ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

CORSO DI LAUREA MAGISTRALE IN SPECIALIZED TRANSLATION

Design, Creation and Evaluation of PodIT: A Corpus of Spoken Italian in the Media

Tesi di laurea magistrale in

CORPUS LINGUISTICS

Relatrice

Prof.ssa Silvia Bernardini

Presentata da

Joanna Giacobbe

Correlatori

Dott. Daniele Polizzi

Prof.ssa Maja Miličević Petrović

Sessione marzo 2026

Anno Accademico 2024/2025

Alma Mater Studiorum Università di Bologna

DIPARTIMENTO DI INTERPRETAZIONE E TRADUZIONE

Corso di Laurea magistrale Specialized Translation (classe LM - 94)

TESI DI LAUREA

in

Corpus Linguistics

Design, Creation and Evaluation of PodIT: A Corpus of Spoken Italian in the Media

CANDIDATA:

Joanna Giacobbe

RELATORE:

Silvia Bernardini

CORRELATORI

Daniele Polizzi

Maja Miličević Petrović

Anno Accademico 2024/2025

Terzo Appello

Table of contents

Introduction

A corpus is large collection of samples of authentic use, selected to be representative of a whole language or language variety. A corpus is the fundamental tool for corpus linguistics, a branch of linguistics that studies language through real-life examples of language use. The aim of the present dissertation is to design and present a sample corpus of general spoken Italian as used in the media: PodIT. PodIT (named after "podcast" and "Italian") is a 100,000 words corpus composed of instances of spoken language collected from Italian podcasts. The corpus was built with a combination of automatic and manual processes, which aimed at achieving the highest efficiency with limited time and resources. The corpus will constitute a part of a general corpus representative of contemporary Italian from the 2020s. PodIT positions itself in the landscape of Italian corpus resources and aims to be a resource for Italian corpus linguistics, to analyze and study phenomena in spoken Italian used in the media. This thesis addresses both theoretical and practical aspects of corpus linguistics and more specifically corpus building.

Chapter 1 serves as a background chapter, as its objective is to provide an overview of the main aspects of corpus building, corpus design and existing corpus resources. It has two parts. The first part describes what a corpus is, focusing on its defining characteristics, such as balance and representativeness. Subsequently, techniques and methods to create a valid corpus are explored, such as sampling and annotation. The second part of the chapter focuses on the Italian language and its corpus resources, in order to review past experiences and to identify the gap that this corpus is going to fill.

Following the literature review, chapter 2 describes the method used in creating PodIT. The chapter addresses corpus design choices that define the features of the end corpus and ensure its representativeness and balance, such as the source of language, the definition of the domain and the sampling frame. Successively, decisions on the matters of text collection, transcription and annotation are described, with a focus on their practical applications, advantages and limitations.

Chapter 3 focuses on the description and analysis of the corpus. First, various statistical analyses on the contents of the corpus are presented, in order to highlight the corpus assets. The chapter proceeds with an analysis of the corpus. The analysis was done to both explore the corpus, its features and potentialities, and prove the corpus suitability as a tool for analysis of both spoken language in the media and general spoken language.

1. Background and literature review

This dissertation describes the process of designing and compiling a corpus of spoken Italian. The present chapter will first offer an introduction on corpus linguistics and choices in corpus design, with a particular focus on past experiences in the field. Later in the chapter, the Italian linguistic landscape and the state of the art of Italian corpora will be presented.

The corpus that this thesis is concerned with is part of a larger PhD project that focuses on constrained varieties of language and issues of authenticity, currently under development at the Department of Interpreting and Translation (DIT) of the University of Bologna. Its objective lies, among others, in the creation of a general corpus representative of contemporary Italian from the 2020s, providing a language resource that can be used as a reference for human-authored texts at a time where AI language increasingly influences the way we construct and understand text[1]. Spoken language, of which the corpus described here represents a section, currently plays a privileged role in this respect, as it constitutes a useful yardstick for comparison, being less extensively affected by AI-generated content than written language (see Yakura *et al.* 2024 for a discussion on LLM's influence on spoken communication through repeated exposure to AI-authored texts).

1.1 Corpus linguistics and corpus design

Corpus linguistics is, as the name suggests, a branch of linguistics that relies on corpora, or large collections of naturally-occurring texts. Corpus linguistics analyzes language based on *real life examples of language usage*, through the use of corpora, therefore, leveraging empirical evidence. Although this kind of analysis is not new, the term for corpus linguistics was only coined in the late eighties of the 20th century. Since then it has benefitted from many technological advantages, such as the development of more powerful computers that could handle larger datasets, i.e. larger corpora (McEnery, Xiao and Tono, 2006).

Henceforth, the term *text* will not solely refer to written texts, as it will also include spoken texts.

1.1.1 What is a corpus?

In corpus linguistics a corpus is described as a *large* collection of samples of *authentic* use, selected to be *representative* of a whole language or language variety (Stefanowitsch, 2020, pp.

[1] Empirical evidence suggests that more than 50% of all data available online is machine-generated, whether through automatic translation or as a result of AI generation (cf. Thompson *et al.*, 2024)

22–23). A corpus is usually also machine-readable and nowadays often annotated, two features that facilitate its consultation. A corpus should be *large* since corpus linguistics is based on quantitative evidence, therefore particular instances of language use are counted and analyzed against each other, instead of accounted for as individual cases. However, there is no ideal size for a corpus: an adequate size is relative to its aim. A corpus that aims to represent English on the web should be much larger in order to grasp the full breadth of its language, compared, e.g., to a corpus aiming to represent Shakespeare's style. A related notion is that of representativeness. A corpus is *representative* when it can be used as the basis for generalizations on a language (or language variety) (Biber, 1993). A corpus does not normally contain all the occurrences of a given language (unless the variety is clearly delimited, as in the case of Shakespeare above): such a resource would be almost impossible to create. Rather it should contain texts that make up a representative sample of that language. The third and most important feature is the *authenticity of language* use: in other words, the instances of language included in the collection have to be "real" or produced with a communicative purpose in a real context, rather than being created ad hoc for the corpus or for exemplification purposes (Stefanowitsch, 2020), for instance in language learning textbooks by linguists and language professionals.

Corpora nowadays are employed for various purposes, such as discourse analysis (Taylor, 2014), translation studies (Xiao and Yue, 2009) and language learning (Boulton, 2017), among others. In this section the main distinctive characteristics of a corpus will be listed in order to present the differences between existing corpora and contextualize the object of this study.

Language(s). One key characteristic that distinguishes one corpus from another is the language it represents; each corpus aims to represent a language or a variety thereof. This is the most important feature of a corpus, as it is the first to be set up and will influence any future decision. Although this aspect might initially seem trivial, choosing the language under study also includes thoroughly understanding and describing it in order to be able to represent it. A corpus that represents one language is called a monolingual corpus; conversely, a corpus that represents more than one language is referred to as a multilingual corpus. A multilingual corpus can be parallel, when it contains both texts and their translations, or comparable, when it contains texts in two languages that are comparable in terms of domain or language variety, but are not direct translation of each other. Parallel translation corpora are popular among translators and translation scholars, as they provide a large collection of source-target

translation instances which could be analyzed to investigate translation hypotheses or used in order to create translation memories or terminology resources.

Comprehensiveness. A second characteristic is whether the corpus is a general or a specialized one. General corpora aim to broadly represent a whole language and they include a rather wide range of topics and text types. Specialized corpora on the other hand, have the aim of representing a specific language variety, differentiated, among others, by topic (medicine, politics, sport) or genre (novels, newspapers).

Mode of communication. Corpora can differ from each other based on the medium of the language they aim to represent. This distinguishes written corpora from spoken corpora. Written corpora usually collect instances of language such as books, periodicals, newspapers, academic articles. Spoken corpora, on the other hand, collect face-to-face conversations, interviews, lectures, or political debates.

Time frame. The time frame of the corpus distinguishes two types of corpora, namely a sample (or snapshot) corpus, and a monitor corpus. A sample corpus aims to represent or "take a snapshot" of a language in a specific span of time (McEnery and Hardie, 2011), e.g., American English from 1990 to 1999. Opposite to the sample corpus, there is the monitor corpus. A monitor corpus is a resource that grows over time, as it has the aim of monitoring language changes across time.

Based on the definitions provided above, the corpus that this thesis is concerned with is a **monolingual general sample corpus of contemporary Italian spoken in the media**. Different language varieties in Italian will be further outlined in the following chapter; however, for the purposes of this thesis, *Italian spoken in the media* refers to a language variety of Italian, which is spoken and produced in media contexts.

Currently, several corpora of written language exist, ranging from general ones, like the Brown Corpus (Kučera and W. Nelson, 1967), to less general ones like the corpora in the TenTen Corpus Family (Jakubíček *et al.*, 2013) (since they only contain language from the web), to highly specific ones, like the American Movie Corpus (Forchini, 2021). However, spoken language is underresourced, as spoken corpora (or spoken parts of general corpora) are usually smaller than the written ones (Knight and Adolphs, 2022). This leads to neglecting a large portion of language, and could result in corpus linguistics focusing solely on written data. Spoken language is time-bound and intangible, unless it is recorded and transcribed. Recording and transcription processes are usually challenging and time consuming, although recent technological advantages, especially in automatic text transcription, have made them somewhat easier. Analyzing spoken language can provide useful insights into how a language is spoken

in actual day-to-day conversations, including dialects and slang, highlighting how grammar changes from textbook to reality. That is because spoken language is for the most part spontaneous and unedited and thus reveals information on how a specific group of people speaks on a daily basis and on phenomena that are only present in oral form. It is the case of linguistic phenomena that present in Italian regional dialects, such as phonological, morphological and syntactical phenomena (Canalis, 2006; Loporcaro and De Angelis, 2009).

Before proceeding with the design and compilation of an Italian spoken corpus, the main steps involved in corpus design will be discussed in the following section, both in theory and by presenting past salient experiences in corpus building.

1.1.2 Corpus design

Corpus design is the first and most crucial step in corpus creation, as it determines the characteristics and end goals of the corpus. Designing a corpus entails choosing, considering time and resource limitations, the type of corpus which is to be built taking into account the language(s) or language variety featured as well as the time frame, text length (whether full or sampled) and overall size in tokens (Atkins, 1992).

To make sure that a corpus will ultimately serve our purpose(s), several other issues must be considered, such as representativeness and balance, sampling and metadata. Due to the absence of a standard fit-all method, corpus design is strictly evaluated on a case-by-case basis; the same choice could be both right and wrong, and it is the compiler's task to decide the best course of action depending on their final objective. The following section will present some of the most important aspects to consider when compiling a corpus.

1.1.3 On representativeness and balance

Representativeness is a key characteristic of a corpus, as it guarantees to the user that their analysis will lead to conclusions which can be then generalized to the entirety of the language represented in the corpus (Biber, 1993). In other words, corpus linguists make conclusions about a language (or language variety) based on the assumption that the corpus is a representative sample of the language under study (also referred to as *population*); if the corpus is not representative, the results will only be true for that specific sample and cannot be generalized. The term *population* in corpus linguistics indicates the entirety of the language that the corpus aims to represent. There is no fixed way to achieve representativeness, as it can only be achieved and evaluated in relation to the aim and research question of the corpus that the compiler aims to build. The whole collection of Dante's works might be fully representative if

its aim is to represent Dante's writing, however it wouldn't be if its aim was to describe Italian literature.

When talking about a representative or balanced corpus, most scholars reference the Brown Corpus (Francis and Kučera, 1964) and the British National Corpus (Burnard, 2007), although, as Leech states, there is no way to demonstrate that even those corpora actually are representative (Leech, 2007). The Brown Corpus was the first computer-readable corpus of modern English; it was built in 1963-1964 and features 1 million words. The texts contained in the corpus are samples of 500 books of English prose published in 1961. It is usually agreed upon that the Brown Corpus is representative of written American English (Biber, 1993; Leech, 2007) because of the precision with which the population (written American English) was sampled. The corpus contains different text types in different proportions, for instance, 48 texts pertain to the "popular lore" category, and 6 texts to "science fiction". These proportions are based on the Brown University Library and the Providence Athenaeum, and, while they are representative and balanced in regard to that population, the different text types are in a way "imbalanced" with respect to each other.

The second corpus that is usually mentioned when discussing representativeness is the British National Corpus. The British National Corpus, also known as the BNC (BNC Consortium, 2001), was built from 1991 to 1994 and first published in 1994. The corpus was built with the aim of representing the British variety of the English language, including spoken language. Its text selection methods are similar to those used in the Brown corpus, meaning that the different categories are not balanced with each other. For instance, 75% of texts are informative texts, and 25% are imaginative texts. Library lending statistics and bestseller lists were used to make decisions on which texts to collect for each category.

Although no rules exist on how to achieve representativeness, it is a corpus compiler's job to "strive" for it, as Sinclair puts it (Sinclair, 2005). The concept of representativeness has been extensively discussed with respect to corpus building, and it is frequently mentioned along with the concepts of balance and sampling (for a discussion on sampling see section 1.1.5). The most widely accepted definitions of representativeness usually tie it to balance. For instance, in McEnery, Xiao and Tono (2006) it is stated that "Representativeness is typically achieved by balancing the corpus through sampling a wide range of text categories which are defined primarily in terms of external criteria."

More specifically, McEnery claims that balance can be defined as the range and proportion of text categories included in the corpus, which should be taken from the target population during sampling. Therefore, according to McEnery, a corpus is not representative

unless it is balanced, and balance is achieved by mirroring in the corpus the proportions of the population through sampling. The issue with this description is the overlap and circularity of the concept of representativeness and balance, which defies the purpose of having two separate concepts. Leech proposes a circular definition as well, as he states: "for a corpus to be balanced is an important aspect of what it means for a corpus to be representative" (Leech, 2007, p. 136). A further common issue is the abstract nature and vagueness of some of the definitions of both representativeness and balance. Sinclair refers to both of the concepts as "not precisely definable nor attainable goals" (Sinclair, 2005). His explanation as to how to achieve balance is the following: "Roughly, for a corpus to be pronounced balanced, the proportions of different kinds of text it contains should correspond with informed and intuitive judgements" (ibid.) which explains very little about the steps one should take when compiling a corpus. To sum up, most definitions of representativeness and balance are either unclear, abstract, or overlapping.

Defining and sampling the population is useful to understand which text types to include in the corpus, however the concept of only using sampling to get the proportions can be problematic, since the proportions, and therefore the sections of the corpus, will not be balanced with each other. Some sort of "internal balance", although it is rarely mentioned in the methodology of corpus building, is also a useful characteristic of a corpus. Conducting analyses on an entire corpus to extract linguistic information is a common and valid type of analysis; however, analyzing the corpus by comparing its sub-corpora is an equally important type of analysis. If the corpus is balanced internally, meaning its sub-corpora are of similar size, the user will be able to conduct a truthful comparative analysis within the corpus.

As mentioned above, the definitions of representativeness and balance usually overlap. In theoretical corpus linguistics, the best examples of representative and balanced corpora are the previously mentioned Brown Corpus and British National Corpus. In addition to the previously mentioned meaning of balance, there seems to be a second meaning, much similar to the concept of "internal balance", which was mentioned earlier.

The Corpus of Contemporary American English (COCA) (Davies, 2008) is often framed as a "balanced" or "genre-balanced" corpus. It was built by Mark Davies in 2008, with the aim of representing American English. COCA is a monitor corpus, meaning it continuously gets updated with new material, and, at the time of writing, contains more than 1 billion words, with data spanning from 1990 to 2019. Similarly to the BNC, COCA also has a spoken section. The content of the corpus is equally divided across years and text types, with 3-4 million words collected per year for each text type category: TV and movies, spoken, fiction, magazine, newspaper and academic. Every text type contains between 3 and 4 million words for each year.

This design choice contradicts the established theoretical concept of balance seen above, and exemplifies another one, a *genre balance*, or *genre-based organization* (Davies, 2010). Davies ensure the previously introduced concept of "internal balance" by balancing each sub-corpus based on genre. Davies claims that the COCA is "the first balanced monitor corpus", and a reliable one for comparative analysis precisely because of its internal balance. In his view, the internal balance of the corpus is a way to ensure reliability of a corpus for comparative analysis.

On the basis of this overview, the concepts of representativeness and balance will be redefined for the purposes of this thesis and for the creation of this specific corpus. Henceforth, *representativeness* will encompass both the concepts of representativeness and balance as described by McEnery (McEnery, Xiao and Tono, 2006) and Leech (Leech, 2007), namely the range of categories of texts that will be included in a corpus, these categories could be represented by features such as text types, genres or topics. *Balance* on the other hand, will refer to the internal balance of the corpus, i.e., the feature whereby the end corpus is structured into comparable sub-corpora, allowing comparative analyses of different text varieties predefined by the corpus creator.

This different definition of balance benefits the user. Representativeness is still key to the effectiveness of corpus use; however, the possibility of conducting several types of analyses broadens the range of uses of the end corpus. While creating a representative corpus is still relevant today, and every corpus compiler should strive for it, a corpus does not exist for the sole sake of representing a language. With the advance of corpus linguistics, most of the times a corpus will be used to conduct linguistic analyses of subsets or varieties, and a way to ensure that these analyses can be done, is by creating a resource that is internally balanced, meaning that its different components are comparable to each other. There is no doubt that the proportions in the COCA corpus do not accurately represent the population of American English texts: for instance, the spoken section of the corpus is equal in size to other sections. However, COCA has proven to be an extremely useful resource for linguistic analysis (Davies, 2010), thanks to its precision in harmonizing sections within the corpus.

In the following chapters, corpus design decisions will be discussed, with internal balance having a fundamental role in the design and collection of the end corpus. The theoretical concepts of representativeness will not be disregarded; however, as Biber puts it (emphasis added):

> "In choosing and evaluating a sampling frame, considerations of **efficiency** and cost effectiveness must be balanced against higher degrees of representativeness."
>
> (Biber, 1993, p. 244).

1.1.4 Describing the domain and operationalizing the domain

A crucial step the corpus compiler must take before attempting to sample the population that the corpus aims to represent, is understanding that population. For this purpose, one must define it. Following Egbert, Biber and Gray (2022, p. 70), in this section, the population will be referred to as the *target domain*, namely "the full domain of language use that exists in the real world". The process of sampling and defining a population or target domain is a difficult task, since they usually include many text categories and types, and they usually do not have clear-cut boundaries. In order to set boundaries and select texts for a corpus, the corpus compiler can define an *operational domain*. The operational domain is a "detailed operational definition for the domain that designates the set of texts that will be candidates for inclusion in the corpus" (Egbert, Biber and Gray, 2022, p. 70). The operational domain will then be sampled following the criteria presented in section 1.1.5.

1.1.5 Sampling

In corpus linguistics, the term *sampling* refers to the way a corpus is derived from a population. To be more specific, sampling is the process of obtaining a scaled-down version of the population under study. Through sampling the corpus compiler understands which text types to include in the corpus and in which proportions, in order for it to be representative (McEnery, Xiao and Tono, 2006).

A useful sampling process is proposed by Sinclair (2005) where he describes the idea of *criteria* and *cells*. To compartmentalize both the population of interest and the end corpus sampled from it, a corpus compiler can use criteria that can vary for each corpus. However, some common criteria include: mode, type, domain, language, location and date (Sinclair, 2005).

From the definition of the starting criteria, it is possible to then create cells from their intersection. The first subdivision of the population must be binary, namely public or private, spoken or written, news or fiction; successively, each category can be split into more categories, in order to capture the greatest variation possible.

This type of sampling is also referred to as *stratified sampling* (Biber, 1993). In stratified sampling, the population is first divided into homogeneous groups (the *strata*), and then the texts inside the groups are randomly sampled. An example of this type of sampling can be found in the Survey of English Usage and in the London-Lund Corpus of Spoken English. For the Survey of English Usage (SEU) written and spoken language data were collected at University College London. Texts were obtained, tagged and made machine-readable by the Survey of Spoken English (SSE) project at Lund-University. The original SEU corpus was structured and described using a tree-like representation (see Figure 1). This diagram provides a visual representation of what sampling a language entails.

