DICI-A

Dizionario delle collocazioni italiane per apprendenti

PRIN 2022 - Synthetic report of the project

https://dici-a.it/

DICI-A
Dizionario delle Collocazioni Italiane per Apprendenti

Partners and project members

Funded by the European Union NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

Università per Stranieri di Perugia

- Stefania Spina (Principal Investigator)
- Irene Fioravanti
- Fabio Zanda
- Luciana Forti (from July 2024 to September 2025)

A.D. 1308 unipg UNIVERSITÀ DEGLI STUDI DI PERUGIA

- Osvaldo Gervasi
- Sergio Tasso
- Damiano Perri

General information

ERC field: SH4

ERC subfields:

- ✓SH4_8 Language learning and processing - first and second languages
- ✓SH4_9 Theoretical linguistics; computational linguistics

- ✓PRIN 2022 n. 2022HXZR5E

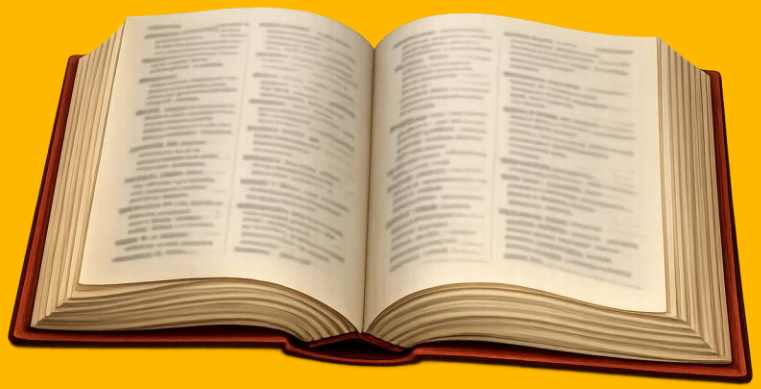
- ✓from 05-10-2023 to 28-02-2026

DICI-A

Dizionario delle collocazioni italiane per apprendenti

Motivation

The project *DICI-A - Dizionario delle Collocazioni Italiane per Apprendenti* falls within the research areas of applied linguistics, lexicography, second language acquisition and phraseology, with a focus on collocations. It is intended to fill a gap in the pedagogical lexicography of Italian, where there is a lack of dictionaries specifically designed for learners of Italian as an L2.



Aims

1

Create a learner dictionary of Italian collocations designed according to criteria that lie at the intersection of lexicography, corpus linguistics, second language acquisition and language teaching. The DICI-A is intended for both classroom and individual use by teachers and learners, as well as for scientific research purposes.

2

Develop *Link*, a database of Italian collocations, based on a sample of the entries of the DICI-A, enriched with data on their measurable and psycholinguistic properties, designed to support research on the associative links of collocations in Italian.

Main features

1. The DICI-A includes **10,596** Italian collocations, each representing a dictionary entry, belonging to seven syntactic types.

2. It has been created according to corpus-based criteria. The collocations included as dictionary entries have been extracted from the **PEC24**, a 47-million-word reference corpus of written and spoken Italian.

3. The selection of dictionary entries has involved a three-step process of **extraction** of candidate collocations combining part-of-speech and parsing methods; **filtering**, through the integration of dispersion, association measures and frequency; **validation**, using two existing collocation dictionaries and human evaluation.

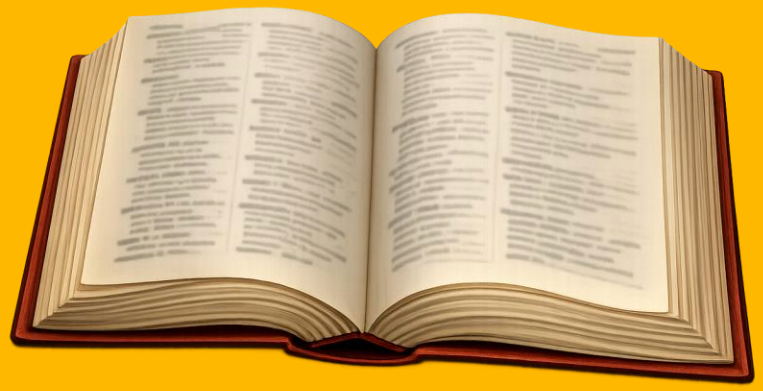
4. Each collocation included in the DICI-A has been assigned to a specific **proficiency label**, using a set of criteria, both quantitative and qualitative, based on the CEFR descriptors.

5. Each collocation has been assigned definitions and examples of use suitable to learners' proficiency, combining **Generative Artificial Intelligence** and human assessment.

6. The DICI-A is freely available online at : https://dictionary.dici-a.it/. It is searchable both using a pc and a mobile device.

