# RIDIRE. Corpus and Tools for the Acquisition of Italian L2

Alessandro Panunzi, Emanuela Cresti, Lorenzo Gregori
University of Florence
alessandro.panunzi@unifi.it, elicresti@unifi.it, lorenzo.gregori@unifi.it

## Abstract

This paper introduces the RIDIRE corpus, built by means of an open source tool (RIDIRE-CPI) for creating specifically designed web corpora through a targeted crawling strategy. The RIDIRE-CPI architecture combines existing open source tools with specifically developed modules, comprising a robust crawler, a user friendly web interface, several conversion and cleaning tools, an anti-duplicate filter, a language guesser, and a PoS-tagger. The RIDIRE corpus is a balanced Italian web corpus (1.5 billion tokens) designed for enhancing the study of Italian as a second language, while also being exploitable for lexicographic purposes. The targeted crawling was performed through content selection, metadata assignment, and validation procedures. These features allowed the construction of a large corpus with a specific design, covering a variety of language usage domains (News, Business, Administration and Legislation, Literature, Fiction, Design, Cookery, Sport, Tourism, Religion, Fine Arts, Cinema, Music). The RIDIRE query system allows research to be carried out on the whole corpus itself or on the sub-corpora. Specifically, available queries comprehend all the functions usually exploited in corpus-based lexicography: frequency lists, concordances and patterns, collocations, Sketches, and Sketch Differences.

**Keywords:** Corpus linguistics; Terminology; Collocations

## 1 Introduction

RIDIRE (acronym for *RIsorse Dinamiche dell'Italiano in Rete*, "Italian Dynamic Resources Online"; Moneglia & Paladini 2010) is a project which produced a large Italian language corpus, and an open-source tool for web corpora building and processing, named RIDIRE-CPI (Panunzi et al. 2012). The corpus - of 1.5 billion tokens - was built using web-crawling techniques and exploited the Italian content of the Internet. The corpus is now available online and is integrated with computational tools for the exploitation of vast corpora to enhance language usage in L2 Italian learners. RIDIRE is designed for use by both teachers and learners, who will be able to profit from access to a database of representative texts which characterize Italian culture. The database collects a massive amount of freely available content, covering a selection of domains which are relevant to Italian identity: law, religion, politics, literature, trade, administration, information, design, food, fashion. To reach this goal, a distributed

crawling infrastructure was created and a targeted crawling strategy pursued. This document will summarize the corpus design for the resource as well as the crawling techniques and processing tools used for deriving language corpora from the web. Also presented are examples of queries that are relevant for both learners and lexicographers.


**Figure 1: The RIDIRE resource home page.**

## 2 Corpus Design Strategy

Different kinds of projects have been carried out to exploit the language data populating the web (Kilgarriff & Greffenstette 2003, Sharoff 2006). Among these, the WaCky initiative (Baroni et al. 2009) and the Italian web corpus ItWaC are important antecedents. More recently a new generation of web corpora have been created and processed with boilerplate cleaning and de-duplication tools and are available through Sketch Engine for a large number of languages (Kilgarriff et al. 2004); these are identified through their target size as the TenTen collection: 10 billion word corpora ($10^{10}$). Such initiatives resulted in the development of dedicated software for crawling (Heritrix), text-processing, cleaning, and the large-scale use of existing technologies for morpho-syntactic annotation (TreeTagger) and online corpus querying (CQPweb). These technologies have been used in RIDIRE and adapted to its specific goals.

The RIDIRE project aimed to build an online database representative of a wide and significant Italian language universe which would have value for sourcing information on the use of Italian in various aspects of life and culture, for linguistic/lexicographic researches, and for didactic purposes. To build such a resource involved two corpus design requirements which did not characterize the web corpora collected in previous initiatives: a) the selection of linguistic resources which document the main domains of usage (life and culture); b) the enrichment of the resource with metadata which enables a perspicuous querying of the database in each specific domain.

The collection focuses on two sets of non-hierarchically structured domains, selected for their pragmatic relevance to the use of the Italian language. The first set is constituted by general non-semantic fields, in which language characterizes its function (up to 400 million words for each domain):

- News
- Business
- Administration and Legislation

The second consists of semantic fields in which Italian excellence is largely recognized (up to 100 million words for each domain):

- Literature
- Fiction
- Design
- Cookery
- Sport
- Tourism
- Religion
- Fine Arts
- Cinema
- Music

The possibility for learners to find specific information on the language usage characterizing a domain should enhance their ability to find the right expressions for it. From a lexicographic point of view, the presence of different domains allows the derivation of specific uses of a word and the description of its semantic variation across the different domains of language use. Table 1 and Figure 2 show the structure of the corpus and the quantitative measures for each domain.

| DOMAINS | # WEBSITES | # PAGES | # TOKENS | # WORDS |
|---|---|---|---|---|
| Functional (total) | 189 | 976,460 | 854,388,230 | 747,268,841 |
| Information | 27 | 550,169 | 216,431,868 | 186,577,769 |
| Economics and Business | 123 | 226,535 | 179,710,476 | 161,377,152 |
| Administration and Law | 39 | 199,756 | 458,245,886 | 399,313,920 |
| Semantic (total) | 816 | 907,374 | 660,243,564 | 566,229,119 |
| Sport | 49 | 138,235 | 98,172,470 | 82,695,548 |
| Architecture and Design | 142 | 136,725 | 93,822,675 | 81,235,939 |
| Cooking | 20 | 123,376 | 52,784,045 | 45,523,096 |
| Cinema | 25 | 122,850 | 51,466,145 | 44,370,692 |
| Music | 195 | 113,015 | 12,906,213 | 106,287,283 |
| Fashion | 103 | 74,584 | 24,645,980 | 21,690,140 |
| Visual Arts | 118 | 70,601 | 56,517,442 | 48,929,903 |
| Religion | 51 | 66,053 | 72,454,492 | 62,291,806 |
| Literature and Theatre | 113 | 61,935 | 85,474,102 | 73,204,712 |
| **Total** | **2,010** | **3,767,668** | **1,514,631,794** | **1,313,497,960** |

