS	France
	SORBONNE	SORBONNE UNIVERSITE (LinkedTP)	France
7	CNAM	CONSERVATOIRE NATIONAL DES ARTS ET METIERS	France
8	NISV	STICHTING NEDERLANDS INSTITUUT VOOR BEELD EN GELUID	Netherlands
9	KNAW	KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETEN-SCHAPPEN	Netherlands
10	DP	DIGITAL PATHS	Italy

Project Summary

European musical heritage is a dynamic historical flow of experiences, leaving heterogeneous traces that are difficult to capture, connect, access, interpret, and valorise. Computing technologies have the potential to shed a light on this wealth of resources by extracting, materialising and linking new knowledge from heterogeneous sources, hence revealing facts and experiences from hidden voices of the past. Polifonia makes this happen by building novel ways of inspecting, representing, and interacting with digital content. Memory institutions, scholars, and citizens will be able to navigate, explore, and discover multiple perspectives and stories about European Musical Heritage.

Polifonia focuses on European Musical Heritage, intended as musical contents and artefacts - or music objects - (tunes, scores, melodies, notations, etc.) along with relevant knowledge about them such as: their links to tangible objects (theatres, conservatoires, churches, etc.), their cultural and historical contexts, opinions and stories told by people having diverse social and artistic roles (scholars, writers, students, intellectuals, musicians, politicians, journalists, etc), and facts expressed in different styles and disciplines (memoire, reportage, news, biographies, reviews), different languages (English, Italian, French, Spanish, and German), and across centuries.

The overall goal of the project is to realise an ecosystem of computational methods and tools supporting discovery, extraction, encoding, interlinking, classification, exploration of, and access to, musical heritage knowledge on the Web. An equally important objective is to demonstrate that these tools improve the state of the art of Social Science and Humanities (SSH) methodologies. Hence their development is guided by, and continuously intertwined with, experiments and validations performed in real-world settings, identified by musical heritage stakeholders (both belonging to the Consortium and external supporters) such as cultural institutes and collection owners, historians of music, anthropologists and ethnomusicologists, linguists, etc.

Executive Summary

The deliverable reports on the annotation and interrogation of the Polifonia Textual Corpus, the plurilingual diachronic corpus focused on Musical Heritage (MH) covering Italian, English, French, Spanish and Dutch. Natural Language Processing (NLP) techniques were used to process the corpus and produce automatic morphosyntactic, semantic and MH-specific annotations. Custom APIs have been developed and released to enable domain experts, scholars and music professionals to leverage the annotations produced to perform advanced structured queries on the corpus. The available interrogation capabilities overcome the basic keyword-based search, offering the possibility of querying the corpus by taking advantage of the advanced semantic and MH-specific information encoded in the annotation.

Document History

Version	Release date	Summary of changes	Author(s) - Institution
V0.1	30/03/2022	Outline released	Rocco Tripodi - UNIBO
V0.2	20/05/2022	First review draft	Rocco Tripodi, Eleonora Marzi, Arianna Graciotti, Ana Pano Alaman, Valeria Zotti - UNIBO Enrico Daga - OU Peter Van Kranenburg - KNAW
V0.3	31/05/2022	New draft during the review	Rocco Tripodi, Eleonora Marzi, Arianna Graciotti
V0.4	09/06/2022	Post-review feedback absorption	Rocco Tripodi, Eleonora Marzi, Arianna Graciotti, Andrea Scharnhorst, René Van Horik, Ana Pano Alaman, Valeria Zotti, Antonella Luporini, Monica Turci
V1.0	30/06/2022	Final version submitted to EU	UNIBO

Table of contents

1 Introduction

This is the second report describing the work developed within Work Package 4 (WP4). This Work Package (WP) aims to provide the project and its pilots with methods and tools to extract Musical Heritage (MH) knowledge from text. To accomplish this objective, an initial focus has been dedicated to building and evaluating a multilingual corpus on MH (Task 4.1), the Polifonia Textual Corpus [1]. This task breaks down into two subsequent parts: (i) building a multilingual textual corpus, which first release was delivered at M10 and reported within the report "D4.1: Plurilingual corpora containing source texts in English, French, Spanish and German (v1.0)" [2], (ii) annotating it and supporting its interrogation through custom APIs - which is the focus of this report.

The work completed within this Deliverable and described in this report contributes to the Polifonia project's general objective of preserving MH by tailoring the construction of a plurilingual corpus of MH-relevant texts enriched with linguistic annotations. As one of the main goals of the Polifonia project is to construct a knowledge graph about music, the The Polifonia Textual Corpus [1] will be linked to the Polifonia's Knowledge Graph to indicate the provenance of the represented facts.

1.1 The Polifonia Textual Corpus and Lexicon

The Polifonia Textual Corpus [1] data, metadata and statistics, along with its annotations and interrogation tools are part of the Polifonia Ecosystem[1] and oblige the Ecosystem's rulebook[2]. They are released through the dedicated Polifonia Corpus GitHub repository[3]. The corpus is released under CC BY license as a set of metadata that allows the reproduction of the whole corpus. We remark that the texts included in the corpus are not published in their integral form because they are subject to heterogeneous licensing.

As Deliverable 4.1 [2] details, the initial intention of producing a textual corpus on MH of around 1 million words for five languages (English, French, German, Italian and Spanish) underwent a substantial adjustment during the unfolding of the project. The large modularized corpus developed overcame the original expectations per dimension (it eventually contains more than 100 million words for each language) and per language coverage (it also includes Dutch). A significant part of the sources of the corpus was only available as images or pdf files. We leveraged Optical Character Recognition (OCR) [3], state-of-the-art technology that addresses this problem, to convert them in a processable format.

Together with the corpus, WP4 also developed a specialized lexicon, extensively described in D4.1 [2]. The Polifonia Lexicon[4] is a linguistic resource representing the MH-specific terminology and concepts in six languages, organized in the style of WordNet [4], as sense-equivalent classes called synsets, each associated with its lexicalizations.

1.2 Annotation and Interrogation of the Polifonia Textual Corpus

This report illustrates the output and outcomes of the work executed to perform the annotation and allow the interrogation of all the modules (Encyclopedic, Books, Periodicals and Pilots) and languages (Dutch, English, French, German, Italian and Spanish) of the Polifonia Textual Corpus [1].

The annotation of this corpus and the development of tools for its interrogation are motivated by the importance of supporting MH scholars, researchers and professionals by providing them with resources they can query to disclose

[1] https://polifonia-project.github.io/ecosystem/
[2] https://polifonia-project.github.io/ecosystem/rulebook/README.html
[3] https://github.com/polifonia-project/Polifonia-Corpus
[4] https://github.com/polifonia-project/Polifonia-Lexicon

novel connections and shed light on lesser-known aspects of MH. Also, it provides the project (especially WP4) with a rich, representative and large body of MH-relevant texts and machine-readable linguistic annotations to favour the application of computational models for extracting facts about MH.

The Polifonia Textual Corpus [1] is an annotated corpus of an unprecedented scale for languages and periods covered. To the best of our knowledge, it is the only MH-related domain-specific corpus which can be interrogated through advanced semantic and MH-specific annotations.

1.3 Contribution and Structure of the Document

The main contribution of this report are:

- Larger and more robust release of data, metadata and statistics of each module of the Polifonia Textual Corpus;
- Advanced morphosyntactic, semantic and MH-specific annotations of each module of the Polifonia Textual Corpus;
- Custom APIs to allow the interrogation of the corpus.

This report is organized as follows. In Section 2.1, we outline the relevant related work, focusing on the annotation layers and interrogation capabilities offered by domain-specific, music-related and sense-annotated corpora. We highlight that, unlike the related work mentioned, the Polifonia Textual Corpus is the only one that can be interrogated through semantic and MH-specific queries. In Section 3.1, we build up to what is described in [2] by sharing updated data, metadata and statistics for each module of the Polifonia Textual Corpus. Where relevant, we account for any expansion of the corpus materials or methodological novelties introduced within the work done for this Deliverable. Then, we enter the core of our corpus annotation endeavour: in Section 4.2, we present the data format that we chose to encode the annotations performed. We give, as an example, a fully annotated sentence (cf. Subsection 4.2.1) extrapolated from the English language Encyclopedic Module (cf. Subsection 3.1.1). In Subsection 4.3.3, we provide an in-depth account of each layer of annotations produced by our NLP pipeline (cf. Table 4.2). For each annotation layer, we introduce the task performed, give an example of its realization taken from the fully annotated example sentence and indicate the model(s) used to perform the task providing references and a concise technical description. Finally, we report the links for downloading the annotations of each language and module of the corpus (cf. 4.3.5). In Chapter 5, we focus on the principal aspects of the work carried out to allow the interrogation of the corpus. In Section 5.1, we focus on the process we followed to elicit, collect and discuss requirements for designing and implementing the corpus interrogation APIs. We give an account of the meetings held with domain experts in the field of corpus linguistics and members of the WP in charge of implementing approaches and interfaces for a successful interaction of humans and MH (WP5). In Section 5.2, we describe how the released APIs can be used to perform advanced queries on the corpus, giving instructions and reporting a few examples of the different queries that an end-user can perform. Chapter 7) encompasses the conclusions of this report, along with the discussion of how we intend to expand on the work done so far.

2 Background and Related Work

2.1 Annotated Polifonia Textual Corpus

2.1.1 Context

The process of building the Polifonia Textual Corpus [1] was extensively described in Deliverable 4.1 [2] submitted at Milestone 10. In particular, Chapter 4 of [2] addresses all the details of the Polifonia Textual Corpus' composition (sub-sections 4.1-4.4), including methodological details and FAIRness and reproducibility issues (sub-section 4.5). Section 2.4 of Deliverable 4.1 [2] outlines the state-of-the-art of music textual corpora, highlighting the advancements brought by Polifonia Textual Corpus [1], thanks to its novel combination of comparable plurilingual, contemporary and historical sources.

The release of the annotated Polifonia Textual Corpus [1], which this report describes in detail (cf. Chapter 4), constitutes a substantial contribution to the offer of domain-specific cultural heritage annotated corpora, especially of musical heritage (MH). Several corpora are available in the field of Music Information Retrieval, as extensively surveyed in Section 2.4 of [2]. However, no multilingual, diachronic corpora concerning MH are currently available in a digitised machine-readable format that permits interrogations based on a wide range of automatic annotation types ranging from morphosyntactic to semantic levels of language analysis, as enabled by state-of-the-art NLP techniques. Morphosyntactic annotations such as tokenisation, lemmatisation and part-of-speech tagging are commonly made available both by general corpora, whose main aim is to sample a language or a linguistic variety, and by domain-specific corpora. Semantic annotations are less frequently provided by both types of corpora. The annotated Polifonia Textual Corpus [1] makes both morphosyntactic, semantic and MH-specific annotations available to its end-users for interrogation via custom APIs (cf. Chapter 5).

2.1.2 Morphosyntax-annotated Corpora

Among the morphosyntax-annotated general corpora, worth-mentioning examples are corpora belonging to the TenTen Corpus Family [1] [5], available in forty-two languages. Their content is crawled from the Web and undergoes tokenisation, lemmatisation, and part-of-speech tagging. Therefore, end-users can perform queries leveraging information related to the morphosyntactic level of analysis of a language.

Equally, domain-specific corpora provide the end-users with layers of morphosyntactic annotations that can be used to perform queries. A pertinent example is a corpus collected and curated within the project *LBC - Lessico dei Beni Culturali* [2]. The corpus is available in Italian, French, Spanish, German, English and Russian language. It contains texts of diverse genres and topics, encompassing art history books, technical documents, travel accounts, and tour guides [6]. Its texts underwent tokenisation, lemmatisation, and part-of-speech tagging thanks to the functionalities offered by TreeTagger[3] [7], a language-independent Part-of-Speech tagger [8]. Another relevant initiative in the scope of corpora creation and curation aimed at cultural heritage promotion and preservation is Corpus UNICittà [4]. This multilingual corpus collects and digitises textual testimonies about the University of Bologna's cultural heritage, published in different eras and belonging to different discursive typologies. Corpus UNICittà provides the end-users with the possibility of browsing the corpus leveraging metadata produced by manual annotations carried out with the aid of the software ATLAS.ti [5]. These manual annotations and the resulting metadata enrich the corpus documents

[1] https://www.sketchengine.eu/documentation/tenten-corpora/
[2] http://corpora.lessicobeniculturali.net/it/
[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[4] https://corpusunicitta.it/
[5] https://atlasti.com/

with labels describing their content, source of origin, and relevant material heritage entities (such as architectonically relevant buildings and art pieces). Other research in the music domain brought to the production of corpora specific to MH. However, most of these studies focus on particular problems and comprise a limited number of selected annotations which are instrumental to the ultimate goal of the research. For example, parallel corpora of music and lyrics have been collected and enriched with semantic information such as crowd-sourced emotion annotations [9]. The construction of monolingual music corpora enabled experiments in topic modelling. An example of this is the Russian text corpus on musicology [10], which covers texts of music encyclopedias, reference books and monographs, ensuring a broad representation of domains such as music, history of music, musical instruments, artists' biographies. In other cases, corpora creation has been functional to the study of the production of a single author through corpus linguistics methodology, as Wagner's librettos [11].

