FROM PEC TO PEC24: A NEW REFERENCE CORPUS FOR ITALIAN

Stefania Spina, Fabio Zanda, Irene Fioravanti[1]

1. CORPUS LINGUISTICS: THE ROLE OF CORPORA IN RESEARCH AND LANGUAGE TEACHING[2]

Corpus linguistics emerged within the area of language sciences around the mid-1960s. It is usually traced back to the publication of the Brown corpus (Francis, Kučera, 1964), the first electronic corpus of written English. According to one of its many definitions, «Corpus linguistics is the investigation of linguistic research questions based on the complete and systematic analysis of the distribution of linguistic phenomena in a linguistic corpus» (Stefanowitsch, 2020: 55). This definition highlights two important and complementary elements. On the one hand, corpus linguistics is an empirical discipline, based on the systematic observation of authentic data, collected and organised in corpora. Corpora, therefore, considered as resources documenting the use of a language or of its varieties, are the core of the discipline. As a consequence, the developments in corpus linguistics have been paralleled by the advancements in the corpora that have been created over the decades, in terms of size and quality, diversification, balancing and annotation of data. On the other hand, if corpus linguistics is «the investigation of linguistic research questions», this implies that «it is not directly about the study of any particular aspect of language. Rather, it is an area which focuses upon a set of procedures, or methods, for studying language» (McEnery, Hardie, 2012: 1). Accordingly, corpus linguistics is an approach to language potentially able to redefine theories as well as methods, but also to bring out new theories from a renewed attitude towards data and tools to analyse them.

Examples of this revitalising effect of corpus linguistics can be observed in the field of second language acquisition (SLA): the interplay between acquisitional approaches and corpus methods have paved the way for the emergence of a new area, learner corpus research (Granger *et al.*, 2015), which has now acquired the status of an autonomous discipline, deeply rooted in the methods and principles of corpus linguistics. Another example are recent theoretical approaches such as usage-based theories (Ellis, 2017), grounded in the prominence of exposure to linguistic use in L1 and L2 acquisition, that have found in corpora as well as in methods and tools to analyse them an ideal environment to develop. Furthermore, the entire area of research of phraseology, which was considered peripheral by the generativist paradigm, had a tremendous boost with the development of corpus linguistics, also due to particularly influential scholars such as John Sinclair (Sinclair, 1991): the possibility of extracting unprecedented amounts of word combinations from corpora allowed the lexico-grammatical phenomenon of phraseology

[1] Università per Stranieri di Perugia.

[2] The research is part of an Italian Ministry of Research PRIN Project – Call 2022 – Prot. 2022HXZR5E, funded by the European Union - Next Generation EU, Missione 4 Componente 1, CUP D53D23009680006. The title of the project is: *DICI-A: A Learner Dictionary of Italian Collocations*. This work is a joint effort by the co-authors. However, Stefania Spina is responsible for sections 1, 2, 4.1, 5 and 6; Fabio Zanda for sections 3, 3.1, 3.2, and 3.3; and Irene Fioravanti for sections 4.2, 4.2.1, 4.2.2, 4.3 and 4.4.

to be analysed in great depth. In the field of discourse studies, a new approach called CADS (Corpus-assisted discourse studies; Baker, 2006; Orrù, 2017) systematically adopts corpus methods to analyse the mutual influence between discourse and society. Finally, since the early 1990s Rundell and Stock (1992) have described the impact of corpora on lexicographic activity as a "corpus revolution": corpora became the main source of empirical evidence in lexicography as well as in linguistics, so that linguistic introspection could be largely replaced or at least complemented.

This brief overview of some of its significant contributions shows that corpus linguistics is a heterogeneous field, spanning the entire spectrum of linguistic research. At the same time, it is a field constantly evolving, in methods as well as in the resources that are used. The methodological core of the discipline is the integration of a quantitative and a qualitative approach. Regardless of the phenomenon under investigation, the first phase of a corpus linguistics study always involves its distribution in a linguistic corpus: it is therefore a matter of measuring a phenomenon, through its frequency or any other type of quantitative dimension. The availability of increasingly larger corpora and of larger amounts of data has led to the systematic adoption of statistical methods to analyse the quantitative dimension of phenomena: the comparison of frequencies or other measures in different corpora, the statistical significance of any differences, the effect of linguistic or extralinguistic variables on a given phenomenon: these and other analyses, based on data extracted from corpora, have become routinised since the development of corpus linguistics (Paquot, Gries, 2020). This phase provides linguistic analysis with robust quantitative support relying on the observation of large amounts of authentic data. Next, this quantitative and measurable evidence is complemented with a qualitative analysis of linguistic data within the authentic co(n)text where they have been produced. The typical tools used for this analysis of linguistic contexts are concordances, which in their electronic version are part of the most established tradition of corpus linguistics, although they are still constantly evolving, as shown for example by recent research on artificial intelligence (AI)-assisted concordancing (Anthony, 2024). As an output generated by a preceding quantitative investigation of data, concordances provide high-quality evidence of language use (Sinclair, 1991). Furthermore, the possibility of ordering the contexts in which the analysed phenomenon occurs, and of dimensioning them to encompass larger or smaller sequences of text, allows researchers to analyse concordance lines for various properties of the search term, and to identify distributional patterns in language (Wulff, Baker, 2020). This ability to uncover patterns, revealing what is systematically recurrent and typical of a language, is one of the most distinguishing features of corpus linguistics.

Along with its contribution in linguistic research, since its beginnings corpus linguistics has fostered the development of concrete applications in the field of language learning, resulting in a pedagogical approach called Data-driven learning (DDL), in which the authentic linguistic data included in corpora are used to guide learners towards an autonomous discovery of linguistic uses and recurrent patterns (Forti, 2023). This approach is also spreading to Romance languages and Italian, giving rise not only to innovative pedagogical practices but also to research into the effects of the new methodologies on language learning (Tyne, Spina, 2025).

2. CORPUS LINGUISTICS IN ITALY

In the Italian context, the use of corpora in linguistic research was introduced relatively early, particularly in the field of frequency dictionaries: the *Lessico di frequenza della lingua italiana contemporanea* (LIF; Bortolini *et al.*, 1972) was built from an early corpus of written Italian. Since the 1980s, Tullio De Mauro's groundbreaking work has led to the design

and compilation of seminal linguistic resources entirely based on the use of corpora of Italian, which were built according to rigorous principles of balancing and sampling. The *Vocabolario di base della lingua italiana* (De Mauro, 1980) and the *Lessico di frequenza dell'italiano parlato* (LIP; De Mauro *et al*., 1993) paved the way for the use of frequency and dispersion as measures capable of classifying Italian vocabulary into different usage bands and gave a strong boost to empirical studies on spoken Italian. A further qualifying element of De Mauro's pioneering work in the area of corpus linguistics is that the resulting linguistic resources were designed for specific educational contexts: each of the different bands of the *Vocabolario di base* (VdB) is linked to the level of schooling required to understand the texts that contain them, ascribing in this to the discipline the crucial role of providing tools for addressing real-world social problems, that has been recognised in much more recent times (McEnery, Brookes, 2024).

Since then, many corpora of Italian have been created, as will be described in the following sections. They are often projects independent of each other, carried out by individual scholars or research teams, as is the case of other languages. What distinguishes Italian from other languages is the lack, to date, of a "national" corpus, a true written and spoken reference corpus, based on a thorough design of structure, data collection and annotation, and «designed to provide comprehensive information about a language [...] large enough to represent all the relevant varieties of the language, and the characteristic vocabulary, so that it can be used as a basis for reliable grammars, dictionaries, thesauri and other language reference materials» (EAGLES, 1996). Reference corpora including both written and spoken texts exist for many languages, including, just to give a few examples: the *British National Corpus* (BNC Consortium, 2007) and the *British National Corpus 2014* (Brezina *et al*., 2021; Love *et al*., 2017), the *Russian National Corpus* (Savchuk *et al*., 2024), the *National Corpus of Polish* (Przepiórkowski *et al*., 2009), the *National Corpus of Irish* (Bhreathnach *et al*., 2024), the *Corpus de Referencia del Español Actual* (Real Academia Española, 1994), the *Corpus del Español del Siglo XXI* (Real Academia Española, 2013), and the *Corpus de référence du français contemporain* (Siepmann *et al*., 2016).

In addition to this, corpus linguistics has not yet reached a complete level of institutionalisation as an academic discipline in Italy: for instance, there are no Italian scientific journals dedicated to corpus linguistics studies[3], just as there are no regularly organised conferences on corpus linguistics[4], nor are there any academic associations of corpus linguists[5]. In terms of higher education, the discipline is taught in only a few master's degree courses (as far as we know, at the universities of Torino, Bologna, Firenze, Bergamo and Catania) and bachelor's degree courses (Perugia Stranieri). Even the systematisation of the principles of the discipline in reference works has been relatively slow: the first Italian handbook of Corpus linguistics, for instance, was only published in 2001 (Spina, 2001).

A similar reluctance can be found at the level of scientific studies on Italian: while data extracted from corpora are now widely used in all areas of linguistics, where they constitute one of the most widespread sources of evidence, the systematic application of the principles and methods of corpus linguistics is still rather rare. On the one hand, the quantitative perspective is often relegated to a mere numerical observation of the

[3] In the English-speaking context journals specifically dedicated to corpus linguistics have been active for several decades, such as *International journal of corpus linguistics* (since 1996), *Corpus linguistics and linguistic theory* (2005), *Corpora* (2006), and *International journal of learner corpus research* (2015). This is also the case in France (*Corpus,* since 2002) and Spain (*Research in Corpus Linguistics*, since 2013).

[4] Narrowing it down to the European context, the biennial *International Corpus Linguistics conference* is now in its 13th edition; the annual Spanish *Congreso internacional de Lingüística de corpus* is currently in its 16th edition.