**Table 1: Number of crawled websites, pages, tokens and words per domain.**

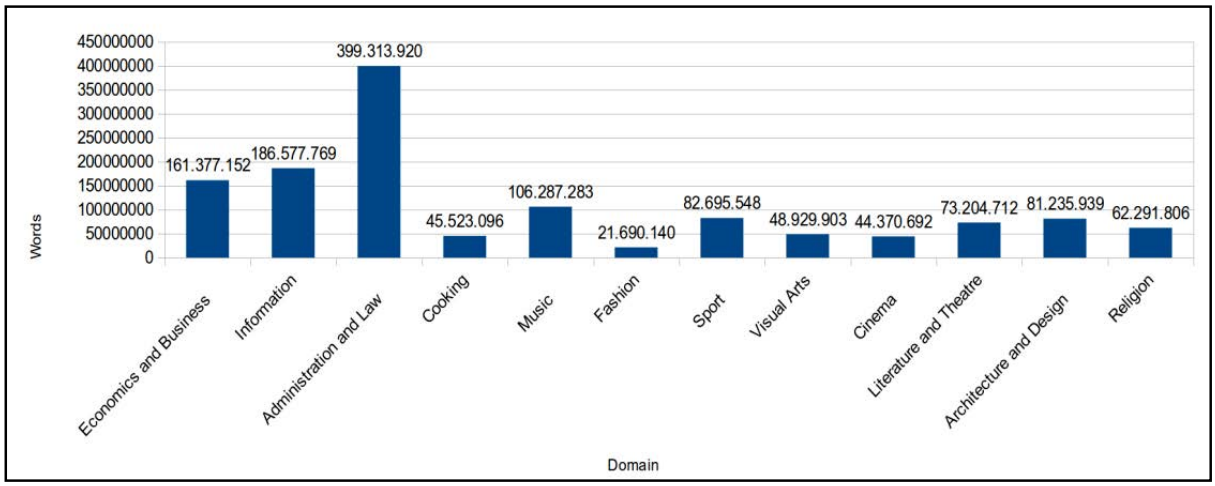



**Figure 2: Words per Domain chart of RIDIRE Corpus.**

## 3 The Crawling Infrastructure

The gathering of specific linguistic data for each sub-corpus requires a targeted crawling strategy performed by different teams of experts. The tool developed within the RIDIRE project for the crawling and the processing of the web resources (RIDIRE-CPI) is now open source and its user-friendly web interface is specifically intended to allow collaboration between users unskilled in web technology and text processing, working in a distributed environment. The application comprises:

- the crawling process
- the mapping of the resource in a MySQL database
- user interaction via web interface

RIDIRE-CPI has a modular architecture (see Figure 3), which is made up of:

- a web crawler
- a web interface for crawling management and validation
- conversion tools
- HTML cleaner tools
- anti-duplicate filters
- a language guesser
- a PoS-tagger

The crawling activity, as in the other cited web corpus initiatives, makes use of the Heritrix web crawler (version 3.1.1). However RIDIRE-CPI configures it via the web interface, making it suitable for use in a distributed environment. The crawling activity itself is structured into "jobs" (fully configured crawling sessions) in which the user determines three sets of parameters. First, the user selects the seed URLs from which the crawling activity starts. Then the formats of the resources that should be

downloaded are specified. In addition to HTML, RIDIRE-CPI is able to process TXT, RTF, DOC, and PDF documents. This feature is crucial, since many linguistically relevant resources from the web are not contained in web pages, but in documents of varying formats. The third set of parameters determine the strategy for the selection of content from websites. This step is important in downloading resources which comply with the representativeness requirement, since the reference unit for text on the web (when representing the language of a particular domain) is the web page rather than the website. As a matter of fact, only a subset of the web pages from a given site give information strictly concerning the specific domain to which the site belongs. Within the step, the user selects and/or discards the "resources" specifying

- which found URLs the crawler has to add to the queue ("URL to be navigated");
- which resources the crawler has to download to the file system ("URL to be saved")

Once all the parameters are defined by the user, the crawler starts from the first seed URL, which is put in the processing queue. The crawler accesses the web page relative to the first URL in the queue, extracts all the links that match the "URL to be navigated" rules and saves them in the queue; then, if the page is a "URL to be saved", the crawler downloads the web page content and stores it on the file system. Finally, it goes back to the first step and proceeds recursively until the processing queue is empty.

To maximize the precision of the process, the user can decide to insert a list of complete URLs, to specify website areas with path substrings (any URL containing one of these strings) or to write a customized regular expression that matches desired page URLs. For instance, in Figure 4 the user decided to crawl the website http://musica.atuttonet.it, getting HTML pages only, and further navigating to any link found (this option is set with a regular expression in the *Pattern* field), downloading any pages that do not contain the word *varie* or *artisti* in the URL.

In this stage no technical competence is required, but a pre-analysis of the website(s) is necessary to ensure only relevant information is retrieved.

