

# The Use of Corpora in Language Education An Overview of the Italian Language Corpora

Moira De Iaco  
Università degli Studi di Bari “Aldo Moro”  
moira.deiaco@uniba.it

## Abstract

This paper supports the power of the use of corpora in language education without giving up on the examination of the critical issues therein. By analysing the different types and functions of corpora, it will put forward the advantages of corpus-based linguistic analysis in foreign language teaching and learning. Furthermore, this contribution will offer a state of the art of Italian language corpora for the purpose of teaching and learning Italian as a foreign language.

## Keywords

language education; corpus-based linguistic analysis; lexicon learning; Italian language corpora

## 1. Introduction

Corpora is used to refer to collections of authentic linguistic material inclusive of written texts and/or transcriptions of oral interactions. Through the analysis of these texts, it is possible to learn how to use the words therein and to get lexicon-level competence in a language directly through real language use, overcoming the idea of a separation between lexicon and grammar. In fact, corpora refer to lexical chunks. These are lexical items and lexical phrases, which favour the acquisition of linguistic patterns in an appropriate context of use, which is a valid support for language learning. Corpus-based analyses are very innovative and important tools in the study of the lexicon of a language.

Foreign language students need to acquire the pragmatic ability, in addition to meta-linguistic and metacognitive skills, to recognise and correctly and fluently use lexical chunks. In the case of native speakers, the lexical chunks are stored in the long-term memory by the unconscious and repeated exposure to the use of the mother tongue. They constitute an internal corpus that has a priming effect which can negatively influence the learning of a foreign language by suggesting to the learner inappropriate linguistic choices in the target language (Hoey 2005). To avoid this problem, it is necessary to learn the foreign language lexicon through an intense and repeated linguistic interaction and by using authentic linguistic material based on the real contexts regarding the use of the words therein.

Therefore corpus-based linguistic analysis is an important tool that can be used:

- To develop the meta-linguistic skills necessary to acquire the lexical peculiarities of a language. This is because the corpora highlight how the words behave in the use of a certain language.

- To facilitate the memorisation of the lexicon of a language thanks to the presentation of it in its constitutive composite, phraseological and idiomatic aspects according to the possible collocations and co-occurrences of the words.

Nevertheless, the data about a language, which can be obtained by researching the corpora, is rough data. This can be used successfully in language learning to solve problems related to the use of words, but it is not enough in isolation to learn the lexicon of a language. The data needs to be contextualised in oral interactions and adapted by the teachers to the specific teaching goals of each language classroom context. Students need to know how to search for information using the interface of the corpus software. They must interpret the results obtained by the linguistic research into the corpus and they must know how to read and contextualise the data they obtain according to their specific learning purposes. To take advantage of the use of corpora, language students need to be trained in the use of them with the support of teachers. Furthermore, they must regulate the use of this tool based on their proficiency, as well as according to the goals that are to be achieved.

## **2. Functions of the different corpora in language teaching and learning**

Corpora provide a description of the real use of words in a language, showing general trends on a statistical basis (Lüdeling, Kytö, 2009; Reppen 2010; Freddi 2014). According to Corino (2014: 233-234, my transl.), corpora are “an observatory equipped to provide a picture of the language authentically used by real speakers and to enjoy the unlimited and full usability of these contexts of use”.

Starting from the real linguistic facts that the corpora show, language teachers and students can get linguistic material for studying the lexicon of a particular language. Teachers can analyse the list on the frequency of words to establish which words of the language students need to learn first. Language teachers can search for a word throughout the corpora to show the students the derived forms. They can use the examples of the contexts of use of the words to teach the different meanings by directly showing the word’s placement in the living language. They can also refer to the authentic linguistic material offered by the corpora to create exercises and tests. Students can use the corpora to answer any lexical or syntactical doubts by observing the semantic nuances assumed by the words in the different contexts in which they occur and to learn with which words it usually co-occurs.

The standard research tools of the corpora allow you to search for words by letter sequences, sometimes replacing the endings with a wildcard to find all occurrences referring to both primitive words and the inflected forms of a verb. The information that one receives in response to the linguistic query of the corpora concerns the frequency of occurrence of the searched words, the concordances of the words with their contexts of use and the co-occurrences, i.e. the other words that statistically, on the basis of the corpus reference, occur together with the searched word.

For each language, it is now possible to find collections of different types of text. A brief distinction between the different types of corpora is useful to understand how important it is to choose the corpora when making the linguistic search according to the specific teaching or study needs.

For instance, a ‘reference corpus for a language’ is a representative sample of a language in its different aspects that offer a general observation of the language since it collects different kinds of texts: written texts, transcriptions of the spoken language, formal and informal text registers, literary texts, and journalistic texts. It permits to get general data such as the most frequent words or lemmas of the reference language, the list of itself and information about its use, the adjectives that co-occur more frequently with a certain noun, the adverbs that usually appear after a certain verb and the prepositions that agree with a certain verb according to the different contexts of use. By consulting this kind of corpus, a language student can measure their lexical competence by verifying, for instance, if he/she knows the words necessary for a daily interaction in the language that he/she is learning.

While ‘a specialised corpus’ only includes texts from a specific sector (e. g. a corpora of medical language), texts of a particular type only (e. g. texts from the spoken language), texts by a particular author or texts from a specific historical period. They are particularly useful for investigating the micro-linguistic aspects of a language, to get the linguistic aspects of a sector lexicon, to deduce the basic lexicon of a language in reference to a certain specialised field.

Then, ‘learner corpora and teacher corpora’ come directly from language education contexts. A learner corpus includes written and/or oral textual material produced by the learners of second or foreign languages. Learner corpora could be used by linguists, teachers, and students. Linguists can use this kind of corpus to study the variety of a language, to detect the real difficulties encountered in the production of second or foreign language learners, and to create a list of the most frequent mistakes among the native speakers of a certain language (Corino 2014: 236, my transl.). Learner corpora can also provide to the expert teachers material that is useful for gathering information for didactic purposes and for preparing exercises and tests. Finally, learner corpora are an important tool for students because they permit them to observe the most common mistakes made by the learners of the language that they are studying, permitting to develop a meta-linguistic competence that allows them to self-correct and to avoid the same mistakes that they became aware of. Instead, the teacher corpora contain the texts used as teaching material by foreign language teachers, i. e. textbooks, various reading texts, transcriptions of oral texts used for training students to listen to during the lessons and previously submitted exercises. They include material to which the student has been exposed that can be reused in the language class or as a basis for creating language textbooks and tests for students.