Our definition of collocation:

- ✓ A co-occurrence of two syntagmatically related words in a *grammatical configuration* and *syntactic relation*
- ✓ Commonly used (*frequency*) in a range of texts (*dispersion*), where the two words are strongly associated (*association measures*) either adjacently or within a distance
- ✓ A combination whose meaning may be more or less *compositional* and *transparent*



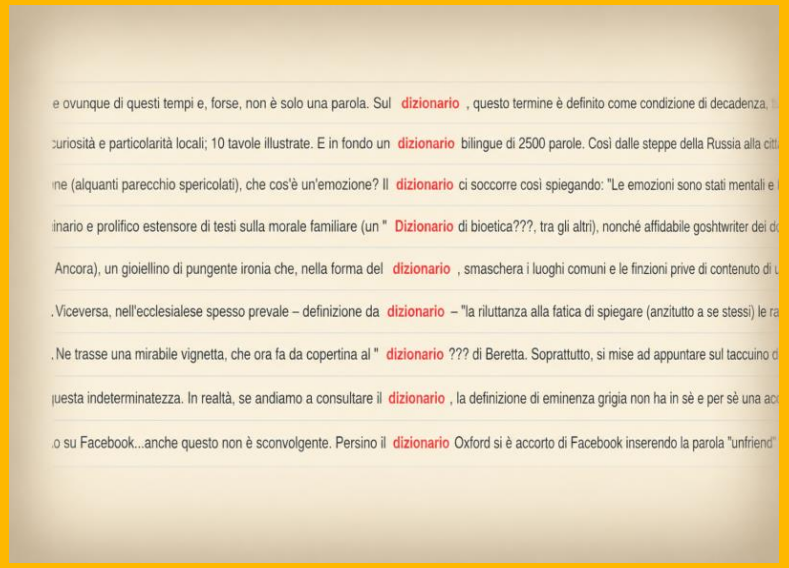
Definition

Seven syntactic types

1. Verb + Direct object (vdobj): *mantenere una promessa*, ‘to keep a promise’
2. Adjective + Noun (amod): *brutta avventura*, ‘bad adventure’
3. Noun + Adjective (amod): *tempo libero*, ‘free time’
4. Verb + Adjective (advmod1): *stare zitto*, ‘to stay quiet’
5. Verb + Adverb (advmod2): *fare presto*, ‘to hurry up’
6. Adverb + Adjective (advmod3): *altamente positivo*, ‘highly positive’
7. Noun + Noun (comp): *banca dati*, ‘data base’

Corpus

The PEC24 corpus has been selected as the most suitable to serve as a reference to extract collocations for a dictionary, as it covers 10 different written and spoken text genres: *academic writing*, *school essays*, *administrative writing*, *literary fiction*, *non-fiction*, *newspapers*, *web texts*, *film dialogues*, *spoken texts*, *tv programs*.
These genres can effectively represent the potential input to which L2 learners are exposed.



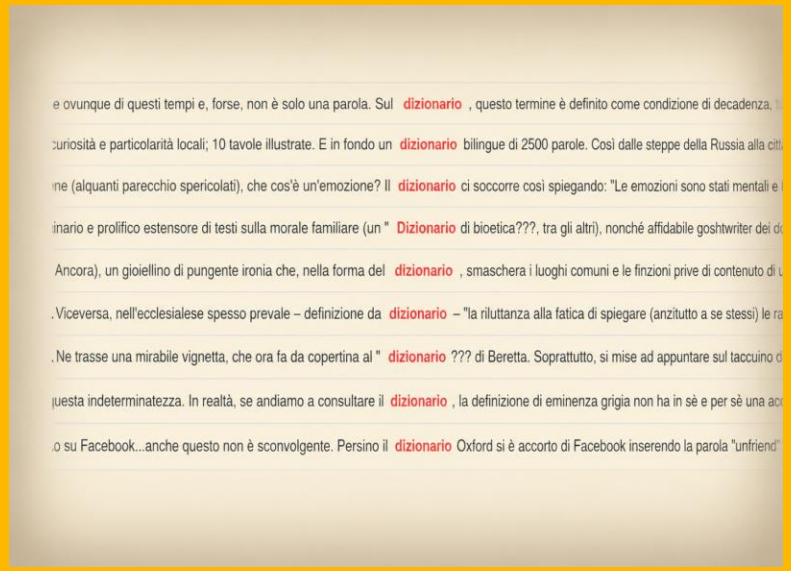
PEC24 is available at Search-it - Italian native and learner corpora

Identification of collocation entries

A three step selection method

Step 1

- ✓ the extraction of candidate collocations combined POS-based and parsing-based methods, to maximise the benefits of each method and minimise the drawbacks
- ✓ the extraction was based on fixed queries performed in the POS-tagged and in the parsed versions of the PEC24 corpus, and on specific rules to make the queries more effective
- ✓ each candidate was associated to its measures of frequency, dispersion and association.