2.1.3 Sense-annotated Corpora

Large sense-annotated corpora are not widespread, as the collection of semantic annotations is an expensive and time-consuming task. However, their demand has increased due to the rising necessity for training deep supervised systems in the past few years. As the survey at [12] summarises, the sense-annotated corpora currently available fall into different categories depending on the underlying resource from which they draw the senses to assign to words in context. The principal resource focused on English for building such corpora is Wordnet [4], while the main multilingual ones are BabelNet[6] [13] and Wikipedia[7]. As the survey reports, sense-annotated corpora can be constructed following a manual approach, as is the case for SemCor [14], one of the first attempts in this sense. Worth mentioning, for its significant number of annotations, is the MASC Word Sense Annotation (MASC-WSA) corpus [15], which is a sense-annotated excerpt of the Manually annotated Sub-Corpus of American English. Both corpora mentioned leverage WordNet as a sense inventory. Semi-automatic strategies are also adopted, such as in the Semantically Enriched Wikipedia (SEW) [8] corpus, which was annotated by propagating information from Wikipedia hyperlinks to Wikipedia pages [16]. Automatic annotation strategies are also applied, as in the case of SenseDefs[9], which builds up to the Princeton WordNet Gloss Corpus project [10] by automatically applying word sense disambiguation to definitions in 263 languages [17]. Research in the field of multilingual entity linking has also offered a proficuous occasion to produce automatically annotated corpora, as is the case for Mewsli-9 (Multilingual Entities in News, linked) [18]. This dataset, whose annotations address entity mentions and links them to WikiData, builds up to the English-only WikiNews-2018 dataset [19] dramatically scaling up in language coverage, including documents in nine languages (Japanese, German, Spanish, Arabic, Serbian, Turkish, Persian, Tamil, English). In addition, work in the space of multilingual word-sense disambiguation led to the production of sense-annotated datasets as a by-product, as in the case of MULTIMIRROR [20], whose results encompassed the creation of sense-annotated datasets for word-alignment between English and one target language between French, German, Italian and Spanish.

2.1.4 The Polifonia Textual Corpus Novelty and Impact

The annotated Polifonia Textual Corpus [1] enriches the scenario of morphosyntax-annotated, sense-annotated and domain-specific textual corpora described in the paragraphs above. It constitutes a remarkable resource for MH, both from the perspective of the variety of periods, languages and cultures covered and for the depth of layers of annotations offered to the end-users to perform advanced semantic queries that overcome keyword-based and morphosyntax-focused searches. In fact, the annotated Polifonia Textual Corpus [1] content is enriched with information derived from automatic annotations that range from NLP foundational tasks centred on morphosyntax (tokenisation, lemmatisation, part-of-speech tagging) to the most advanced NLU analysis (word sense disambiguation, named entity recognition, entity linking). The MH-specific annotation layer, based on the Polifonia Lexicon

[6] https://babelnet.org/
[7] https://www.wikipedia.org/
[8] http://lcl.uniroma1.it/sew/
[9] http://lcl.uniroma1.it/disambiguated-glosses/
[10] http://wordnet.princeton.edu/glosstag.shtml

Tagging (cf. 4.3.4), completes the picture by offering an additional level of specialist information based on the MH concepts contained in the Polifonia Lexicon (cf. [2], Chapter 3). To the best of our knowledge, the annotated Polifonia Textual Corpus [1] is the sole domain-specific corpus dedicated to MH providing the end-users with the possibility to go beyond the information offered by morphosyntactic annotation and to access structured semantic knowledge such as disambiguated word senses, named entities linked to Wikipedia as a knowledge base, and MH-specific information based on a specialised lexicon.

3 The Polifonia Textual Corpus

The Polifonia Textual Corpus [1] is a comparable, multilingual, diachronic, domain-specific corpus on Musical Heritage. It is designed to support different purposes:

- Designing the approach used for APIs development;
- Ensuring the representativeness of the MH discourse in terms of languages (Dutch, English, French, German, Italian and Spanish) and time periods (from 1400 to these days);
- Supporting the development of tools for knowledge extraction from texts;
- Serving case studies applications as expressed by the Polifonia Pilots.

The Polifonia Textual Corpus [1] is at the same time representative of discourse on Musical Heritage in different languages through an historical perspectives, and also specific to single case studies represented by the Pilots. Heterogeneous requirements lead the corpus to be organized into four modules: The Encyclopedic Module, The Book Module, The Periodical Module and The Pilots Module. Each module has its features and development methodologies extensively documented in Deliverable [2].

A significant part of the sources collected for the Polifonia Textual Corpus [1] were available as images or pdf files. As reported in [2], within the work done for Deliverable 4.1 we converted them in a processable format through a custom piece of software released in the Polifonia Textual Corpus Population[1] GitHub repository. This piece of software is based on Tesseract [21] and leverages OCR technology [3] to specifically address the challenges raised by multilingual and historic sources.

[1] https://github.com/polifonia-project/textual-corpus-population

3.1 The Polifonia Textual Corpus: overview

3.1.1 The Encyclopedic Module

General aspects of MH, such as the description of musical instruments or musical genres, are covered by the Encyclopedic module. This was developed by selecting documents from Wikipedia. This part of the corpus allows having encyclopedic and up-to-date information about music concepts.

3.1.1.1 Metadata

We make available the Metadata related to the Wikipedia pages that constitute the Encyclopedic Module of the Polifonia Textual Corpus [1]. As extensively reported in [2], Metadata for this module include, per each Wikipedia page, its Wikipedia ID, BabelNet ID, gloss, resource type (that can be named entity or concept), Lemmata, Sensekey, WikiData ID.

The URLs for direct download of metadata of the Encyclopedic Module of the Polifonia Textual Corpus [1] can be found, for convenience, in Table 3.1.

lang	url
DE	10.5281/zenodo.6671494
EN	10.5281/zenodo.6671510
ES	10.5281/zenodo.6670984
FR	10.5281/zenodo.6671560
IT	10.5281/zenodo.6671571
NL	10.5281/zenodo.6671519

Table 3.1: The URLs to download the metadata of the Encyclopedic Module of the Polifonia Textual Corpus.

3.1.1.2 Data

The data of the Encyclopedic Module of the Polifonia Textual Corpus [1] are Wikipedia pages and is therefore in the public domain. The URLs for direct download can be found, for convenience, in Table 3.2.

lang	url
DE	10.5281/zenodo.6671663
EN	10.5281/zenodo.6671681
ES	10.5281/zenodo.6671673
FR	10.5281/zenodo.6671728
IT	10.5281/zenodo.6671734
NL	10.5281/zenodo.6671738

Table 3.2: The URLs to download the data of the Encyclopedic Module of the Polifonia Textual Corpus.

3.1.1.3 Statistics

The statistics of the Encyclopedic Module of the Polifonia Textual Corpus [1] are provided in Table 3.3.

	Pages	Sents	Tokens	Types	Links	Named Entities
Dutch	36.609	1.246.881	23.539.528	479.962	4.716.170	2.453.332
English	250.413	7.362.272	198.257.649	1.191.901	54.059.979	25.786.043
French	65.970	2.901.295	82.979.944	653.489	19.208.818	6.212.997
German	53.986	1.459.265	44.523.547	9.732.779	12.561.177	2.197.438
Italian	77.986	1.548.981	47.497.487	491.500	14.519.636	2.649.949
Spanish	57.891	1.247.583	36.229.557	537.465	7.171.759	2.996.185

Table 3.3: Overall statistics of the Encyclopedic Module of the Polifonia Textual Corpus.

3.1.2 The Books Module

Interpretative aspects of the MH discourse are covered by the Books module of the Polifonia Textual Corpus [1], including: monographs, essays, novels, biographies and diaries. This module allows to study MH discourse as it is performed by practitioners, scholars or music lovers. It includes themes, receptions and reactions through languages, time, and space about music and it is not restricted to musicologists but looks at music in a broad sense. It was developed selecting historical documents from open digital libraries, therefore available in the open domain.

3.1.2.1 The Books Module Ranking

Linguistic and musicologist experts' further analysis made for the release of the Books metadata and annotations of this deliverable highlighted that the Books collected through the methodology described in sections 4.2 and 4.2.1 of Deliverable 4.1 [2] needed to be filtered. The necessity of filtering was because the Books collected contained noisy texts, namely books that do not belong to the music domain. We produced a ranked list of the Books collected for each language to filter out noisy texts and focus on the most relevant resources for the music domain. The method that we applied to perform the ranking consisted of the following steps:

1. Lemmatising the titles of all the collected Books;
2. Checking the occurrence of Polifonia Lexicon's lemmas in the list of lemmas contained in each title;
3. Assignation of 1 point for each match obtained as a result of the check at point 2.

The steps above allowed us to obtain an ordered ranking of the Books for each language. The order of the ranking reflected the numerousness of the Polifonia Lexicon's lemmas in each book's title. We carried out a sub-selection of Books, making sure to include in it the first three-hundred books in this ranking per language (cf. Table 3.1.2.4) [2].

3.1.2.2 Metadata

We release the Metadata of the Books module of the Polifonia Textual Corpus [1]. According to the availability from the source origin, the Metadata may include the URL from which a text of the Books corpus is accessible, along with the title, the author, the year of publication, and the publisher. Metadata allow for a complete reconstruction of the corpus as we cannot make the actual texts available because they are subject to heterogeneous licensing. The URLs for direct download of the Metadata can be found, for convenience, in Table 3.4.

[2]The sub-selection of Books in Dutch language is particularly exiguous due to several issues we encountered in converting them into processable format via OCR technologies. In future releases, we will expand the sub-selection of Books in Dutch to a more significant number of elements after contextually improving the OCR component.

lang	url
DE	10.5281/zenodo.6772115
EN	10.5281/zenodo.6772136
ES	10.5281/zenodo.6772131
FR	10.5281/zenodo.6772142
IT	10.5281/zenodo.6772137
NL	10.5281/zenodo.6772145

Table 3.4: The URLs to download the metadata of the Books Module of the Polifonia Textual Corpus.

3.1.2.3 Data

The data of the Books Module of the Polifonia Textual Corpus [1], namely the texts that constitute this module of the corpus, cannot be published in their integral form because they are subject to heterogeneous licensing. The respective set of published metadata (cf. 3.4) allows for the reproduction of the whole corpus.

3.1.2.4 Statistics

The statistics of the Books Module of the Polifonia Textual Corpus [1] are provided in Table 3.5.

	Documents	Sents	Types	Tokens
Dutch	83	116.593	539.102	1.779.824
English	360	49.595	185.280	940.232
French	265	633.173	1.305.283	14.354.611
German	237	38.633	121.530	489.225
Italian	12.200	202.730	405.099	2.571.090
Spanish	41.093	731.606	1.852.430	20.180.197

Table 3.5: Overall statistics of the Books module

3.1.3 The Periodicals Module

Musicological discourse is covered by the Periodicals module of the Polifonia Textual Corpus [1]. This module represents the most specialized part of the corpus, representing how music experts have been discussing about MH over the years and across countries. This module includes the most influential musical journals of all times such as *Allgemeine musikalische Zeitung*, to which contributed Robert Schumann and Franz Liszt, or *The Harmonicon*, one of the first British music periodicals, in the form of full articles made possible by free copyright considering the publication date of the periodicals (19th century).

3.1.3.1 Metadata

We release the Metadata of the Periodicals module of the Polifonia Textual Corpus [1]. According to the availability from the source origin, the Metadata may include the URL from which a text of the Periodicals corpus is accessible, along with the title, the author, the year of publication, and the publisher. Metadata allow for a complete reconstruction of the corpus as we cannot make the actual texts available because they are subject to heterogeneous licensing. The URLs for direct download can be found, for convenience, in Table 3.6.

lang	url
DE	10.5281/zenodo.6761779
EN	10.5281/zenodo.6671912
ES	10.5281/zenodo.6761787
FR	10.5281/zenodo.6761794
IT	10.5281/zenodo.6761806
NL	10.5281/zenodo.6761809

Table 3.6: The URLs to download the metadata of the Periodicals Module of the Polifonia Textual Corpus.

