Polifonia: a digital harmoniser for musical heritage knowledge, H2020

D4.1: Plurilingual corpora containing source texts in English, French, Spanish and German (v1.0)

Deliverable information	
WP4	
Deliverable dissemination level	PU Public
Deliverable type	R Document, report
Lead beneficiary	UNIBO
Contributors	UNIBO, KNAW, OU
Document status	Draft (Univer review, Final)
Document version	v1.0
Date	December 23, 2021
Authors	Rocco Tripodi, UNIBO Eleonora Marzi, UNIBO Andrea Poltronieri, UNIBO Valentina Presutti, UNIBO Angelo Pombilio, UNIBO Antonella Luporini, UNIBO Peter van Kranenburg - KNAW Entico Daga - OU

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 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 new 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, understood 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 text-types and disciplines (memoire, reportage, news, biographies, reviews), different languages (English, Italian, French, Spanish, and German), and different centuries.

The overall goal of the project is to realise an ecosystem of computational methods and tools supporting the 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. Therefore, 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, music historians, anthropologists and ethnomusicologists, linguists, etc.

Executive Summary

The deliverable reports on the creation of multilingual corpora for Polifonia covering Italian, English, French and Spanish. Using the methodology of corpus linguistics, a corpus was designed that contains the sub-corpora corresponding to the different pilots. NLP techniques were used for the data collected. Finally, an additional, complementary resource was created - a multilingual aligned lexicon, essential for corpus creation and to guarantee interoperability between the textual resources.

The title of this report follows the one indicated in the GA, even though a more appropriate title for the report that reflects the additional work that we carried out would be: *Plurilingual corpora and lexicons containing source texts and concepts in Dutch, English, French, German, Italian and Spanish*.

Document History

Version	Release date	Summary of changes	Author(s) - Institution
V0.1	17/09/2021	Outline released	Rocco Tripodi, Eleonora Marzi - UNIBO
V0.2	16/11/2021	First review draft	Rocco Tripodi, Eleonora Marzi, Andrea Poltronieri, Angelo Pompilio - UNIBO
V0.3	23/11/2021	Structure of the report changed. Todos based on main comments from meeting with internal reviewers	Valentina Presutti - UNIBO
V0.4	14/12/2021	Integration of reviewers' comments	Rocco Tripodi, Eleonora Marzi, Andrea Poltronieri, Valentina Presutti - UNIBO
V1.0	22/12/2021	Final version submitted to EU	UNIBO

Table of contents

1 Introduction

This is the first report describing the work developed within WP4. This WP is in charge of supporting the project and its pilots as far as Musical Heritage (MH) knowledge extraction from text is concerned. To address this goal, a crucial preparatory development is needed and was planned as the focus of Task 4.1: Building and evaluating a multilingual corpus on musical heritage. This task, in turn, is addressed in two subsequent parts: (i) building a multilingual textual corpus - which is the focus of this report, (ii) supporting its interrogation and annotation for discourse analysis - which will be delivered at M18.

In undertaking the creation of the Polifonia textual corpus, we realised two additional needs that had not been originally planned: the creation of a MH-specialised lexicon, and the coverage of the Dutch language, which is key to the pilot ORGANS.

This report describes the first versions of the Polifonia lexicon and of the Polifonia textual corpus, and the methodologies we followed for their development. Both resources cover six languages - Dutch, English, French, German, Italian and Spanish - and focus on MH as domain of knowledge. The textual corpus contains texts published from 1700 to present and covering a large time span in terms of MH history, approximately from 1400 to present. Additional texts may be added in the next period as result of further analysis of the pilots' needs. We are releasing the Polifonia lexicon under the CC BY license. The Polifonia textual corpus is released under the same license as a set of metadata that allow 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. Both resources are available on the Polifonia GitHub repository[1]. We remark that they are living and evolving creatures: their current versions (along with the figures provided in this document) will be replaced in the next months by larger and more robust ones[2]. We expect further evaluation and enrichment from the feedback-improvement loop, deriving from the interaction with the other WPs, especially WP1 (i.e. pilots), WP2 (ontologies and knowledge graphs) and WP4.

1.1 The Polifonia Lexicon

The Polifonia lexicon is a linguistic resource representing the MH-specific terminology and concepts in six languages, organised in the style of WordNet [1], as sense-equivalent classes called synsets, each associated with its lexicalisations. The construction of the Polifonia lexicon is motivated by several applications. It is used to build the Polifonia textual corpus (cf. Chapter 4.5): it guides the selection of textual sources to make sure that the corpus shows a wide, possibly complete coverage of the MH terminology. The lexicon has high potential impact in enabling the development of cutting edge Music Information Retrieval and MH-specific knowledge extraction (from text) tools. The lexicon will play a key role in supporting the interrogation of the Polifonia textual corpus, based on concepts rather than keywords (cf. Deliverable 4.2 due at M18). The lexicon can be used to categorize musical objects such as scores that contain specific features (such as harmony, melody, or counterpoint), which are defined as concepts in the lexicon. Furthermore, it provides valuable background knowledge for addressing tasks such as entity linking, word-sense disambiguation, and relation extraction tailored to the MH discourse (cf. Deliverable 4.3-6 due at M18/M30). The lexicon may be a key resource in supporting the analysis of music terminology and its evolution over time, languages, and space.

The Polifonia lexicon in its current status includes 3.869 interlinked concepts (i.e. synsets), each associated with an average of 3.8 lexicalisations, for a total of 76.250 terms. It describes the MH domain in all its aspects: from the terms

[1] Polifonia Lexicon https://github.com/polifonia-project/Polifonia-Lexicon, Polifonia Textual Corpus https://github.com/polifonia-project/Polifonia-Corpus

[2] The evolution of the resources can be monitored on the GitHub repository.

used to describe musical forms such as *sonata*, *sinfonia*, *hymn*, *anthem*, *litany*, to terms used to describe musical supports such as *audio-tape*, *cassette deck*, *audio system*, *sound system*.

We claim that the methodology for creating the Polifonia lexicon (cf. Chapter 3.2.4) is a contribution *per se*, as it is generalisable to other domains. It contributes to advancing the state of the art of semi-automatic creation of domain-specific lexical resources. There are valuable examples of methods for lexicon enrichment tailored to specific tasks, such as SENTIWORDNET [2] and WORDNET-AFFECT [3], and BabelNet [4]. They enrich existing synsets with additional information, rather than adding new synsets and terms, while the Polifonia lexicon intends to extend existing lexicons by creating new synsets based on MH-specialised resources, such as dictionaries and classification systems. Figure 1.1 summarises the methodology for building the Polifonia lexicon by showing a running example

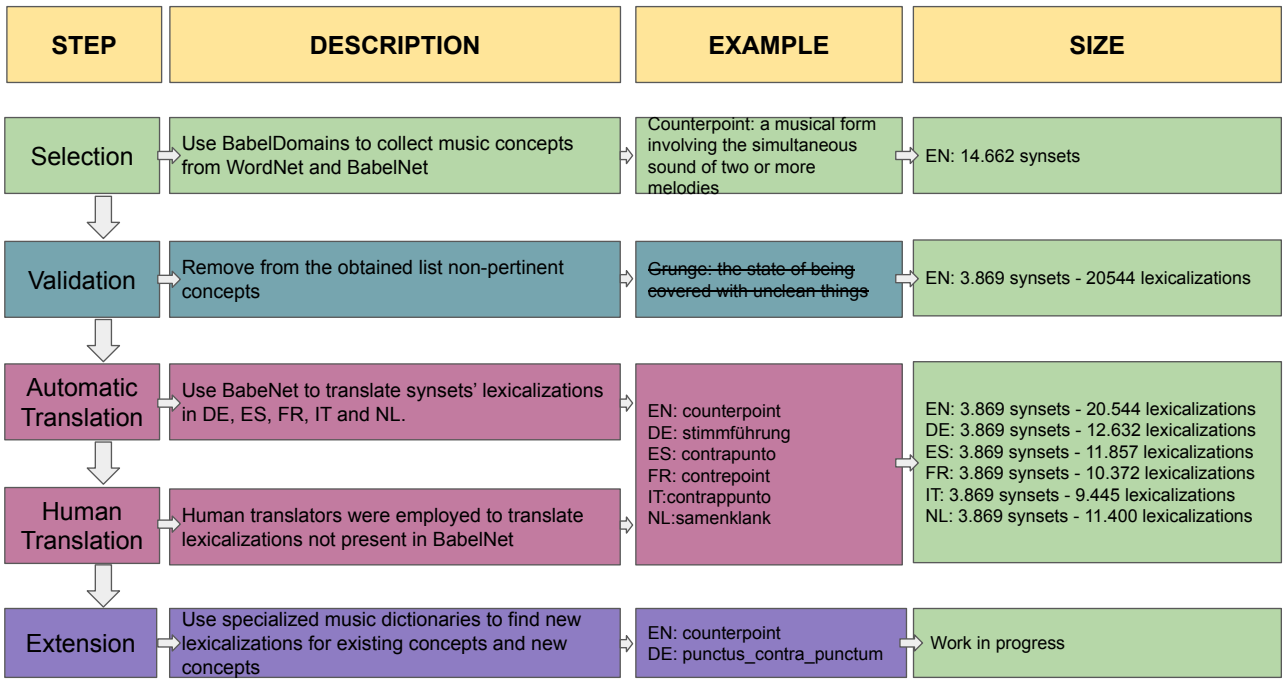


Figure 1.1: A description of the methodological steps carried out for the development of the Polifonia lexicon.

along with the results at each step.

The Polifonia lexicon is built on top of a selected subset from existing lexicons (Selection), such as WordNet [1] and BabelNet [4]. This subset is automatically extracted by using BabelDomains [5] to collect all concepts and terms that are relevant for the MH-domain, focusing on the English language (which is the only one addressed by BabelDomains). The result of the automatic procedure is validated by expert linguists (Validation), who follow specific guidelines designed with the support of expert musicologists. Their task is to identify possible intruders, i.e. terms that are not MH-specific. All terms and concepts included in the lexicon are then translated (Translation) into the other languages addressed by the lexicon, in our case: Dutch, French, German, Italian and Spanish. The translation is done automatically when provided by BabelNet, and manually otherwise. An additional validation activity is performed to evaluate the translations. Currently, we are performing an activity for extending the lexicon to include terms (and concepts) that are not present in either WordNet or BabelNet (Extension). To this end, we are using MH-specialistic vocabularies and terminologies, recommended by expert musicologists: the UNIMARC Bibliographic Resource[3], and the *Terminorum Musicae Index Septem Linguis Redactus* [6]. We use the Polifonia lexicon to select the text sources to be included in the Polifonia textual corpus.

[3]UNIMARC: Form of musical work. Available at https://www.iflastandards.info/unimarc/terms/fom.

1.2 The Polifonia Textual Corpus

The Polifonia textual corpus addresses six languages: Dutch, English, French, German, Italian and Spanish. It covers a large portion of the European modern and contemporary history, approximately from 1400 to the present day, including more than 100.000 wikipedia articles, 100.000 books and 50.000 periodical issues. The corpus collects and homogeneously organises diachronic and multilingual texts from different types of sources having MH topics as their main subjects: encyclopedias, dictionaries, books (historical treatises, monographs, critical essays and diaries), periodicals and newspapers. Such heterogeneity of source types gives the corpus a broad representativeness of the discourse about music: from a general encyclopedic perspective to specialized ones such as reviews and essays, conveyed through books and journal articles.

The creation of this corpus is motivated by the importance of providing scholars, researchers and music professionals with a resource that they can query to support their work, as well as by the need to provide the project (especially WP4) with a representative large body of knowledge for extracting facts about MH. These facts are to be injected in the Polifonia KG and its associated ontologies (cf. Deliverable 2.1), to address MH knowledge use and preservation - some of the core goals of Polifonia. To the best of our knowledge, the Polifonia textual corpus is one of a kind. Existing corpora on music focus on specific aspects of the domain e.g. a specific music genre[7] or e.g. a specific time period [8] or language [9]; the Polifonia corpus instead encompasses all these aspects.

Five of the ten Polifonia pilots require and rely on knowledge extraction from text: MUSICBO, MEETUPS, BELLS, CHILD, and ORGANS. They provide the main requirements to the construction of the textual corpus: text types, period to be covered, recommended sources from which to retrieve the material, specific texts that must be included, the languages to be covered. These requirements have been collected through a Survey submitted to the pilot leaders, and by continuously interacting with the pilot teams (cf. Section 4.4). To support the extraction of pilot-specific sub-corpora, the Polifonia textual corpus is organised in three modules based on source type: Wikipedia, books, and periodicals. One of the challenges we faced during the collection of sources is that a significant part of them was only available as images or pdf files. We need to convert them in a processable format. The state of the art technology that addresses this problem is Optical Character Recognition (OCR)[10]. Nevertheless, existing tools are unable to handle the scale and heterogeneity that characterises our sources. Therefore, we implemented a piece of software based on Tesseract [11] to tackle these issues (cf. 4.3.1.1)[4]. Currently, we have collected all books and periodicals and we are performing the OCR process on them. The result of this process will contribute to enrich the *Books* and *Periodicals* modules of the textual corpus.

The Polifonia textual corpus will be the basis for the development of tools for MH-knowledge extraction from text. The encyclopedic module and annotated parts of the other modules of the Polifonia textual corpus will be used to develop training sets for machine learning algorithms, specialized in the extraction of MH knowledge from text (WP4): Tasks 2 "Automatic extraction of time, space, events, people and musical artifacts from text" and Task 3 "Automatic extraction of socio-cultural and historical context of musical heritage". The books and periodicals modules will be the main target of MH knowledge extraction algorithms, whose output will be injected in the Polifonia knowledge graph.

Figure 1.2 depicts an extract from *The Harmonicon* (March, 3, 1823), an important English periodical included in the Polifonia corpus, recounting Mozart's childhood. The text is annotated with different colors highlighting examples of information that can be extracted from the Polifonia corpus: they are terms that refer to different entities, each color indicating a different entity type. For example, the term *England* is highlighted in blue because it refers to the country England, which is a location. The term *1766* is highlighted in yellow because it refers to a date (a year in this case).