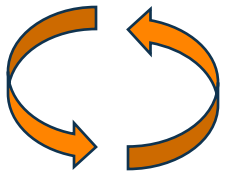
Extraction

Identification of collocation entries

STEP 1: Extraction of candidate collocations > 2 million candidate collocations

POS-based (Part-of-Speech) method

- Assigns each token its POS
- High accuracy in identifying adjacent combinations
- Limited in identifying non-adjacent word pairs syntactically related

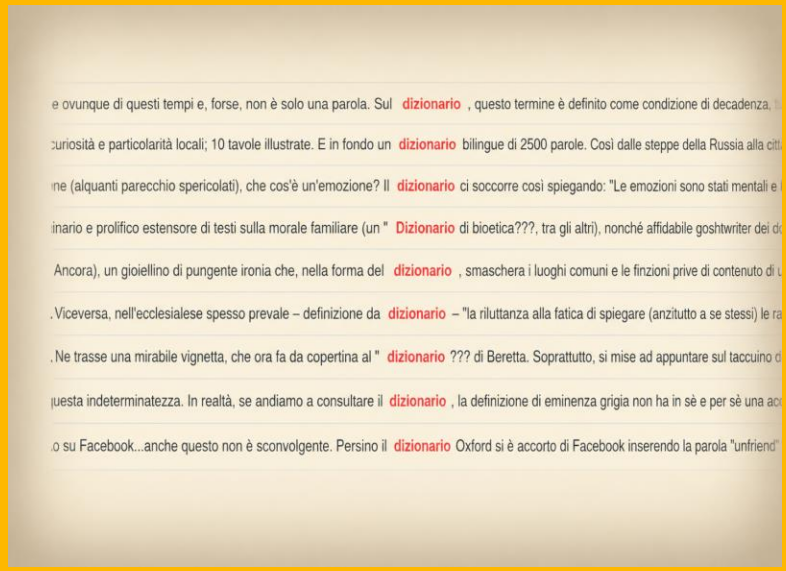


Parsing-based method

- Identifies the syntactic relationships between the lexical items
- Has yet to reach high accuracy
- Capable to extract syntactic relationships between non-adjacent words

Step 2

- ✓ The 2 million candidate collocations were filtered in 3 stages according to the properties represented by each of the selected measures.
- ✓ *Dispersion* was used consistently throughout all stages, with the aim of only retaining the collocations used across a range of texts.
- ✓ In <u>stage 1</u> strongly associated collocations were retained using *Mutual information*.
- ✓ In <u>stage 2</u> low frequencies collocations were removed.
- ✓ In <u>stage 3</u> poorly associated collocations were removed.



Filtering

Identification of collocation entries

STEP 2: Filtering of candidate collocations

- Stage 1: retaining strongly associated collocations
  - DPnorm ≤ 0.55; MI ≥ 7.00; LL=0
    2,097,595 candidates > **12,201 candidates**

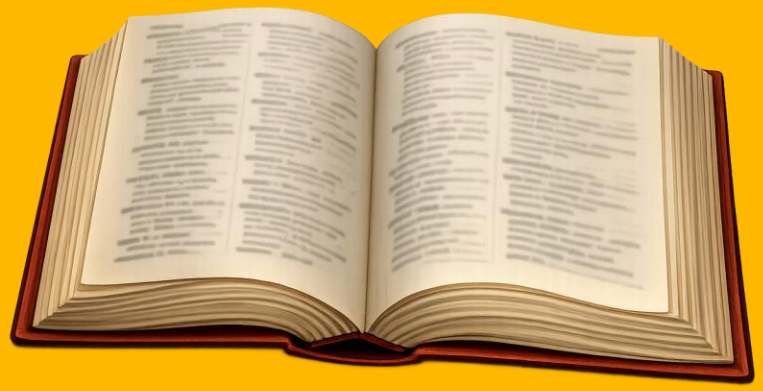
- Stage 2: removing low frequencies collocations
  - DPnorm ≤ 0.55; MI < 7.00; Raw frequency ≥ 30
    2,097,595 canidates > **9,177 candidates**

- Stage 3: removing poorly associated collocations
  - DPnorm ≤ 0.55; LogDIce ≥ 5; LL = 0
    9,177 canidates > **4,619 candidates**

Final set of candidates retained
(Total of Stage 1 +
Total of Stage 3)
= **16,820**

Step 3

- ✓ A comparison against two existing Italian collocations dictionaries not targeted to L2 learners was performed.
- ✓ The candidate collocations that were not found in either dictionary were then evaluated by three human raters.
- ✓ After this evaluation, **2,297** were judged suitable for inclusion in the DICI-A.
- ✓ After a manual check (geographical and technical collocations were removed), a final list of **10,596** collocations to include as dictionary entries was available.



