


Using a learner corpus to design a phraseological syllabus of Italian collocations

Francesca La Russa, Maria Roccaforte
Sapienza Università di Roma



Defining, learning
and teaching
collocations

1. Defining collocations

A collocation is an institutionalized word combination, corresponding to a conventionalized way of saying a certain thing.

A lexical restriction applies, for which the choice of a particular word (the collocate) to express a given meaning is influenced by a second word (the base) to which this meaning applies.

For example: “pay attention”; “heavy rain”

Jezek (2016: 199-200)



1. Learning collocations

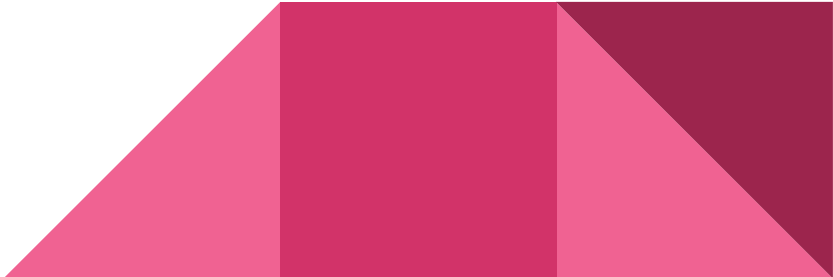
Lexical combinations are central to language learning:

- **processed more quickly** than free combinations (Siyanova-Chanturia, 2015);
- **“islands of reliability”** (Henriksen, 201) on which learners can rely instead of constructing the message word by word.
- **increase fluency** in production (Nesselhauf, 2005).



1. Learning collocations

Research on collocations based on learner corpora shows that acquiring a collocational competence is often a difficult and non linear process.

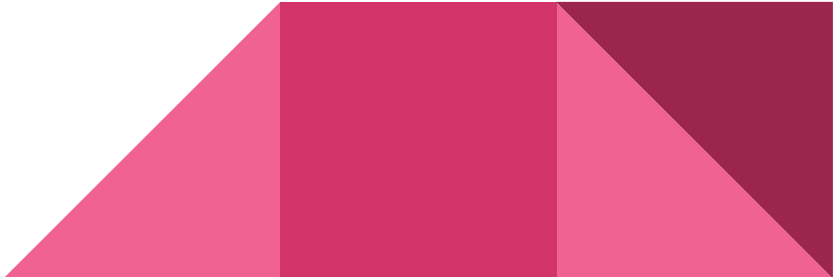
- The production of collocations remains a significant obstacle even for advanced learners (Wang, 2016);
 - Their longitudinal development in learners' interlanguage is slow (Yoon, 2016) and follows a U-shaped pattern (Bestgen & Granger, 2014; Siyanova-Chanturia & Spina, 2019).
 - Compared to native speakers, learners tend to overuse a few very frequent collocations (Durrant & Schmitt, 2009).
- 

1. Teaching collocations

According to Lewis (2000:8):

The single most important task facing language learners is acquiring a sufficiently large vocabulary. We now recognise that much of our 'vocabulary' consists of prefabricated chunks of different kinds. The single most important kind of chunk is collocation. Self-evidently, then, teaching collocation should be a top priority in every language course.

However, unlike other phraseological units (e.g. idioms and proverbs), collocations are usually not emphasized in language courses, so students do not notice and assimilate them as complex lexemes (Bini et al., 2007).



1. Teaching collocations

In Italian L2 syllabuses and profiles collocations do not have much space:

- *Profilo della lingua italiana* (Spinelli & Parizzi, 2010): lexical lists of single words in alphabetical order for levels A1-B2,
- *Sillabo di riferimento per i livelli di competenza in Italiano L2 A1-B2* (AA. VV., 2011): a non-exhaustive list of words is placed alongside each semantic area (for example, famiglia, 'family': padre, 'father'; madre 'mother'; etc.).

More space in ***Dizionario delle collocazioni italiane per apprendenti*** (DICI-A, cf. Spina, 2016): a corpus-based dictionary of Italian collocations specifically targeted to learners. Collocations were extracted from the *Perugia corpus* (PEC, cf. Spina, 2014), ordered by their coefficient of usage (frequency + dispersion through textual genres), assigned to a the beginner proficiency level (A) taking into account also the topic they address.



1. Teaching collocations

-> If every language is formulaic in nature and acquiring collocations is particularly useful for learners, it is crucial to draw their attention not only towards single words, but towards combinations of words, aiming to develop collocational competence.

A syllabus of Italian collocations could therefore be a useful resource for Italian L2 teachers.



The top right corner of the slide features a decorative arrangement of overlapping geometric shapes. It includes a dark pink square, a medium pink square, and a light pink square, all partially overlapping each other and the main background.