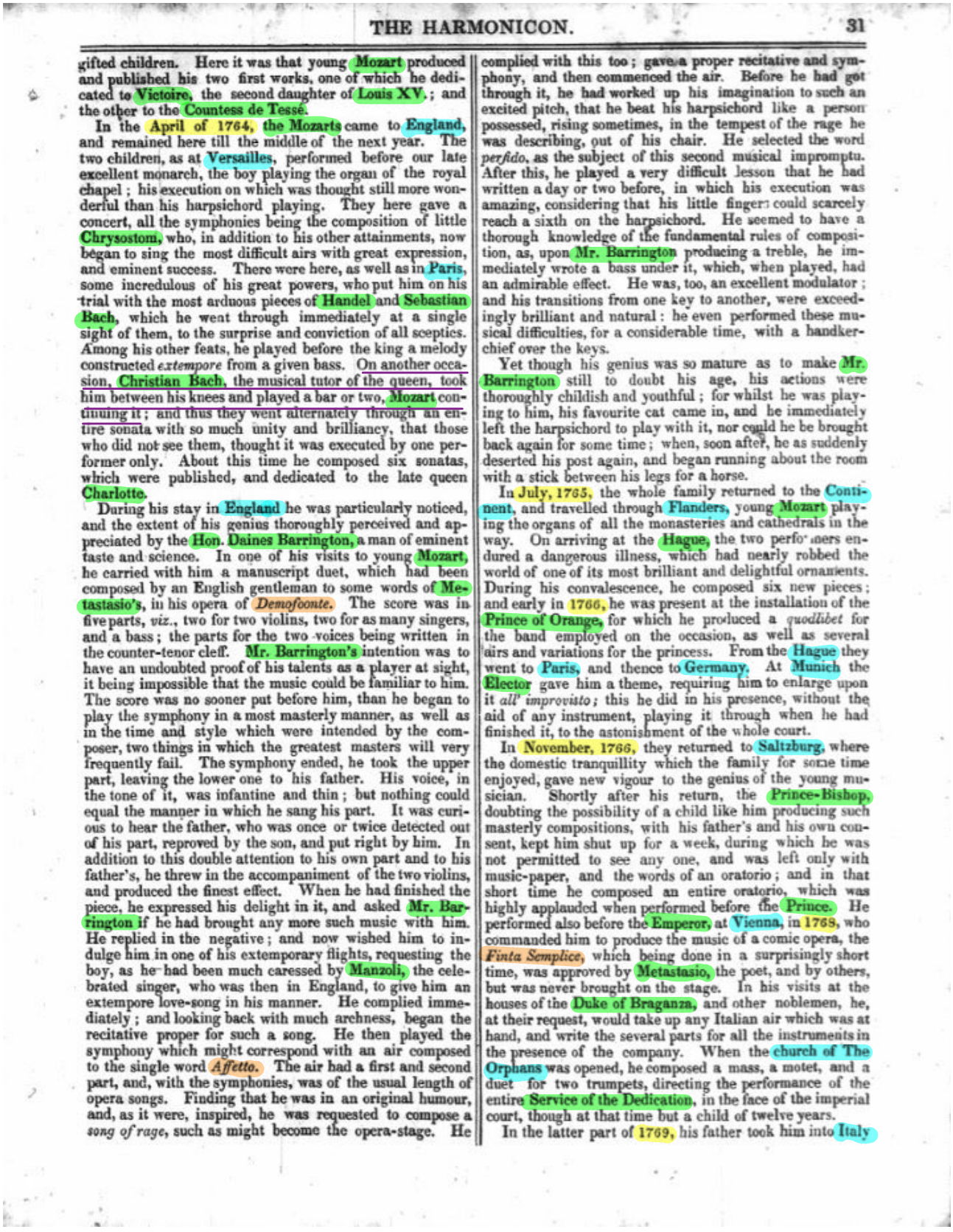
All these terms will be linked to entities (the ones they refer to) in the Polifonia knowledge graph. If they are not present, they will be used to enrich it to support the extraction of facts that are relevant to the Polifonia pilots and, more generally, to the MH domain. Such facts will be used in turn to enrich the Polifonia knowledge graph. For example, in the text depicted in Figure 1.2 we find that the comic opera *Finta Semplice* was composed by Mozart. A fact that the Polifonia knowledge extraction tools will recognise as an occurrence of a concept/synset e.g. *to compose*, that is in turn formally represented in the Polifonia Ontology e.g. by a class `:CompositionSituation` or an object property `:composes`. The result will be a triple (or a set of triples) formalising this fact and injected in

[4]. The code is openly available at: https://github.com/polifonia-project/textual-corpus-population

the Polifonia KG to enrich it, while maintaining a link to the original information source: in this case, *The Harmonicon*. Similarly, and showing a more complex challenge, we can infer that Mozart and Christian Bach met, during Mozart's adolescence, and played together, from the sentence "*On another occasion, Christian Bach, the musical tutor of the queen, took him between his knees and played a bar or two, Mozart continuing it;*"[5] - a fact that is relevant for the MEETUP pilot. To effectively perform knowledge extraction from text, we need a very large textual corpus. The most common approaches to building general textual corpora aim at representing a particular language variety, e.g. 1990s English, radiophonic Italian, Italian spoken in Canada, which is well addressed by using sampling procedures [12]. By contrast, the methodology we apply to build the Polifonia textual corpus aims at preserving as many language varieties as possible, by including as many subcultures and sub-genres as possible, using a large set of data sources. Sampling procedures, in our case, would exclude resources that may be relevant to MH scholars. This aspect is clarified in Section 4.1 that shows how e.g. Wikipedia pages in certain languages represent peculiar aspects of the culture associated with that language. The criteria followed to select the texts to be included in the corpus are therefore external[6] [12]: they are tailored to include resources that contain expressions from the (large) Polifonia lexicon. Section 4.5 describes the vocabulary of the corpus, showing that the corpus is *saturated* at lexical level [13], i.e. if we include additional texts, the vocabulary of the corpus is affected only by negligible modifications, meaning that the curve of lexical growth has become asymptotic. The notion of saturation is considered to be superior to balancing techniques for its measurability [14].

[5] highlighted in purple

[6] External criteria are non-linguistic ones, e.g. we do not select only texts that have particular linguistic features, but select document types e.g. articles and books.



THE HARMONICON. 31

gifted children. Here it was that young Mozart produced and published his two first works, one of which he dedicated to Victoire, the second daughter of Louis XV.; and the other to the Countess de Tessé.

In the April of 1764, the Mozarts came to England, and remained here till the middle of the next year. The two children, as at Versailles, performed before our late excellent monarch, the boy playing the organ of the royal chapel; his execution on which was thought still more wonderful than his harpsichord playing. They here gave a concert, all the symphonies being the composition of little Chrysostom, who, in addition to his other attainments, now began to sing the most difficult airs with great expression, and eminent success. There were here, as well as in Paris, some incredulous of his great powers, who put him on his trial with the most arduous pieces of Handel and Sebastian Bach, which he went through immediately at a single sight of them, to the surprise and conviction of all sceptics. Among his other feats, he played before the king a melody constructed *extempore* from a given bass. On another occasion, Christian Bach, the musical tutor of the queen, took him between his knees and played a bar or two, Mozart continuing it; and thus they went alternately through an entire sonata with so much unity and brilliancy, that those who did not see them, thought it was executed by one performer only. About this time he composed six sonatas, which were published, and dedicated to the late queen Charlotte.

During his stay in England he was particularly noticed, and the extent of his genius thoroughly perceived and appreciated by the Hon. Daines Barrington, a man of eminent taste and science. In one of his visits to young Mozart, he carried with him a manuscript duet, which had been composed by an English gentleman to some words of Metastasio's, in his opera of *Demofoonte*. The score was in five parts, *viz.*, two for two violins, two for as many singers, and a bass; the parts for the two voices being written in the counter-tenor cleff. Mr. Barrington's intention was to have an undoubted proof of his talents as a player at sight, it being impossible that the music could be familiar to him. The score was no sooner put before him, than he began to play the symphony in a most masterly manner, as well as in the time and style which were intended by the composer, two things in which the greatest masters will very frequently fail. The symphony ended, he took the upper part, leaving the lower one to his father. His voice, in the tone of it, was infantine and thin; but nothing could equal the manner in which he sang his part. It was curious to hear the father, who was once or twice detected out of his part, reproved by the son, and put right by him. In addition to this double attention to his own part and to his father's, he threw in the accompaniment of the two violins, and produced the finest effect. When he had finished the piece, he expressed his delight in it, and asked Mr. Barrington if he had brought any more such music with him. He replied in the negative; and now wished him to indulge him in one of his extemporary flights, requesting the boy, as he had been much caressed by Manzoli, the celebrated singer, who was then in England, to give him an extempore love-song in his manner. He complied immediately; and looking back with much archness, began the recitative proper for such a song. He then played the symphony which might correspond with an air composed to the single word *Affetto*. The air had a first and second part, and, with the symphonies, was of the usual length of opera songs. Finding that he was in an original humour, and, as it were, inspired, he was requested to compose a *song of rage*, such as might become the opera-stage. He complied with this too; gave a proper recitative and symphony, and then commenced the air. Before he had got through it, he had worked up his imagination to such an excited pitch, that he beat his harpsichord like a person possessed, rising sometimes, in the tempest of the rage he was describing, out of his chair. He selected the word *perfido*, as the subject of this second musical impromptu. After this, he played a very difficult lesson that he had written a day or two before, in which his execution was amazing, considering that his little fingers could scarcely reach a sixth on the harpsichord. He seemed to have a thorough knowledge of the fundamental rules of composition, as, upon Mr. Barrington producing a treble, he immediately wrote a bass under it, which, when played, had an admirable effect. He was, too, an excellent modulator; and his transitions from one key to another, were exceedingly brilliant and natural: he even performed these musical difficulties, for a considerable time, with a handkerchief over the keys.

Yet though his genius was so mature as to make Mr. Barrington still to doubt his age, his actions were thoroughly childish and youthful; for whilst he was playing to him, his favourite cat came in, and he immediately left the harpsichord to play with it, nor could he be brought back again for some time; when, soon after, he as suddenly deserted his post again, and began running about the room with a stick between his legs for a horse.

In July, 1765, the whole family returned to the Continent, and travelled through Flanders, young Mozart playing the organs of all the monasteries and cathedrals in the way. On arriving at the Hague, the two performers endured a dangerous illness, which had nearly robbed the world of one of its most brilliant and delightful ornaments. During his convalescence, he composed six new pieces; and early in 1766, he was present at the installation of the Prince of Orange, for which he produced a *quodlibet* for the band employed on the occasion, as well as several airs and variations for the princess. From the Hague they went to Paris, and thence to Germany. At Munich the Elector gave him a theme, requiring him to enlarge upon it *all' improvisto*; this he did in his presence, without the aid of any instrument, playing it through when he had finished it, to the astonishment of the whole court.

In November, 1766, they returned to Saltzburg, where the domestic tranquillity which the family for some time enjoyed, gave new vigour to the genius of the young musician. Shortly after his return, the Prince-Bishop, doubting the possibility of a child like him producing such masterly compositions, with his father's and his own consent, kept him shut up for a week, during which he was not permitted to see any one, and was left only with music-paper, and the words of an oratorio; and in that short time he composed an entire oratorio, which was highly applauded when performed before the Prince. He performed also before the Emperor, at Vienna, in 1768, who commanded him to produce the music of a comic opera, the *Finta Semplice*, which being done in a surprisingly short time, was approved by Metastasio, the poet, and by others, but was never brought on the stage. In his visits at the houses of the Duke of Braganza, and other noblemen, he, at their request, would take up any Italian air which was at hand, and write the several parts for all the instruments in the presence of the company. When the church of The Orphans was opened, he composed a mass, a motet, and a duet for two trumpets, directing the performance of the entire Service of the Dedication, in the face of the imperial court, though at that time but a child of twelve years.

In the latter part of 1769, his father took him into Italy

Figure 1.2: A simple example of knowledge extraction from the corpus in which different entity types are depicted with different colours: dates in yellow, people in green, locations in light blue and compositions in orange.

1.3 Contribution and Structure of the Document

The main contributions of this report are:

- a specialised lexicon for the musical heritage domain
- a method for semi-automatic creation of domain-specific lexicons
- a multilingual, diachronic textual corpus specialised on the musical heritage domain

This report is organised as follows. In the next section we give on overview of relevant related work, focusing on domain specific corpora, music-related corpora and lexicons. Section 3 describes the Polifonia lexicon, while Section 4.5 presents the Polifonia textual corpus (and all its modules). Section 5 is dedicated to discuss the main challenges of this work and on how we plan to use the resources developed so far. Conclusions are given in Section 6.

2 Background and Related work

To create the Polifonia textual corpus and the Polifonia lexicon, we used different lexical and ontological resources. These resources, which allowed us to select and organize concepts and textual material about music, will be presented in the next sections.

2.1 WordNet

WordNet[1] [1] is one of the largest and most widely used lexical resources for English (the original paper [1] has over 13,000 citations). It is a general-purpose resource covering concepts of different semantic domains.

WordNet is structured as a graph, whose nodes are *synsets*, i.e., sets of different words (nouns, verbs, adjectives and adverbs) with the same meaning and part of speech, while edges represent conceptual-semantic and lexical relations between synsets. These relations include antonymy (opposite), hyponymy (subordinate), meronymy (part), troponymy (manner), entailment (drive – ride) and derivation (magnetic – magnetism). WordNet 3.0 contains 117,000 synsets in English. WordNet is the basis of many other lexical resources such as BabelNet [4] and the more recent WordNet 2020 [15] that extends and corrects the original resource. WordNet, originally developed only for English, was extended to address many other languages [16].[2] This makes it a perfect starting point for the Polifonia lexicon that aims at supporting multilinguality.

2.2 BabelNet and BabelDomains

BabelNet[3] [4, 17] is a popular lexical-semantic knowledge resource obtained by merging heterogeneous sources such as WordNet, Wikipedia[4] and WikiData[5] [18] into a unified semantic network that, thanks to its multilingual nature, helps to scale tasks and applications to hundreds of languages.

BabelDomains[6] [5] is a unified resource which provides lexical items included in different lexical resources (BabelNet, Wikipedia and WordNet) with information about domains of knowledge. Each synset is associated with a pre-defined domain of knowledge, such as *mathematics*, *biology*, *history*, *education* or *music*. These domains have been selected from the Wikipedia featured articles page.[7] To associate synsets with domains, an automatic hybrid distributional and graph-based method has been developed. This approach allows to classify synsets constructing and evaluating vector representations for synsets and domains. The last version of BabelDomains has been integrated in BabelNet, both into the online interface and in the API. The annotation of lexical items in this resource is automatic and covers thirty-two domains of knowledge.

2.3 Wikipedia

Wikipedia, is a Web-based collaborative encyclopedia. A Wikipedia page presents knowledge about a specific concept (e.g. polyphony) or named entity (e.g. Frank Zappa). The page typically contains hypertext linked to

[1] https://wordnet.princeton.edu
[2] http://compling.hss.ntu.edu.sg/omw/
[3] https://babelnet.org
[4] http://wikipedia.org/
[5] https://www.wikidata.org
[6] http://lcl.uniroma1.it/babeldomains/
[7] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

other relevant pages. The title is composed of the lemma of the concept to be defined plus an optional label in parentheses which specifies its meaning if the lemma is ambiguous. A page also provides inter-language links to a lemma's counterparts in other languages.

2.4 Wikidata

Wikidata [18] is a free and open knowledge base that can be read and edited by both humans and machines. It is document-oriented and focuses on items, which represent any kind of topic, concept, or object. Each item is allocated a unique, persistent identifier, a positive integer prefixed with the upper-case letter Q, known as a "QID". An example of a WikiData item is provided in Figure 2.1, presenting an extract of the information about the composer Erik Satie. From this page we can obtain various information on the French composer; in particular, we can see that he was an avant-garde musician, that he studied at the Conservatoire de Paris and that he was a student of Vincent d'Indy. Wikidata acts as central storage for the structured data of many other projects. For this reason, we decided to use information included in this resource to reference the entities in the Polifonia corpus (see Section 4.1) and in the Polifonia knowledge base (cf. Deliverable 2.1).

Figure 2.1: Extract of the WikiData page of Erik Satie (Q187192).

2.5 Music Textual Corpora

According to [19] and from an analysis of the resources collected by the International Society for Music Information Retrieval (ISMIR) [8], existing musical corpora or datasets mostly address three categories: sounds audio files, biographical information about performers, composer, album or songs, and lyrics considered as a textual type. These three categories often overlap, i.e. an audio dataset is annotated with biographical categories or metadata such as artist, time, genre and language [20]. The Polifonia textual corpus must support the analysis of musical heritage knowledge in a broad sense: its social, historical and cultural context involves text-types that go beyond lyrics or author's biographical information, dealing with musical performances, compositions and receptions while also providing a multilingual dimension.

