

Open Corpus Interface for Italian Language Learning

Verena Lyding^a, Claudia Borghetti^b, Henrik Dittmann^c,
Lionel Nicolas^a, Egon Stemle^a

^aEURAC research, ^bUniversità di Bologna, ^cInstitut Jules Bordet (^{ab}Italy) (^cBelgium)
verena.lyding@eurac.edu, claudia.borghetti@unibo.it, henrik.dittmann@gmx.de,
lionel.nicolas@eurac.edu, egon.stemle@eurac.edu

Abstract

In this article, we present the multi-faceted interface to the open PAISÀ corpus of Italian. Created within the project PAISÀ (Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati) [1], the corpus is designed to be freely available for non-commercial processing, usage and distribution by the public. Hence, this automatically annotated corpus (for lemma, part-of-speech and dependency information) is exclusively composed of documents licensed under Creative Commons (CC) licenses [2]. The dedicated corpus interface is designed to provide flexible, powerful, and easy-to-use modes of corpus access, with the objective to support language learning, language practicing and linguistic analyses.

We present in detail the interface's functionalities and discuss the underlying design decisions. We introduce the four principal components of the interface, describe supported display formats and present two specific features added to increase the interface's relevance for language learning.

The main search components are (1) a basic search that adopts a "Google-style" search box, (2) an advanced search that provides elaborated graphical search options, and (3) a search that makes use of the powerful CQP query language of the Open Corpus Workbench [3]. In addition, (4) a filter interface for retrieving full-text corpus documents based on keyword searches is available. It is likewise providing the means for building temporary sub-corpora for specific topics. Users can choose among different display formats for the search results. Besides the established KWIC (KeyWord In Context) and full sentence views, graphical representations of the dependency relation information as well as keyword distributions are available. These dynamic displays are based on a visualisation for dependency graphs [4] and one for Word Clouds [5], which build on latest developments in information visualisation for language data. Two special features for novice learners are integrated into each search component. The first feature is a function for restricting search results to sentences of limited complexity. Search results are automatically filtered based on formal text characteristics such as sentence length, vocabulary, etc. The second is the supply of pre-defined search queries for linguistic constructions such as sentences in passive voice, questions, etc.

Finally, we show how the PAISÀ interface can be employed in different language teaching tasks. In particular, we present a complete unit of work aimed at learners of Italian (CEFR level A2/B1) and centered on students' direct use of the interface and its functionalities. By doing so, we are giving concrete examples for targeted searches and interactions with the provided language material, as well as an exemplification of how the use of the corpus can be integrated with communicative language activities in the classroom.

1. Introduction

Since the 90s, corpora have gained increasing interest as a resource for language learning and teaching (cf. e.g. [6], [7], [8]). Ever since, their uses have been researched and strategies and approaches for employing corpora in language education have been developed. Yet, for corpora to become a standard and universal tool in language teaching three conditions still have to be fully met: (1) the availability of large and free corpora, (2) easy to use tools and interfaces for analysis and (3) teacher training with respect to the use of corpora in language education.

The PAISÀ project [1] aimed at providing a free resource and interface for language learning and cultural education based on authentic texts in Italian, thus responding to demands (1) and (2). First, the PAISÀ corpus provides a freely available large-scale corpus of Italian. It is composed of more than 380.000 web documents of contemporary Italian, downloaded in 2010. All documents are licensed under CC, hence free for use and sharing-alike by the public community. Overall, the corpus contains about 250 million tokens and is annotated with lemma, part-of-speech and dependency information. Second, a multi-faceted online interface (see section 2) provides different modes for searching and retrieving the corpus data, including visualisation and download facilities. Third, this open resource

can serve as basis for developing best practices for teacher training. In section 3, we provide an example for educational corpus use by presenting an entire teaching unit for A2/B1 learners of Italian.

2. The open corpus interface

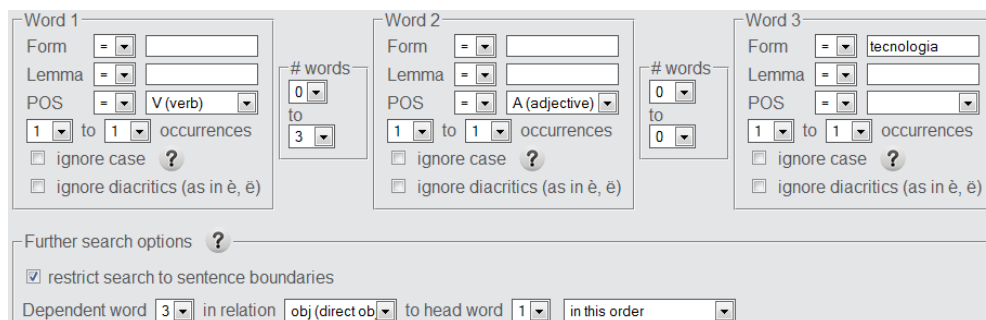
The interface to the PAISÀ corpus is designed to serve a variety of user groups, including language learners and teachers, language researchers and the interested public. Different search modes as well as special display formats and features for supporting learners are provided.