### **3. The Italian language corpora**

The first Italian language corpus was published in 1971 based on the “Lexicon of Frequency of the Contemporary Italian Language” (Lessico di Frequenza dell’Italiano contemporaneo. LIF). It included 500,000 words taken from novels, theatrical texts, film scripts, journal articles and parts of textbooks. The Italian linguist Tullio De Mauro used this corpus to draw up the list of words for his *Basic Vocabulary of Italian* published in 1980.

The more representative corpora of the Italian language online available are the corpora “Lessico di Frequenza dell’Italiano Scritto” (CoLFIS) (Lexicon of Frequency of Written Italian) and the “Corpus di Italiano Scritto Contemporaneo” (CORIS) (Corpus of Contemporary Written Italian. CoLFIS (<http://esploracolfis.sns.it/EsploraCoLFIS/>) is a lemmatised and annotated corpus of over 3 million words based on the ISTAT data of Italians reading trends. It contains texts from periodicals, newspapers, and various kinds of book.

CORIS (<http://corpora.dslo.unibo.it/TCORIS/>) is a larger corpus that counts 100 million words. It includes mostly journalistic and narrative texts, but also academic and juridical-administrative texts that are representative of contemporary Italian. There is a version of CORIS that is periodically updated in order to monitor the evolution of Italian, called CODIS (<http://corpora.dslo.unibo.it/CODIS/>). It is a dynamic and adaptive corpus that, according to the specific needs of the user or query, allows for the selection of one or more sub-corpora through which to search. It must be said that the CORIS/CODIS and CoLFIS interfaces are not user-friendly, although they are in English and they have the advantage of being rich in information and of being online and available for free. They are not immediately usable and limitless. The query methods are presented in a language that is too technical and it involves the selection of options that is not easily understandable. Nevertheless, to learn how to search through these platforms, teachers and students can use the video tutorials available on the CoLFIS website and the guide included on the CORIS/CODIS website.

The screenshot shows the CoLFIS search interface with several tabs: **Forme**, **Lemmi**, **Concordanze**, **Coricorrenze**, **Lista**, and **Opzioni**. The **Opzioni** tab is active, displaying various search filters:

- Forma:** Input field with an **Azzera** button.
- Numero di lettere:** Minimum and maximum input fields, each with an **Azzera** button.
- Dizionario:** A dropdown menu.
- Lemma:** Input field with an **Azzera** button.
- Parte del discorso:** Dropdown menu with an **Azzera** button.
- Sintagmaticità:** Dropdown menu with an **Azzera** button.
- Sintagmaticità della forma:** Dropdown menu with an **Azzera** button.
- Statistiche:** A section header.
- Risultati per pagina:** Input field set to 10 with a dropdown arrow.

At the bottom of the form are **Cerca** and **Azzera tutto** buttons. Below the form, the text reads: **EsploraCoLFIS** and **Laboratorio di Linguistica (SNS) - CELI**.

CoLFIS Interface. Last access: September 2020

The screenshot shows the CODIS (100Mw) - Corpus query form. It is divided into several sections:

- User Authentication:** A purple box with a message: "CODIS access is now free for research purposes (Please, read the footnote carefully)." and a **Query** input field with a **(Query Language Help)** link.
- Subcorpora selection:** A table with columns for **Subcorpus** and **Size (in Mw)**. The table has 5 columns for sizes: 20, 10, 5, 3, 2, 1.
- Concordance Options:** A pink box with radio buttons for "Reduce to max" (30, 100, 300, 1000 lines) and a dropdown menu set to "Unsorted".
- Collocations:** A yellow box with radio buttons for "Get Collocates?" (NO!, Yes) and a dropdown menu for "Sort using" (Log-Likelihood Ratio, Mutual Information, T-score, Raw frequency).

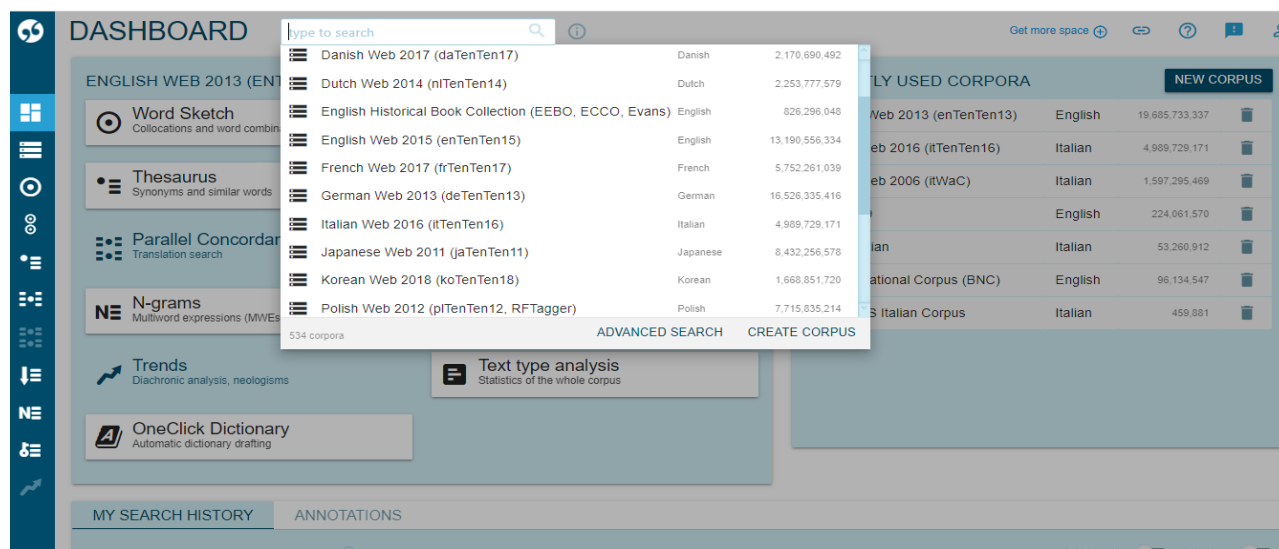
At the bottom are **Esegui** and **Cancella** buttons.