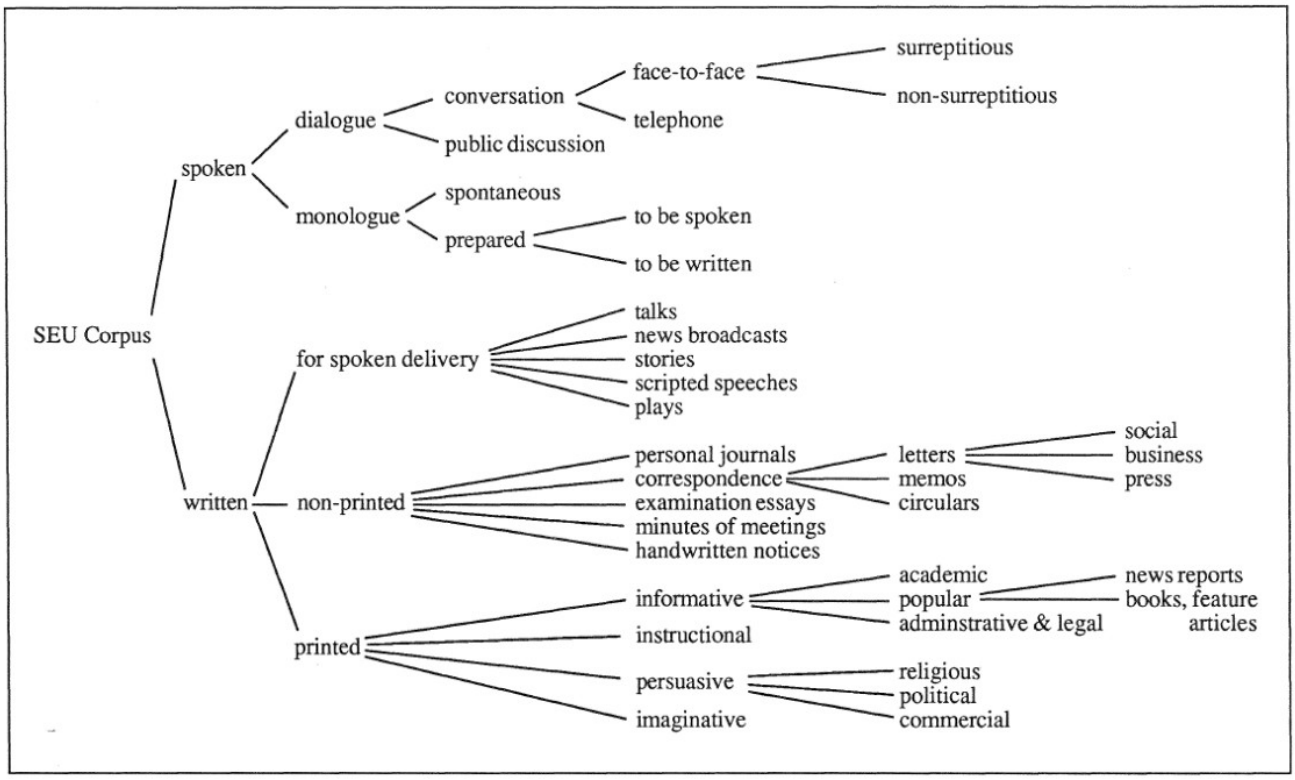

Figure 1. Tree diagram of the SEU corpus (Svartvik, 1990).

Sinclair's cell concept, the stratified sampling and the tree branch diagram of the SEU corpus serve a dual purpose. Primarily, these methods were introduced here as means to describe a population and the corpus itself. These practices are also useful both for documenting the choices made throughout creation and for the user to understand the corpus composition. A second purpose of these methods is to guide the compiler during the creation of the corpus. As a matter of fact, trying to sample a large category (e.g., spoken conversation, Figure 1) top-down is much harder than sampling each outmost branch or section (e.g., telephone or face-to-face conversations, Figure 1) bottom-up in order to fill the general category.

1.1.6 Text sampling

A further step in corpus compiling is text sampling, namely, how much of a single text should be included in the corpus. Scholars have long debated on text sampling, as it is one of the choices that affects size and variability of the corpus.

Text sampling is the process of taking an extract of a given number of words from a text and only including that amount in the corpus. There are several reasons why a corpus compiler might want to include a sample of a text rather than the full piece. One reason, related to licenses, is that a compiler might face copyright issues when including the whole work in their corpus (McEnery, Xiao and Tono, 2006). Other reasons are related to individual corpus design choices. For instance, in the Brown corpus, all the texts were sampled, and only a sample of 2,000 words was included in the corpus. These samples start on a randomized page at the beginning of a sentence, and end at the first sentence ending after 2000 words. Some scholars agree that text sampling is a reasonable and useful practice in corpus compilation. For instance, including chunks of equal length for every text can prevent one speaker's style or topic to be overly present, therefore leading to skewed results. Furthermore, it has been shown (Biber, 1993) that generally, frequent linguistic features tend to show in large enough samples. The optimal length of a sample, according to Biber (1993), is 2,000 words.

However, some scholars advise against text sampling. Sinclair (2005), for example, states that the best decision in terms of text sampling is to not sample at all. The reason behind this choice is that linguistic features are dependent on their position in the text. From this perspective, in order to solve the potential issue of imbalance in the data, the compiler should make a large enough corpus, which could dilute longer texts. In the case where the corpus compiler is dealing with copyright issues, Sinclair suggests they try to reach a compromise with the copyright holder, by including the full texts with occasional omissions, so that the value of the document is diminished.

As discussed in the previous sections, there is no "gold standard", and each decision on corpus design matters depends on the corpus' aim and objectives. For instance, text sampling might be pointless for specific kinds of corpora, such as corpora built for literary analyses of a single writer. Other cases might require sampling, but the length of the starting material does not allow the collection of 2,000 words samples.

1.1.7 Contextual information and linguistic annotation

An essential feature of modern corpora is the presence of contextual information, namely, "data about data". *Contextual information* refers to information that is recorded about the texts contained in the corpus. When compiling a corpus, it is crucial to not only collect and store texts but also to record information about those texts. Recording this kind of information has multiple purposes, for instance, it serves the corpus compiler during the creation of the corpus, as it helps efficiently categorizing the texts in an ordered and clear manner. However, a corpus that is rich in contextual information also benefits the end user, since it can be used to perform contrastive analyses across text types and sociolinguistic factors. Information can be recorded at the level of the text, providing information about the text as a whole, such as the year of publication, text type or topic; additional information can also be recorded at the level of the speaker or author, capturing social backgrounds, such as age, gender or occupation (Love *et al.*, 2019; McEnery and Brookes, 2022). In modern corpora, contextual information is usually stored at the beginning of a text in XML format, and it is referred to as metadata. The following section aims to describe the basic elements of XML and was adapted from the TEI guidelines (TEI Consortium, no date).

XML stands for *eXtensible Markup Language*, which is a set of markup conventions used to encode texts. A textual unit (e.g., text, sentence) is referred to in XML as an *element*. In order to tag, or define, such element, it must be *tagged*. To tag an element, a start- and an end-tag are inserted at its beginning. *Start-tags* are enclosed in angle brackets and *end-tags* are enclosed in angle brackets with the addition of a slash after the opening angle bracket. What is contained inside of an element is known as the *content* of said element. For instance,

```
<sentence>Hello world!</sentence>
```

is an XML element. The content "Hello world!" is enclosed in a start-tag "<sentence>" and an end-tag "</sentence>. This convention is useful for encoding sentences, paragraphs and even whole documents. In addition, elements can have *attributes* and *values* which are used to describe and store information about the element. Adding attributes to elements is a very useful practice in modern corpus building, since it adds meaningful information that can be useful for later analyses. Attributes-value pairs are stored inside of the element's start-tag:

```
<text year="2024" author="Anna" author_age="25">
Hello world!
</text>
```

In this case, the attributes for the element "text" are "year", "author" and "author_age", and their respective values are "2024", "Anna" and "25".

Nowadays, corpora contain a substantial amount of contextual information in the form of metadata, which facilitate its use and exponentially increases possibilities of analyses. By combining the use of metadata and modern tools, such as *Sketch Engine*[2], the researcher can conduct many types of analyses by creating sub-corpora within a corpus. Metadata is particularly useful in spoken contexts, where language can vary greatly based on social characteristics and backgrounds of the speaker. The Desert Island Discs Corpus (Marchi and Ferraresi, 2004) is a great example of a spoken corpus that makes large use of metadata. It contains transcriptions of episodes of a radio program airing on the BBC, namely, "Desert Island Discs". The corpus contains metadata about the texts (e.g., year of publication) and metadata about the speaker (e.g., name, gender, age, profession, political orientation) (Figure 2). By leveraging this kind of metadata, it is possible to conduct contrastive analyses; for instance, comparing if and how gender or political orientation influences the speaker's language use.

Metadata field type	Metadata field name	Metadata field value
On host & guest	Host's name	E.g. Sue Lawley
	Guest's name	E.g. Douglas Adams
	Guest's gender	Female, male, nonbinary, transgender female, transgender male
	Guest's date of birth	E.g. 1952-03-11
	Guest's year of birth	E.g. 1952
	Guest's year of death	E.g. 2001
	Guest's age at date of recording: exact	E.g. 41
	Guest's age at date of recording: generation	E.g. 40
	Guest's place of birth	E.g. Cambridge
	Guest's country of birth	E.g. United Kingdom
	Guest's country of citizenship	E.g. United Kingdom
	Guest's occupation: category	academic, activist, actor, architect, artist, broadcaster, businessperson, chef, comedian, dancer, designer, director, engineer, explorer, farmer, gardener, journalist, lawyer, medical personnel, military, misc, model, musician, photographer, politician, producer, religion personnel, scientist, sportsperson, trainer, writer
	Guest's occupation: first mentioned	E.g. playwright
	Guest's occupation: all	E.g. playwright; screenwriter; novelist; science fiction writer

Figure 2: Portion of the contextual information collected in the Desert Island Discs corpus[3]

Linguistic annotation refers to the process of annotating linguistic information within the text (McEnery and Hardie, 2011, pp. 29–30). This is done to enrich the corpus of

[2] Sketch Engine is a powerful online software built for corpus linguistic analysis. It allows the user to access corpora and use tools to explore them, such as Concordance, Wordlist, Keywords, Term extraction. Available at: https://www.sketchengine.eu/, last visited, February 20, 2026

[3] https://docs.sslmit.unibo.it/doku.php?id=corpora:desert_island_discs_corpus, last visited February 12, 2026

information regarding parts of speech, phrases, sentences and dependencies. A corpus that is linguistically annotated allows its user to perform specific and more complex searches and analysis. For instance, with linguistically annotated corpora, it is possible to analyze language by searching parts of speech, rather than unique words or lemmas. Linguistic annotation can be performed both manually and automatically. As technical innovations make their way in corpus linguistics and corpus building, automatic taggers and parsers become more solid and reliable. Some examples of automatic taggers include CLAWS (Garside, Leech and Sampson, 1987), which was used to tag the BNC corpus, and TreeTagger (Schmidt, 1994), a language-independent part-of-speech (or POS) tagger.

In this section, the main concepts of corpus building were presented; reviewing the concepts has set the basis for the second chapter of this thesis, where the corpus design decisions are presented. The following section is an overview providing some notions on the Italian language, its history and its modern-day characteristics. While not in any way exhaustive, this overview will offer a brief history of the formation of the language and a description of the current features of the language, focusing on its varieties and dimensions.

1.2 The current Italian landscape

1.2.1 The Italian language

The Italian language, due to its history and formation, is a complex and stratified language. Italian has historically been characterized by the presence of many dialects of Latin, called *volgari romanzi* (Romance vernaculars). In the fifteenth century, because of the prestige obtained from its literary tradition, the Florentine dialect was codified as *standard Italian*, while every other vernacular became a geographical dialect of Italian (Berruto and Cerruti, 2011). This contributed to the creation of a complex sociolinguistic landscape which held true for centuries; standard Italian became the language of institutions, while most of the Italian population kept speaking their own dialect, a phenomenon known as *diglossia*. However, since the eighties of the twentieth century, Italian linguists have identified a second shift in language, namely, the evolution of the *neostandard Italian*. Neostandard Italian is the result of the convergence of standard Italian and regional dialects (Ballarè, 2020), mostly caused by industrialization and related migratory flows and the diffusion of media such as radio and television (Zingaro, 2024). According to this configuration, standard Italian was used by speakers of different social backgrounds and in informal settings as well; simultaneously, features of dialects started to be incorporated into spoken Italian. Neostandard Italian can

nowadays be described as a variety of language that is spoken (and written) by educated speakers in moderately controlled contexts. However, neostandard Italian did not replace standard Italian, which is still widely present in controlled and formal social settings, such as academia and bureaucracy (Ballarè, 2020).

1.2.2 The dimensions of the Italian language

The Italian language is also characterized by the presence of language varieties, namely types of language that differ due to social and extralinguistic factors or situational contexts (Berruto and Cerruti, 2011, p. 278). Different language varieties can be described based on the dimension they are related to. These dimensions are the *diatopic*, *diastratic*, *diaphasic* and *diamesic* dimensions (Berruto and Cerruti, 2011). The diatopic dimension is related to geographical context: the language varieties differ based on the geographical origin of the speaker. This dimension is particularly relevant for the Italian language and mainly affects phonetics and lexis (Berruto and Cerruti, 2011, p. 279). Diatopic varieties of the Italian language are the so-called regional Italians (my translation) (ibid., p.279), varieties spoken in different regions which present different linguistics phenomena, for instance, varieties of Italian spoken in Tuscany, Campania, Sicily. The diastratic dimension depends on the social stratum, namely, the social and cultural background and level of education of the speaker(s) involved. For instance, the language used in a conversation which involves graduate students will be different to that used in a conversation which involves uneducated individuals. The diaphasic dimension is influenced by the communicative context in which the language occurs. For instance, diaphasic varieties include high and low registers and technical languages, such as the so-called "burocratese" or "legalese" (legalese) (Treccani, no date), that is, the language used by professionals in legal contexts. The diamesic dimension is connected to the mode and channel of communication. For instance, it distinguishes spoken language from written language.

In the current section, past experiences in the field of Italian corpus linguistics, more specifically corpus building, will be studied to review what has already been done, take inspiration, and create the best possible corpus design. Salient experiences with spoken and written corpora will be presented, with a particular focus on their approaches and methodologies. This section does not aim to offer a comprehensive list of every Italian corpus ever made; rather a list of corpora is provided including the ones that are either salient or relevant for the project that this thesis is concerned with. In order to compile the list, relevant papers on Italian corpora (Barbera, 2013) were read, and thorough research was performed on

search engines like Google Scholar and other aggregator websites. Successively, information about each corpus was accessed on relative papers or on the corpus' web page.

The number of written Italian corpora is not large, compared to those of English, both because of corpus linguistics' disciplinary traditions and because English has a much larger number of speakers and resources, compared to Italian. As a starting step, before designing and create a new corpus, past experiences must be considered in order to have a broader understanding of the Italian current landscape and thus avoid duplication of effort and learn from best practices in the field.

CORIS/CODIS

The CORIS/CODIS[4] is a corpus of written Italian, which was created between 1998 and 2001 by Rema Rossini Favretti and Fabio Tamburini at the University of Bologna (Rossini Favretti, 2000). The corpus gets updated every three years and is available for online consultation. The corpus is made up of two parts: CORIS, which stands for *Corpus di Riferimento dell'Italiano Scritto* (reference corpus of written Italian, my translation) was built as a reference corpus for the analysis of written Italian. As of 2021 the corpus contains 165 million words divided amongst press, academic prose, legal texts and letters. Parallel to that corpus, CODIS was also built; CODIS stands for *Corpus Dinamico dell'Italiano Scritto* (dynamic corpus of written Italian, my translation), which was built in order to let the user select sub-corpora and exclude or include different text types to conduct linguistic analysis.

itWaC

ItWaC[5] is a written Italian corpus compiled at the universities of Trento and Bologna as part of the Wacky project between 2005 and 2007 (Baroni *et al.*, 2009). The corpus contains 2 billion words collected from web pages, which were crawled using seeds taken using medium-frequency words from the *La Repubblica* corpus (a corpus of Italian newspaper texts (Baroni *et al.*, 2004)). ItWaC is part of the Wacky corpus family, a group of corpora which includes corpora derived from the English, French, German and Italian web.

[4] https://corpora.ficlit.unibo.it/coris_ita.html, last visited January 22, 2026
[5] https://docs.sslmit.unibo.it/doku.php?id=corpora:itwac, last visited January 22, 2026

itTenTen

The itTenTen[6] corpora, part of the TenTen corpus family (Jakubíček *et al.*, 2013), are the largest Italian corpora to date. The TenTen corpus family contains three corpora of Italian, compiled in 2010, 2016 and 2020. The texts were collected from the web and, since all corpora were sampled using the same criteria, they are all comparable with each other, diachronically and across languages. The most recent corpus was built in 2020 and it contains 12.4 billion words.

PAISÀ

PAISÀ[7] is part of the PAISÀ project, which stands for "Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati" (platform for learning Italian on annotated corpora, my translation) (Lyding *et al.*, 2014). The corpus only contains texts released under the Creative Commons license, which were downloaded from the web in September and October 2010. Other than being a resource for analyzing the Italian language, the corpus was also designed as a resource for learners of Italian, especially second and third generation Italians. In order to build a resource for learners, the compilers thoroughly annotated the sentences with morphological, lexical and syntactic information. Furthermore, each sentence is annotated with information about structural dependency, allowing the user to generate a graphical representation of dependency relations. Since the corpus is also intended for learners of Italian, it is also possible to restrict query results, in order to only view sentences of limited complexity (Lyding *et al.*, 2014, p. 41).

Araneum Italicum

Araneum Italicum[8] is a corpus of Italian compiled by Vladimír Benko in 2014 as part of a corpus family called *Aranea* (Benko, 2014). The corpora were compiled in Slovakia and include languages spoken and taught in Slovakia. All the texts were derived from the web with the aim of creating comparable corpora for multilingual linguistic research and lexicography. The Italian corpus contains 120 million tokens.

[6] https://www.sketchengine.eu/ittenten-italian-corpus/, last visited January 22, 2026
[7] https://www.corpusitaliano.it/, last visited January 22, 2026
[8] http://ucts.uniba.sk/aranea_about/_italicum.html, last visited January 22, 2026

1.2.3 Corpora of spoken Italian

Although limited, spoken corpora of Italian offer substantial language data and tools to perform linguistic analysis on spoken language. Existing Italian spoken corpora must be studied and taken into consideration in order to identify the gap that this corpus is going to fill.

LIP/VoLIP

The corpus VoLIP[9] is a continuation of a previous project, the LIP corpus project, which was carried out between 1991 and 1994 (Voghera *et al.*, 2018). The LIP corpus contained 500,000 words and it was created with the objective of representing the Italian spoken language, highlighting its varieties. It includes face-to-face conversations, calls, interviews, debates, monologues and radio programs, which were collected in 5 different Italian cities, in order to take diatopic differences into consideration. The LIP corpus was then converted into the VoLIP corpus in 2014; here audios and phonetic transcriptions were associated with the preexisting corpus. In this way the user can look for lemmas, see the word in context and listen to the recording. Furthermore, the user can also navigate the corpus through a wide range of metadata like gender, town, social context, medium. VoLIP is a useful resource for Italian linguistics, as it allows the user to analyze spoken language both as a text and as a recording, allowing researchers to investigate regional pronunciation and intonation.

KIParla Corpus

The KIParla Corpus[10] (Mauri *et al.*, 2019) is one of the most recent projects concerned with spoken Italian. Its compilation began in 2016 at the universities of Bologna and Turin and it was first published in 2019. As of 2025 it contains 4 modules, a total of 2,900,000 tokens, and it is being constantly updated. The recordings for the first modules were made in Bologna and Turin, two highly multicultural cities, which allowed the compilers to collect data from Italian speakers of different regional origin. A large part of the recordings belongs to a single communicative context, namely face-to-face conversations, mainly obtained in university settings. The data was then transcribed, including intonation, pauses and prosodic elements. In addition to texts, KIParla collects contextual information about the conversation(e.g., length, setting) and information about the speaker, (e.g., gender, age group, education and origin) ("Corpus KIParla – L'italiano parlato e chi parla italiano," no date).

[9] https://www.volip.it/, last visited January 22, 2026
[10] https://kiparla.it/, last visited January 22, 2026

It-tok

It-tok is a specialized corpus of spoken Italian, which focuses on the language which occurs on the social media platform TikTok (Troncone, 2025). The corpus was compiled by Luisa Troncone at the University of Salerno. The corpus is an ongoing project started in 2025 and its main aim is collecting a resource to analyze the properties of spoken Italian by focusing on social media. The videos that were selected for the corpus vary in topic, however they are mostly relative to social themes, namely politics, environment and social rights.

What emerges from this brief overview is that the Italian language has limited but useful resources for corpus analysis. Written corpora cover many sources of language, including language from the web, which is of increasing importance. Spoken corpora as well are great tools for corpus analysis, however, except for It-tok (Troncone, 2025), there does not seem to be a large interest in new forms of language and communication. Both VoLIP (Voghera *et al.*, 2018) and KIParla (Mauri *et al.*, 2019) cover private conversations and interactions, but little is done to address public spoken language, the type of language spoken in the media and public entertainment. The corpus that this thesis is concerned with will have the aim of covering this gap in Italian resources and shedding light on more aspects of the Italian language.