Context-related musical information also concerns critical receptions in a historical perspective. In this respect, there are corpora that include music reception during a specific historical period. To the best of our knowledge, there is

[8] https://ismir.net/

a shortage of corpora about music reception and criticism, and existing resources are focused on particular case studies such as [21], which develops a method for extracting information from a corpus collecting one hundred critical reviews of Beethoven's piano sonata recordings, published in the *Gramophone* between August 1934 and July 2010. Another example is the case of the book *Music Criticism in France, 1918-1939: Authority, Advocacy, Legacy* [22], which focuses on the years between the two World Wars; it includes texts on music criticism with the aim of understanding the musical landscape in France. Other attempts to collect texts about music are related to the development of bibliographies that include mainly pointers to specialist resources,[9] or initiatives that make available specialist resources (mainly periodicals) only to subscribers.[10] We notice how, considering the discourse about music in general, corpora are very focused: general knowledge is missing. Examples of corpora of music lyrics as a textual genre are an excellent starting point for a state-of-the-art analysis in two perspectives: what kind of information can be extracted from songs, and what aspects related to multilingualism have been dealt with. A special focus on the emotions that are aroused song lyrics is evident through examples such as MoodyLyrics[23] where Corpus is annotated with emotion categories from Russell's model of valence-arousal. The interest in annotated lyrics corpus is not limited to sentiment analysis, as shown by the Wasabi Song Corpus[11] [24] that includes 1.73 million songs with lyrics and provides annotations of structure segmentation, topics, explicit lyrics content, salient passages of a song and the emotions conveyed. The project Songkorpus[12] [8] considers lyrics as having features of both written and spoken discourse that can provide scholars in various disciplines with extensive linguistic and cultural information. This perspective leads to further annotation of three types of data: TEI-compliant song lyrics as primary data, linguistically and literary motivated annotations, and extra linguistic metadata. Corpora of lyrics could be considered under different perspectives for extracting information, from context to discourse analysis. However, while maintaining the focus on contextual information, the texts that Polifonia wants to include in its corpus are not only those related to songs. With reference to multilingualism, a review of state-of-the art literature shows that multilingual corpora in music are missing. In order to tackle multilingualism in music some cutting edge analyses have been carried out; these include[25] [26], focusing on cultural differences in human perception of music genre.[13] In the context of the study of music, the multilingual aspect is very important because it links up to cultural issues; that music is intrinsically linked to culture is unanimously accepted, as also shown by the Compmusic project[14], which focuses on five music traditions from around the world: Hindustani (North India), Carnatic (South India), Turkish-makam (Turkey), Arab-Andalusian (Maghreb), and Beijing Opera (China). The aim of this project is twofold: to contribute to the automatic description of music based on cultural specificity; to research the field of music information processing through a domain knowledge approach. The five corpora include Hindustani[15] Carnatic[16], Turkish-makam[17] Beijing Opera [18], Arab-Andalusian[19] audio recordings and complementary information that describes the recordings. Projects such as The natural history of song[20] concentrate on the formality, arousal, and religiosity of song events through the construction of a corpus of ethnographic texts on musical behavior from a representative sample of world's societies as well as a discography of audio recordings of music itself.

This analysis of the state of art of multilingual music corpora seems to reveal that the sociocultural aspects of music in an intercultural perspective have received little attention: though some experiments have been carried out, these do not present an original framework. The Polifonia corpus aims at filling this gap by providing a multicultural and multilingual corpus that ranges across different text types, countries, cultures and historical periods.

[9] See for example the Bibliography of Music Literature initiative https://www.musikbibliographie.de.
[10] See for example the Répertoire International de Littérature Musicale https://www.rilm.org.
[11] https://github.com/micbuffa/WasabiDataset,
[12] http://songkorpus.de/
[13] https://github.com/deezer/MultilingualMusicGenreEmbedding;
[14] https://compmusic.upf.edu/;
[15] dunya.compmusic.upf.edu/hindustani
[16] dunya.compmusic.upf.edu/carnatic
[17] dunya.compmusic.upf.edu/makam
[18] dunya.compmusic.upf.edu/jingju
[19] dunya.compmusic.upf.edu/andalusian
[20] https://osf.io/jmv3q/

2.5.1 Construction of Linguistic Corpora

A corpus is defined as a large digital collection of texts which is a sample of a language or a linguistic variety. In the first case it aims at representing language in general, as for example in the Brown Corpus [21]. This includes one million words sampled from 15 different text types; it was the first of the modern, computer readable, general corpora and aims to be a reference for American English. Its counterpart in British English is the Lancaster-Oslo/Bergen Corpus[22], also including one million words sampled from 15 different text types. A specialized corpus, by contrast, aims to focus on specialized discourse or a specific topic, as for example in the Press Corpora: Clarin *ERIC* (The research infrastructure for language as social and cultural data) [23] has a special section dedicated to Newspapers with 34 newspaper corpora, 7 of which are multilingual and 27 monolingual [24]. The construction of a corpus consists of two phases which cannot be separated: 1) design and 2) implementation. The design is based on criteria that depend on the intended use of the corpus (general or specialized); furthermore, what is available and achievable has to be considered. The processes of design and implementation are in constant dialogue and allow the parameters to be adjusted, always keeping track of the content of the corpus. The web is nowadays an important source for collecting material: on the one hand, it functions as a resource portal which, thanks to its numerous libraries, can be exploited for accurate research. On the other hand, the web lends itself as a container for free materials or knowledge archives such as Wikipedia, which are often included in corpora. The Polifonia textual corpus is a domain specific corpus which follows a construction design based on its requirements 4.5 which led to the need to identify three modules: 4.1 for encyclopedic knowledge, 4.3 for more specific context-related knowledge and 4.2 for historical and critical knowledge.

2.6 Music Specialised Lexical Resources

In recent years, music-related vocabulary resources have also been developed. These resources include both music-specific vocabularies and thesauri. In particular, these types of resources focus on specific aspects of the musical domain, providing lists of terms that can be understood to describe a narrow part of the musical domain. An example of such resources is the MusicBrainz Instrument Vocabulary[25] or the Musical Instrument Taxonomies developed by Queen Mary University of London in the context of the Isophonics project[26]. Also concerning musical instruments, some lexical resources have been developed for catalographic purposes and are widely used in museum contexts, such as the world standard SKOS vocabulary for organology [27] [27], which is based on the Hornbostel-Sachs classification. There are also ontological models specifically modelled to describe different types of musical resources. For example, some ontologies describe high-level music-related information (e.g. [28, 29]), while others describe musical theoretical concepts (e.g. [30, 31, 30]), and specific aspects of the musical domain, such as the Chord, Tonality, Temperament, and Segments (e.g. [32, 33]). A complete overview of these resources is proposed in the WP2 Deliverable [34]. These types of resources are of great interest for the activities of the Work Package, as they can be used to describe and model the entities identified within the textual resources.

2.7 Linguistic and Ontological Resources

There are also models for modelling textual resources and aligning the semantics of different vocabularies. These resources are generally based on the integration of Natural Language Processing (NLP) techniques and the Semantic Web. In general terms, these approaches have been described under the term Linguistic Linked Data [35],

[21] https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html
[22] http://korpus.uib.no/icame/manuals/LOB/INDEX.HTM
[23] https://www.clarin.eu/
[24] https://www.clarin.eu/resource-families/newspaper-corpora
[25] MusicBrainz Instrument Vocabulary https://musicbrainz.org/instruments
[26] Isophonics Musical Instrument Taxonomies http://www.isophonics.net/content/musical-instrument-taxonomies
[27] MIMO Vocabulary http://www.mimo-db.eu/

a field of research that concerns the representation of language and its representation in the form of Linked Data. An excellent example of this kind of resource is Framester [36], a frame-based ontological resource that provides interlink between existing linguistic resources. More secifically, Famester acts as a hub between FrameNet [37], BabelNet [4], VerbNet [38], DBpedia [39], Yago [40], and other resources, by interpreting their semantics as a subset of Fillmore's semantics. Other approaches are based on modelling lexical resources by reusing other ontologies, such as CIDOC-CRM, FRBRoo, and Ontolex-Lemon [41].

2.8 Methods for Building Wordnets

There are two main approaches for building wordnets [42], 1. the merge approach – that requires a set of words annotated with sense information to then create synsets with words referring to the same meaning; 2. the expansion approach – that requires existing synsets from a reference WordNet to be used as a guide to create corresponding synsets in new WordNets, by gathering applicable words. These two approaches have been used mainly to construct WordNets for new languages [43, 44].

Other projects such as SENTIWORDNET [2] and WORDNET-AFFECT [3] extended WordNet by including new information into its synsets. In SENTIWORDNET each synset is associated with three numerical scores which indicate how positive, negative, and "objective" (i.e., neutral) the terms contained in the synset are. In WORDNET-AFFECT WordNet synsets are tagged by means of a taxonomy of affective categories (e.g. Behaviour, Personality, Cognitive state). Approaches of this kind do not add or remove synsets from WordNet; only new information is attached to the resource.

In our work we used all these approaches. We started from WordNet synsets, mapped them to other languages, included semantic domain information and we are currently analysing the resources to add new synsets and lexicalizations to WordNet.

3 The Polifonia Lexicon

In computational linguistics, a lexicon, or lexical resource, is a language resource consisting of a list of lexemes[1] in which each lexeme is associated with information such as part of speech, synonyms, sense definitions, related terms and related concepts [45]. Such information is organized in a machine readable format and thus can be used in computational models and applications. Lexical resources are being widely used to disambiguate the meaning of words, associating lexical forms (words or phrases) to concepts (meanings) [46]. For example, the term *beat* has several meanings depending on the context in which it is used, and can be a noun, a verb or an adjective. As a noun it can mean *pulse* according to the definition: "*the rhythmic contraction and expansion of the arteries with each beat of the heart*". Another meaning for *beat* is *musical rhythm*, according to the definition "*the basic rhythmic unit in a piece of music*". As a verb its meaning can be *vanquish* defined as "*to come out better in a competition, race, or conflict*" or *drum* defined as "*to make a rhythmic sound*". Therefore, a lexicon can be used to distinguish among the different meanings of a word, making it possible to study how words take their meaning in particular semantic domains and how these meanings are related [47]. A particular kind of lexicon, (lexical semantic networks or wordnets), organises concepts (groups of terms referring to the same meaning) through a network of semantic relations, such as hyper/hyponymy (tree-oak), meronymy (tree-branch), antonymy (long-short) and various entailment relations (buy-pay, show-see, untie- tie) [48]. For example, *sonata*[2], *fantasia*[3], *bagatelle*[4], and *suite*[5] are defined as more specific concepts than *musical composition* (hypernym relation).

Building a lexicon for the Polifonia project is motivated by two main objectives:

- To provide a specialised lexical resource addressing the musical heritage (MH) domain. Such a resource will be the basis to develop knowledge extraction tools from text/speech with a bias/awareness towards the discourse in this domain;
- To provide background knowledge for the construction of a MH textual corpus, effectively and comprehensively representing the discourse in the musical heritage domain.

3.1 Lexicon Structure and Current Status

The Polifonia lexicon is structured as a domain-specific semantic network [49]. The nodes of the Polifonia lexicon are synsets [1] (see Section 2.1), i.e., groups of words referring to the same concept (synonyms). The edges of the network represent semantic relations between concepts. Each concept is associated with a set of lexicalizations for each of the six languages represented in the Polifonia textual corpus, a definition describing its meaning, and a part of speech label to distinguish their roles as nouns, verbs, adjectives and adverbs. The relations among concepts are inherited from WordNet, the resource from which the synsets are taken.

The Polifonia lexicon can be used to identify music concepts in texts and to organize and navigate the textual corpus by means of concepts instead of keywords.

Figure 3.1[6] depicts and example of the lexicon structure and content for the term *musical composition*[7] with hyponymy, synonymy and meronymy relations depicted with different colours.

[1] A lexeme is a unit of lexical meaning that covers words related through inflection, such as *run*, *runs* and *running*.
[2] Defined as *a musical composition of 3 or 4 movements of contrasting forms*.
[3] Defined as *a musical composition of a free form usually incorporating several familiar themes*.
[4] Defined as *a light piece of music for piano*.
[5] Defined as *a musical composition of several movements only loosely connected*.
[6] Source: https://github.com/aliiae/lexical-graph.
[7] Defined as: *a musical work that has been created*.

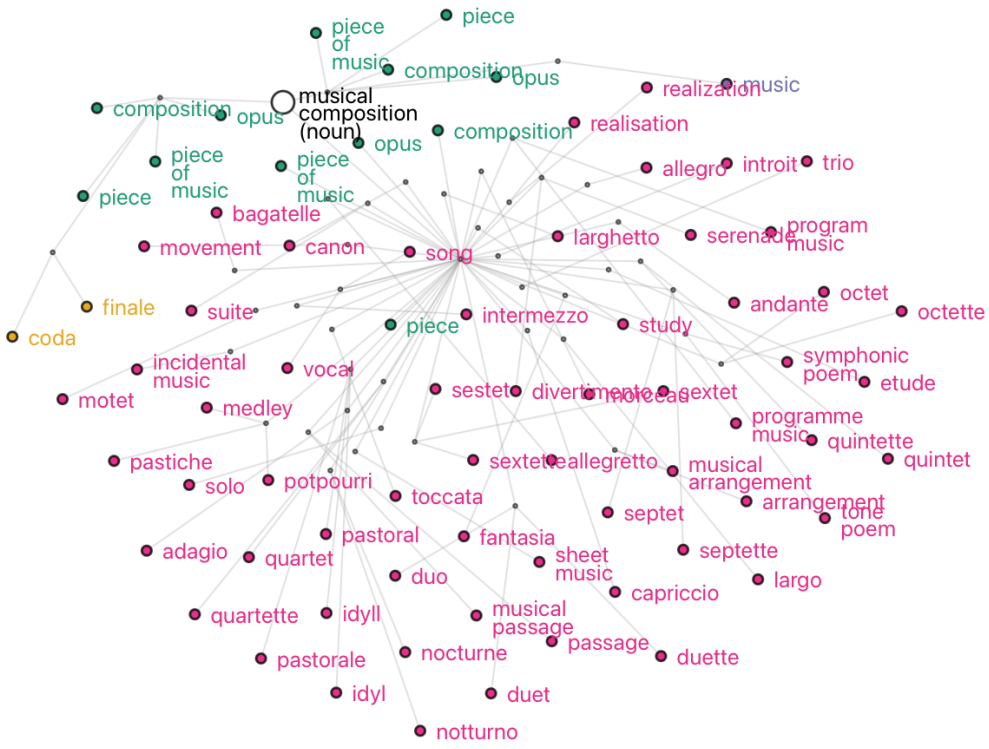


Figure 3.1: An example of the lexicon structure and content for the term *musical composition*. The target word *musical composition* is a noun and is represented as a white node. Pink nodes are the hyponyms of the target concepts, green nodes are the synonyms and yellow nodes are meronyms.

3.2 Methodology

The methodology for building the Polifonia lexicon is a semi-automatic process involving the reuse of existing lexical resources and an analysis and validation effort, performed by human annotators. To address the analysis and validation activity, UNIBO created a dedicated internship program, which is officially part of the Master Degree program in Language Society and Communication of the Department of Modern Languages, Literature and Cultures[8]. At the moment 8 linguists are involved, each having expertise in the English language plus another language represented in the lexicon. All languages are covered by the group of annotators. These annotators are supported by 2 expert musicologists, and are provided with clear guidelines (cf. Section 3.2.4) to perform their tasks. The Polifonia lexicon builds on top of WordNet 3.0 (see Section 2.1) and BabelNet 5.0 (see Section 2.2). We adopted the most commonly used version of WordNet and the last version of BabelNet.

The Polifonia lexicon must reflect the (multilingual) specialised terminology (and its associated meanings) used in the MH domain of discourse. WordNet and Babelnet provide a core basis as far as general or common-sense terms are concerned. Nevertheless, there are two important issues that must be addressed for achieving our goal.

- missing concepts: there are specific meanings that may be missing from these resources
- missing translations: there are terms that are represented only in some of the languages

For example, the term *chord* is a relevant term for the MH domain, it has a specialised meaning in the MH discourse.