Designing a corpus-based syllabus of Italian collocations

2. Designing a corpus-based syllabus of Italian collocations

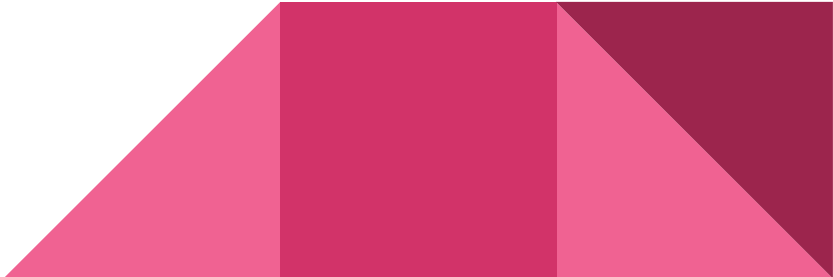
A. Content selection

Following the model of the *English Vocabulary Profile* (EVP), collocations were extracted from a **learner corpus** that provides reliable data on learners' authentic use of the language and shows direct evidence of when collocations are used by learners.

The **CELI corpus** (Spina *et al.*, 2022) is a balanced pseudo-longitudinal corpus that collects 3041 written texts produced by learners of Italian L2 who attended the *Certificati di Lingua Italiana* (CELI) exams (levels B1, B2, C1, C2) administered by the University for Foreigners of Perugia.

The main corpus is made up of four sub-corpora. Each sub-corpus collects the written productions corresponding to a given level.

The automatic extraction from the corpus involved 3 Part Of Speech (POS) sequences:

- noun-adjective: *sistema operativo*, 'operating system';
 - verb-adverb: *tornare indietro*, 'go back';
 - **verb-noun: *prendere una decisione*, 'make a decision'.**
- 

2. Designing a corpus-based syllabus of Italian collocations

A. Content selection

The list of the automatically extracted verb-noun also included some combinations that had **to be removed** with a further screening:

- **To remove free combinations** (e.g. *cercare televisione*, 'look for television') -> **Pointwise Mutual Information (PMI)**, a measure of collocational strength that brings out combinations made up of closely associated words. Since a PMI score of 3 or above generally indicates a significant collocation threshold (Hunston, 2002; Stubbs, 1995), all the collocations with a PMI score below 3 were removed from our initial list.
- **To remove non target-like combinations** (e.g. **utilizzare attenzione*, '*use attention', instead of *prestare attenzione*, 'pay attention') -> **coefficient of usage** (frequency + dispersion) in native speakers' productions. Following the model of the DICL- A (Spina, 2016), all the collocations with a coefficient of usage below 2 were removed from the list.

2. Designing a corpus-based syllabus of Italian collocations

A. Content selection

- **Final judgment made by 5 linguists:** based on the idea that conventional combinations are often extremely useful for L2 learners, it was decided to keep in the final list not only pure collocations but also some word combinations that are highly conventional, such as *aprire la porta*, 'open the door'; *chiudere la porta*, 'close the door'; *lavare i piatti*, 'wash the dishes', etc.

-> Final syllabus list: 952 collocations and highly conventional combinations.



2. Designing a corpus-based syllabus of Italian collocations

B. Compiling methods: procedure

To assign the collocations in the final list to the CEFR levels from B1 to C2, several criteria have been adopted.

Coefficient of usage in native speakers' production	Check if the collocation belongs to the high, medium or low frequency band: <ul style="list-style-type: none">• collocations in the high frequency band should be assigned to level B1 or level B2;• collocations in the medium frequency band should be assigned to level B2 or level C1;• collocations in the low frequency band should be assigned to level C1 or C2.
Number of occurrences in the CELI subcorpora	Between the two proficiency levels indicated by the frequency band, assign the collocation to the level in which it occurs more often.
Italian Profile word lists	When criteria 1 and 2 give contrasting information, check the Italian Profile lexical lists and assign the collocation to the level to which the words that make up the collocation belong.
Topic	Double check if the collocations assigned to a given proficiency level address topics that are relevant to that level.

2. Designing a corpus-based syllabus of italian collocations

B. Compiling methods: procedure

Some examples:

- **trovare lavoro, 'find a job'**: high frequency band -> should be assigned to level B1 or B2. Used 58 times at level B1 and 39 times at level B2 -> assigned to **B1**;
- **avere diritto, 'have right to'**: high frequency band -> should be assigned to level B1 or B2. never used at level B1; used 5 times at level B2; 40 times at level C1; 20 times at C2. The word *diritto* does not appear in the lexical lists of the Italian Profile, we can assume that it is learned at an advanced level. Topic:socio-political structures, more relevant for C level learners ->assigned to **C1**;
- **visitare città, 'visit a city'**: medium frequency band ->should be assigned to level B2 or C1. Used 12 times at level B1, 6 times at level B2 and 7 times at level C1. *Visitare* and *città* belong to the A1 lexical list of the Italian Profile. Topic relates to travel and everyday life ->assigned to **B1**.