2.1 Search modes

Users can choose among three search interfaces that differ both in terms of usability (e.g. graphical vs. command line) and search expressiveness. However, they all provide a uniform set of display options and access the same underlying corpus data. A fourth search allows for the filtered retrieval of texts and the building of sub-corpora on specific topics based on keyword searches.

The **basic search** aims at simplicity. It takes a plain search box as model, as, for example, known from Google Search and found all over the internet. By simply entering search words into a text box and hitting the "submit" button all lines of text matching the search are retrieved. Multiword expressions can be searched by enclosing words in quotation marks (e.g. "lingua madre").

The **advanced search** is designed for the retrieval of complex linguistic structures by means of an easy-to-use graphical interface. Text boxes allow to search for sequences of words or lemmas with optional intervening words. Part-of-speech and dependency relation annotations can be selected from drop-down menus. Figure 1 shows a search for "tecnologia" as direct object, with preceding adjective. It also allows to select what annotations to include in the results. Figure 2 shows a concordance for "tecnologia" with annotation of lemmas (in parentheses) and parts-of-speech (in subscript).



The screenshot shows the advanced search interface with three word boxes. Word 1 is set to 'V (verb)', Word 2 to 'A (adjective)', and Word 3 to 'tecnologia'. The search is configured to find 3 occurrences of 'tecnologia' as a direct object (obj) of a verb, with an adjective preceding it. Further search options include 'restrict search to sentence boundaries' (checked) and 'dependent word' set to 3.

Fig. 1 - Advanced search specification for "ADJECTIVE tecnologia" as direct object

Divenne_V (divenire) una_R leggenda_S vivente_A della_{EA} (di) tecnologia_S e_{CC} giocò_V (giocare) una_R parte_S di decisiva_A (decisivo) nell'_{EA} (in) ulteriore_A sviluppo_S della_{EA} (di) tecnologia_S de_{EA} (di) Motore_S (motore) Diesel_{SP} Dominus_{SP} utilizza_V (utilizzare) la_{RD} (il) sua_{AP} (suo) grande_A tecnologia_S per_E costruir_{-V} (costruire) sic_{CC} un_R

Fig. 2 - Concordance of "tecnologia" with lemma and part-of-speech annotations

The **CQP search** (cf. OpenCWB [3]), aimed at linguists trained in formal query languages, emulates a command line for queries in CQP-syntax [9], thus offering extended control over searches.

The **filter interface** operates on the corpus source documents, instead of retrieving concordances. By specifying formal text characteristics (e.g. text length, type-token ratio, top-level domain of the source URL, etc.) and providing a keyword, specific subsets of documents can be retrieved or compiled into sub-corpora. The targeted retrieval of documents can serve the preparation of teaching materials, and the creation of sub-corpora enables comparisons over different groups of text.

2.2 Display options

All three search modes yield a uniform results display that can dynamically be modified in a number of ways. Initially, search results are returned as concordance lines with adjustable size of context (n words to left/right, one sentence). Then the user can choose to inspect the source text document or access the *extended Linguistic Dependency Diagram* [4] for each sentence of the result. The dependency visualisation can interactively be adjusted according to the user's preferences, by e.g.

highlighting or excluding relation types or visually marking selected words (see Figure 3). This functionality is meant to facilitate the understanding and linguistic analysis of dependency relations.

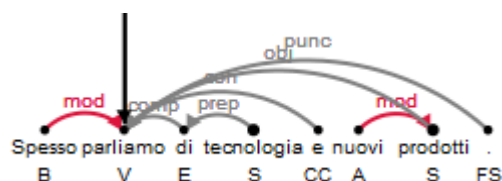


Fig. 3 - Dependency diagram, "modifier" relation in red and POS noun (S) as bigger bullet

In the filter interface, for each subset of documents a Word Cloud [5] illustrating the frequency distribution of co-occurring words is generated (see Figure 5).

2.3 Interface features for learners

The PAISÀ interface offers to restrict searches to sentences of low complexity ("easy sentences") with the aim to provide easy access to sentences that are likely to be understandable also by learners of Italian on an intermediate level. When this option is activated, only those search results are returned that conform to readability criteria, such as a maximum/minimum sentence length, maximum number of words outside a basic vocabulary, and a readability score according to the Gulpease Index [10].