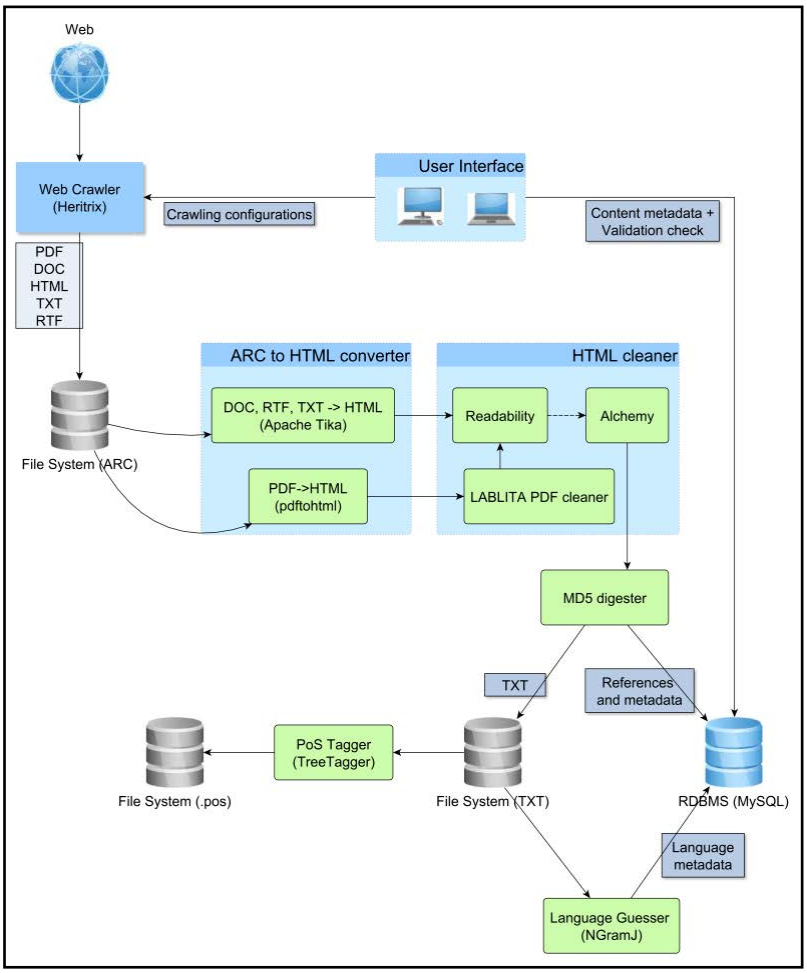



**Figure 3: RIDIRE-CPI Architecture.**

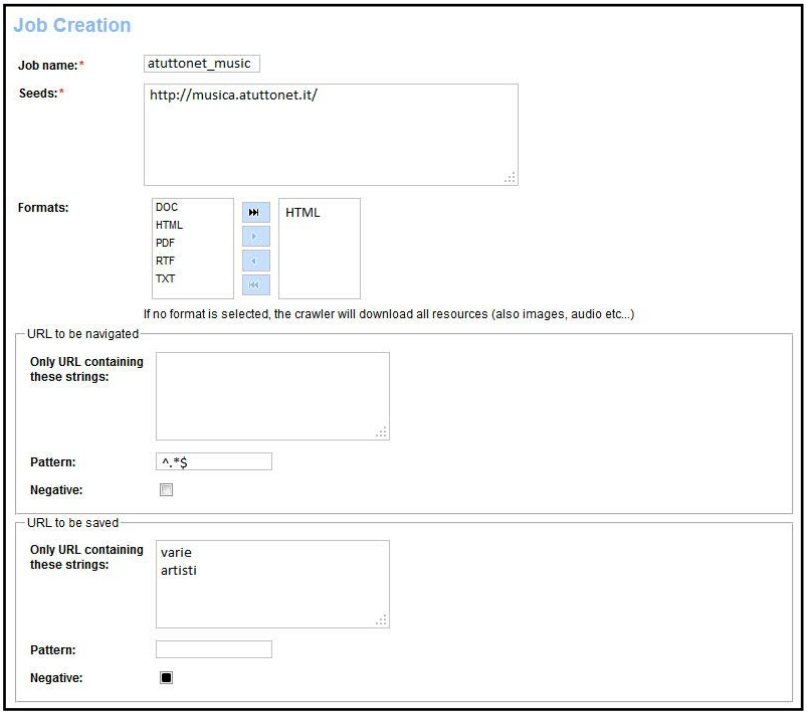



**Figure 4: RIDIRE "Job Creation" page.**

## 3.1 The Mapping Process

To be adequate for linguistic research, the crawled data needs to be processed by a procedure that includes text cleaning, duplicate removal, and PoS-tagging (Baroni et al. 2009). To this end, RIDIRE-CPI uses an automatic processing pipeline on the downloaded resources to extract the running text that will constitute the corpus itself. Web pages, as is well known, contain text that is not relevant for the constitution of a corpus e.g. advertising, navigation menus, disclaimers, credits, etc. (the so called "boilerplate"). Each terminated job is first converted into HTML, which involves several tools depending on the input format. After the conversion, the text cleaning is performed. The boilerplate is removed by means of two external tools freely available for research: Readability and Alchemy API. PDF files are more difficult to clean, so they are treated separately with a dedicated tool - PDF-Cleaner - that performs a deep filtering on the content.
Readability is the first option for the HTML cleaner, but if it won't yield results or outputs an error, the Alchemy API provides a second chance. The plain text documents output from the cleaning stage are then processed by a simple MD5 digester to get their signature, which acts as an anti-duplication system allowing the application to discard resources found with the same signature. The last phase of the mapping procedure is the part-of-speech tagging of the plain text resource. The PoS-tagging is performed by TreeTagger, which is run as an external executable by the main application. TreeTagger creates the PoS-tagged file in the correct file location directly.

