The ICoN Corpus of Academic Written Italian (L1 and L2)

Federica Cominetti, Mirko Tavosanis
Università di Firenze, Università di Pisa
Piazza di San Marco, 4, Firenze, Italy, via Santa Maria 36, Pisa, Italy
fedcominetti@gmail.com, mirko.tavosanis@gmail.com

Abstract

This paper describes the ICoN corpus, a corpus of academic written Italian, some of the directions of research it could open, and some of the first outcomes of research conducted on it. The ICoN corpus includes 2,115,000 tokens written by students having Italian as L2 students (level B2 or higher) and 1,769,000 tokens written by students having Italian as L1; this makes it the largest corpus of its kind. The texts included in the corpus come from the online examinations taken by 787 different students for the ICoN Degree Program in Italian Language and Culture for foreign students and Italian citizens residing abroad. The texts were produced by students having 41 different L1s, and 18 different L1s are represented in the corpus by more than 20,000 tokens. The corpus is encoded in XML files; it can be freely queried online and it is available upon request for research purposes. The paper includes the discussion of preliminary research in the field of collocations, showing that, in the texts included in the corpus, while learners and natives do use multiword expressions in a similar way, learners can overuse relatively infrequent forms of multiword adverbials, or use some adverbials in a non-standard way.

Keywords: Corpus linguistics, Learners, Italian, multiword expressions, collocations

1. Introduction

Learner corpora are useful and widely used in the study of language and language learning. In this paper we describe the building of the largest (to date) learner corpus of academic written Italian as L2 and of a large comparable corpus of academic written Italian as L1, collectively called "The ICoN corpus", already partially described in Tavosanis (2014) and (2016). The corpus can be queried online while its complete contents are available on request for research purposes.[1]

2. The Context of Learner Corpora Research and the Status of Learner Corpora for Written Italian

Being a learner corpus, the ICoN corpus belongs to a class of corpora well understood in theory and much appreciated for its practical uses (for an overview, see Granger, Gilquin and Meunier 2015). Learner corpora today number in the hundreds: the list of learner corpora maintained by Magali Paquot at the Université Catholique de Louvain includes to date 166 entries, while the Learner Corpus Bibliography maintained by the Learner Corpus Association (at the address www.learnercorpusassociation.org/resources/lcb/) includes more than 1,500 entries.

In the research tradition of the field, "in terms of medium and text type, the dominant focus was – and to a large extent still is – on writing, in particular essay writing" (Granger - Gilquin -Meunier 2015, 2). Well known examples of this include for example the ICLE (International Corpus of Learner English) corpus. This corpus (described in detail in Granger, Dagneaux, Meunier, Paquot, 2009) includes 2.5 millions of words coming from 3,640 unabridged essays, with an average length of 700 words, written by students with 11 different L1s. The essays do cover a variety of topics but for the most part are examples of argumentative writing and general use of English (rather than of English for specific purposes).

The ICoN Corpus, as will be described in detail in the following paragraphs, follows closely this tradition and, at least in size and composition, it can be roughly compared to the ICLE corpus. However, similar corpora are still relatively rare in the study of Italian as FL / L2. The most used resources of this kind are arguably the twin corpora VALICO (Varietà di Apprendimento della Lingua Italiana Corpus Online) and VINCA (Varietà di Italiano di Nativi Corpus Appaiato), realized at the University of Torino (Barbera - Marello 2004). Those corpora do include productions elicitated with similar techniques from FL / L2 learners (VALICO) and mother tongue speakers of Italian (VINCA). In both contexts, informants were mainly university students, with levels of L2 located for the most part in the A2-B1 range. VALICO includes 3,804 texts while VINCA has 680 texts (Allora – Colombo - Marello 2011), for a minimum size of 100 words; the web site of the project currently declares an overall size of 568,000 tokens for VALICO. This twin-corpora approach has a self-evident usefulness for comparison between L1 and L2 productions and it was followed in the building of the ICoN corpus.