Validation

Identification of collocation entries

STEP 3: Validation of candidate collocations

Evaluation against two Italian collocations dictionaries (Tiberii 2012 and Lo Cascio 2013)

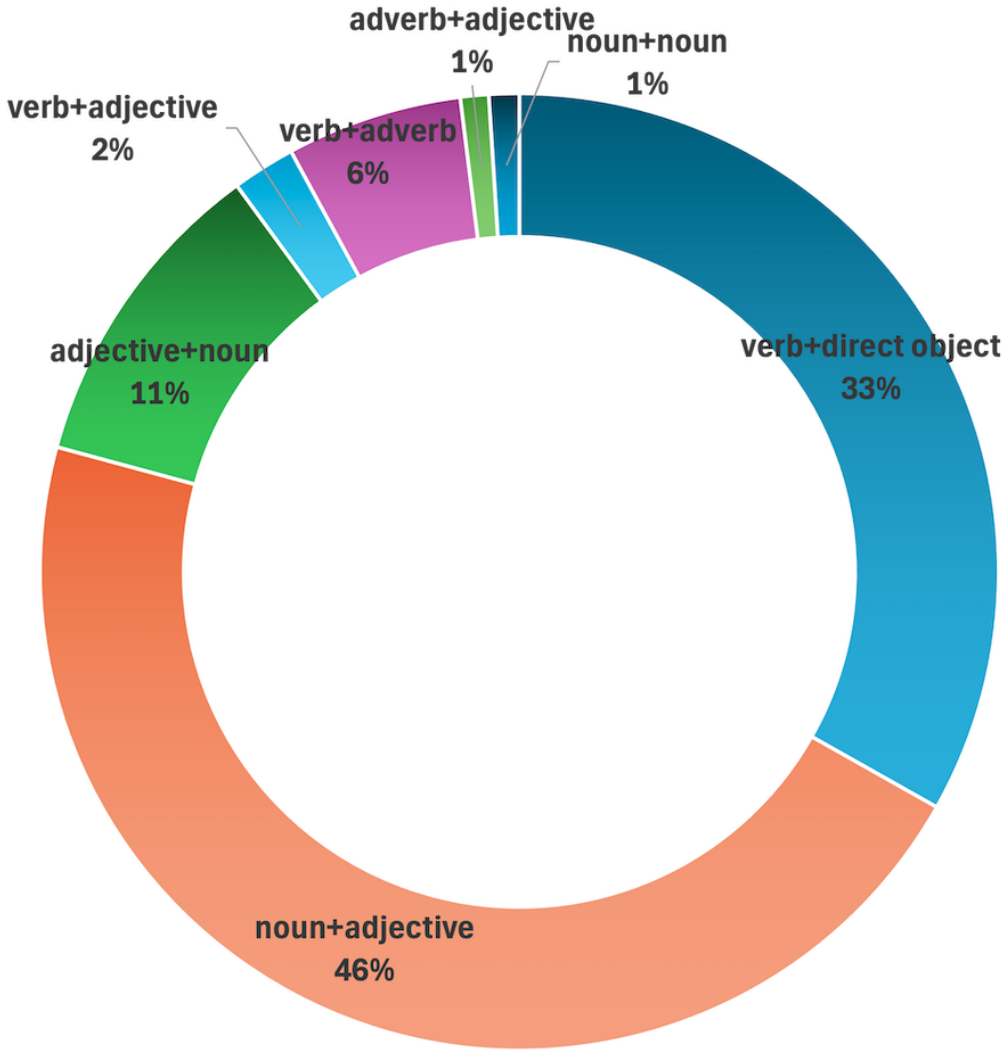
- Found in at least one dictionary: **9,977**
- Not found in either dictionary: **6,843**

Human evaluation of the 6,843 not found in dictionaries

- Judged suitable for inclusion: **2,297**

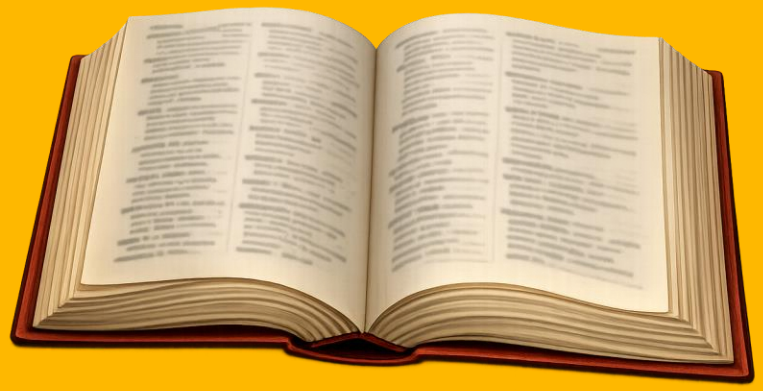
Creation of the final list: **12, 274** candidates > **10,596** after manual check

Final distribution of collocation entries



Proficiency

- ✓ The list of 10,596 dictionary entries served as the starting point for the work on the pedagogical and lexicographical features of the learner dictionary.
- ✓ Each collocation was assigned to an appropriate proficiency label **A** (*beginner*), **B** (*intermediate*) and **C** (*advanced*).
- ✓ Five quantitative and qualitative criteria were defined to perform this task.