To support novice corpus users in carrying out complex corpus searches, pre-defined search queries for linguistic phenomena, such as passive constructions or question types, are provided with the PAISÀ interface. By simply clicking on the examples a corpus search is started and the respective query is displayed in the search box for inspection or modification by the user.

3. Adopting PAISÀ in the language classroom: An example teaching unit

In this section we present a unit of work for A2/B1 learners, aimed at teenagers. The instructional phases of the unit of work are based on a reworking of Balboni [11] and Zorzi [12], and rely on and benefit from targeted searches and interactions with the PAISÀ interface while adopting a communicative language teaching approach. The suggested unit lasts 6-8 class hours, most of which are centred on students' direct use of the interface and its functionalities [13]. The unit aims at developing students' ability to speak and write about their music preferences and give reasons for such preferences in informal and formal ways, to use Italian superlatives and comparatives, and to distinguish and spell the sounds [k] [tʃ].

3.1 Warm up

In order to stimulate students' motivation to learn, teachers may encourage a preliminary discussion in Italian about music preferences asking questions such as "What music do you like/dislike?", "Who is your favourite singer?", etc. They may also ask students to look for relevant articles, blog posts, pictures, videos, etc. on the web.

3.2 Input presentation

The students are then introduced to the central text of the unit of work (see Figure 4). Also, a short overview of what a (web) corpus is will be needed in order to familiarise learners with PAISÀ. Then, teachers can ask students to repeat on the corpus some of the web searches they previously did, using both basic and advanced search options (e.g., singers such as Justin Bieber and Mariah Carey are mentioned in the corpus), before focussing learners' attention on the chosen text.

ciaooo secondo me justin bieber è un fenomeno....le sue canzoni sn mitike... è hanno tt 1 grande successo...lui è bellissimo...e soprattutto molto bravo..... se nn vi piace xro nn dovete scrivere ke è 1 bambino ke nn sa cantare e ke vuole fare il grande xkè qst dimostra ke siete solo gelosi di lui...xkè nn sarete mai belli e bravi cm lui...ma manko sognando...mettetevelo in testa justin è 1 mito e nn si tokka!!! xkè tt qll ke scrivono ste cavolate devono prima contare fino a 10 e riflettere sulle scemenze ke stanno x dire e poi andarsene a fare in ****.... xkè ce mezzo mondo se nn di più ke adoro qst ragazzo....e qualche cretino cm voi ke scrive ste cose da bambinetti su di lui.....nn avete gusto per la musica... andate ad ascoltarvi le vostre canzoni depresse e lasciate in pace il nostro justin!!!!

Fig. 4 - Resulting text for the multiword search: "Justin Bieber"

This text (see Figure 4), evidently a comment on a blog post, can be exploited for several activities aimed at fostering students' reading comprehension and textual skills. Learners may be asked to:

- Search the corpus for further occurrences of "nn", "x", "ke" and decipher the comment;
- Create a sub-corpus using "nn" or "ke" as keywords and analyse in what typology of texts these abbreviations are used;
- Rewrite the same web message adopting a less informal register and correcting the errors.

Moreover, the transcription of "k" as used in the text may be a good opportunity to practice pronunciation and spelling of the sounds [k] [tʃ].

3.3 Focus

When addressing Italian pragmatic, lexical, and morphosyntactic features with students, the PAISÀ interface can help deepen the discovery of the language used in the selected central text one step further. Expressions such as "è un fenomeno" or "è 1 mito" might for example be searched for forms or lemmas within the corpus, if necessary restricting the searches to sentences of low complexity (see 2.3). Then, students may be asked to group (in a separate file) the results obtained on the basis of the meaning such expressions assume in the related contexts (see Table 1). This activity helps to activate students' reading strategies of global comprehension, while reflecting on registers and text types.

| MEANING 1 – Literal meaning | MEANING 2 – Figurative meaning |
|--|---|
| Il razzismo in Italia è un fenomeno storico complesso | Lettermann è un fenomeno! |
| La forfora è un fenomeno molto diffuso che provoca imbarazzo | Visto che sei un fenomeno, spiegami tu [...] |
| La dispersione scolastica" è un fenomeno che interessa sia i paesi ricchi sia i paesi poveri | Guarda che per fare una cosa così bisogna essere dei fenomeni!!!! |

Table 1 - Grouping of results for "è un fenomeno" (lemma search)

Vocabulary learning can be enhanced through the use of the Word Cloud feature: By entering "mito" or "fenomeno" as keywords, it is possible to visualise the distribution of the words which mostly co-occur with them on the basis of their frequencies (see Figure 5):



Fig. 5 - Word cloud resulting from using "mito" as a keyword

It is thus possible to ask students to group words on the basis of their semantic relations (e.g. "mitologia", "leggenda", and "greca") and explain their choices. This represents an additional opportunity for students to notice the idiomatic meaning of "è un mito" and, thus, its rare use. Within a sub-corpus, students may also look for what relations the most or least frequent words have with "mito".

