UNIVERSITÀ CATTOLICA DEL SACRO CUORE

SEDE DI MILANO

DOTTORATO DI RICERCA IN SCIENZE LINGUISTICHE E LETTERARIE

CICLO XXXIII

S.S.D. L-LIN/12

Adverb + adjective collocations in a spoken learner corpus: A quantitative and qualitative approach

Coordinatore:

Ch.ma Prof.ssa Anna Bonola

Tutor:

Ch.ma Prof.ssa Amanda Murphy

Co-tutor:

Ch.mo Prof. Adriano Ferraresi

Tesi di Dottorato di:

Francesca Poli

N. Matricola: 4713478

Anno Accademico 2020/2021

To my Mum, for all the love you have given me.

Acknowledgements

I would like to thank Prof. Amanda Murphy for the opportunity to do a PhD under her supervision, for her guidance, and for offering me precious insight throughout my journey. Her immense knowledge of Linguistics and unwavering enthusiasm for research have kept me motivated and engaged to my project.

A sincere thank you goes to Prof. Adriano Ferraresi, without whom I would still be lost meandering in R and flickering through statistics books in a panic. I am especially grateful for the time he dedicated to my work and his infinite patience.

Prof. Anna Bonola and Prof. Dante Liano, the Phd school coordinators, have always been very helpful and understanding with us students, so thank you for supporting us and making our life easier! Also, thank you for taking the time to organise interesting meetings and lectures with scholars from all around the world. It was much appreciated! And of course, I would not be writing my acknowledgments if it were not for Simona. Thank you for explaining all the nuts and bolts of the school and helping out with bureaucracy. You have been an invaluable source of help and information. Thank you.

I also wish to thank my university, Università Cattolica del Sacro Cuore, for welcoming me as their student once again and for granting me the scholarship. It has been a pleasure to experience the historical campus in Milan and walk down its corridors and cloisters.

I would like to thank all the colleagues that helped out during these three years. You were really helpful and my work could not have been done without you. Thank you Chiara and Costanza for listening to my rants and being patient, sometimes I forgot you also had a lot of work to do.

The PhD school was a fun place to be, despite the struggles and tight deadlines, but I could always find support and moments of good fun with my fellow PhD students. I met Vincenzo, Elena C., Elena P. during my interview in Milan and I was immediately struck by their brightness and wit. We shared many fun moments during these three years, both in person and virtually (especially in the last year). You were my source of strength and, above all, my information hub! Countless forgetful moments were promptly saved by your punctual reminders. Many thanks go to Andrea, who shared my concerns about statistics and supported me during my R frenzies. I could not have done this without all your precious

advice and insights into Lyx.

I would also like to thank my reviewers, Prof. Pierfranca Forchini and Prof. Agnieszka Leńko-Szymańska, who were so kind to accept this burdensome, to say the least, task! I thank you for all the incredibly detailed comments you provided which I believe have made my thesis so much better. Thank you for taking the time to do it, thank you for staying up late to meet the deadline despite all your other impending and much more important deadlines! I really appreciated all your insights, it was a true learning experience.

I dedicate another thankful paragraph to all my friends, you are many and scattered all around the world, Marijke, Marco, Sim, Alice A., Charis, Alice C., Dalila my life saviour and all the other tutor colleagues Giulia, Anna and Vale, Gaby, Luciana, Federica, Beppe, Luca, and all the other ones I am fogetting to mention, you have all been so precious!

I must thank Linda on her own because of her incredible patience and support, you have been so kind and helpful, I apologise for all the rants and emergency crises I threw at you. Thank you for patiently waiting it out and letting me fume over a glass of wine! It has been a rollercoaster ride, but I could not have done it without you.

Thank you to Alessandra and Serafino for their incredible support; Serafino opened my eyes and let me in a world I was not even aware of and for that I thank you. Alessandra, thank you so much for the time you dedicated to me, your support, and your encouraging words. I hope I will be able to return the favour in due time.

I must also acknowledge Stefano and all his help, although it was not strictly Phd related, I feel like you gave me the confidence to accomplish this and many other things! Especially those involving flying tubes.

Last, but certainly not the least, I shall thank my Mother. You have been my rock for so many years now, I do not even know how I could have done this without you. The bond between mother and daughter is a special one, but I feel like ours is even stronger. It is the clichéd love and hate relationship, but it is so much more than that. It is a deeply rooted love and understanding, it is the knowledge and awareness that there is always going to be someone there for you. There are so many things you did for me, things I was not even aware you were doing for me. Of course, there was the standard cooking, cleaning, doing the laundry, food shopping, and so much more which I could not have done because I was stuck at the computer, in my "bunker" all day. Every day you lovingly prepared a meal or washed my clothes because my well-being was more important than your tiredness or your aches. But you also listened to me when you already knew how the story was going to end, you told me what I wanted to hear at the right time, you hugged me even though you were the one needing a hug, you ignored my snappy responses, you indulged my weird quirks, you did so much because of your love for me. I have spent my whole life studying

and learning (lifelong learning, if you know what I mean), but you have been my best and hardest learning experience. I still have so much to learn, I still make mistakes, I still am a grumpy daughter, but if I can say I have learnt something, it is because of you. This work is dedicated to you, because I wish to make you proud and see you smile, and because, after all, I would not be here today if it had not been for you.

Contents

List of Figures

List of Tables

List of abbreviations

BNC	British National Corpus
CA	Contrastive Analysis
CEA	Computer-aided Error Analysis
CECL	Centre for English Corpus Linguistics
CIA	Contrastive Interlanguage Analysis
CIA^2	Reappraised Contrastive Interlanguage Analysis
EA	Error Analysis
EFL	English as a Foreign Language
ELF	English as a Lingua Franca
FLT	Foreign Language Teaching
ICLE	International Corpus of Learner English
IL	Interlanguage
ISLC	Italian Spoken Learner Corpus
LCR	Learner Corpus Research
LINDSEI	Louvain International Database of Spoken English
LOCNEC	Louvain Corpus of Native English Conversation
OED	Oxford English Dictionary
OOPT	Oxford Online Placement Test
PEC	Perugia Corpus

SLA	Second language acquisition
TL	Target language
TLC	Trinity Learner Corpus
WFF	Word Family Framework
WTC	Willingness to Communicate

Chapter 1

Introduction

THIS thesis investigates *adverb* + *adjective* collocations produced by advanced Italian learners of English, both from a quantitative and qualitative perspective. The field of study is learner corpus research (LCR) on English L2. The aim of the research is to investigate whether there are any differences or similarities in terms of adverb + adjective collocations between a corpus of advanced Italian learners of English and a corpus of native-speaker university students. It is hoped that the identification of elements in common and discrepancies between the two corpora will inform university teaching staff and learners about both the gaps to be filled in terms of collocations acquisition and performance, and the areas of strength of the learners.

1.1 Collocations

In recent years, there has been an increase in studies on collocations which have stemmed from the Firthian tradition of looking for lexical combinations which occur more frequently than expected (Firth, 1957; Hoey, 2005; Moon, 1998b; Sinclair, 1991; 2004; Stubbs, 1996; 2001). The results of such studies have provided evidence that phraseology is pervasive to language (Altenberg, 1998; Biber et al., 1999; Cowie, 1991; 1992; Howarth, 1998a). This indicates that good command of collocations is needed if learners aim to achieve a native-like fluency in the L2.

Indeed, research on the learner and native-speaker production of formulaic language has highlighted that collocations play a central role in the acquisition of first and second languages and adult language production (Cowie, 1998; Pawley & Syder, 1983; Peters, 1983), and they are an essential component for the development of fluency (Ellis, 2002; 2003; Ellis et al., 2015; Howarth, 1998a). In particular, Pawley and Syder (1983, p. 191) posited that "fluent and idiomatic control of a language rests to a considerable extent on knowledge

of a body of 'sentence stems' which are 'institutionalised' or 'lexicalised'". Thus, mastery of collocations can facilitate a more native-like idiomatic production and it can also improve language comprehension. Failure to use native-like expressions may "divert the reader's attention from content to form" (Howarth, 1998a, p. 174) and create "an impression of brusqueness, disrespect or arrogance" (Wray, 2002, p. 143). However, despite the widely verified importance of collocations and formulaic language for native-like fluency in learners, mastery of lexical combinations is still one of the most difficult tasks even for advanced learners (Pawley & Syder, 1983). Indeed, a growing body of research has revealed that collocation learning is a hurdle for non-native speakers even at fairly advanced proficiency levels. Studies in this field usually depict L2 learners' collocational difficulties and the main findings of L2 collocation research have highlighted that learners possess a better receptive knowledge of collocations (Biskup, 1990; Gyllstad, 2005) compared to their production. Learners are found to operate more on the open-choice principle rather than the idiom principle (Sinclair, 1991) and use fewer collocations compared with native speakers. In addition, overuse and misuse have also been frequently reported in learners' writings (Bestgen & Granger, 2014; Granger, 1998a; Granger & Bestgen, 2014; Howarth, 2013; Laufer & Waldman, 2011; Nesselhauf, 2005).

Given the difficulties learners encounter with collocational knowledge and production, studies into learners' collocations in L2 are and should continue to be one of the primary concerns of second language learning and teaching. Furthermore, since most of the research has been conducted on written texts, future studies should focus more on spoken language as it is the more direct reflection of the learners' interlanguage[1] (IL) (Myles, 2015).

Therefore, given the importance of phraseological knowledge both in language production and comprehension, the struggle to acquire it even for advanced L2 learners, and spoken language being the litmus test for IL, the present work aims to fill the gaps by examining learners' collocational production in spoken language.

1.2 Scope of the thesis

The majority of second language acquisition (SLA) and LCR studies have focused on discovering whether learners process formulaic language similarly to native speakers, possess

[1]Interlanguage has been defined by Selinker (1972, pp. 213–214) as "the utterances which are produced when the learner attempts to say sentences of a TL [target language]". Since the utterances produced by the learners do not correspond to the target utterances produced by a native speaker of a given L2, there is a "separate linguistic system based on the observable output which results from a learner's attempted production of a TL norm" (Selinker, 1972, p. 214). Therefore, IL is the linguistic system found at the intersection between a learner's L1 and target L2.

their same level of knowledge of phraseology, and whether learners struggle with particular phraseological structures. This has implied the use of both elicited and introspective data, as well as learner corpora; however, especially as regards learner corpora, most research has been conducted on written texts, rather than spoken ones. Thus, little is known as regards learners' collocation performance in spoken language. This study is intended to investigate the use of adverb + adjective collocations by advanced learners of English in spoken language. This work has chosen to analyse collocations from a frequency-based perspective; for this reason, by adopting a *surface co-occurrence* approach (Evert, 2009) association measures have been calculated in order to measure the strength of attraction of collocations. In particular, the choice fell on adverb + adjective collocations for two main reasons: the first is strictly related to the surface co-occurrence approach which entails choosing a span which can range from ± 1 to whatever is chosen by the researcher. However, the choice of a small span which selects adjacent lexical items results in improved performance of statistical measures and it facilitates extraction from the corpus. Secondly, adverb + adjective collocations are very frequent structures (and thus are easily retrievable from a corpus), and little work has been done on intensified adjective collocations (Altenberg, 1991; Granger, 1998b; Granger & Bestgen, 2014; Lorenz, 1999). Furthermore, these collocations are worth studying since intensification is an intricate part of foreign language learning (Lorenz, 1999, p. 26) and, as such, more research work should be dedicated to it. Having outlined the focus of the thesis, the following are the overarching research questions:

1. How do advanced Italian learners of English and native speakers of English compare in their production of different adverb + adjective collocations in spoken language?
2. What is the difference between the syntactic patterns and lexical meaning of the ad-

verb + adjective collocations[2] produced in spoken language by advanced Italian learners of English and native speakers?

3. Does L1 congruency have a transfer effect on the production of infrequent/unattested adverb + adjective collocations generated in spoken language by advanced Italian learners of English?

1.3 The methodology of the research

This work adopts a corpus-based approach to the study of learner collocations in spoken English L2 from a Contrastive Interlanguage Analysis (CIA) perspective. The corpora used for the research are a newly compiled local spoken learner corpus, the *Italian Spoken Learner Corpus* (ISLC), and the well-known sister corpus of the *Louvain International Database of*

[2]In this thesis, the term *lexico-grammatical patterns* will be adopted to refer to the syntactic and lexical patterns of the collocations analysed. The choice of the term stems from the tradition of pattern grammar which can be traced back to Hornby (1954). In the preface to his book, *A Guide to Patterns and Usage in English*, Hornby set out his ground-breaking objective:

> Analysis is helpful, but the learner is, or should be, more concerned with sentence-building. For this he needs to know the patterns of English sentences and to be told which words enter into which patterns. (p. v)

Thus, Hornby's concern was with patterns rather than a grammatical analysis of the sentence; in his description of language, there is a very fine line between lexis and grammar where there once was a clear distinction between the two (Hunston & Francis, 2000, p. 28). This anticipates Sinclair's (1987; 1991) recognition that the two domains are not distinct. They could only be so if language was only based on the open-choice principle (Sinclair, 1987, p. 114), but that would cause an enormous amount of anomalous language. This theory was later reprised by Francis (1993; 1995) who argued that:

> Particular syntactic structures tend to co-occur with particular lexical items, and – the other side of the coin – lexical items seem to occur in a particular range of structures. In short, syntax and lexis are co-selected, and we cannot look at either of them in isolation. (Francis, 1995, p. xx).

In this regard, corpus linguistics has demonstrated that lexis and grammar interact and are inseparable (Francis, 1993, p. 142-146). Collocations are one of the direct manifestations of lexico-grammar as the two domains are intertwined and some co-occurrences may simply be preferential while others are imposed on the words according to the collocates/nodes (Cosme & Gilquin, 2008, p. 260).

As regards learner language, it has been noted that the lexico-grammatical patterns of words are a source of difficulties for learners (Flowerdew, 2001). For example, Gilquin (2008) found that native students perform better than non-native students as regards the lexico-grammar and phraseology of academic writing. In the LCR field, the use of corpora and contrastive approaches to the analysis of learner language has uncovered a wide range of lexico-grammar patterns (Granger, 2003, p. 542).

In conclusion, given that collocations embody the two domains of grammar and lexis in that one lexical choice may influence a syntactic pattern and viceversa, it seems apt to adopt the term *lexico-grammatical patterns* to refer to the both the syntactic and lexical patterns of the collocations analysed. This thesis will investigate whether there are any differences on the syntagmatic level between the native and non-native speakers' collocations and whether there are also different lexical aspects which may or may not be directly linked to the syntax.

Spoken English Interlanguage (LINDSEI), the *Louvain Corpus of Native English Conversation* (LOCNEC). A further reference corpus was used as the basis for the computation of the association measures scores of the collocations: the British National Corpus (BNC). The standard text retrieval software used for the extraction of the collocations is AntConc (Anthony, 2019).