[8]Notice that the internship program includes different activities, relevant for all Polifonia WPs and some of the pilots, and is adopted also by other Master Degrees such as "Digital Humanities and Digital Knowledge" and "Artificial Intelligence".

Nevertheless, its MH-specialised meaning also belongs to, and is commonly used by non-MH experts. In fact, both WordNet and Babelnet include this term and its MH-specific meaning. The term *Midwinter horn*, or *Midwinterhoorn* (in Dutch) refers to a traditional Dutch instrument. This term and its specialised meaning are relevant for the MH domain. However, they are mainly used in specialised contexts e.g., discourse about Dutch traditional music, therefore they are rarely part of common-sense or general discourses. As expected, the term *Midwinter horn* is a *missing term* from both WordNet and Babelnet.

Similarly, the term *tarantella*, a folk dance characterized by a fast upbeat tempo, popular in Southern Italy, is relevant for the MH discourse in Italian and it is commonly used and known by people speaking Italian. Therefore, it is present in the Italian part of WordNet and Babelnet, but its *translation is missing* in the French part.

To address these issues, the methodology for building the Polifonia lexicon consists in four activities: selection, translation, extension and validation.

3.2.1 Selection

The selection activity aims at identifying English concepts relevant to the MH domain. To this end we use BabelDomains (see Section 2.2). We select from WordNet and BabelNet all synsets that are identified as belonging to the music domain, according to BabelDomains. This operation results in 1.493 WordNet synsets and a total of 2.756 lexicalizations for English: $1,8$ lexicalizations per synset, on average. The collection of BabelNet synsets is much larger than the WordNet one, including 13.169 synsets.

BabelDomains was created automatically, hence it may contain some noise. We address this issue with the validation activity (cf. Section 3.2.4, performed by the group of annotators, to ensure that all the identified synsets are relevant to the MH domain. The annotators are provided with clear guidelines, defined with the support of musicology experts.

The validation step results in discarding 565 WordNet synsets (over 1.493), including named entities and synsets of ambiguous words that have at least one sense related to the MH domain. For example, the term *grunge* may refer to a music genre but it was included also with the sense defined as *The state of being covered with unclean things*, the synset associated with this sense is discarded.

On the BabelNet selection, the validation step results in keeping 3.304 synsets. In this case the elimination of synsets from the resource was mainly due to the presence of named entities, which is normal given the nature of BabelNet, a multilingual encyclopedic dictionary. We want to remark here that the exclusion of named entities from the lexicon is due to the fact that we want only concepts to be represented in the lexicon and not instances of the concepts (e.g., the concept *composer* is included in our lexicon but not composers' names such as Antonio Vivaldi or Erik Satie), this choice is in line with current approaches in the development of wordnets [15] and it is also due to the fact that the coverage of named entities in WordNet is scarce. Therefore, we decided to use different resources for the collection of named entities, specifically Wikipedia and WikiData. The synsets obtained from BabelNet enriched the coverage of the Polifonia lexicon especially on concepts introduced very recently[9] and for specialist terms, such as *ottu*[10] or *crunkcore*[11].

For this task we involved professional translators to validate the synsets included in the resource and to translate missing lexicalizations (see Section 3.2.4 for more details). Please refer to Section **??** for a discussion of inter-annotator agreement metrics used to ensure the quality of the annotations.

3.2.2 Translation

The second activity consists in translating all identified synsets from English into the other five *target* languages. We use BabelNet Java APIs [12] to retrieve all the lexicalizations for the identified English synsets in Dutch, French,

[9] We remark here that WordNet is a resource developed in 1990s.

[10] Defined as: *The ottu is a double reed wind instrument, used in Carnatic music of Southern India to provide a drone accompaniment to the similar nadaswaram oboe.*

[11] Defined as: *Crunkcore is a musical fusion genre characterized by the combination of cultural and musical elements from crunk, screamo, pop, electronic and dance music.*

[12] https://babelnet.org/4.0/javadoc/index-all.html

German, Italian and Spanish. This operation resulted in retrieving 80 to 90% of the synsets' lexicalizations depending on the language. To fill the gap of missing synsets' lexicalizations, human intervention was necessary. Annotators integrated missing lexicalizations for synsets not covered by BabelNet for languages other than English (cf. Section 3.2.4).

3.2.3 Extension

The extension activity of the process aims at further enriching the Polifonia lexicon, with new lexicalisations for the existing synsets and possibly by adding new synsets. To this end we identified, in this first iteration, three specialist resources: the UNIMARC[13] Bibliographic Resource [14], and the *Terminorum Musicae Index Septem Linguis Redactus* [6] (*Terminorum*, henceforth), a multilingual curated vocabulary in seven languages and the lexicon of the Dutch encyclopedia of organs (*Orgelencyclopedie*). The UNIMARC bibliographic resource was selected because it is a standard resource to organize music material. *Terminorum* was selected because, to the best of our knowledge, it is the largest project aiming at providing a precise terminology for the music domain in different languages. *Orgelencyclopedie* was selected because it is very detailed (1712 terms) and because it is fundamental for the ORGANS Polifonia pilot.

The first two resources cover all the languages addressed by the Polifonia lexicon, except Dutch, for which we plan to employ additional translators. UNIMARC contains terms used by bibliographic resources to organize musical material. *Terminorum* is a large vocabulary covering musical terminology in seven languages (English, French, German, Hungarian, Italian, Russian and Spanish). We went through a digitisation step to obtain a digital version of *Terminorum*. After scanning the book, we processed the resulting images by using Optical Character Recognition (OCR) technologies (cf. Section 4.3.1.1). *Orgelencyclopedie* is also available only as a printed book; therefore, in this case too we resorted to OCR techniques to acquire its text.

The structure of these three resources is incompatible with the organisation based on synsets and semantic relations, used in the Polifonia lexicon. They are collections of lexical items, in different languages without a definition. Some of their terms may be already present in the Polifonia lexicon, therefore we first perform a matching task to identify and discard all of them. The remaining terms can be added to the Polifonia lexicon in two ways. If a term is semantically close to an existing synsets, it is integrated as additional lexicalisation for the concept represented by the synset (e.g. *punctus contra punctum* and *counterpoint*). For example *counterpoint* is expressed in German with the word **kontrapunkt** but, in this language, it can be also expressed with more technical terms such as *stimmführung* or *stimmführungsregeln*. If a term cannot be assigned to any existing synset, a new synset is created and the term is added to its corresponding lexicalisations. The new synset is also put in context with the other synsets through the relations of hyperonymy, hyponymy and meronomy.

3.2.3.1 Enriching Existing Lexicalizations

We model this problem as a Word Sense Disambiguation (WSD) task [50]. We create a vector representation of all synsets from the Polifonia lexicon, i.e. sense embedding, by using ARES[15] [51]. With the same approach we obtain a vector representation for all terms from UNIMARC *Terminorum* and *Orgelencyclopedie*. We compute the cosine similarity (see Equation **??**) between the synset vectors and the lexical items vectors. The cosine similarity between two vectors A and B is computed as:

$$cos(\theta) = \frac{A \cdot B}{||A||||B||} \tag{3.1}$$

where $\cdot$ represent the dot product and $||\cdot||$ represent the vector norm operation. The result of this process provided us with a ranked list of candidate synsets, for each lexical item. These ranked lists are then processed by the group of annotators who confirm or reject the belonging of a lexical item to a specific synset.

[13] UNIversal MAchine Readable Cataloguing
[14] https://www.ifla.org/publications/unimarc-formats-and-related-documentation/
[15] http://sensembert.org.

3.2.3.2 Adding New Synsets

The lexical items that have not been assigned to existing synsets are analysed by annotators with the help of the musicologists, to define new synsets and add them to the Polifonia lexicon according to the guidelines described in Section 3.2.4.

3.2.4 Analysys and Validation

To support the analysis and validation activity, the group of annotators is provided with guidelines that have been defined with the help of expert musicologists.

Guidelines for the selection activity and for the extension activity (enriching existing synsets).

To establish whether a term shall be part of the Polifonia lexicon, either in the selection or in the extension activity (in this last case, when adding new lexicalisations to existing synsets), the group of annotators is provided with the following instructions:

1. For each term, check whether it belongs to one of the relevant conceptual categories from table 3.1, which provides examples for each category.
2. If the term belongs to one of these categories, it will be included in the lexicon;
3. Discard all terms referring to specific people or compositions (i.e., Louis Armstrong, Bach Johann Sebastian, Leonard Bernstein)
4. These guidelines proved useful for discarding irrelevant definitions (i.e. senses) associated with relevant terms, which was found to happen very frequently. For example, the term *orchestration* was associated with the following definitions:
   - (n) orchestration (an arrangement of a piece of music for performance by an orchestra or band)
   - (n) orchestration, instrumentation (the act of arranging a piece of music for an orchestra and assigning parts to the different musical instruments)
   - (n) orchestration (an arrangement of events that attempts to achieve a maximum effect) "the skillful orchestration of his political campaign".

   The third meaning falls out of the scope of a MH specialised lexicon, therefore it can be discarded.

Guidelines for the translation activity.

1. check the definition
2. check the part of speech tag (noun, verbs, adjective, verbs)
3. find the correct translation using specialized dictionaries.

Guidelines for the extension activity (new synsets).
To provide guidelines for this activity, we adapted the instructions for the creation of new synsets, from English WordNet 2020 [15]:

- Significance; it should be possible to easily find at least 5 examples of the usage of the word with this meaning in the Polifonia corpus.
- Non-compositionality; the meaning of the term should not be derivable from its components, e.g., *italian singer* could be tagged with the synsets for *italian* and *singer*; in contrast *civic organ* refers not to an organ that is civic, but it is a special kind of pipe organ
- Distinction; the concept should be distinct from other concepts in the Polifonia lexicon and care should be taken to check relevant synonyms
- Well-defined; it should be possible to easily write a definition for this concept
- Linked; the synset should be easily linked to other synsets

Inter-annotator Agreement To assess the quality of the Polifonia lexicon, we plan to have at least three annotators for each item in the lexicon. This allows us to compute the inter-annotator agreement score for each item, which will be reported explicitly in the lexicon.

In this document we report the computation of the inter-annotator agreement for a subset of 100 synsets ($\approx 10\%$ of the MH subset of WordNet), while in the next weeks the subset will be constantly growing.

We used the Krippendorff's alpha coefficient [52], α to compute inter-annotator agreement because it supports ignoring/missing data entries, handles different sample sizes, and applies to any measurement type. It takes values in the interval $0 - 1$, 0 indicates extreme disagreement while 1 indicates perfect agreement. Commonly $\alpha \geq 0.8$ is considered reliable, values $0.8 > \alpha \geq 0.667$ are used only to draw tentative conclusions, and values $\alpha < 0.667$ are discarded [53].

We computed the inter-annotator agreement on the task of selecting concepts from BabelDomains (see Section 3.2.1). For this task we selected three annotators using the same criteria described in Section 3.2. We obtained a very high α, 0.956, among the four annotators, indicating that the guidelines described are clear and that it is easy to distinguish among relevant and non-relevant concepts of the MH.

Upon publication, we plan to conduct the same validation conducted for the previous steps asking musicologists of the Polifonia project to validate a subset of the new lexicalizations and synsets (10% each).

Category	Definition	Exemple
Performer	a person who performs for an audience in a concert, by Voice or instrument	*pianist, singer*
Medium of performance	the instrument or the composition of persons with or without instrument that perform in a concert	*chorus trio, bongo, trombon*
Genre	a particular type or style that you can recognize because of its special features	*quickstep, a cappella*
Performance technique	a set of both human (phonetic) and mechanical (phonic) physical techniques used to generate sounds or music	*hum, nasality, resonators, crooning, head register, head voice, head tone*
Performance mood	The type of expressiveness that characterises a musical performance	*affettuoso, dolce, energico, espressivo, maestoso, cantabile*
Musical form	Morphology of musical composition	*sonata, sinfonia, hymn, anthem, litany*
Agogic	of or relating to agoge or agogics especially to variations in tempo within a piece or movement	*andante, moderato, presto, allegro, allegro con brio*
Dynamics	variation and contrast in force or intensity	*piano, fortissimo, forte, diminuendo*
Musical notation	visual record of heard or imagined musical sound, or a set of visual instructions for performance of music	*crome, semicrome*
Musical grammar	includes morphology and syntax, the former being defined as the identification of different categories of musical structures, the latter as that of rules connecting morphological unities	*contrappunto, dissonanza,consonanza, intervallo*
Supports	machine that allows to record and play sounds or music	*Audio-tape, cassette deck, audio tape, audio system, sound system, digital, i-Pad*
Dances	sequential configurations of defined musical patterns accompanying defined movement patterns (choreography)	*cake walk, boogie dance, danse macabre, conga line, courante, boogie, twist, waltz, mazurca, samba, jive*
Sounds	any kind of recognizable sound produced by an animate or inanimate being	whistle, grate, knell, sough, purl, dub
Production and reproduction of music	any kind of situation (formal or informal) during which music is produced or reproduced	*recording studio, cover, royalty*
Human perceptions and reactions	every form of human reaction to music both during listening and as a long-term behavior	*groupie, hand, handclap*
Musical Events	Any kind of human gathering characterized by the common interest in producing, reproducing and listening to music	*Rave, gig, cabaret, concert, musical*

Table 3.1: Lexicalization and concepts.

4 The Polifonia Textual Corpus

The Polifonia textual corpus is a comparable, multilingual, diachronic, domain specific corpus on musical heritage. It is designed to support different purposes:

- it must be representative of the MH discourse in terms of languages (Dutch, English, French, German, Italian and Spanish) and time periods (from 1400 to these days);
- it must support the development of tools for knowledge extraction from text;
- it must serve for case studies applications expressed by the Polifonia Pilots.

The aim of the Polifonia textual corpus is to collect and organize texts that can be representative of MH in different languages and in a historical perspective.

The first difficulty encountered in the development of such a corpus was related to the construction of a resource that can be general and that can be used for different purposes. We recall that the Polifonia project is driven by Pilots, case studies characterised by specific requirements studying cultural heritage from different perspectives and with different levels of specialization, BELLS and ORGANS for example study very specific topics, while MEETUPS and CHILD are more general. Furthermore, we recall that one of the aims of the Polifonia project is to develop a large knowledge graph revolving around music heritage and that this can only be achieved by collecting general and specialist knowledge. In order to solve the problem of defining the boundaries of the MH domain and to select relevant documents for it, we used different methodologies to collect specialist, descriptive and interpretative documents.

Analyzing the requirements of the project, we identified two kinds of data that the corpus must include. On the one hand, the project needs data that can be easily processed and organized to develop machine learning algorithms to extract knowledge from text and to populate the Polifonia knowledge graph, according to the ontology developed in WP2. On the other hand, the project needs historical data to preserve and study cultural heritage from the perspectives of the Pilots.

To satisfy these requirements we decided to modularize the corpus, proposing three different modules each of which has its own features and development methodologies. This organization responds to heterogeneous requirements and makes the corpus open to a wide and transversal use. The first module is the Encyclopedic Module (see Section 4.1), the second one is the Books Module (see Section 4.2) and the third one is the Periodicals module (see Section 4.3). Each module covers different aspects of the MH domain and has its own sources of information.

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

Interpretative aspects of the MH discourse are covered by books, 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.

Musicological discourse is covered by the Periodicals module. This represents the most specialized part of the Polifonia textual 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.

The three modules of the Polifonia textual corpus represent a big data resource on MH. The size of this corpus is well beyond the initial objective of the project, that was to collect a corpus of around 1 million words in five languages (English, French, German, Italian and Spanish). In fact, it contains more than 100 million words for each language.