## 3.2 Validation and Corpus Creation

RIDIRE-CPI integrates a validation interface dedicated to the evaluation of the crawled resources, which ensures that they belong to the specific domain they should represent. The validation procedure creates a random sample of the resources found and the user can check whether they are adequate with respect to the corpus design or content restrictions. A job can be considered "valid" if it contains non adequate resources under a given percentage (less than 10%, in principle). Since a manual revision is required for a high quality result, but checking the whole corpus is not an option due to its size, the validation process implemented in RIDIRE is a good trade-off between a clean corpus and a fast check. Figure 5 shows how the interface presents a random sampling of one crawled job, allowing direct access to a selection of pages whose adequacy in representing the given domain can be verified.

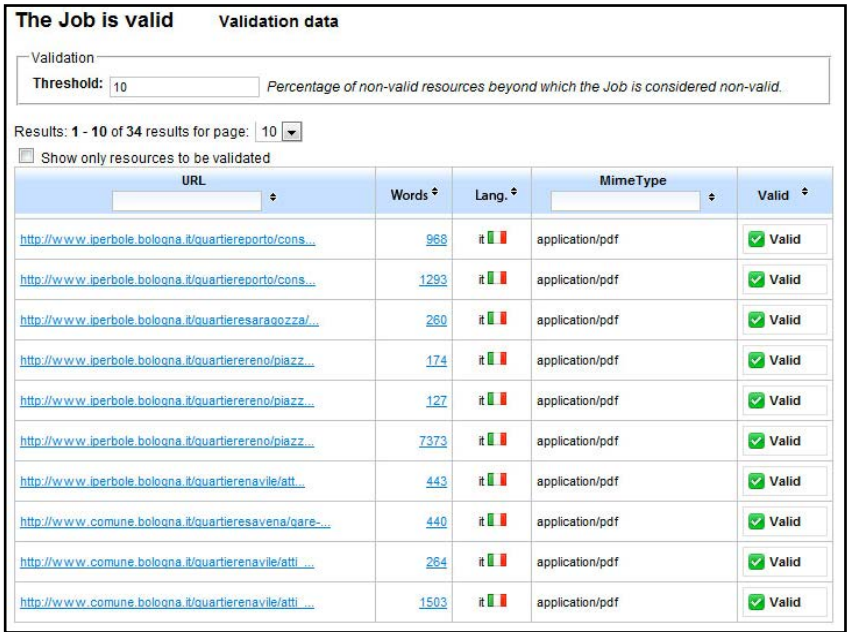



**Figure 5: Validation sampling.**

Through the content selection, metadata assignment, and validation procedures, the RIDIRE-CPI allows the gathering of linguistic data from the web with a supervised strategy that allows a high level of control. The frequency lists of the various domains provide direct evidence that the crawling performed within expectations. The nouns (i.e. the referred entities) that ranked highly identify each domain (Religion, Fashion and Cookery) quite well, and are shown in Table 2.

# 4 Methods for the Extraction of Linguistic Information from Corpora in L2 Acquisition and Lexicography

Various experiences in trying to use corpora for second language acquisition purposes clearly show that both learners and teachers are scared by the complexities of techniques involved in corpus linguistics and that the resultant data is difficult to appreciate (Kilgarriff 2009). Concordances provide a large amount of fragmented information that is difficult to read, especially for second language learners. Despite the fact that corpora contain information that is needed and that the tools are pretty powerful (Sinclair 2004; Conrad 2006), the way to use these tools is undefined and the information retrieved is difficult to interpret, with the overall process being felt as time consuming. The challenge for corpus linguistics in the field of second language acquisition is to provide a simple way to link the actual needs of learners to corpus data.

| Religion | | Fashion | | Cooking | |
|---|---|---|---|---|---|
| **Lemma** | **Freq.** | **Lemma** | **Freq** | **Lemma** | **Freq** |
| vita | 210,420 | collezione | 56,685 | ricetta | 135,498 |
| uomo | 169,995 | moda | 50,381 | iscritto | 104,610 |
| amore | 110,831 | anno | 49,369 | località | 93,692 |
| fede | 100,514 | colore | 32,777 | acqua | 82,492 |
| mondo | 98,913 | abito | 30,085 | farina | 81,695 |
| pagina | 95,462 | mondo | 28,816 | volta | 81,274 |
| parola | 92,532 | donna | 28,657 | pasta | 75,144 |
| cuore | 92,351 | stile | 26,815 | zucchero | 67,609 |
| tempo | 82,891 | linea | 26,026 | minuto | 66,579 |
| giorno | 76,190 | pelle | 20,962 | impasto | 65,074 |
| figlio | 70,231 | capo | 20,619 | forno | 61,672 |
| persona | 69,251 | euro | 19,199 | olio | 59,151 |
| anno | 69,054 | modello | 18,947 | cucina | 56,065 |
| popolo | 66,595 | articolo | 18,747 | gr | 55,079 |
| modo | 65,716 | tempo | 18,307 | burro | 52,101 |
| preghiera | 64,907 | prodotto | 17,365 | uovo | 49,057 |
| cosa | 57,020 | marchio | 16,968 | cosa | 48,276 |
| santo | 52,341 | vita | 16,388 | tempo | 47,712 |
| fratello | 51,370 | accessorio | 16,268 | messaggio | 47,453 |
| famiglia | 51,234 | stilista | 16,254 | parte | 46,829 |

**Table 2: The 20 most frequent nouns, taken from 3 different domains.**

The types of queries available in RIDIRE are inspired by those from the Sketch Engine and are available for both the general corpus and each sub-corpus:

- frequency lists
- concordances and patterns of words (ranked according to raw frequency)
- collocations (general and restricted to specific PoS)
- Sketches and Sketch Differences (between two words or domains) of collocates for the most relevant patterns of a word

The key strategy adopted in RIDIRE is to give a clear picture of the subset of problems that a learner can solve through corpora access, providing each problem area with a predetermined search path which leads to satisfactory results.