In order to address the overarching research questions, the present work examines the collocations from a two-fold perspective: the quantitative approach can better inform whether learners tend to overuse or underuse adverb + adjective collocations. The concept of overuse and underuse has amply been described and used in the literature and, despite some of the criticism it has received, it is still a valuable concept to be reported, especially for teaching staff and in English as a Foreign Language (EFL) contexts. Indeed, although the terms overuse and underuse may be interpreted as a prescriptive approach to the analysis of learner language, these are merely used to identify and describe patterns of language use by language learners. The qualitative perspective instead provides a more fine-grained approach to the analysis of the different adverb + adjective collocations since their individual lexico-grammatical patterns (of both types and tokens) are examined and discussed. This can be helpful to teaching staff and learners themselves to better understand instances of native-like use and misuse: by consulting a detailed description of the grammatical patterns of collocations, teachers and learners can be better informed on the areas of strength or weakness which require consolidation or further studies.

1.4 Structure of the thesis

Chapter 2 introduces the framework within which this thesis is set, namely that of LCR. The Chapter provides the background of LCR, from early studies in SLA to the development of computer learner corpora. Furthermore, a description of the main approaches to LCR are provided, namely Computer-aided Error Analysis (CEA) and CIA. Chapter 3 reviews the literature on phraseology starting from the multiple definitions of formulaic language and the different approaches to the identification of formulaic sequences. Then, the chapter focuses on the two main approaches to the study of collocations, that is, the phraseological and the frequency-based approach. The last section is dedicated to literature findings on learner collocations. Chapter 4 reports on the research aims and presents the rationale for the corpus-based quantitative and qualitative analyses, as well as the research questions. In particular, the chapter illustrates the compilation of the ISLC by providing corpus design and criteria; then, it discusses the extractions of the collocations from the corpus, and lastly,

it explores the main metadata as regards the learners and some corpus descriptive information. Chapter 5 presents the results of the quantitative analysis: four sub-analyses were carried out in order to address the first research question (*How do advanced Italian learners of English and native speakers of English compare in their production of adverb + adjective collocations in spoken language?*) starting from the number of combinations produced on average by individual speakers, the number of infrequent/unattested collocations produced on average by individual speakers, the number of grey area collocations produced on average by individual speakers, and a comparison between the average association measure scores between learners and native speakers. Chapter 6 deals with the qualitative analysis which consists in investigating the lexico-grammatical patterns of 11 frequent collocations, while Chapter 7 analyses nine frequent infrequent/unattested collocations for L1 transfer effects. Both sets of collocations are analysed within their context and inferences are made as regards their use by the learners of the ISLC. In addition, the rationale to the choice of these specific sets of collocations is given. Chapter 8 summarises the findings of the research chapters, discusses the contribution this thesis has made in learner corpora studies, and addresses some pedagogical implications. Finally, the limitations of the research are presented and future learner language studies are proposed.

Chapter 2

Learner Corpus Research

THIS chapter provides a review of the research framework within which the present study is set, LCR. The first Section focuses on the early approaches to learner language, namely those of SLA theory and the developments that led to the widespread use of learner corpora; the second Section introduces key points regarding the collection methods for the investigation of learner language, such as experimental data, introspective data, and natural language data; the third Section dives deeper into the field of LCR and introduces the main elements characterising the compilation of a learner corpus, its variables, the typology, and some core issues; the fourth Section focuses on the methodological approaches to the analysis of learner corpora, namely those of CIA and CEA; the fifth Section presents the long-standing debate surrounding the likelihood of LCR being able to bridge the gap between learner corpora and SLA.

2.1 Early approaches to the study of learner language

A learner corpus is an "electronic collection(s) of natural or near-natural data produced by foreign or second language (L2) learners" (Granger et al., 2015, p. 1). In other words, it is a collection of foreign language learner's IL. Learner corpora are compiled in order to carry out two main functions: to contribute to SLA research with better descriptions of IL and provide further data for its understanding, and to help the creation of new teaching tools and practices to target the learners' needs (Granger, 2008, p. 259). Research which employs learner corpora for their investigation falls under the broad term LCR. LCR usually adopts two main methodologies for the study of learner corpora, CIA and CEA. CIA consists in a comparison between L1 and L2, or between ILs, whereas CEA requires the annotation of the learner corpus for errors (cf. Section 2.4.3). Both these methodologies require the aid of computers for the extraction and analysis of data and this is the reason why LCR only

began to develop at the beginning of the 1990s, when computers came into general use. However, learner language had long been studied prior to the advent of learner corpora and LCR. This section will provide an overview of the previous approaches to learner language and will highlight their shortcomings as these represent the rationale for the birth of learner corpora and LCR. Before computer learner corpora, many studies into SLA focused on learner language and employed other methodologies. SLA is a relatively new field, only dating back about 70 years. It was defined by Ellis (1997, p. 3) as the "systematic study of how people acquire a second language", but a broader definition was provided by Gass and Selinker (2001):

> Second language acquisition is the study of how second languages are learned. It is the study of how learners create a new language system with only limited exposure to a second language. It is the study of what is learned of a second language and what is not learned; it is the study of why most second language learners do not achieve the same degree of proficiency in a second language as they do in their native language; it is also the study of why only some learners appear to achieve native-like proficiency in more than one language. (p. 1)

This ample definition contains the essential points of this research branch: according to the authors, SLA aims to understand how learners develop their L2 knowledge, thus suggesting a focus on the learning processes. Crucially, this definition includes the study of "what is learned [. . .] and what is not learned", thus addressing one of the issues that SLA had to deal with during its development: namely the absence of errors. Indeed, SLA was concerned with what the learners acquired during their learning process without focusing on what they were not learning (and thus producing). As it will be explained in this Section, the main criticism directed at SLA was precisely the lack of interest in the absence of errors in learner language. Lastly, the definition introduces the concept of fossilisation (see Selinker, 1972; Han, 2004; among others), as described by the last point which recognises that only some learners achieve native-like fluency, while others remain "stuck" and cannot improve or reach native-like proficiency. Thus, the main goal of SLA research is to identify the underlying processes of language learning and explain them.

During the 1960s and 1970s, SLA studies mainly investigated learner language from a pedagogical perspective and through the analysis of errors, which were considered to represent the gaps or failures in the learner's acquisition of the language. Indeed, errors were traditionally seen as a sign that the learner had not mastered the rules of the *target language* (TL) and therefore it was deemed necessary to correct those errors until they disappeared. This was a simplified view of learner language and progress was only made

in 1967 when Corder published an article, *The significance of learners' errors*, in which he suggested that errors should no longer be seen as negative signs of acquisition, but rather signals of the state of IL. For Corder, the acquisition of a second language was practically identical to the process of acquisition of the mother tongue. Thus, errors in the IL should have been regarded as similar to the errors made by L1 children during their L1 learning processes. Error Analysis (EA) became the preferred methodology for investigating learner language, errors were to be analysed not as the mere product of imperfect learning or a simple matter of wrong imitation, but rather as the learners' attempts to understand the new language system (Gass & Selinker, 2001, p. 78). EA was based on the idea that language is a rule-governed system, and by comparing the errors made by the learners in the IL to the TL itself, it aimed to collect useful insights into learner language processes (Gass & Selinker, 2001, p. 79). EA also distinguished between two types of errors: interlingual and intralingual (Richards, 1971), thus working with the concept of language transfer and reflecting on learning processes. EA dominated SLA studies for two decades, until the beginning of the 70s, when the first criticisms to the methodology began to emerge. In 1974, Schachter published an article in which she drew the attention to some inadequacies of EA, in particular the significance of non-errors and the absence of this aspect in EA analyses. In her study, she showed that the lack of errors was a significant phenomenon as it implied that the learners were not confident with a particular grammatical structure (in this case it was restrictive relative clauses) and therefore rarely used it and when they did, they did so carefully and accurately. Later, Schachter and Celce-Murcia (1977) further questioned EA and the concept of errors by pointing out six areas of EA weakness:

> 1) The analysis of errors in isolation; 2) the classification of identified errors; 3) statements of error frequency; 4) the identification of points of difficulty; 5) the ascriptions of causes to systematic errors; 6) the biased nature of sampling procedures. These altogether limit the usefulness of error analysis in describing the acquisition process of the second language learner. (p. 441)

At the time, it was common in EA to select the errors and then discard the remaining text, so that the errors would then appear in isolation. This greatly impaired the research and the reliability of the results, since the exclusion of the context also eliminated non-errors from the analysis. An example of this is given by Schachter and Celce-Murcia (1977) as regards Andersen's (1977) research into the erroneous use of articles by Spanish learners of English. The errors did not provide any insightful data into the learning processes until the whole sentence was analysed. The context not only enriched the interpretation of the analysis, but also lead to the conclusion that the learners were producing the English

equivalent articles of the required Spanish one in specific contexts only. EA also tended to categorise the types of errors belonging to one category or another, but this was shown by the authors to be rather difficult, as one error may belong to one or more categories and, since classification is no simple task, the process is bound to generate mistakes. Another weak area of EA which is worth commenting on is the biased nature of sampling procedures. If a sample is not representative, opposite findings can emerge and impair the overall analysis. The assumption that no error equals correct competence (Schachter, 1974) also greatly undermined the theoretical foundations of EA.

At the end of the 70s, EA lost its credibility among the research community for its lack of empirical data[1], which, at the time, was acquiring growing importance in the linguistic field for its reliability in the field of lexicography. In this regard, during the 1980s the COBUILD[2] project was set up at the University of Birmingham and this was the first lexicographic attempt to use large quantities of empirical data, namely corpus data, thus engendering a new tradition of dictionaries based on corpora. Following the dismissal of EA, SLA studies acquired new research methods (i.e., Performance Analysis[3] and classroom process research[4]). In the following years, SLA developed more sophisticated approaches and data collection methods. The latter are particularly relevant for their connection with LCR and the development of learner corpora. Thus, the following paragraphs present the three main approaches to the collection of learner language data, their advantages, and drawbacks.

2.2 Learner language data collection

In terms of learner language data collection, Ellis's (1994, p. 670) categorisation of data collection methods is the most common and known in the SLA and LCR fields: learner language can be collected through the acquisition of language use data, metalingual judgements, self-report data.

The first data type illustrated in Figure 2.2.1 is language use, which can be either natural or elicited, the latter referring to the method of collection, either clinical or experimental. As regards natural language data, this may consist in recordings of learners speaking the L2 or transcriptions of learners' L2 writings (i.e., learner corpora are composed of natural language data). On the other hand, elicited language data consists in administering specific

[1] For a relegitimised version of EA, see Thewissen (2005).

[2] The first edition of the dictionary was published in 1987 with the title *Collins COBUILD English language dictionary*. For more information on the project, see https://collins.co.uk/pages/elt-cobuild-reference

[3] See Brown (1973); Dulay and Burt (1974).

[4] See Gaies (1983a; 1983b).

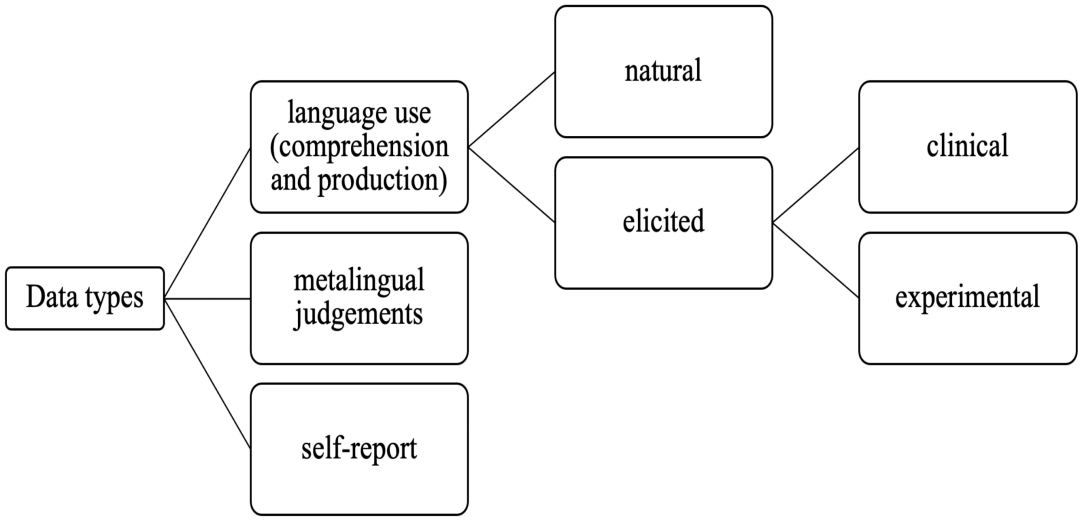

Figure 2.2.1: Data types used in the collection of learner language in SLA studies (Ellis, 1994, p. 670).

tasks to the learners in order to obtain language data. Elicited data is further divided between clinical and experimental: clinical data attempt to elicit spontaneous language features by means of tasks that resemble natural conditions, but with the added advantage of exercising the control of variables such as task type, settings, learners' L1, country of origin, proficiency level, and so on. Experimental data instead consists in more controlled tasks, such as reading aloud, fill-in-the-blanks, imitation and translation tasks. Experimental data fill the gaps of clinical data by collecting specific samples of learner language which may contain examples of structures that may be infrequent and hard to obtain through natural or clinical data collection.

The second type of data type shown in Figure 2.2.1 are metalingual judgements. This is an umbrella term for sets of questionnaires administered to the learners whose function is that of testing the learners' competence on some linguistic items. One of the most common forms of metalingual judgements are acceptability tests, which require the learners to judge the grammaticality (or overall meaning) of a sentence.

The last tile of the chart in Figure 2.2.1 is self-report data, which can be collected by means of written questionnaires or oral interviews. Their aim is that of exploring the learners' introspective point of view on linguistic items or structures. As regards language use data, Granger (2002, p. 5) underlined how much of the SLA research ignores natural lan-

guage use in favour of more elicited data which allows for a greater degree of variable control. Natural language can provide the researchers with a plethora of data about the learners' language development and it is free from any conscious application of learnt rules (Myles, 2015, p. 314). Furthermore, it can inform about the frequency of certain structures so as to better devise experimental data to tap into those rarer occurrences. However, natural language is difficult to collect and time-consuming to transcribe. In addition, if there is no variable control during the collection process, one risks having a large amount of insignificant data.

The use of elicited data may be necessary to investigate specific and infrequent structures which would not otherwise occur in natural language. Elicitation also allows for a greater degree of variable control compared to natural language data. However, Ellis (1994, p. 672) noted that elicitation risks biasing the data towards a less authentic collection of learner language. As far as biases are concerned, when it comes to experimental data, it is difficult to design the elicitation so as to avoid any biases from the learners. Biases should always be taken into consideration when designing corpora, especially learner corpora, since learners (in some countries more than others) are used to a more assessment-centred interaction context.