The corpus was organized using the metadata retrieved from each data source (Wikipedia, digital libraries, online music periodicals). We used the Polifonia lexicon to organize the documents corpus. We used different techniques to filter the documents according to the terminology of the lexicon, in particular we implemented a keyword search that allows to select only documents that contain unambiguous music keywords (words that have only a sense according to WordNet and the Polifonia lexicon). As planned for the second release of the corpus (at M18, Deliverable 4.2), a semantic search function will be provided, which will allow to search documents according to concepts (senses in the Polifonia lexicon) and not only keywords. This will allow to browse the corpus according to specific concepts and will require disambiguating all the terms of the corpus in order to connect them with the concepts in the lexicon. The Polifonia pilots will benefit from this feature since it will allow to restrict the searches to relevant documents only, for example it will be possible to retrieve from the corpus only documents that contain a specific named entity linked to the Polifonia knowledge graph or a specific music concept, linked to the Polifonia lexicon. For example, it will be possible to search documents that contain the concept *variation*, defined as *a repetition of a musical theme in which it is modified or embellished* discarding from the search documents containing the word **variation** with other senses, such as *fluctuation*, defined as *an instance of change; the rate or magnitude of change*.

We preprocessed the documents of the Polifonia corpus using the Stanza[1] [54] NLP library and developing sentences annotated with morphological, syntactic and semantic information.

Field	Example
PolifoniaCorpusID	PC10
BabelNetID	bn:00056176n
Description	Prolific Austrian composer and child prodigy; master of the classical style in all its forms of his time (1756-1791)
WikiData ID	Q254
Type	Named Entity
Lemmata DE	wolfgang_amadeus_mozart, joannes_chrysostomus_wolfgangus_theophilus_mozart, w._a._mozart,...
Lemmata EN	w._a._mozart, johann_chrysostom_wolfgang_amadeus_mozart,...
Lemmata ES	juan_crisóstomo_wolfgang_teófilo_mozart, w.a._mozart, w_a._mozart,...
Lemmata FR	johannes_chrysostomus_wolfgangus_theophilus_mozart, w.a._mozart,...
Lemmata IT	wolfgang_amadeus_mozart, joannes_chrysostomus_wolfgangus_theophilus_mozart,...
Lemmata NL	joannes_chrysostomus_wolfgangus_theophilus_mozart, wolfgang_amadeus_mozart,...
Wikipedia Title DE	Wolfgang Amadeus Mozart
Wikipedia Title EN	Wolfgang Amadeus Mozart
Wikipedia Title ES	Wolfgang Amadeus Mozart
Wikipedia Title FR	Wolfgang Amadeus Mozart
Wikipedia Title IT	Wolfgang Amadeus Mozart
Wikipedia Title NL	Wolfgang Amadeus Mozart
BabelNet Categories	Austrian_classical_composers, 18thcentury_male_musicians, Organ_improvisers, ...

Table 4.1: The metadata fields of the encyclopedic module with an example about Wolfgang Amadeus Mozart. Due to space limitations we inserted only a few lemmata and categories.

[1] https://stanfordnlp.github.io/stanza/

	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 4.2: Overall statistics of the Polifonia encyclopedic corpus.

4.1 The Encyclopedic Module

We decided to use Wikipedia to create a general corpus about MH that contains information about musical principles, such as rhythm, harmony or beat; musicians such as John Sebastian Bach or Jimi Hendrix; instruments such as strumming guitar or handpan; or general facts about music. Wikipedia, differently from other specialized resources such as Grove Music Online[2], is freely available and extensively used in many scientific projects [51, 55]. Its popularity is also due to the fact that it contains structured information that is linked to many other resources such as WikiData or BabelNet. Furthermore, it is constantly updated with information about recent facts and new discoveries.

4.1.1 Methodology

We used the same approach used for the construction of the lexicon to select music pages from Wikipedia. Specifically, we adopted BabelDomains and selected all the pages in this resource classified as belonging to the music semantic domain. The select list contains 302.185 Wikipeda titles, including pages about music compositions, artists and music places. We used the BabelNet Java API [3] to enrich the metadata about each page, including the resource type that can be *named entity* or *concept*, the BabelNet categories and the information sources. All the metadata fields used are listed in Table 4.1.

After collecting this rich set of metadata, we used the Python wikipedia API (version $1.4.0$) [4] to retrieve the textual data of each page. We used this resource because it allows to download the full text of Wikipedia articles together with other metadata such as the Wikipedia categories associated to each page.

From the original list of pages obtained from BabelDomains, we collected 250.413 pages in English, 77.986 pages in Italian, 65.970 pages in French, 57.891 pages in Spanish, 53.986 pages in German and 36.609 pages in Dutch.

The overall statistics about the Polifonia encyclopedia corpus are presented in Table 4.2. In this table we can see that just the Wikipedia corpus is well beyond the initial aim of the project of constructing a corpus of 1 million tokens per language. The choice of collecting more data was necessary to cope with the rising use of machine learning algorithms that need large amounts of data to be trained. Furthermore, as already noted, it allows to discover and preserve heterogeneous information about MH, seeing music form different perspectives. The Wikipedia corpus, in fact, lends itself well to be used as training corpus for machine learning algorithms, especially for Named Entity Recognition (NER), Entity Linking (EL) and Relation Extraction (RE) [56, 57] models that will be the focus of the next WP4's tasks: T4.2, automatic extraction of time, space, events, people and musical artifacts from text; T4.3, automatic extraction of socio-cultural and historical context of musical heritage; and T4.4, evaluation of automatic knowledge extraction methods. In fact, we can exploit a large set of information coming from Wikipedia pages, in particular, we can use hyperlinks to associate mentions in the text with named entities codified in the Wikipedia and WikiData (see Section 2.4). This approach will allow to develop training sets for machine learning algorithms focused on the music domain that will be used in Task 4.2, 4.3 and 4.4 of the Polifonia project (see Section 5 for more details).

[2] https://www.oxfordmusiconline.com/grovemusic
[3] https://babelnet.org/guide
[4] https://pypi.org/project/wikipedia/

4.1.2 Overall Statistics of the Corpus

We preprocessed the documents of the Polifonia corpus using the Stanza[5] [54] NLP library and developing sentences annotated with morphological, syntactic and semantic information. Specifically, we split each document of the corpus into sentences. Each sentence was then tokenized[6] and multi-word expression were marked[7]. Each token of the documents was labeled with a part of speech tag and lemmatized. Table 4.2 reports the overall statistics of the corpus, including the number of Wikipedia pages per language together with the number of sentences, tokens, links and the length of the vocabulary of each subcorpus.

	MISC	LOC	PER	ORG
Dutch	822.646	389.373	889.448	351.865
French	1.836.146	1.205.995	2.199.417	971.439
German	981.729	877.628	186.762	719.436
Italian	–	554.550	1.716.888	378.511
Spanish	956.120	448.652	1.183.330	408.083

Table 4.3: Number of named entities in the Polifonia encyclopedic corpus as classified by Stanza NLP for Dutch, French, German, Italian and Spanish.

Category	Quantity
NORP	903.024
PERSON	6.092.895
ORG	3.477.811
LANGUAGE	95.027
CARDINAL	3.557.947
GPE	2.033.071
DATE	4.167.189
FAC	248.414
ORDINAL	831.737
WORK_OF_ART	3.379.321
LOC	221.055
EVENT	225.014
PRODUCT	267.873
TIME	90.810
QUANTITY	80.258
LAW	17.749
MONEY	58.318
PERCENT	38.530

Table 4.4: Number of named entities in the Polifonia encyclopedic corpus as classified by Stanza NLP for the English language.

[5] https://stanfordnlp.github.io/stanza/
[6] This operation consists in splitting a text into words (tokens) or subwords (subtokens).
[7] This operation consists in identifying words that are composed by two or more tokens, e.g. *holiday season*.

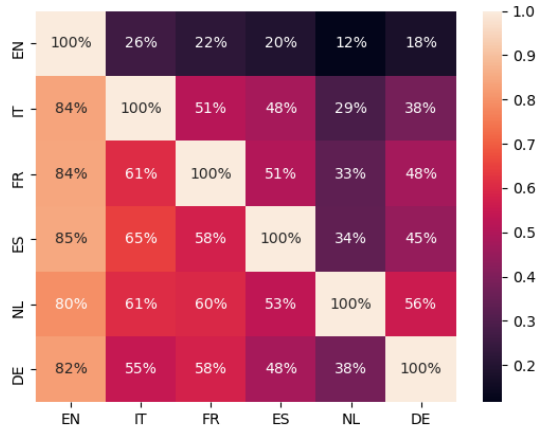


Figure 4.1: Percentages of pages shared among language pairs.

4.1.3 Named Entities in the Corpus

To have an idea of the number of named entities that can be discovered in our corpus we pre-processed the texts using the Stanza [54] NLP library. We can see in Table 4.2 that the number of entities that can be discovered in our corpus is very high, ranging from 5% to 10% of the number of tokens. The training sets used by the Stanza library are different for each language, as we can notice also by looking at the different entity types reported in Table 4.3 and 4.4. From these tables, we can have a detailed view of the entity types used for each language and their distribution. As we can see, the entity types in which mentions are collected are very general and do not allow to distinguish among very different properties. For example, from these categories, we cannot infer occupations such as *performer* or *producer* (in the music sense) because they are conflated in the same general category: PERSON. Even a more specific category, WORK_OF_ART, is not useful for the Polifonia project, because we want to distinguish among different types of piece of work, as we have shown in Section 3.2.4 especially on Figure 3.1, where we presented different types of musical composition, such as *song*, *bagatelle*, *vocal* or *pastorale*. All these specific pieces of information can be easily obtained from the corpus using the BabelNet categories introduced above in Table 4.1.

4.1.4 Wikipedia Pages and Languages

It is not possible to find all the pages in the list constructed using BabelDomains. This is due to the fact that many pages have changed title since the development of BabelDomains or have been removed by Wikipedia editors. In other cases, the some concepts or entities are covered only for some languages. This aspect indicates that the corpus that we are collecting does not contain the same pages for all our six languages.

The overall number of different pages in all languages is 276.592 and the 3.5% of these pages can be found in all the languages of the corpus. These shared pages are about famous composers or very general music concepts. For example, the Wikipedia page for the Symphony No. 4 by Sergei Prokofiev con be found in all our six languages. The number of shared pages is low if compared with the English subcorpus but it is very high if compared with other languages (e.g., 40% of the pages in Dutch can be found in all other languages). The corpus is not fully parallel because it contains country-specific information. We argue that this aspect makes the various language-specific subcorpora comparable because they have been collected following the same criteria and also because they reveal peculiar aspects of different music cultures. In fact, we noticed that many entities have a Wikipedia page only in their original language. The statistics about the percentages of pages that are shared between language pairs can be found in Figure 4.1. Each row of this heatmap indicates the percentages of pages of the language on the vertical

axis that are shared with the language in the horizontal axis. For example, we can see that 26% of the pages in the English part of corpus are also present in the Italian part or that more than 80% of the pages in all languages are also in the English part.

The metadata of each page, including the BabelNet identifier and the language, were stored into a SQLite database (see Table 4.1) while the textual data were downloaded directly in HTML format. Having the data in HTML format is very convenient because it allows to extract information that is already encoded in the Wikipedia pages such as hyperlinks or tables.

4.1.5 Vocabulary Saturation

Vocabulary saturation [13] has demonstrated to be a convenient analytic tools to study sub-languages [58] and to evaluate the quality of a corpus [59]. The intuition behind this technique is that if we analyze how the vocabulary size of a corpus changes with respect to the size of the corpus (expressed as number of tokens) we can observe the degree of saturation of the corpus. A corpus, considered as a linearly ordered partition into segments of equal size, is considered saturated at lexical level if at a certain point its vocabulary size starts to increase slowly when the size of the corpus is increased.

The plots in Figure 4.2 show the relationship between the size of the vocabulary (vertical axis) and the size of the corpus (horizontal axis) for all the languages of the Polifonia textual corpus. We can observe that the grow of the vocabulary starts to slow down when its size arrives around 250.000 for all our languages. We want to remark here that this particular corpus includes names of musicians and musical works that can be different for each document in it.

Language	Source	URL	Documents
Dutch	Delfer	https://www.delpher.nl	47.670
English	JSTOR - Constellate	https://constellate.org	19.065
French	Gallica	http://gallica.bnf.fr/	40.961
German	DDB	https://www.deutsche-digitale-bibliothek.de	8.076
German	MGG	http://mgg-online.com/	17.000
Italian	Internet Culturale	http://internetculturale.it/	1.735
Spanish	BNE	http://bne.es/	15.894

Table 4.5: Overall statistics of the Polifonia book corpus.

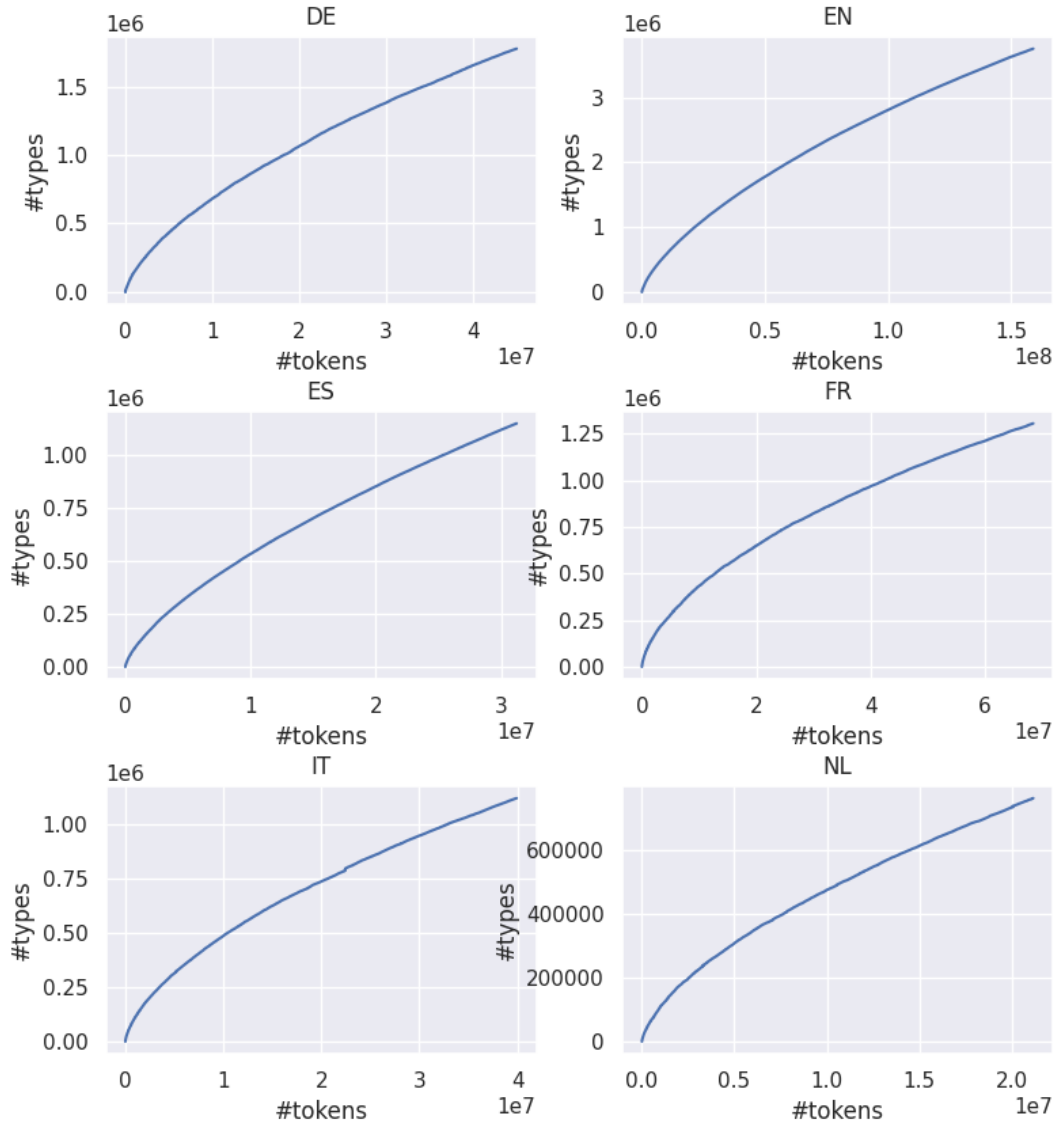


Figure 4.2: Relation between the size of the vocabulary (vertical axis) and the size of the corpus (horizontal axis) expressed as number of tokens. The axes are in logarithmic scale.