[5] The *Spanish Association for Corpus Linguistics* was founded in 2008 at the University of Murcia; the *Learner Corpus Association* was founded at the Université catholique de Louvain (Belgium) in 2013.

occurrence of phenomena, without being based on a robust statistical methodology; on the other hand, the qualitative approach is rarely a systematic analysis of concordances, and thus an exhaustive and in-depth investigation of phenomena in the authentic context in which they occur.

In a nutshell, despite some pioneering work, in the Italian context corpus linguistics is slow to be institutionalised as an academic discipline, and lags in systematically applying the principles and methods of the discipline to the analysis of Italian.

3. ITALIAN CORPORA

In this section, we review three types of corpora available in the Italian language panorama: written corpora, spoken corpora, and web corpora. Corpus selection criteria focus on free online searchability, large size, monolingual scope, and a primary focus on contemporary language, while excluding learner corpora. Although this review may not be exhaustive, it highlights the main resources available online for Italian corpus-based research on present-day language.

3.1. *Written corpora*

Written corpora of contemporary Italian can be distinguished into two main strands: corpora including a diversified range of texts of general written Italian, and corpora focusing on a specific written variety.

Following the pioneering work of the LIF (Bortolini *et al.*, 1972) reported in the previous section, the first tool that was presented in the internet era as a large written corpus of general Italian is the *Corpus e Lessico di Frequenza dell'Italiano Scritto* (CoLFIS; Laudanna *et al.*, 1995; Bertinetto *et al.*, 2005). It is composed of over 3 million words drawn from contemporary Italian written texts, divided into three main sections of newspapers, magazines, and books. Each section and subsection consist of different text genres, which were selected on the basis of a national survey reflecting the reading preferences of the Italian readership (ISTAT, 1993). CoLFIS established itself as a reference source for research in psycholinguistics and neurolinguistics (Bambini, Trevisan, 2012), while being employed as the lexical source for a number of open-access linguistic tools, such as a phonological lexicon of Italian word forms (Goslin *et al.*, 2014) and an annotated lexicon of Italian derivatives (Talamo *et al.*, 2016). CoLFIS has been freely accessible and searchable through a dedicated platform, the *EsploraCoLFIS* (Bambini, Trevisan, 2012)[6].

Another large freely available corpus created in the following years is the *CORpus di Italiano Scritto* (CORIS; Rossini Favretti *et al.*, 2002). The first version of CORIS was released in 2001 and included around 100 million words. Since then, it has been updated every three years by means of a monitor version, reaching 165 million words in 2021 while maintaining the original design and proportions (Tamburini, 2022). In fact, CORIS still consists of six main subcorpora, corresponding to different textual macro-varieties categorised as *Press* (38% of total texts), *Fiction* (25%), *Academic prose* (12%), *Legal and administrative prose* (10%), *Miscellanea* (10%), and *Ephemera* (5%). Alongside CORIS, which is described by its authors as a 'defined model', a 'dynamic model' was proposed, the *COrpus Dinamico dell'Italiano Scritto* (CODIS; cf. Tamburini, 2002). CODIS contains exactly the same texts as CORIS but allows corpus users to combine or recompose the sections according to their specific research needs. Subsequently, CORIS was accompanied with a

[6] https://linguistica.sns.it/esploracolfis/home.htm.

diachronic extension from 1861 to 2001 which allows for historic enquiries, covering the same macro-varieties as CORIS (with the exception of Ephemera), the DiaCORIS (Onelli *et al.*, 2006). CORIS and its derived corpora can be freely queried on a dedicated webpage[7].

Most written corpora of Italian tend to focus on a specialised variety of language, such as the legal (e.g., *Jus Jurium*, in progress; cf. Barbera *et al.*, 2022) or the academic (e.g., *Athenaeum*; cf. Barbera *et al.*, 2007) ones, which cannot all be covered here. Among the largest specialised corpora, the corpus *La Repubblica* (Baroni *et al.*, 2004) is particularly notable for its size and coverage, with texts gathered from all the issues published by the newspaper *La Repubblica* between 1985 and 2000, for a total of over 5,000 issues, 500,000 documents, and more than 300 million words. It is searchable through the freely accessible NoSketch Engine platform[8] (Rychlý, 2007).

A promising written corpus appears to be *Univers-ITA* (Grandi *et al.*, 2023a), which includes over 2,000 *ad hoc* texts authored by university students in various academic areas and multiple locations in Italy, for a total of ca. 800,000 words. It is annotated with rich socio-linguistic metadata related to text authors which allows for advanced searches and text analyses. Within the same project, two other non-*ad hoc* corpora were designed: the *Univers-ITA-ProUniv* (Grandi *et al.*, 2023b), which compiles students' dissertations and reports (ca. 5.5 million words), and the *Univers-ITA-ProGior* (Grandi *et al.*, 2023c), made up of texts drawn from university newspapers (ca. 1.5 million words). All the three corpora originated within the UniverS-ITA project are searchable on a freely accessible page based on the Sketch Engine platform[9].

3.2 *Spoken corpora*

The first spoken corpus developed to create a reliable frequency-based lexicon of the most frequent 3,000 lemmas in spoken Italian is the *LIP* (De Mauro *et al.*, 1993). It consists of about 60 hours of recording (ca. 500,000 words) for a total of 469 transcriptions, included in five equally represented sections corresponding to just as many diaphasic situations: face-to-face conversations, telephone conversations, interviews and debates, monologues and TV/radio. Moreover, relevance was given to another balance criterion, i.e. diatopic variation. In fact, LIP's recordings were collected in four Italian cities: Milan, Florence, Rome, and Naples, with 25,000-token transcribed texts for each one of the five main sections, for a total of ca. 125,000 tokens per city. The LIP transcriptions were tagged by part of speech (henceforth, *POS*) and made available for queries online via the tool *VoLIP* (Voce del Lip; Voghera *et al.*, 2014)[10], which also provides access to the audio recordings.

The *Corpora e Lessici di Italiano Parlato e Scritto* (CLIPS; Savy, Cutugno, 2009) is a spoken corpus consisting of about 100 hours of recorded speech collected in 15 cities in Italy, selected on the basis of geo-linguistic, socio-linguistic, and socio-economic criteria (cf. Sobrero, Tempesta, 2007). CLIPS is made up of five main subcorpora corresponding to several diamesic varieties, namely elicited dialogues, read-aloud speech, radio and TV recordings, telephone conversations, and ortho-phonic recordings. CLIPS data is accessible, searchable, and downloadable upon registration[11].

[7] https://corpora.ficlit.unibo.it/.
[8] https://bellatrix.sslmit.unibo.it/noske/public/#dashboard?corpname=repubblica.
[9] https://corpora.ficlit.unibo.it/CUSP/crystal/index.html#dashboard?corpname=UniverS-Ita.
[10] https://www.volip.it/.
[11] http://www.clips.unina.it/home.

The *LABLITA* corpus (Cresti *et al.*, 2022) is a corpus of spontaneous speech recorded in various diaphasic situations in Tuscany since 1965. It can be considered as an 'open' corpus, as it consists of three sub-corpora produced for different aims: the *GRIT subcorpus* of the LABLITA resources, which in turn includes the *Corpus of Spoken Italian* (Cresti, 2000), the Italian sub-section of the integrated corpora for spoken romance languages, the *C-ORAL-ROM* corpus (Cresti, Moneglia, 2005), and the first spoken corpus of Italian, the *Stammerjohann corpus* (Stammerjohann, 1971). These sub-corpora were merged to create a ca. 700,000 token spoken corpus, with 422 transcripts involving 1,000 different speakers and their corresponding metadata (education, sex, age, profession, and origin). The LABLITA corpus includes non-balanced diamesic samples of face-to-face conversations, telephone conversations, and media broadcasting. Its transcripts are annotated according to the Language into Act Theory (Cresti, 2020) and they are accessible through the platform Orféo (*outils pour l'étude du Français écrit et oral*; Benzitoun *et al.*, 2016) which allows to perform various searches including channel, type of interaction, social context, while accessing to text-aligned original audios[12].

KIParla (Mauri *et al.*, 2019) is a spoken corpus designed for the study of diatopic, diaphasic, and diastratic variation (Ballarè *et al.*, 2022). The KIParla corpus was built to be integrated over time, featuring a modular and expandable structure. Consequently, its design incorporates independent 'modules' that share a common set of metadata, the transcription method, and searchability through an open NoSketchEngine platform[13], which «guarantees a high level of mutual comparability» (Mauri *et al.*, 2019). At the time of writing, KiParla includes several different subcorpora (or 'modules') designed to explore multiple dimensions of linguistic variation and collect data in different geographical locations. The foundation units of the KIParla are the *KIP* module (cf. Goria, Mauri, 2018), focused on academic settings in Bologna and Turin (ca. 70 recorded hours), and the *ParlaTO* (Cerruti, Ballarè, 2021) which is composed of semi-structured interviews on various topics recorded in Turin, involving speakers of different age groups, origin, education, and occupation (ca. 50 hours). The core modules of the KiParla have been integrated over time with two additional subcorpora: the *KIPasti* module (Mauri *et al.*, 2024a) which includes 40 hours of recordings of spontaneous speech during meals with family and friends throughout Italy; and the *ParlaBO* module (Mauri *et al.*, 2024b), consisting of 50 hours of semi-structured interviews recorded in Bologna, with a parallel design to ParlaTO.

A peculiar type of spoken corpora is 'transmitted language' corpora, i.e. corpora derived from media broadcast, which enable the analysis of oral communication in mass media. For instance, the *Lessico dell'Italiano Radiofonico* (LIR; Maraschio *et al.*, 1997; 2003) is a database composed of two corpora collected in 1995 and 2003. It combines 141 hours of Italian radio speech, amounting to ca. 960,000 tokens. The LIR is searchable through a dedicated platform[14], allowing the selection of metadata filters – such as radio station, genre and speech types – while accessing audio recordings and annotated textual data.