Alhough the data may have been elicited with extreme care, often the sole use of elicited data is not sufficient to feed comprehensive research into learner language, and as such it should be employed as one of multiple resources in a triangulating method perspective[5]. Metalingual judgements are adopted in SLA for a similar reason to elicited language use, to be able to test the performance of learners on infrequent structures. However, they may be wrongly considered the appropriate tool to access the learners' competence sphere, where in fact they do not provide a "direct window into competence" (Ellis, 1994, p. 673) any more than natural language. Furthermore, they are subject to biases, misinterpretation, and variability among the learners performing the task (Ellis, 1994, p. 674). Lastly, self-report data are the preferred tools for researchers for the investigation of individual variability (something which may be lost in learner corpora). However, self-report data come with limitations, mainly regarding the level of introspection of the learner (are the learners aware of their cognitive processes?) and whether the learners remember a particular choice exactly (do they remember why they used this structure or should have used another in this case?).

[5]For studies which use triangulating methods, see Baker and Egbert (2018), Callies (2009), Gilquin (2007), Lozano and Mendikoetxea (2013), Meunier and Littré (2013).

2.3 LCR approaches to learner language

From an LCR perspective, natural language use seems the most obvious choice for investigating unbiased learner language, provided the SLA natural language data pitfalls are avoided. It is at this intersection between LCR and SLA that the former slowly made its way attempting to cater for those gaps in SLA by proposing the use of learner corpora. Well-compiled learner corpora can address all the limitations identified by SLA studies about natural learner language, and they can improve the other approaches by offering a wider view on IL and providing a great deal of metadata. It is in this framework that the first systematic learner corpus was compiled, the *International Corpus of Learner English* (ICLE), launched in the early 1990s by Sylviane Granger. ICLE was built with the strict design criteria identified by Granger (1998a) and consists of written texts, produced by young adult learners (approximately 20 years old), in an EFL context. The proficiency level was assigned based on external criteria, so the learners were all undergraduates in English Language and Literature in their third or four year. The texts are all argumentative essays, approximately 500 words long, they were not subject to any modifications, and they cover a variety of non-technical topics. The corpus includes several L1 backgrounds following its expansion from one L1 to a total of 21 to date[6] for approximately 3 million words. Metadata such as age, gender, mother tongue, region, other foreign languages are also provided and the corpus is both Part-of-speech (POS-) tagged (see Paragraph 2.3.1.4) and error-tagged (see Section 2.4.3). The corpus was made available to the public in 2002 and until recently most studies on English learner language were based on ICLE (Granger, 1998b). Since ICLE, other learner corpora have been compiled (e.g. the *Longman Learners' Corpus*, the *Hong Kong University of Science and Technology* (HKUST) *Learner Corpus*) and gradually introduced and used in SLA research. Arguably, the majority of SLA studies are still carried out via the use of elicited data (cf. Bachman & Cohen, 1998; Hulstijn et al., 2010; Norris & Ortega, 2003; among others), but it is encouraging to see that the use of learner corpora is increasing, though it is still far from Granger's prediction (2009, p. 17) that "this new resource will soon be accepted as a bona fide data type in SLA research" (as cited in McEnery et al., 2019; Myles, 2015).

[6] ICLE continues its expansion, for up-to-date information see http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm

2.3.1 Learner corpus design

In terms of compilation, though learner corpora have been categorised as "special" corpora (Sinclair, 1996), as for all other types of corpora "very careful planning is necessary for designing a structured corpus" (Kennedy, 1998, p. 65). Indeed, strict design criteria are essential when dealing with the heterogeneity of IL (Gilquin, 2015, p. 16), thus Granger (1998a, p. 8) identified and distinguished between language-related and learner-related criteria. Language-related criteria include: the medium; task type (or genre); topic; technicality; task setting. Learner-related criteria include: age; gender; L1; region; other foreign languages; proficiency level; learning context; practical experience (for variations on the classification see Granger, 2003; 2008; Tono, 2003). In the following paragraphs, these criteria will be described and the variable of proficiency will be analysed on its own due to its particular relevance in LCR. This Section will end with a reflection on two main issues regarding learner corpus design: authenticity, size and representativeness.

2.3.1.1 Language-related criteria

The first criterion to be considered is that of the medium: the medium refers to whether the corpus is written or spoken. Currently there is a wide range of written and spoken learner corpora available, such as ICLE, LINDSEI, and the *Longidutinal Database of Learner English* (LONGDALE). In the past, there was an overabundance of written learner corpora, but the trend is slowing down and there is now an increasing number of spoken learner corpora. T*he Centre for English Corpus Linguistics* (CECL) at Université Catholique de Louvain provides a detailed account of available learner corpora[7]: currently, there are 177 compiled learner corpora in the LCR community, 114 of which are written corpora, 59 are spoken, and 3 are multimedial. The reason why there is a higher number of written corpora lies in the difficulty of collecting spoken data, both in terms of obtaining the subjects' approval and recording and transcribing the data. The compilation process is highly time-consuming and requires a great deal of resources, neither of which are always available to researchers. Indeed, large written corpora are much more common, with CECL listing 13 written corpora below 100,000, 35 between 100,000 and 500,00 words, 16 with over 500,000, and 36 with over 1 million words. As far as the spoken corpora are concerned, there are 8 below 100,000, 5 between 100,000 and 500,000, 4 over 500,000, and only 3 with over 1 million words[8].

[7]For a further and updated list of available corpora see: Centre for English Corpus Linguistics. (2019). *Learner corpora around the world*. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

[8]Data regarding size is not available for all corpora in the list at CECL.

Within the medium, we can distinguish different task types (or genres), such as argumentative essay, narrative writing, semi-structured interview, informal interview, among others. Genre is a particularly important variable for learner corpora since it has been shown to influence learners' output (cf. Alexopoulou et al., 2019; Bouwer et al., 2015; Caines & Buttery, 2018; Foster & Skehan, 1996; Gablasova et al., 2015; Michel et al., 2019; Plonsky & Kim, 2016; Tracy-Ventura & Myles, 2015; Yoon & Polio, 2016; among others). Topic and technicality may also affect learners' production (for topic effects cf. He & Shi, 2012; Hinkel, 2009; Tedick, 1990; Yang, 2014; Yang et al., 2015; Yoon, 2017; among others), for example in terms of vocabulary or range of grammatical structures. Lastly, task setting should be recorded in order to know whether the learners were producing language under a particular set of circumstances, such as a timed essay, or an exam (Granger, 1998a, p. 8).

2.3.1.2 Learner-related variables

Usually a learner corpus may contain learner-related variables (age, gender, other information regarding the learner, such as level of motivation, country of origin) and other metadata, such as L1, parents' L1, L3, school year. These variables are collected because they may provide insights into any potential influences the learner may have been subject to (Granger, 1998a, p. 8). Perhaps, one of the most researched variables is the mother tongue, namely how the L1 can affect (either positively or negatively) learner language (Borin 2004; Crosthwaite, 2016; Jarvis, 2000; Murakami & Alexopoulou, 2016; Paquot, 2013; Wang, 2016; among others). Region is a useful piece of information when dealing with learners and the same L1: for instance, there could be two groups of German-speaking learners, but one group comes from Germany and the other from Austria. In this case, the region/country of origin is important to be able to distinguish different L1 varieties, and consequently their potential transfer effect on learner language. Other foreign languages may also play a crucial role as regards transfer on the learners' IL, though regrettably research has so far mainly focused on the influence of L2 on L3 production (cf. Bardel, 2015; Gut, 2010; Lindqvist, 2009; 2010; Martins & Pinharanda, 2013; Mutta, 2014; among others).

The learning context variable aims to distinguish between English as a Second Language (ESL) and EFL learners, a difference which, in the case of SLA studies, is often a disregarded detail (Granger, 1998a, p. 9). However, practical experience is an umbrella term (Granger, 1998a, p. 9) to indicate all other relevant information about the learners, such as number of years studying the L2, experience abroad, learning materials, or other languages. These variables, if controlled for, could be shown to have an effect on learner language, such as in the case of experience abroad (see Tracy-Ventura et al., 2016;) and other foreign

languages (for an SLA perspective on this see Cheung et al., 2011).

2.3.1.3 Proficiency

Another significant variable in learner corpus design is proficiency: proficiency has been defined, not without challenges (Leclercq & Edmonds, 2014, p. 3), as the ability to use a language regardless of how, where, or when that language was acquired[9] (Bachman, 1990, p. 16). However, though the definition may seem rather straightforward, proficiency is hard to define and often multiple approaches are employed to assign a defined proficiency level to a corpus, thus in fact preventing researchers from comparing multiple corpora. It is still an under-investigated field, though the few available studies have proven its importance in LCR and SLA (cf. Carlsen, 2009; 2012; Granger & Thewissen, 2005; Hultsijn et al., 2010; Jarvis & Pavlenko, 2007; among others).

Since it is a multidimensional phenomenon, there are several models of proficiency scales, among which the *Common European Framework of Reference*[10] (CEFR), which identifies the field of proficiency, that is, the areas that are to be evaluated (e.g., pronunciation, grammar, vocabulary, communicative functions). The CEFR is widely used in Europe, especially in Italy, and as such the vast majority of available learner corpora employ it as their scale. However, as useful as the CEFR descriptors may be[11] for the assessment of students during exams, for instance, they may often be misattributed to certain groups of learners, resulting in unreliable research. This usually occurs when researchers adopt learner-centred rather than text-centred criteria[12] Carlsen (2012) to define the proficiency level of their corpus texts.

One of the most common learner-centred criteria for proficiency is year of study: if a learner is attending the final year of high school or university, the researcher assumes the institutional level of expected proficiency for that school or university. The advantage of this approach is certainly that of reducing the time of data collection and obtaining larger quantities of data since the texts can be collected indiscriminately from a whole class. This can also encourage teachers to quickly build viable learner corpora which can be used for data-driven learning or other teaching/research practices. However, the major drawback for learner-centred criteria is the unreliability of data. Simply put, year of study, age, or number of years studying the L2 are not reliable sources of information regarding proficiency

[9] For other definitions of proficiency cf. Canale and Swain (1980), Carroll (1961), Hulstijn (2010; 2011), Lado (1961), Thomas (1994), among others.

[10] See Carlsen (2012) for more about proficiency scales.

[11] For criticism towards the adoption of proficiency scales, see Brindley (1998).

[12] Atkin et al. (1992) defined them as *external* and *internal criteria* (p. 5).

(Carlsen, 2012; Callies et al., 2013; Callies & Zaytseva, 2013; Gilquin et al., 2010; Gráf, 2017; Mukherjee, 2009; Thomas, 1994; 2006; Tono, 2003; among others) and ICLE is the proof of the unreliability of this criterion (Thewissen, 2015, p. 62). This is due to the high level of variability among institutions and learners. Indeed, as regards the latter, it is hard to imagine that proficiency does not depend on other variables that are not usually accounted for by institutions, such as motivation, willingness to communicate (WTC), classroom anxiety, and many others. Therefore, text-centred criteria should always be preferred as proficiency indicators whenever possible.

Text-centred criteria are based on learner language and as such are far more reliable as they mainly focus on text features rather than external elements. Carlsen (2012, p. 166), mentions three different methods for assigning a proficiency level based on text-centred criteria: teacher's opinion, standardised language tests, raters' assessment. Teacher's opinion may be more reliable than institutional status or year of study; however, it does not seem appropriate for research based on proficiency level since the word "opinion" itself implies an impressionistic evaluation. Subjective evaluations are not scientific in their approach to the object of evaluation and, as such, they are subject to high levels of variability. On the other hand, standardised language tests and raters' assessment have been used to assign corpora a CEFR proficiency level, such as in the *Cambridge Learner Corpus* (Nicholls, 2003) and part of ICLE.

As regards standardised language tests, Carlsen (2012) distinguished between a learner's total score on a standardised language test and the learner's partial score on the same test. The author argued that the total score is a learner-centred method of measuring proficiency since it contains information not only about the learner's performance on one type of text (e.g. written or spoken), but also other skills, such as reading, listening, grammar use. On the contrary, if only the written composition is taken into account, then the partial score of the standardised test can be said to be a text-centred method of measuring proficiency. However, this view may be challenged as different scholars may have different perspectives of what constitutes proficiency. One could argue that when partial competence (such as written competence, but not oral) is tested, it is unlikely that the score is representative of the learner's proficiency, or that when the full scores are not reported, or the average score is provided without ranges or standard deviation (Thomas, 1994, p. 324) this is not sufficient to define the proficiency level. Undoubtedly, standardised language tests may be more reliable than the impressionistic judgement of teachers since they provide specific scores which represent recognised benchmarks. However, as with all assessment methods, there are some limits, which in this case are represented by the different approaches to test score (partial or total) and by the constrained output the learners produce during the test

(Callies et al., 2014, p. 72).

Lastly, raters' assessment is the third proposed method. It is perhaps the most flexible of the three as it can be adapted to any learner production, written or spoken (or multimedia), but it is the most labour intensive and demanding. Indeed, in order to obtain a high level of inter-rater reliability (without which it is impossible to acquire consistent data), the raters need to receive extensive CEFR (or any other proficiency scale model) training before they can assess the texts. Furthermore, two raters might not suffice if they do not agree on a significant number of texts, thus a third rater is needed.

Recently, new procedures and assessment methods are being tested: Thewissen (2015) proposed a new stratified proficiency assessment. Hilton (2014) has shown that temporal fluency measures are a reliable method to assess oral proficiency; Wulff and Gries (2011) tested a probabilistic analysis of lexico-grammatical association patterns for measuring proficiency; Tracy-Ventura et al. (2014) demonstrated how elicited imitation can be a fairly valid and reliable measure, while Leńko-Szymańska (2020) has provided a comprehensive account of performance-based assessment of L2 lexical proficiency. All these procedures are crucially text-centred[13], thus underlining once again the importance of not trusting learner-centred criteria. Nevertheless, it is not always feasible to employ the most appropriate methods due to several constraints; the researcher should therefore opt for what is most practical, where practicality is the balanced ratio between available resources and required resources (Bachman & Palmer, 1996, p. 35). Carlsen (2012) indeed pointed out that the researcher should aim for a compromise between the purpose and the intended use of their corpus and choose whether a more resource-expensive level assignment is needed or not (p.179). In any case, as Jones (2011) adequately put it, it is worth remembering that it is impossible to "perfectly measure(d)" proficiency.

2.3.1.4 Corpus typology: some considerations

In the last few years, there has been an international surge of learner corpora. There are several types of learner corpora, such as academic, commercial, or global corpora. In this Section, we will briefly present the different types of corpora and some considerations will be provided regarding written vs spoken corpora, annotated vs unannotated corpora, and

[13] As regards the elicited imitation measure, this could be considered either textual or arbitrary. Indeed, since elicited imitation consists in providing the learners with a stimulus and eliciting an imitation of that stimulus, one could interpret this as textual criteria since the proficiency is assessed based on those elicited texts and no arbitrary information (such as age of participant, year of study, etc.) are taken into consideration. However, a different perspective may perceive elicited imitation as arbitrary criteria since there are multiple variables at play during the assessment, such as the length of the stimulus sentence or the participant's working memory (Tracy-Ventura et al., 2014, p. 146).

cross-sectional vs longitudinal corpora.