Proficiency-label attribution: 5 criteria

1. The **frequency** and **dispersion** values derived from the PEC24 corpus. The rationale behind this criterion is that a frequent and dispersed collocation should be easier to understand and use for learners.

2. The **presence of collocation constituents** in the word lists of the *Profilo della lingua italiana* (Spinelli & Parizzi, 2010), whose aim is to define the Italian vocabulary included in the *Common European Reference for Languages* (CEFR) levels from A1 to B2.

3. The **semantic transparency** of the collocation. Each collocation has been assessed as *transparent*, *semi-transparent*, *figurative* or *idiomatic*.

4. The **linguistic register.** Each collocation has been assessed as *neutral*, *colloquial*, *formal* and *technical*.

5. The **vocabulary range descriptor** established by the CEFR at each proficiency level (CEFR, Council of Europe, 2020).

Proficiency-label attribution: procedure

1. **Step 1:** each combination was annotated independently by two annotators

2. **Step 2:** cases of disagreement were identified

3. **Step 3:** a third annotator resolved the cases with no agreement

verb + direct object collocations:

✓ Inter-annotator agreement between the first two annotators: **80%**

✓ Re-annotated collocations: **20%**

Proficiency-label attribution: example

fare colazione (‘to have breakfast’)

1. High frequency and high dispersion
2. Word 1 in *Profilo*: A1 level; Word 2 in *Profilo*: A1 level
3. Semantic transparency: *transparent*
4. Register: *neutral*
5. CEFR Descriptor A2: *basic communicative needs*

Label

A

Proficiency-label attribution: example

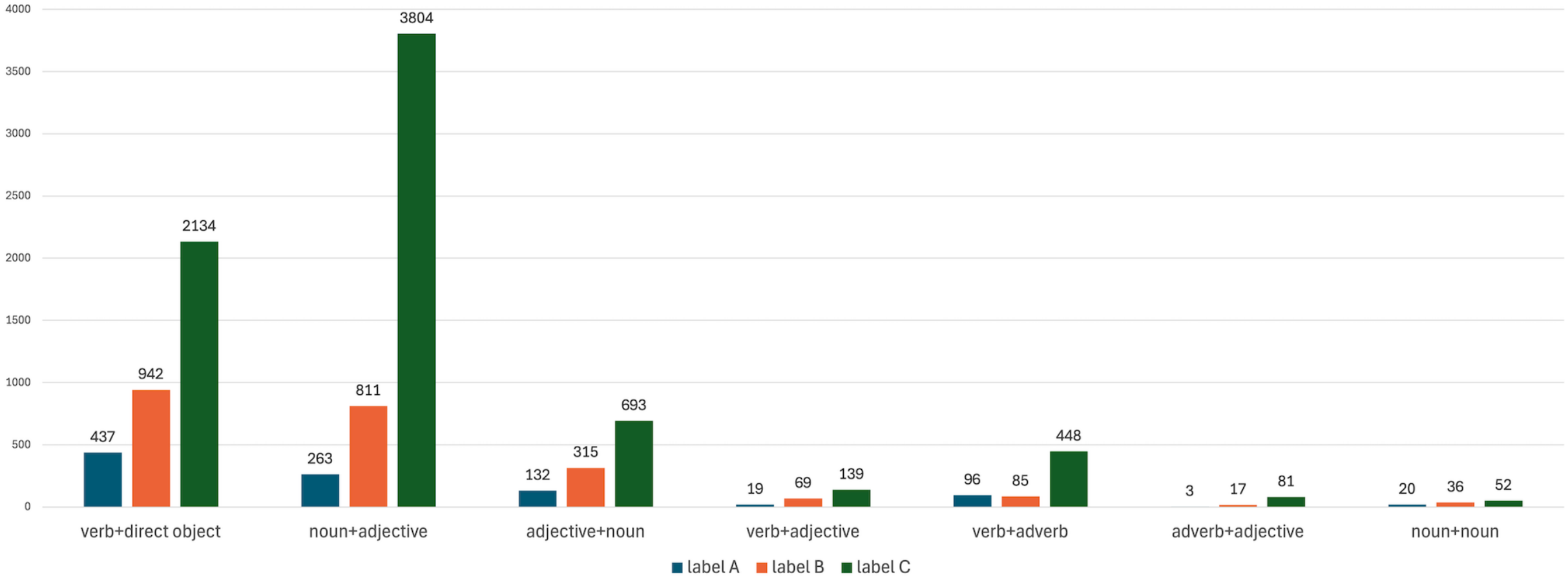
gettare la spugna (‘to give up’)

1. Low frequency and high dispersion
2. Word 1 in *Profilo*: absent; Word 2 in *Profilo*: absent
3. Semantic transparency: *idiomatic*
4. Register: *neutral*
5. CEFR Descriptor C1: *good command of common idiomatic expressions*