2. Corpus creation

The corpus that this dissertation is concerned with is a monolingual corpus representative of contemporary spoken Italian in the media. More specifically, the corpus contains transcriptions of samples of podcasts spread across different topics as well as speakers of different origins and backgrounds, from the year 2020 to the beginning of 2026. Because language data is solely taken from podcasts, the corpus can and should be taken as a reference for public spoken Italian or Italian spoken in the media. However, because of the general spontaneity of language and colloquial tone of most podcasts, it can also be used as a reference for spoken Italian as a whole (see section 2.3).

In this chapter, the theoretical aspects of corpus design presented in the previous chapter are used to motivate and discuss the decisions made in creating this corpus. These design choices were all guided by the final objective of producing a representative and topic-balanced corpus suitable for corpus linguistics and corpus-assisted discourse analysis.

2.1 Domain

As discussed in chapter 1 compiling a representative corpus requires a thorough description and subsequent operationalization of its language domain. The domain under study for this corpus is **contemporary spoken Italian in the media**. According to a 2015 study[11] by ISTAT (the Italian National Institute of Statistics), around 90% of Italian residents speak Italian as their native language. Most Italians use it in combination with regional dialects, to varying degrees based on their social background and communicative setting. As a result of the convergence between standard Italian and regional dialects, the aforementioned neostandard has emerged as a new, dynamic variety and the *de facto* language of everyday communication. In line with this development, a recent study on spoken regional Italian signals the rise of a "composite Italian" (my translation), a new variety born from the convergence of multiple regional varieties (Cerruti, 2018).

New phenomena are also emerging through the use of the Italian language in the media and online. Studies show that written language use changes in online settings, such as blogs or text messages, acquiring characteristics that are typical of spoken language (Bonomi, 2010; Pistolesi, 2018). Spoken Italian in radio and in television are described by linguists as unique varieties, which resemble spontaneous speech while displaying medium-specific characteristics

[11] https://www.istat.it/en/press-release/the-usage-of-italian-language-dialects-and-other-languages-in-italy-year-2015/, last visited February 21, 2026

(Sabatini, 1982; Spina, 2006; Maraschio, 2011). Table 1 sums up the defining characteristics of the domain as well as the sub-categories that it includes.

	Domain
Sources	Italian (Berruto and Cerruti, 2011) Neostandard Italian (Ballarè, 2020; Zingaro, 2024) Italian in the media (Sabatini, 1982; Spina, 2006; Bonomi, 2010; Maraschio, 2011; Pistolesi, 2018)
Description	Spoken Italian which occurs in the media from the year 2020 to 2026. A variety of the Italian language which is characterized by its spoken nature and by the context in which it occurs, namely, the media.
Categories	Italian used in television Italian used in radio programs Italian used in podcasts Italian used in social media

Table 1: Description of the domain, spoken Italian in the media

2.2 Operational domain

In order to create a corpus adequate for the envisaged uses, it is crucial to define an operational domain, namely, define the source of language data that will be used to compile the corpus. The operational domain is a clear and definite pool of language data, from where selected specimen will be obtained through sampling. Prior to selecting and defining the best operational domain, in this section corpora that contain language data from media sources are reviewed, with a view to identifying best practices in the field. Examples from both English and Italian are considered.

The Spoken English Corpus (SEC) was built around 1984-1985 by IBM and the University of Lancaster (Taylor and Knowles, 1988). The corpus is made up of recordings of speakers from the BBC, including commentaries, news broadcasts, lectures, poetry, and propaganda; all of the recordings are spoken in an accent which is as close as possible to Received Pronunciation (RP). The Corpus of North American Spoken English (CoNASE) (Coats, 2023), built in 2023, was entirely made up of transcripts of YouTube videos posted by administrative identities. The spoken sub-corpus of the COCA corpus by Mark Davies collects transcripts of recordings of TV and radio programs. Ultimately, the corpus It-tok (Troncone, 2025) is a corpus of spoken Italian which collects language from the social media platform Tik-

Tok. The corpus collects both videos addressing themes of interest for the public debate (e.g., politics, environment), and videos with no specific themes.

Nowadays, spoken media communication happens through several different channels, such as radio, television and social media. Therefore, a corpus compiler that aims to collect language spoken in the media has a large array of strata and data sources to choose from. For the corpus that this thesis is concerned with, the chosen source of language data was **Italian podcasts**. This corpus design choice will be addressed in the following section, where its advantages and limitations will be presented and discussed. The language of podcasts shows characteristics that are generally present in "public" language, more specifically of language occurring in digital media (for the analysis see section 3.4). For this reason, podcasts were deemed representative of the entirety of language spoken in the media.

Collecting language data for corpus building from podcasts has several advantages. Podcasts were chosen as a source for the corpus due to their increasing popularity among listeners and their rise in the media landscape. Compilers of corpora are often interested not solely in the language that is being produced, but also in the language that is being received (Stubbs, 1996). A recent study by NielsenIQ (NIQ) reveals that, as of 2025, 18 million Italians listen to podcasts; 62% listen to them weekly, while 16% listen to them daily; furthermore, in the past years, the number has risen (TG24, 2025). Because of their increasing popularity, analyzing podcasts could give useful insights into the kind of language that media consumers are exposed to.

A further reason why collecting language data from podcasts has advantages is their availability. A wide range of podcasts are available for listening on the internet, either provided by broadcasting services such as radios and TVs, or by music streaming services, as well as aggregator sites and platforms. Statistics on Italian podcasts are not available; however, recent statistics show that over 4.5 million podcasts are available worldwide in hundreds of languages.

In addition to availability, a positive asset of podcasts is variability; similarly to other types of media contents, podcasts differ widely from one another in terms of topic, format, and style. Variability is also increased because virtually everyone can create a podcast. Unlike radios and TV programs, which are usually created and hosted by radio and TV hosts or entertainment professionals, podcasts are free to post on most streaming services such as Spotify[12] or Apple Podcasts[13]. This broadens the range of possible social backgrounds of the speaker, thus representing a broader portion of the population.

[12] https://open.spotify.com/, last visited January 23, 2026
[13] https://podcasts.apple.com/us/new, last visited January 23, 2026

Following the choice of podcasts as the source for language data for the corpus, the name PodIT was chosen, derived from the word *podcast* and the word *Italian*.

2.2.1 Definition of the operational domain

Considering what has been presented in the previous section, the source of language for the creation of PodIT was chosen to be Italian podcasts. Nowadays, the term *podcast* has several different definitions, especially since the medium has evolved through time with new technologies; some of these definitions will be presented in this section. In order to define clear-cut boundaries and facilitate the compilation process, an operational definition of podcast was created. The definition was created starting from available dictionary and encyclopedia definitions and scholarly articles. The term *podcast* was coined by Ben Hammersley in 2004, by combining the word *pod*, from apple's mp3 player *iPod*, and *cast* which was taken from the word *broadcast*. The Merriam-Webster[14] dictionary defines a podcast as:

> A program (as of music or talk) made available in digital format for automatic download over the Internet.

Collins English dictionary[15] proposes this definition, which uses the radio as a comparison example:

> A podcast is an audio file similar to a radio broadcast, that can be listened to on a website or app on your phone, computer, etc.

The Italian encyclopedia Treccani[16] defines "podcast" as (my translation):

> A digital audio file shared through the internet and available on a laptop or MP3 player […].

From this brief dictionary overview, it is possible to conclude that two of the constant characteristics of a podcast are its audio format and its availability on electronic devices. However, these definitions lack some features of modern podcasts. As a matter of fact, the medium has evolved in the past years, and is often now more than just an audio file (Balanuta, 2021). Similarly, some aspects of the previously presented definitions are obsolete; the Merriam-Webster dictionary defines a podcast as a piece of media content that can be downloaded automatically from the internet, which is no longer true, since the majority of podcasts nowadays are only available for online streaming. Furthermore, alongside audio, nowadays some podcasts include captioning for accessibility purposes, and it is becoming

[14] https://www.merriam-webster.com/dictionary/podcast?src=search-dict-hed, last visited February 21, 2026
[15] https://www.collinsdictionary.com/dictionary/english/podcast, last visited February 21, 2026
[16] https://www.treccani.it/enciclopedia/podcast/, last visited February 21, 2026

increasingly popular to attach a video content element to the audio podcast (Balanuta, 2021), sometimes also leading to a change in the name itself: from *podcast* to *vodcast* or *video podcast*. This poses a challenge for the definition of podcast, as it might no longer be distinguishable from a visual blog (vlog) or a generic video product. A usable definition of podcast was proposed by Balanuta (2021), who defined a podcast as:

> An on-demand listening experience, mediated through audio or video platforms, which involves heterogeneous formats and generous thematic designs that can be authored by producers of multiple backgrounds.

This definition is more inclusive of modern podcasts; therefore, it was adopted as the starting point for the operational definition of podcast that was used for the corpus. By combining the aforementioned definition and some aspects of the previously cited dictionary entries, the resulting operational definition of podcast for corpus compilation reads as follows:

> A podcast is an on-demand listening experience, mediated through audio or video format, accessible for streaming or download through the internet, which can be authored by producers of multiple backgrounds.

This definition manages to encompass the large majority of podcasts available on the internet.

2.2.2 Description of the operational domain

In this chapter, the operational definition will be segmented and further analyzed, in order to set clear boundaries, that is, inclusion and exclusion criteria for what constitutes a podcast. At the same time, this allows to provide an initial mapping of the degree of variability within the operational domain itself.

On-demand listening experience. *On-demand listening experience* refers to one of the fundamental assets of podcasts, namely, that it can be listened to at any time and place. This is true for most podcasts; however, in some cases, a podcast might consist of recordings of TV or radio programs (e.g., No Spoiler[17], Auditorium[18]), which are made available by broadcasters to listen to on-demand. While these do technically classify as podcasts, the language used in live television might differ from the one that is used in podcasts. In order not to introduce potential

[17] https://www.deejay.it/podcast/no-spoiler/stagione-2026-di-no-spoiler/hbo-max-i-record-di-zalone-le-origini-di-gomorra-e-la-nuova-tomb-raider/, last visited February 4, 2026
[18] https://www.raiplaysound.it/programmi/auditorium, last visited February 4, 2026

noise in the corpus, recordings of live radio and TV programs were excluded from the operational domain.

Mediated through audio or video format. This part of the definition refers to the type of medium used to mediate the content of a podcast. Although the original file type used to share podcast was an audio file, there have been changes in recent years. More and more podcasts are being uploaded on platforms that allow or require the attachment of a video component, like YouTube[19], Facebook[20], and Twitter (now X[21]) (Sullivan, 2019). Furthermore, in recent years, existing platforms for audio streaming, such as Spotify, have started to allow the upload of videos as well. By including podcasts mediated through video, the operational domain becomes broader and more inclusive. However, some rules and boundaries had to be set in order not to mistakenly include in the corpus content that, following our definition, is not classifiable as a podcast. In order to be eligible for selection, a video podcast has to be understandable also when deprived of its visual component; this is quite difficult to assess, if not by watching the whole video. Therefore, during the collection process, various methods were used to determine whether or not a video fit the definition of video podcast. The first one relies on cross-checking between media platforms. Video podcasts were mostly collected from video platforms like YouTube or YouTube Music, which are platforms where thousands of regular videos get uploaded daily. A video was classifiable as a video podcast if it was uploaded on audio platforms as well, e.g., Spotify or Apple Podcasts. If a video podcast only appeared on video platforms, a second method was used to identify video podcasts that fit the operational domain. Video podcasts were only included if the video consisted of a static image, e.g., the cover of the podcast, a photo of the speaker, or if the video consisted in one or more people speaking in a physical or virtual environment (see figure 3 for an example), with minimal or no location changes. During the collection, if the validity of a video as video podcast was dubious, it was excluded from the sampling, in order not to compromise the corpus.

[19] https://music.youtube.com/, last visited January 26, 2026
[20] https://www.facebook.com/, last visited February 4, 2026
[21] https://x.com/, last visited January 4, 2026

Figure 3: Example of a video meeting the operational domain for a video podcast[22]

Accessible for streaming or download through the internet. In order to be included in the operational domain, a podcast had to be freely accessible through the internet. This part of the definition was added because sample collection only took place on the internet, since podcasts can be listened to on numerous platforms and websites. A single podcast is usually available from many websites, like media streaming platforms, as well as official radio and TV channels or podcast aggregators. An online podcast is typically composed of several platform-independent structural elements. Figure 4 shows the web page of a podcast on the web version of Spotify. A web page of a podcast usually shows the title, the author or producer, a cover image and a brief description of the podcast. Similarly, Figure 5 shows the web page of a podcast episode, which includes the title of the episode, the title of the podcast it belongs to, its date of publication and the cover image, which can be either the cover used for the whole podcast or a unique cover for that specific episode. The episode as well is often complemented with a description, in which the podcast creators briefly describe its content. The interface layout is unique to each platform; however, the content often remains the same.

[22] https://open.spotify.com/episode/3QAtbolQSwUmpcJBB4c7tI, last visited January 2, 2026

Figure 4: Web page of a podcast on Spotify[23]

Figure 5: Web page of a podcast episode on Spotify[24]

Authored by producers of multiple backgrounds. This part of the definition clarifies that the operational domain has no limits in terms of authorship of the podcast. As a matter of fact, as previously stated, the fact that podcasts can be created by virtually everyone is a positive aspect for the corpus, as it ensures the most possible variation in terms of social backgrounds of the speaker(s). Therefore, the operational domain includes podcasts produced both by experts, such as radio and TV hosts or content creators, and by amateurs.

[23] https://open.spotify.com/show/4uV5zn8IEenL2YbIj6ehpK, last visited January 2, 2026
[24] https://open.spotify.com/episode/3FaxJcXzvWwFbyMBQZRUmr, last visited January 2, 2026

2.3 Representativeness and balance

The PodIT corpus was built in order to be, and it is deemed to be representative of spoken Italian in the media, since podcasts contain language whose characteristics reflect those of public spoken Italian and Italian spoken in the media. Furthermore, in line with similar initiatives (e.g., the spoken section of the COCA corpus, which is made up of radio programs), PodIT can be used as a general reference for spoken Italian. As argued by Davies (Davies, no date) responding to claims about the unnaturalness of radio programs' conversations, no recorded conversation is entirely natural, unless the speaker is unaware of being recorded. Therefore, language collected from sources like radio programs, can be considered as a valid source of spoken language data as any other.

As discussed in the previous chapter, representativeness and balance are two crucial concepts that ensure quality and usefulness of a corpus. After outlining the characteristics of the domain, it is the compiler's job to understand how to sample it in order to achieve representativeness and balance. The concepts of representativeness and balance follow the definitions presented in the previous chapter (section 1.1.3). Representativeness refers to the range of categories of texts that will be included in the corpus; balance refers to the internal balance of the corpus. In order to achieve representativeness, the main descriptive criteria were devised following an analysis of the domain and operational domain. The criteria are:

1. type of conversation: monologic or dialogic;
2. type of content: professional or non-professional;
3. topic addressed: general.

These criteria were then used as a basis to create the sampling frame (section 2.4). The criteria were used as a guide to keep track of which and how many texts to include.

Similar to the COCA corpus, this corpus is internally balanced, which means that the proportions in which the texts were collected do not reflect the actual proportions of the operational domain. Rather, the corpus was built by balancing its sections; to be more specific, monologic and dialogic podcasts, and professional and non-professional ones, have similar proportions. Moreover, the corpus also has a *topic-balance*, that is, topics are proportional to each other.

2.4 Sampling

As presented in the previous chapter, sampling is one of the most crucial aspects of corpus design, which ensures that the corpus is representative of the chosen language or language

variety. The sampling frame that was used to describe the corpus was inspired by the London-Lund Corpus sampling frame (Svartvik, 1990). By using the criteria presented in the previous section, it is possible to create a tree representation of the sampling frame (Figure 6). The corpus has an initial division between monologic and dialogic podcasts; then each branch is further split into professional and non-professional podcasts. These branches are then split into topics. The corpus includes no more than one episode from each podcast, in order not to collect evidence of language use from the same speaker more than once. For the same reason, episodes belonging to different podcasts, but featuring a speaker that was already present in another samples podcast episode, were also excluded. A further limit was added to the operational domain, which concerns the year of publication of the podcast episode. Since the aim of this corpus is to collect instances of contemporary spoken language, only episodes published between 2020 and 2026 were considered.

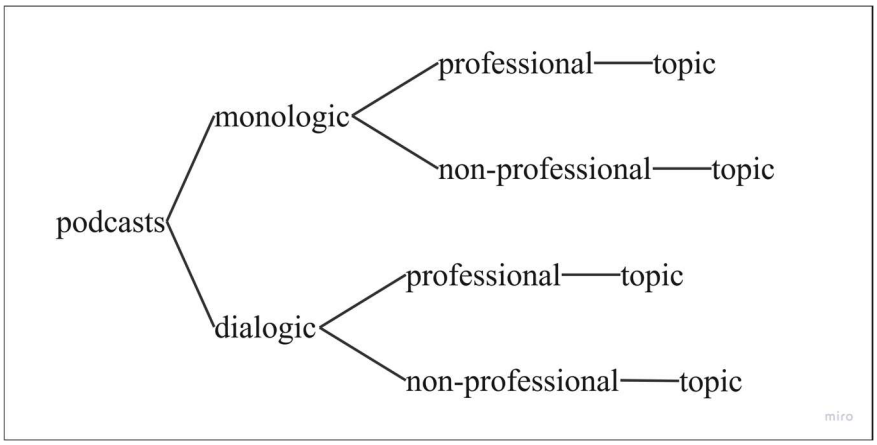

Figure 6: Sampling frame

The monologic/dialogic variable refers to the number of speakers in the podcast. Instances of both monologues and dialogues were included in the corpus in order to achieve representativeness. Other corpora (such as the BNC or the COCA corpus) have collected spoken language both in typically monologic formats (news broadcasts, sermons) and in typically dialogic formats (face-to-face or phone conversations), signaling a necessity to analyze both kinds of settings. Monologic and dialogic podcasts were balanced; therefore, the two sections of the corpus will contain roughly the same number of texts.

The professional/non-professional descriptive variable was added due to the nature of the chosen medium. Due to time and resource limitations, it was not possible to investigate each podcast included in the corpus to assess if the creator is a professional or an amateur. However, a trend was observed, which was then turned into a categorization rule. Most video and audio streaming platforms, such as Spotify, Apple Podcasts or YouTube have little to no restrictions

as to who can publish a podcast. Since it is possible to find podcasts authored both by professionals and by amateurs, the fact that a podcast is present on one of those platforms was not considered an indicator of it being professional. At the same time, many radio and tv channels provide podcasts (e.g., RaiPlay Sound[25], Radio DeeJay[26]), many creators are backed by podcast producer companies (e.g., OnePodcast[27], Chora Media[28]), and more and more companies and institutions, such as newspapers or publishing houses (e.g., Il Messaggero[29], Feltrinelli[30]), also publish podcasts. Podcasts created by entertainment professionals or experts tended to be backed up or published by established institutions, while amatorial podcasts were usually self-published on the streaming platforms mentioned above.

For the purposes of this corpus, professional podcasts were defined as the ones that appear on either tv or radio channels or that are produced by podcast producer companies or professional companies and institutions. Similarly, non-professional podcasts were defined as those that appear exclusively on media streaming platforms that allow self-publishing, and are not backed up by companies or institutions. This categorization rule was used to facilitate the collection process and is not an indicator of the general professional nature or validity of a podcast. Furthermore, the categorization is solely based on production context, and does not imply that the speakers of "non-professional" podcasts lack experience or competence in the field of entertainment and podcast production.

2.4.1 Topic selection and description

In this section, choices made in terms of topics to include will be discussed. Topic variation is an important matter in corpus building, as it contributes to ensuring a higher degree of representativeness (section 1.1.3). Corpora that have a significantly high concentration of one topic risk not being representative of a whole language, rather of the language used when addressing that specific topic. PodIT is a *topic-balanced* corpus, which means that the different topics appear in similar proportions to each other. The process of selecting topics will be described in the current section.

Most podcast streaming platforms use categories to sort their podcasts and episodes by creating a page, or a section of a page, that divides them by topic. This division not only

[25] https://www.raiplaysound.it/, last visited January 26, 2026
[26] https://www.deejay.it/, last visited January 26, 2026
[27] https://www.onepodcast.it/, last visited January 26, 2026
[28] https://choramedia.com/, last visited January 26, 2026
[29] https://www.ilmessaggero.it/, last visited January 26, 2026
[30] https://www.lafeltrinelli.it/, last visited January 26, 2026

improves the listening experience for the user, as it simplifies the exploration of various themes, but also simplified the collection process of podcasts for this project.