CODIS Interface. Last access: September 2020

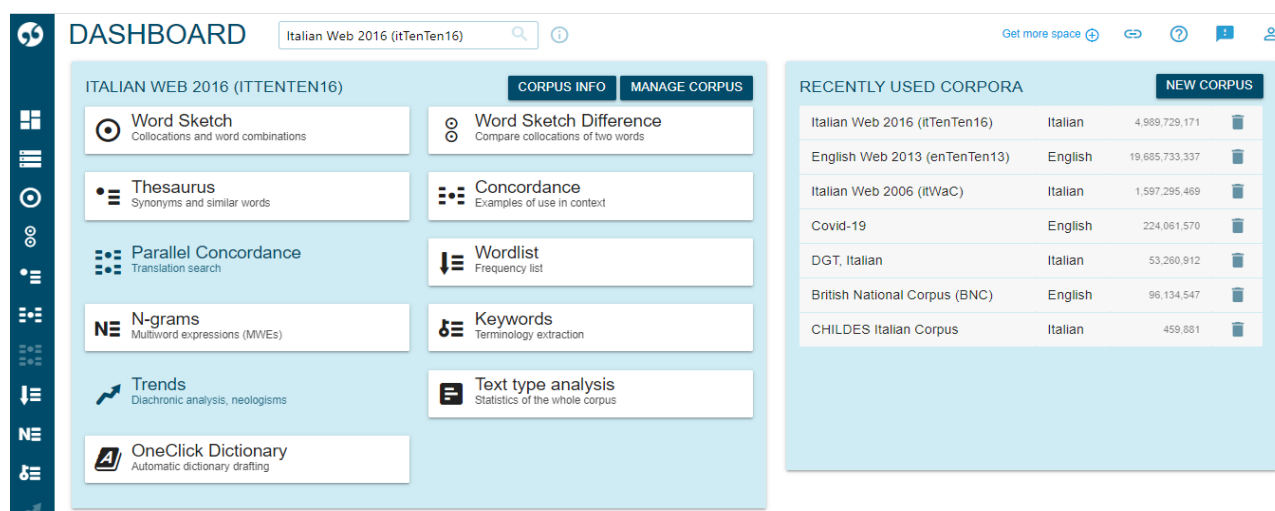
The corpus of the Italian newspaper “La Repubblica” is also considered to be a representative corpus for the Italian language since it was not born as a corpus for Italian because it did not set out to collect different types of texts. It only collects the journalist articles of “La Repubblica.” However, as it includes so many tokens (about 380 million) and it is an annotated corpus, it is an important tool for the study of Italian because it allows for an advanced search through the metadata, lemmas, and parts of the discourse. In addition, it has a user-friendly interface and is easily accessible. It is possible to select it among the corpora available on the free platform “NoSketchEngine” ([https://corpora.dipintra.it/public/run.cgi/first\\_form](https://corpora.dipintra.it/public/run.cgi/first_form)).

Nevertheless, the biggest corpus for Italian is now the Italian Web Corpus (itWaC) (est. 1.5 billion of words) that includes texts automatically collected by the web. It is possible to search for it on the platform “Sketch Engine”, where you can also find metadata like general information, word counts, lexicon size, text type, common tag, and sub-corpora. Sketch Engine (<https://www.sketchengine.eu/>) has made available several different kinds of corpus for several languages. It includes, for instance, spoken language corpora such as the “British Academic Spoken English Corpus” and learner corpora such as the “Arabic Learner Corpus.”

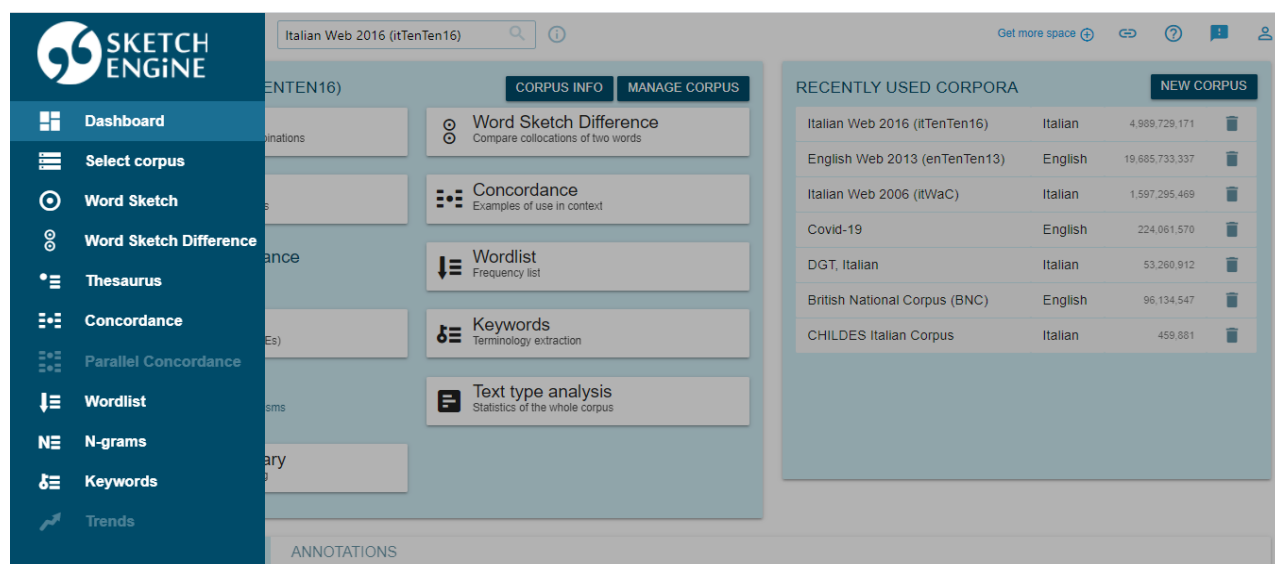
Searching on the Sketch Engine platform it makes possible to get word sketches that show, through examples, the use of the grammatical behaviour of words in terms of collocations and co-occurrences. It also allows for the user to obtain frequency lists of words and to investigate the synonyms by showing the differences in the use of similar words by selecting the thesaurus function. Sketch Engine’s interface is intuitive, and it is easy to understand. After selecting the corpus to query in the drop-down menu at the top, it is necessary to choose one of the different search options in the window. Subtitles which briefly explain the search function of each option are included. The selection of the corpus and the search setup can also be done through the toolbar located on the right of the page.