An extension of the concordances search function is the pattern search, where a user can view the concordances of a sequence of words (rather than a single one) specified by a form, lemma or PoS attribute; then, grouping the results together, he can see the more frequent usages of the sequence and

what the allowed syntactic structures are. In Figure 6 we searched the occurrences of the Italian verb *sperare* immediately followed by a preposition and we can see that there are five returned sequences (we excluded the rare occurrences): *sperare di* (68.37%), *sperare in* (13.88%), *sperare per* (4.24%), *sperare nel* (3.7%), *sperare nella* (3.26%). In this way a language learner can understand which prepositions may follow *sperare* and how they may be used by scrolling the occurrences list and looking at the different application contexts.


**Figure 6: Pattern search grouped results.**

RIDIRE is furthermore characterized by a set of sub-corpora representing Italian usage in different semantic and functional domains. The way in which a concept can be characterized in a given domain is partly a function of idiosyncratic usage conventions and corpus data can show this to the learner. In language this is reflected in particular by adjectives and adverbs, which show preferential meaning and associations and which vary across language usages. For instance, the variety of objects which are modified by the adjective *forte* ("strong") vary when the context of usage is Religion or Cookery. The learner should wonder whether or not this adjective, learned in general, has specific meaning in a domain when applied to its particulars. Here, RIDIRE exploits its corpus variation. Corpus queries based on collocations demonstrate the possible choices, highlighting the adjective's variation across domains.

The collocations in Figure 7 highlight the vastly different meanings conveyed by this adjective in each domain. In Religion, internal state is intensified (*fede*, "faith"; *tentazione*, "temptation"), while in Cookery flavours and smells are augmented. The meaning in one domain cannot automatically be extended to another.

**Religion**

Risultati: **1** - **20** di **4520**

| Parola | Punteggio | Freq. nel corpus | Freq. collocata | Freq. collocata attesa |
|---|---|---|---|---|
| richiamo | 0,0077 | 47888 | 245 | 4.2191 |
| grido | 0,0071 | 20125 | 129 | 1.7731 |
| tentazione | 0,0069 | 17810 | 116 | 1.5691 |
| fede | 0,0065 | 163796 | 586 | 14.431 |
| supplica | 0,0053 | 2467 | 49 | 0.2174 |
| legame | 0,0051 | 47612 | 161 | 4.1948 |
| lacrima | 0,0037 | 18831 | 64 | 1.6591 |
| vento | 0,0036 | 49803 | 117 | 4.3878 |
| impulso | 0,0036 | 19318 | 63 | 1.702 |
| invito | 0,0032 | 66481 | 134 | 5.8572 |

**Cookery**

Risultati: **1** - **20** di **2049**

| Parola | Punteggio | Freq. nel corpus | Freq. collocata | Freq. collocata attesa |
|---|---|---|---|---|
| farina | 0,169 | 8023 | 1254 | 0.2976 |
| sapore | 0,0138 | 49409 | 388 | 1.8329 |
| odore | 0,0098 | 16632 | 115 | 0.617 |
| aroma | 0,0068 | 9707 | 56 | 0.3601 |
| abbraccio | 0,0066 | 12520 | 64 | 0.4644 |
| drago | 0,0055 | 4226 | 30 | 0.1568 |
| gr | 0,0043 | 60518 | 149 | 2.245 |
| botta | 0,0043 | 7613 | 31 | 0.2824 |
| emicrania | 0,0043 | 644 | 16 | 0.0239 |
| ricotta | 0,0042 | 15348 | 46 | 0.5694 |

**Figure 7: The first 10 collocations (lemmas) of the adjective *forte* in the Religion (left) and Cookery (right) domains.**

Despite the versatility of the collocation extraction procedure and its implementation in linguistic applications, a basic knowledge of corpus querying techniques is required for correct usage. RIDIRE collocations across domains can also be extracted with the Sketches tool, which provides a more intuitive way to obtain linguistically relevant information. In other words, Sketches are more suitable for language learners that do not have high competence in corpus linguistics tools, as it provides them with an explicit language acquisition path.

A Sketch is a selection of relevant lemmas that co-occur with the key lemma in a specific syntactic pattern. The relevance of lemmas in each Sketch is determined by a lexical association measure (log-Dice in the RIDIRE implementation). Each Sketch corresponds to a precise grammatical relation[1]; for example, Figure 8 shows the *e_o* Sketch for the adjective *forte* in all domains i.e. the first ten adjectives that co-occur with *forte*, linked to it by a copulative (*e*, "and") or disjunctive (*o*, "or") conjunction:

| e_o | 39731 | |
|---|---|---|
| deciso | 586 | 8,15 |
| coraggioso | 381 | 7,73 |
| chiaro | 1122 | 7,4 |
| sano | 297 | 7,09 |
| forte | 1148 | 6,92 |
| debole | 313 | 6,87 |
| potente | 266 | 6,71 |
| incisivo | 159 | 6,61 |
| competitivo | 200 | 6,46 |
| intenso | 263 | 6,42 |
| radicato | 118 | 6,41 |

**Figure 8: Example of a Sketch.**

1 RIDIRE Sketches (including both the lexical queries and the visualization layout) are realized with the rules of SketchEngine, that is considered the reference web application for corpus linguistics studies.