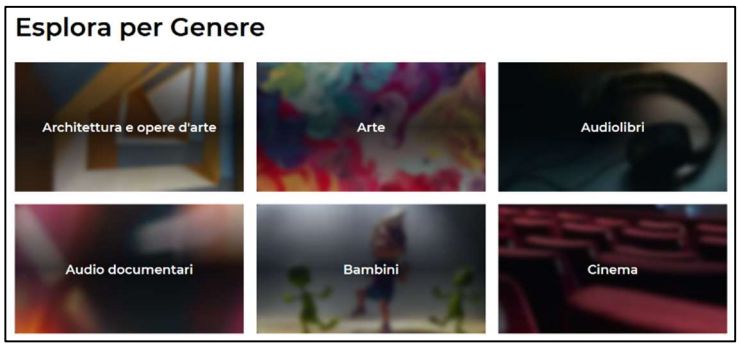

Figure 7: Snapshot of the categories on RaiPlay Sound[31]

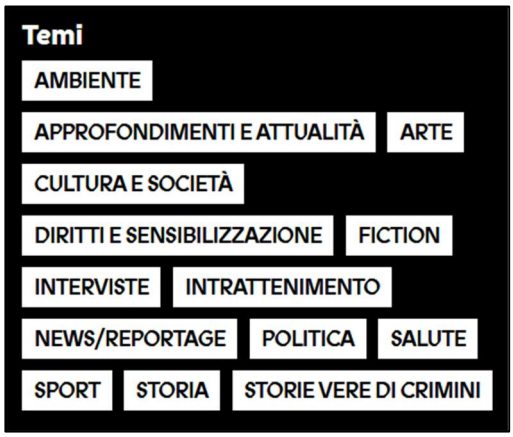

Figure 8: Snapshot of the categories on Chora Media[32]

In order to create a comprehensive list of topics to include in the corpus, lists from five podcast sites were compared, namely: Podcastiamo[33], Podcast Italia[34], Radio DeeJay, Chora Media, RaiPlay Sound. Podcastiamo is an Italian website that stores and makes available a large number of Italian podcasts for listening. Podcast Italia is a website that selects and ranks the most popular podcasts by topic and redirects the user to their original website. Radio Deejay is an Italian radio that also produces podcasts. Chora Media is an Italian podcast producer company. RaiPlay Sound is the radio and podcast section of RAI, the Italian national public broadcasting company.

After a brief analysis of the five podcast sites, it was found that the division by topics was not standardized, and each site sorted their podcasts in different manners. The decision was made that if a topic appeared in three or more sites, it was to be included in the sampling. This was done to make sure that the most widely recognized topics were included in the corpus. In

[31] https://www.raiplaysound.it/generi, last visited January 16, 2025
[32] https://choramedia.com/podcast/, last visited January 16, 2025
[33] https://www.podcastiamo.it/, last visited January 26, 2026
[34] https://podcast-italia.com/, last visited January 26, 2026

order to make the selection process easier, topics were merged into topic categories. Almost all the categories were selected from the available topics; however, a few modifications were made to the final list. Some categories were removed; other categories were merged based on conceptual proximity, such as “fiction” and “cinema”, which were converted to “cinema and tv”, or “science” and “technology”, which were merged into “science and technology”. The categories included in the corpus will be presented and described below.

Art. This category includes podcasts about visual arts, such as painting, sculpture and architecture, it includes interviews to experts in the field and commentaries by art enthusiasts. This category does not include podcasts about literature, since they are included in a different category.

Chat. The category “Chat” was one of the most problematic to define. While investigating the domain on various platforms, it was noticed that a large number of podcasts feature people talking about their lives and narrating their thoughts and experiences. Therefore, the category “chat” was created, which includes personal interviews with famous or non-famous people addressing life events and personal experiences.

Cinema and TV. This category was created by combining the topics of film and tv series. It includes commentaries on film awards, critiques and analyses of films and TV series from both experts and amateurs.

Culture and society. “Culture and society” includes podcasts about literature, history, historical figures and general knowledge. It also includes podcast episodes that address society and societal issues.

Crime. Crime podcasts, according to a recent study by NIQ, are the most listened to podcasts by the Italian public (TG24, 2025). These podcasts include narrations of crimes, often murders, that took place either in Italy or abroad, including descriptions of events, investigations and trials. The creators are often true crime enthusiasts, lawyers, criminologists or psychologists.

Economy. The category “Economy” was chosen and used as a broad container of podcast episodes about economy, finance, investments and the job market. Episodes included in this category can be interviews with experts in which they talk about the economy, or tutorials, for instance on how to invest, deal with economic issues or save money.

Health. A large number of podcasts about several aspects of health are present and popular among listeners. Podcast episodes in this category include talks from doctors and popularizers about nutrition, mental and physical health.

Music. "Music" includes podcasts that address various aspects of music, such as music theory, interviews with artists presenting their music, analysis and commentaries of songs, albums and music awards.

Politics. The category "Politics" features podcast episodes that address international and Italian news, such as elections and major political events; it also includes general political theory, political debates and interviews with politicians and diplomats.

Science and technology. This category includes episodes that discuss science (e.g., scientific subjects, such as physics, astronomy, astrophysics, geophysics) and technology (e.g., computer science, new media and new technologies). The category includes interviews with science experts or popularizers.

Sports. The category "Sports" includes episodes about the history of sports or sports competitions, alongside with commentaries on matches or athletes by sports enthusiasts, commentators or journalists. Interviews with athletes were included as well.

Among these topics, podcasts which were destined to a restricted, sectorial public were not included. This was done because the aim of PodIT is to collect instances of language destined to the general public. For this reason, podcasts destined to specific audiences were excluded (e.g., learners, children).

While these categories were built to collect instances of language that are distinct from one another in terms of topic, there might be some level of overlap, since categories do not always have clear-cut boundaries. For instance, overlaps might be present among similar categories, like "Art" and "Culture and society" or "Science and technologies" and "Health". Given that some podcasts might be classified as belonging to more than one category, an educated guess had to be made in order to choose one among the plausible candidates. Furthermore, categories indicate the general topic of the podcast episode; however, podcast speakers often wander off when discussing topics, briefly talking about their lives or their ideas. Division by topic does not fully ensure that the speakers will only address one topic, or that the language used is specific to that particular topic only.

2.5 Text sampling

As presented in the first chapter of this thesis (see section 1.1.6), text sampling, or the selection of an extract of given words from a text to be included in the corpus, is a common practice in corpus linguistics, although some scholars advise against it. The previously presented corpus of spoken Italian KIParla and VoLIP does not employ sampling, as every audio sample has a different length. The reason behind this choice probably derives from the type of language data

collected (i.e., recorded conversation), therefore there is a need to not "waste" language data, since its collection and acquisition is costly and time consuming.

For the creation of this corpus, text sampling was preferred, as opposed to including the whole podcast episode. The choice to only include samples was made in order to eliminate length differences between episodes, and to avoid thematic imbalance.

Although many scholars agree that the optimal sample to capture linguistic features is composed of 2,000 words (Biber, 1993), these measures refer to written language; therefore they are not necessarily applicable to spoken language. In order to find the optimal length of a language sample taken from a podcast, the domain had to be analyzed. From what has been observed during the collection process, most podcasts range from 5 to 60 minutes, with only few being shorter or longer. Length also usually differs based on the topic, with interviews and narrative podcasts being longer than, for instance, book commentaries.

A sample size of minimum 5 minutes was chosen as the standard length of a text sample. This length allows the collection of a good amount of text for each episode, and also does not limit the number of podcast episodes that can be collected. If a longer sample was selected (e.g., 10 minutes) many podcasts would have to be excluded from selection. Decisions about text sampling are not limited to the length of the sample; for instance, the part of the episode from where to extract the language sample was to be established. Collecting all the samples from the start of the corpus could create bias, since the corpus would contain language that is typical of introductions or beginning of conversations. Similarly, collecting language from the end of the podcast episode could result in collecting a large number of set phrases that are used when concluding an episode or a conversation. Collecting language from the middle of the episode was considered as a viable option; however, having variation ensures that the entirety of the medium's language is represented. Therefore, following what has been done in the BNC (BNC Consortium, 2001), samples were collected in equal numbers from the start, the middle, and the end of episodes.

A further decision had to be made, regarding the exact starting point of the audio sample. During the compilation of written corpora, sentence boundaries given by punctuation (periods, exclamation and question marks) are used as guides to mark the beginning and end of a language sample, as done in the Brown corpus (Francis and Kučera, 1964). Spoken language, however, does not have such clear boundaries, and randomly choosing the starting time of the sample might result in it beginning in the middle of a word. To avoid this, the samples all start at the beginning of a sentence-like stretch of text, in order to keep as much meaning as possible;

once the recording passed the 5 minutes mark, it was stopped at the first long pause or end of a sentence.

Considerations were also made about whether to include information that classifies as boilerplate. Boilerplate refers to text that could constitute noise in the corpus (McEnery and Brookes, 2022), such as repeating sentences or expressions that do not carry any linguistic information. It is standard practice in corpus linguistics to exclude this type of textual content from texts before inclusion in a corpus. Some podcasts have some sort of recurring sequence at the beginning or at the end of each episode, which is also where most of the advertisement, if present, is located. An example of boilerplate as a repeating sequence at the beginning and ending of an episode is "Questo è "Amare parole", io sono Vera Gheno, sociolinguista, e saluto chi mi sta ascoltando" (This is "Amare parole", I am Vera Gheno, a sociolinguist, and I greet those who are listening to me" (Gheno, 2024) and "Vi saluto e vi aspetto alla prossima puntata" ("Goodbye, see you in the next episode") (ibid., my translation). These sentences are identical or almost identical in every episode of the series, and were therefore classified as boilerplate. However, not every starting or ending portion of a podcast should be considered boilerplate, since some podcasts either do not have introductions and conclusions, or they are unique to each episode. For this corpus, a starting or ending section of a podcast was considered boilerplate if repeated in every episode. Commercials were also considered boilerplate, since they are usually snippets of pre-recorded commercials for radio or TV that are included in the episode and rarely voiced by the podcast speaker.

After considering what counts as boilerplate, the decision was made to not include it in the corpus; therefore, the recordings were made with an effort to exclude these elements.

2.6 Contextual information

One of the strengths and advantages of modern corpora is the possibility to filter information based on features of the text or text producer. Modern corpora make extensive use of contextual information, stored in the form of metadata, both to record information that would otherwise be lost, and to facilitate more fine-grained analyses. As seen in the previous chapter, including metadata allows the user to perform all kinds of contrastive analyses inside of the corpus. Although recording metadata does benefit the user, it is also good practice for corpus compilation, as it ensures that the corpus is well-constructed, detailed and transparent. For this reason, during data collection, information about the text and the speakers was collected in an excel spreadsheet, which served as the corpus database, to be later appended to the texts. In this

section, contextual information that was collected about the texts will be presented by listing each datum, describing it and defining its purpose.

ID. Each audio and each transcription has a unique ID, which can be traced back to the database of collected podcasts. While the ID does not serve any analytical purpose, it is crucial during corpus collection, as it keeps the data ordered. The ID naming convention was devised by combining the name of the topic category, or its abbreviation, to sequential numbers, for instance "art_01", "art_02", "crime_01", "crime_02".

Year. "Year" refers to the year of publication of the episode. Due to corpus design decisions, only episodes from 2020 to 2026 were collected.

Type of conversation and number of speakers. For every sample it was indicated whether it was a monologue or a dialogue, in order to specify the situational context in which language occurs. The number of speakers involved in the conversation was recorded to further specify what kind of conversation occurs in the episode.

Professional/non-professional. Information about the professional nature of the podcast was also recorded; rules on how this was assessed are described in section 2.4. Language occurring in professional podcasts might vary from that occurring in non-professional ones due to differences in production practices. Professional podcasts can have multiple writers and producers, and might undergo several review and editing stages before being published, which might increase planning, reduce disfluencies and promote standardization.

Topic category. The category signals the main topic category of the podcast episode. The criteria that were used to select categories were described in the section 2.4.1. This datum will allow the corpus user to employ it to create sub-corpora and make comparative analyses.

Length of sample. This information was stored mostly for technical purposes. It is good practice to record information about the sample of language collected. Although it does not serve a straightforward purpose for corpus analyses, the length of the sample could be used for other types of analyses (e.g., to calculate speech rate).

Podcast and episode name and source. The name of the podcast and the name of the episode were recorded for information purposes. The user might want to listen to the whole episode or investigate the podcast as a whole. For the same reason, the source to the web page of the episode was kept as well, so that the user can directly access it. For most podcasts, the source is a Spotify link, which was chosen for two main reasons; first, it is one of the most intuitive media players with a clear display and overall design; second, on its web version, it does not require users to have an account to stream podcasts.

Sample starting point. The sample starting point refers to the part of the episode from where the sample was taken. By recording this information, the user is aware of which portion of the podcast episode was sampled, namely the start, the middle or the end.

Name of the speaker. In addition to information about the text, data about the speaker was also collected, which can be used to perform sociolinguistic analyses. The name of the speakers was recorded solely for transparency and information purposes, so that, if needed, the user can easily look for information about them.

Age, gender and profession. Data about age, gender and profession of the speaker was recorded to allow the user to perform diastratic analyses, namely analyze how language changes across age groups, genders and occupations. When recording age and profession, the preferred source of information was web blogs, web magazines and the speaker's personal profiles. When conflicting information about age and profession was found, the information that appeared on the highest number of websites was selected and recorded in the database. When researching the profession of some speakers, especially celebrities, it happened that more than one profession was listed in biographies. To overcome this issue, if more than one profession was listed, the first one was arbitrarily chosen as the one to be recorded in the database. An example from the corpus is Gianluca Gazzoli. A radio company which publishes some of his works describes him as (my translation): a radio and TV host, a video maker and a storyteller[35]; in the data frame, he therefore appears as a radio host.

Origin and "nativeness". Information about the origin of the speaker was recorded in order to allow the user to perform diatopic analyses. If the speaker was born in Italy, their region of origin was recorded; conversely, if the speaker was not born in Italy, their country of birth was recorded. Furthermore, the database contains information about whether the speaker is a native speaker of Italian or not.

Status in the podcast. "Status" refers to the role that the speaker has in the podcast episode, namely, that of a host or a guest, as this might influence some features of their language.

Recording contextual information has some limitations in terms of data availability. In some cases data about age, occupation and origin of the speaker were not available on the internet. The first and most common reason is that creators of non-professional podcasts do not often share many personal details, like their age or their occupation. In some cases, in both

[35] https://www.deejay.it/conduttori/gianluca-gazzoli/, last visited January 20, 2026

professional and non-professional podcasts, the episode guest is an expert in a field, such as a professor, but they are not active online. In those cases, the field was left empty.

The technical aspect and the addition of metadata in PodIT is addressed in section 2.8.1. Additionally, a table showing the attributes and values as they appear in the corpus is presented in the third chapter of this dissertation.

2.7 Collection and transcription

In this section, the process of creating the corpus will be described. In order to create a corpus of spoken data, various ways are employed to obtain a text version of a spoken event, depending on resources and time constraints. For instance, for the compilation of the spoken section of the BNC, the texts were transcribed manually (Burnard, 2007). A different approach was used for the COCA corpus, where ready-made transcripts of radio and tv programs were obtained by the compiler. While Spotify and Apple Podcasts do offer automatic transcriptions of podcasts, the method of obtaining data by downloading the transcriptions was discarded for two main reasons. Firstly, both Apple Podcasts and Spotify have restrictions on the download of the transcription; Apple Podcasts only allows the user to copy up to 200 words, while Spotify does not allow download at all. Secondly, recordings store meaningful linguistic information (such as linguistic prosody) and will benefit the corpus further development; for instance, recordings could be associated to transcripts for easier consultation.

Having considered the advantages and disadvantages of data collection methods, the process to obtain corpus data was defined: audio samples were collected, automatically transcribed and annotated. In the following sections, these steps will be described.

2.7.1 Collection of language samples

The process of collecting the language samples from podcasts can be divided into two sections, the first one concerns the selection of texts, the second one concerns the collection of the audio sample. The sites (presented in section 2.4.1) were investigated by topic, podcast episodes were chosen manually, and information about them was recorded in the database. The selection process began with choosing a topic and using the query section to search for it in the sites mentioned above; successively, podcasts were investigated individually by reading their titles and their description. A podcast was selected if it fit the definition of podcast (e.g., not an audiobook, not a recording of a live event), if it fit the category that was being selected, and if it was longer than 6 minutes (for sampling reasons). Once the episode was chosen, relevant details were recorded in the database. Podcast episodes were selected while making a conscious

effort to keep a balance between monologic and dialogic ones. The same effort was put into keeping the balance between professional and non-professional podcasts and between male and female speakers.

The following step to the selection process was the collection of audio samples. The length chosen for the audio sample, as discussed in section 2.5 of this chapter, is 5 minutes, while the format chosen was MP3. The MP3 is a widely used format for audios, which, owing to the compression process, stores audio data in a small file (Denegri-Knott, 2015). While the compression process does involve some loss in terms of audio quality, the small size of the files is an advantage. Lighter files are generally easier to share and take less time to upload and process, which facilitates workflow during multiple stages of corpus compilation.

Various methods were tested to record the audio sample in MP3 format from a web player, such as Chrome extensions and apps for screen recording. Chrome extensions (Chrome Audio Capture[36] and Web Audio Recorder[37]), at first, seemed to be intuitive and produce great results. However, the quality of the audio resulted to be lower than other services tested, leading to their exclusion. Two apps for screen recording were also tested and both resulted in high-quality audio samples. First, OBS Studio[38] was used and, while its results in terms of quality were excellent, it could only record the full screen in video format and it did not allow conversion from its video file (MKV) to an MP3 file. Successively, the app BANDICAM was tested and ultimately chosen for the recording process. BANDICAM[39] is a screen recording software for Windows that provides a wide range of services, such as high-quality screen, game and audio recording. The app has a simple and intuitive interface and outputs a high-quality MP3 file. The collection process started by selecting a starting point in the podcast episode that coincided with the beginning of a sentence; the recording on BANDICAM was started using the Start/Stop button (Figure 9). After the interface timer (Figure 9) signaled a recording time of 5 minutes, the recording was stopped at the first sentence ending or long pause. The audio file was then saved as an MP3 file, stored in a folder and renamed after the naming convention described in section 2.6, namely the name of the topic or its abbreviation in combination with sequential numbers.

[36] https://chrome-audio-capture.en.softonic.com/chrome/extension, last visited January 22, 2026
[37] https://chromewebstore.google.com/detail/web-audio-recorder-record/pddjlpangidimliafldcpfkbkifmegbd, last visited January 22, 2026
[38] https://obsproject.com/, last visited January 22, 2026
[39] https://www.bandicam.com/, last visited January 22, 2026

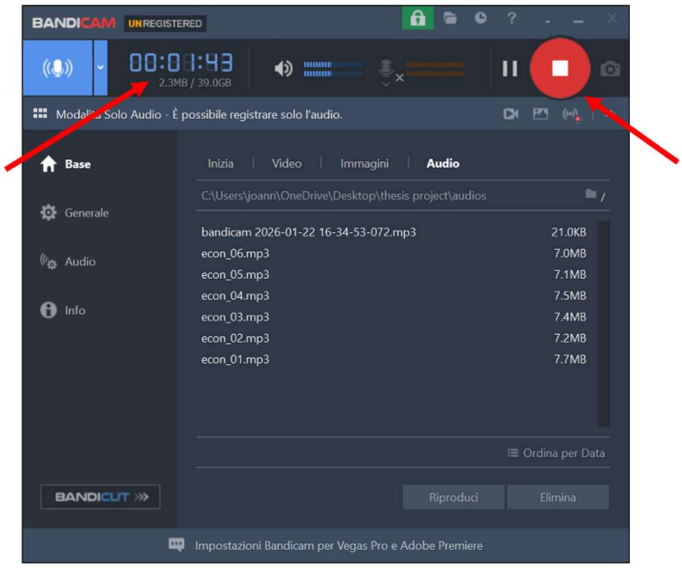

Figure 9: Timer and Start/Stop button in the BANDICAM interface

2.7.2 Transcription

In order to transform an audio sample (in MP3 format) into a text file (in TXT format) to be included in a corpus, transcription had to be employed. Manual transcription was employed in older corpora such as the BNC (BNC Consortium, 2001), however, this kind of transcription is often not accessible due to resources and time constraints.

In order to convert audio files to text files, automatic transcription was employed. Automatic transcription was preferred to manual transcription due to its high-quality output and time efficiency. Transcriptions are made thanks to *Automatic Speech Recognition (ASR)*, which is a technology that converts spoken language into text using machine learning models. Recent development in AI models have also led to improvements in ASR technologies, making speech-to-text conversion accessible. Studies show that open-source models (such as OpenAI's Whisper) provide high-quality transcriptions, which can be refined with minimal editing (Russell *et al.*, 2024, p. 13). In September 2022 OpenAI launched Whisper (Radford *et al.*, 2022), an open-source multilingual ASR system employing an encoder-decoder transformer architecture. Whisper was trained on 680,000 hours of audio, the majority of which cover the English language, while 117,000 hours cover the other 96 languages featured. Whisper offers six models (compared in Table 2), each model has a different size, which influences the time and computing resources needed to produce the transcription, as well as the quality of the latter. In other words, smaller models take less time and resources, however the accuracy of the transcription is lower than that of larger models. In this context, *computing resources* refers to the central processing unit (CPU), the graphics processing unit (GPU) and the random-access memory (RAM).