Selection of the corpus on Sketch Engine. Last access: September 2020.



Search setup on the interface of Sketch Engine. Last access: September 2020.

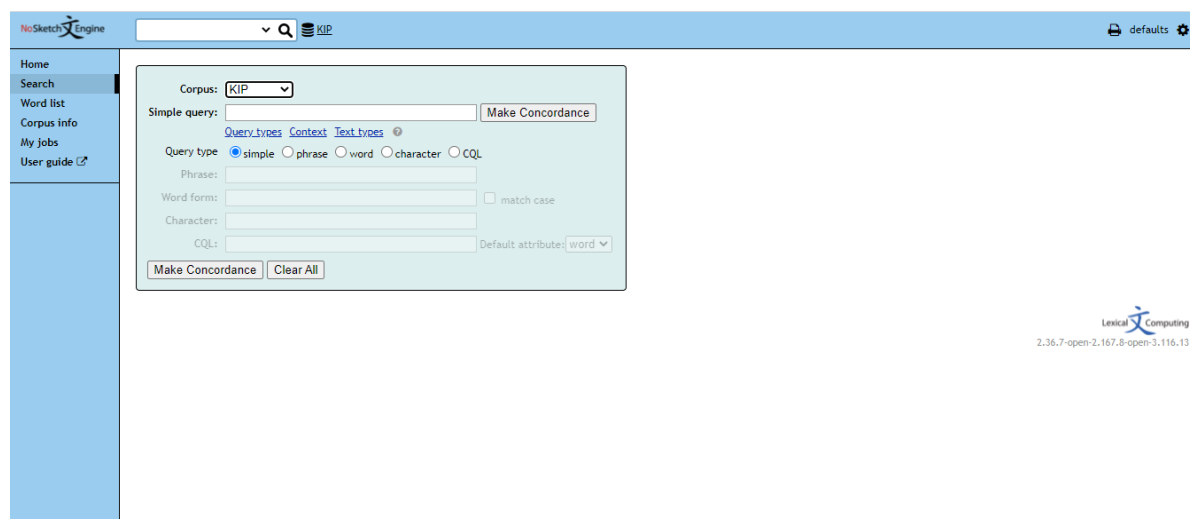


Selection of the corpus and search setup using the retractable toolbar on the right. Last access: September 2020.

The corpus of the “Lexicon of Frequency of Spoken Italian” is available on the open and free access platform BADIP (“Banca Dati dell’Italiano Parlato” 2003-2019, <http://badip.uni-graz.at>). It is a corpus created between 1990 and 1992 by a group of linguists directed by Tullio De Mauro. It is annotated and includes the transcription of recordings from four different Italian cities (Milan, Florence, Rom, and Naples) and has about 500,000 words total. It is one of the most used corpora for linguistic research. It has a user-friendly interface. The setting up of the query is guided by a clickable window that provides simple and clear instructions. The data obtained from the query can be easily exported and it is possible to set up the search by selecting the texts on the basis of their origin (Milan, Florence, Rome or Naples) and by the type of text showing the peculiarities of the different registers of native speakers (Guidetti, Lenchi, Storchi 2012).

The corpus of spoken Italian “Corpora e Lessici dell’Italiano Parlato e Scritto” (CLIPS), based on about 100 hours of speech, is equally divided between male and female voices. It is partly transcribed, segmented, and annotated from a segmental phonetic point of view and it is characterised by a double stratification, specifically the geographical variation and the style and register. The geographical variation was sampled through a preliminary sociolinguistic survey by the University of Lecce that involved the entire national territory. The points of collection of the material are representative both from the point of view of the variety of Italian, as well as from that of the demographic and socio-economic significance of the localities. The selected locations are the cities of Bari, Bergamo, Bologna, Cagliari, Catanzaro, Florence, Genoa, Lecce, Milan, Naples, Palermo, Parma, Perugia, Rome, and Venice. The variation of style and register linked to the variation of the communicative situations of the speakers is represented by the different types of collected materials: radio and television speeches, news, interviews, talk shows, dialogues collected directly during the interactions, spoken readings and telephone speech. On the basis of these variables, the corpus is divided into 5 folders corresponding to the sub-corpora: radio-television, dialogic, reading, telephone and speech. Each sub-corpus is divided into 15 folders corresponding to the 15 localities where the material was collected from. Registration on the site <http://www.clips.unina.it/> is required to access the corpus. Spoken language corpora are particularly important because they permit the observation of the strong linguistic variability that manifests itself in the spoken language (Mcenery, Wilson 1996).

A more recent corpus of spoken Italian is “Corpus KIParla. L’italiano parlato e chi parla italiano” (<http://kiparla.it/>). It collects more than 100 hours of partially structured interviews spontaneous conversations, and university lessons, exams, and talks between students and professors registered in Bologna and Turin. The perspective of the project is to increase the collections points of the material in order to offer an increasingly varied and extensive corpus. It is important to remark that each registered talk was transcribed and the transcripts are aligned with audios. Users also find metadata about each registration. The Corpus KIParla is accessible on the platform NoSketchEngine.



Last access: June 2021

Among the Italian specialised corpora, we find the “Corpus OVI dell’Italiano Antico” (Corpus OVI of Ancient Italian) (<http://gattoweb.oivi.cnr.it/>). It is a corpus that collects ancient Italian texts in the vernacular and it includes about 22 million words. As specified on the website of this corpus, for ancient Italian it means here the Italian of texts dating back to before 1400. It is a corpus that can be consulted for diachronic research on the lexicon and it has not a direct use for learners of Italian as a foreign language, but it could be useful for linguists and also teachers of Italian with lexicographic interests. It is open access and free: registration is not required.