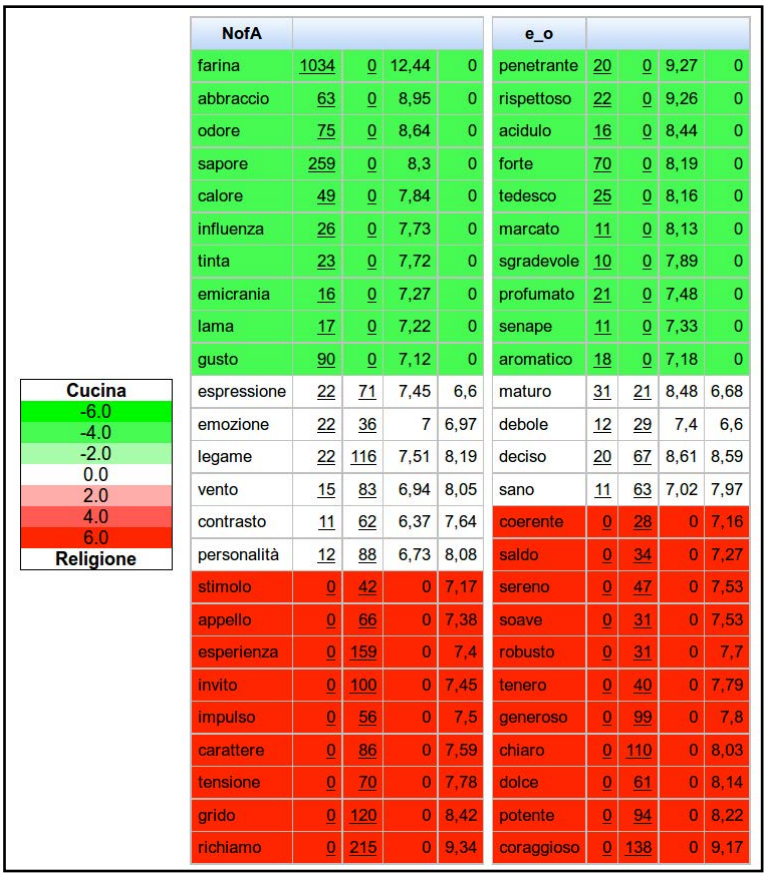
RIDIRE provides two extensions of the Sketch tool: Sketch Difference and Domain Sketch. The Sketch Difference tool shows the difference between the collocational behavior of two lemmas within the same syntactic pattern: we can see the words usable with the first lemma, with the second and with both of them.

In Figure 9 we see the difference between the Italian adjectives *forte* and *resistente* ("resistant") in the Fashion domain; specifically, we select two important Sketches: *e_o*, as in Figure 8, and *NofA*, which selects the nouns related to the adjective. From this example we can see that *forte* has a more varied usage in Fashion and is often related to the characterization of personality traits, while *resistente* is more specific and used for the technical specifications of clothing and accessories.

| forte |
|---|
| -6.0 |
| -4.0 |
| -2.0 |
| 0.0 |
| 2.0 |
| 4.0 |
| 6.0 |
| resistente |

| NofA | | | | | e_o | | | | |
|---|---|---|---|---|---|---|---|---|---|
| impatto | 365 | 0 | 10,52 | 0 | deciso | 159 | 0 | 10,38 | 0 |
| personalità | 291 | 0 | 10,02 | 0 | chiaro | 59 | 0 | 8,31 | 0 |
| legame | 137 | 0 | 9,21 | 0 | sicuro | 38 | 0 | 8,14 | 0 |
| tinta | 169 | 0 | 9,18 | 0 | indipendente | 23 | 0 | 8,05 | 0 |
| pezzo | 254 | 0 | 9,15 | 0 | determinato | 17 | 0 | 7,93 | 0 |
| carattere | 122 | 0 | 8,87 | 0 | sensuale | 38 | 0 | 7,91 | 0 |
| crescita | 174 | 0 | 8,85 | 0 | riconoscibile | 16 | 0 | 7,7 | 0 |
| identità | 108 | 0 | 8,85 | 0 | grintoso | 19 | 0 | 7,67 | 0 |
| appeal | 73 | 0 | 8,36 | 0 | sano | 18 | 0 | 7,65 | 0 |
| espansione | 79 | 0 | 8,35 | 0 | coraggioso | 12 | 0 | 7,46 | 0 |
| richiamo | 71 | 0 | 8,26 | 0 | emerso | 10 | 0 | 7,44 | 0 |
| presenza | 99 | 0 | 8,24 | 0 | vivace | 19 | 0 | 7,37 | 0 |
| colore | 374 | 0 | 8,19 | 0 | pratico | 0 | 13 | 0 | 7,29 |
| segnale | 63 | 0 | 8,16 | 0 | aerodinamico | 0 | 3 | 0 | 7,31 |
| emozione | 70 | 0 | 8,12 | 0 | protettivo | 0 | 6 | 0 | 7,34 |
| contrasto | 94 | 0 | 8,09 | 0 | funzionale | 0 | 10 | 0 | 7,6 |
| punto | 152 | 0 | 7,96 | 0 | duttile | 0 | 4 | 0 | 7,76 |
| messaggio | 51 | 0 | 7,67 | 0 | elastico | 0 | 10 | 0 | 7,82 |
| connotazione | 39 | 0 | 7,61 | 0 | leggero | 0 | 56 | 0 | 7,94 |
| impronta | 40 | 0 | 7,57 | 0 | comodo | 0 | 22 | 0 | 7,94 |
| contenuto | 49 | 0 | 7,53 | 0 | robusto | 0 | 9 | 0 | 8,54 |
| carica | 38 | 0 | 7,43 | 0 | capiente | 0 | 13 | 0 | 8,63 |
| polimero | 0 | 2 | 0 | 7,73 | impermeabile | 0 | 24 | 0 | 8,96 |
| guaina | 0 | 3 | 0 | 8,14 | flessibile | 0 | 19 | 0 | 9,44 |
| ultra | 0 | 4 | 0 | 9,24 | ultra | 0 | 11 | 0 | 9,51 |