Model	Relative Speed	Accuracy	Memory Usage	Disk Space
tiny	10x	Lowest	~1GB	~150MB
base	7x	Low	~1GB	~300MB
small	4x	Medium	~2GB	~1GB
medium	2x	High	~5GB	~3GB
large	1x	Highest	~10GB	~6GB

Table 2: Comparison table of Whisper models (adapted from Whisper API, 2025)

In order to choose the best model for the transcriptions of the audios, computational resources and expected quality were considered. The required quality of the transcription had to be high, in order to reduce the amount of noise present in the corpus. Therefore, the tiny, base and small models were excluded. When choosing between the medium and the large models, computational resources and time limitations were considered. The large model offers professional transcriptions with minimal errors; however, it requires high computational power, slowing down the transcription process. A great balance between output quality, time limitations and computational power can be achieved using the medium model, since it is preferred for "[h]igh-quality transcriptions where accuracy is important and you have decent computing resources" (Whisper API, 2025).

Whisper was released as an open-source library, which allows it to be implemented in various computational environments. For this project, Whisper was accessed through Google Colaboratory (Colab)[40], which is a hosted Jupyter Notebook service; Jupyter Notebook is an open-source web application developed by Project Jupyter[41]. By hosting the Jupyter Network, Google allows its users to access a cloud-based interactive coding environment through a web browser. Colab was chosen as a platform because of its cost effectiveness and generally simple interface. In the following paragraphs, the code used to automatically transcribe the audio files will be presented and explained.

The code[42] was written in the Python programming language and it involves a simple pipeline that transcribes the audios using Whisper's library. As a first step, the audios were uploaded to the programming environment in a zipped folder and the notebook was connected to a GPU runtime. The starting portion of the code installs the required packages (Whisper and

[40] https://colab.google/, last visited January 26, 2026
[41] https://jupyter.org/, last visited January, 27, 2026
[42] https://colab.research.google.com/drive/14rPhJqJnbwMgEv0UG1rCBSujreK_n5YP?usp=sharing

Torch[43]) imports the necessary libraries (whisper, os[44], zipfile[45], shutil[46] and files[47]), and loads the "medium" model. The second part of the code creates two directories, one for the extracted audio files and one for the transcriptions, and extracts the contents of the zipped folder. Successively, the code iterates over each audio file in the directory, transcribes it, saves the output as a TXT file and names the file after the corresponding audio source[48]. The last part of the code compresses the folder containing the transcriptions into a ZIP archive and downloads it to the local device. After the transcriptions were obtained, minimal post-editing was performed using a spell-checker, which identified misspelled words and proper nouns. These were then corrected also with the support of the audio recording. On a sample of 22 documents, 5 errors per text were found on average, demonstrating that automatic transcription was in fact efficient. Table 3 shows the most common errors in automatic transcription. Most were spelling errors, due to overlapping speakers or presence of music. In some cases, it produced errors related to language, as it either translated words, or wrongfully identified the language.

Error type	Audio	Wrong transcription
Spelling	Appassionato	Apasionato
	avrebbero	Averbero
Spelling proper names	Michielin	Michelini
	Max Herst	Max Ernst
	Priaruggia	Piarugia
Unnecessary translation	X-Factor	Ex-fattore
	Seicento	XVI century
Language	Veemente	Vehemente

Table 3: Common errors in transcriptions

2.8 Annotation

2.8.1 Metadata annotation

The annotation of texts through the use of metadata is the most used method for storing contextual information in electronic corpora. In modern corpora metadata is usually stored in

[43] Refers to PyTorch. PyTorch is an open-source machine learning framework, which is a crucial dependency for the whisper library
[44] https://github.com/python/cpython/blob/3.14/Lib/os.py, last visited January 27, 2026
[45] https://github.com/python/cpython/tree/3.14/Lib/zipfile/, last visited January 27, 2026
[46] https://github.com/python/cpython/blob/3.14/Lib/shutil.py, last visited January 27, 2026
[47] https://colab.research.google.com/notebooks/io.ipynb, last visited January 29, 2026
[48] The same naming convention was used for audios and transcriptions, so that the ID could be used as an identifier of the audio-transcription pair

an XML string at the beginning of the text (a description of the basic elements of XML is provided in section 1.1.7). As addressed in the previous sections, a corpus that is rich in metadata is a corpus that offers more possibilities of analyses to its users. However, adding metadata strings manually to each text is often time-consuming and could lead to inconsistencies due to typing and/or spelling mistakes. Conversely, an automatized process speeds up the annotation because it can process multiple texts at once. Moreover, it works with IDs and automates text copying and pasting, which minimizes potential transfer errors. In order to add the relative metadata to the file, the excel database (introduced in section 2.6), was leveraged.

Similarly to the previously described one, the code[49] was written in Python in a Google Colab Notebook. After importing the necessary libraries (pandas[50], zipfile, os, shutil and files), the Excel database was turned into a comma separated file (CSV), which was then uploaded to the Notebook and turned into a Pandas DataFrame (a table-like data structure in Python), along with the folder containing the transcriptions. Each transcription is matched to the corresponding row of the DataFrame, which is identified through the file ID. After the match is established, the script creates an XML start-tag called "doc", to which attributes and values are added by turning the title of each column (e.g., "year") into an attribute, and the values inside of the row into values of the attributes (e.g., "2025"). Whenever a datum was missing, the value was set to "N/A".

Once created, the string is added at the beginning of the transcription, and an end-tag is added at its end. The resulting metadata string and the attributes and value used will be shown in chapter 3.

2.8.2 Structural annotation

Structural annotation is a type of textual annotation that concerns the internal structure of the text. Structural annotation usually tags paragraphs and chapters or turns in conversations. SketchEngine automatically performs a preliminary structural annotation by tagging sentences inside of the documents by detecting sentence boundaries.

Since this corpus contains dialogic podcast episodes, structural annotation was taken into consideration to identify and tag the different speaker turns in the conversation. The concept of automatically transcribing conversations and partitioning the transcript into segments according to the speaker is referred to as *speaker diarization*. At the time of writing,

[49] https://colab.research.google.com/drive/1yTtCJnjGh0VghDAcvSFJ5mLdObJqkXWz?usp=sharing
[50] https://github.com/pandas-dev/pandas/blob/v3.0.0/pandas/__init__.py, last visited January 29, 2026

Whisper, the software used to transcribe texts, does allow speaker diarization[51]; however, it was not optimal for the creation of this corpus. The first issue concerns time: transcribing and identifying different speakers took longer than a normal transcription. Furthermore, the quality of the resulting transcription was lower and contained a larger number of errors, mainly caused by overlapping speakers, which required a more meticulous manual correction. Longer transcription times and effort required to manually correct mistakes made speaker diarization not feasible for the purposes of this thesis. In the current version of the corpus, further structural annotation is not present; therefore, texts are not tagged according to turn. However, further development of the corpus will include speaker diarization and a higher matching degree between the audio and the transcription.

2.8.3 Linguistic annotation

Along with metadata annotation, linguistic annotation was also applied to the corpus. Annotation can be performed both manually and automatically; for this corpus, automatic annotation was preferred. Manual annotation can ensure a smaller number of errors; however, it usually requires resources and time that were not available to this project. Linguistic annotation as well relies on XML tags (in this example, np stands for noun phrase, whereas v stands for verb):

```
<np>The kid</np> <v>eats</v> <np>the cake</np>
```

Linguistic annotation (part-of-speech tagging and lemmatization) was performed automatically by uploading the corpus on SketchEngine, compiling it and downloading it. SketchEngine automatically tags corpora in several languages using specific tagsets for each language[52]. A tagset is "a list of part-of-speech tags (POS tags for short), i.e. labels used to indicate the part of speech and sometimes also other grammatical categories (case, tense etc.) of each token in a text corpus" (Sketch Engine, no date b). Initially, the tagset used by SketchEngine for the Italian language was the TreeTagger tagset. The associated grammar (Word Sketch), however, was later rewritten for compatibility with a new tagset used since 2022, namely, the FreeLing tagset (Figure 10). FreeLing was developed by EAGLES, and "intends to be able to encode all existing morphological features for most European languages" (TALP Research Center, no date).

[51] https://github.com/MahmoudAshraf97/whisper-diarization/blob/main/Whisper_Transcription_%2B_NeMo_Diarization.ipynb, last visited January 3, 2026
[52] https://www.sketchengine.eu/tagsets/, last visited January 30, 2026

SKETCH GRAMMAR FOR ITALIAN

italian-fl-1.1.wsdef.m4

```
# == Word sketch grammar for Italian (FreeLing tagger) ==
# version 1.1
# 2022-02-02 v1.0 rewritten from the TreeTagger grammar v2.0 by Jan Kraus and Emma Romani
# 2024-04-09 v1.1 cleaned up by Marek Blahuš
#
# based on the word sketch grammar for Italian (TreeTagger tagger):
# 2006-03-26 v1.0 initial version by Marco Baroni
# 2008-07-14 v1.1 modified at the Lexicom Workshop in Barcelona by Francesca Masini and Valentina
# 2008-07-24 v1.1.1 modified again by Valentina Efrati and Francesca Masini
# 2008-10-03 v1.1.2 modified again by Valentina Efrati and Francesca Masini
# 2014-07-07 v1.1.3 UNIMAP added by Vít Baisa
# 2017-07-11 v1.2 WSPOSLIST added by Michal Cukr
# 2017-11-23 v2.0 improved by Lexical Computing intern Ludovica Lanini <ludovica.lanini@uniroma1.
```

Figure 10: Documentation provided for the Word Sketch Grammar used for Italian corpora

FreeLing's tagset consists of alphanumeric tags of different lengths that identify both the part of speech and its morphological features. Table 4 shows the tag(s) used to identify adjectives. The first letter of each tag indicates the part of speech of the element (e.g., noun, adjective, verb), while the following letters (or numbers) indicate its morphological features (e.g., number, gender). Each position past the 0 (the part of speech) is optional and can be substituted with a zero if underspecified. For instance, the Italian adjective *piccola* (feminine singular declination of *small*) would be tagged as <AQ0FS00>.

Position	Atribute	Values
0	category	A:*adjective*
1	type	O:*ordinal*; **Q**:*qualificative*; **P**:*possessive*
2	degree	S:*superlative*
3	gen	F:*feminine*; **M**:*masculine*; **C**:*common*; **N**:*neuter*
4	num	S:*singular*; **P**:*plural*; **N**:*invariable*
5	possessorpers	1:*1*; **2**:*2*; **3**:*3*
6	possessornum	S:*singular*; **P**:*plural*; **N**:*invariable*

Table 4: Adjective tag attributes and values (TALP Research Center, no date)

In corpus analysis, these tags are especially useful when building CQL searches in SketchEngine. *CQL* stands for *Corpus Query Language*, which is a special code used in SketchEngine to match complex grammatical or lexical patterns or to restrict searches to specific parts of speech or positions.

3. The corpus: PodIT

The third and final chapter of this thesis addresses the final PodIT corpus by describing it, its advantages and limitations. The corpus is explored, its features displayed, and some examples of analysis are illustrated. This chapter will also include the description of additional materials in support of the corpus, which will be useful for its future development and use. Furthermore, methods to address the issue of copyright will be presented.

PodIT is a **monolingual sample corpus of contemporary Italian spoken in the media**. It was built with the aim of creating a resource that would allow the analysis of an important section of modern spoken language, namely that spoken on the media. The corpus was built using a combination of manual and automatic processes, employed in order to maximize quality and efficiency. The texts were selected and collected manually, along with contextual information; successively, the texts were automatically transcribed, and ultimately, automatically annotated.

In order to obtain data, graphs and figures used in this chapter, the corpus was uploaded to Sketch Engine, where textual data and metadata were processed. Following the upload, it was possible to use the platform's tools to visualize data and analyze the corpus.

3.1 Data

3.1.1 Text data

PodIT is a collection of 133 texts, each collected by transcribing approximately 5-minute audio samples extracted from Italian podcasts published between 2020 and 2026. The corpus contains a total of 105,523 words and 117,134 tokens (that is, words and punctuation, numbers etc.). The texts in the corpus have an average length of 782 words; the shortest text is 515 words long, and the longest one is 1081 words long. Figure 11 represents the most common text lengths, showing that the majority of the texts contain between 700 and 800 words.

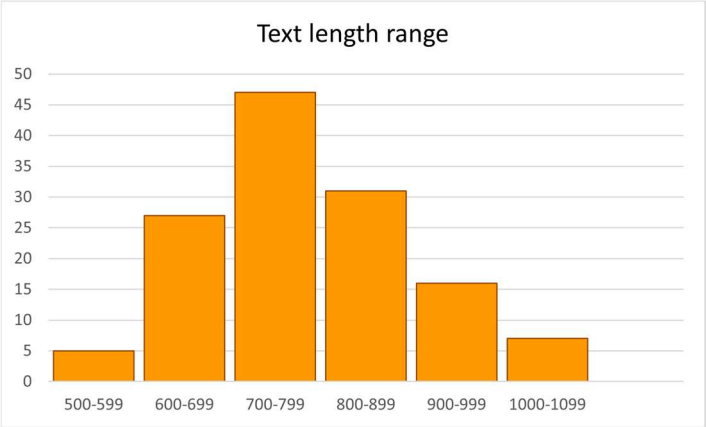

Figure 11: Distribution of text length in the corpus.

3.1.2 Audio data

The total registration time amounts to 11 hours, 50 minutes and 9 seconds, and the audios feature 223 speakers. The audio samples have an average length of 5 minutes and 20 seconds; the shortest audio sample has a length of 5 minutes and 1 second, while the longest audio sample has a length of 6 minutes and 20 seconds. Table 5 summarizes information about the audios of the corpus, namely, the number of speakers, the number of words of the transcription and the total registration time for each topic.

Topic	Speakers	Words	Total time
art	20	8.525	0:59:56
chat	24	12.721	1:16:11
cinema and tv	21	11.256	1:09:45
crime	15	8.445	1:02:59
culture and society	23	9.993	1:09:34
economy	18	7.931	0:58:16
health	19	9.753	1:08:27
music	17	8.332	0:57:32
politics	20	8.356	0:58:54
science and technology	21	9.403	1:04:21
sports	25	9.301	1:03:51

Table 5: Total length of audios for each topic

At the time of writing, widely used platforms for the consultation of corpora (e.g., Sketch Engine, AntConc[53]) do not allow straightforward video or audio linkage, and only deal with text data. Therefore, the user can not access both the text and the audio file using a single platform. However, it is still possible for the user to access the original audio/video source for a transcript, by following the source link that is present in the metadata. Due to copyright

[53] https://www.laurenceanthony.net/software/antconc/, last visited February 11, 2026

reasons (addressed in depth in section 3.6) the audios were not made publicly available. However, it is anticipated that in the later stages of the project, users should be able to access the audio files upon request to the corpus owners.

3.2 Corpus statistics

In this section, the final corpus will be described through the use of graphs, which were obtained from the *Text type analysis* function on Sketch Engine. A positive asset of the corpus, which will be highlighted in this section, is its balance. PodIT can be defined as a balanced corpus, since the different variables that define texts appear in similar proportions.

Figure 12 shows the distribution of monologic and dialogic podcast episodes. From this graph, it is possible to observe that monologic and dialogic podcast episodes appear in an almost equal number. A total of 71 texts pertain to the dialogic category, making up 53.4% of the corpus; conversely, 62 texts pertain to the monologic category, making up 46.6% of the corpus.

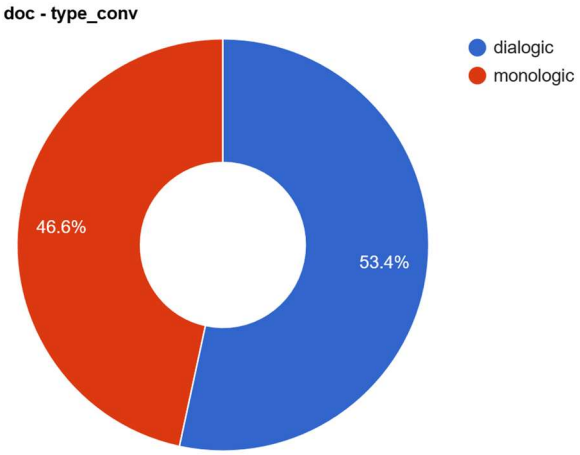

Figure 12: Distribution of monologic and dialogic podcast episodes in PodIT

The corpus shows a similar distribution for the professional/non-professional variable (figure 13); texts pertaining to each category have nearly equal proportions: 51.1% of texts (68 texts) are classified as non-professional, while 48.9% (65 texts) were classified as professional.

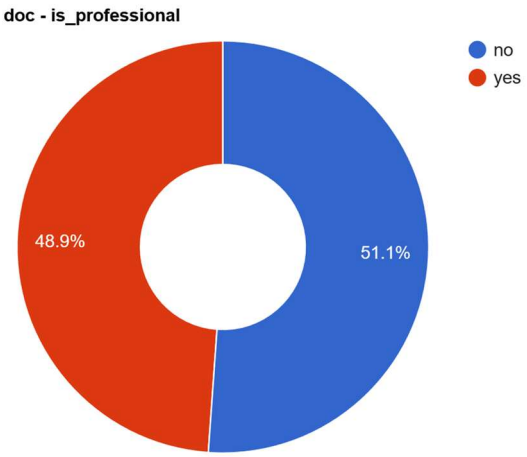

Figure 13: Distribution of professional and non-professional podcast episodes in PodIT

Similarly to the previous two variables, topics are balanced across the corpus. The topics, discussed in depth in section 2.4.1, have similar proportions. Every topic category contains a minimum of 11 texts and a maximum of 14 texts; their proportions and numbers are shown in figure 14 and table 6.

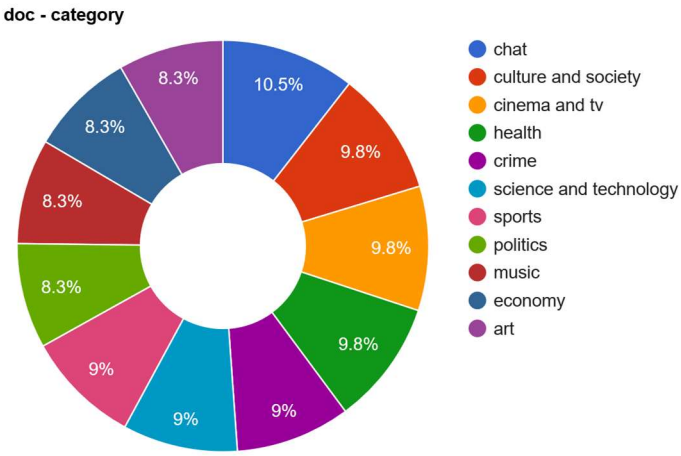

Figure 14: Distribution of topic categories in PodIT in percentages

Category	Number of texts
Art	11
Chat	14
Cinema and tv	13
Crime	12
Culture and society	13
Economy	11
Health	13
Music	11
Politics	11
Science and technology	12
Sports	12

Table 6: Distribution of topics in PodIT in numbers

While collecting the texts, a conscious effort was made to balance the gender of the speaker. The corpus features 223 speakers, 109 of which are women and 113 are men. One speaker identifies as non-binary.

Since the variables inside the corpus are all balanced with each other, it is possible to conduct contrastive analysis and have comparable results without the need for standardization.

3.3 Annotation

The present section will focus on the annotation that was performed on the texts that compose PodIT. This section does not delve into the methods that were used to apply annotation, as its only aim is to present the annotation and discuss its advantages and limitations. For a discussion on choices and rationales behind annotation, see section 2.6.

3.3.1 Metadata

Table 7 presents and briefly describes the attributes of the XML tags and their possible values. This table has solely the aim of showing the attributes as they appear in the metadata string while providing a brief description of their meaning. For an in-depth discussion on the reasoning and rationale behind the choice of each attribute, see section 2.6.