There is also the collection of children’s corpora that provide data that is useful when observing the Italian learnt by the children. The CHILDES Italian Corpus is part of the large collection of CHILDES corpora which includes the corpora of children of different languages. They mostly consist of transcriptions of recordings of spontaneous conversations and they are included on the Sketch Engine platform.

The Corpus “Varietà Apprendimento Lingua Italiana Corpus Online” (VALICO.org <http://www.valico.org/valico.html>) is a portal that offers free and open access to an Italian learner corpus annotated by according to the part of the discourse it is from and the type of text. It collects texts written by the learners of Italian as second language and includes about 570,000 words. It is a tool for language teaching and linguistic research. The querying of the corpus gives the opportunity to get:

- Information about the variations in the writing of learners of different ages and mother tongues.

- Methodological and teaching ideas based on the analysis of the material produced by the students.

- Raw material to be developed as exercises and tests for students of Italian as a second or foreign language.

- Data on the behaviour of words in the contexts of the use of the language and information about the common mistakes made by learners, which is useful for developing meta-linguistic competence.

- An observation of the study of the variations of Italian and the problems of learning Italian as a foreign language.

VALICO.org also contains a paired corpus of texts created by Italian native speakers: “Varietà di Italiano di Nativi Corpus Appaiato” (VINCA). This paired corpus was initially thought of as a control corpus for VALICO but it has become a real support for studies about language teaching and applied teaching (Corino, Marengo 2009; 2017).

On the PAISÀ platform (“Piattaforma per l’Apprendimento dell’Italiano su corpora Annotati”), we can find a fully annotated Italian corpus of authentic texts from the web created in 2010 by Marco Baroni. It is a large corpus (it includes about 250 million tokens) for learners and teachers but as Barbera claims (2013, p. 56, my transl.), it “transcends the language teaching purposes for which it declares itself to be born.” The collected texts can be reused, and the corpus is queried through a very friendly interface that facilitates learners who want to take advantage of the use of this tool.

#### **4. Advantages of the use of corpora in teaching the lexicon of a foreign language**

The importance assumed by words for language learners depends on the specific stage of their language learning and its purpose. Nation (2001) divides the lexicon into four levels: high frequency words, intellectual lexicon, technical lexicon, and low frequency words.



If a teacher of Italian as foreign language intends to know what the most commonly used words are in Italian and therefore which words need to be taught first, the teacher can obtain this data by consulting a corpus like the Italian Web Corpus *itWAC*.

If the type of task proposed by the teacher or the goals which the student intends to achieve focus more on the spoken language, then it is preferable to use corpora like the *CLIPS* or the more recent *KIParla*. These types of corpora allow learners of Italian as a foreign language to know directly from concrete context of conversation regional varieties both from the point of view of expression and from that of pronunciation.

If there are learning needs from a lexical point of view, specialised lexicon learning is required and it is thus possible to use specialised corpora like, for instance, the *EUR-Lex Italian 2/2016 Corpus* available on the *Sketch Engine* platform. It collects European Union legislative documents currently translated into 24 European languages including Italian. Therefore, it is a corpus aligned with the same type corpora in the other 23 European languages and it offers useful data both for the analysis of micro-linguistic aspects of the legal lexicon that a learner of Italian as a foreign language with legal training could be interested in investigating and for translation studies.

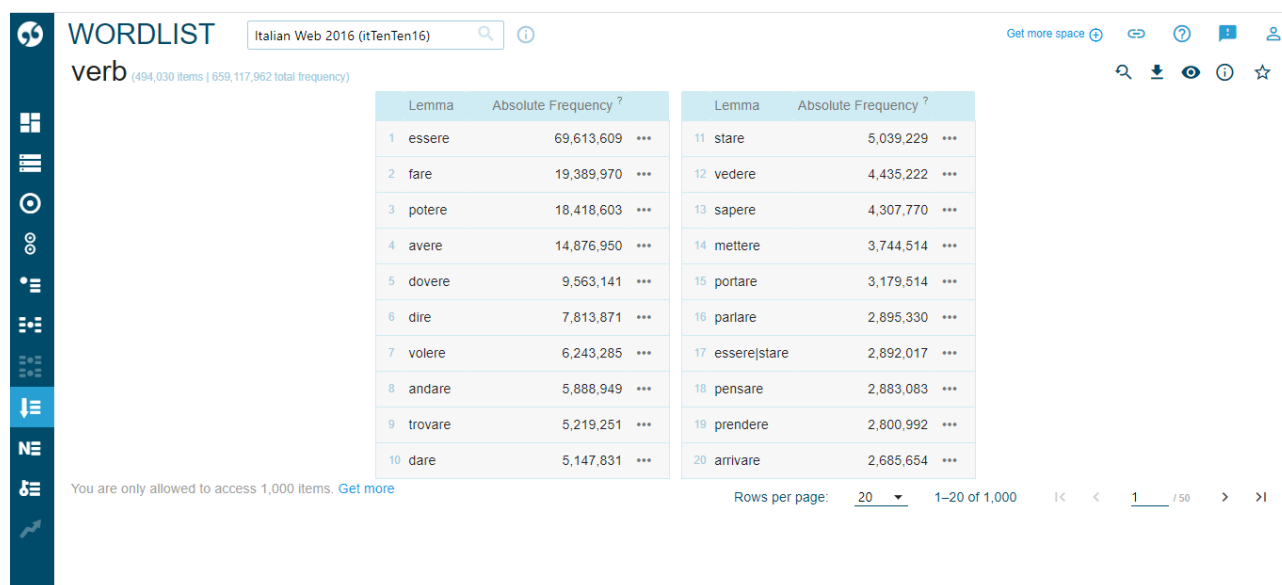
Specialised corpora can also be useful to project the content of the language courses (*CLIL*, Content and Language Integrated Learning) to be developed in relation to the learning of a certain subject in a foreign language. For instance, law in English at the university. The access to the lexicon of a sector through specialised corpora is certainly a useful support for preparing adequate content and material in the lessons, considering the lexical skills necessary to deal with the specific learning.