The initial distinction between corpora is academic vs commercial. Academic corpora are corpora compiled by researchers of a university, generally for their own research purposes. Commercial corpora are usually much larger, contain a wider range of L1 backgrounds and other variables, and have been compiled by publishing houses or in collaboration with commercial entities and universities. Quantitatively, there is a higher number of academic corpora available, while there are only few commercial corpora, such as the *Cambridge Learner Corpus* (Nicholls, 2003) or the *Longman Learner's Corpus*[14].

Another difference among corpora is the TL: the majority of available corpora adopt English as their TL (e.g., ICLE). However, other TL corpora have been compiled and are available to researchers, such as EVA, a corpus of learners of Norwegian (Hasselgren, 1997), LANGSNAP (Tracy-Ventura et al., 2016) a corpus of learners of Spanish, or LOCCLI, the *Longitudinal Corpus of Chinese Learners of Italian* (Spina & Siyanova-Chanturia, 2018). Another variable strictly related to the TL regards the L1: corpora can also include multiple L1s, which are very useful for studying the cross-linguistic influences on TL (one such corpus is LINDSEI containing data from 11 L1 backgrounds); but others can be mono-L1 corpora, such as LOCCLI (whose participants are all L1 Chinese speakers).

Another aspect that can be identified among different types of corpora is genre: some corpora include a variety of genres, while others have a more restricted set of genres. For example, ICLE contains essays written by learners of English, while LINDSEI contains three genres: a monologue, a dialogue, and a picture description. Within genre, there is also the distinction between general learner corpora, that is corpora which include learner language produced in general learning settings, and language for specific purposes corpora (Gilquin, 2015). The latter include genre-specific texts produced within the framework of language for specific purposes, such as the *Active Learning of English for Science Students* (ALESS) learner corpus (Allen, 2009).

A further element of distinction lies between global and local learner corpora: global corpora are large-scale projects and the data is collected by interviewing or collecting learners' material (Gilquin, 2015). On the other hand, local learner corpora are usually collected by teachers among their students, who become both contributors and users of the corpus. Local corpora allow the researcher to investigate the specific features and needs of their group of learners and "provide tailor-made solutions to their problems" (Gilquin, 2015, p. 15). Local learner corpora are usually compiled in order to have access to data collected in a specific context (region, school, teacher, etc.) or because the available corpora do not satisfy the

[14] See http://www.pearsonlongman.com/dictionaries/corpus/learners.html

researcher's needs. This is the case of the corpus compiled for this study, whose design will be further discussed in Chapter 4.

Written vs spoken As mentioned before, the majority of learner corpora fall into the written category (114 written vs 59 spoken according to CECL). Leech (1998: xviii) stated that "writing is an exceedingly important skill for most foreign language learners, and well deserves the expenditure of effort to collect corpora of written learner language". Although he was aware that humans primarily process events and situations through speech, he drew attention to the fact that computers are built for written texts (Leech, 1998: xviii). The immediate advantage of written language is the clean reflection of the language without any interruptions, repetitions, reformulations. However, spoken language better reflects the thinking and learning processes of the learners, it contains the more spontaneous output of language and as such it has been indicated as the preferable resource for SLA research (Myles, 2015). Recently, there seems to be an inversion of trend in the literature, with LCR shifting more towards spoken learner language, and SLA switching from "oral language as a privileged site for L2 learning" to the more academic and written registers side of learner language (Ortega & Byrnes, 2008, p. 284).

Annotated vs unannotated corpora A corpus is a very useful resource by itself, as we have seen so far, but its research potential can be greatly enhanced if further information is attached to it. Corpus annotation is "the practice of adding interpretative (especially linguistic) information to an existing corpus of spoken and/or written language by some kind of coding attached to, or interspersed with, the electronic representation of the language material itself" (Leech, 1993, p. 275). This added information can aid the researcher in the automatic retrieval, computing, comparison and other research activities. The annotation can be via automatic, semi-automatic, or manual processes. One of the most widespread automatic annotations is POS-tagging: the annotation assigns to each token of the corpus a tag representing the part of the speech to which that token belongs. There are several sets of tags, also called tagsets, and they can vary from 50 to 250 items. The higher number of tags guarantees a greater level of granularity and specificity in the research. As far as non-native speakers corpora are concerned, the researchers can count on a number of automatic POS-taggers and parsers which have been shown to have a high level of reliability. However, since these taggers have been trained on native language, they perform significantly less well when applied to learner language. In the majority of cases, it is necessary for the researcher to manually re-check or implement the tags, thus the tagging becomes semi-automatic (for a study that looks at adapting automated part-of-speech tagging to learner English, see Na-

gata et al., 2018). The taggers have a higher rate of performance reliability when used on advanced learner language (due to the lower number of errors or disfluencies); indeed Van Rooy and Schäfer (2002) conclude that the CLAWS[15] tagger achieves a 98% of accuracy on advanced written texts (provided spelling errors are first removed), thus practically equalling the level of accuracy reached when tagging a native corpus. This obviously changes when the tagset is used on lower levels of proficiency (de Haan, 2000; Díaz-Negrillo et al., 2010), so a more manual approach is involved. POS-tagging can aid SLA and Foreign Language Teaching (FLT) researchers by allowing them to perform selective searches of specific parts of speech in learner language (Granger, 2002, p. 16). Furthermore, they are especially useful when comparing language varieties to extract data about the frequency or infrequency of certain structures.

Another type of annotation is error-tagging: a tag is assigned to the learner errors in the IL. This procedure is fully manual, though a few studies are attempting to automatise the process (see Section 2.4.3), and as such it is highly labour-intensive and time-consuming. In order for the error-tagging to be effective and reliable, strict error categorisation is required. There are a few error-tagging systems available, such as the *Université Catholique de Louvain Error Editor* (UCLEE) (Hutchinson, 1996) or TagEditor (Izumi et al., 2003). These systems usually include an interface which allows the researcher to select the desired tag and insert it in the text. Furthermore, it is also possible to add corrections or other types of information. Once the tags have been assigned, they can be retrieved with software retrieval tools, such as AntConc (Anthony, 2004) or Wordsmith Tools (Scott, 2004). The main purpose error-tagging serves is CEA. CEA will be further presented and described in Section 2.4.3. However, the literature warns that too much annotation may clutter the corpus and hinder the research by not allowing the researchers to "very closely look at the physical evidence" (Sinclair, 1991, p. 29), thus it is best to evaluate the priorities of the research and consider working both on a raw (clean) corpus and annotated one, "a mixture of plain text and annotation" (Hunston, 2002, p. 94).

Cross-sectional vs longitudinal corpora A further perspective in learner corpus typology which deserves some attention is the temporal difference between corpora: there are cross-sectional (or synchronic) and longitudinal corpora. Cross-sectional corpora include data collected from different learners at the same point in time, while longitudinal corpora are compiled with data from a defined set of learners collected at several points in time. Depending on the collection points, there are denser and larger corpora which can span

[15]CLAWS (Garside, 1987) is a POS-tagger devised developed at the beginning of the 80s and version CLAWS4 was employed to tag the BNC.

from a few collection points across a few months to multiple collection points over a year or more. This second type of corpus is ideal for studying the SLA processes of the learners, but it is very difficult to compile since it requires researchers to follow the same group of learners for months or years. While there is a wide range of cross-sectional corpora, few longitudinal corpora are available: the most important longitudinal project was started in 2008 at the University of Louvain, Belgium and it is LONGDALE. There is also the *Corpus of Young Learner Interlanguage* (Housen, 2002), which contains EFL spoken data from the European School pupils at different stages of L2 development, LONGSNAP (Tracy-Ventura et al., 2016) which is a corpus of L2 Spanish, and, more recently, Università per gli Stranieri di Perugia is working on an L2 Italian longitudinal corpus with three collection points.

Since it is rather difficult collecting longitudinal corpora, a compromise has been reached among LCR, namely quasi-longitudinal corpora (Granger, 2002): these are corpora collected at different points in time from different groups of learners. In this case, variable control is key to maintain consistency. The NICT JLE Corpus (*National Institute of Information and Communications Technology Japanese Learner English Corpus*; Izumi et al. 2004) is one such example.

There is a clear SLA benefit from having more longitudinal corpora available, but the resources required for their compilation often outnumber the potential advantages. Future research should undoubtedly focus on collecting more longitudinal and diverse data to promote a stronger interaction between the two fields.

2.3.1.5 Core issues

Authenticity As with standard corpora, one issue which is noteworthy of attention is that of authenticity. It stems from Sinclair's (1996) statement that corpora should contain the language of people "going about their normal business". This clearly clashes with learner language, especially with EFL students who cannot simply "go about their business" speaking their L2. Indeed, Granger (2002, p. 5) pointed out that learner language is "rarely **fully** natural" and she reiterated this in her definition of learner corpus in 2015, which is reported at the beginning of Section 2.1. If Sinclair's definition of authentic language is followed to the letter, learner language collected in a corpus is rarely authentic since the tasks carried out by the students in a classroom fulfil pseudo-communicative purposes and do not take place in real-life contexts (Gilquin, 2015; Tono, 2016, p. 123). However, one could argue that there are varying degrees of naturalness: fill-in the gaps exercises and read aloud tasks rank very low on the naturalness scale, but informal interviews or free compositions may resemble authentic language more closely (Granger, 2008, p. 261). Unnatural language

data such as elicited tasks have been widely used in SLA studies to demonstrate theories which required the analysis of infrequent language features or controlled settings, but in the case of learner corpora, researchers strive to collect the most natural data possible, i.e., open-ended productions (free writing, non-guided conversations). Thus, although a certain degree of artificiality is inherent to learner language since the learners are subject to varying levels of elicitation, it may be argued that, despite elicitation, informal interviews and free writing are the appropriate context for naturally occurring learner language (Tono, 2016, p. 124) (see also Section 4.2.2).

Corpus size and representativeness Another consideration about learner corpora regards size and representativeness. Representativeness is usually understood in the literature as the balanced sampling of a target population. The target population is the language the researcher wants to investigate (e.g., advanced Italian learners of English L2; elementary French learners of Italian L2). Since it is usually impossible to collect the whole production of a target population, corpora aim to collect balanced samples, thus becoming representative of that specific population. Corpora, by definition, are large collections of data, so one should attempt to "gather data in quantity" (Sinclair, 1995, p. 21). Hundt et al., (2007) argued that although the standard size of reference corpora now is approximately 100 million words, this might not be enough for certain research questions. Indeed, big corpora allow for generalisations of results and, intuitively, large corpora may seem far more representative of the target population than small ones. In addition, statistical tests, without which the corpus linguists' sought-after generalisations cannot be made, can only yield meaningful data if they are run on a high number of tokens (Brezina, 2018). However, if the sample is not representative, even large quantities of data are meaningless.

Thus, the choice of the sample needs to be carefully planned: Gries (2013, p. 21) provided a useful example of how balanced a sample should be and warned that although the considerations about the target population may be rough estimates, the sample should nevertheless reflect as closely as possible the chosen population. It follows that sample size does not necessarily matter and should not be one of the primary concerns of the researcher, but rather we should worry about obtaining a clear definition of the target population and sampling methods (Biber, 1993, p. 243). This is all true for standard corpora, but for learner corpora it can be different. First of all, compiling learner corpora is a "painstaking process" (Granger, 1998a, p. 10) and even though technological advances have simplified most of the compilation steps, in countries such as Italy, until recently, it has been difficult to find written texts already in digital format. This means that in order to collect 100 essays the researcher will likely have to transcribe each one of them, with the added effort of in-

terpreting individual handwriting. In addition, the researcher may encounter the issue of legal permission when dealing with minors. This can easily be solved if compiling a small non-commercial corpus which will not be shared within the research community, but not when attempting to collect data for a large corpus to be made available to the public. These problems become even more significant when trying to compile a spoken learner corpus: this involves the transcription of the audio recordings which cannot be aided by software because more often than not there is a high degree of variability in terms of pronunciation, word order, and errors. Consequently, corpus size is not a black or white matter and each case should be carefully considered on the basis of the only factor that truly matters: the research question (de Haan, 1992). For certain investigations, perhaps in the field of SLA, a small corpus may be sufficient and representative of a population, while for other analyses larger corpora are needed to obtain reliable data (Granger, 1998a, p. 11).

2.4 LCR methodological approaches

In this section, the three main approaches to the analysis of learner corpora will be introduced. The first paragraph proposes a brief history of CIA together with some criticism and rebuttals; the second paragraph provides a description of the new version of CIA, CIA2; and the last paragraph is a short introduction to CEA. CIA has been defined as "the heart of LCR" (Granger, 1998b, p. 12) and this still holds true as most of the contemporary studies in LCR employ a corpus-based[16], quantitative, cross-sectional and comparative approach (Callies, 2015, p. 38). Following major criticism to CIA, Granger decided to upgrade CIA to CIA2, thus addressing some of the objections that the model had been subject to. The three approaches lend themselves to either corpus-based or corpus-driven[17] analyses, although, so far, the majority of studies have employed corpus-based methods (Callies, 2015). However, as more resources and tools are becoming available to researchers, it is likely that future research will move towards more corpus-driven approaches. Furthermore, since available learner corpora have become rather large and LCR is inheriting the quantitative spirit of cor-

[16] As regards the corpus-based approach, Storjohann (2005) defines it as a method that uses a corpus as an inventory of linguistic data: "from this repository, appropriate material is extracted to support intuitive knowledge, to verify expectations, to allow linguistic phenomena to be quantified, and to find proof for existing theories or to retrieve illustrative samples. It is a method where the corpus is interrogated and data is used to confirm linguistic pre-set explanations and assumptions. It acts, therefore, as additional supporting material" (p. 9).

[17] Although many definitions have been provided for the terms *corpus-driven* and *corpus-based*, the description provided by Storjohann (2005) is one of the most well-rounded. She defined the corpus-driven approach as the methodology whereby the corpus is used as an empirical basis from which researchers can extract data and identify linguistic phenomena "without prior assumptions and expectations" (p. 5).

pus linguistics, all these approaches are adopting increasingly quantitative methods. This results in generalisations of linguistic phenomena which can also be useful for SLA studies, provided that the data have been extracted from well-balanced and sampled corpora with strict variable control.