Attribute	Value	Description
ID	e.g., health_01, news_03, crime_12	naming convention composed of the topic name or its abbreviation, followed by a sequential index.
year	2020, 2021, 2022, 2023, 2024, 2025, 2026	the year of publication of the episode
type_conv	monologic, dialogic	indicates if the episode is monologic or dialogic
number_speakers	1, 2, 3	the number of speakers involved
is_professional	yes, no	indicates if the episode is regarded as professional
topic	art, chat, crime, cult, econ, film, heal, music, pol, sci, sport	name or abbreviation of the topic of the episode
length sample	e.g., 00:05:16	the duration of the sample in hh:mm:ss format
episode_name	e.g., Ep.4 - Diabete mellito	the full name of the episode
podcast_name	e.g., Med 101 - Pillole di divulgazione	indicates the full name of the podcast
source	e.g. https://...	link to the episode
sampling	start, middle, end	indicates whether the sample was taken from the start, middle or ending of the episode
speaker_name	e.g., Elisa De Marco	the full name of the speaker
speaker_age	e.g., 36, 43	the age of the speaker
speaker_gender	female, male, non-binary	the gender of the speaker
speaker_profession	e.g., content creator	the profession of the speaker
speaker_origin	e.g., Piemonte	the region or country of origin of the speaker.
speaker_nativeness	yes, no	indicates if the speaker is a native speaker of Italian
speaker_host_guest	host, guest	indicates if the speaker is a host or a guest

Table 7: Metadata attributes, values and brief description

The data was then inserted at the beginning of each document as an XML header through the method described in section 2.8.1. Figure 15 shows the XML header of a document present in the corpus (ID art_11) opened in the text editor Notepad++. The structure <doc> is further described by attributes (in orange) and their values (in purple). From this string, it is possible to observe that data about the speakers (e.g., speaker_name, speaker_age) was inserted three times with the addition of a number, in order to differentiate between speakers. Therefore, every attribute that stores data about the speakers appears in the metadata string as “speaker1”, “speaker2”, “speaker3”. As any other missing value, if a podcast featured one or two speakers, the attributes relative to the missing speaker(s) had “N/A” as a value.

<doc ID="art_11" year="2023" type_conv="dialogic" number_speakers="3" is_professional="no" category="art" length_sample="00:05:26" episode_name="Gioconde S.2 Ep.4 'Allegoria dell'Umana Fragilità' di Salvator Rosa con Caterina Volpi" podcast_name="Gioconde" source= "https://open.spotify.com/episode/1nIe6U1mLBiqAlL4NhMC7V" sampling="end" speaker1_name="Michela Giraud" speaker1_age="35" speaker1_gender="female" speaker1_profession="actor" speaker1_origin= "Lazio" speaker1_nativeness="yes" speaker1_host_guest="host" speaker2_name="Maria Onori" speaker2_age="N/A" speaker2_gender="female" speaker2_profession="N/A" speaker2_origin="N/A" speaker2_nativeness="yes" speaker2_host_guest="host" speaker3_name="Caterina Volpi" speaker3_age= "N/A" speaker3_gender="female" speaker3_profession="professor" speaker3_origin="N/A" speaker3_nativeness="yes" speaker3_host_guest="guest">LF

Figure 15: Screenshot of the metadata string present at the beginning of document "art_11", visualized in the text editor Notepad++

As can be observed from Figure 15, the XML header is fairly long and contains a large amount of data; this risks making metadata consultation difficult and time-consuming. Uploading the corpus on a software for corpus analysis allows the user to bypass the issue and filter contextual information. When a corpus is uploaded to Sketch Engine, the software processes the information contained in the string and orders it in a list, which can be accessed when exploring the corpus through concordances[54] (Figure 16). Concordances are a powerful tool for corpus linguistic analysis, which allows the user to look for linguistic elements (e.g., words, lemmas, phrases) and display them in the form of a concordance (Sketch Engine, no date a), namely a list of words surrounded by their immediate context.

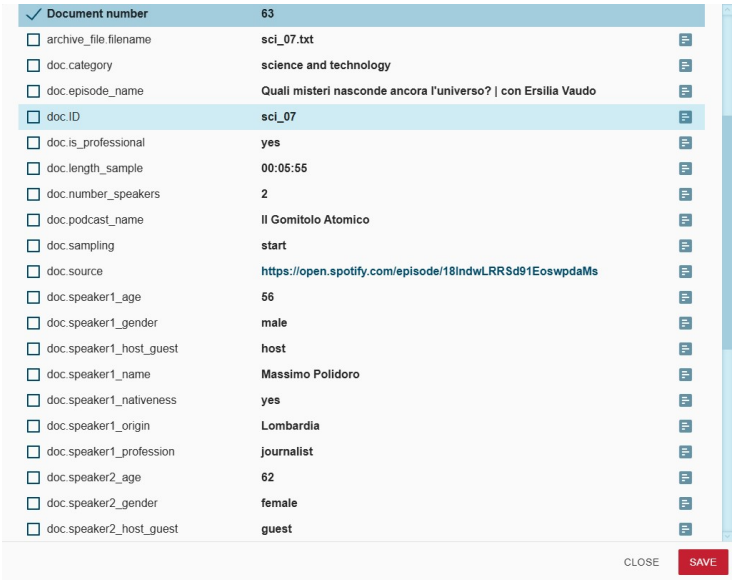

Figure 16: Screenshot of metadata of text "sci_07" displayed in Sketch Engine

Through the use of tools and software for corpus analysis, it is possible to leverage metadata in order to filter information based on the needs of the user. For instance, on Sketch Engine, it is possible to perform linguistic analysis on selected sub-sections. Figure 17 shows an example of a corpus query that searches for the lemma “creare” (*to create)* in 3 specific topic categories, namely art, cinema and tv, and culture and society.

[54] More information can be found at https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/, last visited February 11, 2026

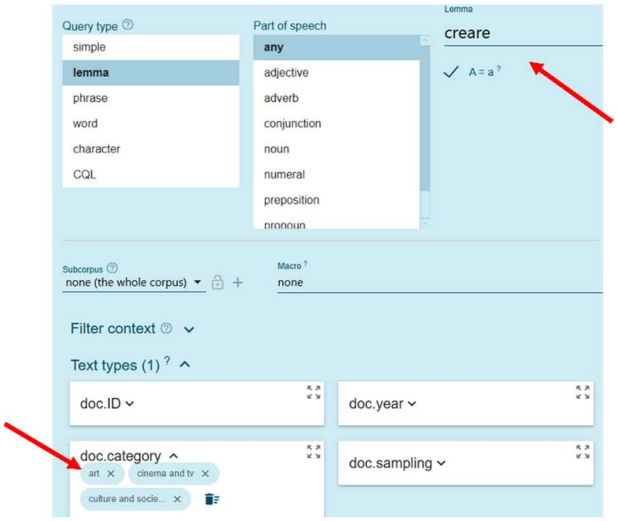

Figure 17: Query for the lemma "creare" in selected categories (art, cinema and tv, culture and society)

The corpus also allows the user to perform sociolinguistic analyses by leveraging data such as age, gender, profession and origin. However, at the time of writing, the absence of speaker diarization (section 2.8.2) limits analysis based on the speaker to monologic podcast episodes. This happens because the turns have not yet been tagged; therefore, the datum "speaker1_gender=male" in a dialogic text indicates that one of the speakers is male, however, it does not imply that the whole text was produced by a male speaker. In order to perform this type of analysis, the user can take one of two routes. The first method consists in applying a double filter, one that restricts the conversation type to "monologic", and the second that restricts the gender of speaker 1 to "male". In this way, each matched text is produced in its entirety by a male speaker. The second method consists in including both monologic and dialogic podcasts, and applying filters to speakers. Gender of speaker 1 has to be set to "male", while the gender of speaker 2 and 3 has to be either set to "male" or to not available ("N/A"). In this way, the resulting texts will either be monologic podcasts featuring a male speaker, or dialogic podcasts featuring multiple speakers who are all male.

3.3.2 Linguistic and structural annotation

As anticipated in section 2.8.3, linguistic annotation was applied automatically by Sketch Engine, using the FreeLing's tagset. Linguistic annotation applies POS-tagging and lemmatization, allowing the users to perform complex searches. For instance, it is possible to look for specific parts of speech or words appearing in specific positions. Figure 18 shows a search query that looks for feminine qualificative adjectives, while figure 19 shows the resulting concordance lines (selection).

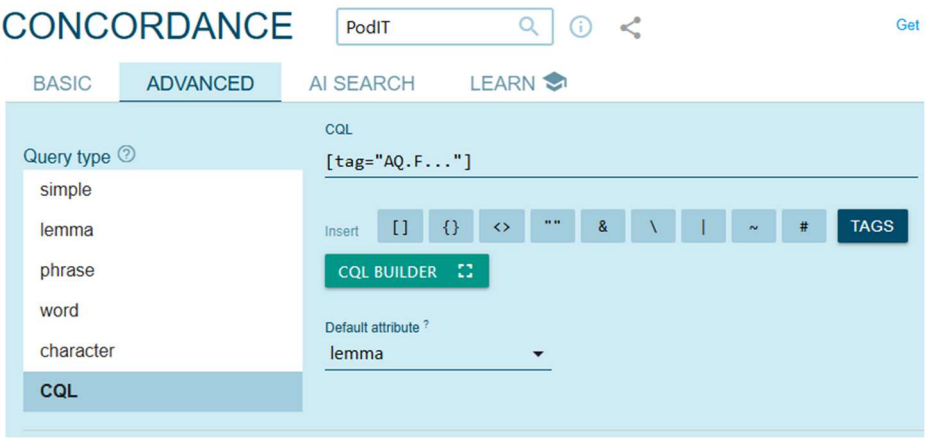

Figure 18: Screenshot of a search query by tag which looks for feminine qualitative adjectives

1	doc#83	e dalla centinatura della tavola, ossia dal fatto che nella parte	alta	la tavola è curvata in un rettangolare.</s><s>Nonostante le ta
2	doc#82	vorrebbe dire che sostanzialmente il 92% di tutta la criosfera	alpina	sarebbe destinata a scomparire da qui a fine secolo considera
3	doc#48	ce e ver sarei, ma può burlarmi ancor.</s><s>Anche qui rime	alternate	, vorrei, sarei, cor ancor.</s><s>Poi a mano a mano che l'ope
4	doc#129	s'è?</s><s>No il suo videocorso per imparare la divulgazione	scientifica	andate su data sapiens punto academy o in descrizione al link
5	doc#72	sono confuso.</s><s>Ovviamente tutte le volte che dirò nelle	prime	due stagioni c'era una trama che poi è cambiato nella terza e
6	doc#33	ıi confronto spesso col dottor Stingi e ci troviamo proprio sulla	stessa	linea, ricordare che esistono sì cibi con una capacità antiinfiar
7	doc#49	il primo principio da rispettare è cercare strumenti a gestione	passiva	che si adattano all'andamento dei mercati che a loro volta altri
8	doc#77	ziativa sembra aver perso un po' lo slancio iniziale.</s><s>La	stessa	sorte poi è toccata a un'altra proposta curiosa e che secondo
9	doc#10	› che erano orgogliose di me, di quello che avevo fatto, erano	contente	.</s><s>Solo le persone che ti hanno accompagnato nel perc
10	doc#82	ıante e anche inventando cioè avendo delle delle innovazioni	tecnologica	molto molto interessante</s><s>Oggi parleremo di un altro gr

Figure 19: Screenshot of the first 10 concordance lines resulting from a query looking for feminine qualitative adjectives

Linguistic annotation performed by Sketch Engine can be accessed by downloading a vertical file (VRT), that stores each token on a separate line, associating every single token with the corresponding POS-tag and lemma. Figure 20 shows three lines from the vertical file. The line starts with the token found in the corpus, its tag, its corresponding lempos[55], its corresponding lowercased lemma, a repetition of the tag, the lemma and the gender and number of the element.

```
risulta VMIP3S0 risultare-v risultare   VMIP3S0 risultare   0   0
poco    RG  poco-r  poco    RG  poco    0   0
pulita  AQ0FS00 pulito-j    pulita  AQ0FS00 pulito  F   S
```

Figure 20: Three lines of the vertical file visualized on Notepad++

Structural annotation as well was performed automatically by Sketch Engine. As explained in section 2.7.5, in this case structural annotation is limited to identifying and tagging sentences. This kind of annotation can be observed in Figure 19, Where <s> and </s> tags signal the opening and the closing of sentences.

[55] https://www.sketchengine.eu/glossary/lempos/, last visited February 18, 2026

3.4 Analysis

This section contains analyses that were performed on the corpus on the Sketch Engine platform, in order to highlight characteristics of Italian spoken in the media. To do so, a comparative analysis between PodIT and a corpus of written Italian was performed. First, a general exploration of the corpus was performed using keywords and wordlists; successively, some linguistic features of spoken neostandard Italian will be briefly explored through the corpus.

Several considerations had to be done in order to find the best corpus of written Italian to act as a *reference corpus*. A reference corpus is a corpus used in analyses to act as a source of comparison for the *focus corpus*, which is the corpus that the analysis focuses on (PodIT). Initially, the CORIS/CODIS corpus was considered, since it contains instances of written language and is constantly updated. However, it was eventually excluded, since its platform does not provide the kind of analytical tools that Sketch Engine provides. Corpora of general written Italian that were available on Sketch Engine were the following: itWaC, itTenTen20, PAISÀ and Araneum Italicum. Both itWaC and PAISÀ were excluded because of their time frame. PodIT is a spoken corpus which contains texts from the year 2020 to the year 2026, while itWaC and Araneum Italicum collect data respectively up to 2006 and 2010. Comparing two corpora that contain texts produced more than a decade apart could introduce bias and produce skewed results. The remaining two more recent corpora were itTenTen20 and Araneum Italicum, which were both built by collecting texts from the web, and they were created respectively in 2019-2020 and in 2014. For the reason mentioned above regarding the time frame, the corpus itTenTen20 was preferred.

The corpus itTenTen20 is part of the itTenTen corpora, which in turn are part of the TenTen family of comparable corpora, built by collecting linguistically valuable content from the web (Sketch Engine, no date c). The corpus itTenTen20 was created by crawling the web in 2019 and 2020; all the texts were first downloaded from the web, then cleaned, tagged and annotated. The majority of the texts contained in itTenTen (94.8%) do not have a topic classification due to the resources required to manually tag a large corpus; however, the remaining texts (5.2%) are classified by topic. Table 8 lists the topic categories present in itTenTen20 with their relative word count.

Category	Frequency
arts	41,23
business	15,215
games	31,976
health	3
home	4,936
news	266,534
recreation	8,483
reference	1,138,965
society	22,005
sport	5,89
sports	10,566
technology	54,31

Table 8: Distribution of categories in itTenTen20

For the purposes of the following analysis, a sub-corpus of itTenTen was created by only selecting categories that were either equivalent or similar to the ones present in PodIT. This was done in order to ensure a higher degree of comparability across the two corpora. The classification of texts in the original corpus was performed according to the classification used by DMOZ (now Curlie[56]), a multilingual human-edited directory of the web. The site has the aim of sorting website into macro-categories which are then expanded and divided into sub-categories to facilitate web-browsing. In order to assess which categories best matched the ones present in PodIT, each macro-category was analyzed on the Curlie website. Out of the categories listed in table 7, "games", "home" and "reference" were excluded, as their composition, analyzed on the original website, did not match any of the categories included in PodIT. The resulting sub-corpus constitutes around 1.5% of the original corpus and contains around 187,400,000 words and exactly 218,534,262 tokens.

In order to highlight salient characteristics of PodIT, the *Keywords* tool will be employed. The Keywords tool on Sketch Engine generates a list that shows words that appear more times in the focus corpus (PodIT) than in the reference corpus (itTenTen sub-corpus), highlighting salient differences between the two.

The Keywords tool was used for a preliminary exploration of the corpus. Subsequently, separate analyses on specific parts of speech were conducted using the *Wordlist* tool. The Wordlist tool allows the user to generate frequency lists of many linguistic elements: attributes (words, lemmas and tags), parts of speech (e.g., adjectives, verbs, conjunctions) or character sequences (e.g., words beginning or ending with specific characters). Because the analysis aims

[56] https://curlie.org/, last visited February 13, 2026

at observing surface patterns, only the first 10 instances of the resulting wordlists were considered.

A further exploration of the corpus was performed by generating and analyzing lists of the most frequent N-grams. N-grams are "recurring sequences of n words" (McEnery and Hardie, 2011), namely, sequences of 2 or more words, also termed "lexical bundles", that frequently co-occur. For this analysis, the *Keywords* tool was also used, providing a list of the most salient n-grams.

Specific patterns of Italian spoken language, obtained from relevant literature on Neostandard and spoken Italian, were also analyzed. Analyses of specific patterns were carried out using queries in the *Concordance* tool. The Concordance tool allows to search for words or patterns and display the results in the form of a KWIC (Key Word In Context) concordance. A KWIC concordance (figure 19) displays the results of the query in a list where the search item is surrounded by its immediate left and right context (Stefanowitsch, 2020, p. 50).

Findings of the previously mentioned analyses are summed up and discussed in the conclusion.

3.4.1 Keywords extraction

In this preliminary step of the analysis, a keywords list was produced for the corpus PodIT. The reference corpus was set to itTenTen, and the reference sub-corpus was set to the previously mentioned itTenTen sub-corpus. The keywords were set to be displayed as lemmas with a minimum frequency of 10. Sketch Engine also allows the user to set a value, on a scale from 0.001 to 1,000,000, which will determine if the keyword list should focus more on rare or common words. If the focus is set to rarer words (values closer to 0.001), the keywords will display words that are rare in general language and in the reference corpus; conversely, if the focus is set to more common words (values closer to 1,000,000), the keywords will list words that are more common in general language or in the reference corpus. For this specific analysis, following a piloting stage with different settings, the balance between rare and common words was set to 100. Setting the value closer to rare words gave skewed results, as the keywords identified were uncommon words like people's last names or misspelled words. Conversely, if the value set was too high, towards common words, it failed to identify features typical of spoken language or spoken language of the media. By setting the value to 100, the resulting keyword list was balanced, since it listed function words as well as content words which are typical of spoken language. Table 9 shows the rank of the item, the item, the relative frequency (per million tokens) for both the focus and the reference corpus, and a score. The score is a

keyness score assigned by Sketch Engine, which measures the relevance (or *keyness*) of a specific token for the focus corpus.

	Item	Relative frequency (focus)	Relative frequency (reference)	Score
1	cioè	2,416.04	211.68	8.1
2	no	2,373.35	229.36	7.5
3	sì	1,622.07	133.28	7.4
4	io	3,713.70	416.89	7.4
5	magari	1,314.73	92.85	7.3
6	quindi	4,951.59	600.21	7.2
7	appunto	1,161.06	83.69	6.9
8	tu	1,186.68	97.90	6.5
9	ti	2,031.86	233.30	6.4
10	perché	6,368.77	964.27	6.1
11	po	1,989.17	262.81	5.8
12	però	3,295.37	507.02	5.6
13	me	2,040.40	289.68	5.5
14	cosa	4,695.48	840.67	5.1
15	ok	546.38	31.78	4.9
16	lì	862.26	107.68	4.6
17	c'	3,252.69	647.83	4.5
18	vabbè	341.49	2.63	4.3
19	dicare	529.31	48.52	4.2
20	veramente	776.89	109.25	4.2
21	qua	409.79	26.57	4.0
22	te	631.76	90.79	3.8
23	adesso	870.80	153.15	3.8
24	ovviamente	700.05	112.88	3.8
25	poi	3,978.35	986.74	3.8
26	comunque	1,229.36	256.99	3.7
27	insomma	614.68	92.32	3.7
28	qualcosa	956.17	188.85	3.7
29	dire	5,848.00	1,585.87	3.5
30	podcast	256.12	1.12	3.5

Table 9: Keywords list generated from the comparison of PodIT with the itTenTen sub-corpus

Through the analysis of the keywords list, it is possible to visualize tokens that are specific to spoken language. The word *cioè* in PodIT (translatable as *that is* or *namely*) has a relative frequency which is 10 times higher than the relative frequency of the itTenTen sub-corpus. In Italian spoken language, *cioè* is not only used as a simple conjunction, but as a discourse marker (Khachaturyan, 2011; Ghezzi and Baido, 2024), and as a reformulation marker (Mereu and Negro, 2025). The expression *vabbè* (*anyway*), a contraction of the Italian

expression *va bene* (*it is all right*), is typical of spoken language, and has entered written language in the last decades through the media (Dardano, 2012).

Similar proportions can be found when comparing the relative frequencies of *sì* (*yes*) and *no* (*no*) and *ok*. In PodIT, the first two particles have a relative frequency which is 10 times larger than the relative frequency found in itTenTen, while *ok* has a relative frequency that is 17 times larger. These large differences are probably due to the fact that *sì*, *no* and *ok* are usually used in conversations to express agreement or disagreement; therefore, they are typical of conversations, whether they are spoken or written (e.g., texts messages, social media comments, blog posts).

The keyword list also highlights the different use of pronouns in the two corpora, namely *io* (*I*), *tu* (*you*), *ti* (object and reflexive form of *you*), *me* (*me*), *te* (object form of *you*). The large use of personal pronouns, especially first and second person pronouns, is typical of conversations, since they are used to address oneself or the other speaker.

The most frequent part of speech in the keyword lists is adverbs. Some adverbs featured in the list are used purely as adverbs, meaning they carry semantic information about time, place or manner. This is the case for the adverbs *lì* (*there*), *qua* (*here*), *adesso* (*now*), *ovviamente* (*obviously*), *poi* (*then*), *veramente* (*really*). Some other adverbs are also commonly used as discourse markers, like *appunto* (*precisely*), *comunque* (*anyway*), *insomma* (*anyway*, *basically*), *magari* (*if only*, *maybe*), *poi* (*then*).