Among other learner corpora of written Italian, in addition to a wealth of spoken corpora, we would also like to cite: ISA (Italiano Scritto di Americani), cf. Rastelli 2006; ADIL2 (Archivio Digitale di Italiano L2), realized at the Università per Stranieri of Siena, 432,606 tokens of which two-third are spoken data, cf. Palermo 2009; CAIL2 (Corpus di Apprendenti di Italiano L2), realized at Università per Stranieri di Perugia, cf. Bratankova 2015 (see also Andorno – Rastelli 2009 for an overview of Italian research on the topic). Among those, however, the ICoN corpus is the only one specifically devoted to upper-intermediate and advanced varieties.

3. The ICoN Degree Program and the ICoN Examinations

The ICoN corpus has been assembled at the University of Pisa in close cooperation with ICoN (Italian Culture of the Net), a consortium composed of nineteen Italian

[1] The paper was written jointly. However, for attribution of value, we declare that paragraphs 1-5 and 7 were the work of Mirko Tavosanis and paragraphs 6 and 8-10 were the work of Federica Cominetti.

universities whose aim is the promotion of the Italian language and culture all over the world through e-learning technologies (Tavosanis 2003). Among other educational initiatives, ICoN offers a three-year degree program in Italian Language and Culture, reserved to foreign students and Italian citizens resident abroad. The corpus is composed by short essays composed by the students during the examinations.
Each exam of the ICoN degree program includes the writing of a short text on a specific topic relevant to some aspect of the Italian humanistic culture (e.g. "The Baroque restoration practices of Maratta and the nineteenth century restoration practices of Cavalcaselle: compare two different attitudes towards conservation of the work of art"; "Analyze the relationship between Petrarch and Humanism"; "Illustrate the concept of equivalence and its role in the method of interlinguistic confrontation"). The corpus ICoN collected all such texts written by the ICoN students from 2001, year of the start of the degree program, until 2014.
Examination procedures in all cases required students to choose one topic for the essay among three different given topics. The task had to be completed in less than 90 minutes, but during this time students also had to answer to 30 fixed-choice questions (inside the limit of 90 minutes, students could freely choose how much of their time give to the questions or to the composition of the text). The use of supporting materials, such as textbooks and notes, was forbidden and enforcement of this rule was trusted to proctors in the examination rooms, set in the residence countries of students.
As for the text, the reference length given to students was 300 words (corresponding to an average of around 2200 characters), with penalties for texts both too long and too short. Moreover, during the examination, the number of words was shown to the student by a counter inserted in the writing interface: not surprisingly, the average length of the corpus texts is of 302 tokens.

4. Composition of the ICoN Corpus

The corpus in its published form is composed of 12,556 essays written by 787 different students. A single student then produced 16 texts on average. The inclusion of texts written by students that dropped out of the program and the fact that many students were still completing the program at the time of corpus generation are the reasons why the average is below the 21 texts required by the degree course. On the other hand, some students took the same examination more than once, and therefore some students are credited in the corpus with more than 21 texts.
The total corpus tokens are 3,794,000, of which about 2,115,000 written by Italian L2 students and approximately 1,769,000 written by Italian L1 students. The sum of the two sub-corpora is higher than the total, because some students were registered as "bilinguals" and their work has been included in both corpora. However, only 16 students are eligible for this, corresponding to a total of only 90,000 tokens (2.37% of the total).
The texts included in the two subcorpora were composed by the two groups of students in identical circumstances. This makes it possible to compare the sub-groups according to the consolidated VALICO / VINCA model. In this perspective, two strengths of the ICoN corpus are undoubtedly:

• its size, far superior to the size of the other comparable corpora
• the fact that its contents have not been realized for linguistic purposes but represent actual and evaluated exams; since it was created in a real-life context outside of dedicated language instruction, this kind of output can been considered as somewhat more than "semi-natural" (in the sense of "a pedagogical task that is natural in the context of the language learning classroom": Granger 2015: 10).