2.4.1 CIA

The advent of CIA is closely linked to the compilation of ICLE, which was not only fertile ground for the beginning of LCR studies, but also turned into the framework for approaching learner corpora. Indeed, CIA lies "at the core of the ICLE project" (Granger, 1996, p. 43), as it is not difficult to imagine how the initial 11 sub-corpora naturally spurred a series of comparisons. Comparing languages was part of the methodological approach of Contrastive Analysis (CA), but the new CIA took a step further and did not compare only different languages, but rather different varieties of language. Although CIA has attracted much criticism regarding its underlying concept of comparing, Hunston (2002) expressed how it is in the nature of learner language to be compared "between corpora produced by different sets of learners, and between corpora produced by learners and those produced by native or expert speakers" (p. 206). Indeed, the process of learning a foreign language is in itself a comparison to a given norm, since learners do not acquire their IL without inevitably looking at a TL. Therefore, it seems impracticable to only analyse IL by itself, though this kind of investigation undoubtedly has its benefits and should not be disregarded (Granger, 2002; 2017).

CIA envisages two types of comparison:

1) Native language vs interlanguage;

2) Interlanguage vs interlanguage.

The first comparison aims to shed light on the features and patterns of learner IL, such as non-nativeness (i.e., errors, disfluencies, overuse/underuse), both in terms of quantitative and qualitative aspects. This type of comparison is the one that has been subject to the highest level of criticism (Granger, 2017, p. 5) due to its underlying assumption that learner language should be compared to a norm. However, this comparison has been able to provide interesting findings on learner language. For instance, Hasselgren (1994) found that learner vocabulary is less varied than that of native speakers, while De Cock et al. (1998) identified learners' recurrent word combinations. Both these studies were based on L1 vs L2 comparisons; other approaches not involving a direct comparison to the L1 would not be able to yield these findings.

The second type of CIA involves the comparison of different varieties of IL: for example,

it is possible to compare L1 French learners of English L2 with L1 Italian learners of English L2, which is an interlingual analysis. Another comparison is the intralingual one, that is, the analysis of intermediate Italian learners of English L2 vs advanced Italian learners of English L2. The aim of this kind of analysis is to gain insight into the nature of IL and evaluate the effect of variables such as age, proficiency, task type (Callies, 2015, p. 40). Comparing interlanguagae vs interlanguage allows the researchers to verify whether certain features of learner language characterise a specific group of learners or it is attributable to a wider developmental pattern of SLA (Granger, 2002). This type of CIA did not receive much criticism from SLA researchers (because the IL is investigated on its own, without the comparison to a norm) and has been able to provide the research community with some insightful findings: for example, it has been shown that advanced learners are greatly influenced in their written production by informal spoken language (De Cock, 2004; Gilquin & Paquot, 2007; Götz, 2013; among others), and Osborne (2008) demonstrated that learners transfer their L1 syntactic patterns to written English (cf. Murphy & Poli, 2018).

2.4.1.1 The comparative fallacy

One of the loudest arguments against the CIA has been that of the comparative fallacy (Bley-Vroman, 1983; Larsen-Freeman & Long, 1991): learner language should be analysed on its own, rather than in comparison to native speakers'. Indeed, the main problem that needs to be addressed when employing CIA is to what should learner language be compared. Many questions arise at this point, among which are: should we compare learner language to native-speaker language? If so, which native-speaker variety? Should the corpora have the same task types? Is it fair to compare bi- (or multi-)lingual learners to monolingual native speakers? Should they be peers? Part of these questions have been addressed by the reappraised version of CIA, which will be examined in the next section (2.4.2).

However, although many solutions to the comparability issue have been found (e.g., better reference corpora), SLA theorists may still argue that by comparing IL to a norm, researchers are still focusing on the deviations from the TL, and as such they perceive learners are more or less successful, rather than truly examining the underlying learning processes. On the other hand, as Granger pointed out, even SLA researchers implicitly compare IL to the TL whenever they mention proficiency (since it is by definition measured in comparison to an L1 norm) or when they administer grammar acceptability judgements. She even proposed that there is "comparative hypocrisy" among the SLA research community (Granger, 2009, p. 18–19) as also sustained by Sung Park (2004) that "any SLA study implicitly has a built-in notion of IL with the target language lurking in the background" (as

cited in Granger, 2009, p. 19). Undoubtedly SLA analysts are aware that IL is never investigated purely on its own, but there is always a notion of TL. Indeed, Myles (2015) did not negate the influence of L1, but contextualised it within a more complex framework of cross-linguistic differences that go beyond apparent overt contrasts (p. 311).

Perhaps observations coming from SLA studies aim to address the explicit analysis of learner corpus studies of IL vs TL, which from the SLA perspective may not include the most recent theories of language learning. However, as already mentioned above, many interesting findings in LCR could not have been demonstrated without CIA. Furthermore, CIA is an appropriate method when investigating learner language from a pedagogical perspective: the majority of teachers perceive deviation of their learners' IL from the TL as areas of work, thus not with a negative lens, but with a teaching aim in mind to bring the learners closer to their TL. In addition, one element that has never been widely debated as much as the others when arguing for or against CIA[18] is that the learners themselves may have a specific norm in mind. It may be that, as Hunston (2002) sustained, CIA "assumes that learners have native speaker' norms as a target" (p. 211–212) and as such not all learners may have a target norm in mind. However, some learners may indeed have clear objectives and base their learning on the comparison between their current level and their desired target (most likely an L1 norm). This is supported by Ellis and Barkhuizen's claim (2005, p. 360) that "learners are typically targeted on native-speaker norms and as such themselves perform 'cognitive comparisons' in the process of learning an L2", even though the authors also stressed that IL should be studied on its own. It is the duty of the researcher to make sure that the corpus they are compiling will reflect the target norm of that particular set of learners[19]. Thus, it appears futile to debate further about the legitimacy of CIA, especially since CIA is not the only method for investigating learner language and most of the SLA concerns about it have already been amply addressed.

2.4.1.2 Overuse and underuse

The other widely debated issue arising directly from employing CIA in learner corpus studies is the common occurrence of terms such as *underuse* and *overuse*. The problem with using these terms, according to the SLA literature, is that they stress deviations or deficiencies of the learner language from the L1 norm (Aston, 2011). Leech (1998, p. xix–xx) warns linguists that the terms should not be used "in a judgmental spirit", but rather as mere descriptors of the findings of the corpus, as it is easy to fall into the prescriptive trap

[18] To the author's knowledge.
[19] The revised version of CIA addressed this pitfall (see Section 2.4.1).

when analysing learner language and attempting to provide useful data for pedagogical implications. The learner corpus literature has always made clear that these terms were used neutrally, without any sort of judgement beyond a simplistic frequency count (cf. Gilquin & Paquot, 2008). In the evaluation of the learner corpus literature, no works which employed these terms in a negative perspective were found. The terms were rather simply used to describe learner language in comparison to the TL in terms of quantitative results. Moreover, as with all quantitative results, more often than not frequency may not be a sufficient indicator of learner performance and may even hide errors or disfluencies; thus, the quantitative data extracted from the comparison, that is, the underuse and overuse, are usually followed and contextualised by a more qualitative analysis of their occurrences (cf. Guo, 2006 for more on underuse and overuse). However, should the criticisms regarding the use of these terms persist and become stronger, one feasible solution to the issue is represented by Callie's (2015) proposal of replacing the terms with the more neutral and descriptive alternative ones of *underrepresentation* and *overrepresentation*. This seems like a plausible alternative, if it were not for the fact that the terms are now widely spread among the research community and their replacement seems unlikely, as also discussed by Granger (2015, p. 19).

2.4.2 CIA: the updated version

As it has been described above, over the years CIA has been the subject of multiple reviews and comments about its underpinnings and practices. Consequently, in 2015, Granger proposed a revised version of CIA, called CIA^2. The following paragraphs will present CIA^2 and its main core features.

CIA^2 was developed by Granger (2015) in order to address criticisms levelled at some CIA's issues. One of these problems has always been the terminology. In her first presentation of CIA, Granger (1996) used the terms *native language* and *interlanguage*, and, to exemplify her model, she used English as the second/foreign language.

The diagram in Figure 2.4.1 shows the two traditional types of comparison in CIA, native language (NL) vs interlanguage (IL), for which Granger provided the example of English as native language (E1) vs English as a second language (E2), and interlanguage (IL) vs interlanguage (IL). The example for this second comparison once again uses English as the second language and several different native languages to mark the different types of IL (E2F is French learners of English L2, E2G is German learners of English L2, E2S is Spanish learners of English L2, and E2J is Japanese learners of English L2). Despite updating the terminology a few years later with *native speaker*, *non-native speaker*, and

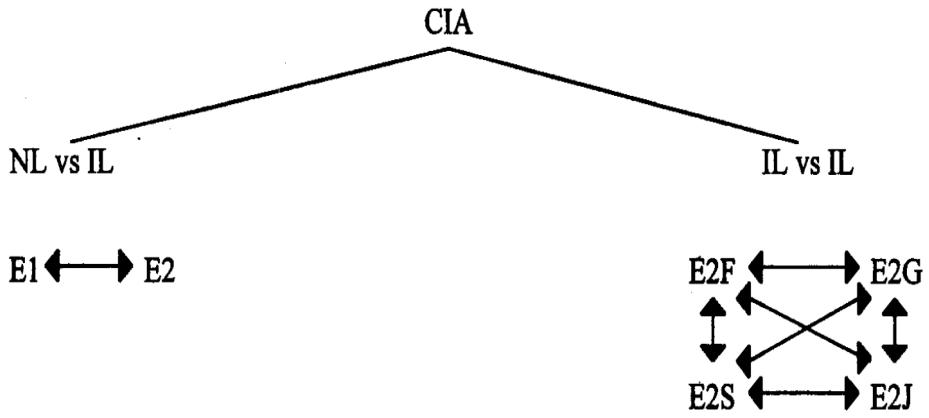

Figure 2.4.1: Contrastive Interlanguage Analysis (Granger, 1996).

L1 or L2 (Granger, 2012a; 2012b), criticism from the SLA community continued, and still continues for that matter, to address CIA and its comparative fallacy (Bley-Vroman, 1983; Larsen-Freeman, 2014; Selinker, 2014). However, despite not all SLA scholars agreeing about the alleged shortcomings of CIA (Lardiere, 2003; White, 2003), Granger felt the need to take stock of the flaws in her model and implement the recent SLA theories on language learning in a revised version of CIA.

CIA[2] changes the terminology and broadens it to include a wider spectrum of (inter) language varieties. Indeed, since there is a tendency in LCR to compare L1 vs L2, the new model allows for more focus on the L2. This is provided by changing the concept of L1, which has so far acquired a negative status and has been perceived as "one single monolithic norm".

The new version of CIA promotes the idea of varieties, thus on the left of the diagram (Figure 2.4.2) we read reference language varieties (RLV). With this new term, Granger challenged the idea of norm, which is no longer associated with the L2 (perhaps English L2), but with virtually any potential variety of L2 which represents the TL for a particular set of learners. In this regard, researchers are encouraged to use any variety of L2, starting from the more traditional English varieties, to the more contemporary World Englishes[20] (WEs) or English as a Lingua Franca[21] (ELF). Furthermore, researchers should also be

[20]The term World English is used to refer to all different varieties of English spoken in various contexts across the world. Such varieties include both native-speaker varieties (British English, American English, Australian English) and second-language ones (Indian English, Nigerian English etc.). For more on World Englishes, see Kachru (1984).

[21]The term English as a Lingua Franca is used to refer to the teachingand learning of English as a means of

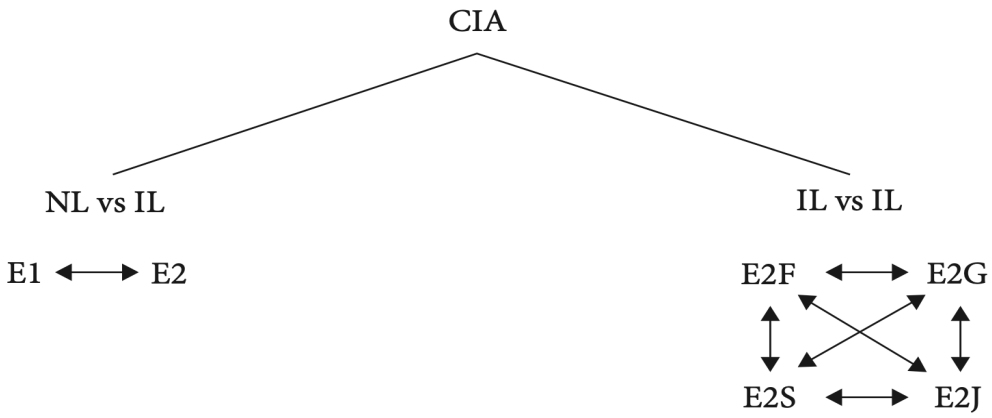

Figure 2.4.2: CIA^2 (Granger, 2015: 17).

aware of the diatypic[22] and dialectal[23] variables of their reference corpora so as to be able to compare like with like. Multiple comparisons among reference varieties are also suggested by Granger (2015) when she cited the study of Lee and Chen (2009), who adopted CIA^2 to investigate expert academic writing.

On the right of Figure 2.4.2, the new terminology for IL is found: it acquires the term varieties (ILV). This choice reflects the SLA scholars' objections (and LCR too, see Callies, 2015) about the poor consideration that language variability has received in LCR studies. By introducing the term varieties, Granger prompted the readers to focus more on the L2 in its own right and all its uncountable facets. One criticism levelled at CIA and LCR studies in general is that too much attention is being paid to the L1 variable, whereas CIA^2 promotes the investigation of new variables, such as task or study abroad experiences (which have been grouped under the terms task and learner variables). Granger (2015, p. 18) noted that there is plenty of metadata available for the majority of learner corpora, so it is up to the researchers to exploit the data and expand their research to other less investigated areas of learner language.

Granger (2015, p. 14) suggested that researchers could start with a comparison between L1 vs L2 (which is undoubtedly more relevant in the field of LCR whose aim is pedagogical

mutual communication and intelligibility between different L1 speakers. For more on ELF, see Jenkins (2013).

[22]The term diatypic refers to the fact that language may vary according to the type of situation the speakers encounter. Most commonly, diatypic varieties reflect changes in register.

[23]The term dialectal refers to all those social, geographic, individual, and group factors that may influence the production of language. In particular, these often affect a given speech community (country) or region.

applications, rather than uncovering new theories of SLA) and, once they have identified characterising features of the learner language, they may proceed with a further exploration of such elements.

2.4.3 CEA

CEA is another approach to learner corpora which has received a warm welcome in the LCR community due to its relevance for teaching practices. CIA and CEA are probably two of the most common approaches in LCR, with CEA perhaps used slightly less due to its labour-intensive and time-consuming preparation and application. An extensive review of the CEA literature is beyond the scope of this work, thus this Section will briefly present the key principles of CEA and some of its implications.

CEA was first introduced in 1998 by Dagneaux et al. as a new approach to EA and learner corpora. It addressed the main methodological issues of EA (i.e., errors in isolation, no clear error categorisation) and spurred a series of studies and innovations. The main aim of CEA is the consistency of the analysis (Dagneaux et al., 1998, p. 166).