Label

C

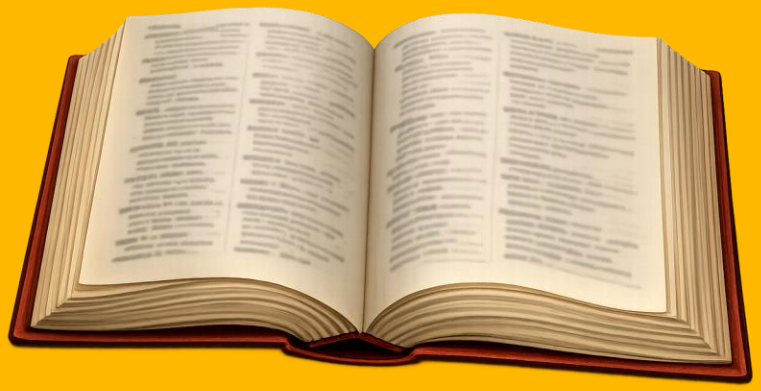
Proficiency-label attribution



Distribution of the 10,596 collocations by type and proficiency label

Integrating GenAI and human evaluation

- ✓ As the DICI-A is aimed at learners, the challenge of creating definitions and examples lies in the need to provide texts that are easy to understand. In this task, we used **generative artificial intelligence** (GenAI) to support human decision-making.
- ✓ Recent lexicographical research showed that GenAI can support the process of creating high-quality dictionaries, by generating sense definitions in line with the style of dictionaries and by producing good examples of use.
- ✓ Based on these studies, we have developed and applied a method for the creation of lexicographical definitions and examples that **combines the output of GenAI with human evaluation**.



Definitions and examples

GenAI and human evaluation: prompt

- ChatGPT 4o: optimised zero-shot prompting strategy

- You are given a list of Italian collocations to develop a learner dictionary. For each collocation, your task is to generate:
  - definition (in Italian, max 30 words, simple syntax, learner-friendly vocabulary)
  - if the collocation has multiple distinct meanings (e.g., literal and figurative, or domain-specific), you must output one separate row per meaning
  - describe the meaning of the full collocation (not individual words)
  - avoid repeating words from the collocation, unless strictly necessary
  - use clear, simple vocabulary suitable for learners of Italian

GenAI and human evaluation: assessment

- ✓ Assessing the quality of the AI-generated definitions with L1 and L2 speakers of Italian

- ✓ **Method**:
  - ✓ elicitation task submitted to 150 L1 and L2 speakers of Italian
  - ✓ 75 collocations of three different types as stimuli
  - ✓ two parts:
    1. form recognition task: starting from the AI-generated definition, it was required to identify the collocation to which it referred among four
    2. judgment task: the same definitions were rated on a Likert scale from 1 to 7

- ✓ **Results**:
  - ✓ AI-generated definitions helped learners identify the intended collocation
  - ✓ ChatGPT produced clear definitions: all three collocation types received high clarity ratings across speaker groups

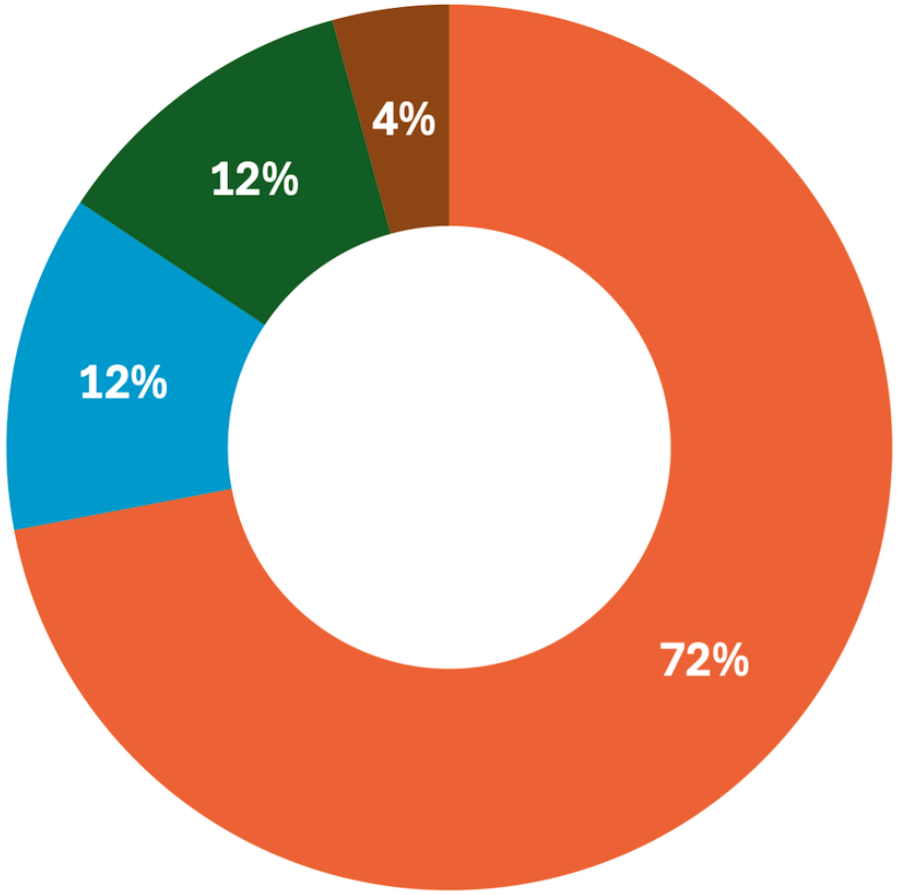
GenAI and human evaluation: human check

- ✓ **Two human evaluators checked the quality and validity of all the AI-generated definitions and examples**