4.2 The Books Module

The aim of the book corpus is to support music scholars and practitioners to find historical texts about music. The book section includes monographs, essays, novels, biographies and diaries. This type of text allows the collection of contextual information, characterised by a historical perspective. Critical essays can report on the discourse on MH,

historical treatises can inform about the evolution of the musical discourse, monographs contain information about relevant characters and their relationships, novels can reveal the social impact of music.

4.2.1 Methodology

The first step for the collection of documents was the identification of suitable sources of information. For each language of our corpus we selected different digital libraries. The requirements for the selection were: to be freely available on the web, to have a permissive licence (e.g. Creative Commons[8]) to have APIs that can be used to select and download documents, to contain a reasonable number of digitised documents.

We used Delpher [9] for Dutch, JSTOR [10] for English, Gallica - the digital library of the Bibliothèque nationale de France - [11] for French, the Deutsche Digitale Bibliothek [12] for German, Internet Culturale [13] for Italian and Biblioteca Digital Hispánica [14] for Spanish. We plan to extend the searches also to Wikisource [15] since it is multilingual and contains mainly literary texts that can be integrated to the ones collected using the reported sources.

Querying web APIs can be very expensive in terms of time and web resources used, for this reason, we decided to use a simple strategy to download documents from digital libraries. Even if each digital library has its own APIs with different parameters, we tried to use the same filters for all languages. We selected documents that are categorized as monographs (hence discarding periodicals), whose full text contains at least one occurrence of the word *music* (in each language), published between 1700 and 2021. Once these general searches are done in all languages, other refinements can be conducted on the corpus, using the musical terminology collected in the Polifonia Lexicon (see Section 3.2.4). However, we noticed that the downloaded material contains many spelling errors that are intrinsic to the OCR techniques used to extract text from digitized documents. For this reason, at this stage, we do not provide information about the length of the corpus in terms of number of tokens and types. In fact, we plan to apply a correction method to the texts using new tools specifically tailored to historical documents [60]. Please refer to Section 4.3.1.1 for more details on this activity.

The statistics about the books module can be found in Table 4.5. The metadata about this part of the corpus, including author title, year of publication, url of the resource on the digital library and words in the Polifonia Lexicon that are contained in the text of the resource[16], can be found on the GitHub repository of the corpus.[17]

4.3 The Periodicals Module

The intellectual history of an epoch is strikingly reflected in its periodicals. This applies not only to general journals but also to musical ones, the articles and reports of which offer a variety of material for study as regards the outlook of the era. Within the European context, the mid-1800s is the period in which the first generalist musical periodicals were published. They were dedicated to historical essays and biographies of contemporary composers, above all they provided information about musical performances that were held in the main European capitals. Periodicals are, in this sense, an excellent support for the study of music in the historical, social and cultural context.

The emergence of musical life in the public domain instead of its confinement within a more exclusive social group leads to the creation of an audience increasingly interested in the music of their time, and in the institutions that promote it. With the increasing development of musical life, the large general periodicals and the more locally oriented journals were joined. This trend began in the 1840s and gained momentum in the second half of the century.

[8] https://creativecommons.org/licenses/by/2.0/
[9] https://www.delpher.nl
[10] https://constellate.org
[11] https://gallica.bnf.fr/
[12] https://www.deutsche-digitale-bibliothek.de
[13] http://www.internetculturale.it
[14] http://www.bne.es/es/Catalogos/BibliotecaDigitalHispanica/Inicio/index.html
[15] https://wikisource.org/
[16] We would like to remark that this information is temporary since the texts will be processed to correct OCR errors.
[17] https://github.com/polifonia-project/Polifonia-Corpus

For example, the foundation of choral societies and the organization of music festivals gave rise to publications largely devoted to choral singing and vocal music. Developments in instrument manufacture demanded the appearance of specialist periodicals. Reforms in music education led to numerous periodicals dealing with general educational matters or more specific ones such as piano or singing pedagogy, or eurhythmics on the Jaques-Dalcroze system, or notation (e.g. John Curwen's Tonic Sol-fa method) Liturgical reforms, the revival of hymns and sung services and the founding of church music organizations encouraged the appearance of many more. The increasing cultivation of instrumental and vocal music also gave rise to periodicals dealing with particular genres, including chamber music, popular music (for such instruments as the guitar, mandolin and zither), and even light music, including that of the music halls and smoking concerts.

Since the early 19th century, and particularly since 1850, many music periodicals have been issued by music publishing firms with the propagation of their own music publications as an important objective; when such factors are permitted to influence the editorial policy, such periodicals may give a highly partial view of the musical scene. Many others, for example the official organs of institutions or learned societies, or independently owned periodicals, are unaffected by such factors. Commercial considerations over the years have increasingly dictated the necessity for periodicals to include advertising material; the nature of such advertising may often provide clues as to a periodical's readership.

The first significant general music periodical in Italy was the *Gazzetta musicale di Milano* (1842–8, 1850–62, 1866–1912; from 1903 *Musica e musicisti*, from 1906 *Ars et labor*, and revived later as *Musica d'oggi*) which gave special attention to Italian opera, in Italy and abroad, and deserves special mention for its chronological lists of productions at La Scala and La Fenice. It also provided biographical, historical and bibliographical contributions and reported on Italian and other main musical centres.

Besides smaller-scale journals, some of which were locally oriented and ephemeral such as the *Gazzetta musicale di Firenze* (1853–9), founded by E. Picchi, which espoused Meyerbeer's cause, and the *Gazzetta musicale di Napoli* (1852–68) or *L'Italia musicale* (1847/8–59), which offered detailed critiques of recent operas, there were others such as *L'arpa* (1853/4–1902) or *Il trovatore* (1854–1913), which enjoyed a wide circulation by covering the arts in general while according music a leading place.

The *Revue musicale* (1827–35), founded by F.-J. Fétis, was the first significant French music periodical of the 19th century: in addition to historical essays and biographies of contemporary composers, it contained detailed notices of performances in Paris. Its amalgamation with the *Gazette musicale de Paris*, founded in 1834, resulted in the *Revue et gazette musicale de Paris* (1835–80), which enjoyed particular success. It had the services of notable writers, including Liszt and Berlioz as well as Fétis; Berlioz published his articles on Rameau and on Beethoven's symphonies in this journal, and Meyerbeer's operas were prominently featured. Of equal stature were *Le ménestrel* (1833/4–1940) and La France musicale (1837/8–1870). The former developed into a journal of considerable renown, concerned both with historical matters and with contemporary events and offering valuable reports on Paris and other European musical centres. *The Harmonicon* : (1823–33) contained significant articles on London musical life. *The Musical World* (1836–91) was England's first comprehensive music periodical; in some aspects it was modeled on its German and French predecessors. It contained not only historical articles but also detailed reports from the main European musical centers, together with critiques of publications and performances.The Musical Times was founded in 1844 and it was originally published under the title *The Musical Times e Singing Class Circular*, before changing to the shorter *The Musical Times*. It is actually the oldest of all musical journals with a continuous record of publication. Within the English-speaking area, we point to *Dwight's journal of music: a paper of art and literature* founded in 1852 by the american critic John Sullivan Dwight with the purpose to bring awareness of European classical music to American readers.

The first Spanish music periodicals appeared in the 1840s. *La Iberia musical y literaria* (1842–5), founded by M. Soriano Fuertes and edited by J. Espín y Guillén, was the first significant one; it contained articles on various topics – instrumentation, the activities of performers, and Spanish and foreign music. Most of the periodicals that followed lasted only a short time, but among them the locally oriented *Gaceta musical de Madrid* (1855–6), issued by H. Eslava, strove to elevate the level of Spanish musical life. Among the periodicals of the 1860s, *La España musical* (1866–74) contains essays on Wagner's operas and Liszt's symphonic poems.

The next most important and comprehensive music periodical in Germany was the *Neue Zeitschrift für Musik*,

founded by Robert Schumann (with Friedrich Wieck, Ludwig Schunke and Julius Knorr) in 1834. Its contents were organized on the same lines as those of the *Allgemeine musikalische Zeitung*, but its intellectual outlook, dominated by Schumann until 1844, was entirely different from the earlier journal's rationalism. Schumann's aim was to use his periodical – which he saw as the standard-bearer of the Romantic movement in music – to improve the musical resources of Germany, depleted after the deaths of Weber, Beethoven and Schubert, and affected by increasing superficiality and the domination of the virtuoso, and to restore the 'poetry of art' to its rightful position. To this end he used the journal as a forum for the creative artist, excluding the dilettante. Another very important journal was the *Berliner allgemeine musikalische Zeitung* that was founded by the well-known music theorist Adolf Bernhard Marx and was focused on detailed analyses of musical works. As far as the Dutch language area is concerned, the periodical *Het Orgel* - founded in 1886 is the oldest in Europe about organs, and also it is very important for its diffusion and longevity the *Tijdschrift der Vereeniging voor Noord-Nederlands Muziekgeschiedeni* founded in 1882 which became *Tijdschrift van de Vereniging voor Nederlandse Muziekgeschiedenis* in 1960 and then *Tijdschrift van de Koninklijke Vereniging voor Nederlandse Muziekgeschiedenis* still active today.

Title	Language	Years of Activity
L'Arpa	italian	1853-1880
Gazzetta musicale di Napoli	italian	1852-1868
L'Italia musicale	italian	1847-1859
Il trovatore	italian	1854-1910
La nuova musica	italian	1899-1919
Rivista Nazionale di Musica	italian	1920-1943
La cronaca musicale: piccola rivista di musica	italian	1896-1917
Giornale delle belle arti e della incisione, antiquaria, musica e poesia	italian	1784-1788
Revue et Gazette musicale de Paris	french	1835-1880
Le ménestrel	french	1833-1940
La France musicale	french	1837-1870
The Harmonicon	english	1823–1833
The Musical World	english	1836-1891
The Musical Times	english	1844-1900
Dwight's journal of music : a paper of art and literature	english	1852-1881
La Iberia musical y literaria	spanish	1842-1849
Gaceta musical de Madrid	spanish	1855-1878
La España musical	spanish	1866-1867
Almanaque musical de teatros	spanish	1866-1867
Ilustración musical hispano-americana	spanish	1888–1894
La música ilustrada hispano-americana	spanish	1898–1902
Berliner allgemeine musikalische Zeitung	german	1824-1830
Neue Wiener Musik-Zeitung	german	1852-1860
Allgemeine musikalische Zeitung	german	1798-1848
Tijdschrift der Vereeniging voor Noord-Nederlands Muziekgeschiedeni	dutch	1882-1959
Tijdschrift van de Vereniging voor Nederlandse Muziekgeschiedenis	dutch	1960-1994
Tijdschrift van de Koninklijke Vereniging voor Nederlandse Muziekgeschiedenis	dutch	1995-2018
Mens en Melodie	dutch	1946-2012
Het Orgel	dutch	1886-present
De Prestant	dutch	1952-1971
De Schalmei	dutch	1946-1950
Orgelkunst	dutch	1978-2020

Table 4.6: List of downloaded journal.

4.3.1 Methodology

We developed dedicated software to access web APIs and download the periodicals that we identified with the help of music historians. Information about each periodical we have been able to download at the current state of progress - including the title, the language, the years of activity - is presented in Table 4.6. On this material will be conducted OCR processing and post-correction as described in the following section.

4.3.1.1 OCR

OCR stands for Optical Character Recognition. OCR is a technology used to convert PDF files, scanned or sometimes photographed images of machine printed characters into digital text data that is searchable or editable in standard desktop applications.

Today, OCR technologies are widely used alongside other Data Capture technologies to maximise accuracy in text recognition and conversion. Often Data capture is part of a wider automation or optimisation project. One of the main problems that OCR techniques have to face is the great variety of materials to be processed, not only due to very different graphic characteristics (particular layouts, presence tables, historically obsolete characters), but also due to the bad conservation state of some documents belonging to ancient historical periods.

In our project we used OCR technology to process material for the Polifonia textual Corpus, which covers a very large period of time, from the 18th century to the present, and is made up of four languages that have specific typefaces. These two characteristics confronted us with two open problems in OCR techniques: the diachronic and the multilingual aspect of texts.

4.3.1.2 OCR Pipeline

For the conversion into text of the corpus documents, we defined a pipeline for the OCR (see Figure 4.3). This pipeline has the objective of defining the operations needed to perform the OCR on a large amount of textual data. Consequently, the pipeline has set the requirements for the implementation of software needed to perform these tasks[4]. The pipeline consists of the following steps:

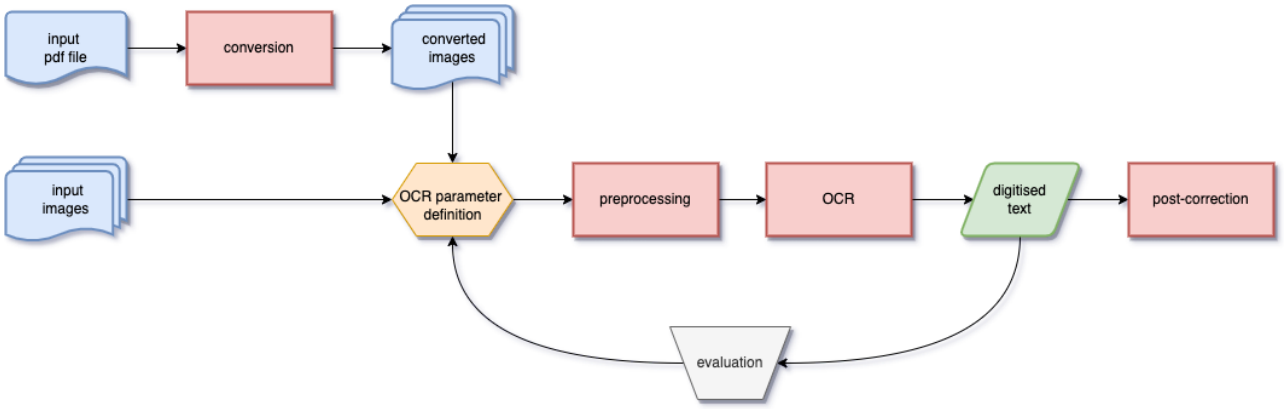


Figure 4.3: Pipeline of the OCR process.

File Preprocessing Before the actual OCR of the content, the files are pre-processed. This step is necessary because of the technical constraints imposed by the technological solutions adopted in the next phases. Moreover, the files to be converted into text are in different formats. In order to process the files, it is necessary to have files in image format. In the case of an input file in pdf format, this is converted into an image file, one for each page of the pdf.

Review's Title	Language	Pages considered	Accuracy
L'Arpa	italian	12	97.62
Il Trovatore	italian	4	82.62
Le Ménestrel	french	4	87.64
Revue et Gazette musicale de Paris	french	7	96.04
The Harmonicon	english	5	92.93

Table 4.7: Evaluation metrics for some of the journal considered.