Similarly, the *Lessico Italiano Televisivo* (LIT; Biffi, 2010) is a corpus consisting of 168 hours of televised speech, collected in 2006 from the Italian major networks. LIT's searchable transcripts can be filtered by metadata – e.g., TV channel, genre, type of speech, speaker characteristics – and are accompanied by the corresponding video and audio recordings[15]. To enable the diachronic study of the Italian televised speech, the *Lessico Italiano Televisivo in Diacronia* (DIA-LIT; Cialdini, 2016) was developed as an

[12] http://corpus.lablita.it/.
[13] https://search.corpuskiparla.it/corpus/crystal/#open.
[14] http://lir.accademiadellacrusca.org/lir2/.
[15] http://lit.accademiadellacrusca.org/lit2/.

extension of the LIT, maintaining the same architecture while adding another 40 hours of video transcriptions, which are searchable through a dedicated interface[16].

3.3. *Web corpora*

Corpora composed exclusively of texts retrieved from the web have been around for the last two decades. Usually, web corpora allow for large-scale collection of text data with reasonable effort. Nevertheless, the trade-off is that it is generally difficult to ensure balance and quality of automatically web-crawled data, which requires pre- and post-processing noise cleaning (e.g., Jakubíček *et al.*, 2020) and additional work for assigning metadata typically present in non-web corpora, such as genre or register information of the extracted texts (Laippala *et al.*, 2020).

Many web corpora built for linguistic use have been created using web crawling tools (e.g., Baroni *et al.*, 2006) and are part of larger scope multilingual projects. For instance, this is the case for the *Italian Web corpus* (itWac), a «very large Italian general-purpose Web corpus» (Baroni, Ueyama, 2006) developed in the early stages of the *Web-As-Corpus Kool Yinitiative* (WaCky) project (Baroni *et al.*, 2009), which today encompasses over fifty large web corpora of various target languages[17]. ItWac contains almost two million texts (over 1.5 billion words) retrieved from the .it domain using specific words and word pairs as seeds. It has been POS-tagged, lemmatised, and is freely searchable on NoSketch Engine.[18]

The *Newsgroups UseNet Corpora* (Nunc; cf. Barbera *et al.*, 2022) include several corpora composed of texts retrieved from online open discussion groups. The Nunc are part of a multilingual project, but the Italian section is the most developed one. In fact, it consists of two large general corpora, each containing over 100 million tokens, and several smaller specialised corpora focusing on topics such as cooking, automotive, photography, and cinema. Nunc corpora have been POS-tagged and lemmatised, and are searchable on a dedicated CQP-based webpage[19].

PAISÀ (Lyding *et al.*, 2014) is a corpus of authentic contemporary Italian written texts gathered from the web. It includes over 250 million tokens and ca. 380,000 texts, most of which were harvested from Wikipedia, while the rest were crawled using specific seeds based on the Italian basic vocabulary word list (i.e. the VdB list; De Mauro, 1980) and on specific bigrams (cf. Lyding *et al.*, 2014). PAISÀ has been lemmatised, POS-tagged, and annotated for syntactic dependencies, and is freely searchable on its website through a dedicated interface[20].

Although the above-mentioned web corpora are the main freely searchable ones, it is important to note that many other web corpora exist online. However, they have been overlooked in this review due to their specialised focus or because they are accessible only through paid platforms. In the latter case, some of the largest web corpora of Italian were crawled within the context of the *TenTen Corpora Family* (Jakubíček *et al.*, 2013), a project which aims to build 10-billion-word corpora for each language included. The TenTen corpora are available exclusively via the subscription-based online corpus management software *Sketch Engine*[21] (Kilgarriff *et al.*, 2014). To date, the most recent Italian corpus belonging to this project, the *itTenTen20*, consists of over 30 million documents (almost

[16] http://193.205.158.203/dialit/.
[17] https://www.sketchengine.eu/wac-corpora/.
[18] https://bellatrix.sslmit.unibo.it/noske/public/#dashboard?corpname=itwac_full.
[19] https://www.corpora.unito.it/index_nunc.php.
[20] https://www.corpusitaliano.it/it/access/simple_interface.php.
[21] https://www.sketchengine.eu/documentation/tenten-corpora/.

15 billion tokens) harvested from various web domains, although it is only partially annotated by topic.

The linguistic resources surveyed in this section encompass most of the corpora publicly available for research on contemporary Italian. While these tools vary in scale, scope, and metadata availability, they are all freely searchable, notwithstanding the clear differences in access conditions and query interfaces, which cannot be examined in greater depth here. Within this broader landscape, we introduce the updated version of the Perugia Corpus (PEC; Spina, 2014), describing its design, composition and annotation (§ 4), and its role as a reference corpus within the wider context of Italian language corpora (§ 5).

4. THE PEC24 CORPUS: DESIGN, COMPOSITION AND ANNOTATION

4.1. *What's new in PEC24 after 10 years from PEC 2014*

Against the background described in the previous paragraphs, the PEC (Spina, 2014), in its first version, was published and made available for querying in 2014. Since then, the purpose of the PEC was to address the lack of a reference corpus of Italian, and to provide the research community with a general corpus of wide coverage of the Italian language. The rationale behind the design of the PEC was to privilege the differentiation of textual genres, including spoken ones, at the expense of corpus size.

Failing a broad and structured project with several academic partners on which to ground a large reference corpus such as those mentioned in par. 2 for other languages, priority was given to the reuse of already existing and available resources (Zampolli, 1991), which were often fragmented and difficult to access, according to the fourth principle of the FAIR paradigm on Findability, Accessibility, Interoperability, and Reuse of scientific data (Wilkinson *et al.* 2016). To this body of existing data, new data was added, trying to balance the different sections of the corpus as much as possible. From its very first version, the PEC can therefore be considered a low-cost reference corpus (Spina, 2014), rather small in size but with a good representativeness of different written and spoken varieties of Italian.

A major update of the PEC, with a significant increase in its size, was started as early as 2021, but this activity was intensified in 2023: the identification and extraction of collocations of various syntactic types to be included in a *Learner dictionary of Italian collocations* (Spina *et al.* under review) required the availability of a reference corpus with a wide coverage of written and spoken textual genres, but larger in size. In the following paragraphs we describe the result of this upgrade, the PEC24 corpus, that has been released at the end of 2024 and made publicly searchable in the CQPweb-based site *Search-it* [22].

The PEC24 corpus preserves the same structure as his predecessor, being divided into 10 textual genres, representing some of the main written and spoken varieties: seven written genres (literary fiction, non-fiction, newspapers, academic writing, school essays, administrative writing, and web texts), and three spoken genres (tv programs, film dialogues, and spoken texts).

Table 1 shows the additions made to PEC24 in terms of number of texts, number of tokens and percentage of tokens to the overall number. In terms of total size, PEC24 contains 46,949,410 tokens, 20,457,799 more than PEC, and is therefore 43.6% larger. This increase in size is distributed through all the 10 sections, each of whom has been

[22] https://lt.eurac.edu/cqpweb/.

upgraded, but is mainly concentrated in some of them. Overall, the spoken texts have been increased by 3.4%, and the entire spoken part, also including tv programs and film dialogues totals 17.2% of the PEC24. The written part of the corpus also has a different distribution in terms of percentage to the total number of tokens: compared to its predecessor, in PEC24 three of the most prominent written genres, i.e. literary fiction, newspapers and school essays, are represented with the same percentage (around 14%), while web texts are still the most prominent (24%). The upgrade to the new version has led in particular to the newspapers occupying a proportionally smaller place, and school essays being conversely more numerous. Overall, the PEC24 is a less newspaper-centric reference corpus, with a better balance between the different textual genres.

Table 1. *A comparison between PEC and PEC24 in terms of texts and tokens*

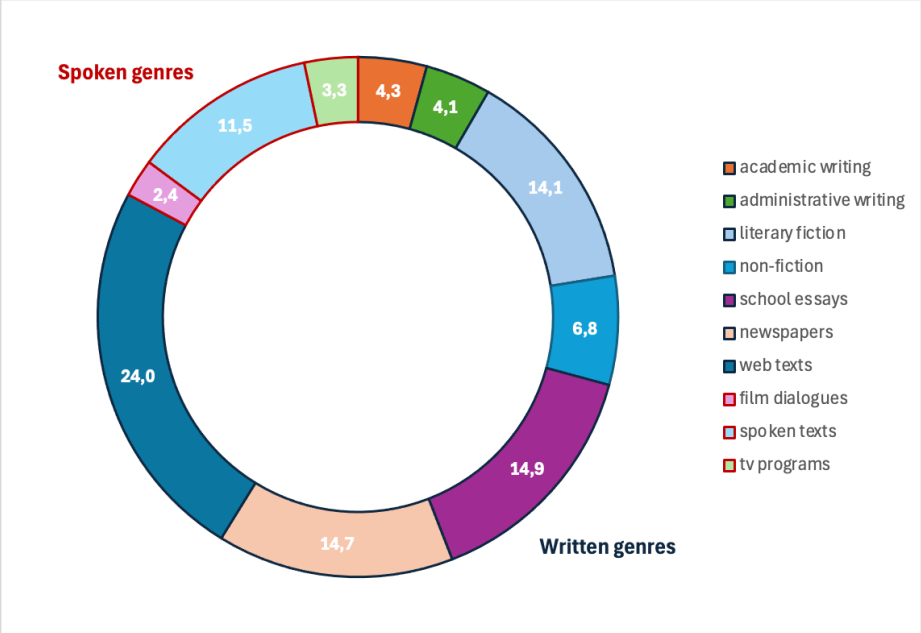
Text genres	PEC texts	PEC24 added texts	Total PEC24 texts	PEC tokens	PEC24 tokens	%PEC	%PEC24
Written genres							
Academic writing	240	75	315	1,113,606	2,003,969	4.2	4.3
Administrative writing	119	75	194	1,156,368	1,914,625	4.4	4.1
Web	27,343	78,624	105,967	7,359,077	11,266,857	27.8	24
Literary fiction	60	30	90	3,545,430	6,623,695	13.4	14.1
Non-fiction	79	28	107	2,355,003	3,172,781	8.9	6.7
School essays	4,051	21,086	25,137	1,258,031	6,989,768	4.7	14.9
Newspapers	8,232	2,201	10,433	5,722,001	6,902,520	21.8	14.7
Spoken genres							
Tv programs	127	69	196	1,147,251	1,556,098	4.3	3.3
Film dialogues	66	50	116	626,289	1,107,452	2.4	2.3
Spoken texts	1,041	1,335	2,376	2,158,555	5,431,647	8.1	11.5
TOTAL	41,358	103,573	144,931	26,491,611	46,969,412		