3.1.3.2 Data

The data of the Periodicals Module of the Polifonia textual Corpus [1], namely the actual periodicals issues that constitute this module, cannot be published in their integral form because they are subject to heterogeneous licensing. The respective set of published metadata (cf. 3.6) allows for the reproduction of the whole corpus.

3.1.3.3 Statistics

The statistics of the Periodicals Module of the Polifonia Textual Corpus [1] are provided in Table 3.7.

	Documents	Sents	Types	Tokens
Dutch	125	187.350	716.506	3.880.499
English	2.868	4.400.015	7.342.527	76.180.398
French	349	329.166	696.427	5.111.915
German	705	121.113	544.376	2.405.289
Italian	1.251	433.465	992.902	7.879.459
Spanish	455	87.025	677.041	3.170.480

Table 3.7: Overall statistics of the Periodicals module

3.1.4 The Polifonia Pilots Module

The ten pilots that make up Polifonia are practical case studies that translate specific interests thematically or chronologically: five of them are based on textual corpora.The pilots therefore provide specific data at the corpus level that integrate with the more general ones of the three aforementioned modules. Below are the details of each pilot and the research requirements that motivate the data collection.

3.1.4.1 MusicBo Pilot

The MusicBo Pilot aims to investigate the role that music played in the life of the city of Bologna from a historical and a social perspective. The history of the performances is intertwined with that of the musicians or composers or clients, but also with the public and with the buildings that hosted them: the city is therefore described and told over the centuries not only through the history of musical performances but also through the history of music criticisms, encounters or the lives of musicians who have passed through Bologna. The MusicBo Pilot Corpus is made up of 135 texts in 4 languages: Italian (primarily, considering given the nature of the corpus), English, French, and Spanish, published between 1800 and today and referring to a period between 1600 to today. To meet the needs of the Pilot the following textual genres have been selected:

- essays;
- historical texts;
- biographies;
- autobiographies;
- correspondences;
- media; catalogs of performances.

Following the copyright licence, they cannot be disclosed in their entirety, but metadata that allows for the reproduction of the corpus is released (cf. Table 3.8). The metadata contains, per each document of the corpus, author, title, publication date, textual genre information and source link.

3.1.4.2 Bells Pilot

The Bells Pilot aims to represent the intangible phenomena of sound bells in urban and rural areas. It could inform about a sound landscape, performing the function of a marker of the daily and festive, ritual time. Also, it constitutes

a complex heritage, made by knowledge, practices and discourses in a social dimension. The core of the corpus of the Pilot Bells is made up of 76 texts in Italian. Texts are related to:

- General principles of bell acoustics, characteristics and production techniques;
- Performing techniques of Bell Ringers in Liguria and Emilia Romagna;
- Performing techniques in other Italian regions with an important tradition of bell-ringers;
- Major Bells and Bells concerts in Italy

The text types and their sources are specified as follows:

- 16 texts are edited volumes –scientific literature
- 36 texts are extracted from the websites of bell ringers' associations and campanologists' associations, in PDF and .txt format.
- 9 texts about Campanology and performing traditions are extracted from Wikipedia and Treccani website
- 11 catalogue sheets describing performing techniques in various regions are already published on the General Catalog of Cultural Heritage website[3].

Though some of the texts are resources on general concepts of Campanology applicable to the entire national territory, the corpus focus from a geographical point of view on two regions of Northern Italy (Liguria, Emilia Romagna) well known for the concentration of bells. The methodology used to collect the corpus mentioned above is scalable and applicable to other geographical areas such as other regions of Italy or outside Italy.

Following the copyright rule, the texts collected cannot be disclosed in their entirety, but metadata that allows for the reproduction of the corpus is released (cf. Table 3.8). The metadata contains, per each document of the corpus, author, title, publication date, textual genre information and source link.

3.1.4.3 Child Pilot

The Child Pilot aims to explore the role that music has played in children's lives through education, play and family and community practices, with special attention being paid to how perspectives and experiences change across time, culture and gender. Child Corpus is made up of 30 texts in english focused on a period from the eighteenth century to the early twentieth century. To meet the needs of the Pilot following textual genres have been selected:

- biographies,
- education books (to teach music to children),
- conduct literature

The texts collected are royalty-free, so they are available on Polifonia Github[4]. Links are reported, for convenience, in Table 3.9. Metadata contain author, title, publication date, textual genre information and sources link are made available (cf. Table 3.8).

3.1.4.4 Meetups Pilots

This pilot focuses on supporting music historians and teachers by providing a Web tool that enables the exploration and visualisation of encounters between people in the musical world in Europe from c.1800to c.1945, relying on information extracted from public domain books such as biographies, memoirs and travel writing, and open-access databases. For this specific knowledge, we are focusing on biographies from Wikipedia and selected 1002 Wikipedia pages regarding personalities relevant to the music domain. In perspective, the pilot aim at collecting sources for studying *musical meetups*, focusing on the dimensions of people, places, events and time entities, including. Relevant textual genres include:

- biographies,

[3] https://www.catalogo.beniculturali.it/
[4] https://github.com/polifonia-project/documentary-evidence-benchmark/tree/main/corpus

- memoirs,
- travel writing, and
- open-access databases.

The current texts collected are royalty-free, so they are available on Polifonia Github[5]. Links are reported, for convenience, in Table 3.9. Metadata contains ID page of Wikipedia, musician date of birth and sources link are made available (cf. Table 3.8).

3.1.4.5 Organs Pilots

The Organs pilot aims to represent the history of organs and organ building. This includes history and technical descriptions of individual instruments, as well as activities of historical agents, such as organ builders, organists, architects, administrators, etc. The corpus contains:

- The full text of Het Historische Orgel in Nederland (1997-2010), a 15 volume, 4,500+ pages encyclopaedia containing histories and images of almost 2,000 Dutch organs, published by Nationaal Instituut voor de Orgelkunst (NIvO).
- The full text of Het Orgel (volumes 1950-2022), the main Dutch journal for organ professionals.
- The full text of De Schalmei (1946-1950). Flemish journal for organists.
- Index of De Praestant (1952-1971)
- Table of contents of Orgelkunst (1978-2020)

[5] https://github.com/polifonia-project/meetups_pilot/tree/main/cleanText

3.1.4.6 Metadata

We release the Metadata of the Pilots module of the Polifonia Textual Corpus [1]. According to the availability from the source origin, the Metadata may include the URL from which a text of the Periodicals corpus is accessible, along with the title, the author, the year of publication, and the publisher. Metadata allow for a complete reconstruction of the corpus as we cannot make the actual texts available because they are subject to heterogeneous licensing. The URLs for direct download can be found, for convenience, in Table 3.8.

lang	url
BELLS	10.5281/zenodo.6672061
CHILD	10.5281/zenodo.6672093
MEETUPS	10.5281/zenodo.6672133
MUSICBO	10.5281/zenodo.6672165
ORGANS	10.5281/zenodo.6672193

Table 3.8: The URLs to download the metadata of the Pilots Module of the Polifonia Textual Corpus.

3.1.4.7 Data

The data of the Pilots Module of the Polifonia textual Corpus [1] collected for Bells, MusicBo and Organs pilots cannot be published in their integral form because they are subject to heterogeneous licensing. The respective set of published metadata (cf. 3.8) allows for the reproduction of the whole corpus. Texts collected for Child and Meetups Pilots are royalty-free, therefore we report links to access them in Table 3.9.

lang	url
CHILD	https://github.com/polifonia-project/documentary-evidence-benchmark/tree/main/corpus
MEETUPS	https://github.com/polifonia-project/meetups_pilot/tree/main/cleanText

Table 3.9: The URLs to retrieve the data of the royalty-free Pilots Module of the Polifonia Textual Corpus.

3.1.4.8 Statistics

The statistics of the Pilots Module of the Polifonia Textual Corpus [1] are provided in Table 3.10. Any inconsistencies with regard to the number of documents in the Pilot corpus compared to the number of annotated documents can be traced back to the annotation process with very noisy text. In such cases, the system may generate errors and move on to the next document without being able to create annotations. This problem is related to a small percentage of documents, but with a corrective action of the initial version of the texts we aim to correct these imperfections.

	Documents	Sents	Types	Tokens
BELLS	59	18.481	128.061	434.439
CHILD	30	157.815	361.550	3.442.840
MEETUPS	19.476	822.861	1.631.371	21.536.135
MUSICBO	46	51.781	289.247	1.412.456
ORGANS	1.660	25.647	45.298	368.439

Table 3.10: Overall statistics of the Polifonia Pilots module

4 Annotation of the Polifonia Textual Corpus

4.1 Annotated Polifonia Textual Corpus

The Polifonia Textual Corpus [1] (cf. Chapter 3) was processed using top-notch Natural Language Processing technologies in order to automatically extrapolate the implicit morphosyntactic, semantic and domain-specific information contained in it. This process, called *corpus annotation* [22], consists in encoding such information in a structured form to enable analysts (linguists, musicologists, scholars and the general public) to interrogate the corpus by performing sophisticated queries.

In case of sources available as images or pdfs file, the NLP technologies employed to produce the annotations were applied to the version of the documents converted into a processable format by the custom piece of OCR software[1] released within Deliverable 4.1 [2].

4.2 Data Format of the Annotations

To ensure a smooth machine-readability, we decided to encode the data produced by annotating the Polifonia Textual Corpus [1] in a "stand-off" fashion [23] following the CoNLL-U format [2], designed within the Universal Dependency (UD) project [24]. As the UD project's foundational principle is to build a consistent encoding resource that could be valid for many languages, we found the CoNLL-U format particularly suitable for our case, given the multilingual composition of Polifonia Textual Corpus [1]. In CoNLL-U format, each word/token is reported in tab-separated columns on one line, and sentence boundaries are indicated by blank lines. An example of a sentence extrapolated from the Encyclopedic Module (cf. 3.1.1) of the Polifonia Textual Corpus [1], annotated and encoded following the CoNLL-U format, is shown in Table 4.1.

4.2.1 Example

Given an input sentence from the Polifonia Textual Corpus' English Encyclopedic Module (cf. 3.1.1) such as:

> *James H. Mathis Jr. (born August 1967), known as Jimbo Mathus, is an American singer-songwriter and guitarist, best known for his work with the swing revival band Squirrel Nut Zippers.*

The resulting annotation will start with metadata information:

> #polifonia_doc_id = 32607842_bn___02615097n.html

The metadata information in the quote above provides a unique identifier for the document from which the sentence is extrapolated. In this case, it is composed of two identifiers: the first one is the BabelNet id of the corresponding Wikipedia page (32607842_bn), the second part is the Wikipedia identifier of the page (02615097n).

> #polifonia_sent_id = sent_0

Then, there is a progressive number for each sentence of the document, as displayed in the quote above.

> #sent = James H. Mathis Jr. (born August 1967), known as Jimbo Mathus, is an American singer-songwriter and guitarist, best known for his work with the swing revival band Squirrel Nut Zippers. And then there is the text of the sentence.

[1] https://github.com/polifonia-project/textual-corpus-population
[2] https://universaldependencies.org/format.html

Then, there is the text of the sentence, as displayed in the quote above. After the metadata, there is the sentence annotation, as we report in Table 4.1.

Token ID	Word Form	Lemma	POS	WordNet sense	NER class	NER BIO tag	Entity Linking	Is a musical concept?
token_0	James	James	PROPN		PERSON	B	James H. Mathis Jr.	0
token_1	H.	H.	PROPN		PERSON	I	0	0
token_2	Mathis	Mathis	PROPN		PERSON	I	0	0
token_3	Jr.	Jr.	PROPN		PERSON	I	0	0
token_4	(	(	PUNCT			O	0	0
token_5	born	bear	VERB	wn:02518161v		O	0	0
token_6	August	August	PROPN		DATE	B	August 1967	0
token_7	1967	1967	NUM		DATE	I	0	
token_8	)	)	PUNCT			O	0	0
token_9	,	,	PUNCT			O	0	0
token_10	known	know	VERB	wn:01426397v		O	0	0
token_11	as	as	ADP			O	0	0
token_12	Jimbo	Jimbo	PROPN		PERSON	B	Jimbo Mathus	0
token_13	Mathus	Mathus	PROPN		PERSON	I	0	0
token_14	,	,	PUNCT			O	0	0
token_15	is	be	AUX			O	0	0
token_16	an	an	DET			O	0	0
token_17	American	american	ADJ	wn:02927512a	NORP	B	United States	0
token_18	singer	singer	NOUN	wn:10599806n		O	0	1
token_19	-	-	PUNCT			O	0	0
token_20	songwriter	songwriter	NOUN	wn:10624540n		O	0	1
token_21	and	and	CCONJ			O	0	0
token_22	guitarist	guitarist	NOUN	wn:10151760n		O	0	1
token_23	,	,	PUNCT			O	0	0
token_24	best	well	ADV	wn:00011093r		O	0	0
token_25	known	know	VERB	wn:00596644v		O	0	0
token_26	for	for	ADP			O	0	0
token_27	his	his	PRON			O	0	0
token_28	work	work	NOUN	wn:05755883n		O	0	0
token_29	with	with	ADP			O	0	0
token_30	the	the	DET			O	0	0
token_31	swing	swing	NOUN	wn:07066042n		O	0	1
token_32	revival	revival	NOUN	wn:01047338n		O	0	0
token_33	band	band	NOUN	wn:08240169n		O	0	1
token_34	Squirrel	Squirrel	PROPN		ORG	B	Squirrel Nut Zippers	0
token_35	Nut	Nut	PROPN		ORG	I	0	0
token_36	Zippers	Zippers	PROPN		ORG	I	0	0
token_37	.	.	PUNCT			O	0	0

Table 4.1: Sentence annotation taken from Encyclopedic Module of Polifonia Textual Corpus following CoNLL-U format.