Image Preprocessing As next step, it is essential to preprocess each image file in order to obtain the cleanest possible image and avoid noise in the OCR phase. For these purposes, we used OpenCV[18] [61], one of the best known libraries employed in the field of computer vision. Using this library we have included in our software the option of performing different types of prepocessing on corpus images, including:

- *Gray scale*: the images are converted to grey scale in order to avoid possible noise caused by colours;
- *Noise removal*: denoising algorithms estimate the original image by suppressing noise from the image;
- *Tresholding*: separates out regions of an image corresponding to objects which we want to analyze;
- *Dilate*: dilates the source image using the specified structuring element that determines the shape of a pixel neighborhood over which the maximum is taken;
- *Erosion*: erodes the source image using the specified structuring element that determines the shape of a pixel neighborhood over which the minimum is taken;
- *Edge detection*: an image-processing technique, which is used to identify the boundaries (edges) of objects, or regions within an image;
- *Skew correction*: given an input file containing some rotated text, this algorithm allows to detect the block of text in the image; to compute the angle of the rotated text; to rotate the image to correct for the skew.

OCR For the OCR phase, after a thoughtful analysis of the state-of-the-art technologies in the field, we opted for Tesseract[19] [11], an open-source OCR engine developed at Hewlett-Packard Co, Greeley Colorado between 1985 and 1994. Tesseract, now in its 4th version, can recognise more than 100 languages "out of the box". The latest version is based on Long Short Time Memory (LSTM) artificial recurrent neural network (RNN) architecture and focuses on line recognition, but also still supports the legacy Tesseract OCR engine of Tesseract 3 which works by recognising character patterns. All pre-processed images are provided to Tesseract, which will then convert them into text.

Output handling Once a file has been processed by the OCR software, the software implemented allows to save the resulting text in *.txt* format, taking into account the volume or edition number to which it belongs. If the original files contain the necessary metadata, it is in fact possible to save specific files for each number of a journal, thus combining the OCR performed on several images that referred to a common resource.

Evaluation OCR results produced using different preprocessing parameters are finally evaluated in order to obtain the best result. These parameters are specific to each resource. Before proceeding with the OCR of the whole resource, a small portion of the resource (4-10 pages) is taken, on which the accuracy of the output produced is calculated and compared with a manually validated ground-truth file. In order to evaluate automatically the quality of the OCR, we used a tool called OCREval[20] [62]. This library consist of 17 tools for measuring the performance of and experimenting with OCR output. In particular, we used the Accuracy metric that the library provides. Once we had found the parameters that gave us the best results, we proceed with OCR on the full resource.[21]. Table 4.7 shows some of the accuracy metrics related to OCR performed on some of the

[18] https://opencv.org/releases/
[19] https://github.com/tesseract-ocr/tesseract
[20] https://github.com/eddieantonio/ocreval
[21] At this state of the project we acquired the text for three languages English, French and Italian, as it can be seen in Table 4.7. The same methodology will be applied also for Dutch, German and Spanish.

resources in the corpus.

Although the results obtained so far are far from perfection, state-of-the-art post-processing models have been shown to provide excellent results from accuracy rates around 90%, or at least higher than 80% [63].

The OCReval tool was always used for error assessment. This tool allowed us to analyse the most common errors in the samples examined and allowed us to compare them with manually annotated files. In the project's GitHub repository, it is possible to consult the files containing the accuracy evaluation and the errors found in each of the files. The detailed analysis of these errors will allow us to develop post processing software that is able to correct most of the recurrent errors that were produced during the OCR phase (see Section 4.3.1.3). Currently, error analysis was conducted on the sample of assessed texts from Table 4.7. On these texts, the following were evaluated: (i) the most common errors (ii) the percentage of errors for each character in the texts. The most common errors found in all texts were generally attributable to noise generated by images, graphics, or defects in the image on which the ocr was performed. This type of error is easily identifiable, since in most cases it consists of long strings of characters and symbols. The remaining common errors are swaps between similar characters (e.g. *'e'* with *'è'*). On the other hand, considering the accuracy per character, an aggregate analysis was conducted taking into account all samples of all languages. From this analysis emerged an average accuracy value of 88.7%, with a standard deviation of 15.04. Figure 4.4 shows the accuracy rate for each character. From this bar chart it can be deduced that the characters that are confused most often are special characters, such as *'°'*, *'Ç'*, and *'ï'*. In addition, uppercase letters and numbers are confused more often than lowercase letters.

4.3.1.3 Next Steps

In order to further improve the accuracy of the software produced, we plan to conduct more experiments on the different texts that are part of the Polifonia textual corpus. In particular, we plan to expand the experiments carried out using other preprocessing techniques and to do experiments with different fine-tuning parameters.

After this, we plan to to extend the analysis on the most common errors found in the files generated by OCR. Finally, we will implement post-processing algorithms that can automatically correct the most common errors found in each file.

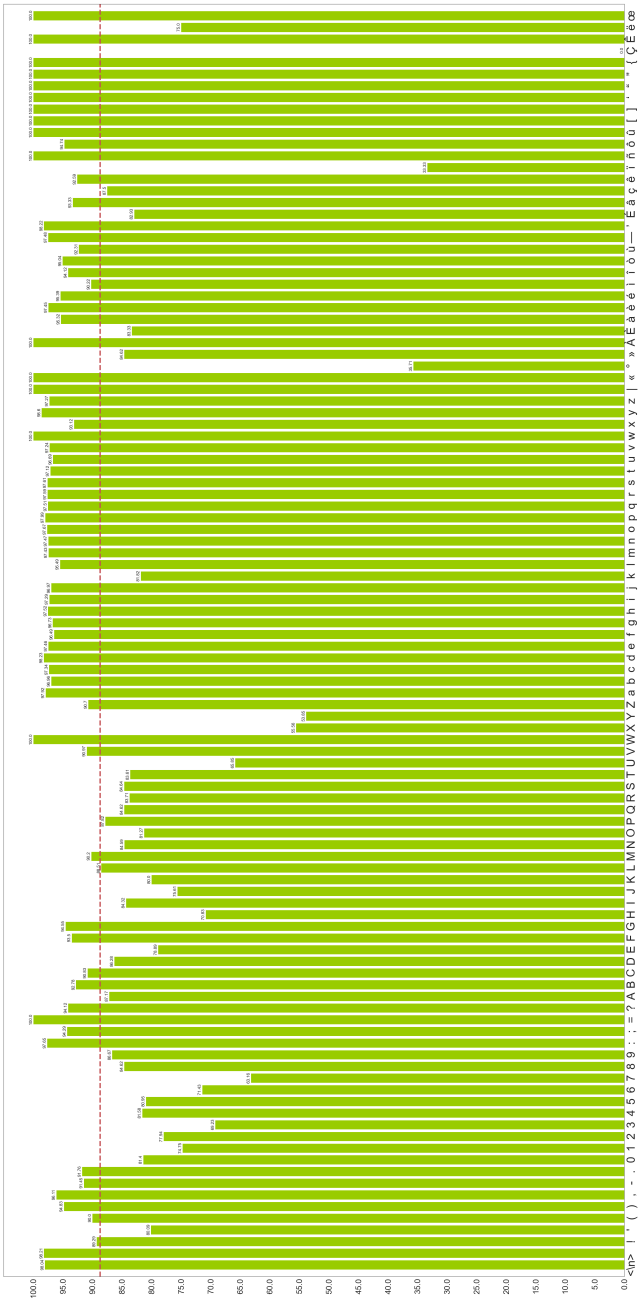


Figure 4.4: Analysis of the accuracy rate for each character of the evaluated corpus.

4.4 Pilot Corpora

The pilots that make up Polifonia are practical case studies that translate specific interests thematically or chronologically. Within the construction of the general corpus these are used for two purposes: on the one hand they serve to select sources from a balancing perspective, which is possible due to the specific requirements each of them has to fulfil. On the other hand, the requirements highlight the need for additional documents that are thus identified and integrated into the corpus. The pilots are therefore the result of research on the Polifonia textual corpus and in order to collect necessary requirements for investigations a survey was submitted to each Pilot leader, aimed at collecting the necessary requirements by qualitative and quantitative information. The survey was organised in 7 questions (2 closed question and 5 open questions) + 1 field for comments.

The questions were related to:

- the time period covered by the pilot;
- the typology of the textual genres useful for the pilot;
- the sources used to collect material;
- the languages covered by the pilot;
- suggested data sources.

The survey thus made it possible to identify the following specific features.

Concerning textual genres we had a preference for historical treatises (e.g. book about popular music), technical treatises (e.g. documents about constructions of organs or bells) and essays (e.g. travelers' testimonials, scientific literature) followed by - with fewer preferences - correspondences, commercial and notarial documents and prescriptive texts and fiction. As regards time period, Pilots focused mainly on contemporary texts and a strong preference can be seen also for XIX and XX centuries. Renaissance (XIV century) is also taken into account by two pilots. Five Pilots deal in particular with textual sources: Bells, MusicBo, Child, Organs and Meetups.

The Bells Pilot aims to represent intangible phenomena, such as the diffusion and impact of sound in space, or sound variations caused by wall structures, to support landscape planning and architectural restoration design. The sound of bells, both in urban and rural areas, constitutes a cardinal element of the landscape, contributes to the definition of a sound landscape, performs a function of marker of the daily and festive, ritual time. Bells constitute a complex heritage, made by knowledge, practices, discourses in a social dimension. For this specific knowledge the following textual genres were selected:

- Catalogue sheets;
- Campanology magazines or websites
- Scientific literature
- Diocesan decrees

Texts relevant for the reconstruction of a chronology of reference and for the extraction of vocabularies relating to performance techniques, local variants in repertoires and playing practices. Texts about local sound practices; sound events and performances and informal transmission of knowledge.

The MusicBo Pilot aims to investigate the role that music played in the life of the city of Bologna through a historical perspective in terms of performances, encounters between musicians, composers, critics and historians. The pilot focuses on the testimonies of scholars, journalists, travelers, writers and students from medieval to modern times through published documents showing diverse discourse styles such as stories, letters, reports, news, reportage, For this specific knowledge the following textual genres were selected:

- historical treatise
- critical essay
- Media (blog, specialist or general press)
- Correspondences
- Commercial documents

- Legal documents
- Lyrics (about Bologna)
- Travel Guides

The Organ Pilot focuses on the history of pipe organs, which is rich and diverse, and highly interrelated with economic, religious and artistic contexts. For centuries, pipe organs have been the most complex extant musical instruments and the Pilot aims to build a knowledge graph about their existence out of the text of the *Orgelencyclopedie*(1997-2010), in order to access information about building practices and characteristics of Dutch pipe organs. The *Orgelencyclopedie*(1997-2010) is a 4.500+ pages containing histories and images of almost 2.000 Dutch organs, published by NIvO.

The Child Pilot aims to explore a historical perspective on the part 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. The Pilot aims to build a knowledge graph of the historical experience of music in childhood, using life writings and other historical texts as sources for adults' reflections on music heard in childhood, third-party observations on children's engagement with music, and children's own first-hand accounts. For this specific knowledge the following textual genres were selected:

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

The Meetup 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. Such encounters were particularly significant for cultural and musical exchange and dissemination, and meetings that were catalysts for musical change. For this specific knowledge the following textual genres were selected:

- biographies
- memoirs
- travel writing
- open-access databases

4.5 FAIRness and Reproducibility

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

As described, the central data used in this research served to develop the Polifonia Lexicon and the Polifonia Textual Corpus. For the development of the Polifonia Lexicon we used resources that have been widely used in previous work (i.e. WordNet, BabelNet and BabelDomains) and we are trying to improve them with the help of translators and the use of publicly available resources (i.e. UNIMARK and *Terminorum*) that will not be reproduced but only used as reference vocabularies. Regarding the development of the Polifonia Textual corpus, we collected data from different sources (i.e. Wikipedia and open digital libraries with Creative Commons[22] or similar licences[23]). On the data obtained from digital libraries, cleaning and conversion operations will be applied, especially to convert images into text and to correct possible errors involved in this operation. The texts obtained from digital libraries will not be shared publicly and will be used only by the Polifonia partners. However, we will share the Polifonia Lexicon under a permissive Creative Commons license.[24] We will also share the code that we used to construct the corpus and the metadata of the corpus for findability, accessibility and to make our work reproducible.[25]

[22] https://creativecommons.org/licenses/by-nc-sa/3.0/it/deed.it.
[23] https://gallica.bnf.fr/edit/und/conditions-dutilisation-des-contenus-de-gallica
[24] https://github.com/polifonia-project/Polifonia-Lexicon
[25] https://github.com/polifonia-project/Polifonia-Corpus

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.

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

5 Discussion: main challenges and next steps

5.1 Main Challenges

One of the challenges that we faced during the development of the Polifonia textual corpus was related to the collection of materials in different formats. The purpose of the textual corpus of Polifonia is the preservation of the MH, which implies that a part of the textual material composing it has special material characteristics. For example historical treatises or periodical issues, often dating back to the 18th Century, are obviously not necessarily in the best material condition: they may have been damaged by time, which makes reading and transforming them into machine-readable formats difficult. In addition, a major difficulty was to standardise the content of texts: due to their linguistic and temporal heterogeneity they have different formatting (images, columns, etc.) and the transformation of sources into a machine-readable format assumes the problem of format inhomogeneity.

During the creation of the Polifonia lexicon, we found some critical issues related to multilingual resources which also show how important the contribution of Polifonia is. These critical issues have been considered for improved translation guidelines according to the following aspects. The translation of a term from a source language (sl) into a target language (tl) is rarely unique. Because of the phenomenon of polysemy (coexistence of different meanings within the same symbol) a single term in the *sl* can have several meanings in the *tl*. For example the word *roll* in English (source language) has a literal translation in *rouler* in French (target language), however, depending on the context of use, it can be translated with other meanings of the term (related to the musical domain) such as *retentir*, *résonner*. Another critical issue is that of synonymy, characterised by the equal meaning of two terms (they can be replaced in a sentence without modifying its overall meaning). This is the case of the terms *sans bruit*/ *silencieusement*, which can be used in the same context without changing the overall meaning of the text. A term like *songlike* in *sl* can be translated in target language as *ayant une melodie* and *chantant*: both are correct translations but each brings emphasis to a different aspect of the meaning, both being connotations of the same term. Such complexity allows us to refine the first version of the translation guidelines, in particular according to the following aspects:

- avoid multiwords expressions as much as possible;
- provide for synonymous translations that do not take stop words (i.e. grammar prepositions, articles) into account for the purposes of evaluation.

5.2 Next Steps

One of the main objectives of the Polifonia project is to construct a knowledge graph about music. One way to produce this kind of resources is to process text and extract from them knowledge, i. e. facts of the form subject, predicate, object that tell us basic information about entities involved in the music discourse. Nowadays, information extraction techniques are mainly based on large neural networks [64, 65] that require large amount of data to be trained. For this reason, we plan to develop training sets for information extraction models specialized on the music domain. Before producing these datasets it is necessary to process the data and to define the taxonomies for the classification of named entities. This step is facilitated by the fact that Wikipedia pages are connected to WikiData entities and from these entities it is possible to extract many features, such as gender, nationality, occupation, instrument played and so on. For example, Miles Davis[1] in WikiData [2] has different occupations, such as composer, trumpeter, musician and jazz musician. From all these occupations we can select a set of categories to be used to classify mentions in the text, for example, we can decide to use coarse categories such as artist or fine-grained

[1] https://en.wikipedia.org/wiki/Miles_Davis
[2] https://www.wikidata.org/wiki/Q93341

categories such as trumpeter. This step is necessary because current datasets for training NER [66, 67, 68] models are based on few general categories such as LOC (location), PER (person), and ORG (organisation). With the new training sets, we will develop algorithms specialized on the music domain that will be used on the other parts of the corpus to extract information and to populate the Polifonia knowledge graph.