4.2. *Corpus design: 10 textual genres*

As already mentioned above, the new PEC24 presents the same design as the PEC corpus (Spina, 2014). It is therefore made of ten sections, which are divided in turn into subsections. The written section covers 82.8% of the total corpus (142,243 texts amounting to approximately 39 million tokens); the three spoken sections cover the remaining 17.2% of the corpus with 2,688 texts, which amount to approximately 8 million tokens.

Figure 1 shows the distribution of the ten sections in the corpus. In the following paragraphs, we describe the internal composition of each section in 47 subsections.

Figure 1. *Distribution of the ten sections in the PEC24*



4.2.1. *Written sections*

Literary fiction

The section including literary fiction has been expanded by 30 texts compared to the PEC corpus, reaching a total of 90 texts (approximately 7 million tokens). The add-on has the same structure as the previous version: the 30 novels are included through a sample of 100.000 tokens each. This section includes texts from 90 contemporary Italian novels published between 1991 and 2020 by 66 Italian writers.

Non-fiction

The non-fiction section includes essays on various topics, organised into four main themes: current affairs; biography; politics; and free time. The essays were published by Italian authors between 1990 and 2023. This section comprises 107 texts, totaling more than 3 million tokens. The 28 texts added in the new PEC24 corpus are samples of 25.000 tokens each.

Newspapers

The newspapers section comprises 10,433 texts, amounting to approximately 7 million tokens, with an increase of 2,201 texts in PEC24. The articles were extracted from 7 Italian daily newspapers (Corriere della Sera, Repubblica, Il Fatto Quotidiano, Il Sole 24 Ore, Avvenire, Domani and Huffington Post), and from a weekly newspaper (Espresso), and were published between 2011 and 2024. They cover nine different newspaper subsections, with news and politics being the most represented. Table 2 shows the token distribution across the nine subsections of the newspaper section.

Table 2. *The nine subsections of the newspapers section*

Subsection	tokens	%
Economics	648,073	9.4
Foreign affairs	906,788	13.1
Culture	721,685	10.5
Letters	120,099	1.7
News	1,849,412	26.8
Sports	462,079	6.7
Entertainment	396,548	5.7
Editorial	512,098	7.4
Politics	1,285,738	18.6

Academic writing

The academic writing section includes five subsections (scientific articles, handouts, PhD theses, handbooks, and dissertations) which belong to three thematic areas: humanities (33.4% - 670,225 tokens); law and economics (33.6% - 673,976 tokens); and science (33% - 659,768 tokens). The academic writing section comprises 315 texts, totaling more than 2 million tokens, with an increase of 75 texts compared to the PEC corpus, all belonging to the new subsection with PhD theses. The PhD theses were also evenly divided between the three thematic areas (25 per area), and a sample of 25,000 tokens of each was included. Further, the *Corpus di Italiano Accademico* (Spina, 2010) has been fully integrated into this section. Table 3 reports the total number of tokens for each subsection.

Table 3. *The subsections of the academic section*

Subsection	tokens	%
Articles	261,271	13
Handouts	549,914	27.4
PhD theses	890,362	44.4
Handbooks	134,561	6.7
Dissertations	167,861	8.4

School essays

A total of 25,137 texts were included in the section on school essays. This section has been significantly expanded compared to the PEC corpus, which contained 4,051 texts for the same section. The total number of tokens amounts approximately to 7 million, consisting of two subsections: essays of lower secondary school (37.4% - 2,618,984 tokens) and essays of higher secondary school (62.6% - 4,370,784 tokens). The school essays were automatically extracted from *Repubblica Scuola* and were produced between 2008 and 2020 by lower secondary and higher secondary students on 75 different topics

(e.g. the role of internet in our society, the problem of immigration, the last book you have read, why do you like travelling, etc.).

Administrative writing

The administrative section contains 194 texts, with an increase of 75 texts compared to the PEC corpus. It includes the two subsections of laws, covering 64.6% (1,236,132 tokens) of the entire section, and regulations, representing 35,4% (678,493 tokens) of the administrative section.

Web

The web section represents the largest part of the PEC24, with an increase of 78,624 texts: 9,397 posts from 10 blogs, and 69,227 tweets randomly selected among those sent in 2012 and 2013 and composed of more than 19 words. This choice was made in an attempt to avoid those texts consisting of only a few verbal elements and containing mainly emojis or reactions to other texts. The web section totals 105,967 texts and 11,266,857 million tokens. It is further divided into five subsections: chat, blogs, forums, social media, and Wikipedia entries. Regarding blogs, only texts were considered, excluding the comments. The Wikipedia texts were extracted from the full Italian version and selected randomly. The social media subsection includes posts extracted from Facebook and Twitter. Table 4 shows the token distribution across the five subsections.

Table 4. *The token distribution across the five sections*

Subsections	**tokens**	**%**
Chat	119,005	1.1
Blogs	5,006,467	44.4
Forums	171,092	1.5
Social media	2,317,412	20.6
Wikipedia	3,652,880	32.4

4.2.2. *Spoken sections*

Film dialogues

The film section has been increased by 50 texts, totaling 116 texts and 1,107,452 million tokens. It comprises the full transcription of dialogues from 116 Italian movies produced between 1986 and 2022. Subtitles websites were used to obtain the full transcriptions, which were subsequently manually checked.

Spoken texts

A part of the PEC spoken section was represented by texts from existing corpora, already available in 2014: the texts from the LIP corpus (De Mauro *et al.*, 1993); the Italian section of the *Saccodeyl* corpus (Pérez-Paredes, Alcaraz-Calero, 2007); and some transcriptions from the CLIPS corpus (Savy, Cutugno, 2009). The texts had been re-annotated according to the PEC corpus criteria.

Additional 1,335 transcriptions were added to the PEC24 spoken section, bringing the total to 2,376 texts (5,431,647 million tokens). Examples of these adds-on are the

CIP-PG corpus (*Corpus pilota dell'italiano parlato*; Spina *et al.*, 2020), the Italian section of the TEDx multilingual corpus (Salesky *et al.*, 2021), with the transcriptions of 496 TEDx conferences, a sample of the ParlaMint-IT corpus (28 sessions of the Italian Parliament from 2013 to 2020; Erjavec *et al.*, 2024), and more than 700 lyrics of Italian songs from 1990 to 2020, that are part of the COCI corpus (*Corpus della canzone italiana* 1958-2022; Coccia, 2023). The spoken section maintains the subdivision into the two sections of the PEC corpus: dialogic speech, covering 23.5% (1,281,082 tokens), and monologic speech, representing 76.5% (4,150,565 tokens) of the section. Both dialogic and monologic speech are in turn distinguished in different subsections, which are shown in table 5.

Table 5. *Dialogic and Monologic Speech sections and their subsections*

Subsections	**tokens**	**%**
DIALOGIC SPEECH		
Interviews	526,283	9.7
Conversations	327,303	6
Phone calls	283,652	5.2
Classroom interactions	62,011	1.1
Interrogations	45,257	0.8
Meetings	36,576	0.7
Total Dialogic Speech	*1,281,082*	*23.5*
MONOLOGIC SPEECH		
Institutional speeches	1,744,061	32.2
Conferences	1,244,716	23
Songs	396,424	7.3
Lectures	256,729	4.7
Trials	174,730	3.2
Sermons	168,917	3.1
Political speeches	158,346	2.9
Oral exams	6,642	0.1
Total Monologic Speech	*4,150,565*	*76.5*

Tv programs

The section on television programs has been expanded with 69 texts (41 talk shows, 27 episodes of TV series and a government press conference broadcast on tv), bringing the total to over 1,5 million tokens (1,556,098 tokens). It contains data from the *Corpus di Italiano Televisivo* (Spina, 2005), already included in the PEC corpus. The television programs belong to seven subsections (see Table 6), with talk shows being the most represented one.

Table 6. *The token distribution across the seven subsections*

Subsections	tokens	%
Fiction	261,207	16.8
Information	235,619	15.1
Talk shows	830,296	53.4
Advertising	30,330	1.9
Sports commentaries	58,446	3.8
Sports news	54,748	3.5
Entertainment	85,452	5.5

4.3. *Linguistic annotation*

The PEC24 corpus features a two-level annotation system: the first level concerns the annotation of text structure, while the second level involves linguistic annotation.
The text structure has been annotated using XML (eXtensible Markup Language), a markup language that allows for defining a document's structure and text content by assigning labels or tags.

As a reference corpus, and given the heterogeneity of the texts included in PEC24, which do not derive from a centralised project adopting a single annotation scheme common to all genres, PEC24 follows recommendations for a standard and minimal annotation approach (Burnard, Bauman, 2014; Hardie, 2014): each corpus text has been assigned a unique identifier and labeled with its text genre, thus specifying the section of the corpus to which it belongs, and the following additional information, which can be used as search filters: text type (one of the 47 subsections constituting the 10 corpus sections), year (1982-2024) and channel (written and spoken).