4.3 Annotation

4.3.1 Annotation Pipeline

We annotated each text of the corpus with a NLP pipeline composed of seven steps, as reported in 4.2.

Step #	Description
1	Sentence splitting
2	Tokenization
3	Lemmatization
4	Part-of-speech tagging
5	Word Sense Disambiguation
6	Named Entity Recognition
7	Entity Linking

Table 4.2: NLP pipeline implemented to annotate Polifonia Textual Corpus.

4.3.2 Annotation Tools

Sentence Splitting, Tokenization, Lemmatization, Part-of-Speech tagging and Named Entity Recognition (steps 1-4 and 6 of the NLP pipeline reported in Table 4.2) were conducted using SpaCy[3]. For the task approached with SpaCy, we used a dedicated SpaCy model for each language of the corpus, as summarised in Table 4.3. Among the several models that SpaCy makes available for each language, we chose those that allowed us to balance accuracy and speed of execution, with an advantage in favour of accuracy.

lang	model name
DE	de_core_news_lg
EN	en_core_web_trf
ES	es_core_news_lg
FR	fr_core_news_lg
IT	it_core_news_lg
NL	nl_core_news_lg

Table 4.3: The language covered in Polifonia Textual Corpus and the respective SpaCy Model employed for producing annotations.

Word Sense Disambiguation and Entity Linking (steps 5 and 7 of the NLP pipeline reported in Table 4.2) require more sophisticated technologies. For this reason, we used EWISER[4] for step 5 for English. The other languages of the corpus have been annotated using a new system developed *ad hoc* for the project. It exploits recent advantages on lexical semantics and in particular on the representation of word senses (ARES[5]) and a powerful WSD model (WSD-games[6]). This model ensured to work accurately, fastly and efficiently on different languages. For step 7 we used ExTenD[7] for English. For the other languages of the corpus we adapted WSD-games to work on the entity linking task. Also in this case the model is accurate, fast and can work efficiently on different languages.

[3] https://spacy.io/
[4] https://github.com/SapienzaNLP/ewiser
[5] http://sensembert.org/
[6] https://github.com/roccotrip/wsd_games_emb
[7] https://github.com/SapienzaNLP/extend

4.3.3 Annotation Layers

4.3.3.1 Sentence Splitting

Breaking down the texts of a corpus to their fundamental segments, the sentences, is usually one of the earliest steps tackled in an NLP pipeline. Approaching the problem of splitting sentences and ensuring an accurate sentence boundaries recognition is critical for successfully applying subsequent steps in NLP pipelines.

To perform sentence splitting for all languages of the Polifonia Textual Corpus [1], we used SpaCy's *Sentencizer* [8].

4.3.3.2 Tokenization

Tokenization is the task of segregating natural language text into its basic units. It is recognized as the first step in NLP [25]. Addressing it by recognizing its significance and being mindful of any language-specific complexities is crucial for enabling a successful application of subsequent steps in NLP pipelines.

As we can notice from the annotation of the example sentence reported in 4.2.1 in Table 4.1, the tokenization step of our NLP pipeline (cf. Table 4.2) breaks the input sentence down into its thirty-eight atomic components (tokens). Each of them is listed in the second column ("Word Form") of Table 4.1. Each token is associated to an ID, which is listed in the first column ("Token ID") of Table 4.1. For example, the token *born* is listed as a word form and linked to a specific ID *token_5*. Also punctuation markers are considered as single tokens.

To perform tokenization, we used specific SpaCy models for each language of the Polifonia Textual Corpus [1], as reported in Table 4.3. SpaCy's component dedicated to tokenization is *Tok2Vec* [9].

4.3.3.3 Lemmatization

Lemmatization is the process that allows the mapping of a token to its corresponding base form. It usually corresponds to linking the inflected variants of a word to their canonical form, namely the form that can be found in a dictionary.

As we can notice from the annotation of the example sentence reported in 4.2.1 in Table 4.1, the word form *known* (corresponding to *token_10* and *token_25*) is reconducted by the lemmatization step of our NLP pipeline (cf. Table 4.2) to the lemma *know*. In fact, "known" is an inflected variant of the verb "to know". Lemmas identified in the lemmatization step of our pipeline are listed in the third column ("Lemma") of Table 4.1.

To perform lemmatization, we used specific SpaCy models for each language of the Polifonia Textual Corpus [1], as reported in Table 4.3. SpaCy's component dedicated to associating base forms to tokens is *Lemmatizer* [10].

4.3.3.4 Part of Speech Tagging

Part-of-speech (POS) tagging consists of identifying the part-of-speech of each word occurring in the texts of a corpus [26]. It addresses the morphosyntactic level of language analysis and is usually one of the first steps performed in an NLP pipeline [27], as it paves the way for any semantic analysis tasks happening at a later stage of a pipeline. A part-of-speech tagging algorithm takes a two-fold input, consisting in a tagset and in each word of a sentence of a corpus. The algorithm's output is the association of a given word to the most appropriate tag of the tagset. As we rely on CoNLL-U as an annotation format, as a tagset we leverage the Universal POS Tags (UPOS) tagset[11], designed within the Universal Dependency (UD) project [24]. Part-of-speech (POS) tagging consists of identifying the part-of-speech of each word occurring in the texts of a corpus [26]. It addresses the morphosyntactic level of language analysis and is usually one of the first steps performed in an NLP pipeline [27], as it paves the way for

[8] https://spacy.io/api/sentencizer
[9] https://spacy.io/api/tok2vec
[10] https://spacy.io/api/lemmatizer
[11] https://universaldependencies.org/u/pos/index.html

any semantic analysis tasks happening at a later stage of a pipeline. A part-of-speech tagging algorithm takes a two-fold input, consisting of a tagset and each word of a corpus sentence. The algorithm's output is the association of a word to the most appropriate tag of the tagset. As we rely on CoNLL-U as an annotation format, as a tagset, we leverage the Universal POS Tags (UPOS) tagset[12], designed within the Universal Dependency (UD) project [24]. As we can notice from the example reported in 4.2.1 in Table 4.1, the word form *swing* (corresponding to *token_31*) is correctly assigned by the POS tagging step of our NLP pipeline (cf. Table 4.2) to the tag *NOUN*, which stands for the open class word noun, intended for common nouns and typically denoting a person, place, thing, animal or idea. This example shows that POS tagging algorithms deal with the challenges raised by the ambiguity of natural languages on a morphosyntactic level. In fact, "swing", in the English language, can also be a verb, for example, with the sense described by the WordNet glossa "hit or aim at with a sweeping arm movement" or with the metaphorical sense described by the WordNet glossa "alternate dramatically between high and low values".

To perform POS tagging, we used specific SpaCy models for each language of the Polifonia Textual Corpus [1], as reported in Table 4.3. SpaCy's component dedicated to predicting morphological features such as part-of-speech following UPOS tags is *Morphologizer* [13].

4.3.3.5 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is the task of resolving the intrinsic ambiguity of natural language by automatically associating the most appropriate sense to its utterances [28]. WSD addresses aspects related to the semantic level of language analysis and is enabled by the preliminary NLP tasks such as tokenization, lemmatization, POS tagging. The senses chosen and assigned to words at this step of our NLP pipeline (cf. 4.2) correspond to the meaning of the canonical form (lemma) of a token in a given context and is usually chosen from a pre-defined inventory of senses [12] stored in a lexical knowledge base. In our case, the lexical resource from which senses are chosen is WordNet[14].

As we can notice from the example reported in 4.2.1 in Table 4.1, the lemma *swing* (which corresponds to *token_31*) is linked by the WSD step of our NLP pipeline (cf. Table 4.2) to the Wordnet sense *wn:07066042n*. According to the corresponding Wordnet glossa, that sense accounts for the meaning of *swing* as "a style of jazz played by big bands popular in the 1930s; flowing rhythms but less complex than later styles of jazz". The lemma *swing*, considered for its morphosyntactic role of common noun (as identified by the POS tagging step of our NLP pipeline, which assigns to it the POS *NOUN*), is polysemic in the English language. In fact, according to Wordnet lexical database [4], it can lexicalize in nine different senses, among which the ones described by the Wordnet glossas "a state of steady vigorous action that is characteristic of an activity", "mechanical device used as a plaything to support someone swinging back and forth", "a sweeping blow or stroke" or "changing location by moving back and forth". Our WSD model chose the correct sense from the roster of candidates available on Wordnet. WSD information are stored in the fifth column ("WordNet sense") of Table 4.1.

To perform WSD on English language documents of the Polifonia Textual Corpus [1] we used EWISER (Enhanced WSD Integrating Synset Embeddings and Relations) [15]. EWISER is implemented as a supervised neural architecture that integrates explicit knowledge by embedding relational information from the WordNet knowledge graph [29].

To perform WSD on documents in languages other than English, we used a new system developed within the project. It exploits recent advantages in lexical semantics, in particular in the representation of word senses (ARES[16], [30]). ARES is used in combination with a powerful WSD model (WSD-games[17], [31]), which combines game dynamics with word and sense embeddings. The advantages of this new system are based on the fact that it is accurate, fast and can work efficiently on different languages.

[12] https://universaldependencies.org/u/pos/index.html
[13] https://spacy.io/api/morphologizer
[14] https://wordnet.princeton.edu/
[15] https://github.com/SapienzaNLP/ewiser
[16] http://sensembert.org/
[17] https://github.com/roccotrip/wsd_games_emb

4.3.3.6 Named Entity Recognition

Named Entity Recognition (NER) is the task of identifying and extracting mentions of named entities in the texts of a corpus [27], which usually correspond to tokens whose POS tag is *PROPN* (proper noun). People, locations and organizations are commonly considered named entities. Also, dates, times, money and percentages are considered as such.

As we can notice from the example reported in 4.2.1 in Table 4.1, the information retrieved by the NER step of our NLP pipeline (cf. Table 4.2) are stored in the sixth ("NER class") and seventh ("NER BIO tag") columns. The "NER class" column stores information about the type of named entity occurring in the text. For example, the word forms *James*, *H.*, *Mathis*, *Jr.* (which correspond respectively to *token_0*, *token_1*, *token_2*, *token_3*) have all been assigned the "NER class" *"PERSON"*. The "NER BIO tag" column reports BIO notation tagging format [27]. According to this tagging format, each token which constitutes the beginning of a named entity string is assigned the label 'B' (an abbreviation of 'beginning'). Each token which falls within a named entity string is assigned the label 'I' (an abbreviation for 'inside'). Each token standing outside of a named entity string is assigned the label 'O' (an abbreviation for 'outside'). For example, the word form *James*, which corresponds to *token_0*, has *B* as a "NER BIO tag", while the rest of the tokens that constitute the entity (*H.*, *Mathis*, *Jr.*, corresponding respectively to *token_1*, *token_2*, *token_3*) all have "I" as a "NER BIO tag".

To perform NER, we used specific SpaCy models for each language of the Polifonia Textual Corpus [1], as reported in Table 4.3. SpaCy's component dedicated to NER is *EntityRecognizer* [18], which approaches the task by implementing a transition-based algorithm.

4.3.3.7 Entity Linking

Entity Linking (EL) is the task of correlating named entities occurring in a text to the actual entity they refer to, which is usually identified by choosing among a set of possible candidates generated from a reference knowledge base [32]. The information required to perform such a choice is retrieved from the context in which the named entity occurs. In our NLP pipeline (cf. Table 4.2), the EL step leverages Wikipedia[19] as a reference knowledge base.