As for the encoding, texts were exported and collected as XML files with UTF-8 encoding. Personal data were thoroughly anonymized by the ICoN consortium: birth date was given as year of birth; country of residence was the only information given regarding residence. The encoding was carried out including each text as plain text inside a <doc> element including as attribute values the information regarding the student, as in this sample:

```
<doc id = "pre2008_esamistudente_9489.xml"
idstudente="9489" risposta="RISPOSTA 4"
Area_Titolo_Studio="Linguistica"
Condizione_Lavorativa="In Altra Condizione"
Data_Nascita="1976" Lingue_Conosciute="Rus
Eng Pol"  Madrelingua="Cze" Naz="Rep. Ceca"
Posizione_Lavorativa="" Provincia=""
Sesso="F" Titolo_Studio="Laurea 1°
Livello">
```

The XML files generated in this way were then used both for direct search and as sources for the search engine with Web interface described below.

5. Level of Knowledge of Written Italian

Students enrolled in the ICoN degree program are supposed to have reached a level at least equal to B2. In many cases, their level is definitely higher, while in some rare cases we feel it would be more appropriate to classify their production ability at level B1. Texts included in the corpus are therefore relatively homogeneous productions. At the moment no attempt has been made to evaluate more accurately the level of real competence of each individual student.
An example of a full text is this (Greek L1):

Dialoghi con Leucò è una raccolta di poesie scritte da Pavese e pubblicata nel 1947.L'opera consiste nei 27 componimenti in forma di dialogo fra personaggi mitici.Probabilmente l'autore si ispirò dalla Tesalique Mitologie dell' autrice Paula Phillipson.Nell'opera traspare una teatralità intesa non come visione ma come forma dialettica, basata appunto nei dialoghi che si svolgono fra i vari personaggi a volte in maniera tragica a volte in maniera ironica. Nel centro d'ispirazione del poeta e la sacralità del mito inteso come quel sostratto culturale insostituibile che accomuna tutta l'umanità. Sono gli antichi miti della cultura greca che appartengono a un' era remotta ma che tornano nel nostro tempo con una ciclicità come il continuo alternarsi della vita e della morte, come simboli privilegiati della transcendenza

umana. In tutte le culture,e specialmente in quella classica, sono i miti che hanno dato conforto alle angoscie umane della morte e dei fenomeni inspiegabili.
Pavese e reduce di letture antropologiche, psicologiche ed etnologiche che nel doppoguerra lo avevano appassionato. Autori come Forbenius, Fraser, Levy-Bruhl, opere dello Jung e Levy Strauss. In questo aspetto molto lo influenzò anche la sua conoscenza con Bianca Garufi.Questi temi gia apparsi nella sua opera Feria d'Agosto diventano più marcati, mentre l'idea della donna, dea e belva insieme, riempie la concezione della femmina.La donna appare come forzza della terra e della fertilità,la stessa che da vita e morte,in tutta la sua matterialità e mai come una transcendenza.
L' autore considerò questi componimenti come la sua opera più completa, non a caso una coppia fu trovata accanto allo scrittore quanto morì suicida nel 1950.

Another factor to take into account is the frequency of imitation, sometimes almost literal, from the teaching materials used by the students. This phenomenon is obvious and natural in all circumstances; it is even more so in a situation in which both didactic materials and final papers take the form of written texts displayed on computer screens.

6. L1s of Students

Regarding geographic distribution, students graduated with ICoN come from more than 60 countries around the world. Their L1s reflect this variety: in the registration phase students have indicated in fact 41 different L1s, including Italian. The L1s whose speakers have produced more than 20,000 tokens in total are 18 (as shown in Table 1).