The main differences between CEA and traditional EA lie in the use of computers to aid the analysis, the high level of standardisation of errors, and the analysis of errors within their own context (Granger, 2002, p. 10). In order to carry out CEA, the texts need to undergo a rigorous series of steps: the researcher(s) (Dagneaux et al., 1998, recommended two researchers: one native speaker of the learners' L1, and one native speaker of the learners' TL) manually check and correct the texts; they assign the appropriate error tag using an error editor[24]; they analyse the texts using text retrieval software.

The third step can move in two directions: the first is a more corpus-informed/corpus-based approach; the second is fully corpus-driven. The first method consists in the selection of an error which has been shown in the literature (or predicted by intuition) to have a tendency to occur frequently in a given context. The researcher can then search the corpus and find instances of the error quickly via a text retrieval software programme. This is a very fast and intuitive method, though it limits the findings to the researchers' prior selection of error(s). The second method is more time-consuming and can often become a bottleneck for the analysis of data (Rayson & Baron, 2011, p. 109), but it opens up to a wide and comprehensive exploration of learner data and errors. The researchers devise a standardised set of error tags which they then assign to the errors found in the corpus. This translates to a fully-tagged corpus ready to be investigated for any potential error patterns, even those that

[24]CECL has devised its own error editor, which is accessible on its website (https://uclouvain.be/en/research-institutes/ilc/cecl) with its accompanying error tags manual.

had not been initially hypothesised by the researchers.

More recent studies have tried to devise semi-automatic or fully-automatic tagging of errors. For example, De Felice (2008) worked on the automatic detection of determiners and prepositions in learner writing (De Felice & Pulman, 2009), Rayson and Baron (2011) have attempted to implement a hybrid system based on natural language processing for the tagging of spelling errors, and Klyachko et al. (2013) have pilot tested how the use of *n*-grams can perfect corpus error annotation on a small sub-set of the *Russian Learner Corpus of Academic Writing* (RULEC). It is encouraging to see that CEA is leading to some interesting teaching practices and applications, as shown by a TaLC2018 workshop on error annotation in learner corpora led by Olga Vinogradova, Stefania Spina, Luciana Forti, Ivan Torubarov, and Nikita Login. The researchers demonstrated the promising potential of REALEC (the collection of English essays written by Russian university students), the error tagging scheme, and the error editor BRAT[25].

2.5 Core issues in LCR: will LCR ever meet SLA?

One of the two main functions of LCR, as mentioned in Section 2.1, is to contribute to SLA with better data for the description and understanding for IL. As promising as this may sounds, few pitfalls emerge as regards this collaboration. Despite Granger predicting that LCR will meet SLA and both will benefit from the implementation of one into the other and the great number of studies, efforts, and new compiled corpora, this has not yet occurred (Gries, 2015; Myles, 2015). In this Section the main issues preventing the interaction of the two disciplines are described along with some future directions.

Learner corpora are undoubtedly great tools with a strong potential for better corpus-informed SLA research, but the two disciplines oftentimes seem to be misaligned. SLA aims to explain and understand the processes of language learning and to do so it mainly relies on samples of natural language (but not systematically compiled corpus data) or elicited data. On the other hand, LCR aims to contribute to SLA and to teaching practices/materials and it does so with corpus data. Issues arise when SLA perceives corpus data as non-viable for research into language acquisition processes due to the lack of variable control (and balance of the sample) and when LCR discards SLA results as they are drawn from small samples of learners and are thus non-generalisable. Thus, on the one hand SLA studies do not rely on corpus data as learner corpora are often not designed for SLA research; while on the other LCR aims to contribute to SLA studies, but at times fails to take into account the needs of

[25]https://brat.nlplab.org/index.html

SLA researchers.

First of all, if LCR is to help SLA research with the corpora contribution, one must assume that corpus compilers have an extensive background in SLA studies. It would only seem feasible to compile a useful learner corpus for SLA if one knew the research questions of this discipline. Yet, in the LCR literature, corpus development has not been a primary issue as much as analysis and interpretation of data (Brezina, 2018), thus there are still several corpora which do not meet SLA research needs: the majority of the corpora are written, there is no strict variable control, the data are not well distributed, the size is still too small, and the annotation is sparse (Lessard, 1999), to cite a few.

One may argue that corpora such as EFCAMDAT (Alexopoulou et al., 2017) or the Cambridge Learner Corpus (Nicholls, 2003) can cater for SLA research needs as they are rather large corpora (83 and 50 million words respectively) and offer a wide range of variables that can be combined for multifactorial analysis, for instance. However, as McEnery et al. (2019) pointed out, these corpora are only apparently well-balanced, as when one digs deeper into the variables, it is easy to discover that the data is sparse and there is not an even distribution, thus in fact hindering potential SLA research questions. This is an important aspect that corpus compilers should take into careful consideration: the size of the corpus does not necessarily imply a wide range of potential research questions.

There are always limitations and every corpus design decision should be taken carefully. However, the limit of learner corpora are not the limitations themselves, but rather the potentially SLA-weak design choices and especially the lack of communication about these choices to the research community. McEnery et al. (2019, p. 80) explained this: "corpus builders [...] need to be open about the design decisions that shaped that collection of data. If they are not, corpus linguistics may be seen, on occasion, to overpromise and ultimately disappoint". The misalignment of goals between the two disciplines causes several issues as regards LCR methodology. Among them, the following ones may be particularly relevant:

a) lack of varied corpora in terms of variables (e.g., proficiency, tasks, sociolinguistic contexts, and so on);

b) surplus of written corpora to the detriment of spoken corpora;

c) overconcentration of studies on the L1 variable to the detriment of other potentially more SLA-relevant variables;

d) studies' main focus on frequency counts and descriptive accounts of differences about IL vs TL.

The literature has drawn attention to the first issue, the lack of varied corpora, by pointing out the fact that for an extensive SLA analysis, rich learner data with comprehensive backgrounds and contexts are needed (Myles, 2015) and, naturally, if the SLA researchers cannot have access to a wide and rich range of learner data, they will resort to experimental methods (Callies, 2015; McEnery et al., 2019). However, it is encouraging to see that the latest project by Gablasova et al. (2019) is addressing this issue and providing the research communities with a well-balanced, wide-range learner corpus, the *Trinity Lancaster Corpus* (TLC): the corpus comprises 4.2 million words of transcribed spoken interaction between exam candidates. The L2 data are produced by speakers coming from different cultural and linguistic backgrounds and with varied sociolinguistic characteristics. Other extensive metadata regarding the learners is attached to the corpus. Furthermore, the compilers have chosen to be transparent about their design choices, thus fully informing both the research communities of the new corpus's potential.

The TLC also addresses the second issue, namely the lack of spoken corpora. Indeed, the TLC is the largest spoken learner corpus available now and is the result of the collaboration between the Lancaster University's Centre for Corpus Approaches to Social Science (CASS) and Trinity College London (Gablasova et al., 2019). For more such corpora to be compiled and made available to the research communities, more collaborations are called for. Naturally, such a feat is not possible if a wide range of resources are not available to the researchers, which still remains the major problem for most of us. Myles (2015) reminds us that SLA needs spoken data because it is richer from the point of view of potential insights into learning processes and it is especially free of the conscious application of rules and control. Indeed, spoken data can provide useful information regarding the cognitive processes of foreign language use (Tomasello, 2003), or insights into the use of stance adverbs (Pérez-Paredes & Díez-Bedmar, 2019). The *International Journal of Learner Corpus Research* has seen a surge in papers investigating spoken learner language, which means that this issue is probably going to be solved very soon, due to the availability of rich spoken learner corpora such as TLC and a keener interest by LCR researchers into the processes of spoken language.

The last two issues arising from the clash between SLA and LCR, namely the overconcentration of studies on L1 as main variable and the preference of frequency counts rather than more qualitative and explanatory investigations, can be addressed concurrently. There is still a high number of studies in the literature taking the L1 as the most interesting variable and quick explanation of any differences between IL and TL (Callies, 2015). These studies should not be disregarded because they are not analysing other variables, on the contrary they should be valued as interesting and enriching perspectives. However, as shown in Ädel

(2008) and Gilquin and Paquot (2008) there are other variables that should be taken into consideration (cf. Gilquin, 2019). The reason why researchers concentrate on L1 is strictly linked to issue a), that is the lack of well-balanced corpora: if other variables are not controlled for, the only safe variable that can be analysed is the L1, and in turn, if the only variable that can be safely analysed is the L1, what other approach can be adopted if not CIA and therefore a more frequency-based account of the differences and similarities of the IL and TL?

The answer to these issues lies in the compilation of more varied corpora, such as TLC, which allow researchers to focus on multiple variables without the concern for data sparsity and can encourage other types of exploratory analyses not based on frequency. However, though TLC is one step forward towards meeting SLA, it is one corpus and the L2 is still English. Other L2s should also be investigated systematically (Myles, 2015) and with corpora similar to TLC. Perhaps, as we wait for more corpora to be compiled and become available, we should reconsider the two main functions of LCR and include a third one: provide frequency-based descriptions about the IL which can aid the study of specific learning processes by the SLA research community. Indeed, LCR should recognise its current limits and not necessarily push towards a cooperation with SLA earlier than what is actually feasible. New and more SLA relevant corpora will come, but for the moment perhaps it is enough to readjust the expectations and perform quantitative and qualitative analyses on the available data which can then be used by SLA researchers, if need be, to explain or find confirmation of their theories.

2.6 Summary

This chapter has offered a bird's-eye view of LCR, from the the early approaches to the study of learner language of the 1970s to more recent developments in LCR. During the 1970s the first pedagogical perspectives on learner language emerged through the analysis of errors. SLA studies in those years were influenced by Corder's 1967 work on learners' errors which led to the establishment of the notions of IL and TL. However, EA was highly criticised by Schacter and Celce-Murcia (1977), who questioned whether it was appropriate to analyse errors in isolation, produce frequency counts of errors and categorise such errors. In addition, they also objected to the more general means of collecting data through sampling procedures which were not balanced and systematic, but rather quite biased (e.g., the participants were a small group of students and did not represent the learner population well). On the other hand, SLA studies continued to collect learner language through elicited

collection methods, such as clinical or experimental data. It was not until the 90s that natural learner language became a feasible alternative to clinical and experimental data. With the advent of computer corpora and the pioneering project of Sylviane Granger (ICLE), LCR effectively earned its own field and a series of works began to populate the research field. The Chapter introduced the essential concepts of LCR, namely learner corpus design, variables collection, the notion of proficiency, and corpus typology, that is written vs spoken corpora, annotated vs unannotated corpora, and cross-sectional vs longitudinal corpora. Furthermore, by delving into the learner corpus compilation, a few core issues, such as authenticity and corpus size, emerged.

In terms of methodology, the literature review reports three main approaches to the analysis of learner language: CIA, the reappraised version of the CIA, and CEA. CIA aims to compare the learners' IL with the TL in order to shed light on features and patterns of IL. However, CIA has drawn much criticism from scholars who believe that comparing IL to a TL norm is unfair; following this, Granger (2015) has reviewed CIA and has proposed a reappraised version called CIA. The aim of CIA is to avoid the so-called comparative fallacy and offer researchers the possibility to compare IL varieties and reference language varieties, both from the perspective of IL varieties vs reference language varities, but also from the point of view of IL varieties vs IL varieties and reference language varieties vs reference language varieties. The third approach is CEA, which has received a warm welcome in the LCR community do to its relevance for teaching practices. The approach consists in tagging errors in a learner corpus (either manually or automatically) and then retrieving the tags in order to better understand learners' errors.

Despite the great deal of contributions that LCR has offered to the scientific community, both from the point of view of corpus linguistics and SLA approaches, the long-existing gap between SLA studies and LCR still has not been bridged. This is due to several reasons, the majority of which have to do with the different perspectives that the disciplines adopt as regards learner language. However, for the moment being, it is argued that LCR should readjust its expectations and perform more qualitative and quantitative analyses which, if necessary, SLA researchers can then exploit to explain or find confirmation of their theories.

Chapter 3

Phraseology

THE aim of this chapter is to present a bird's-eye view of phraseology. Phraseology can be defined in multiple ways and it is a multi-dimensional and vast field of studies, so that the present account will be a short overview of its key concepts in order to depict the landscape against which this research is set. The first Section introduces the first approaches to phraseology, namely those of the Russian scholars and the British researchers who then gave rise to the two main persepectives on formulaic language: the phraseological approach and the frequency-based approach. The second Section introduces key elements as regards phraseology, that is how formulaic sequences can be identified according to several parameters that have been proposed in the literature; the third Section describes the importance of phraseology for language since formulaic sequences are pervasive and processed differently compared to other language structures; the fourth Section zooms in on one specific set of formulaic sequences: collocations. These are the topic of the present research and the Section presents the main two approaches to the analysis of collocations and delves deeper into the frequency-based approach, which is the chosen perspective for the analysis. The fourth Section focuses on their importance for language learners and illustrates some of the most important findings stemming from psycholinguistic and LCR on collocations.

3.1 Historical approaches to phraseology

Phraseology has been defined as "the study of the structure, meaning and use of word combinations" (Cowie, 1994, p. 3168), but the concept is far from being easy to discern and categorise. Indeed, phraseology is inherently multidisciplinary and it involves multiple branches of linguistic research, such as morphology, syntax, semantics, discourse analysis (Granger & Paquot, 2008, p. 29–35), and first and second language acquisition. Phraseology has been the subject of numerous studies (cf. Allerton et al., 2004; Cowie, 1998; Granger &

Meunier, 2008; Nattinger & De Carrico, 1992; Nesselhauf, 2005; Schmitt, 2004; Sinclair, 1991; Wray, 2002; among others), which have generated several definitions, classifications, and approaches, and the most relevant ones will be discussed in the following Sections.

Phraseology was a restricted field of study until the 1970-1980s and the majority of studies prior to that mainly focused on formulaic phrases, such as idioms. However, collocations were one of the first formulaic sequences to be studied within the field of phraseology, with early work dating back to the 1950s. Firth (1951; 1957) provided one of the initial definitions of collocations (see Palmer, 1933 for the first definition; see also Section 3.4.1) as he described them as "actual words in habitual company" , or perhaps with the better well-known quote: "You shall know a word by the company it keeps" (Firth, 1957, p. 11–12). Firth not only defined collocations, but he also distinguished between different types of collocations based on their frequency: in his work he mentioned habitual collocations, that is, the accompanying words of other words which are most common or "most characteristically embedded", for instance *March hare* (Firth, 1957, p. 12). Habitual collocations and their individual components are in contrast with other collocations whose individual components may not be so restricted, and as such must be considered by themselves rather than inserted into the whole collocation, such as *light*, which may collocate with many other words (Firth, 1957, p. 12). Firth's approach was crucially based on frequency and was later developed by other scholars, such as Halliday, Mitchell, Greenbaum, Sinclair, and Kjellmer, who developed it into the frequency-based approach, which will be described in Section 3.4.1.2.