As we can notice from the example reported in 4.2.1 in Table 4.1, the information retrieved by the EL step of our NLP pipeline (cf. Table 4.2) are stored in the eight ("Entity Linking") column of Table 4.1. For each named entity mentioned in the sentence, the column reports the title of the corresponding Wikipedia page. For example, the word forms *Squirrel*, *Nut*, *Zippers* (corresponding respectively to *token_34*, *token_35*, *token_36*, are identified as constituting the named entity whose "NER class" is *ORG* and whose Wikipedia page title is *Squirrel Nut Zippers*[20]. As a convention, the "Entity Linking" column cell reporting the Wikipedia page title related to a given named entity corresponds to the row which stores the information about the first token ("NER BIO tag": *B*) of the named entity string.

To perform EL for English language we used ExtEnD (Extractive Entity Disambiguation) [21]. ExtEnD approaches EL as a text extraction task and is implemented as a Transformer-based architecture [32]. For languages other than English, we adapted WSD-games[22] [31] to work on the EL task. Also in this case the model is accurate, fast and can work efficiently on different languages.

4.3.4 Polifonia Lexicon Annotations

The process of building the Polifonia Lexicon was extensively described in Deliverable 4.1 [2]. In particular, Chapter 3 of [2] addresses all the details of the Polifonia Lexicon structure and methodology (sub-sections 3.1-3.2), including further information regarding concept selection, translation activities, specialist sources identified to extend the

[18] https://spacy.io/api/entityrecognizer
[19] http://wikipedia.org
[20] https://en.wikipedia.org/wiki/Squirrel_Nut_Zippers
[21] https://github.com/SapienzaNLP/extend
[22] https://github.com/roccotrip/wsd_games_emb

lexicon, and the approach implemented to enrich existing lexicalizations. The Polifonia lexicon has been structured as a domain-specific semantic network whose ultimate objective is to enable the automatic identification of music concepts in texts to allow scholars to carry out queries centred on MH knowledge on the Polifonia Textual Corpus [1].

As we can notice from the example reported in 4.2.1 in Table 4.1, the information retrieved by applying the automatic identification of music concepts by means of Polifonia lexicon are stored in the ninth ("Is a musical concept?") column of Table 4.1. For each "WordNet sense" identified in the sentence, the column reports the value *1* if the sense is also part of the Polifonia lexicon and, therefore, relevant to the MH domain. Otherwise, the column reports the value *0*. For example, the word forms *singer*, *songwriter*, *guitarist*, *swing* and *band* (corresponding respectively to *token_18*, *token_20*, *token_22*, *token_31* and *token_33*, are identified as musical concepts, therefore their corresponding cells in the ninth column report the value *1*.

4.3.5 Annotations Download

4.3.5.1 The Encyclopedic Module

The annotations of the Encyclopedic Module (cf. 3.1.1) of Polifonia Textual Corpus [1] are released within this Deliverable. The URLs for direct download can be found, for convenience, in Table 4.4.

lang	url
DE	10.5281/zenodo.6702689
EN	10.5281/zenodo.6759156
ES	10.5281/zenodo.6759021
FR	10.5281/zenodo.6759025
IT	10.5281/zenodo.6759017
NL	10.5281/zenodo.6757537

Table 4.4: The URLs to download the annotations produced for the Encyclopedic Module of the annotated Polifonia Textual Corpus.

4.3.5.2 The Books Module

At this stage, the annotations of the Books module are only available to the Polifonia consortium members in accordance with the heterogeneous licensing of the Books sources. Interested parties may contact us, and we will evaluate the sharing of the annotated data.

4.3.5.3 The Periodicals Module

At this stage, the annotations of the Periodicals module are only available to the Polifonia consortium members in accordance with the heterogeneous licensing of the Periodicals sources. Interested parties may contact us, and we will evaluate the sharing of the annotated data.

4.3.5.4 The Polifonia Pilots Module

At this stage, the annotations of the Pilots module are only available to the Polifonia consortium members in accordance with the heterogeneous licensing of the Pilots sources. Interested parties may contact us, and we will evaluate the sharing of the annotated data. Texts collected for Child (cf. 3.1.4.3) and Meetups (cf. 3.1.4.4) Pilots are royalty-free, therefore we release their annotations that can be downloaded from the table below:

lang	url
CHILD	10.5281/zenodo.6759261
MEETUPS	10.5281/zenodo.6759272

Table 4.5: The URLs to download the annotations produced for the Polifonia Pilots Module of the annotated Polifonia Textual Corpus.

5 Interrogation of the annotated Polifonia Textual Corpus

We developed custom software interfaces (APIs) to ensure an effective interrogation of the annotated Polifonia Textual Corpus [1]. To design and implement the APIs, we collected requirements by involving professors and senior researchers in the field of linguistics with a long experience in building, annotating and interrogating digital corpora to accomplish their research objectives. The requirements collected guided the development of our interrogation APIs. The following paragraphs describe the requirements that have been elicited, collected and formalised. Then, we will explain the process of interrogating the APIs to carry out specific kinds of searches, providing some examples and finally we will show an example of mapping between the types of interrogation and Persona's competency questions used to collect requirement for knowledge graph modeling as explained in Deliverable 1.1 [33], and Deliverable 2.1 [34].

5.1 Interrogation Requirements

5.1.1 Requirements Collections

In order to ensure successful delivery of the annotated Polifonia Textual Corpus [1] interrogation features, we led round tables and brainstorming sessions to elicit corpus interrogation requirements from domain experts. The sessions involved also WP1 and WP5 team members, respectively responsible for Polifonia's Pilots and Web Portal and the aspects of the project related to the human interaction with MH. The main objective of the sessions was to receive input from Linguistics Professors and Senior Researchers to ensure that the design of our APIs for corpus interrogation would meet the needs of scholars who pursue their research objectives by querying digital corpora. To achieve our goal, we first shared it with our audience to align everyone on the purpose of the sessions. In order to clarify our message further, we made explicit that the information that we needed to collect could be thought as answers to the following questions:

- What kind of interrogation would an domain expert expect to be able to perform on such APIs?
- What input would an domain expert expect to give as a query, and what output would the domain expert expect to receive as search results?

Another objective of the sessions was leveraging WP1 and WP5 members' expertise to evaluate the feasibility of implementing a web-based User Interface to allow end-users to query the corpus.

To achieve the first objective, we previewed the annotations of the Polifonia Textual Corpus [1] by showcasing the content of the related GitHub repository[1] to facilitate the initiation of a brainstorming session on the possible queries and searches that our APIs would need to allow. The brainstorming session raised the following draft use-cases:

1. Scholars in terminology would find it helpful to leverage tools that facilitate the investigations of specific musical domain terminology. For example, it would be convenient for them to construct queries which combine word senses and Polifonia Lexicon annotations to more easily recall concepts related to musical instruments or genres, such as retrieving all the musical instruments used in acoustic music;
2. Scholars in discourse analysis may want to construct queries combining morphosyntactic tags to word senses and Polifonia Lexicon annotations to spot common words that are used metaphorically in music contexts, such as in the expression "the singer sang like a bird";
3. Linguistics scholars may need to combine different sources of information to correlate the information on concepts offered by the Polifonia Textual Corpus [1] annotations to their related frequency, concordances or multimodal data, available in repositories external to the Polifonia's ecosystem.

[1] https://github.com/polifonia-project/Polifonia-Corpus/tree/master/annotations

WP4 team members, along with WP1 and WP5 team members, addressed all the use-cases raised by the domain experts.

WP4 team highlighted that the first draft use-case raised by domain experts in the sessions could, at least partially, already be satisfied by leveraging the information returned by the current implementation of the annotation pipeline (cf. 4.2). For example, it is possible to build queries by combining word senses returned by the WSD step of our pipeline (cf. 4.3.3.5) and the Polifonia Lexicon Annotations. Future expansions to the annotation layers 4.3.3 and the APIs features may be considered to make available a wider variety of querying methods aimed at retrieving concepts that are somehow related by means of semantic relations, as hypernymy, hyponymy, meronymy, etc. These relations may be expanded to encompass, for example, the creation of tailored music-focused semantic links between the Polifonia Lexicon concepts and WordNet concepts.

With regard to the second use-case raised in the sessions, WP4 team commented that further analysis would be needed to understand the feasibility of implementing the improvements needed, as they would imply a dramatic expansion of the current annotations capabilities offered. Possibly, the music knowledge graph that will be constructed as an end goal of the Polifonia project will meet those requirements. In that case, the second draft use-case would be satisfied by the capabilities offered by Polifonia's music knowledge graph, therefore it may not be necessary to expand the corpus annotations capabilities to that extent.

As far as the third use-case is concerned, WP4 team and the domain experts agreed that, at this stage, integration with third parties corpora resources should be considered a low priority and that the WPs work should concentrate on intra-corpus queries.

The sessions audience also discussed the possibility of developing a User Interface to query the annotated Polifonia Textual Corpus [1], as several attendees believed that such a tool would be crucial to ensure that the usage of the annotated Polifonia Textual Corpus [1] could bring the maximum of its potential benefits. This User interface could be integrated in the Polifonia Web Portal, but WP1 and WP5 members raised concerns about the possibility of integrating such an interface into it. In fact, the Polifonia Web Portal is not targeted at linguistics experts. At the same time, an interface that allows querying a corpus based on advanced linguistics annotations would aim at satisfying only the needs of a sectorial segment of possible users (linguistics scholars). In conclusion, the participants to the sessions agreed that implementing a User Interface, even a minimal version of the tool designed and developed to target linguistics scholars specifically, is an essential byproduct of the annotated Polifonia Textual Corpus [1] release. However, as its implementation has not been considered feasible in the time-frame dictated by the M18 deadline, the audience committed to holding further discussions in the next months to bring such an interface to life at a later stage of the project.

5.2 APIs for Interrogating the Corpus

5.2.1 Interrogation Functionalities

The APIs that we developed for interrogating the Polifonia Textual Corpus [1] allow for performing structured queries on the corpus texts based on the annotations described in Chapter 4. We release the code to replicate the interrogation of the corpus, along with detailed instructions to download the scripts and to satisfy its dependencies, in the *interrogation* section of the Polifonia Textual Corpus GitHub repository[2].

The main script to use to interrogate the corpus is:

```
interrogate.py
```

The script has seven different parameters that can be used to select, navigate and store sentences of the corpus that satisfy a specific query. The parameters are introduced at the points (1-7) in the numbered list below.

1. The following parameter allows the end-user to indicate the path were the annotations databases are stored:

   ```
   --annotations_path
   ```

[2] https://github.com/polifonia-project/Polifonia-Corpus/tree/master/interrogation

Its default value is:

`annotation/db/`

2. The following parameter allows the end-user to indicate what module of the corpus has to be queried:

   `--corpus`

   Its possible argument is one of the Polifonia Textual Corpus' modules (cf. Chapter 3): *Wikipedia*, *Books*, *Periodicals* or *Pilots*.

3. The following parameter allows the end-user to indicate the language to use for the interrogations:

   `--lang`

   Its possible argument is one of the Polifonia Textual Corpus' languages (cf. Chapter 3): DE, EN, ES, FR, IT or NL.

4. The following parameter allows the end-user to indicate the type of interrogation to conduct:

   `--interrogation_type`

   It can be *Keyword*, *Concept* or *Entity*. Each interrogation type will be explained in the next sections (cf. Subsection 5.2.2).

5. The following parameter allows the end-user to indicate the input to use for the interrogation:

   `---query`

6. The following parameter allows the end-user to indicate the number of sentences to get at each interrogation:

   `--sent_n`

7. The following parameter allows the end-user to indicate if the results of the interrogations have to be saved to a file. The default value of this parameter is 'No':

   `--save_to_file`

5.2.2 Interrogation Types

5.2.2.1 Keyword Search

The keyword search functionality can be used to select the sentences in the corpus that contain the keyword specified in the query. Setting the parameter at point 4 of the numbered list in Subsection 5.2.1 to "keyword" tells the system to interrogate the corpus by searching the word that the users can provide as the argument of the parameter at point 5 of the same numbered list:

```
> python interrogate.py --interrogation_type keyword  --query swing
```

Example

The prompt provided below as an example of keyword search (cf. 5.2.2.1) will retrieve, from the Encyclopedic Module 3.1.1 of the Polifonia Textual Corpus [1], sentences in English that contain the keyword "swing".

```
> python interrogate.py --annotations_path ../annotations/db
--corpus Wikipedia --lang EN --interrogation_type keyword
--query swing --sent_n 100 --save_to_file Yes
```

As an output, the prompt will show to the users up to 100 sentences at time (cf. Figure 5.1). The users will then be asked if they want to repeat the query to retrieve other sentences. All the sentences that have been shown will be saved to a file in the "out" folder of the repository.