Below, we provide an example of XML annotation for a text from the academic section, in which the following elements have been labeled: ***genre***, specifying both the text genre and the corpus section; ***type***, indicating the type of academic production (in this case, a doctoral dissertation); ***subject***, referring to the thematic area; ***author***, ***title*** of the dissertation; ***year*** of degree completion; and the written ***channel*** to which it belongs.

<text id="acc2550" genre="accademico" type="ACCADEMICO_phd" subject="Scienze_umane" author="$NAME$" title="Apogeo e crisi della politica culturale comunista" year="2022" channel="s">

Subsequently, both the written and the spoken data were annotated for grammatical category and lemmatised using *TreeTagger* (Schmid, 1994), one of the most widely used programs for POS tagging, with a tagset specifically created for the annotation of the PEC corpus (Spina, 2014)[23]. The tagset has an articulated structure and consists of 54 labels, allowing for a detailed annotation of Italian parts of speech. The grammatical category with the highest level of differentiation is that of verbs, which are classified into verbs (VER), auxiliaries (AUX), and modal verbs (VER2). Each category is further subdivided into verbs of finitive mode (e.g., VER:fin), infinitives (e.g., VER:infi), past participles (e.g., VER:ppast), present participles (e.g., VER:ppre), and gerunds (e.g., VER:geru), with additional distinctions based on the presence or absence of clitic pronouns (e.g., VER:fin

[23] The tagset can be accessed at the following link https://osf.io/b6p3g.

vs. VER:fin:cli). The adjective category includes a distinction between descriptive adjectives (ADJ), possessive adjectives (DET:poss), demonstrative adjectives (DET:demo), and indefinite adjectives (DET:indef). Pronouns are similarly categorised into demonstrative pronouns (PRO:demo), indefinite pronouns (PRO:indef), personal pronouns (PRO:pers), and possessive pronouns (PRO:poss). Additionally, the tagset includes labels for non-textual elements such as emoticons (EMO) and hashtags (HASH), as well as non-linguistic items (NOCAT). Lemmatisation was performed using a large lexicon containing approximately 600,000 entries.

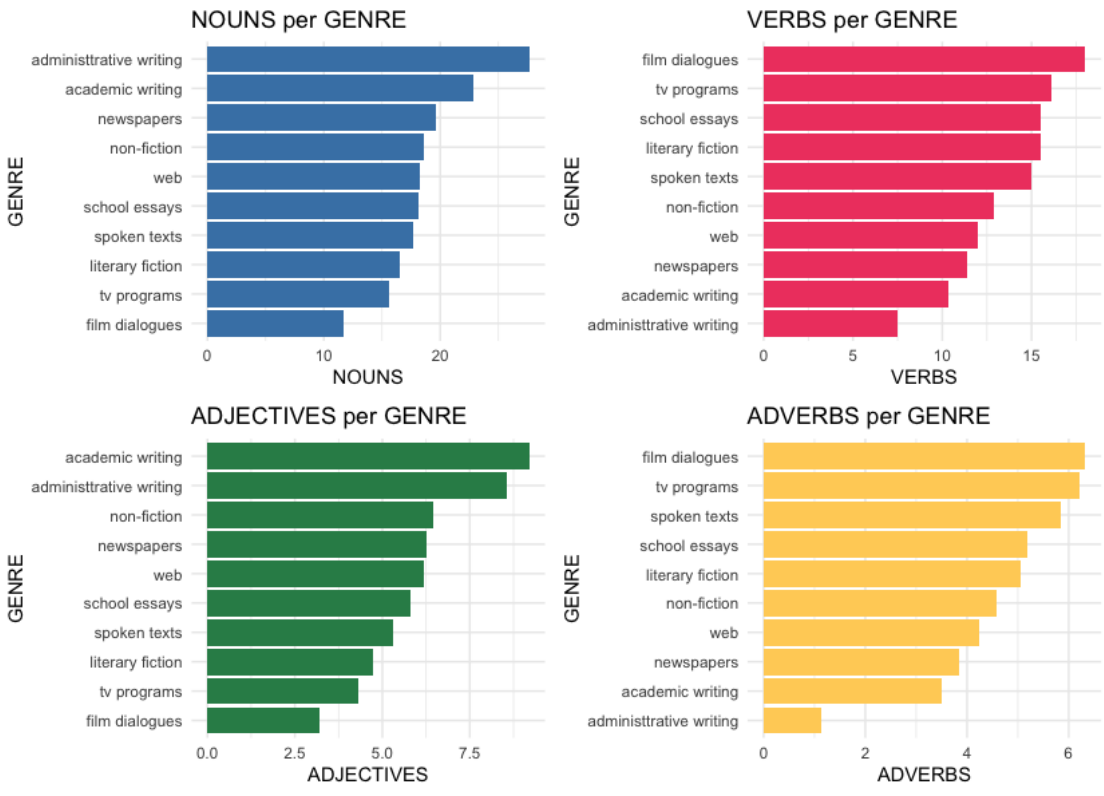
Once POS-tagging and lemmatisation were completed, the data underwent a semi-automatic editing using regular expressions. This process made it possible to identify and correct a significant portion of the errors made by the tagger, particularly those related to forms with high grammatical ambiguity, which the tagger often struggled to assign the correct part of speech to, frequently failing to identify the corresponding lemma. An example is *perché* ('why/because'), which can function both as an interrogative adverb and as a subordinating conjunction.

Additionally, this editing allowed for the identification of lexically ambiguous forms that could correspond to two different lemmas. In such cases, the tagger assigns the correct grammatical category but records both possible lemmas, as in *indica*, which can refer either to the present indicative of *indicare* ('to indicate') or the present subjunctive of *indire* ('to call').

Finally, a further manual correction was carried out for forms that were not recognised by the tagger and classified as *unknown*. These were primarily words not included in the lexicon, such as *chiusosi*, *cyber*, and *remixa*.

As already shown in a previous paper (Spina, 2014), the combination of automatic annotation and lemmatisation with a semi-automatic post-editing phase resulted in a high level of accuracy (98%). Figure 2 shows the distribution of the four major grammatical categories (nouns, verbs, descriptive adjectives and adverbs) across the ten textual genres.

Figure 2. *Distribution of nouns, verbs, descriptive adjectives and adverbs across genres (per 100 tokens)*



4.4. *How to run queries in PEC24*

The PEC24 is currently available and searchable online on a web platform called *Search-it. Italian native and learner corpora* (https://lt.eurac.edu/cqpweb/). It has a publicly accessible CQPweb interface (Hardie, 2012) and can be queried using the Corpus Query Processor (CQP) language. With the CQPweb interface, users can search the corpus using queries of varying complexity.

At a basic search level, CQPweb allows for standard queries, which involve searching for words or sequences of words throughout the entire corpus. The search results are displayed in the form of concordances, where the target word (KWIC – Key Word in Context) appears within concordance lines, showing all its occurrences along with their immediate context. In this format, the query is aligned at the center and is preceded and followed by a portion of surrounding text.

At a more advanced search level, CQPweb allows users to restrict searches to specific sections of the corpus. In the case of PEC24, searches can be refined by *channel* (spoken vs. written), *genre*, which corresponds to the ten main sections of the corpus, *type*, which refers to the subsections of these main sections, and *year*, indicating the year of text production.

By leveraging the CPQ syntax, it is possible to perform more complex queries. First, searches can be carried out both by *lemma* and by *Part-of-Speech* (POS), corresponding to the two levels of annotation in PEC24. The following examples illustrate how to retrieve all forms of the lemma *scrivere* ('to write') and all possessive adjectives within the corpus.

– Lemma search:
[lemma = "scrivere"]

– Part-of-Speech search:
[pos = "DET:poss"]

The CQP syntax also allows for disambiguation between forms. For instance, the word *parto* is ambiguous, as it can refer either to the verb *partire* ('to leave') or to the noun *parto* ('childbirth'). If we want to search specifically for the lemma *partire*, CQP syntax enables us to disambiguate the query as follows:

– Search for the word *parto* when it is verb:
[word = "parto" & pos = "VER*"]

– Search for the word *parto* excluding forms of the verb *partire*:
[word = "parto" & lemma! = "partire"]

Moreover, CQP syntax allows for the exploration of more complex linguistic phenomena, running queries that combine POS and lemma. For instance, we can use the following queries to examine cases where the lemma *stare* ('stay') is followed by gerunds and occurrences of nouns followed by adjectives ending in *–oso*.

– Search for the lemma *stare* followed by gerunds
[lemma = "stare"][pos = "VER:geru"]

– Search for nouns followed by adjectives ending in –oso
[pos = "NOUN"][pos = "ADJ" & lemma = ".*oso"]

CQP syntax proves to be particularly useful in the search of lexical combinations, as it enables the retrieval of word sequences exhibiting specific syntactic patterns. For example,

in the case of Verb + Noun (direct object) combinations, a query could be run to retrieve all collocates of a given verb (in this case, *pubblicare, 'to publish'*) followed by an article and by a noun:

– Search for all the noun following the verb *pubblicare* and preceded by an article:
[lemma="pubblicare" & pos="VER.*"][pos="ART"][pos="NOUN"]

A widely studied type of lexical combination in corpus linguistics is collocations, which often appear in adjacent form – that is, the base is immediately followed by its collocate. However, cases exist in which an adjective or adverb is inserted between the base and the collocate. For instance, we might want to investigate instances where, in the collocation *dare una mano ('give help')*, the verb (*dare*) is modified by an adverb, or the noun (*mano*) is modified by an adjective.