During the same time and following years, in the former Soviet Union and much of Eastern Europe, other scholars, such as Vinogradov (1947), Amosova (1963) and Mel'čuk (1988) were concerned with phraseology, in particular with conceptualising a descriptive framework of categories of formulaic sequences that was comprehensive and systematic. They distinguished between *sentence-like units* (Cowie, 1998, p. 4) or idioms, whose meaning is opaque (i.e., not deducible from the single components of the collocation) and whose function is pragmatic, and *word-like units* (Cowie, 1998, p. 4) or restricted collocations, in which the meaning of one word depends on its relationship to the other and whose function is syntactic. Their approach was later developed by Cowie and became the second other major approach to phraseology: the phraseological approach (see Section 3.4.1.1).

3.1.1 The importance of phraseological sequences for second language learners

Research in the Anglo-Saxon world picked up the pace at the beginning of the 1980s, after Pawley and Syder (1983) published an influential paper which postulated the importance

of formulaic language for second language fluency and native-like proficiency. The authors suggested that language is acquired in chunks and some sequences are retrieved as a whole from the speakers' repertoire (cf. Ellis, 1996). The concept was expanded in 1991 by Sinclair: he proposed the idiom principle and the open-choice principle. The open-choice principle is a simple "slot-and-filler model", whereby upon completing a unit of meaning the speaker is faced with a potentially infinite number of choices only restricted by "grammaticalness" (Sinclair, 1991, p. 109). However, these choices should not just be made on a mere grammatical level, but rather on several different levels and all at the same time: for instance, there is register, which requires a choice based on social conventions (Sinclair, 1991). On the other hand, the idiom principle is based on the idea that there is no randomness in the real world and hence there should be no randomness in language. Very often things that happen together tend to be mentioned together and Sinclair (1991) argued that every speaker has a set of "semi-preconstructed phrases" (p. 110) which are readily available and constitute the actual choices, rather than the single lexical or grammatical item that should be selected with the open-choice principle. Therefore, he concluded that the open-choice principle alone cannot explain the quasi-infinite choices that a speaker has to make on several language levels; one also has to account for the idiom principle which works together with the open-choice one in order to aid the speaker in their language choices. In other words, if a speaker were solely relying on the open-choice principle, they may not be able to produce language in a timely manner, unless the idiom principle is employed. It is therefore easy to conclude that if a(n) (L2) speaker has a series of prefabricated chunks of language ready to use, the processing effort is significantly lower, thus improving effective communication (Nesselhauf, 2005, p. 2).

3.2 Identifying formulaic sequences

Later studies on phraseology became concerned mainly with two issues: the identification of formulaic sequences and the classification of such identified formulaic sequences. The identification of formulaic sequences refers to which chunks of language can be considered phraseological or formulaic. For example, there are some sequences which are transparent and easy to identify, such as idioms, phrasal verbs, compounds, collocations, but there are some others which represent grey areas and need further investigation. One such area is represented by discontinuous expressions: should they be regarded as formulaic sequences nevertheless, or be excluded from the classification? Firth had proposed that discontinuous expressions such as *dark* and *night*, which may occur with several other words in-between,

should be considered collocations, but he did not provide further criteria, such as maximum number of words between the two elements, to classify them as a collocation.

One approach to the identification of formulaic sequences which addresses Firth's lack of criteria has been that of frequency in corpus linguistics. Indeed, the extraction of frequency counts can reveal the most common, and therefore fixed, patterns of specific sequences. Frequency is a salient and determining factor in the identification of formulaic sequences (Wray, 2002, p. 25), but it should not be regarded as the sole criterion. The extraction of frequencies requires the researchers to select a number of search criteria, such as number of co-occurring words, the distance between the two words (the span), and the length of the string. These criteria are usually set arbitrarily, according to the researcher and the corpus size, and although they may be well suited for written texts, they do not account for overlaps, interruptions, false starts in spoken texts. These are only a few of the issues that can arise when identifying formulaic sequences through frequency counts (for an in-depth analysis of frequency procedures and measures, see Gries, 2008).

Other approaches to the identification of formulaic sequences have relied on structure and form: for example, Moon (1998a; 1998b) defined formulaic sequences as the set of multi-word strings listed in a particular dictionary, but this approach did not inform researchers on the nature of formulaicity (Wray, 2002, p. 32). Another approach was proposed by Butler (1997) who identified the first occurring invariable word in a repeated sequence as a function word or discourse marker. For example, if the beginning of a sequence is a fixed preposition, then a noun or verb must follow. The main issue with this approach though is that it is incompatible with corpus searches, as the content of an open class slot will vary and a corpus search will not be able to recognise it as a recurrent sequence (Wray, 2002, p. 32). Another approach has been that of compositionality, that is, a formulaic sequence is identified on the basis of its internal composition which is completely free from the effects of grammar or lexicon. This approach stems from the observation that a sequence of words, once it is formulaic, is no longer subject to the effects of grammar and lexicon. In other words, the sequence is no longer obliged to follow grammatically regular or semantically logical rules. The sequences become "frozen, or fossilized" (Wray, 2002, p. 33) and may often retain words or grammatical forms which are no longer in current use, such as *If I were you*. However, as Wray (p. 33) rightly pointed out, this approach is limited to an exclusive set of sequences and does not extend to other regular sequences, which are formulaic but are still subject to grammar and semantic rules. This approach is also strictly related to the criterion of fixedness for the identification of formulaic sequences: fixed expressions are only a small sub-set of formulaicity and the dynamic nature of language does not allow for fixedness to be the sole criterion. Indeed, one fixed expression today, could be jokingly changed and

become lexicalised tomorrow, in addition to all other semi-fixed expressions that are present in the language. As regards formulaic sequences in spoken language, Pawley (1986) proposed fluency as another approach, postulating that "pauses within lexicalised phrases are less acceptable than pauses within free expressions, and after a hesitation the speaker is more likely to restart from the beginning of the expression" (p. 107). This criterion seems too subjective and dependent on the speaker to be able to provide rigorous schemes for the identification of formulaic sequences and it would only be applicable in spoken texts, thus excluding all written texts (which account for the majority of corpora today).

The brief overview of these approaches has underlined their insights and limitations which inevitably lead to share Wray's (2002) conclusion that a combined approach, which takes into consideration all these factors, is the best solution to the identification of formulaic sequences. However, the main issues with these studies and their approaches is that there is now a plethora of definitions and different categories of formulaic sequences[1] which are based on diverse or even contrasting criteria. Perhaps, a viable solution can be found in Gries (2008), who having confirmed the lack of a universal definition of formulaic sequences, proposed the following six parameters for the identification and extraction of phraseological units:

1. The *nature* of the elements involved in a phraseologism;
2. The *number* of elements involved in a phraseologism;
3. The *number of times* an expression must be observed before it counts as a phraseologism;
4. The permissible *distance* between the elements involved in a phraseologism;
5. The degree of *lexical* and *syntactic flexibility* of the elements involved;
6. The role that *semantic unity* and *semantic non-compositionality/non-predictability* play in the definition[2]. (p. 4).

In a way, these six dimensions conflate all the issues raised by the previous approaches into one explicit scheme that could be feasibly adopted in the identification of phraseologisms. As far as the first criterion is concerned, Gries explained that a phraseologism is the co-occurrence of any linguistic element or grammatical pattern. Secondly, phraseologisms can contain more than two words and, thirdly, the phraseologism has to occur more frequently

[1] For a comprehensive list, see Wray (2002).
[2] Gries (2008) proposed yet another label for phraseological sequences: *phraseologisms.*

than would be expected by chance. As regards the distance or span of the phraseological sequence, Gries adopted a wider perspective and recognised discontinuous sequences as phraseologisms. As far as the lexical and syntactic flexibility is concerned, Gries' definition of phraseologism distinguishes between studies which involve completely inflexible patterns (such as *by and large*), relatively flexible patterns (such as *kick the bucket*, which allows for different tenses but no passivization), partially lexically-filled patterns (such as *into*-causative), completely lexically unspecified, maximally flexible expressions (such as the English ditransitive pattern). All the patterns identified in these fields of studies are accepted by Gries except for those which include completely lexically unspecified expressions as this would not respect the first parameter of having at least one lexically specified element. Lastly, for a sequence to be identified as phraseological, Gries proposed that semantic unity is required, but not non-compositional semantics. The six dimensions lead to the definition of phraseologism as the co-occurrence of any linguistic items of various kinds which "function as a semantic unit in a clause or sentence and whose frequency [...] is larger than expected on the basis of chance" (Gries, 2008, p. 6). This definition of phraseological sequence is very broad and may lead to the identification of a high number of formulaic sequences. However, frequent co-occurring sequences such as *of the*, or *in the* are automatically excluded from identification because they do not function as a semantic unit. Gries' approach seems to be a practical solution to the identification of formulaic sequences as it conflates all the relevant issues of phraseology and combines into a neat scheme, which certainly selects a high number of phraseologisms, but provides a rigorous starting point.

The classification of identified formulaic sequences poses similar problems and is far from being a methodical and rigorous process. There are currently many categories of formulaic sequences, such as collocations, lexical bundles, idioms, proverbs, phrasal verbs, and they can be broadly divided into two categories: sequences that perform a pragmatic function (*sentence-like formulae*), and sequences that perform a syntactic function (*lexeme-like*) (Pawley, 2007). Pragmatic sequences are mainly concerned with conveying a message to the hearer and, on occasions, the individual meaning of the words may not always reflect the overall meaning of the sentence, for example it may depend on the interlocutors, the topic of the message, and most importantly the context. This category may include but not be limited to idioms, proverbs, catchphrases, sayings, and others. On the other hand, syntactic sequences establish a grammatical relationship between one element of the sequence and another and function at a syntactic level, that is, within a sentence, rather than within a wider context, such as that of pragmatic sequences. Sequences in this category may include but not be limited to collocations, lexical bundles, and phrasal verbs.

As regards pragmatic sequences, which have not been as widely investigated as syntactic

ones, Pawley (2007) suggested that they are usually a combination of the following components:

- Segmental phonology
- Music (tempo, rhythm, melody, voice quality)
- Grammatical category
- Grammatical structure
- Idiomaticity constraints
- Literal meaning
- Pragmatic function
- Body language

For example, he mentioned *I'm pleased to meet you* as a sequence with a pragmatic function (response to interlocutor), which should be spoken with a bright tone, the main stress is on *meet*, eye contact is made with the interlocutor and it is usually accompanied by a handshake. From the point of view of grammar or lexicon, it is fairly restricted as *pleased* can be replaced by a limited number of synonyms. Although previous work on pragmatic formulae had focused on the characteristics of these sequences (Pawley, 1986; 1991; 2001; Pawley & Syder, 1983), a more recent attempt has expanded the pragmatic function window and has gone as far as distinguishing three main pragmatic functions of formulae: i) processing, whereby the formulae are used to reduce the cognitive and processing impact of the speaker; ii) interaction, which is reflected in the language choices according to the interlocutor (also called the manipulation of the hearer); iii) discourse marking, whereby the sequences are used to mark and organise the discourse structure (Wray, 2002). Wray's model proposes a different perspective on formulaic language which becomes even more interesting when applied to SLA and is a further categorisation that could be adopted in the study of formulaic sequences.

Another study into the pragmatic functions of formulae has identified another sub-set of formulaic sequences, lexical phrases (Nattinger & De Carrico, 1992). Lexical phrases can either be strings of specific lexical items which may follow the grammar, or generalised frames which include category symbols and specific lexical items. Four criteria can help identify lexical phrases: i) the length and the grammatical status; ii) the canonical or non-canonical shape; iii) variability or fixedness; iv) continuous or discontinuous string of lexical

items. Four major categories of lexical phrases are thus identified: i) polywords, which function as a single word, are fixed and do not allow for lexical insertions and include two-word collocations; ii) institutionalised expressions, which are sentence-like, invariable, and mostly continuous (*nice to meet you*); iii) phrasal constraints, which allow variations of lexical or phrase categories and are mostly continuous (*a _ _ _ ago*); iv) sentence builders, which represent the less fixed category, they help the speaker build sentences and have fillable slots (*I think that*).

A further category of formulaic sequences which has drawn attention is that of lexical bundles (Biber et al., 1999). Lexical bundles are combinations of three or more words that are identified in a corpus by means of corpus analysis. Indeed, Biber et al. (1999) investigated re-occurring sequences of language in a corpus by setting the boundaries for the analysis of co-occurrences appearing at least ten times per million words across at least five different registers. Lexical bundles are particularly frequent in academic writing and uniquely inherent to specific disciplines (Cortes et al., 2002). Biber (2006) found that lexical bundles are abundant in university language, but they are not evenly distributed: they are more frequent in the hard science than in the soft sciences. Furthermore, *referential bundles* (that is sequences that refer to an entity in the text or outside the text) are the most common, *stance bundles* (which express attitudes or opinions) are the second most common, while *discourse organisers* are the least frequent. Lexical bundles are inherent to academic discourse and there have been several studies focusing both on the written register (Ädel & Erman, 2012; Biber & Conrad, 1999; Chen & Baker, 2010; Cortes, 2004; Durrant, 2017; Hyland, 2008; among others) and the spoken register (Biber & Barbieri, 2007; Biber et al., 2004; Conrad & Biber, 2005; Nesi & Basturkmen, 2006; Wang, 2017; among others).

3.3 The pervasiveness and processing of phraseology

Before coming to the end of this brief overview of key studies on phraseology, two further points should be mentioned: the pervasiveness of phraseology, and the processing of formulaic sequences. The literature on these topics is vast and a full account would be beyond the scope the present work. However, in this Section we propose a few essential studies on the pervasiveness and processing of phraseology which are mainly based on corpus linguistics (as regards pervasiveness) and experimental data (as regards the processing).

In the 1970s, Becker (1975) and Bolinger (1976) reached the conclusion that phraseology was pervasive to language, but their intuition was only confirmed later with the advent of corpora and corpus-based studies. Indeed, several scholars have demonstrated the pervasive

nature of phraseology in written and spoken language (Altenberg, 1998; Bestgen & Granger, 2014; Biber et al., 1999; Cowie, 1991; 1998; Granger & Meunier, 2008; Howarth, 1998a; Kjellmer, 1994; Meunier & Granger, 2008; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983; Sinclair; 1991; Stubbs, 2007; among others) and its central role in the acquisition of first and second languages and adult language production (Cowie, 1998; Pawley & Syder, 1983; Peters, 1983; among others). An interesting study on written language showing the pervasiveness of phraseology was conducted by Stubbs (2001): he counted the frequency of phraseological units which had a word-form beginning with the letter *f* in a sample of 1,000 content words. He found that all 47 words beginning with *f* were inserted in a phraseological unit. As far as spoken language is concerned, Altenberg (1998) examined word combinations extracted from the *London-Lund Corpus of Spoken English* and concluded that "over 80% of the words in the corpus form part of a recurrent word-combination in one way or another" (Altenberg, 1998, p. 102). The main limitation of Altenberg's study lies in the retrieval of the combinations from the corpus: he did not qualify word combinations. He only extracted uninterrupted strings of words occurring more than once, without providing any classification or distinction.