Furthermore, regarding the learner corpora, according to a study by Corino and Marengo on the *VALICO* texts (2017), it is interesting to highlight the possibilities that these kinds of corpora offer when it comes to studying the errors committed by foreign language learners. Corino and Marengo (2017: 281) describe two experiments. One analyses the errors present in the *VALICO* texts of Spanish-speaking learners produced by Spanish native speakers aspiring to be teachers of Italian and the other one presents the analysis undertaken by Italian students of a foreign language consisting of the most frequent errors present in the *VALICO* texts written by French-speaking, English-speaking, and Spanish-speaking learners. Through the description of the results of the two experiments, Corino and Marengo have shown how learner corpora can be used for creating distractors in multiple choice exercises. Acquiring meta-cognitive competence regarding the most frequent errors of learners is advantageous both for the self-learning of students and for the self-training of teachers. It allows the teachers to produce tests based on real and specific problems of different native speaker learners. We have said before that one of the main types of data that we can derive from the analysis of the corpora is the frequency of use of a certain lemma in various linguistic contexts. When teaching a foreign language, having statistical data on the frequency of the use of words allows you to deduce the important information about which words need to be taught first to allow the learners to acquire a basic vocabulary. Since the data on the frequency of use of words obtained from the corpora will not only be of a quantitative type but also of a qualitative type, you will acquire information such as the degree of polysemy of the high-frequency words. This information may be submitted by the teacher to the students' analysis to develop in them the meta-linguistic competence necessary to use words appropriately in different contexts. The data on the frequency of the use of words collected through the corpora permits the foreign language teacher:

- To deduce important information about which words need to be taught first to allow the learners to acquire a basic vocabulary.
- To get quality information such as the degree of polysemy of the high-frequency words.

- To prepare teaching material that is adequate according to the language proficiency level of the learners and to check the lexical density of the texts to be used in the classroom. In fact, according to the didactic goals, it is necessary to choose the most appropriate type of text since each text includes words with a variable frequency of use. Therefore, it can be more or less adequate according to the purpose that the analysis aims to achieve.

We can put forward an example. If we take into consideration the list of frequency of Italian verbs obtained by searching through the itWAC corpus on Sketch Engine, we can observe that the verb *fare* (to do) is the statistically most frequent verb in the Italian language after the auxiliary *essere* (to be).



WORDLIST Italian Web 2016 (itTenTen16)

verb (494,030 items | 659,117,962 total frequency)

	Lemma	Absolute Frequency ?		Lemma	Absolute Frequency ?
1	essere	69,613,609 ...	11	stare	5,039,229 ...
2	fare	19,389,970 ...	12	vedere	4,435,222 ...
3	potere	18,418,603 ...	13	sapere	4,307,770 ...
4	avere	14,876,950 ...	14	mettere	3,744,514 ...
5	dovere	9,563,141 ...	15	portare	3,179,514 ...
6	dire	7,813,871 ...	16	parlare	2,895,330 ...
7	volere	6,243,285 ...	17	essere stare	2,892,017 ...
8	andare	5,888,949 ...	18	pensare	2,883,083 ...
9	trovare	5,219,251 ...	19	prendere	2,800,992 ...
10	dare	5,147,831 ...	20	arrivare	2,685,654 ...

You are only allowed to access 1,000 items. [Get more](#)

Rows per page: 20 1-20 of 1,000 1 / 50

By continuing the search and selecting other functions, we can get other important information concerning the qualitative aspects of the verb *fare*, including a series of co-occurrences in which the verb *fare* assumes different meanings from the generic “to do” in the meaning of carrying out an action (*fare ginnastica, fare chiarezza, fare compre, fare matematica*) through to the more specific meaning of *costruire, fabbricare* (to build, to manufacture) and idiomatic meanings such as *far quadrare, fare miracoli* and *non fare una piega*. The verb *fare* also performs different functions like replacing a repetition to make the speech more fluent, as in the sentence “*voglio dirglielo, ma non so come fare*” (I want to tell him, but I don't know how to do it) or the causal function: e. g. when *fare* accompanies the verb *ridere* (laugh) to make the sentence *mi fa ridere*, referring to someone or something that provokes the act of laughing. Other examples are *far riflettere, fare conoscere, fare scattare, fare funzionare, fare notare, fare emergere, fare piangere, fare tendenza* etc.

**WORD SKETCH** Italian Web 2016 (itTenTen16) fare as verb 19,389,970x

objects of "fare"	subjects of "fare"	modifiers of "fare"	prepositional phrases with nouns	prepositions after "fare"	pronominal subjects of "fare"
<b>parte</b> fa parte	<b>anno</b> anni fa	<b>si</b> far sì che	"fare" in	<b>per</b> fare per	<b>tu</b>
<b>riferimento</b> fa riferimento	<b>giorno</b> giorni fa	<b>bene</b>	"fare" a	<b>con</b> a che fare con	<b>io</b>
<b>cosa</b>	<b>fine</b> fine ha fatto	<b>male</b> fa male	"fare" per	<b>da</b>	<b>noi</b> noi facciamo
<b>lavoro</b>	<b> mese</b> mese fa	<b>più</b>	"fare" di	<b>di</b> fatto di	<b>lui</b>
<b>conte conto</b> fare i conti con	<b>tempo</b> tempo fa	<b>anche</b>	"fare" con	<b>in</b>	<b>lei</b> lei fa
<b>attenzione</b> fare attenzione	<b>cosa</b>	<b>sempre</b> fa sempre	"fare" da	<b>a</b> fare a	<b>loro</b> loro fanno
<b>passo</b>	<b>settimana</b> settimana fa	<b>solo</b>	"fare" del	<b>senza</b>	<b>voi</b> voi fate
<b>giro</b>	<b>volta</b>	<b>su</b>	"fare" della	<b>ad</b>	<b>egli</b>
<b>domanda</b>	<b>Dio</b>	<b>così</b>	"fare" al	<b>su</b>	<b>essi</b> essi fanno
<b>piacere</b> fa piacere	<b>governo</b>	<b>molto</b>	"fare" nel	<b>durante</b> fatto durante	<b>essa</b> essa fa
<b>differenza</b> fare la differenza	<b>uomo</b>	<b>prima</b>	"fare" dal	<b>dopo</b>	<b>esso</b> esso fa
<b>uso</b>	<b>persona</b>		"fare" ad	<b>attraverso</b>	