– Search for *dare una mano* with adverbs modifying the verb:
[lemma="dare" & pos="VER.*"][pos="ADV.*"][word="una"][word="mano"]

Figure 3. *Results of the query 'dare una mano' with adverbs modifying the verb*

No , era una mela annurca caduta dal carrello . Che cosa ?	Davo solo una mano	. Come ogni galantuomo dovrebbe , con una signora . Ma quale
Italo capì che l' aveva fatto consapevolmente , convinto di aver	dato così una mano	al trionfo della Legge e dell' ordine . Italo a quel punto
sia presente una dose di farnetico è credenza comune , ma gli	diamo volentieri una mano	, contribuendo perfino , ciascuno per quel che può , all' acquisto
prima ragazza . Ero contenta di averla qui . La sera mi	dava persino una mano	a preparare la tavola . E poi chiacchieravamo un po' , così ,
sorriso di circostanza . « Senti un po' , Steven ... mi puoi	dare ancora una mano	? » . « E certo » rispose il ragazzo versando il
ıdo dei sistemi sociali distribuiti dove le blockchain ci potranno	dare più una mano	, dove forse potrà essere maggiore il loro contributo . Forse più
il tempo sarà mio fratello , e , come lui , mi	darà sempre una mano	, mi darà tempo per andare lontano . E come Ulisse cercherò
neeship ehr e insomma dato che parlo e scrivo bene l' inglese le	do sempre una mano	e però di recente mi ha detto che cioè due o tre giorni
le altre persone Mhm In montagna ci si saluta sempre ci si	dà sempre una mano	fra tutti Sì certo Ehr la volta peggiore abbiamo dormito in un
mpa . Poi aggiunge sibillino : " Sono comunque disponibile per	dare ancora una mano	, se mi sarà richiesto " . Un' ora dopo , Sergio

– Search for *dare una mano* with adjective modifying the noun:
[lemma="dare" & pos="VER.*"][word="una"][pos="ADJ"][word="mano"]

Figure 4. *Results of the query 'dare una mano' with adjective modifying the noun*

Beh , dottor Pereira , la verità è che Marta mi ha	dato una buona mano	, in parte li ha fatti lei , le idee fondamentali sono sue
Mille anni sono tanti . Ora uno se lo potrebbe dimenticare e	dare una bella mano	di bianco , oppure ripararlo e cercare di farlo durare per altri
Qui la tecnologia , negli ultimi anni , sicuramente ci ha	dato una grossa mano	. Esiste infatti un elemento che accomuna tutte le specie , ma
che il campo , comunque il lavoro sul campo mi ha	dato una grande mano	perché impari anche molto a condividere , impari ad ascoltare , facendo
In tutto questo allora io mi aspetto che la scienza mi	dia una forte mano	per risolvere questo problema . Non è così semplice , in realtà .
me contano simpatia e semplicità , e ovviamente la fama mi ha	dato una grossa mano	. Dopo Damaris ho avuto solo un' altra fidanzata prima di Carolina
gli abiti eleganti o casual - chic mai fuori moda gli	davano una gran mano	. Ma ad aumentare il suo fascino contribuivano in particolar modo le
ario politico . Per onestà occorre tuttavia riconoscere che anche lui ha	dato una bella mano	a creare questo clima . " Per esempio ? " Continua a rappresentare
perdite di passaggi importanti . E che , ancora , avrà	dato una buona mano	al redattore del verbale - nel caso in cui abbia affidato tale
motivo c' è bisogno dell' aiuto di tutti ... chiunque riesca a	dare una piccola mano	. Abbiamo tutti il dovere di dare ciò che abbiamo , con

5. The role of the PEC24 as an Italian reference corpus

The previous paragraphs have described the design and composition of the PEC24, a written and spoken corpus of contemporary Italian, including texts produced from 1982 to 2024. As mentioned above, the corpus, also in this expanded and updated version, aims to address the lack of a 'national' reference corpus, able to offer a broad representation of written and spoken Italian through the inclusion of different textual genres.

Diversification is thus one of the strengths of the PEC24: in line with the choices made for the national corpora of other languages, its ten sections ensure a reasonable representativeness of the most widespread textual typologies that Italian speakers produce and are exposed to. As an example, the written part of the PEC24 includes data from press, literature, education (school and university), administration and digital environments such as social media and blogs, which can be considered the textual genres in which a speaker of Italian would typically interact in his or her L1. The way in which these textual genres are balanced within the corpus makes them «acceptably representative» of contemporary Italian (Ädel, 2020): the aim is to ensure a reasonable level of representativeness, bearing in mind that this is a statistical ideal which applies only imperfectly to languages (Evert, 2006), all the more so since representativeness must be achieved by taking into account «both production, by sampling a wide variety of distinct types of material, and reception, by selecting instances of those types which have a wide distribution» (Aston, Burnard, 1998: 28).

Another strength of PEC24 is the significant presence of the spoken component, which accounts for 17.2% of the entire corpus. Since the collection of spoken data is highly complex and time-consuming, even large and extremely elaborate corpora such as the BNC include lower percentages of spoken data (the BNC spoken section represents 10% of the entire corpus). Such a proportion of both monologic and dialogic data allows the spoken varieties to be assigned a significant place in analyses or descriptions of present-day Italian as a whole, or in comparisons between written and spoken language.
Furthermore, the rather small size of PEC24, which is undoubtedly one of its limitations, can also represent another of its strengths: it allows for a high degree of accuracy in linguistic annotation by combining, in a further stage of POS-tagging, automatic and manual operations to correct known and recurring annotation errors.

Finally, the PEC24 corpus can be easily accessed and searched through a widespread and powerful query system (CQPweb) and a dedicated platform (*Search-it*), where it is freely available together with other well-known native and learner Italian corpora adopting the same query language and interface. The system is particularly suitable for querying POS-tagged data, and for combining different criteria (grammatical categories, lemmas) with regular expressions and fine-tuned/specific metadata selection, in order to perform even very sophisticated searches. This makes the corpus widely useable by any scholar, teacher or student, even without specific computer skills, and is an added value for the international research community.

6. Conclusions: limitations and further development

In this paper we have introduced and described the PEC24, a 47-million-word reference corpus of spoken and written contemporary Italian. PEC24 is the evolution and update of the previous PEC corpus, published in 2014 and used by researchers, teachers and students for the past decade. The original structure of the corpus, divided into 10 sections corresponding to ten textual genres, has been preserved. By prioritising genre variety over corpus size, we have chosen to provide a resource that reasonably represents

the language varieties to which, on average, Italian speakers are most exposed. All sections have been expanded by the addition of new texts (43% more than in the original version), and the spoken sections now account for 17.2% of the entire corpus. Among the other strengths of the PEC24, we have mentioned its linguistic annotation and lemmatisation, which makes corpus searches much more powerful and linguistically oriented. The limitations of the corpus stem mainly from its being a low-cost resource, built outside of a structured and financially supported project, which necessarily favoured the reuse of existing data over the possibility of collecting new ones. As a result, while the various sections of the corpus are reasonably balanced with each other, some are less so internally, being composed of subsections that do not always achieve an ideal balance. Therefore, we once again call upon the scientific community to establish a joint project, including several partner universities, with the aim of creating and making available a large national reference corpus of Italian, carefully balanced between written and spoken varieties and organised into sections with strong internal coherence, according to what already exists for many other languages. In the meantime, PEC24 provides a reasonable alternative, capable of plausibly representing contemporary Italian, for the realisation of reference works such as the *Learner Dictionary of Italian collocations* (DICI-A; Spina *et al.*, under review), which relies on the PEC24 for the identification and extraction of collocations to be considered as dictionary entries.

REFERENCES

Ädel A. (2020), "Corpus Compilation", in Paquot M., Gries S. Th. (eds.), *A Practical Handbook of Corpus Linguistics*, Springer, Cham, pp. 3-24 : https://doi.org/10.1007/978-3-030-46216-1.

Anthony L. (2024), "Breaking new ground – AI-enhanced concordance analysis", in *Reading Concordances in the 21st Century (RC21) Blog*: https://blog.bham.ac.uk/rc21/2024/10/28/laurence-anthony-breaking-new-ground-ai-enhanced-concordance-analysis/.

Aston G., Burnard L. (1998), *The BNC handbook: Exploring the British National Corpus with SARA*, Edinburgh University Press, Edinburgh.

Baker P. (2006), *Using Corpora in Discourse Analysis*, Continuum, London-New York.

Bambini V., Trevisan M. (2012), "EsploraCoLFIS: Un'interfaccia Web per ricerche sul Corpus e Lessico di Frequenza dell'Italiano Scritto", in Ricci I., Bertini C. (eds.), *Quaderni del Laboratorio di Linguistica della Scuola Normale Superiore*, Scuola Normale Superiore, Pisa, XI, pp. 1-16.

Barbera M., Corino E., Onesti C. (eds.) (2007), *Corpora e linguistica in rete*, Guerra Edizioni, Perugia.

Barbera M., Corino E., Marello C., Onesti C. (2022), "Corpora.unito.it", in Cresti E., Moneglia M. (eds.), *Corpora e Studi Linguistici*. Atti del LIV Congresso Internazionale di Studi della Società di Linguistica Italiana, SLI, Officinaventuno, Milano, pp. 199-205: https://doi.org/10.17469/O2106SLI000013.

Baroni M., Bernardini S., Comastri F., Piccioni L., Volpi A., Aston G., Mazzoleni M. (2004), "Introducing the La Repubblica Corpus: A Large, Annotated, TEI(XML)-compliant Corpus of Newspaper Italian", in Lino M. T., Xavier M. F., Ferreira F., Costa R., Silva R. (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, European Language Resources Association (ELRA), Lisbona, pp. 1771-1774: https://aclanthology.org/L04-1128/.