**Figure 9: The Sketch Difference for the adjectives *forte* and *resistente* in the Fashion domain.**



| Cucina |
|---|
| -6.0 |
| -4.0 |
| -2.0 |
| 0.0 |
| 2.0 |
| 4.0 |
| 6.0 |
| Religione |

| NofA | | | | | e_o | | | | |
|---|---|---|---|---|---|---|---|---|---|
| farina | 1034 | 0 | 12,44 | 0 | penetrante | 20 | 0 | 9,27 | 0 |
| abbraccio | 63 | 0 | 8,95 | 0 | rispettoso | 22 | 0 | 9,26 | 0 |
| odore | 75 | 0 | 8,64 | 0 | acidulo | 16 | 0 | 8,44 | 0 |
| sapore | 259 | 0 | 8,3 | 0 | forte | 70 | 0 | 8,19 | 0 |
| calore | 49 | 0 | 7,84 | 0 | tedesco | 25 | 0 | 8,16 | 0 |
| influenza | 26 | 0 | 7,73 | 0 | marcato | 11 | 0 | 8,13 | 0 |
| tinta | 23 | 0 | 7,72 | 0 | sgradevole | 10 | 0 | 7,89 | 0 |
| emicrania | 16 | 0 | 7,27 | 0 | profumato | 21 | 0 | 7,48 | 0 |
| lama | 17 | 0 | 7,22 | 0 | senape | 11 | 0 | 7,33 | 0 |
| gusto | 90 | 0 | 7,12 | 0 | aromatico | 18 | 0 | 7,18 | 0 |
| espressione | 22 | 71 | 7,45 | 6,6 | maturo | 31 | 21 | 8,48 | 6,68 |
| emozione | 22 | 36 | 7 | 6,97 | debole | 12 | 29 | 7,4 | 6,6 |
| legame | 22 | 116 | 7,51 | 8,19 | deciso | 20 | 67 | 8,61 | 8,59 |
| vento | 15 | 83 | 6,94 | 8,05 | sano | 11 | 63 | 7,02 | 7,97 |
| contrasto | 11 | 62 | 6,37 | 7,64 | coerente | 0 | 28 | 0 | 7,16 |
| personalità | 12 | 88 | 6,73 | 8,08 | saldo | 0 | 34 | 0 | 7,27 |
| stimolo | 0 | 42 | 0 | 7,17 | sereno | 0 | 47 | 0 | 7,53 |
| appello | 0 | 66 | 0 | 7,38 | soave | 0 | 31 | 0 | 7,53 |
| esperienza | 0 | 159 | 0 | 7,4 | robusto | 0 | 31 | 0 | 7,7 |
| invito | 0 | 100 | 0 | 7,45 | tenero | 0 | 40 | 0 | 7,79 |
| impulso | 0 | 56 | 0 | 7,5 | generoso | 0 | 99 | 0 | 7,8 |
| carattere | 0 | 86 | 0 | 7,59 | chiaro | 0 | 110 | 0 | 8,03 |
| tensione | 0 | 70 | 0 | 7,78 | dolce | 0 | 61 | 0 | 8,14 |
| grido | 0 | 120 | 0 | 8,42 | potente | 0 | 94 | 0 | 8,22 |
| richiamo | 0 | 215 | 0 | 9,34 | coraggioso | 0 | 138 | 0 | 9,17 |

**Figure 10: Domain Sketches (Cooking *vs.* Religion) for the adjective *forte*.**

As the Sketch Difference function displays the contrast between the lexical associations of two lemmas in one corpus domain, the Domain Sketch tool shows the variation of a single lemma between two different domains.

In the Figure 10 we used the Domain Sketch tool to search the differences in usage for the lemma *forte* in two domains: Cooking and Religion. The general difference between these domains (*forte* is applied to flavours and smells in the Cooking domain and to feelings in the Religion domain) has already been demonstrated with the collocation search (Figure 7); however the result here is more fine grained, as it is divided into sketches, giving a more comprehensive overview of the lemma usage.

## 4.1 Lexicographic applications

Corpora have been widely used as data source in lexicography (Kilgarriff 2013). As a matter of fact, each of the researches presented in the previous section provide very relevant information for the lexical description of a word. Moreover, large corpora can be used as test-beds in order to decide what words and meanings should be inserted in a dictionary.

One of the main application field of corpora in lexicography is the detection of neologism by means of automatic or semi-automatic comparative analysis between an older word lists, taken from a dictionary or from a previous reference corpus, and the newer one, derived from an up-to-date corpus

(O'Donovan & O'Neill 2008). In this respect, web corpora are particularly interesting, since the web can be nowadays considered as the main access to written language, both in comprehension and in production, for a large part of the population.

The dimension and the structure of the RIDIRE corpus make it particularly attractive for lexicographic purposes. For instance, its data have been explored by Carla Marello for the study of Latin loanwords in Italian. The results showed that, in this respect, the corpus is richer than the modern dictionaries: all the Latinisms that are frequent in Italian monolingual dictionaries are frequent also in the corpus, but the corpus contains also various frequent Latinisms that are not reported in the dictionaries (but they probably should be).

The availability of very large corpora gave also a new perspective in the studies of collocations. Starting from these data. for example, it becomes possible to determine the input to which the learners are exposed while reading, and to select the collocations that should be considered during the compilation of monolingual and learner's dictionaries (Marello 2013). The use of sketches, that are a sort of quick synopsis of the grammatical and collocational behavior of a word, makes available a wide range of usage pattern that should be considered during the dictionary creation process.