pronominal objects of "fare"	adjectives after "fare"	usage patterns	"fare" and/or ...
<b>io</b>	<b>presente</b> fa presente	<b>poter "fare"</b>	<b>dire</b>
<b>la</b>	<b>proprio</b> fatta propria	<b>dover "fare"</b>	<b>fare</b>
<b>l'</b>	<b>politico</b> fare politica	<b>voler "fare"</b>	<b>essere</b>
<b>mi</b> mi fa	<b>salvo</b> fatte salve le	<b>stare per "fare"</b>	<b>andare</b>
<b>le</b>	<b>felice</b>		<b>avere</b>
<b>gli</b>	<b>breve</b> Per farla breve		<b>vedere</b>
<b>ti</b>	<b>grande</b> fatto grande		<b>dare</b>
<b>li</b>	<b>vivo</b>		<b>pensare</b>
<b>loro</b>	<b>freddo</b> fa freddo		<b>prendere</b>
<b>si</b> Si fa	<b>bello</b>		<b>disfare</b> fare e disfare
<b>vi</b>	<b>caldo</b> fa caldo		<b>mettere</b>
<b>se</b> Se fai	<b>franco</b> di farla franca		<b>cercare</b>

CONCORDANCE Italian Web 2016 (itTenTen16) Get more space

cql [lempos=="Tare-v"] 19,389,970 (3.308.33 per million)

Details Left context KWIC Right context

1	cavallieri.it	lo... </s><s> Il Residence "Mini House" offre le bellezze di Roma senza	farà	rimpiangere le comodità e l'indipendenza di casa vostra... </s><s> Dalla
2	liberliber.it	azione dei redditi? </s><s> Puoi aiutarci anche con il 5 per mille. Non ti	fa	pagare più tasse, ma fa sì che una piccola parte dei tuoi soldi venga usa
3	toscana-mare.it	dente, nota con il nome di Granducato di Toscana. </s><s> Da allora ha	fatto	parte del Regno di Sardegna, del Regno d'Italia ed oggi della Republic
4	libero.it	nell'aula scolastica severa, passa la mano sui libri ruvidi e grandi che gli	faranno	compagnia per, cinque anni di grammatica, due di retorica e due di filost
5	libero.it	itirà presto parlare di questo santo plebeo, e sulla strada da lui tracciata	farà	un lungo cammino. </s><s> Un ciuffo di capelli per tracciare una strada
6	libero.it	'erma Joseph Lortz – un dilleggio del comandamento cristiano, e spesso	fanno	apparire la professione cattolica un'ipocrisia ". </s><s> Ma proprio menti
7	libero.it	' Paoli. </s><s> Nei 18 mesi che passa in seminario, Jean Baptiste può	farsi	un quadro completo della vita che conducono i ragazzi del popolo. </s>·
8	libero.it	ri e direttivi. </s><s> Una severa lezione che temprò il suo carattere e lo	fa	diventare "adulto" in brevissimo tempo. </s><s> C'è il problema del sar
9	libero.it	evi cercarlo nella tua famiglia. </s><s> Ma se ti accorgerai che possono	fare	senza di te, allora Dio continua a chiamarti per la strada del sacerdozio.
10	libero.it	o del suo vescovo, Jean Baptiste De La Salle è sacerdote. </s><s> Si è	fatto	un uomo alto e slanciato, Jean Baptiste. </s><s> Ha la fronte spaziosa,
11	libero.it	> Nei salotti della nobiltà si sussurrava anche delle sue stranezze: si era	fatta	scolpire una statua-manichino dalle perfette proporzioni del suo corpo, è
12	libero.it	ere ne ebbe pietà. </s><s> Senza che la padrona ne sapesse niente, lo	fece	riposare sulla paglia asciutta, in un angolo della scuderia. </s><s> Durai
13	libero.it	lla la mente da quel giorno: " il mendicante ha rifiutato la mia elemosina	fatta	con rabbia, e mi ha rimandato dall'altra vita il lenzuolo ". </s><s> Lei ha
14	libero.it	altra vita il lenzuolo ". </s><s> Lei ha respinto Cristo in persona </s><s>	Fa	chiamare padre Barré, un santo religioso conosciuto in tutta la Normand
15	libero.it	uno dei suoi servi. </s><s> Ma nel Vangelo, Cristo ha detto che ciò che	facciamo	ai piccoli, ai miserabili, lo facciamo a lui. </s><s> Respingendo in modo
16	libero.it	Vangelo, Cristo ha detto che ciò che facciamo ai piccoli, ai miserabili, lo	facciamo	a lui. </s><s> Respingendo in modo villano quel mendicante, lei ha resp
17	libero.it	ine e di bambini per le strade, senza scuola, che imparano a rubare e a	fare	il male. </s><s> – Mi presenti un progetto preciso, padre. </s><s> Non t

The information that the corpus-based analyses can provide about collocations are another advantage gained from the language-teaching point of view. Collocations are a widespread phenomenon in the language and they are difficult to frame and convey to students as a precise rule. This is because they often have a paradigmatic in nature and they depend on the use of the language itself. They can take on different forms. Examples include “boarding pass” [noun + noun], “hard-earned money”, “low cost”, [adjective + noun], “save time” [verb + noun] and “a great number of” [article + adjective + noun+ preposition] etc. Corpora allow not only for the viewing of the collocations and the memorising of them in the different contexts of use in which they occur, but also the increased awareness of their frequency.

The concordance lists provide many examples and information about the regular uses of the searched word, string of words or sentences (NATION 2014). They allow for the observation of the tendency of the lexical elements to connect into typical structures such as idiomatic sentences, the meaning of which is difficult to explain through a rule. The occurrence of words in specific sequences has oriented linguists to allow them to describe the language in phraseological terms, so we understand the meaning of some expressions only as they are part of a sentence. If we think, as Guidetti, Lenzi and Storchi (2012) suggested, of the difficulty of grasping the meaning of Italian idiomatic expressions such as *tagliare corto*, *vuotare il sacco*, *alzare il gomito*, the question will arise of how it is possible to teach them without showing their context of use. Language, as Sinclair (1991, 2004) points out, is configured by a set of lexicalised expressions and not as a sum of lexical units separated by grammatical units. The meaning is contained in the entire sentence as the co-occurrences obtained through corpora demonstrate.