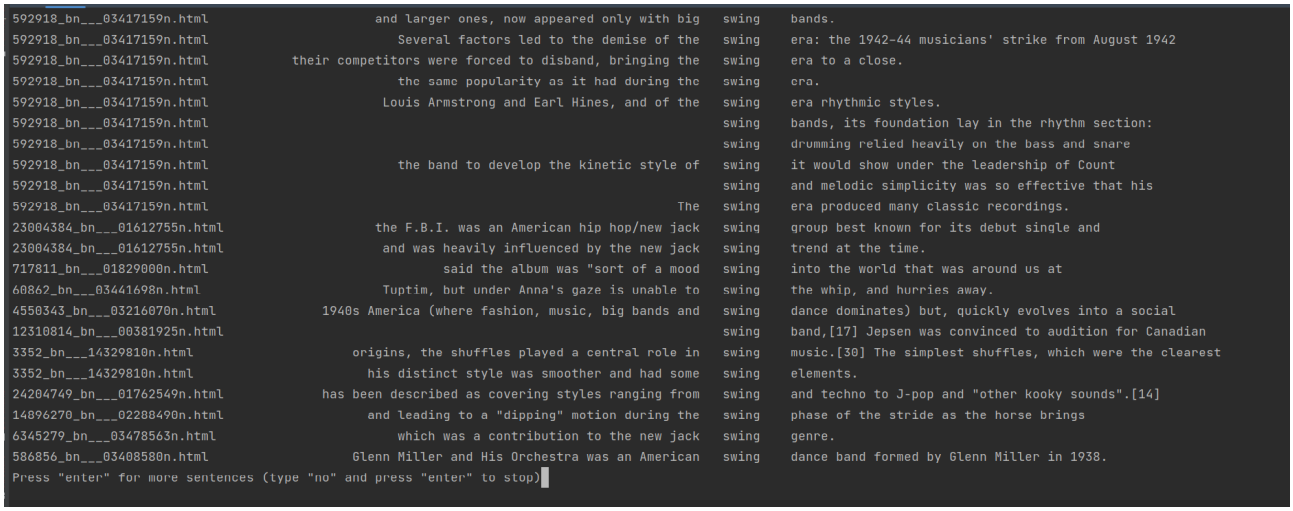

Figure 5.1: An example of output of a keyword search. The input keyword is "swing".

5.2.2.2 Concept Search

The concept search can be used to select the sentences in the corpus that contain an occurrence of a word that is recognized by the Word Sense Disambiguation (cf. Subsection 4.3.3.5) step of our NLP pipeline (cf. Table 4.2) as linked to a specific WordNet sense. Setting the parameter at point 4 of the numbered list in Subsection 5.2.1 to "concept" tells the system to interrogate the corpus by searching the sentences annotated with the WordNet sense specified. The end-users can input a word base form (*lemma*) as the argument of the parameter at point 5 of the same numbered list.

```
> python interrogate.py --interrogation_type concept  --query swing
```

Example

The prompt provided below as an example of concept search (cf. 5.2.2.2) will retrieve, from the Encyclopedic Module 3.1.1 of the Polifonia Textual Corpus [1], sentences in English that contain a concept that has "swing" as its *lemma*.

```
> python interrogate.py --annotations_path ../annotations/db
--corpus Wikipedia --lang EN --interrogation_type concept
--query swing --sent_n 100 --save_to_file Yes
```

To select the concept the system will provide a list of concepts as shown at Figure 5.2. Entering "4", the system will show to the users up to 100 sentences at a time annotated with the corresponding sense and ask the users if they want to repeat the search to retrieve other sentences. All the sentences that have been shown will be saved to a file in the "out" folder of the repository.

As shown at Figure 5.3, the results of the query are presented sentence by sentence. In each line of the results there is the document ID of the sentence and the keyword in context. If the results are saved, the entire sentences are saved, not only the snippets of the keyword context.

```
The word swing is associated with 22 concepts
Please select one concept from the list below indicating its number:
0. n - a state of steady vigorous action that is characteristic of an activity
1. n - mechanical device used as a plaything to support someone swinging back and forth
2. n - a sweeping blow or stroke
3. n - changing location by moving back and forth
4. n - a style of jazz played by big bands popular in the 1930s; flowing rhythms but less complex than later styles of jazz
5. n - a jaunty rhythm in music
6. n - the act of swinging a golf club at a golf ball and (usually) hitting it
7. n - in baseball; a batter's attempt to hit a pitched ball
8. n - a square dance figure; a pair of dancers join hands and dance around a point between them
9. v - move in a curve or arc, usually with the intent of hitting
10. v - move or walk in a swinging or swaying manner
11. v - change direction with a swinging motion; turn
12. v - influence decisively
13. v - make a big sweeping gesture or movement
14. v - hang freely
15. v - hit or aim at with a sweeping arm movement
16. v - alternate dramatically between high and low values
17. v - live in a lively, modern, and relaxed style
18. v - have a certain musical rhythm
19. v - be a social swinger; socialize a lot
20. v - play with a subtle and intuitively felt sense of rhythm
21. v - engage freely in promiscuous sex, often with the husband or wife of one's friends
```

Figure 5.2: An example of concept selection given the input *lemma* "swing".

5.2.2.3 Entity Search

The entity search can be used to select the sentences in the corpus that contain an occurrence of a word being recognized by the Named Entity Recognition (cf. Subsection 4.3.3.6) and Entity Linking (cf. Subsection 4.3.3.7) steps of our NLP pipeline (cf. Table 4.2) as a named entity linked to a specified Wikipedia entity. Setting the parameter at point 4 of the numbered list reported in the "Interrogation Functionalities" Subsection above 5.2.1 to "entity" tells the system to interrogate the corpus by searching the sentences annotated with the specified Wikipedia entity. The end-user can input a word through the parameter at point 5 of the same numbered list. Then, our APIs will ask the end-user to select an entity related to the provided word.

```
> python interrogate.py --interrogation_type entity --query bach
```

Example

The prompt provided below as an example of entity search (cf. Subsection 5.2.2.2) will retrieve, from the Encyclopedic Module (cf. Subsection 3.1.1) of the Polifonia Textual Corpus [1], sentences in English that contain a mention to a named entity that has "bach" as its *lemma*.

```
> python interrogate.py --annotations_path ../annotations/db
--corpus Wikipedia --lang EN --interrogation_type entity
--query bach --sent_n 100 --save_to_file Yes
```

To select the specific named entity the system will provide a list of named entities based on the input *lemma* as shown in Figure 5.4.

Entering "0", the system will show to the users up to 100 sentences at a time annotated with the corresponding named entities and ask the users if they want to repeat the search to retrieve other sentences. All the sentences that have been shown will be saved to a file in the "out" folder of the repository.

```
The research is based on the concept defined as:
swing:  a style of jazz played by big bands popular in the 1930s; flowing rhythms but less complex than later styles of jazz
36792451_bn___03041639n.html                 It also paraphrases the lyrics to the Western   swing    standard " Cocaine Blues",[9 ] and touches on
12131261_bn___00354925n.html                      appearing on many of his albums in a   swing    trio that includes pianists Ray Kennedy and Larry
88771_bn___00054669n.html           steady beat of a metronome neglects the characteristic   swing    of samba.[40][41 ] A style of performance that
3225847_bn___02580499n.html               But the song declined in popularity during the   swing    era , as it had become something of
26208253_bn___01977493n.html                  band to perform tracks from his album of   swing    songs , as well as distinguished classics from
2668859_bn___17163963n.html                    mixed country , rock and roll , Western   swing    , rockabilly , and jump blues together on
16441868_bn___00881503n.html                 hot melodic variations and cool , laid back   swing    gave the whimsical " Fran Dance " (
1214126_bn___01219773n.html                       with elements of disco , reggae , 40s   swing    music , 70s jazz pop and motown.[104 ]
37371769_bn___15875559n.html                           is described to be of " new jack   swing    " style , which was popular during the
1754447_bn___00994470n.html                      , and The Selecter , recruited ska /   swing    auters Cherry Poppin ' Daddies and Reel Big
10017860_bn___00119198n.html                      East Coast Family . " Two new jack   swing    groups , Another Bad Creation and Bell Biv
32427_bn___01564673n.html                          big band music . With the rise of   swing    music , however , from 1935 to 1945
159031_bn___00023352n.html                    record company 's label ) . During the   swing    era , when a bobby soxer went looking
3578999_bn___02924669n.html             Orchestra — a Nazi - sponsored German propaganda   swing    ensemble derisively nicknamed " Goebbels ' band"[31 ]
15983857_bn___00836340n.html        exploring Latin American rhythms and American Big Band   swing    made him very popular in South America ,
149223_bn___03551905n.html                 have different durations . First , the mood   swings   of BPD and bipolar disorder tend to have
11524102_bn___00281919n.html                       he plays to his crowd with " Texas   Swing    " ( you can almost hear George Strait
3610274_bn___15112125n.html                resurfaced in 2014 , reinventing himself as a   swing    guitarist to play with Lola Lamour And Her
11566003_bn___00286779n.html                          film . He sings this song on a   swing    set he built similar to the one in
185514_bn___00057215n.html                     the latter , we find the " familiar   swing    and syncopation " of the African American.[42 ]
52116_bn___00010096n.html                  bassist Billy Taylor . The band included such   swing    era musicians as the trombonist Jack Teagarden ,
15419061_bn___00760379n.html                 You " . Notable compositions include " Java    Jive    " , " I 'll Take Romance "
1611629_bn___01790379n.html                    nod to the Foy Willing - led Western   swing    combo from the 1940s , Riders of the
294084_bn___02302620n.html               legendary Victoria , Graham Fox gives plenty of   swing    on the drums and sizzle cymbals and Louis
```

Figure 5.3: An example of output of a concept search. The input *lemma* is "swing" and the selected concept is *4* as seen in Figure 5.2

```
The word bach is associated with 2 entity
Please select one entity from the list below indicating its number:
0. Johann Sebastian Bach - German baroque organist and contrapuntist; composed mostly keyboard music; one of the greatest creators of western music (1685-1750)
1. Bach (surname) - Bach is a surname of German-language origin.
$
```

Figure 5.4: An example of entity selection given the input *lemma* "bach".

```
The research is based on the concept defined as:
bach:   German baroque organist and contrapuntist; composed mostly keyboard music; one of the greatest creators of western music (1685-1750)
25169847_bn___01863631n.html                          , possibly for the lute , written by   Johann   Sebastian Bach shortly after he moved from Köthen
214019_bn___01421260n.html      founders of modern musicology.[1 ] His publications include   Johann   Sebastian Bach : His Life , Art ,
713599_bn___03459637n.html                          by his son Carl Philipp Emanuel . The    Bach    family was of importance in the history of
713599_bn___03459637n.html                            went into a more modern world . The    Bach    family never left Thuringia until the sons of
19092781_bn___01181523n.html                          1988).[7 ] Born in Zagreb , Croatia ,    Bach    graduated as a civil engineer from the University
3441041_bn___02799329n.html                                 , BWV 248 , is an oratorio by   Johann   Sebastian Bach intended for performance in church during
20264440_bn___01300585n.html                set of chorale preludes for organ prepared by   Johann   Sebastian Bach in Leipzig in his final decade
13833886_bn___00587914n.html                     from the 69 Sacred Songs and Arias that   Johann   Sebastian Bach contributed to Musicalisches Gesangbuch by Georg
8522408_bn___00438104n.html        technical accuracy and adaptation of classical works by    Bach    and Mozart into his own blend of neo
36000374_bn___02949278n.html            includes five John Lewis arrangements of pieces by   Johann   Sebastian Bach , interspersed with four original blues
713599_bn___01894386n.html                          by his son Carl Philipp Emanuel . The    Bach    family was of importance in the history of
995303_bn___03893506n.html                       a Passion , a sacred oratorio written by   Johann   Sebastian Bach in 1727 for solo voices ,
14358658_bn___00643651n.html       minor for transverse flute and obbligato harpsichord by   Johann   Sebastian Bach ( BWV 1030 ) is a
1045209_bn___00164287n.html                        of Josquin des Prez , Carlo Gesualdo ,   Johann   Sebastian Bach , Arnold Schoenberg , as well
4685487_bn___03236738n.html                a society devoted to performing the music of   Johann   Sebastian Bach ( 1685-1750 ) with small ,
2021679_bn___01656711n.html                               for solo violin ( BWV 1004 ) by   Johann   Sebastian Bach was written between 1717 and 1720
```

Figure 5.5: An example of output of an entity search. The input *lemma* is "bach" and the selected named entity is *0* as seen in Figure 5.4

5.2.3 Mapping between Polifonia Ecosystem's *Personas* and Interrogation Types

The API developed, and the resulting interrogation types, were mapped onto the competency questions of the *Personas*[3] described in Deliverable 1.1 [33] and Deliverable 2.1 [34] to identify requirements for KG modelling. In

[3] https://polifonia-project.github.io/ecosystem/personas.html

general terms, the keyword search is useful when one wants to research the behaviour of the exact lexicalization of a concept, for example: what are the contexts in which the word *swing* appears. The Concept search takes advantage of word sense disambiguation, which allows for a more complex semantic level query: because of the polysemy of words (a term with multiple meanings of the same sense, or a term that has several different senses between them) sense disambiguation is useful, which also allows for in-depth searches, often related to the interpretive aspect of language. For example, one might be interested in comparing two different types of speech in which swing appears for its musical and non-musical sense. The third type, entity search, allows searches using the features (time and location) of known entities such as places, people, and institutions. For example, one might be interested in knowing all the pieces of work of a given author belonging to a given period that are attributable to the *swing* genre. In Table 5.1, mapping has been made based on *Personas*, their research goal and the aforementioned three levels of knowledge extraction by aligning with the *Personas*' competency question, depending on whether they are related to the search for an exact lexicalization, a second meaning or an interpretation process, or any entity features. However, the *Persona* creation process is ongoing; the table below will be updated constantly. In addition, future research will be performed to respond to a level of discourse analysis, preliminary work of which has already been done and will be discussed in the 6.3 future works section.