The only verb that appears in the analyzed part of the keyword list is the verb *dire* (to say); it appears twice, as in some instances it was incorrectly lemmatized as *dicare*. In PodIt, the verb has a frequency of 5,848.00 per million tokens, while in the itTenTen sub-corpus its frequency per million is equal to 1,585.87. The verb appears more than three times as much in the spoken corpus. More specifically, the verb frequently appears in the form "dire" (*to say*) and *diciamo* (*let's say*), forms that are usually employed as part of discourse markers (Khachaturyan, 2011) like *come dire* (*how to say, you know*), *vuol dire* (*it means*), *diciamo che* (let's say that). The large difference in frequency could be due to the use of the verb *dire* both as a simple verb (since podcasts are speech events, therefore based around the processes described by the verb *dire*) and as a discourse marker in spoken language. Since no other verb was present in the keywords list, verbs are analyzed separately in section 3.4.2.

A separate analysis was also conducted on adjectives (section 3.4.3), since they do not appear in the first 30 instances of the keywords list. Furthermore, because the only conjunctions in the keywords list are *perché* (*why*) and *quindi* (*therefore*), conjunctions are analyzed separately in section 3.4.4, as this analysis could shed light on the type of sentences that

speakers of podcasts usually build. A separate analysis on nouns will not be included in this section; following a preliminary analysis, it was assessed that nouns tend to reflect corpus-specific characteristics rather than general discourse patterns, which is the main focus of this analysis.

3.4.2 Verbs

Wordlists of verbs were created to further analyze the most frequent verbs of PodIT and compare them to the most frequent ones in itTenTen. Table 10 and 11 show the 10 most frequent verbs in PodIT and in the itTenTen sub-corpus.

	Verb	Frequency	Relative frequency
1	essere	4,750	40,551.85
2	avere	1,910	16,306.11
3	fare	955	8,153.06
4	dire	685	5,848.00
5	potere	535	4,567.42
6	andare	365	3,116.09
7	stare	324	2,766.06
8	dovere	293	2,501.41
9	volere	265	2,262.37
10	sapere	251	2,142.84

Table 10: First 10 verbs ranked by frequency in the PodIT corpus

	Verb	Frequency	Relative frequency
1	essere	5,448,201	24,930.65
2	avere	2,442,065	11,174.75
3	fare	739,685	3,384.76
4	potere	657,751	3,009.83
5	dovere	359,245	1,643.88
6	dire	346,557	1,585.82
7	venire	318,398	1,456.97
8	andare	253,451	1,159.78
9	stare	222,228	1,016.90
10	volere	212,178	970.91

Table 11: First 10 verbs ranked by frequency in the itTenTen sub-corpus

Through the analysis of these wordlists, it is possible to observe that the spoken corpus has a higher density of verbs compared to its written counterpart; in line with the observed lower lexical density of spoken language, compared to a higher verbal density (Halliday, 1990). A higher use of modal verbs can be observed in PodIT, *potere* (*can*), *dovere* (*must*), and volere (*to want*). Specifically, the verb *volere* (*to want*) appears twice as frequently in the spoken corpus. This phenomenon could highlight a tendency to use the verb *volere* to express personal needs and desires. The verb *volere*, in addition to expressing desire, is part of lexicalized expressions like *voglio dire* (*I mean*), *vuol dire* (*it means*), used to address the meaning of a sentence.

The verbs *essere* (*to be*) and *fare* (*to do*) have a higher relative frequency in PodIT than they do in the itTenTen sub-corpus. Although they convey meaning like any other verb, they are rather general ones; for this reason, it is possible that they are preferred in spoken language, but avoided and substituted with more complex verbs in written language. This is in line with the assumption that in spoken language, speakers tend to prefer efficient expressions, that is expressions which convey meaning while requiring the least possible costs (articulation,

processing and time) (Levshina, 2023). This inevitably leads to a simplification of language, which was also identified in neostandard Italian (Grandi, 2019).

3.4.3 Adjectives

As mentioned in section 3.4.1, adjectives were analyzed separately since they were not present in the keywords list. Two wordlists showing the frequency and frequency per million in the itTenTen sub-corpus and in PodIT were generated and compared. The first 10 instances for both corpora are showed in table 12 and 13.

	Adjective	Frequency	Relative frequency
1	primo	189	1,613.54
2	grande	167	1,425.72
3	stesso	128	1,092.77
4	mia	123	1,050.08
5	suo	123	1,050.08
6	loro	122	1,041.54
7	sua	121	1,033.00
8	mio	119	1,015.93
9	ultimo	103	879.33
10	bello	102	870.80

Table 12: First 10 adjectives ranked by frequency in the PodIT corpus

	Adjective	Frequency	Relative frequency
1	primo	347,686	1,590.99
2	nuovo	312,583	1,430.36
3	sua	293,183	1,341.59
4	suo	272,966	1,249.08
5	grande	259,286	1,186.48
6	loro	233,055	1,066.45
7	stesso	230,328	1,053.97
8	ultimo	167,165	764.94
9	italiano	130,525	597.27
10	diverso	130,265	596.09

Table 13: First 10 adjectives ranked by frequency in the itTenTen sub-corpus

By analyzing these two wordlists, it is possible to note some slight differences. For instance, the possessive adjectives *mia* and *mio* (feminine and masculine singular of *my*) appear respectively in the 4th and 8th position, while they do not appear in the first 10 positions of itTenTen. The adjectives were searched for in the itTenTen wordlist, and it was found that *mia* had a frequency of 356.81 per million tokens, while *mio* had a frequency of 383.66 per million tokens. By comparing these numbers, it was noted that possessive adjectives appear almost three times as much in the PodIT corpus. Through empirical observation of podcasts, it was found that speakers tend to reframe conversation by addressing personal topics, opinions and experiences. This tendency could justify the higher positions of the possessive adjectives *mia* and *mio* in the tables.

A further difference can be found in the frequency of *bello* (masculine singular of *beautiful*). The adjective *bello* is 10th in the frequency ranking for PodIT, with a frequency of 870.80 per million tokens. In the written corpus, it appears in the 45th position, with a frequency of 255.63 per million tokens, more than three times lower than in PodIT. Similar to the discussion on possessive adjectives, the high occurrence of *bello* could be tentatively aligned

with the tendency of speakers to express personal opinions through evaluative adjectives that reflect their subjectivity. Moreover, it could be related to the previously mentioned principle, by which speakers tend to use simpler expressions that minimize cognitive work and guarantee meaning.

3.4.4 Conjunctions

A further investigation was performed on conjunctions, as they can give insights on the sentences that speakers build and on how discourse is organized. Wordlist generated for both PodIT and the itTenTen sub-corpus and are shown in tables 14 and 15.

	Conjunction	Frequency	Relative frequency
1	e	2,480	21,172.33
2	che	991	8,460.40
3	ma	704	6,010.21
4	se	433	3,696.62
5	o	271	2,313.59
6	ed	97	828.11
7	mentre	57	486.62
8	oppure	37	315.88
9	sia	32	273.19
10	ovvero	26	221.97

Table 14: First 10 conjunctions ranked by frequency in the PodIT corpus

	Conjunction	Frequency	Relative frequency
1	e	4,992,307	22,844.50
2	che	969,146	4,434.76
3	ma	762,587	3,489.55
4	o	463,212	2,119.63
5	se	406,281	1,859.12
6	ed	299,831	1,372.01
7	mentre	138,779	635.04
8	dopo	51,730	236.71
9	sia	42,569	194.79
10	né	38,336	175.42

Table 15: First 10 conjunctions ranked by frequency in the itTenTen sub-corpus

In the two tables, the conjunctions that appear are roughly the same, with the exception of positions 8 and 10; however, the frequency with which they appear differs widely across the two corpora. While other coordinating conjunctions like *o* (*or*) and *e* (*and*) have nearly the same frequency per million tokens, differences can be found in the use of the conjunction *ma* (*but*). In PodIT, the conjunction *ma* has a frequency of 6,010.21 per million tokens, which is almost double the normalized frequency found in itTenTen (3,489.55). The higher normalized frequency of *ma* can signal both a tendency of Italian speakers to use more adversative clauses as well as a tendency of using *ma* as a discourse marker to change topic.

Similarly, in PodIT, the subordinating conjunctions *se* (*if*) and *che* (*that/which*) have a normalized frequency that is almost twice as high as the normalized frequency in itTenTen. In the spoken corpus, the conjunction *se* has a frequency of 3,696.62 per million tokens, while in the written corpus, it has a frequency of 1,859.12 per million tokens; similarly, *che* has a frequency of 8,460.40 per million tokens in PodIT, and of 4,434.76 in itTenTen. Based on these differences, it is possible to infer that speakers tend to form more subordinate clauses with the

conjunctions *se* and *che* in spoken language than in written language. More specifically, the higher frequency of *che* can be due to multiple reasons. It could indicate a tendency of Italian speakers to form more relative clauses, however, it could signal the presence of the of the *che polivalente* (*polyvalent che*). The conjunction *che* becomes *polyvalent* when its uses are extended to clauses that would normally require other conjunctions (Fiorentino, 2010); this phenomenon can be mostly observed in spoken neostandard Italian.

The presence of *ovvero* (*namely*) in the first 10 conjunctions in PodIT is somewhat surprising, since it appears to be a more formal conjunction. In the spoken corpus, the conjunction has a frequency of 221.97 per million tokens, compared to a frequency of 94.83 per million tokens in the written corpus. The higher frequency per million tokens in the spoken corpus is an unexpected result which would be worth exploring in further works.

3.4.5 N-grams

As anticipated in section 3.4 a further exploration of the corpus was carried out through the use of N-gram lists. Since n-grams are uninterrupted sequences of words that frequently occur together, they also provide evidence of lexicalized expressions. Through the use of Keywords, lists of key bi-grams and tri-grams, namely n-grams composed of two or three words, were generated. Tables 16 and 17 list the first 5 occurrences for every type of n-gram. Similarly to the previous keyword analysis, the balance between rare and common n-grams was set to 100 and the reference corpus was set to the itTenTen sub-corpus.

Table 16 lists the first 5 salient bi-grams for PodIT. The most salient n-gram of the PodIT corpus is *c'è* (there is), which is rather uncommon in the written corpus. In Italian spoken language the *presentational c'è*, is used to give new information and emphasize a particular utterance (Giampieri, 2025, p. 181). Examples from the corpus include:

1) *comunque* ***c'è*** *qualcuno che guarda il mio profilo* (*anyway there is someone that is looking at my profile*);
2) ***c'è*** *un pilota che ha fatto molto molto meglio* (*there is a driver who did much, much better*).

The n-gram *un po* (correct spelling *un po'*), contraction of *un poco* (a little), is used in the Italian language in several ways. Generally, it is used before or after adjectives or verb phrases to mitigate their meaning, for instance, in these examples from the corpus:

3) *è una ragazza* ***un po'*** *timida,* ***un po'*** *imbranatella* (*she's kind of a shy and clumsy girl*);

4) *riprendendo **un po'** quello che abbiamo detto* (*to pick up shortly on what we've said*).

Potentially, this expression could be regarded as a hedging technique, used in order to minimize the strength of what is being said. Hedging techniques are rhetorical strategies used to limit commitment to the semantic meaning of an expression (Fraser, 2010). The high frequency of this expression in the PodIT corpus could be due to a tendency of podcast speakers to distance themselves from strong statements or claims. The bi-gram *che è* (which/that is) is used to form relative clauses, which, as observed in section 3.4.4, are fairly common in the corpus. *Che è* is also used to clarify concepts and add details, for instance:

5) *[p]oi il MMCA **che è** il museo nazionale di arte moderna* (*Then, the MMCA, which is the national museum of modern art*)

6) *una storia **che è** proprio una storia d'amore* (*a story that is truly a love story*).

Lastly, the bi-gram *quello che* (similar to *what/that*) is used to introduce relative clauses, to specify details and introduce argumentations and points. For instance, in:

7) *è anche **quello che** più ci interessa* (*it's also what interests us more*).

	Bi-gram	Relative frequency (focus)	Relative frequency (reference)	Score
1	c' è	2,450.19	34.81	18.9
2	un po	1,972.10	14.11	18.2
3	che è	1,946.49	33.24	15.4
4	è un	1,946.49	38.74	14.8
5	quello che	1,289.12	19.26	11.6

Table 16: First 5 key bi-grams in PodIT

A list of the key tri-grams was also created in order to identify salient three-word patterns in the PodIT corpus. Table 17 shows the first 5 key tri-grams in the PodIT corpus. The first instance in the table is *non lo so* (*I don't know*); its higher relative frequency could be traced back to the findings discussed above. The use of *non lo so* could be part of hedging techniques used by podcasts speakers when making assumptions or claims. Examples from the corpus include:

8) ***[n]on lo so**, sì secondo me esiste* (*I don't know, I believe it exists*);

9) *[t]rovo abbastanza, **non lo so**, simboliche* (*I find [them] rather, I don't know, symbolic*).

It is used to express uncertainty or to attempt to mitigate claims or statements. Both the tri-grams *una cosa che* (*a thing/fact that*) and *è una cosa* (*it is a thing/something*) are used to

introduce explanations, or address concept or ideas that were either shortly mentioned or about to be mentioned by substituting it with the placeholder *cosa* (*thing*). For instance:

10) ***[e] una cosa*** *che ti dicono spesso gli insegnanti è che* (*something that teacher often tell you is that […]*).

The last tri-gram, *nel senso che* (*in the sense that*) is used to clarify the preceding word or sentence; for instance:

11) *persone che ho perso,* ***nel senso che*** *sono andati a vivere da un'altra parte* (*people I have lost, in the sense that they went to live someplace else*)

12) *l'esempio più facile,* ***nel senso che*** *è più facile da spiegare* (*the easiest example, in the sense that it's easier to explain*).

	Tri-gram	Relative frequency (focus)	Relative frequency (reference)	Score
1	non lo so	281.73	0.49	189.4
2	una cosa che	264.65	0.75	151.9
3	è una cosa	256.12	0.96	131.4
4	è un po	247.58	0.9	130.6
5	nel senso che	290.27	1.4	121.4

Table 17: First 5 key tri-grams in PodIT

From this analysis of the key n-grams in PodIT, some characteristics of podcast discourse have emerged. Two major tendencies were found, namely explanation and hedging. By simple empirical observation of the domain, it was found that podcast creators tend to make sure that concepts are well-understood and clear to the audience. On the other hand, other n-grams point to a tendency, or technique used, namely, linguistic hedging.

3.4.6 Analysis of specific patterns

As discussed in section 1.2.1, neostandard Italian is a variety of the Italian language which is spoken and written by Italian speakers in moderately controlled contexts, which was born from the influence of regional dialects on standard Italian. Characteristics of neostandard Italian are mostly present in spoken language; therefore, in order to prove the suitability of PodIT as a tool to analyze general spoken Italian, relevant patterns presented in literature were investigated in the corpus. One of the common patterns of spoken neostandard Italian is the use of cleft sentences with *che* and *a* (Zingaro, 2024). The cleft sentence is used to highlight a part of the sentence that is of greatest importance to the speaker. Some examples from the spoken corpus include:

1) ***[s]iamo noi che*** *stiamo sbagliando* (*we are the ones who are wrong*);

2) ***è quello che*** *secondo me poi lo spinge* (*it's that what I think it pushes him*).

In order to check if this characteristic was specific to PodIT when compared to a written corpus, the pattern was searched for in PodIT and in the previously generated itTenTen sub-corpus. Since the pattern is complex and includes multiple items, CQL (Corpus Query Language) was used. The following CQL query was employed to match a pattern that includes the lemma of the verb *be*, followed by any pronoun, followed by *che/a*.

```
[lemma="essere"][tag="P.*"][word="che"|word="a"]
```

In the PodIT corpus, the pattern has a relative frequency of 461.01, while in the itTenTen sub-corpus, it has a relative frequency of 118.11. The pattern appears more than three times as frequently in the spoken corpus.

A second feature that was searched for in the corpus is related to the use of *Passato remoto* (tense similar to past simple), a verb tense used to talk about past events. Linguistic studies have shown a general reduction of the use of *passato remoto* in spoken language; it tends to be substituted with *passato prossimo* (compound tense, similar to Present Perfect). The difference in use is mainly of a diatopic kind; in northern regions it is more frequently substituted with *passato prossimo*, while in southern regions it is still in use (Bertinetto and Squartini, 1996; Accademia della Crusca, 2004). This difference only shows in spoken Italian, as *passato remoto* is still used in the written language. Regardless of regional variants, however, a reduction in the use of *passato remoto* and equivalent tenses is envisioned in future developments of the language as a whole (Wiberg, 2011). The use of *passato remoto* was examined in PodIT, and in the itTenTen sub-corpus, in order to identify any difference between speaking and writing. In order to do so, linguistic annotation was leveraged; the following CQL captures every verb in the indicative mode and *passato remoto* tense:

```
[tag="V.IS.*"]
```

The results obtained were in line with the previously reported claims; in PodIT, *passato remoto* has a relative frequency of 1,263.51, while in the itTenTen sub-corpus, it has a relative frequency of 3,208.47, double compared to the spoken corpus. From this brief analysis, it is possible to infer that the PodIT corpus shows the tendency found in spoken Italian, namely, a decrease in the use of *passato remoto*.

The analysis section has shown that PodIT displays characteristics which are traceable both to neostandard and spoken Italian and traceable to public language or language spoken in the media. A further discussion is present in the Conclusion chapter.

3.5 Copyright

Copyright issues have long been a concern for corpus linguistics scholars (McEnery, Xiao and Tono, 2006), since the creation of corpora includes acquisition and sharing of works that might be protected by copyright. One of the most common ways in which corpus creators deal with copyright issues is to only include a portion of the original text in their corpus, as done by the Brown corpus compilers, or to collect works that are in the public domain, as done by the creators of the corpus PAISA, or the Gutenberg corpora[57].

Since this corpus contains copyrighted material, relevant literature and European directives were reviewed in order not to breach EU copyright and intellectual property laws. Articles 3 and 4 of the Directive (EU) 2019/790 of the European Parliament (European Parliament and Council of the European Union, 2019) regulate text and data mining in the European Economic Area (EEA), introducing exceptions that allow researchers and research institutions to reproduce, extract and process copyrighted materials for the purposes of scientific research, including data analysis. In order to be compliant with the directive, however, these materials have to be lawfully accessible (for instance, by being freely accessible online), and, if redistributed, they should not constitute a substitute for the original work.

The PodIT corpus was compiled by paying careful attention to the regulations contained in the Directive (EU) 2019/790. The current paragraph clarifies why the compilation and use of this corpus can be considered compliant with EU copyright laws. The corpus contains transcriptions of short snippets (around five minutes each) collected from podcasts which were freely available online and did not require a paid subscription nor an account to be accessed. Since the corpus is composed of transcriptions of short extracts, it does not function as a substitute for the original works. Given that the original podcasts are freely available online, it is also unlikely that anyone would opt for reading a transcription of a short extract of a podcast instead of listening to the original product. Furthermore, the project is entirely non-commercial and serves no other purpose than providing a resource for scientific research.

At the time of writing, it is not possible to access the audio recordings. In future stages of the project, it has been envisioned that audios could be available on request to verified researchers and solely for non-commercial, scientific research.

In order to make this information public, a copyright compliance statement (Appendix 1) was written and included in the corpus materials.

[57] https://www.sketchengine.eu/gutenberg-corpora-2020/, last visited February 15, 2026

3.6 Resources

All the corpus files and relevant additional materials have been uploaded to a GitHub repository[58] called "PodIT-corpus". This section will describe the resources that are available for consultation and download.

Corpus. The whole corpus was made available on the GitHub repository for download in three formats: as a folder with clean .txt files, with contextually annotated transcriptions and as a vertical file containing linguistic annotation. The folder with clean .txt files contains the transcriptions subject to minor manual revisions, without any annotation; a corpus user might want to add their own annotation or use a corpus query tool that does not process metadata. The folder with the annotated files contains the manually corrected transcriptions along with their metadata (presented in section 3.3.1). This folder can be uploaded and analyzed as a corpus on text analysis platforms like Sketch Engine. Finally, the vertical file containing linguistic annotation (presented in section 3.3.2) can be used to visualize or modify the linguistic annotation performed automatically by Sketch Engine. The vertical file can also be uploaded and used as a corpus on Sketch Engine.

Metadata tables. During the compilation of the corpus, two databases were built: the *text database* and the *speaker database*. The text database is composed of 133 rows (one per document). It provides metadata about each text, and was used to build the XML header. The speaker database contains the same type of information as the text database but at a more fine-grained level. It features 223 rows, one per speaker. This speaker-oriented data frame will be useful in the next processing stages of the corpus, centered around turn identification.