6 Conclusions

In this report we presented the work carried out on the Polifonia project for the first deliverable of Work Package 4. The initial idea for this deliverable was to produce a textual corpus on music heritage of around 1 million words for five languages: English, French, German, Italian and Spanish. During the unfolding of the project and thanks to many meetings with the Polifonia partners we eventually realised that we needed to do more. This was mainly due to the recognition that if we really want to preserve our past and if we want the project to contribute and have an impact also outside its original boundaries, we cannot discard any document about music and we have to try to do our best to collect as many resources as possible. With this spirit in mind we started using existing resources (e.g. Wikipedia, WordNet, BabelNet, etc.) and collecting documents from many other open digital libraries. This effort allowed us to develop a large modularized corpus that went well beyond the initial idea in terms of dimension (it contains more than 100 million words for each language), in terms of language coverage (it also includes Dutch) and in terms of products (together with the corpus we developed also a specialized lexicon). All these aspects make it possible to use the Polifonia textual corpus for different purposes, from computational studies aimed at analyzing how the language of music evolved over time and across countries to more specialized studies, for example the life and influence of the Italian musician Gian Antonio Perti.

One of the main contributions of our work was to create a unified repository in which documents in different languages and from different sources can be easily discovered by means of pointers to the original resources (i.e. Wikipedia or open digital libraries). During the next months we will prepare the second release of the corpus (Deliverable 4.2). It will be further improved, including other documents coming from the Polifonia pilots and cleaning the texts obtained using OCR technologies. We will also provide automatic and manual linguistic annotations that will make the corpus even more useful since it will be possible to use the annotations to create structured queries and to extract information more easily.

Bibliography

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

[2] A. Esuli and F. Sebastiani, "SENTIWORDNET: A publicly available lexical resource for opinion mining," in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). Genoa, Italy: European Language Resources Association (ELRA), may 2006.

[3] A. V. Carlo Strapparava, "WordNet affect: an affective extension of WordNet," in Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon, Portugal: European Language Resources Association (ELRA), 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/369.pdf

[4] 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.

[5] J. Camacho-Collados and R. Navigli, "BabelDomains: Large-scale domain labeling of lexical resources," in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia, Spain: Association for Computational Linguistics, apr 2017, pp. 223–228. [Online]. Available: https://aclanthology.org/E17-2036

[6] H. L. (edited by), "Terminorum musicae index septem linguis redactus," 1980.

[7] S. Oramas, F. Gómez, E. Gómez, and J. Mora, "Flabase: Towards the creation of a flamenco music knowledge base," in Proceedings of thr 16th ISMIR Conference. Malaga, Spain: ISMIR, oct 2015.

[8] R. Schneider, "A corpus linguistic perspective on contemporary german pop lyrics with the multi-layer annotated "songkorpus"," in Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020). Marseille, France: European Language Resources Association, may 2020, pp. 842–848. [Online]. Available: https://www.aclweb.org/anthology/2020.lrec-1.105

[9] M. A. G. Rodrigues, A. de Paiva Oliveira, and A. Moreira, "Development of a song lyric corpus for the english language," Natural Language Processing and Information Systems, vol. 11608, pp. 376–383, 2019.

[10] 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

[11] 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, sep 2007, pp. 629–633. [Online]. Available: https://doi.org/10.1109/ICDAR.2007.4376991

[12] G. Leech, The state of the art in corpus linguistics. Routledge, 2014, pp. 20–41.

[13] T. McEnery, R. Xiao, and Y. Tono, Corpus-based language studies: An advanced resource book. Taylor & Francis, 2006.

[14] W. Teubert, "Corpus linguistics-a partisan view," TELRI-Trans-European Language Resources Infrastructure II, no. 8, 1999.

[15] J. P. McCrae, A. Rademaker, E. Rudnicka, and F. Bond, "English WordNet 2020: Improving and extending a WordNet for English using an open-source methodology," in Proceedings of the LREC 2020 Workshop on Multimodal Wordnets (MMW2020). Marseille, France: The European Language Resources Association (ELRA), may 2020, pp. 14–19. [Online]. Available: https://aclanthology.org/2020.mmw-1.3

[16] F. Bond and K. Paik, "A survey of wordnets and their licenses," in Proceedings of the 6th Global WordNet Conference (GWC 2012). Matsue, Japan: Tribun EU, jan 2012, pp. 64–71.

[17] R. Navigli, M. Bevilacqua, S. Conia, D. Montagnini, and F. Cecconi, "Ten years of BabelNet: A survey," in

Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, Z.-H. Zhou, Ed. Montreal, Québec: Elsevier, aug 2021, pp. 4559–4567, survey Track.

[18] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," Communications of the ACM, vol. 57, no. 10, pp. 78–85, 2014.

[19] P. Savage, "An overview of cross-cultural music corpus studies," Oxford Handbook of Music and Corpus Studies, estimated 2021-2022.

[20] M. Domingues, R. B. Mangolin, L. Catharin, V. D. Feltrim, J. Donini, Y. M. e Gomes da Costa, and I. A. P. Santana, "Music4all: A new music database and its applications," in International Conference on Systems, Signals and Image Processing. Niterói, Brazil: IEEE, jul 2020, pp. 1–6.

[21] E. Alessandri, V. J. Williamson, H. Eiholzer, and A. Williamon, "Beethoven recordings reviewed: a systematic method for mapping the content of music performance criticism," Frontiers in Psychology, vol. 6, p. 57, 2015.

[22] B. Kelly and C. Moore, Music Criticism In France, 1918-1939: Authority, Advocacy, Legacy. Boydell and Brewer, 2018.

[23] C. Erion and M. Maurizio, "Moodylyrics: A sentiment annotated lyrics dataset," in International Conference on Intelligent Systems, Metaheuristics and Swarm Intelligence. Hong Kong, Hong Kong: ACM-Association for Computing Machinery, mar 2017, pp. 118–124.

[24] M. Fell, E. Cabrio, E. Korfed, M. Buffa, and F. Gandon, "Love me, love me, say (and write!) that you love me: Enriching the wasabi song corpus with lyrics annotations," in Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020). Marseille, France: European Language Resources Association (ELRA), may 2020, pp. 2138—-2147.

[25] E. V. Epure, G. Salha, M. Moussallam, and R. Hennequin, "Modeling the music genre perception across language-bound cultures," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, nov 2020, pp. 4765 – 4779.

[26] E. V. Epure, G. Salha, and R. Hennequin, "Multilingual music genre embeddings for effective cross-lingual music item annotation," in Proceedings of the 21st ISMIR Conference. Montreal, Canada: ISMIR, oct 2020, pp. 803–810.

[27] E. I.Dolan, "Review: Mimo: Musical instrument museums online," Journal of the American Musicological Society, vol. 70, no. 2, pp. 555–565, 2017. [Online]. Available: https://doi.org/10.1525/jams.2017.70.2.555

[28] Y. Raimond, S. Abdallah, M. Sandler, and F. Giasson, "The music ontology," in Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007). Vienna, Austria: ISMIR, sep 2007, pp. 417–422.

[29] P. Lisena and R. Troncy, "Doing reusable musical data (DOREMUS)," in Proceedings of Workshops and Tutorials of the 9th International Conference on Knowledge Capture (K-CAP2017), ser. CEUR Workshop Proceedings, vol. 2065. Austin, Texas, USA,: CEUR-WS.org, dec 2017, pp. 1–4.

[30] S. M. Rashid, D. D. Roure, and D. L. McGuinness, "A music theory ontology," in Proceedings of the 1st International Workshop on Semantic Applications for Audio and Music. Monterey, CA, USA: Association for Computing Machinery, oct 2018, pp. 6—-14.

[31] J. Jones, D. de Siqueira Braga, K. Tertuliano, and T. Kauppinen, "Musicowl: The music score ontology," in Proceedings of the International Conference on Web Intelligence (WI 2017. New York, NY, USA: Association for Computing Machinery, aug 2017, pp. 1222—-1229.

[32] G. Fazekas, Y. Raimond, K. Jacobson, and M. Sandler, "An overview of semantic web activities in the omras2 project," Journal of New Music Research, vol. 39, no. 4, pp. 295–311, 2010.

[33] B. Fields, K. R. Page, D. D. Roure, and T. Crawford, "The segment ontology: Bridging music-generic and domain-specific," in Proceedings of the 2011 IEEE International Conference on Multimedia and Expo, ICME. Barcelona, Catalonia, Spain: IEEE Computer Society, jul 2011, pp. 1–6.

[34] Polifonia, "Deliverable D.2.1, Ontology Engineering," Dicember 2021.

[35] P. Cimiano, C. Chiarcos, J. P. McCrae, and J. Gracia, Linguistic Linked Data - Representation, Generation and Applications. Springer, 2020. [Online]. Available: https://doi.org/10.1007/978-3-030-30225-2

[36] A. Gangemi, M. Alam, L. Asprino, V. Presutti, and D. R. Recupero, "Framester: A wide coverage linguistic linked data hub," in Knowledge Engineering and Knowledge Management - 20th International Conference, EKAW 2016, ser. Lecture Notes in Computer Science, E. Blomqvist, P. Ciancarini, F. Poggi, and F. Vitali, Eds., vol. 10024. Bologna, Italy: Springer, nov 2016, pp. 239–254. [Online]. Available: https://doi.org/10.1007/978-3-319-49004-5_16

[37] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. Montreal, Quebec, Canada: Association for Computational Linguistics, aug 1998, pp. 86–90.

[38] K. K. Schuler, VerbNet: A broad-coverage, comprehensive verb lexicon. University of Pennsylvania, 2005.

[39] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in The semantic web, Proceedings of the 6th International Semantic Web Conference. Busan, Korea: Springer Berlin Heidelberg, nov 2007, pp. 722–735.

[40] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in Proceedings of the 16th international conference on World Wide Web. Alberta, Canada: ACM-Association for Computing Machinery, may 2007, pp. 697–706.

[41] F. Khan and A. Salgado, "Modelling lexicographic resources using cidoc-crm, frbroo and ontolex-lemon," in Proceedings of the International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage co-located with the Bolzano Summer of Knowledge 2021 (BOSK 2021), ser. CEUR Workshop Proceedings, A. Bikakis, R. Ferrario, S. Jean, B. Markhoff, A. Mosca, and M. N. Asmundo, Eds., vol. 2949. Online (Bolzano, Italy): CEUR-WS.org, sep 2021. [Online]. Available: http://ceur-ws.org/Vol-2949/paper7.pdf

[42] S. Neale, "A survey on automatically-constructed WordNets and their evaluation: Lexical and word embedding-based approaches," in Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

[43] P. Vossen, "Eurowordnet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingualindex," International Journal of Lexicography, vol. 17, no. 2, pp. 161–173, 2004.

[44] E. Pianta, L. Bentivogli, and C. Girardi, "Multiwordnet: developing an aligned multilingual database," in Proceedings of the 1st Global Wordnet Conference (GWC 2002). Mysore, India: Tribun EU, jan 2002, pp. 293–302.

[45] S. K. Sarma, D. Sarmah, B. Brahma, H. Bharali, M. Mahanta, and U. Saikia, "Building multilingual lexical resources using wordnets: Structure, design and implementation," in Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon. Mumbai, India: The COLING 2012 Organizing Committee, dec 2012, pp. 161–170. [Online]. Available: https://aclanthology.org/W12-5113

[46] R. Navigli, "Word sense disambiguation: A survey," ACM computing surveys (CSUR), vol. 41, no. 2, pp. 1–69, 2009.

[47] M. Sigman and G. A. Cecchi, "Global organization of the wordnet lexicon," Proceedings of the National Academy of Sciences, vol. 99, no. 3, pp. 1742–1747, 2002.

[48] J. P. McCrae, C. Chiarcos, F. Bond, P. Cimiano, T. Declerck, G. D. Melo, J. Gracia, S. Hellmann, B. Klimek, and S. Moran, "The open linguistics working group: Developing the linguistic linked open data cloud," in Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia: European Language Resources Association (ELRA), may 2016, pp. 2435–2441.

[49] J. F. Sowa, "Semantic networks," Encyclopedia of Cognitive Science, 2006.

[50] M. Bevilacqua, T. Pasini, A. Raganato, and R. Navigli, "Recent trends in word sense disambiguation: A survey," in Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21, Z.-H. Zhou, Ed. Montréal, Québec: International Joint Conferences on Artificial Intelligence Organization, 8 2021, pp. 4330–4338. [Online]. Available: https://doi.org/10.24963/ijcai.2021/593

[51] 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 (EMNLP). Online: Association for Computational Linguistics, nov 2020, pp. 3528–3539. [Online]. Available: https://aclanthology.org/2020.emnlp-main.285

[52] A. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," Communication methods and measures, vol. 1, no. 1, pp. 77–89, 2007.

[53] K. Krippendorff, "Reliability in content analysis: Some common misconceptions and recommendations," Human communication research, vol. 30, no. 3, pp. 411–433, 2004.

[54] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," 2020.

[55] F. Shi, M. Teplitskiy, E. Duede, and J. A. Evans, "The wisdom of polarized crowds," Nature human behaviour, vol. 3, no. 4, pp. 329–336, 2019.

[56] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. D. Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, V. Plachouras, T. Rocktäschel, and S. Riedel, "KILT: a benchmark for knowledge intensive language tasks," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, jun 2021, pp. 2523–2544. [Online]. Available: https://aclanthology.org/2021.naacl-main.200

[57] 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, nov 2020, pp. 7833–7845. [Online]. Available: https://aclanthology.org/2020.emnlp-main.630

[58] R. Panocová, The vocabulary of medical English: A corpus-based study. Cambridge Scholars Publishing, 2017.

[59] R. I. Kittredge, "Semantic processing of texts in restricted sublanguages," in Computational linguistics. Pergamon, 1983, pp. 45–58.

[60] 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, 05 2021. [Online]. Available: https://doi.org/10.1162/tacl_a_00379

[61] I. Culjak, D. Abram, T. Pribanic, H. Dzapo, and M. Cifrek, "A brief introduction to opencv," in 2012 Proceedings of the 35th International Convention, MIPRO 2012. Opatija, Croatia: IEEE Computer Society, may 2012, pp. 1725–1730. [Online]. Available: https://ieeexplore.ieee.org/document/6240859/

[62] E. A. Santos, "OCR evaluation tools for the 21st century," in Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers). Honolulu, Hawaii: Association for Computational Linguistics, feb 2019, pp. 23–27. [Online]. Available: https://www.aclweb.org/anthology/W19-6004

[63] D. van Strien, K. Beelen, M. C. Ardanuy, K. Hosseini, B. McGillivray, and G. Colavizza, "Assessing the impact of OCR quality on downstream NLP tasks," in Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 1, A. P. Rocha, L. Steels, and H. J. van den Herik, Eds. Valletta, Malta: SCITEPRESS, feb 2020, pp. 484–496.

[64] S. K. Manaal Faruqui, "Multilingual open relation extraction using cross-lingual projection," in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics, may-jun 2015, pp. 1351–1356.

[65] Z. Zhong and D. Chen, "A frustratingly easy approach for entity and relation extraction," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021, pp. 50–61. [Online]. Available: https://aclanthology.org/2021.naacl-main.5

[66] E. F. T. K. Sang and F. D. Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003,

W. Daelemans, Ed. Edmonton, Canada: Association for Computational Linguistics, may-jun 2003, pp. 142–147.

[67] X. Pan, B. Zhang, J. May, J. Nothman, K. Knight, and H. Ji, "Cross-lingual name tagging and linking for 282 languages," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics, jul 2017, pp. 1946–1958.

[68] A. Rahimi, Y. Li, and T. Cohn, "Massively multilingual transfer for ner," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, jul 2019, pp. 151–164. [Online]. Available: https://www.aclweb.org/anthology/P19-1015