- ✓ Four possible ratings:
  1. acceptance of both definitions and examples
  2. acceptance of definitions only
  3. acceptance of examples only
  4. acceptance of neither definitions nor examples
- ✓ In the event of non-acceptance, the evaluator supplemented or rewrote the non-accepted part.

- ✓ **Results**: 72% of the definitions and examples generated by AI was assessed as accurate ad suitable for learners by human evaluators.

Human evaluation of GenAI output: results

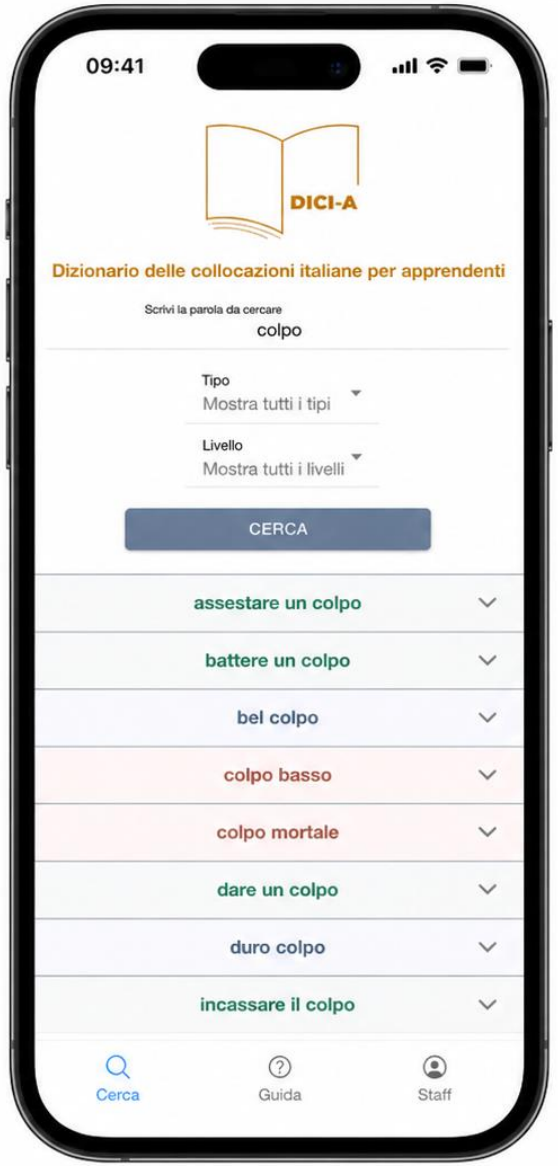


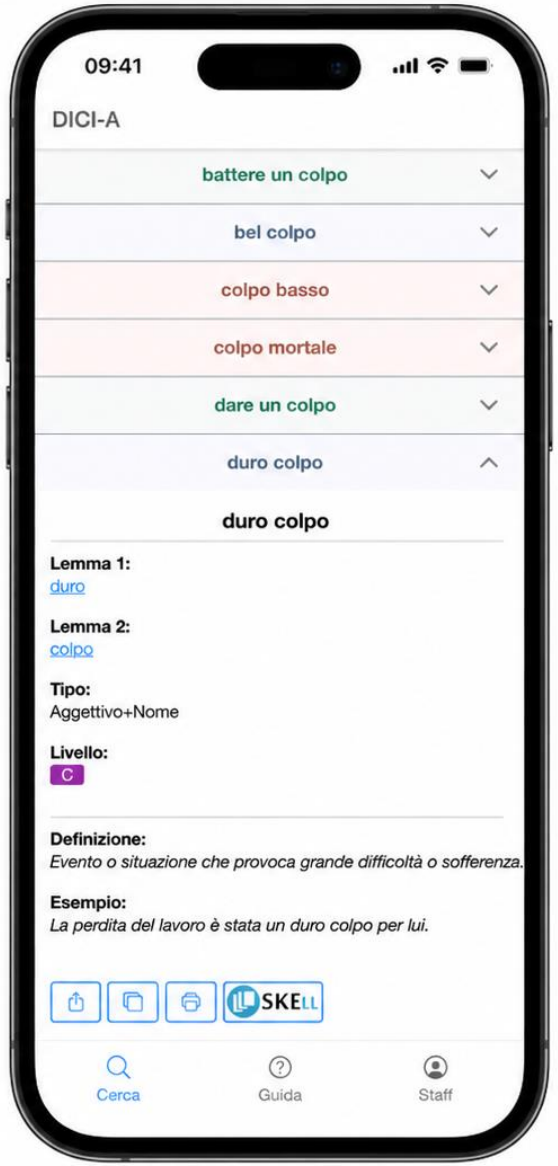
- ✓ The DICI-A is a **digital**, open and updatable resource, **freely accessible** and **searchable**, both on desktop and portable devices through a dedicated digital infrastructure.