The central text presents some instances of superlatives and comparatives, and some features of the PAISÀ interface can help focus students' attention on these morph-syntactic features. Learners may be asked, for example, to:

- Look for ".*issimo" (and ".*issim[a|e|i]") using the advanced or the CQP searches, possibly selecting the option "show POS" in order to distinguish adjectival from adverbial uses;
- Make several advanced searches to progressively distinguish relative superlative from comparative of majority (see Table 2 for some examples);
- Explore syntactically some results analysing the related Dependency Diagrams (see 2.2).

All these examples allow students to formulate hypotheses on possible results and test them with further searches and the teacher's help.

| Search pattern | più + ADJECTIVE + di | più + ADJECTIVE + ARTICULATED PREPOSITION | DETERMINATIVE ARTICLE + più + ADJECTIVE | DETERMINATIVE ARTICLE + FROM 0 TO 2 WORDS + più + ADJECTIVE |
|----------------|---|--|--|--|
| | ha un'orbita molto più grande di quella di Plutone | - | - | - |
| | - | la terza città più popolosa del paese | - | la terza città più popolosa del paese |
| | - | la cultura del tempo celebrava l'amicizia come il più importante dei legami | la cultura del tempo celebrava l'amicizia come il più importante dei legami | la cultura del tempo celebrava l'amicizia come il più importante dei legami |

Table 2 - Examples of searches and respective results for comparative and superlative

3.4 Practice

The analysis conducted in the previous phase can be exploited by asking students to write personal comments where they defend their favourite singers from supposed verbal attacks. Alternatively, they might write a role play where one student is the Bieber fan who wrote the central text, the other the attacker. In both cases, learners are encouraged to use the interface to check their hypotheses about idiomatic expressions, discourse markers, etc. and make their texts communicatively effective.

3.5 Reflection and games

To conclude the unit, the teacher can organise some team games which require students to consult the corpus: challenging students to identify (multiple) meanings of given expressions/words or to rank true and fake Italian collocations by exploring the corpus through the advanced search.

4. Conclusions

In this article we showed how the dedicated PAISÀ interface supports the use of corpora in language learning. By providing multiple modes of access and specialised display functionalities and search options, the interface particularly encourages both learners and teachers in the use of corpora for Italian language learning/teaching purposes. Moreover, the unit of work presented demonstrates how PAISÀ supports both forms of 'discovery' (see [14]) and communicative language learning.

References

- [1] www.corpusitaliano.it, co-financed by the Ministero dell'Istruzione, dell'Università e della Ricerca
- [2] <http://creativecommons.org/>
- [3] Evert, S. & Hardie, A. (2011). 'Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium'. In: Proc. of the Corpus Linguistics 2011. Birmingham, UK.
- [4] Culy, C., Lyding, V., & Dittmann, H. (2011). 'xLDD: Extended Linguistic Dependency Diagrams'. In: Proc. of the 15th International Conference on IV, 12-15. July 2011, London, UK.
- [5] <http://code.google.com/p/visapi-gadgets/>
- [6] Barlow, M. (1992) 'Using Concordance Software in Language Teaching and Research'. In: Shinjo, W. et al. (eds.), Proc. of the 2nd Int. Conf. on Foreign Language Education and Technology. Kasugai, Japan: LLAJ & IALL, pp. 365-373
- [7] Sinclair, J. (ed.) (2004) How to use corpora in language teaching. Amsterdam: John Benjamins.
- [8] Aijmer, K. (ed.) (2009) Corpora and language teaching. Amsterdam: John Benjamins.
- [9] Evert, S. and 'The OCWB Development Team' (2010). The IMS Open Corpus Workbench (CWB), CQP Query Language Tutorial, <http://cwb.sourceforge.net/>
- [10] Lucisano, P. e Piemontese, M. E. (1988). 'GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana'. In: Scuola e città, 3, 31, marzo 1988, La Nuova Italia.
- [11] Balboni, P.E. (1994). Didattica dell'Italiano a Stranieri. Roma: Bonacci.
- [12] Zorzi, D. (1995). 'Introduction'. In: Marinolli, G. & M.G. Zanetti (eds.), Tocca a Te. Bolzano: Athesia-Tappeiner, pp. 6-20.



- [13] Leech, G. (1997). 'Teaching and language corpora: A convergence'. In: A. Wichmann et al. (eds.) Teaching and Language Corpora. London: Longman, pp. 1-23.
- [14] Bernardini, S. (2004). 'Corpora in the classroom: An overview and some reflections on future developments'. In: Sinclair, J. (ed.), How to use corpora in language teaching. Amsterdam: John Benjamins, pp. 15-36.