Final results and examples

3. Final results and examples

Final result

952 collocations organized according to :

- **proficiency level:**
 - B1:221 collocations;
 - B2: 369 collocations;
 - C1:299 to level C1;
 - C2: 63 to level C2.
- **topic:** all the collocations were distributed among 70 topics. It is possible to search all the collocations related to a specific topic.



Spesa - Prezzi e strumenti di pagamento
Expenses - Prices and payment instruments

Collocation	Proficiency level
pagare prezzo (pay the price)	B1
abbassare prezzo (lower the price)	B2
aumentare prezzo (raise the price)	B2
fare soldo (make money)	B2
mantenere famiglia (feed a family)	B2
pagare affitto (pay rent)	B2
pagare bolletta (pay the bills)	B2
pagare tassa (pay taxes)	B2
risparmiare soldo (save money)	B2
spendere soldo (spend money)	B2
buttare soldo (waste money)	C1

The top right corner of the slide features a decorative arrangement of overlapping geometric shapes. These include a dark pink square, a medium pink square, and a light pink square, all partially overlapping each other and the main pink background.


Conclusions,
limitations and
future directions

4. Conclusions, limitations and future directions

Limitations:

- The **topic of the exam tasks** affects the type of collocations produced. The CELI exam includes tasks on a wide range of topics but some topics are inevitably absent while others are overrepresented and so are the collocations related to those topics.
- **Absence of A levels** due to the fact that the CELI corpus collects productions corresponding to levels from B1 to C2. This limit could be overcome in the future creating a reference corpus for the initial levels.

Nonetheless, we hope that our syllabus and the methodology adopted for its creation can be a starting point and a model for the creation of new syllabuses that include collocations belonging to other syntactic patterns or other types of phraseological units.



Thank you for your attention!

Francesca La Russa, francesca.larussa@uniroma1.it
Maria Roccaforte, maria.roccaforte@uniroma1.it

Acknowledgments

The research leading to these results received funding from PRIN-PHRAME Project (20178XXKFY) Phraseological Complexity Measures in learner Italian-Integrating eye-tracking, computational, and learner corpus methods to develop second language pedagogical resources.

References

- Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, 26, 28-41.
- Bini M., Pernas A. & Pernas, P. (2007), Apprendimento e insegnamento collocazioni dell'italiano. Con i NUNC più facile, in M. Barbera, E. Corino & C. Onesti (a cura di), *Corpora e linguistica in rete*, pp. 323–333. Perugia, Guerra Edizioni.
- Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Strasbourg: Council of Europe Publishing.
- Durrant, P. & Schmitt, N. (2009). *To what extent do native and non-native writers make use of collocations?*, 47(2), 157-177.
- Henriksen B. (2013). Research on L2 learners' collocational competence and development-a progress report, in L2 vocabulary acquisition, knowledge and use. New perspectives on assessment and corpus analysis. *Eurosla Monographs Series*, 2, 29-56.
- Hunston, S. (2022). *Corpora in applied linguistics*. Cambridge, Cambridge University Press.
- Ježek, E. (2016). *The lexicon: An introduction*. Oxford: Oxford University Press.
- Lewis M. (Ed.). (2000). *Teaching collocation: Further development in the lexical approach*. Oxford: Oxford University Press.
- Nesselhauf, N. (2005) *Collocations in a Learner Corpus*. Amsterdam: Benjamins.
- Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, 11(2), 285-301.
- Siyanova-Chanturia, A., & Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning*, 70(2), 420-463.
- Spina, S. (2014). The dictionary of Italian collocations: Design and integration in an online learning environment. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 3202-3208.
- Spina, S. (2016). Learner corpus research and phraseology in Italian as a second language: The case of the DICI-A, a learner dictionary of Italian collocations. *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*, 219-244.
- Spina, S., Fioravanti, I., Forti, L., Santucci, V., Scerra, A. & Zanda, F. (2022). Il corpus CELI: una nuova risorsa per studiare l'acquisizione dell'italiano L2. *Italiano LinguaDue*, 14(1), 116-138
- Spinelli B. & Parizzi F. (2010), *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2*, Milano, La Nuova Italia
- Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of language*, 2(1), 23-55.
- Wang, C. (2016). Phraseological Analysis of the Problem-Solution Pattern in EFL Learners' Persuasive Speech Writing. *Proceedings of The Fifth Northeast Asia International Symposium on Language, Literature and Translation*, 702-.
- Yoon, H. J. (2016). Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. *Journal of Second Language Writing*, 34, 42-57.