Language	Tokens
Spanish	704,643
Portuguese	233,275
Serbian and Croatian	173,279
Russian	147,605
Greek	128,307
Polish	122,878
German	68,430
Albanian	66,470
French	45,694
English	44,721
Rumanian	43,518
Bulgarian	35,165
Japanese	33,552
Latvian	25,652
Lithuanian	25,158
Hungarian	24,618
Turkish	23,797
Czech	21,568
Sum	1,968,330

Table 1: Tokens according to the L1s of students

Languages with fewer tokens are, in descending order, Arabic, Maltese, Luo, Macedonian, Finnish, Leonese, Slovenian, Ukrainian, Dutch, Bosnian, Estonian, Maithili, Faroese, Azerbaijani, Tagalog, Slovenian, Vietnamese, Urdu, Kohati Hindko (the latter, along with the Leonese, has no ISO 639-2 code).
Serbian and Croatian were merged, not only because most of the linguists consider Serbian and Croatian a single language, but also because in many cases the students themselves (both Serbians and Croats) stated that their mother tongue was “Serbian and Croatian”.
In addition to the above mentioned case of 16 students declaring to be bilinguals in Italian and another language, one student declared to be bilingual in Russian and Belarusian, one in Spanish and Catalan and one in Italian and Venetian dialect, while three native speakers of Maltese are bilingual with English. In all cases, speakers were counted as L1 for all the languages in which they had indicated this competence. In the currently on-line interface it is also possible to search in relation to these speakers also “Belarusian”, “Catalan” and “Venetian”, which (unlike Maltese) are not represented by speakers who have these languages as exclusive L1.

7. Publication and Web Interface

The corpus has been put online at the Web address http://corpusicon.fileli.unipi.it/ and it is hosted by the University of Pisa. Its current search interface, created by Net7 s.r.l., is shown in Figure 1.

Figure 1: The interface of the ICoN Corpus

In its current form, the interface allows to search for single words. The search can be filtered by L1, country of birth, country of residence, current nationality, sex, age (by age brackets), academic qualifications and area of academic qualifications. Results are displayed as shown in Figure 2.
It is planned to replace the current interface with a more flexible interface by the first half of 2018. In a further

step, the contents of the corpus could be POS-tagged to improve the quality of research.

Figure 2: Sample of search results

8. Preliminary Research on Collocations

The ICoN corpus has already been extensively used in our didactic practice at the University of Pisa. It has been also used in researching the use of Italian definite articles by learners (Cominetti - Tavosanis forthcoming) and the use of focalizing adverbs by learners (Roy - Tavosanis forthcoming).

A further research topic we are currently applying the ICoN corpus to regards the field of collocations (see among others Firth 1957, Sinclair 1991, Evert 2005, 2009, Masini 2009).

Collocations can be defined as frequent word combinations whose syntactic and semantic properties cannot be entirely derived from those of the components. Collocations can be characterized by different levels of cohesion, ranging from fully fixed idioms (i.e. *burn the midnight oil*) to lexical preferences in contexts where other paradigmatic choices are available (i.e. *brush teeth* is preferred, but *wash teeth* is acceptable).

In recent years, the study of collocations and combinations of words has proved to be of great interest not only to lexicology and lexicography, but also to their applications to other fields, including first and second language acquisition. In L2/FL research, the study of collocations allows addressing questions of great theoretical value, such as the Sinclairian dynamics between open choice principle and idiom principle. Learners' behavior has an important heuristic value in suggesting how the mental lexicon is stored, and how important should the formulaic language be considered in comparison to paradigmatic choices (cf. among others Granger 1998, Nesselhauf 2005, Bratankova 2015).The ICoN corpus thus provides an ideal resource, allowing the comparison between native (ITA sub-corpus) and non-native (STRA sub-corpus) productions.

Up to now, a qualitative analysis of the corpus has allowed detecting some aspects of language formulaicity particularly problematic to learners, such as the preposition selection (1), the support verb selection (2) and the paradigmatic fixity (3, 4):

(1) Il romanticismo, realismo e naturalismo dell' Ottocento vengono sostituiti *per* movimenti che si trovano piu' in sintonia colle nuove preocupazioni filosofiche di questo periodo (L1 Spanish).

'Romanticism, realism and naturalism of the nineteenth century have been substituted *from* movements that are more attuned to the new philosophical concerns of this age'.