Moreover, Sketches are useful not only for the detection of collocations, but also to give a quick picture of the distinct meanings of a word, since different meanings often select different collocates (Kilgarriff & Rundell 2002). It has to be noticed that the significance of this "extraction procedure" grows proportionally to the corpus dimension. If detecting meanings and collocations from very large corpora by means of concordance scanning could be very hard and time consuming, for the automatic collocation extraction procedures the bigger is the corpus, the better are the sketches (both in quantitative and in qualitative terms). Finally, the Sketch Differences tool is specifically interesting for comparing a word with its (near) synonyms and antonyms, in a pure lexicographic perspective.

## 5 Conclusions

Large scale corpora representing a language's domain of usage offer a unique source of data to both learners and lexicographers in accessing information about how the language is actually used. The computational tools now available, including those for web based infrastructures, allow the selection of the relevant information in a simple manner, overcoming significant difficulties encountered by corpus linguistics in meeting second language acquisition needs. Learners, teachers, and lexicographers, however, must be aware of the information required for a proper language acquisition that are up to usage conventions. On the basis of this understanding, corpus querying can be used to solve specific problems and be accepted as a modern method for use in the language acquisition process and in the dictionary creation.

# 6 References

Alchemy API. Accessed at: http://www.alchemyapi.com/ [06/04/2014].

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.

Conrad, S. (2006). Challenges for English Corpus Linguistics in Second Language Acquisition Research. In Y. Kawaguchi, S. Zaima, T. Tackagaki, Y. Tsuruga, M. Usami (eds) *Lingusitcs Informatics and Spoken Language Corpora*. Amsterdam/Philadelphia: John Benjamins.

CQPweb. Accessed at: http://cwb.sourceforge.net/cqpweb.php [06/04/2014].

Heritrix. Accessed at: http://crawler.archive.org/ [06/04/2014].

Kilgarriff, A. (2009). Corpora in the classroom without scaring the students. In *Proceedings of 18th Internat. Symposium on English Teaching, Taipei.* Accessed at: http://www.kilgarriff.co.uk/Publications/2009-K-ETA-Taiwan-scaring.doc [06/04/2014].

Kilgarriff, A. (2013). Using corpora as data sources for dictionaries. In H. Jackson (ed.), *The Bloomsbury Companion to Lexicography*. London: Bloomsbury, pp. 77-96.

Kilgarriff, A., Greffenstette, G. (2003). Introduction to the Special Issue on Web as Corpus. In *Computational Linguistics*, 29(3), pp. 1-15.

Kilgarriff, A., Rundell, M. (2002). Lexical Profiling Software and its Lexicographic Applications: A Case Study. In A. Braasch, C. Povlsen (eds), *Proceeding of the Tenth Euralex Conference, Copenhagen, 13-17 August 2002*. Copenhagen: University of Copenhagen, pp. 807-818.

Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. (2004). The Sketch Engine. In G. Williams, S. Vessier (eds) *Proceeding of the Eleventh Euralex Conference, Lorient (France), 6-10 July 2004*. Lorient: Université de Bretagne-Sud, pp. 105-116.

Marello, C. (2013). Sembra che e subordinate soggettive. Primi sondaggi in italiano L2 scritto. In F. Geymonat (ed.) *Linguistica applicata con stile. In traccia di Bice Mortara Garavelli*. Alessandria: Edizioni dell'Orso, pp. 79-94.

Moneglia, M., Paladini, S. (2010). Le risorse di rete dell'italiano. Presentazione del progetto "RIDIRE.it". In E. Cresti, I. Korzen (eds) *Language, Cognition and Identity*. Firenze: Firenze University Press, pp. 111-128.

O'Donovan, R., O'Neill, M. (2008). A Systematic Approach to the Selection of Neologisms for Inclusion in a Large Monolingual Dictionary. In E. Bernal, J. DeCesaris (eds) *Proceeding of the Thirteenth Euralex Conference, Barcelona, 15-19 July 2008*. Barcelona: Universitat Pompeu Fabra, pp. 571-579.

Panunzi, A., Fabbri, M., Moneglia, M., Gregori, L., Paladini, S. (2012). RIDIRE-CPI: an Open Source Crawling and Processing Infrastructure for Supervised Web-Corpora Building. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds) *Proceedings of Eighth Language Resources and Evaluation Conference (LREC 2012), Istanbul, 23-25 May 2012*. Paris: ELRA, pp. 2274-2279.

Readability. Accessed at: http://www.readability.com/ [06/04/2014].

RIDIRE Corpus Online. Accessed at: http://www.ridire.it [06/04/2014].

RIDIRE-CPI. https://github.com/lablita/ridire-cpi [06/04/2014].

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni, S. Bernardini (eds), *Wacky! Working papers on the Web as Corpus*. Bologna: Gedit, pp. 63-98.

Sinclair, J. (ed.) 2004. *How to use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Sketch Engine. Accessed at: http://www.sketchengine.co.uk/ [06/04/2014].

TreeTagger. Accessed at: http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [06/04/2014].

WaCky. Accessed at: http://wacky.sslmit.unibo.it/doku.php [06/04/2014].

**Acknowledgments**

The RIDIRE Project is funded by MIUR – FIRB 2007 and is promoted and maintained by SILFI (Società Internazionale di Linguistica e Filologia Italiana). The web application RIDIRE-CPI was developed by LABLITA and the corpus creation involved six Italian university departments: University of Florence (Dip. Italianistica and Dip. Sistemi e Informatica), University of Turin (Dip. Scienze Letterarie e Filologiche), University of Siena (Dip. Studi Aziendali e Sociali), University of Rome – Roma 3 (Dip. Italianistica), University of Naples – Federico II (Dip. Filologia Moderna).