Baroni M., Ueyama M. (2006), "Building general- and special-purpose corpora by web crawling", in *Proceedings of the 13th National Institute for Japanese Language International Symposium: Language Corpora Their Compilation and Application*, National Institute for Japanese Language, Tokyo, pp. 31-40.

Baroni M., Kilgarriff A., Pomikalek J., Rychlý P. (2006), "WebBootCaT: Instant domain-specific corpora to support human translators", in Hansen V., Maegaard B. (eds.), *Proceedings of the 11th Annual Conference of the European Association for Machine Translation*, European Association for Machine Translation, pp. 247-252: https://aclanthology.org/2006.eamt-1.31/.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E. (2009), "The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora", in *Language Resources and Evaluation*, 43, 3, pp. 209-226: https://doi.org/10.1007/s10579-009-9081-4.

Benzitoun C., Debaisieux J. M., Deulofeu H. J. (2016), "Le projet orféo: un corpus d'étude pour le français contemporain", in *Corpus,* 15: https://doi.org/10.4000/corpus.2936.

Bertinetto P. M., Burani C., Laudanna A., Marconi L., Ratti D., Rolando C., Thornton A. M. (2005), *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*: https://linguistica.sns.it/CoLFIS/Home.htm.

Bhreathnach Ú., Měchura M., Ó Cleircín G., Ó Meachair M., Ó Raghallaigh B., Scannell K., Uí Dhonnchadha E. (2024), *Corpas Náisiúnta na Gaeilge – National Corpus of Irish*, DCU: https://www.corpas.ie/en/cng/.

Biffi M. (2010), "Il LIT – Lessico Italiano Televisivo", in Mauroni E., Piotti M. (eds.), *L'italiano televisivo 1976-2006*, Accademia della Crusca, Firenze, pp. 35-70.

BNC Consortium (2007), *The British National Corpus, XML Edition*, Oxford Text Archive: http://hdl.handle.net/20.500.14106/2554.

Bortolini U., Tagliavini C., Zampolli A. (1972), *Lessico di frequenza della lingua italiana contemporanea*, Garzanti, Milano.

Brezina V., Hawtin A., McEnery T. (2021), "The Written British National Corpus 2014 – Design and comparability", in *Text & Talk*, 41*,* 5-6, pp. 595-615: https://doi.org/10.1515/text-2020-0052.

Burnard L., Bauman S. (2014), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Text Encoding Initiative Consortium, Charlottesville.

Cerruti M., Ballarè S. (2021), "ParlaTO: Corpus del parlato di Torino", in *Bollettino dell'Atlante Linguistico Italiano (BALI)*, 44, pp. 171-196.

Cialdini F. (2016). "L'aggiornamento della banca dati LIT e il DIA-LIT", in Alfieri G., Biffi M., Giuliano M., Motta D. (eds.), *Il portale della TV e la TV dei portali.* Atti del Convegno Firenze, Accademia della Crusca, 8 marzo 2013*,* Bonanno Editore, Acireale-Roma, pp. 31-47.

Coccia D. (2023), *«Ah beh, sì beh». I Segnali Discorsivi nella lingua della canzone italiana e nell'insegnamento dell'italiano L2*, Tesi Magistrale UniStraPg.

Cresti E. (2000), *Corpus di Italiano Parlato,* Accademia della Crusca, Firenze.

Cresti E. (2020), "The pragmatic analysis of speech and its illocutionary classification according to the Language into Act Theory", in Izre'él S., Mello H., Panunzi A., Raso T. (eds.), *In search of basic units of spoken language: A corpus-driven approach*, John Benjamins, Amsterdam, pp. 181-219.

Cresti E., Moneglia M. (eds.) (2005), *C-ORAL-ROM. Integrated reference corpora for spoken romance languages*, John Benjamins, Amsterdam.

Cresti E., Gregori L., Moneglia M., Nicolas Martinez C., Panunzi A. (2022), "The LABLITA speech resources", in Cresti E., Moneglia M. (eds.), *Corpora e Studi*

Linguistici. Atti del LIV Congresso Internazionale di Studi della Società di Linguistica Italiana, SLI, Officinaventuno, Milano, pp. 85-108: https://doi.org/10.17469/O2106SLI000005.

De Mauro T. (1980), *Guida all'uso delle parole*, Editori Riuniti, Roma.

De Mauro T., Mancini F., Vedovelli M., Voghera M. (1993), *Lessico di frequenza dell'italiano parlato*, Etaslibri, Milano.

EAGLES (1996), *Preliminary recommendations on corpus typology*. EAGLES Document EAG-TCWG-CTYP/P: https://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Ellis N. C. (2017), "Cognition, corpora, and computing: Triangulating research in usage-based language learning", in *Language Learning*, 67, S1, pp. 40-65: https://doi.org/10.1111/lang.12215.

Erjavec T., Kopp M., Ljubešić N. *et al.* (2024), "ParlaMint II: advancing comparable parliamentary corpora across Europe", in *Lang Resources & Evaluation*, 2024: https://doi.org/10.1007/s10579-024-09798-w.

Evert S. (2006), "How random is a corpus? The library metaphor", in *Zeitschrift für Anglistik und Amerikanistik*, 54, 2, pp. 177-190.

Forti L. (2023), *Corpus Use in Italian Language Pedagogy: Exploring the Effects of Data-driven learning*, Routledge, London-New York: https://doi.org/10.4324/9781003137320.

Francis N. W., Kučera H. (1964), "A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers", Brown University, Providence.

Goria E., Mauri C. (2018), "Il corpus KIParla: Una nuova risorsa per lo studio dell'italiano parlato", in Masini F., Tamburini F. (eds.), *CLUB Working Papers in Linguistics*, Vol. 2, Alma Mater Studiorum Università di Bologna, Bologna, pp. 96-116.

Goslin J., Galluzzi C., Romani C. (2014), "PhonItalia: a phonological lexicon for Italian", in *Behav Res*, 46, pp. 872-886: https://doi.org/10.3758/s13428-013-0400-8.

Grandi N., Ballarè S., Chiusaroli F., Gallina F., Pascoli M., Pistolesi E. (2023a), *Corpus UniverS-Ita*. University of Bologna, Bologna: https://corpora.ficlit.unibo.it/CUSP/.

Grandi N., Ballarè S., Chiusaroli F., Gallina F., Pascoli M., Pistolesi E. (2023b), *Corpus UniverS-Ita-ProUniv,* University of Bologna, Bologna: https://corpora.ficlit.unibo.it/CUSP/.

Grandi N., Ballarè S., Chiusaroli F., Gallina F., Pascoli M., Pistolesi E. (2023c), *Corpus UniverS-Ita-ProGior,* University of Bologna, Bologna: https://corpora.ficlit.unibo.it/CUSP/.

Granger S., Gilquin G., Meunier F. (eds.) (2015), *The Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, Cambridge: https://doi.org/10.1017/CBO9781139649414.

Hardie A. (2012), "CQPweb – combining power, flexibility and usability in a corpus analysis tool", in *International Journal of Corpus Linguistics*, 17, 3, pp. 380-409: https://doi.org/10.1075/ijcl.17.3.04har.

Hardie A. (2014), "Modest XML for Corpora: Not a standard, but a suggestion", in *ICAME Journal*, 38, pp. 73-103: https://doi.org/10.2478/icame-2014-0004.

ISTAT (1993), *Indagine multiscopo sulle famiglie Anni 1987-1991*. Vol. 7: *Letture, Mass Media e Linguaggio*, ISTAT, Roma.

Jakubíček M., Kilgarriff A., Kovář V., Rychlý P., Suchomel V. (2013), "The TenTen corpus family", in *Abstract book of the 7th international corpus linguistics conference CL2013*, Lancaster, UK, pp. 125-127: https://www.sketchengine.eu/wp-content/%20uploads/The_TenTen_Corpus_2013.pdf.

Jakubíček M., Kovář V., Rychlý P., Suchomel V. (2020), "Current challenges in web corpus building", in Barbaresi A., Bildhauer F., Schäfer R., Stemle E. (eds.),

Proceedings of the 12th Web as Corpus Workshop, European Language Resources Association (ELRA), pp. 1-4: https://aclanthology.org/2020.wac-1.1.pdf.

Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V. (2014), "The Sketch Engine: Ten years on", in *Lexicography ASIALEX*, 1, pp. 7-36: https://doi.org/10.1007/s40607-014-0009-9.

Laippala V., Rönnqvist S., Hellström S., Luotolahti J., Repo L., Salmela A., Skantsi V., Pyysalo S. (2020), "From web crawl to clean register-annotated corpora", in Barbaresi A., Bildhauer F., Schäfer R., Stemle E. (eds.), *Proceedings of the 12th Web as Corpus Workshop*, European Language Resources Association (ELRA), pp. 14-22: https://aclanthology.org/2020.wac-1.3.pdf.

Laudanna A., Thornton A. M., Brown G., Burani C., Marconi L. (1995), "Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente", in Bolasco S., Lebart L., Salem A. (eds.), *III Giornate internazionali di Analisi Statistica dei Dati Testuali*, Volume I, Cisu, Roma, pp. 103-109: https://www.istc.cnr.it/sites/default/files/uploads/jadt95.pdf.

Love R., Dembry C., Hardie A., Brezina V., McEnery T. (2017), "The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations", in *International Journal of Corpus Linguistics*, 22, 3, pp. 319-344: https://doi.org/10.1075/ijcl.22.3.02lov.

Lyding V., Stemle E., Borghetti C., Brunello M., Castagnoli S., Dell'Orletta F., Dittmann H., Lenci A., Pirrelli V. (2014), "The PAISÀ corpus of Italian web texts", in Bildhauer F., Schäfer R. (eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9)@EACL 2014*, European Chapter of the Association for Computational Linguistics (EACL), pp. 36-43: https://aclanthology.org/W14-0406/.