(2) Monteverdi esordi' molto giovane con il Primo libro dei madrigali, il genere molto in voga nel Cinquecento, che lo *fece* famoso. (L1 Russian).

'Monteverdi made his debut very young with his first madrigal book, that *rendered* him famous'.

(3) Carducci *al suo turno* fece un ampio studio del autore e dopo molti anni di ricerca pubblicò la sua opera che conteneva dettagliate annotazioni del Canzoniere. (L1 Luo).

'Carducci *on his time* made a large study of the author and after many years of research published his work that included detailed notes of the Canzoniere'.

(4) La situazione porta a volte il carattere drastico per la gioventù che, spinta dal malessere, entra nei *cerchi criminali*. (L1 Russian).

'The situation carries sometimes the drastic character to the young that, driven by unease, enter the *crime circles*'.

In (1) the preposition *per* 'by' is selected instead of *da* 'from, by'. In (2) the support verb *rendere* 'make' is wrongly substituted by *fare* 'make, do'. In (3) and (4) respectively, the paradigmatic fixity of idioms *a sua volta* 'on his turn' and *giri criminali* 'crime rings' is violated, since the correct words *volta* and *giri* are substituted by synonyms *turno* and *cerchi*.

A quantitative analysis reveals that, if the three mentioned kinds of errors are relatively frequent, the specific examples prove to be rare. For example, the wrong expression *al suo turno*, found in (3), is attested only 4 times in the corpus, while the standard form *a sua volta* counts 93 occurrences in the L2 sub-corpus (a very similar figure is found in the L1 sub-corpus, where 114 occurrences are attested).

A more accurate statistical analysis on the topic has been conducted using EXTra (Passaro - Lenci 2016), a term recognition system that evaluates the association measure of structured POS-sequences. Both effect-size measures (i.e. Mutual information) and significance measures (i.e. t-score, log-likelihood) are used, so that the dimension of the corpus and the particularly high frequency of some tokens (like articles, prepositions) do not affect the results. In particular, the following kinds of multiword lexicon and word collocations have been investigated up to now: multiword nouns with structure Noun Prep Noun (i.e. *punto di vista* 'point of view') and Noun Adj (i.e. *essere umano* 'human being'), multiword verbs (Verb Noun, Verb Prep Noun, Verb Noun Adj) and adverbials (Prep Noun, Prep Adj Noun, Prep Noun Adj) and prepositional locations (Prep Noun Prep).

Such preliminary analysis has shown that the most frequent multiword expressions included in the L1 and L2 sub-corpora tend to be the same. See, for example, in Table 2 (cf. Annex), the list of the 20 multiword nouns (structure Noun Prep Noun and Noun Adj) characterized

by the highest LMI in L1 and L2 sub-corpora. Table 3 includes the adverbials characterized by the highest LMI.
In general, both in the case of nouns and adverbials, the two sub-corpora share the same forms. As for nouns, the datum that learners' behavior is not far from natives' is interesting and could not be easily assumed. As for adverbials, however, being the adverbials a grammatical class, the similarity of the L1 and L2 lists is less surprising. Actually, any significant difference can in this case highlight untypical or distorted uses. For example, in STRA, alongside the very cohesive *allo stesso tempo* 'at the same time', the synonym *nello stesso tempo* is found, while the latter does not appear in the ITA list. In fact, 117 occurrences of *nello stesso tempo* are found in STRA and only 32 in ITA, while *allo stesso tempo* is found 172 times in STRA and 151 in ITA. The ITA data are compatible with standard Italian as testified by a Google search: *nello stesso tempo* has 1,5 million occurrences while *allo stesso tempo* appears 13,3 million times. However, the same search on the CORIS corpus (Rossini Favretti 2000) finds 1771 occurrences for *nello stesso tempo* and just 1908 for *allo stesso tempo*, so it is necessary to be careful in drawing conclusions from rough data.
Another interesting example is provided by the adverbial *dall'altra parte* 'on the other side', quite frequent in STRA (87 occurrences, LMI 3088) but rare in ITA (8 occurrences, LMI < 1000). In this case, the surprisingly high frequency in STRA detects a mistake: the learners use *dall'altra parte* in place of the correct form *d'altra parte* 'after all, moreover' as it is clear in the following example:

(5) Allora potevano lavorare di più d'estate quando il tramonto del sole veniva molto più tardi dell'inverno e *dall'altra parte* potevano cominciare a lavorare molto più presto di mattina come il sole d'estate veniva molto più prima. (L1 Czech).

'Then they could work more in the summer when the sunset arrived much later than in winter and *on the other side* they could start working much earlier in the morning since the summer sun rises earlier'.

Moreover, in the STRA LMI list the adverbial *da un'altra parte* 'somewhere else' appears as well (25 occurrences), while it does not appear in ITA's list (only 1 occurrence). Also in this case, the surprisingly high frequency reveals a non-standard use by the learners: in fact, an analysis of the occurrences in STRA reveals that *da un'altra parte* is often used in contrast to *da una parte* 'on one side', in a context where natives normally use *dall'altra parte* 'on the other side'. In other cases, *da un'altra parte* is wrongly used instead of *da un altro lato* 'on the other side'.
In the mentioned examples, the data extracted from the two sub-corpora testify that ITA more accurately reflects the use of the adverbials typical of standard Italian, while in STRA two tendencies emerge: relatively infrequent forms can be overused, while some adverbials can be used in a non-standard way.

9. Future Work

To date, the corpus has been used for didactic purposes and for traditional research, especially for the use of Italian definite articles. It is expected that use will continue.
However, approaches with a stronger computational side are also envisaged. Preliminary results as those seen in § 8 are promisingly showing that collocations are one of the features – although subtle – that allow detecting non-native productions even in presence of an advanced L2 competence and native-like grammatical behaviors. For this reason, we are willing in particular to explore the possible application of this feature to the setting of tools for the task of Automatic Native Language Identification, which has already been tested to a limited extent on the corpus ICoN.

10. Acknowledgements

The corpus work was done within the Italian national research program PRIN "SCRIBE / Scritture Brevi" (2013-2016, coordinated by Pietro Trifone - University of Rome "Tor Vergata"). The ICoN Consortium has generously contributed to the realization of the corpus through the provision of materials and dedicated funding.

11. Bibliographical References

Allora, A. - Colombo, S. - Marello, C. (2011). I corpora VALICO e VINCA: stranieri e italiani alle prese con le stesse attività scritte. In Maraschio, M. - De Martino, D. (eds.) *La Piazza delle lingue. L'italiano degli altri*. Firenze, Accademia della Crusca, pp. 49-61

Andorno, C. - Rastelli, S. (2009). Corpora di italiano L2. Tecnologie, metodi, spunti teorici. Perugia, Guerra.

Barbera, M. - Marello, C. (2004). VALICO (Varietà di apprendimento della lingua italiana Corpus Online): una presentazione. *Itals* II, 4, pp. 7-18.

Blanchard, D. - Tetreault, J. - Higgins, D. – Cahill, A. – Chodorow, M. (2013). TOEFL11: A Corpus of Non-Native English, Educational Testing Service, Research Report ETS RR–13-24, http://www.ets.org/Media/Research/pdf/RR-13-24.pdf

Bratankova, L. (2015). Le collocazioni Verbo + Nome in apprendenti di italiano L2. Tesi di dottorato, Università per stranieri di Perugia.

Bruni, F. - Alfieri, G. - Fornasiero, S. - Tamiozzo Goldmann, S. (2006). Manuale di scrittura e comunicazione. Bologna, Zanichelli.

Cominetti, F. - Tavosanis, M. (forthcoming). Interferenza della L1 nell'apprendimento degli articoli in italiano L2: una ricerca sul corpus ICoN. In *Atti del convegno SILFI 2016*, Madrid.