Copyright compliance statement. The directory also contains a copyright compliance statement (Appendix 1), written to inform users of the corpus' compliance to EU copyright and intellectual property laws pursuant to Directive (EU) 2019/790. Moreover, the copyright compliance statement holds the user accountable for any misuse of the materials provided. The document is directed to podcast creators as well, which can request the corpus owners to remove the transcription of their podcast episode from the corpus.

[58] https://github.com/joannagiacobbe/PodIT-corpus

Conclusion

This dissertation has dealt with the design and creation of PodIT, a corpus of spoken Italian in the media. The corpus was built through a combination of manual and automatic processes in order to maximize the quality of the corpus while minimizing necessary resources. The corpus filled a gap in Italian resources for corpus linguistics, by providing a resource to analyze spoken Italian in the media, as well as enriching the existing resources for Italian spoken language as a whole.

Chapter 1 has highlighted the most salient concepts of corpus building and the importance of representativity and balance in a corpus described the state of the art regarding both written and spoken Italian corpora, and identified the gap that this corpus fills.

Chapter 2 has described the application of theoretical concepts to the practical steps of corpus building. The chapter has presented the design choices that were made to ensure representativeness and balance, as well as choices regarding the domain, its description and composition. Chapter 2 has also considered the advantages and limitations of combining manual and automatic collection processes. The manual selection and collection (recording) of samples was time-consuming, yet it allowed more control on the contents of the corpus, thus introducing less noise. The use of automated processes of transcription and annotation was time efficient, but, despite some manual correction, did not ensure the total absence of errors.

Chapter 3 has focused entirely on the analysis of PodIT. First, the corpus is described and its numbers and statistics are reported. The advantages of the annotation of the corpus are highlighted, focusing on the type of analyses which can be performed. A large section of the third chapter is dedicated to the analysis of the corpus. The analysis aims to not only explore and further describe PodIT, but also to prove the suitability of the corpus for the analysis of general spoken language. The analysis of keywords and wordlists has highlighted the presence of items and patterns which are generally considered typical of spoken Italian, such as the presence of discourse markers. Discourse and reformulation markers are widely present in spoken language (Ghezzi and Baido, 2024); their higher frequency in PodIT could support the assumption that the corpus includes valid spontaneous speech and therefore can be considered as a valid source for the analysis of general spoken Italian . This assumption is confirmed by the presence of patterns identified as characteristic of spoken Italian, particularly of the Neostandard variety. Conversely, the analysis of n-grams has highlighted patterns that can be identified as strategies employed in public speaking or while speaking in controlled contexts, such as radio or TV programs or podcasts. These patterns include some form of linguistic

hedging and expressions used to explain, clarify and add details to ensure understanding from the audience.

Finally, Chapter 3 includes the description of some technical aspects, namely the steps taken to comply with EU copyright laws and the resources available with the corpus.

Although, as shown by the analysis, the final corpus is a valid resource for corpus analysis of the Italian language, future developments should increase its value and useability. Automatization of some processes and time and resource restrictions have introduced some limitations to PodIT. In its current version, PodIT does not feature the structural annotation of turns. Therefore, sociolinguistic analyses that leverage data about speakers are limited when investigating dialogic podcasts. A further development of the corpus would concern the application of this kind of annotation, which would improve the overall quality of the corpus and increase the possibilities of analysis.

A manual process which would benefit the corpus and increase its quality is the manual correction of audio files. As explained in chapter 2, the transcription was performed by means of OpenAI's Whisper and major errors were corrected using a spell checker. Minor spelling errors or misinterpretations were not corrected due to time limitations. In some cases, especially when dealing with dialogic podcasts where speakers often interrupt each other and their speech overlaps, some parts of the conversation are not transcribed and therefore lost. A manual correction of minor errors in addition to a validation of the transcription by aligning it with the source audio would increase the quality and accuracy of the corpus.

Future work includes using PodIT as a tool to analyze spoken Italian and spoken Italian in the media. Furthermore, resources on spoken Italian in the media could be expanded by collecting language occurring in different media outlets, such as radio and TV programs or social media. This could lead to the construction of a well-structured family of comparable corpora of media language, which would ultimately enrich the Italian landscape of corpus resources. The analysis of this corpus could eventually lead to a better and broader understanding of media language, which is fundamental in today's society.

References

Accademia della Crusca (2004) “Sull’uso del passato remoto,” *Accademia della Crusca*, 12 March.

Atkins, S. (1992) “Corpus Design Criteria,” *Literary and Linguistic Computing*, 7(1), pp. 1–16.

Balanuta, L. (2021) “Podcast–towards an inclusive definition,” *Acta Universitatis Danubius. Communicatio*, 15(2), pp. 31–41.

Ballarè, S. (2020) “L’italiano neo-standard oggi: stato dell’arte,” *Italiano LinguaDue*, 12(2), pp. 469–492.

Barbera, E.F. (2013) *Linguistica dei corpora e linguistica dei corpora italiana. Un’introduzione.* Qu. ASAR.

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. and Mazzoleni, M. (2004) “Introducing the La Repubblica Corpus: A Large, Annotated, TEI (XML)-compliant Corpus of Newspaper Italian.,” *LREC*.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009) “The WaCky wide web: a collection of very large linguistically processed web-crawled corpora,” *Language resources and evaluation*, 43(3), pp. 209–226.

Benko, V. (2014) “Aranea: Yet another family of (comparable) web corpora,” *International Conference on Text, Speech, and Dialogue*. Springer, pp. 247–256.

Berruto, G. and Cerruti, M. (2011) *La linguistica: un corso introduttivo*. 1. ed. Novara: UTET università.

Bertinetto, P.M. and Squartini, M. (1996) “La distribuzione del Perfetto Semplice e del Perfetto Composto nelle diverse varietà di italiano,” *Romance Philology*, 49(4), p. 383.

Biber, D. (1993) “Representativeness in Corpus Design,” *Literary and Linguistic Computing*, 8(4), pp. 243–257.

BNC Consortium (2001) “The british national corpus, version 2 (bnc world),” *Distributed by Oxford University Computing Services* [Preprint].

Bonomi, I. (2010) “Tendenze linguistiche dell’italiano in rete,” *Informatica umanistica*, 3, pp. 17–29.

Boulton, A. (2017) “Corpora in language teaching and learning,” *Language Teaching*, 50(4), pp. 483–506.

Burnard, L. (2007) *Reference Guide for the British National Corpus (XML Edition)*, *natcorp.ox.ac.uk*. Available at: http://www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#spodes (Accessed: October 9, 2025).

Canalis, S. (2006) “Pugliese e salentino: alcuni fenomeni fonologici,” *Federico Damonte, Jacopo Garzonio (Hg.): Studi sui dialetti della Puglia. Padova* [Preprint].

Cerruti, M. (2018) “Il parlato regionale oggi: un italiano composito?,” *Lid’O: lingua italiana d’oggi: XV, 2018*, pp. 15–31.

Coats, S. (2023) “Dialect Corpora from YouTube,” in B. Busse, N. Dumrukcic, and I. Kleiber (eds.) *Language and Linguistics in a Complex World*. De Gruyter, pp. 79–102.

“Corpus KIParla – L’italiano parlato e chi parla italiano” (no date). Available at: https://kiparla.it/ (Accessed: October 12, 2025).

Dardano, M. (2012) “Vabbè, embè e compagnia bella,” *InNoio volevàn savuàr. Studi in onore di Edgar Radtke del sessantesimo compleanno, ed. by Silvia Natale, Daniela Pietrini, Nelson Puccio, and Till Stellino*, pp. 27–40.

Davies, M. (2008) *The Corpus of Contemporary American English (COCA)*. Available at: https://www.english-corpora.org/coca/ (Accessed: January 15, 2026).

Davies, M. (2010) “The Corpus of Contemporary American English as the first reliable monitor corpus of English,” *Literary and linguistic computing*, 25(4), pp. 447–464.

Denegri-Knott, J. (2015) “MP3,” *Consumption Markets & Culture*, 18(5), pp. 397–401.

Egbert, J., Biber, D. and Gray, B. (2022) *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. 1st ed. Cambridge University Press.

European Parliament and Council of the European Union (2019) *Directive (EU) 2019/790 on copyright and related rights in the digital single market*.

Fiorentino, G. (2010) “Che polivalente,” *Treccani*. Available at: https://www.treccani.it/enciclopedia/lingua-parlata_(Enciclopedia-dell'Italiano)/ (Accessed: February 23, 2026).

Forchini, P. (ed.) (2021) *The American Movie Corpus: a tool for the development of spoken lexico-grammatical competence*. Milano: EDUCatt.

Francis, W.N. and Kučera, H. (1964) “A standard corpus of present-day edited American English, for use with digital computers,” *Brown University, Providence*, 2.

Fraser, B. (2010) “Pragmatic Competence: The Case of Hedging,” in G. Kaltenböck, W. Mihatsch, and S. Schneider (eds.) *New Approaches to Hedging*. BRILL, pp. 15–34.

Garside, R., Leech, G. and Sampson, G. (eds.) (1987) “The CLAWS Word-tagging System,” *The Computational Analysis of English: A Corpus-based Approach.* London: Longman.

Gheno, V. (2024) *Ep. 83 - Diamo il benvenuto a “brainrot”, parola dell’anno per gli Oxford Dictionaries…*. Il post. Available at: https://open.spotify.com/episode/5E4asMIiBjhvcZkrJhb0Wm.

Ghezzi, C. and Baido, M.C.L. (2024) “Discourse markers in Italian,” *Manual of Discourse Markers in Romance*, 37, p. 479.

Giampieri, P. (2025) “Multifunctional che, cleft sentences and presentational c’è in corpora of spoken Italian,” *Research in Language*, 23, pp. 180–194.

Grandi, N. (2019) “Che tipo, l’italiano neostandard!,” *Atti dei Congressi SLI*, (2), pp. 59–74.

Halliday, M.A.K. (1990) *Spoken and written language*. 2. ed., 2. impr. Oxford Univ. Press (Language education).

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. and Suchomel, V. (2013) “The TenTen corpus family,” *7th international corpus linguistics conference CL*. Valladolid, pp. 125–127.

Khachaturyan, E. (2011) “Una classificazione dei segnali discorsivi in italiano,” *Oslo Studies in Language*, 3(1).

Knight, D. and Adolphs, S. (2022) “Building a spoken corpus: what are the basics?,” *The Routledge handbook of corpus linguistics*. Routledge, pp. 21–34.

Kučera, H. and W. Nelson, F. (1967) *Computational analysis of present-day American English*. Providence, Brown University Press.

Leech, G. (2007) “New resources, or just better old ones? The Holy Grail of representativeness,” *Language and Computers*, 59, p. 133.

Levshina, N. (2023) *Communicative efficiency: language structure and use*. Cambridge, United Kingdom New York, NY, USA Port Melbourne, VIC, Australia New Delhi, India Singapore: Cambridge University Press.

Loporcaro, M. and De Angelis, A. (2009) “Opposizioni di caso nel pronome personale: i dialetti del mezzogiorno in prospettiva romanza.”

Love, R., Brezina, V., McEnery, T., Hawtin, A., Hardie, A. and Dembry, C. (2019) “Functional variation in the Spoken BNC2014 and the potential for register analysis,” *Register Studies*, 1(2), pp. 296–317.

Lyding, V., Stemle, E., Borghetti, C., Brunello, M., Castagnoli, S., Dell’Orletta, F., Dittmann, H., Lenci, A. and Pirrelli, V. (2014) “The PAISÀ corpus of Italian web texts,” in F. Bildhauer and R. Schäfer (eds.) *Proceedings of the 9th web as corpus workshop (WaC-9)*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 36–43.

Maraschio, N. (2011) “Le nuove fonti della lingua: radio e televisione,” *V. Coletti (a cura di), L’italiano dalla nazione allo Stato, Le Lettere, Firenze*, pp. 161–171.

Marchi, A. and Ferraresi, A. (2004) “A Babel of voices: Making the Desert Island Discs Corpus.” *Corpora & Discourse International Conference*, Innsbruck.

Mauri, C., Ballarè, S., Goria, E., Cerruti, M. and Suriano, F. (2019) “KIParla corpus: a new resource for spoken Italian,” *CEUR WORKSHOP PROCEEDINGS*. SunSITE Central Europe, pp. 1–7.

McEnery, T. and Brookes, G. (2022) “Building a written corpus: what are the basics?,” *The Routledge handbook of corpus linguistics*. Routledge, pp. 35–47.

McEnery, T. and Hardie, A. (2011) *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

McEnery, T., Xiao, R. and Tono, Y. (2006) *Corpus-based language studies: An advanced resource book*. Taylor & Francis.

Mereu, D. and Negro, S.D. (2025) "Pragmatic functions and phonetic reduction: Cioè and cè in contemporary spoken Italian," *Journal of Pragmatics*, 236, pp. 1–14.

Pistolesi, E. (2018) "L'italiano in rete: usi, varietà e proposte di analisi," *Aggiornamenti*, 13, pp. 17–26.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. (2022) "Robust Speech Recognition via Large-Scale Weak Supervision." arXiv.

Rossini Favretti, R. (2000) "Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS," *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*, pp. 39–56.

Russell, S.O., Gessinger, I., Krason, A., Vigliocco, G. and Harte, N. (2024) "What automatic speech recognition can and cannot do for conversational speech transcription," *Research Methods in Applied Linguistics*, 3(3), p. 100163.

Sabatini, F. (1982) "La comunicazione orale, scritta e trasmessa : la diversità del mezzo, della lingua e delle funzioni / Francesco Sabatini," *La comunicazione orale, scritta e trasmessa : la diversità del mezzo, della lingua e delle funzioni*. Roma: s.n.

Schmidt, H. (1994) "Probabilistic part-of-speech tagging using decision trees," *Proceedings of International Conference on New Methods in Language Processing*. *International Conference on New Methods in Language Processing*, Manchester, UK.

Sinclair, J. (2005) "Corpus and Text — Basic Principles," *Developing linguistic corpora: a guide to good practice*. Oxbow Books [u.a.] (AHDS literature, languages and linguistics).

Sketch Engine (no date a) *Concordance – a tool to search a corpus*, *Sketch Engine*. Available at: https://www.sketchengine.eu/guide/concordance-a-tool-to-search-a-corpus/ (Accessed: February 11, 2026).

Sketch Engine (no date b) *List of part-of-speech tagsets available in Sketch Engine*, *Sketch Engine*. Available at: https://www.sketchengine.eu/tagsets/ (Accessed: February 3, 2026).

Sketch Engine (no date c) *TenTen Corpus Family*, *Sketch Engine*. Available at: https://www.sketchengine.eu/documentation/tenten-corpora/ (Accessed: February 13, 2026).

Spina, S. (2006) "L'italiano della televisione: una varietà intermedia tra scritto e parlato. Il caso delle dislocazioni," *Lingua e mass media in Italia. Dati, analisi, suggerimenti didattici*, pp. 153–179.

Stefanowitsch, A. (2020) *Corpus linguistics. A guide to the methodology*. Berlin: Language Science Press (Textbooks in language sciences).

Stubbs, M. (1996) *Text and corpus analysis: Computer-assisted studies of language and culture*. Blackwell Oxford.

Sullivan, J.L. (2019) “The Platforms of Podcasting: Past and Present,” *Social Media + Society*, 5(4).

Svartvik, J. (ed.) (1990) *The London-Lund corpus of spoken English: description and research*. Lund : Bromley: Lund Univ. Press ; Chartwell-Bratt.

TALP Research Center (no date) *Tagset it - FreeLing User Manual*, *FreeLing User Manual*.

Taylor, C. (2014) “Investigating the representation of migrants in the UK and Italian press: A cross-linguistic corpus-assisted discourse analysis,” *International journal of corpus linguistics*, 19(3), pp. 368–400.

Taylor, L. and Knowles, G. (1988) *MANUAL OF INFORMATION TO ACCOMPANY THE SEC CORPUS. The Machine-Readable Corpus of Spoken English.*, *icame.info*. Available at: https://icame.info/icame_static/manuals/SEC/INDEX.HTM (Accessed: October 8, 2025).

TEI Consortium (ed.) (no date) *Guidelines for Electronic Text Encoding and Interchange. [4th September 2025]*. Available at: http://www.tei-c.org/P5/ (Accessed: January 22, 2026).

TG24, S. (2025) *Podcast in Italia, 18 milioni di ascoltatori nel 2025: i dati*. Available at: https://tg24.sky.it/lifestyle/2025/09/30/audible-podcast-ascolti-italia-dati (Accessed: January 5, 2026).

Thompson, B., Dhaliwal, M., Frisch, P., Domhan, T. and Federico, M. (2024) “A shocking amount of the web is machine translated: Insights from multi-way parallelism,” *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1763–1775.

Treccani (no date) *Burocratese*, *Treccani*. Available at: https://www.treccani.it/enciclopedia/burocratese_(Enciclopedia-dell%27Italiano)/ (Accessed: January 22, 2026).

Troncone, L. (2025) “Building it-tok: an italian TikTok corpus,” Clic-It2025.

Voghera, M., Iacobini, C., Savy, R., Cutugno, F., Alfano, I. and Rosa, A. (2018) “VoLIP: a searchable corpus of spoken italian.”

Whisper API (2025) *Which Whisper Model Should I Choose?* Available at: https://whisper-api.com/blog/models/ (Accessed: January 27, 2026).

Wiberg, E. (2011) “Passato remoto,” *Treccani*. Available at: https://www.treccani.it/enciclopedia/passato-remoto_(Enciclopedia-dell'Italiano)/ (Accessed: February 23, 2026).

Xiao, R. and Yue, M. (2009) “Using corpora in translation studies: The state of the art.”

Yakura, H., Lopez-Lopez, E., Brinkmann, L., Serna, I., Gupta, P., Soraperra, I. and Rahwan, I. (2024) “Empirical evidence of Large Language Model’s influence on human spoken communication,” *arXiv preprint arXiv:2409.01754* [Preprint].

Zingaro, A. (2024) “Neostandard o substandard? Criteri metodologici di orientamento,” *Cuadernos de Filología Italiana*, 31, pp. 229–247.

Appendix 1

Copyright and Legal Compliance Statement of the PodIT corpus

1. Copyright compliance

The PodIT corpus was developed within the academic environment of the Department of Interpreting and Translation (DIT) at the University of Bologna as part of a non-commercial scientific research project.

As the corpus contains material derived from works that remain under copyright protection, relevant legislation and European directives were carefully reviewed to ensure compliance with applicable copyright and intellectual property laws. In particular, Articles 3 and 4 of Directive (EU) 2019/790 of the European Parliament and the Council of the European Union regulate text and data mining activities by introducing specific exceptions that permit the reproduction and extraction of copyrighted works for the purposes of scientific research. These provisions allow researchers and research institutions to copy and process lawfully accessible materials in order to analyze them as data.

In accordance with the Directive, all materials included in the PodIT corpus were lawfully accessible at the time of collection (i.e., freely available online without the need for payment or account registration). The corpus consists exclusively of transcriptions of short excerpts (approximately five minutes each) from podcasts. These excerpts are limited in scope and do not constitute a substitute for the original works. Given that the original podcasts remain freely available online, the corpus does not replace, compete with, or diminish access to the source material.

All relevant right holders are acknowledged, and full source references are provided. For each excerpt, the corresponding source link is made available so that users may access and listen to the original episode in its entirety.

At the time of writing, the corresponding audio recordings are not accessible and are therefore not distributed with the corpus. In future stages of the project, access to audio materials may be granted upon request to verified researchers strictly for non-commercial scientific research purposes.

If any podcast creator whose work has been included in the corpus believes that the inclusion of a transcription excerpt infringes their rights or causes damage to their intellectual property, they are invited to make contact via e-mail at joannagiacobbe@gmail.com. Upon notification and verification, the relevant transcription snippet will be promptly removed from the corpus.

2. User responsibility

While the corpus has been compiled with careful attention to the applicable legal framework, responsibility for any use made of the corpus rests solely with the user. Users are required to ensure that their use complies with all applicable copyright laws and licensing conditions. Any infringement resulting from the use, reproduction, distribution, or modification of the corpus or its contents shall be the sole responsibility of the user.

Abstract

A corpus is large collection of samples of authentic use, selected to be representative of a whole language or language variety. A corpus is the fundamental tool for corpus linguistics, a branch of linguistics that studies language through real-life examples of language use. The present dissertation addresses the design and collection of PodIT, a corpus of spoken Italian in the media. The corpus was designed with the objective of creating a representative and balanced corpus; it was created by combining manual and automated processes, in order to achieve the best possible result while acknowledging time and resource constraints. The corpus contains 100,000 words, it is POS-tagged and lemmatized, and the texts all contain relevant metadata that can be leveraged to conduct contrastive analyses within the corpus. PodIT was also analyzed in order to explore its possibilities and advantages. The analysis has also proven the validity of the corpus as a resource to analyze both spoken Italian in the media and general spoken Italian.