Persona(1)	Goal/scenario (2)	Keyword search (3)	Concept Search (4)	Entity Search (5)
Laurent : Music Journalist	Discover and explore Archives, Historical and Research Resources for the extrapolation of content to be included in his musical newsletter	CQ : What types of resources can I find?	CQ: Is the music resource X complete or incomplete?	CQ : Can I search for a musical content by applying filters (genre, historical period . . .)?
Amy : Organologist / Musicologist / Music historian	Amy wants to discover artistic and technical trends of organs and how these developed which may indicate wider social trends.	CQ : What is left from the original organs built by XX organ builder?	CQ: What are the arthistorical features of the front of the organ? In which style is the organ built?	CQ: What are geographically distinct features? (e.g., in what differ 17th century Southern German organs from 17th century Spanish organs?).
Carolina : Music historian, Researcher	Carolina has to prepare a conference for the anniversary of the birth of the composer Giacomo Antonio Perti (baroque Italian composer) and she needs to collect some information about his career.	CQ: In which scores is there evidence of Perti's musical composition?	CQ: Who has spoken about the musical composition X?	CQ: Where was the musical composition X performed? CQ: In which buildings was musical composition X performed? CQ: Where was the musical composition X performed for the first time?
Ralph : musicologist	Ralph is developing a database of librettos with the goal of relating words to music, particularly in terms of how they create emotion.	CQ: What words are associated with terms that indicate emotions?	CQ: How do libretto and music relate, e.g. in describing an emotion?	CQ: Who is the poet (librettist) associated with a text?

Table 5.1: Alignment between *Personas*, *Competency Questions* and Interrogation Types.

6 Discussion: Main Challenges and Future Work

6.1 Annotation

As outlined in Deliverable 4.1 [2], and as remarked in Section 1.1 of this report, the Polifonia Textual Corpus [1] includes scanned documents (image or pdf) whose text needs to be extracted through OCR models. OCR is noisy, especially for historical documents, and it is challenging to identify OCR post-correction models that ensure high-quality performance in different languages. Issues in the text extracted due to low-quality OCR processing impact the quality of the linguistic annotations. We plan to re-use state-of-the-art approaches to perform post-OCR correction, as recurrent (RNN) and deep convolutional network (ConvNet) employed in [35]. We believe that an overall improvement to the quality of the OCRed texts will result in higher performance of the models used for producing the annotations (described in Chapter 4) and, eventually, higher quality of the linguistic annotations.

Furthermore, the linguistic annotations that we provide (cf. Chapter 4 for a detailed description) can be considered "silver" annotations, as they are produced automatically by state-of-the-art models. We plan to design and implement a process to produce human-validated gold-standard annotations. This process may provide the corpus' end-users with the possibility of checking the annotations' accuracy and applying corrections. Gathering a set of gold-standard annotations may allow for evaluating the quality of automatically produced annotations against manually-validated ones.

6.2 Interrogation

The interrogation APIs released within this Deliverable's report provides MH and linguistics scholars with novel capabilities such as performing queries based on specific word senses or accessing MH-specific information thanks to the layer of annotations based on the Polifonia Lexicon (cf. Sub-section 4.3.4). A summary of novel contributions and impact of the Polifonia Textual Corpus is reported in Sub-section 2.1.4. However, the current interrogation APIs capabilities can be further expanded to cover a broader range of linguistics- and terminology-related use-cases. Furthermore, interrogating the corpus can be made more user-friendly to favour the interaction by linguistics and MH scholars, which may lack the technical knowledge to query the corpus using command-line instructions comfortably. During the brainstorming sessions with the expert linguists, questions emerged related to discourse analysis, which takes into account features such as the socio-cultural representation of locutours or the use of rhetorical devices that are functional to some specific socio-cultural representation of the world. The examples that emerged from the brainstorming were regrouped into four types of possible research that we will list below. This grouping is useful for clear representation, but this work will continue in the coming months.

- Metaphor and figurative language. One may want to be able to automatically extract information as to what kind of metaphors/hyperboles are used in the description of a given work or musician. Moreover, conversely, how is a work/ musician/ piece of music used to create figurative expressions (e.g. *He is the Mozart of the chessboard*)?;
- Word behaviour beyond the commonly available co-occurrences analysis. One may want to automatically extract from the corpus all the adjectives (but also nouns, verbs, and adverbs) that enact an evaluation of a given work or musician and to see if such evaluation is positive or negative and along what lines (e.g. is it a question of a piece of music being described as relaxing, or rather of a musician being described as a sensitive person?);
- Investigating gender bias in the history of music. One may be interested in understanding what is the difference according to the gender of the author of the production and reception of musical works throughout history. In this way, it is possible to study the trend of artistic customs from a gender perspective;

- Terminological and lexical aspects. One might be interested in studying the translations of a term over a medium to an extended period (e.g., two centuries), for example, of a musical instrument. In this way, a diachronic analysis of terms belonging to the specialized language of music would be possible, which might also show the influence of other cultures in the creation or dissemination of the instrument.

During the next months, we will hold other sessions with linguistics scholars and MH domain experts to elicit and capture additional use-cases and more granular contributions to the requirements collections. Based on this input, we will improve the interrogation APIs released within this Deliverable (cf. Section 5.2 to allow for performing more complex queries on the Polifonia Textual Corpus [1]. We will continue our discussion with WP1 and WP5 members to translate the use-cases into actionable requirements to guide the design of a custom User Interface whose objective is to facilitate the access and query of the annotated Polifonia Textual Corpus [1] to a broader audience, especially targeting linguistics scholars. We will also resume discussing the suitability of including this User Interface in the Polifonia Web Portal.

6.3 Knowledge Extraction

In the immediate future, within the scope of the work in progress for the Deliverable 4.5 ("Software for knowledge extraction from text – context – 1st version"), we will concentrate on transforming the unstructured texts of the Polifonia Textual Corpus [1] into Abstract Meaning Representation (AMR) [36] graphs. We will start by producing a sub-sample of the corpus. We decided to create a sub-sample of the Polifonia Textual Corpus [1] for several reasons. It allows for iteratively testing the effectiveness of the algorithms developed to minimise the loss of information that segregation of texts into sentences (required to enable text-to-AMR parsing) inherently brings. It also favours a more manageable Quality Assurance of the AMR graphs produced in the text-to-AMR step output. When the results obtained from processing the samples are considered satisfactory, we will expand the text-to-AMR parsing procedure to the entire corpus. The language scope of the sample is English. Regarding sample sizing, we took the Encyclopedic Module of the corpus as a basis, considering one-thousand Wikipedia pages. We calculated the total number of sentences present in the chosen Wikipedia pages and selected a comparable number of sentences from the documents contained in the Books and Periodicals modules. To address the challenges of multilingual and historical corpora textual material, we will apply tailored pre-processing strategies to this mini sample, such as co-reference resolution and post-OCR correction. Lastly, we will process this mini sample through automatic semantic parsing based on AMR [36] formalism, leveraging state-of-the-art *seq2seq* models [37]. This parsing step will harvest the socio-cultural and historical context of MH from textual sources. The structured knowledge automatically extracted through the AMR parsing will then need to be encoded according to formal logic-based representations of knowledge [38], to serve as a compatible input for the ultimate goal of the Polifonia project, which is the modelling and population of the Polifonia's knowledge graph.

7 Conclusions

This report presented the work carried out on the Polifonia project for the second Deliverable of WP4 due at M18. The objective sought in realising the products of this Deliverable was three-fold: (i) releasing a larger and more robust version of the Polifonia Textual Corpus' data, metadata, and statistics, (ii) providing a richer version of the Polifonia Textual Corpus [1] by producing advanced morphosyntactic, semantic and MH-specific annotations employing cutting-edge NLP tools and techniques, (iii) support the work of MH scholars and domain experts by developing tailored APIs to facilitate the Polifonia Textual Corpus' interrogation.

We documented the release of a larger and more robust version of the Polifonia Textual Corpus [1] in Chapter 3. We reported links to the Polifonia Textual Corpus' data (the actual texts included in the corpus) when possible. However, due to the heterogeneous licensing, we could not publish the data that make up *Books*, *Periodicals* and *Pilots* modules. To mitigate the licensing-related issues, we released metadata for each module and language of the corpus to allow the end-user for its reconstruction.

We reported on the automatic Polifonia Textual Corpus' annotation process and format and the NLP techniques used in Chapter 4. We made the annotations available in the annotation section [1] of the Polifonia Corpus repository on GitHub.

We described how we pursued the aim of effectively supporting the work of MH scholars and domain experts in Chapter 5, in which we outlined the process that led us to develop tailored APIs to facilitate the Polifonia Textual Corpus' interrogation. We reported how we elicited and captured requirements to guide the development in *ad-hoc* meetings participated by professors and senior researchers in linguistics and members of WP1 and WP5, the work packages whose work focuses on the human interaction with MH.

The annotations produced and the APIs developed allow for taking advantage of the Polifonia Textual Corpus [1] for different purposes, from computational studies aimed at carrying out in-depth terminological analysis on certain MH-specific words senses or investigating the contexts in which certain named entities relevant for the MH, like Richard Wagner, may occur in a given textual genre corresponding to the Polifonia Textual Corpus' modules (as *Encyclopedic*, *Periodicals*, or *Books*).

The principal contributions of our work were to create a textual corpus of an unprecedented scale for languages and periods covered and to make it interrogable not only through structured queries based on morphosyntactic annotation, as most of the related work surveyed in Chapter 2 allow, but also by leveraging advanced semantic annotation such as word senses (cf. 4.3.3.5), linked named entities (cf. 4.3.3.6) and MH specific flags based on Polifonia Lexicon (cf. 4.3.4).

During the next months, we will concentrate on addressing the challenges and progressing with the future work outlined in Chapter 6. We will implement an *ad hoc* strategy to improve the quality of the text extracted by OCR technology from scanned documents to improve the performance of the state-of-the-art models used to produce linguistic annotations. We will design a strategy to allow the corpus' end-users to evaluate and correct the automatically produced annotations to gather gold-standard annotations and eventually evaluate the automatically produced ones. We will collect further requirements from linguistics and MH scholars, and we will use them to expand the interrogation APIs capabilities. We will design and develop a custom User Interface to favour access to the corpus to non-technical end-users. Finally, we will move ahead in our knowledge extraction work by performing sentence-to-AMR parsing of a subsample of the Polifonia Textual Corpus [1]. This work will lay the basis for the implementation of a new methodology for constructing knowledge graphs based on AMR parsing and will provide automatically extracted facts to integrate into the Polifonia knowledge graph.

[1] https://github.com/polifonia-project/Polifonia-Corpus/tree/master/annotations

8 FAIR Protocol

Our work in this chapter complies with the FAIR[1] principles and is in accordance with the Polifonia First Data Management Plan (D7.1).