- ✓ The interface has been designed to meet educational needs: the dictionary is a teaching tool that can be used both in the classroom, by teachers and students, and individually, by learners of Italian whose language skills are sometimes very limited. The interface has therefore been made particularly simple and intuitive to facilitate its use, especially for learners.

Interface

The DICI-A can be searched by lemma (or part of lemma), syntactic type of collocation, proficiency level A, B or C, or by combining all these criteria.

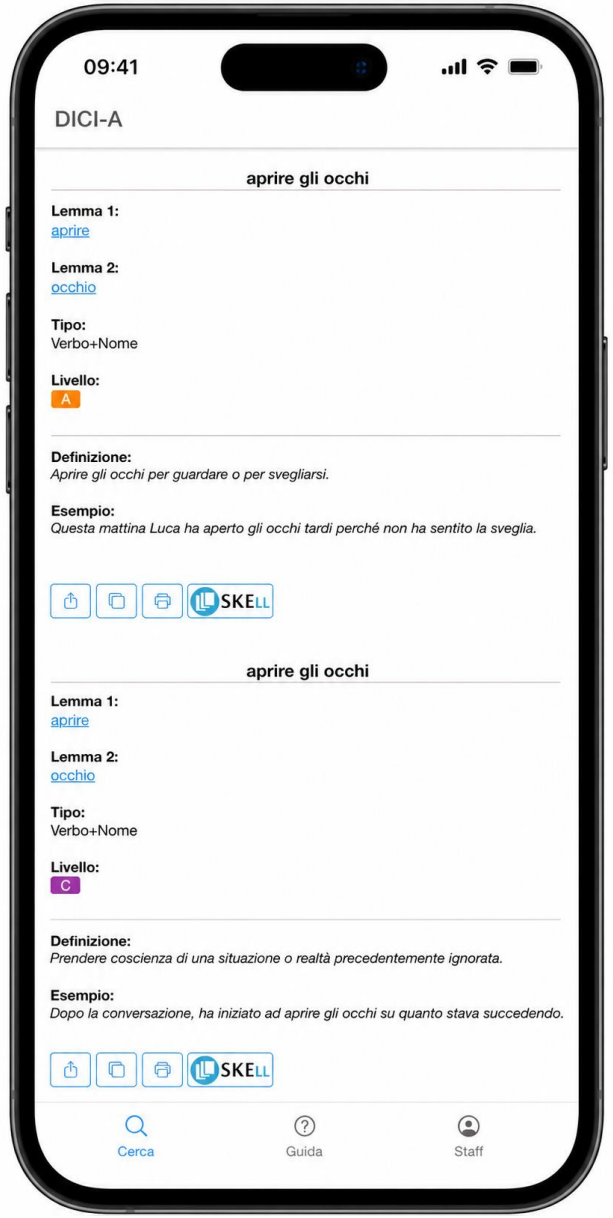




Each entry shows the syntactic type, the proficiency label, the definition(s) and the example(s).

Each entry can be copied, saved, printed and shared.

When a collocation has more than one sense, all the senses are displayed within the same entry, each with its own labels, definitions and examples



Each entry includes a **Skell** button, which allows users to view further real-life examples of the collocation used in different contexts. SKELL (https://www.sketchengine.eu/skell/) is a free for observing how words and word combinations are used in context, and to this purpose it has been integrated in the DICI-A.

***Link*: a resource to support research**

- ✓ ***Link*** is a database of Italian collocations based on a sample of 240 entries of the DICI-A, enriched with data on their statistical and psycholinguistic properties, created for research purposes.

- ✓ A balanced sample of three of the most common collocation types, associated with the statistical measures acknowledged by research on phraseology, has been assessed by native Italian speakers in terms of strength of association and familiarity.

- ✓ The database is an open resource, updatable over time, constituting a benchmark for the study of collocations in Italian, in the fields of phraseology, second language acquisition and psycholinguistics.

References

How to cite the DICI-A:

Spina S., Fioravanti I., Zanda F., Perri, D. & Gervasi O. (2026). *DICI-A. Dizionario delle collocazioni italiane per apprendenti*, Dizionario online, https://dictionary.dici-a.it/, ISBN 979-12-243-2944-2

The dictionary:

https://dictionary.dici-a.it/

The web site of the project:

https://dici-a.it/

All publications on the project:

https://dici-a.it/publications

DICI-A

More info:
Prof. Stefania Spina
info@dici-a.it