Dell'Orletta, F. - Montemagni, S. - Venturi, G. (2011). Read-it: Assessing Readability of Italian texts with a View to Text Simplification. In *Proceedings of the 2nd Workshop on Speech and Language processing for Assistive Technologies*, Edinburgh, pp. 73-83.

Dell'Orletta, F. - Montemagni, S. - Venturi, G. (2016). Esplorazioni computazionali nello spazio dell'interlingua: verso una nuova metodologia di indagine. In *Atti del XLVIII Congresso Internazionale della Società di Linguistica Italiana*. Roma, Bulzoni, pp. 143-161.

Evert, S. (2005). The Statistics of Word Cooccurences. Word pairs and Collocations. Ph.D. Thesis, University of Struttgart.

Evert, S. (2009). Corpora and collocations. In Ludeling, A. - Kyto, M. (eds.) *Corpus Linguistics: An International Handbook*, Volume 2. Berlin, New York, de Gruyter, pp. 1212-1248.

Ferris, D. R. - Hedgcock, J. (2013) Teaching L2 Composition: Purpose, Process, and Practice. 3rd edition. Routledge.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-55. In *Studies in Linguistic Analysis, Philological Society*. Oxford, pp. 1-32, reprint in F. R. Palmer (ed.) (1968). Selected papers of J.R. Firth 1952-1959. Harlow, Longman, pp. 168-205.

Gilquin, G. (2015). From design to collection of learner corpora. In Granger, S. - Gilquin, G. - Meunier, F. (eds.) *The Cambridge Handbook of Learner Corpus Research*. Cambridge, Cambridge University Press, pp. 9-34.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In Cowie A. P. (ed.). *Phraseology: Theory, Analysis and Applications*. Oxford, Oxford University Press, pp. 145-160.

Granger, S. - Dagneaux, E. - Meunier, F- - Paquot, M. (2009). *International corpus of learner English*. Louvain, Presses universitaires de Louvain.

Granger, S. - Gilquin, G. - Meunier, F. (2015). Introduction: learner corpus research – past, present and future. In Granger, S. - Gilquin, G. - Meunier, F. (eds.) *The Cambridge Handbook of Learner Corpus Research*. Cambridge, Cambridge University Press, pp. 1-5.

Köpke, B., - Schmid, M. S., - Keijzer, M. - Dostert, S. (eds.) (2007). Language Attrition: theoretical perspectives. Amsterdam, John Benjamins.

Koppel, M., - Schler, J. – Zigdon, K. (2005). Determining an author's native language by mining a text for errors. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD '05)*, pp. 624-62.

Masini, F. (2009). Combinazioni di parole e parole sintagmatiche. In Lombardi Vallauri, E. - Mereu, L. (eds.). *Spazi linguistici. Studi in onore di Raffaele Simone*. Roma, Bulzoni, pp. 191-209.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. In *Applied Linguistics*, 24, pp. 223-242.

Nesselhauf, N. (2005). Collocations in a Learner Corpus, Amsterdam-Philadelphia, John Benjamins Publishing Company.

Palermo, M. (ed.) (2009). Percorsi e strategie di apprendimento dell'italiano lingua seconda: sondaggi su ADIL2. Perugia, Guerra Edizioni.

Passaro L. C. - Lenci A. (2015). Extracting terms with EXTra. In Proceedings of EUROPHRAS 2015. (pp. 188–196). Málaga (Spain), July 2015.

Rastelli, S. (2006). ISA 0.9. Written Italian of Americans: syntactic and semantic tagging of verbs in a learner corpus. *Studi Italiani di Linguistica teorica e Applicata* 1, pp. 73-99.

Rossini Favretti, R. (2000). Progettazione e costruzione di un corpus di italiano scritto: CORIS/CODIS. In Rossini Favretti R. (ed.), *Linguistica e informatica. Multimedialità, corpora e percorsi di apprendimento*. Bulzoni, Roma, pp. 39-56.