## 6. Conclusions

Although recently corpora have been studied and appreciated as a teaching methodology, they are not yet widely used in language teaching. Teachers and students do not use them because they think that the use of corpora requires complex technical knowledge since they are tools created by computational linguistics. Teachers believe that in order to use corpora in the classroom or to teach their students how to use them, it is necessary to acquire and transmit a very specialised competence. This prejudice leads the teachers and learners to give up the advantages that linguistic research through corpora can offer when learning the lexicon of a language. To try to overcome the preliminary obstacle of acquiring technical competence concerning corpora, Zanca (2018) suggests that teachers to introduce corpora to the students directly by using them to solve concrete linguistic problems, thus showing the possibilities that they offer. Zanca proposes first introducing the use of corpora in the classroom through better known tools such as online dictionaries like “Reverso Context.” Zanca then suggests moving on to software, bringing the students into the more technical dimension of using corpora. The transition to the use of software that allows access to the corpora based on authentic and annotated material is an important opportunity to warn students of the risk of approaching the web as a corpus. The material with which one comes into contact on the web often presents with grammatical and spelling errors that a learner, especially one in the early language learning levels, is unable to recognise. Thus, a student could learn the wrong form of the language, replicating the errors in his writing and speaking of the foreign language.

Studies on the use of corpora as a teaching methodology usually distinguish between the indirect and direct use of corpora (MCENERY, HARDIE 2012; ZANCA 2018). We speak of indirect use when teachers and scholars get materials from corpora such as texts and exercises to be used in the classroom. In this regard, we can think of the possibilities concerning the creation of multiple-choice tests getting data from the VALICO texts as proposed by Corino and Marengo (2009). Alternatively, we can obtain useful information for writing language textbooks and for producing warning sections about rules and common errors in the dictionaries. Instead, when students use corpora for learning and studying aspects of a language or to check the correctness of their speaking and writing, we speak of the direct use of corpora. This is an open rather than predetermined use which can be set up by each student according to their personal and contingent language learning or language use needs. However, the direct use of the corpora by learners implies that the students are taught how to use the corpora and that they are taught the technical aspects in order to allow them to be able to explore the resources adequately. It is also required by the learners that there is a medium level of proficiency when correctly interpreting the data obtained from the research through the corpora, allowing for the appropriate reuse of the results. However, corpora can also be used in the classroom by the teachers to extract language patterns from concordance lists or frequency lists in order to observe, confirm or deduce rules, to make hypotheses on the behaviour of words, and to draw conclusions from the observed facts of the language. In this way, the teacher can assume the role of a “learning facilitator” transforming the language lessons in the laboratory and making it so then the language learning is more motivating (CORINO 2014: 236).

The advantages of the use of corpora in language teaching and the learning of the lexicon of a foreign language are not negligible. Nevertheless, this use is not devoid of limits of which language teachers and students must be aware of in order to start and continue to use the corpora optimally. It should be taken into consideration that the material obtained by querying the corpora from the web can include errors that certainly do not benefit students. They need to be warned of this possibility. Instead, if the corpora have a more controlled origin, the teachers should provide the students with a guide to allow them to perform effective functions in their language learning. Teachers should try to train their students in the use of corpora so then they can learn to use them independently, aware of the power of the tool but also aware of their possible limits. In short, corpora are not enough by themselves to learn the lexicon of a language, but they are certainly useful tools for a communicative didactic approach because they offer a teaching methodology that is able to encourage the learning of the lexicon directly in the socio-pragmatic contexts in which it appears, creates, recreates and changes.

## References

- Barbera M., 2013, "Linguistica dei corpora e linguistica dei corpora italiana: Un'introduzione". Online available: [http://www.bmanuel.org/man/Barbera\\_IntroduzioneCL\\_2013=Ver1-60.pdf](http://www.bmanuel.org/man/Barbera_IntroduzioneCL_2013=Ver1-60.pdf), 2013.
- Freddi M., 2014, *Linguistica dei corpora*, Carocci, Roma.
- Hoey M., 2005, *Lexical priming. A new theory of words and language*, London, Routledge.
- Corino E., 2014, "Didattica delle lingue corpus-based", in *EL.LE*, v. 3, n. 2, pp. 231-258.
- Corino E., Marellò, C., 2009, "Didattica con i corpora di italiano per stranieri", in *Italiano Lingua Due*, n. 1.
- Corino E., Marellò, C., 2017, *Italiano di stranieri. I corpora VALICO e VINCA*, Perugia, Guerra.
- Gandin S., 2009, "Linguistica dei corpora e traduzione: definizioni, criteri di compilazione e implicazioni di ricerca dei corpora paralleli", in *AnnalSS*, n. 5, pp. 133-152.
- Guidetti M. G., Lenzi G., Storchi S., 2012, "Potenzialità e limiti dell'uso dei corpora linguistici per la didattica dell'Italiano LS", in *Supplemento alla rivista EL.LE*, <https://www.italy.it/>.
- Lüdeling A., M. Kytö (Eds.), 2008-2009, *Corpus linguistics. An international handbook*. Berlin, Mouton de Gruyter.
- Mcenery T., Wilson A., 2001, *Corpus Linguistics. An Introduction*, Edinburgh, Edinburgh University Press.
- Mcenery T., Hardie A., 2012, *Corpus Linguistics: Method, Theory and Practice*. Cambridge, Cambridge University Press.
- Nation P., 2012, "What do you need to know to learn a foreign language?", in *School of Linguistics and Applied Language Studies*, Victoria University of Wellington, New Zealand.

Reppen R., 2010, *Using corpora in the language classroom*, Cambridge, CUP.

Sinclair J., 1987, *Collins COBUILD Dictionary*, London, Collins Publishers.

Sinclair J., 1991, *Corpus Concordance Collocation*, Oxford, OUP.

Sinclair J. (Ed.), 2004, *How to Use Corpora in Language Teaching*, Amsterdam, John Benjamins.

Zanca C., 2018, "Corpora, Google e roba simile. Per quale ragione gli studenti di una lingua straniera dovrebbero perderci tempo?", in *InTRAlinea*. Online available: <http://www.intralinea.org/specials/article/2296>.