The products released as a result of this work have been implemented following the FAIR data principles. The Polifonia Textual Corpus [1] data, metadata and statistics, along with its annotations and interrogation tools, are released under the CC BY license. We share the code that we used to construct, annotate and interrogate the corpus, along with the metadata of the corpus for findability and accessibility of its texts, through the dedicated Polifonia Corpus' GitHub repository[2]. The Polifonia Textual Corpus' GitHub repository has been released on Zenodo[3]. Its persistent URI (DOI) is `10.5281/zenodo.6772453`[4]. The DOI ensures the long-term persistence of the resource. The royalty-free parts of the Polifonia Textual Corpus' data, metadata and annotations have been uploaded on Zenodo and are therefore publicly accessible for download (links are reported in Chapter 3 and Chapter 4). Data and annotations of the Polifonia Textual Corpus modules subject to heterogeneous licensing are only available to the Polifonia consortium members. Interested parties may contact us, and we will evaluate the sharing of the annotated data.

The practices implemented to ensure FAIRness of the products released as a result of the work described in this report are described in the list below:

- **The Polifonia Textual Corpus.** The central data used in this research served to develop the annotated Polifonia Textual Corpus [1] (cf. Chapter 3). Regarding the development of the Polifonia Textual Corpus, we reported its FAIRness and Reproducibility details in D4.1 [2]. We collected data from different sources (i.e. Wikipedia and open digital libraries with Creative Commons[5] or similar licences[6]). The sources from which we downloaded the data allowed the non-commercial usage of their data. The texts obtained from digital libraries will not be shared publicly and will be used directly only by the Polifonia partners. The Polifonia Textual Corpus [1] data and metadata are hosted on public and private Zenodo repository and therefore some of them are publicly accessible for download other can be accessible only upon request. Links are reported in Chapter 3, Tables 3.1, 3.2, 3.4, 3.6, 3.8, 3.9.
- **Annotations of the Polifonia Textual Corpus.** For the annotation of the Polifonia Textual Corpus [1] (cf. Chapter 4) we used state-of-the-art Natural Language Processing technologies such as Spacy's[7] trained pipelines (cf. Table 4.3); EWISER[8]; ARES[9]; WSD-games[10]); ExTenD[11]. To ensure machine-readability and consistent encoding across different language we decided to encode the annotations following an open standard, namely the CoNLL-U format [12], designed within the Universal Dependency (UD) project [24]. An example of a sentence of the Polifonia Textual Corpus [1], annotated and encoded following the CoNLL-U format, is shown in Table 4.1 in Chapter 4. The Polifonia Textual Corpus [1] annotations are hosted on public and private Zenodo repositories. The links of the public repositories are reported in Chapter 4, Tables 4.4.
- **Interrogation of the annotated Polifonia Textual Corpus.** We designed and implemented custom software interfaces (APIs) to ensure an effective interrogation of the annotated Polifonia Textual Corpus [1] to allow for

[1] https://www.go-fair.org/fair-principles/
[2] https://github.com/polifonia-project/Polifonia-Corpus
[3] https://zenodo.org/
[4] https://doi.org/10.5281/zenodo.6772453
[5] https://creativecommons.org/licenses/by-nc-sa/3.0/it/deed.it.
[6] https://gallica.bnf.fr/edit/und/conditions-dutilisation-des-contenus-de-gallica
[7] https://spacy.io/
[8] https://github.com/SapienzaNLP/ewiser
[9] http://sensembert.org/
[10] https://github.com/roccotrip/wsd_games_emb
[11] https://github.com/SapienzaNLP/extend
[12] https://universaldependencies.org/format.html

performing structured queries on the corpus texts based on the annotations described in Chapter 4. Interrogation requirements have been collected by involving professors and senior researchers in the field of linguistics, as described in Section 5.1. We release the code to interrogate the corpus, along with detailed instructions to download the scripts and to satisfy its dependencies, in the dedicated section of the Polifonia Textual Corpus GitHub repository[13].

The resources described in the bullet list above are part of the Polifonia Ecosystem[14] and oblige the Ecosystem's Rulebook[15]. The Rulebook provides guidelines regarding how to track changes and progress issues, along with guidance regarding naming conventions, branches and releases. Compliance with the Rulebook guidelines on how to contribute to the Polifonia Ecosystem ensures the sustainability of the resource by providing tools for documenting and monitoring strategies for medium and long-term maintenance.

As outlined in Chapter 6, the future work on the products released as a result of the work described in this report will focus on stimulating the engagement of the interested personas, especially by expanding the corpus interrogation APIs capabilities and by developing a User Interface to make the corpus interrogation more user-friendly.

The data is a corpus, not a dataset in the terms of D7.1 Section 2.1.2, since it has not been converted to a Knowledge Graph or Linked Data format. However, this is envisaged as part of future work, as outlined in Section 6.3.

No personal or sensitive information is involved in this strand of research, so there are no security or privacy concerns.

[13] https://github.com/polifonia-project/Polifonia-Corpus/tree/master/interrogation
[14] https://polifonia-project.github.io/ecosystem/
[15] https://polifonia-project.github.io/ecosystem/rulebook/README.html

Bibliography

[1] R. Tripodi, A. Graciotti, and E. Marzi, "polifonia-project/polifonia-corpus: v0.1.3," Jun. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6772453

[2] UNIBO, "D4.1: Plurilingual corpora containing source texts in English, French, Spanish and German (v1.0)," December 2021.

[3] N. Islam, Z. Islam, and N. Noor, "A survey on optical character recognition system," *CoRR*, vol. abs/1710.05703, 2017. [Online]. Available: http://arxiv.org/abs/1710.05703

[4] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[5] M. Jakubícek, A. Kilgarriff, V. Kovár, P. Rychlý, and V. Suchomel, "The tenten corpus family," in *Corpus Linguistics*. Lancaster University, United Kingdom: UCREL, 2013, pp. 125–127.

[6] R. Billero and M. C. Nicolás Martínez, "Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue lbc," *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, vol. 4, no. 2, pp. 203–216, February 2018. [Online]. Available: https://revistas.uam.es/chimera/article/view/8779

[7] R. Billero, A. Farina, and M. C. N. Martínez, Eds., *I Corpora LBC*. Firenze University Press, 2020. [Online]. Available: https://doi.org/10.36253%2F978-88-5518-253-9

[8] H. Schmid, *Improvements in Part-of-Speech Tagging with an Application to German*. Dordrecht: Springer Netherlands, 1999, pp. 13–25. [Online]. Available: https://doi.org/10.1007/978-94-017-2390-9_2

[9] C. Strapparava, R. Mihalcea, and A. Battocchi, "A parallel corpus of music and lyrics annotated with emotions," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, N. Calzolari, K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. European Language Resources Association (ELRA), 2012, pp. 2343–2346. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2012/summaries/730.html

[10] O. Mitrofanova, "Probabilistic topic modeling of the russian text corpus on musicology," in *Language, Music, and Computing*, P. Eismont and N. Konstantinova, Eds. Cham: Springer International Publishing, 2015, pp. 69–76.

[11] J. M. Molina Mejía, "Hacia un análisis de la obra de richard wagner a través de la lingüística computacional," *Forma y Función*, vol. 32, no. 1, pp. 125–148, January 2019. [Online]. Available: https://revistas.unal.edu.co/index.php/formayfuncion/article/view/77419

[12] T. Pasini and J. Camacho-Collados, "A short survey on sense-annotated corpora," in *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, May 2020, pp. 5759–5765. [Online]. Available: https://aclanthology.org/2020.lrec-1.706

[13] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[14] G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker, "A semantic concordance," in *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993. [Online]. Available: https://aclanthology.org/H93-1061

[15] N. Ide, C. Baker, C. Fellbaum, and R. Passonneau, "The manually annotated sub-corpus: A community resource for and by the people," in *Proceedings of the ACL 2010 Conference Short Papers*. Uppsala, Sweden: Association for Computational Linguistics, July 2010, pp. 68–73. [Online]. Available: https://aclanthology.org/P10-2013

[16] A. Raganato, C. D. Bovi, and R. Navigli, "Automatic construction and evaluation of a large semantically enriched wikipedia," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016, p. 2894–2900.

[17] J. Camacho-Collados, C. Delli Bovi, A. Raganato, and R. Navigli, "SenseDefs: A multilingual corpus of semantically annotated textual definitions," *Language Resources and Evaluation*, vol. 53, no. 2, pp. 251–278, June 2019. [Online]. Available: https://doi.org/10.1007/s10579-018-9421-3

[18] J. A. Botha, Z. Shan, and D. Gillick, "Entity Linking in 100 Languages," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, November 2020, pp. 7833–7845. [Online]. Available: https://aclanthology.org/2020.emnlp-main.630

[19] D. Gillick, S. Kulkarni, L. Lansing, A. Presta, J. Baldridge, E. Ie, and D. Garcia-Olano, "Learning dense representations for entity retrieval," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, November 2019, pp. 528–537. [Online]. Available: https://aclanthology.org/K19-1049

[20] L. Procopio, E. Barba, F. Martelli, and R. Navigli, "Multimirror: Neural cross-lingual word alignment for multilingual word sense disambiguation," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Z.-H. Zhou, Ed. Montreal: International Joint Conferences on Artificial Intelligence Organization, August 2021, pp. 3915–3921, Main Track. [Online]. Available: https://doi.org/10.24963/ijcai.2021/53

[21] R. Smith, "An overview of the tesseract OCR engine," in *9th International Conference on Document Analysis and Recognition (ICDAR 2007)*. Curitiba, Brazil: IEEE Computer Society, September 2007, pp. 629–633. [Online]. Available: https://doi.org/10.1109/ICDAR.2007.4376991

[22] T. McEnery and A. Hardie, *Corpus Linguistics: Method, Theory and Practice*, ser. Cambridge Textbooks in Linguistics. Cambridge University Press, 2011.

[23] H. Thompson and D. McKelvie, "Hyperlink semantics for standoff markup of read-only documents," in *SGML Europe 97*. USA: Graphical Communications Association, 1997.

[24] J. Nivre, M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Hajič, C. D. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, and D. Zeman, "Universal Dependencies v1: A multilingual treebank collection," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 1659–1666. [Online]. Available: https://aclanthology.org/L16-1262

[25] J. J. Webster and C. Kit, "Tokenization as the initial phase in NLP," in *Proceedings of the 14th Conference on Computational Linguistics - Volume 4*, ser. COLING '92. USA: Association for Computational Linguistics, 1992, pp. 1106–1110. [Online]. Available: https://doi.org/10.3115/992424.992434

[26] D. Jurafsky and J. H. Martin, *Speech and Language Processing (2nd Edition)*. USA: Prentice-Hall, Inc., 2009.

[27] J. Eisenstein, *Introduction to Natural Language Processing (Adaptive Computation and Machine Learning series)*, illustrated ed. The MIT Press, 2019. [Online]. Available: http://gen.lib.rus.ec/book/index.php?md5=258080C1A3C31A7161236E4EEF4E7833

[28] R. Navigli, "Word sense disambiguation: A survey," *ACM Computing Surveys*, vol. 41, no. 2, February 2009. [Online]. Available: https://doi.org/10.1145/1459352.1459355

[29] M. Bevilacqua and R. Navigli, "Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 2854–2864. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.255

[30] B. Scarlini, T. Pasini, and R. Navigli, "With More Contexts Comes Better Performance: Contextualized Sense Embeddings for All-Round Word Sense Disambiguation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Online: Association for Computational Linguistics, 2020, pp. 3528–3539.

[31] R. Tripodi and R. Navigli, "Game theory meets embeddings: a unified framework for word sense disambiguation," 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)*.

Hong Kong, China: Association for Computational Linguistics, November 2019, pp. 88–99. [Online]. Available: https://www.aclweb.org/anthology/D19-1009

[32] E. Barba, L. Procopio, and R. Navigli, "ExtEnD: Extractive entity disambiguation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 2478–2488. [Online]. Available: https://aclanthology.org/2022.acl-long.177

[33] CNRS, "D1.1: Roadmap and pilot requirements (v1.0)," June 2021.

[34] KCL, "D2.1: Ontology-based knowledge graphs for music objects (v1.0)," December 2021.

[35] L. Lyu, M. Koutraki, M. Krickl, and B. Fetahu, "Neural OCR Post-Hoc Correction of Historical Corpora," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 479–493, May 2021. [Online]. Available: https://doi.org/10.1162/tacl_a_00379

[36] 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*. Sofia, Bulgaria: Association for Computational Linguistics, August 2013, pp. 178–186. [Online]. Available: https://aclanthology.org/W13-2322

[37] M. Bevilacqua, R. Blloshmi, and R. Navigli, "One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 14, pp. 12 564–12 573, May 2021. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17489

[38] A. Gangemi, V. Presutti, D. R. Recupero, A. G. Nuzzolese, F. Draicchio, and M. Mongiovì, "Semantic web machine reading with FRED," *Semantic Web*, vol. 8, no. 6, pp. 873–893, 2017. [Online]. Available: https://doi.org/10.3233/SW-160240