Roy, T. - Tavosanis, M. (forthcoming). Il focalizzatore *anche* nei testi scritti di studenti con lingue indoarie come L1. In *Atti del convegno SLI 2017*, Napoli.

Sinclair, J. M. (1991). Corpus, Concordance, Collocation. Oxford, Oxford University Press.

Spirito, R. (1999). La scrittura accademica. In Palloti G. (ed). *Scrivere per comunicare*. Milano, Bompiani, pp. 205-219.

Tavosanis, M. (2003). Insegnamento di lingua e cultura italiana a stranieri: l'esperienza di ICoN. In *Italiano e italiani nel mondo. Italiani all'estero e stranieri in Italia: identità linguistiche*. Roma, Bulzoni, vol. 2, pp. 1-13.

Tavosanis, M. (2014). Corpus ICoN: una raccolta di elaborati di italiano L2 prodotti in ambito universitario. In *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014*. Pisa, Pisa University Press, pp. 370-373.

Tavosanis, M. (2016). L'italiano L2 negli elaborati universitari del corpus ICoN: esempi di analisi. In Manco, A., - Mancini, A. (eds.), *Scritture brevi: segni, testi e contesti. Dalle iscrizioni antiche ai tweet*. Napoli, Università degli studi di Napoli «L'Orientale», pp. 177-187.

	STRA	LMI	ITA	LMI
1	punto di vista (1)	1928	punto di vista (1)	1213
2	punto di riferimento (6)	579	essere umano (3)	352
3	essere umano (2)	544	mezzo di comunicazione (5)	324
4	stato d'animo (4)	445	stato d'animo (4)	316
5	mezzo di comunicazione (3)	440	messa in scena (15)	311
6	movimento culturale (14)	365	punto di riferimento (2)	308
7	centro urbano (7)	356	centro urbano (7)	306
8	opera d'arte (8)	350	opera d'arte (8)	258
9	punto di partenza (12)	346	presa di coscienza (-)	241
10	classe sociale (-)	329	classe dirigente (11)	239
11	classe dirigente (10)	266	pena di morte (-)	216
12	tasso di natalità (17)	248	punto di partenza (9)	199
13	rito di passaggio (19)	226	diritto di voto (16)	174
14	posto di lavoro (-)	208	movimento culturale (6)	164
15	messa in scena (5)	202	sala cinematografica (-)	146
16	diritto di voto (13)	195	servizio di leva (-)	139
17	storia d'amore (18)	175	tasso di natalità (12)	136
18	capo del governo (-)	159	storia d'amore (17)	107
19	mezzo di trasporto (-)	154	rito di passaggio (13)	106
20	gioco di parole (-)	141	via d'uscita (-)	102

Table 2: Multiword nouns (Noun Prep Noun, Noun Adj) in STRA and ITA ordered by decreasing LMI.

	STRA	LMI	ITA	LMI
1	per la prima volta (1)	11383	per la prima volta (1)	8541
2	allo stesso tempo (2)	11217	allo stesso tempo (2)	6341
3	per esempio (3)	7516	ad esempio (4)	5933
4	nello stesso tempo (-)	5235	per esempio (3)	2946
5	ad esempio (3)	5124	al tempo stesso (-)	2704
6	dall'altra parte (-)	3088	in un certo senso (8)	1952
7	in gran parte (10)	2112	in primo luogo (-)	1940
8	in un certo senso (6)	2068	allo stesso modo (13)	1826
9	a volte (13)	1538	a livello internazionale (-)	1359
10	in primo piano (14)	1448	in gran parte (7)	1338
11	nel primo caso (-)	1336	in particolare (-)	1201
12	nello stesso modo (-)	1235	in modo particolare (-)	1176
13	allo stesso modo (8)	1109	a volte (9)	1108
14	al primo posto (-)	1105	in primo piano (10)	1091
15	da un'altra parte (-)	1048	in seguito (-)	1070

Table 3: Preposition-headed adverbials in STRA and ITA ordered by decreasing LMI.