Maraschio N., Antonini A. Bellucci P., Fanfani M., Stefanelli S., Avesani C., Pratesi M. (1997), "Il progetto LIR. I lessici di frequenza dell'italiano radiofonico", in *Bollettino d'informazioni*, VII, 1-2, pp. 53-94.

Maraschio N., Stefanelli S., Buccioni S., Biffi M. (2004), "Dal corpus LIR: Prove e confronti lessicali", in Albano Leoni F., Cutugno F., Pettorino M., Savy R. (eds.), *Il parlato italiano.* Atti del Convegno nazionale di Napoli (13-15 febbraio 2003), M. D'Auria, Napoli, pp. 1-36.

Mauri C., Ballarè S., Goria E., Cerruti M., Suriano F. (2019), "KIParla corpus: A new resource for spoken Italian", in Bernardi R., Navigli R., Semeraro G. (eds.), *Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it 2019)* (Vol. 2481), CEUR-WS.org: https://dblp.org/rec/conf/clic-it/MauriBGCS19.html.

Mauri C., Ballarè S., Zucchini E. (2024a), *Modulo KIPasti*, Università di Bologna, Bologna: https://kiparla.it/kipasti/.

Mauri C., Ballarè S., Zucchini E. (2024b), *Modulo ParlaBO*, Università di Bologna, Bologna: https://kiparla.it/parlabo/.

McEnery T., Hardie A. (2011), *Corpus Linguistics: Method, Theory and Practice*, Cambridge University Press, Cambridge: https://doi.org/10.1017/CBO9780511981395.

McEnery A., Brookes G. (2024), "Corpus Linguistics and the Social Sciences", in *Corpus Linguistics and Linguistic Theory*, 20, 3, pp. 591-613: https://doi.org/10.1515/cllt-2024-0036.

Orrù P. (2017), *Il discorso sulle migrazioni nell'Italia contemporanea. Un'analisi linguistico-discorsiva sulla stampa (2000-2010)*, FrancoAngeli, Milano.

Paquot M., Gries S. Th. (eds.) (2020), *A Practical Handbook of Corpus Linguistics*, Springer International Publishing, Cham: https://doi.org/10.1007/978-3-030-46216-1.

Pérez-Paredes P., Alcaraz-Calero J. M. (2009), "Developing annotation solutions for online Data Driven Learning", *ReCALL*, 21, 1, pp. 55-75: https://doi.org/10.1017/S0958344009000093.

Przepiórkowski A., Górski R., Łaziński M., Pęzik P. (2009), "Recent Developments in the National Corpus of Polish", in Levická J., Garabík R. (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar Research,* Proceedings of the Fifth International Conference, Slovko 2009, Smolenice, Slovakia, 25-27 November 2009, Tribun, Bratislava, pp. 302-309: http://korpus.juls.savba.sk/~slovko/2009/Proceedings_Slovko_2009.pdf.

Real Academia Española (1994), *Corpus de referencia del español actual (CREA)*, Real Academia Española: https://www.rae.es/banco-de-datos/crea.

Real Academia Española (2013), *Corpus del Español del Siglo XXI*, Real Academia Española: https://www.rae.es/banco-de-datos/corpes-xxi.

Rossini Favretti R., Tamburini F., De Santis C. (2002), "A corpus of written Italian: a defined and a dynamic model", in Wilson A., Rayson P., McEnery T. (eds.), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Lincom-Europa, Munich.

Rundell M., Stock P. (1992), "The Corpus Revolution.", in *English Today*, 8, 3, pp. 21-32: https://doi.org/10.1017/S0266078400006520.

Rychlý P. (2007), "Manatee/Bonito – A modular corpus manager", in Sojka P., Horák A. (eds.), *Proceedings of the First Workshop on Recent Advances in Slavonic Natural Language Processing* (RASLAN 2007), Masaryk University, Brno, pp. 65-70: https://nlp.fi.muni.cz/raslan/raslan07.pdf.

Salesky E., Wiesner M., Bremerman J., Cattoni R., Negri M., Turchi M., Oard D. W., Post M. (2021), "The multilingual tedx corpus for speech recognition and translation": https://doi.org/10.48550/arXiv.2102.01757.

Savchuk S. O., Arkhangelskiy T., Bonch-Osmolovskaya A. A., Donina O. V., Kuznetsova Y. N., Lyashevskaya O. N., Orekhov B. V., Podryachikova M. V. (2024), "Russian National Corpus 2.0: New opportunities and development prospects", in *Voprosy Jazykoznanija*, 2, pp. 7-34: https://doi.org/10.31857/0373-658X.2024.2.7-34.

Savy R., Cutugno F. (2009), "CLIPS: diatopic, diamesic and diaphasic variations of spoken Italian", in Mahlberg M., González-Díaz V., Smith C., *Online Proceedings of the 5th Corpus Linguistics Conference*, July 20-23, 2009, University of Liverpool, Liverpool, UK: http://ucrel.lancs.ac.uk/publications/cl2009/.

Schmid H. (1994), "Probabilistic part-of-speech tagging using decision trees", in Proceedings of the International Conference on New Methods in Language Processing, Manchester, U.K.

Siepmann D., Bürgel C., Diwersy S. (2016), "Le Corpus de référence du français contemporain (CRFC), un corpus massif du français largement diversifié par genres", in Neveu F., Bergounioux G., Côté M.-H., Fournier J.-M., Hriba L., Prévost S. (eds.), *SHS Web of Conferences, Volume 27: 5e Congrès Mondial de Linguistique Française*, Tours, France, 4-8 juillet 2016, EDP Sciences: https://doi.org/10.1051/shsconf/20162711002.

Sinclair J. (1991), *Corpus, concordance, collocation*, Oxford University Press, Oxford.

Sobrero A., Tempesta I. (2007), *Definizione delle caratteristiche generali del corpus: informatori, località*, CLIPS project document retrieved at: http://www.clips.unina.it/docs.

Spina S. (2001), *Fare i conti con le parole. Introduzione alla linguistica dei corpora*, Guerra, Perugia.

Spina S. (2005), "Il Corpus di Italiano Televisivo (CiT): struttura e annotazione", in Burr E. (ed.), *Traditione & Innovazione. Il parlato: teoria – corpora – linguistica dei corpora*, Atti del VI Convegno Internazionale della SILFI, Franco Cesati Editore, Firenze, pp. 413-426.

Spina S. (2010), "AIWL: una lista di frequenza dell'italiano accademico", in Bolasco S., Chiari I., Giuliano L. (eds.), *Statistical Analysis of Textual Data, Proceedings of the 10th Conference JADT*, Editrice universitaria LED, Milano, pp. 1317-1325.

Spina S. (2014), "Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione", in Basili R., Lenci A., Magnini B. (eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014*, Pisa University Press, Pisa, pp. 354-359.

Spina S., Forti L., Zanda F. (2020), "Verso un corpus di riferimento dell'italiano parlato dialogico: il modello BNC2014", in *Rivista italiana di dialettologia*, XLIV, 44, pp. 89-106.

Spina S., Fioravanti I., Zanda F., Forti L., Perri D., Gervasi O. (under review), "A multi-method approach to the development of a Learner Dictionary of Collocations: corpus-based measures and human evaluation", in *Corpus linguistics and linguistic theory*.

Stammerjohann H. (1970), "Strukturen der Rede: Beobachtungen an der Umgangssprache von Florenz", in *Studi di Filologia Italiana*, 28, pp. 295-397.

Stefanowitsch A. (2020), *Corpus linguistics: A guide to the methodology*, Textbooks in Language Sciences 7, Language Science Press, Berlin: https://doi.org/10.5281/zenodo.3735822.

Talamo L., Celata C., Bertinetto P. M. (2016), "DerIvaTario: An annotated lexicon of Italian derivatives", in *Word Structure*, 9, 1, pp. 72-102: https://doi.org/10.3366/word.2016.0087.

Tamburini F. (2002), "A dynamic model for reference corpora structure definition", in González Rodríguez M., Suarez Araujo C. P. (eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC2002)*, European Language Resources Association (ELRA), Las Palmas, pp. 1847-1850.

Tamburini F. (2022), "I corpora del FICLIT, Università di Bologna: CORIS/CODIS, BoLC e DiaCORIS", in Cresti E., Moneglia M. (eds.), *Corpora e Studi Linguistici*. Atti del LIV Congresso Internazionale di Studi della Società di Linguistica Italiana, SLI, Officinaventuno, Milano, pp. 189-197: https://doi.org/10.17469/O2106SLI000012.

Tyne H., Spina S. (eds.). *Applying corpora in teaching and learning Romance languages*, John Benjamins, Amsterdam

Voghera M., Iacobini C., Savy R., Cutugno F., De Rosa A., Alfano I. (2014), "VoLIP: A searchable Italian spoken corpus", in Veselovská L., Janebová M. (eds.), *Complex Visibles Out There*, Proceedings of the Olomouc Linguistics Colloquium 2014, Language Use and Linguistic Structure, Palacký University, Olomouc, pp. 627-640.

Wilkinson M., Dumontier, M., Aalbersberg I. *et al.* (2016), "The FAIR guiding principles for scientific data management and stewardship", in *Scientific Data*, 3, 160018: https://doi.org/10.1038/sdata.2016.18.

Wulff S., Baker P. (2020), "Analyzing Concordances", in Paquot M., Gries S. Th. (eds.), *A Practical Handbook of Corpus Linguistic*, Springer, Cham, pp. 161-179: https://doi.org/10.1007/978-3-030-46216-1_8.

Zampolli A. (1991), "Towards reusable linguistic resources", in Kunze J., Reimann D. (eds.), *Proceedings of the Fifth Conference of the European Chapter of the Association for Computational Linguistics (EACL 1991)*, Association for Computational Linguistics: https://aclanthology.org/E91-1001/.