As regards the mental processing of phraseological sequences, it is known that our brain contains representations of individual words and other chunks of language, which it conveniently retrieves when needed (Pawley & Syder, 1983). Furthermore, the literature has shown that formulaic language is processed differently from non-formulaic language and several studies have confirmed Sinclair's (1991) idiom principle whereby our brains store sequences of language in the long-term memory which are then retrieved automatically without the need to compose them online through the open-choice principle (cf. Conklin & Schmitt, 2012). However, the evidence of phraseological processing is mixed for non-native speakers as it will be described in Section X.

The vast majority of research into phraseological processing has been conducted on idioms and other non-idiomatic formulaic sequences (e.g., lexical bundles, binomials, etc.). Research on idioms has mainly focused on the processing speed of figurative vs literal meaning of idioms and the processing of idioms vs novel sequences. One such study was done by Swinney and Cutler (1979) who proposed a model for the processing of formulaic language: the lexical representation hypothesis. According to the authors, idioms are stored in the mental lexicon as morphologically complex words. When we encounter an expression, our brain simultaneously computes the literal and figurative meaning of the sequence, but since the literal meaning requires more processing time, the figurative meaning is activated first. Consequently, they found that formulaic sequences are processed faster than non-formulaic sequences. This was also confirmed by the studies of Gibbs (1980) and Van

Lancker et al. (1981).

As regards other non-idiomatic formulaic sequences, Sosa and MacFarlane (2002) and Kapatsinski and Radicke (2009) tested the comprehension of formulaic language in native speakers through an auditory word-monitoring task. The former tested the response rates of participants in identifying the particle *of* inserted in two-word frequent (*kind of*) and infrequent (*piece of*) sequences. The slower response rates for frequent sequences indicated that participants probably process and store these sequences holistically and as such the particle is fused within the combination (Siyanova-Chanturia & Van Lancker-Sidtis, 2018, p. 40) and harder to identify as opposed to more infrequent sequences. Similar results were obtained by Kapatsinski and Radicke (2009) in a more controlled study testing the response times of the identification of the particle *up*.

3.3.1 The processing of formulaic language in non-native speakers

As described in Section 3.3, the literature generally agrees that formulaic sequences are processed faster than non-formulaic ones by native speakers. Cieslicka's (2006) study on non-native speakers about non-defining sentences containing familiar idioms required the participants to perform a lexical decision task on figurative and literal targets. The faster response times on the literal targets suggested that non-native speakers process the literal meaning faster than the figurative. However, a study by Conklin and Schmitt (2008) which employed a self-paced line-by-line reading scheme and involved native and non-native speakers, found that formulaic language was read (and hence processed) faster than non-formulaic language, both in native and advanced non-native speakers. Similar studies confirmed this finding (Carrol & Conklin, 2014; 2017; Swinney & Cutler, 1979;), but Gyllstad and Wolter (2016) found that non-native speakers process idioms slower than non-formulaic language, although they crucially did not take the L1 variable into consideration and relied on meaningfulness judgement tasks. The latter may not be appropriate for the investigation of formulaic language processing of learners as there is a cognitive gap between what is processed online and what is reflected upon during the judgement task.

Thus, it seems that, unsurprisingly, proficient learners of a second language process formulaic language similarly to native speakers, whereas lower proficiency users tend to process words individually and struggle with opaque sequences such as idioms (Conklin & Schmitt, 2012). Perhaps an explanation for the learners' struggle with formulaic language may be found in the higher cognitive load required of learners since they already have their L1 repertoire of formulaic language which may interfere with the processing of the L2. Indeed, as far as idioms are concerned, Conklin & Schmitt confirm that processing is faster

for learners when it comes to congruent sequences, that is phraseological units that share the same form and meaning in the L1, whereas the process is slowed down by incongruent idioms, that is expressions that do not share the same form and meaning in the L1. Since the L1 has been demonstrated to interfere with L2 processing, this variable will be taken into consideration in the present study and will be touched upon in Section X. Although idioms have been shown to be processed more quickly than non-formulaic language by native speakers, it is still unclear whether the figurative meaning is activated before the literal meaning.

More recent research into phraseological processing has adopted eye-movement tracking technology, which eliminates environmental variables (such as pushing a button) and is a closer approximation of natural reading processes (Conklin & Schmitt, 2012, p. 49). Underwood et al. (2004) investigated the recognition of formulaic sequences in texts by native and non-native speakers. They found that native speakers lingered less on the final part of the sequence as they had presumably recognised the formulaic unit, and this was partially shared by the non-native speakers. Similarly, Siyanova-Chanturia et al. (2011a) found that non-native speakers read formulaic sequences faster than non-formulaic ones. Although some of the research findings point towards different directions, as far as idioms are concerned, there is a tendency for non-native speakers to process formulaic language faster than non-formulaic language, except when the meaning of such sequences is more figurative than literal.

With regard to lexical bundles and non-native speakers, Jiang and Nekrasova (2007) tested the response rate and accuracy of native and non-native speakers during a grammaticality judgement task on frequent and infrequent lexical bundles. Both native and non-native speakers performed faster and more accurately on frequent lexical bundles and the authors argued that this was due to the fact that frequent phrases were not subject to a full syntactic analysis by the participants as opposed to infrequent ones. Although this was one of the first studies testing non-native speakers' processing of frequent phrases, it could be argued that the authors' conclusion is not based on any empirical evidence, since no syntactic analysis was actually required in the task (Edmonds, 2014; Siyanova-Chanturia, 2015). A similar study was conducted by Siyanova and Schmitt (2008) on native and non-native speakers' response rates to phrases ranging from high frequency to low frequency. As regards native speakers, the findings confirmed that they are faster at processing frequent vs infrequent phrases, and the same effect is maintained when dealing with high frequency vs medium frequency phrases. Non-native speakers behave similarly when it comes to frequent and infrequent phrases, but there is no processing time difference between high frequency and medium frequency phrases.

Other interesting and insightful studies have been conducted over the years on a number of other non-idiomatic formulaic sequences, such as *n*-grams (Forchini & Murphy, 2010; Hernandéz et al., 2016), speech formulas (Tremblay et al, 2011), binomials (Arnon & Snider, 2010; Siyanova-Chanturia et al., 2011), compounds (Badecker, 2001; Badecker & Allen, 2002; Juhasz, 2007; Libben, 1998; among others), and formulaic language processed by impaired participants (such as native speakers suffering by aphasia or other left- or right-brain injuries) (Van Lancker & Kempler, 1987; Van Lancker-Sidtis & Postman, 2006), but a comprehensive overview of this would be beyond the scope of this Section and work. See the works of Conklin and Schmitt (2012) and Siyanova-Chanturia and Pellicer-Sánchez (2018) for a more in-depth analysis.

3.4 Collocations

As seen in the previous Sections, there is an abundance of phraseological sequences in the literature – idioms, lexical bundles, phrasal verbs, binomials – among which there are collocations. Collocations are the object of this study, and as such deserve an introduction and explanation. This Section provides a summary of L1 studies on the definitions and classifications and presents the two most common approaches to collocations: the phraseological and the frequency-based approach. Furthermore, some of the most important empirical LCR reseach findings on collocations are illustrated (i.e., overuse/underuse, misuse, L1 congruency, productive knowledge gap); these are followed by a Section on statistical tests for measuring the strength of association of collocations.

3.4.1 Collocations and the two approaches

The term collocation has been widely used in theoretical and applied linguistics. There are many diverse definitions and approaches to collocations, but two main views have been particularly relevant in the literature (Nesselhauf, 2005): the phraseological approach and the frequency-based approach. A few scholars (Granger & Paquot, 2008; Nesselhauf, 2005) agree about the duplicity of the definition according to the theoretical perspectives of these two main approaches to the study of collocations (and other word combinations in general). Perhaps the only common denominator of all definitions is the fact they all refer to a syntactic relation of words (Nesselhauf, 2005, p. 11).

The concept of collocation can be traced back to the 1930s when Palmer (1933) defined it as “a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts”. Palmer adopted the term collocation to refer

to all combinations of words. His attested definition of collocation precedes that of Firth (1951; 1957) (see Section 3.1) and already contained the building blocks of the concept of collocation, namely the idea that two or more words occur together at a syntagmatic level, they are somewhat fixed, and should be learnt as a whole, this last point being of particular relevance for language learners, as it will be described in Section 3.5. Palmer conceived collocations with a pedagogical intention and did not provide a clear explanation or categorisation about the relationships between two or more words.

In the 1950s, the notion of collocation as we know it was introduced by Firth, the attested father of collocations, who provided the famous quote: "You shall know a word by the company it keeps" (Firth, 1957, p. 11). Following Firth's concept, later researchers proposed a reprised version of collocation: "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991, p. 7). This means that collocations are words that collocate with each other and co-occur more often than what would be predicted (Jones & Sinclair, 1974, p. 19) (cf. Gries, 2008)[3]. Firth's concept of collocation was later developed by other scholars, who became known as the Neo-Firthian school and who gave rise to the *statistically oriented approach* by Herbst (1996, p. 380) or the *frequency-based approach* by Nesselhauf (2005). This approach is usually employed by researchers in the field of corpus linguistics.

Concurrently with the development of the frequency-based approach, collocations came to be studied from another perspective, which viewed them as semi-compositional and lexically determined combinations (Evert, 2007). In this view, collocations are semi-fixed combinations and this approach has been called the *significance oriented approach* by Herbst, or the *phraseological approach* by Nesselhauf (2005). This approach is usually adopted by researchers in the field of pedagogy and/or lexicography.

A further approach to the definition of collocation has been that of Leech (1974), who stressed that "a word acquires [psychological association] on account of the meanings of words which tend to occur in its environment" (p. 20), thus introducing a more psychological perspective to the concept of collocations.

3.4.1.1 The phraseological approach

The most representative scholar of the contemporary phraseological approach is the British researcher A.P. Cowie: as regards collocations, he defined them as a type of word combination which is distinct from idioms and from free combinations, thus placing them in the middle of the continuum and terming them "restricted collocations" (Cowie, 1981; 1994).

[3] cf. the definition of *colligation* in Section 3.4.1.2.

His classification of word combinations is one of the most precise (Nesselhauf, 2005, p. 14): he distinguished between composites and formulae. The latter are combinations that function at a pragmatic level, while the former (which include collocations) have a syntactic function. Composites are further divided based on the transparency of meaning (criterion of transparency) and the commutability of the items (criterion of commutability). Transparency refers to whether the combination has a literal meaning, that is, if the meaning is inferable from the meaning of its own items (the so-called semantic compositionality), while the possibility to replace an item of the combination indicates the measure to which this is restricted.

Transparency is measured in terms of whether the meaning of the combination can be deduced from the literal meaning of the components of the combination. This criterion is particularly efficient for the distinction between idioms and non-idiomatic expressions (Hausmann, 1989): for idioms the meaning is opaque, the components do not individually make up the overall meaning of the combination (i.e., *go bananas*, *to give the sack*); while the meaning of non-idiomatic expressions can easily be inferred from the individual components. However, a problem arises with non-idiomatic expressions, because they may include collocations or free combinations (e.g., *commit a crime*, a collocation, or *control the crime*, a free combination). A further criterion is needed to aid the differentiation between collocations and free combinations: commutability distinguishes between free combinations, in which one element can be substituted with another without changing the second element's meaning, and collocations in which the commutability of their elements is restricted (Aisenstadt, 1979; Cowie, 1992; Howarth, 1996).

From this distinction and criteria, Cowie (1998) proposed the following four types of combinations which can be found on a phraseological continuum:

– Free combinations (e.g., *drink tea*): the items of the combinations can be freely replaced on the basis of their semantics and are used in their literal sense.

– Restricted collocations (e.g., *perform a task*): substitutions are allowed, but there are restrictions. At last one element of this combination is used with a non-literal meaning, and at least one is used in its literal sense. Thus, the commutability is restricted because of one of the two terms is specialised. The combination is transparent.

– Figurative idioms (e.g., *do a U-turn*, used in the sense of changing policy or behaviour): substitution is rarely possible; the combination has a figurative meaning but the possibility of literal interpretation is preserved, such as in *to*

do a U-turn, which not only refers to a driving manoeuvre, but also to politics or behaviour.

– Pure idioms (e.g., *blow the gaff*): no substitution is allowed; the combination has a figurative meaning and there is no possibility of a literal translation since its meaning is not compositional.

Free combinations and restricted collocations are sometimes both used by Cowie to refer to collocations, but when this happens, he uses the term "open collocations" (free combinations) and "restricted collocations". There are variations of Cowie's terminology as identified by Nesselhauf (2005): on occasions Cowie categorised combinations such as *foot the bill* as restricted collocations (Cowie, 1998, p. 221) since at least one item of this combination has a specialised meaning (in this case *foot*), but other times he proposed a novel category for these combinations which could be placed between idioms and collocations (Cowie, 1981, p. 228). Furthermore, Cowie assumed that all elements of a collocation are lexemes, but in 1994 he suggested that these items may comprise inflectional and derivational forms of a word (p. 3169). The variations in terminology in Cowie's works reflect the high degree of terminology variations in the phraseological approach literature. Indeed, other authors adopt different terms: for example, Benson et al. (1997) and Bahns (1993) used "collocation" to refer to Cowie's restricted collocations, while they term "free combinations" or "co-creations" Cowie's open collocations. The terminology also overlaps or differs on the other end of the continuum, with more restricted and opaque combinations: Cowie distinguished between figurative and pure idioms, while Benson et al. (1986) proposed a fifth additional category in-between collocations and idioms, "transitional combinations" or "transitional collocations" (p. 254).

Cowie's two criteria of transparency and commutability have also been used by other scholars, such as Aisenstadt (1981) and Hausmann (1989): on the basis of these criteria, they distinguished between collocations and free combinations, and between collocations and idiomatic expressions. Hausmann (1989) considered restricted commutability as the main distinguishing element between collocations and free combinations, and transparency as the main factor for the difference between collocations and idioms (p. 1010). Aisenstadt (1981) instead only relied on the criterion of commutability to distinguish between restricted and free collocations, while she employed both criteria for the distinction between restricted collocations and idioms.

However, as useful for the classifications of combinations as Cowie's criteria of commutability and transparency are, what was lacking in the commutability principle was a description or classification of the components of the combination: Cowie proposed that

one of the two terms can be commuted, but he did not specify which of the terms or their syntactic relationship. Hausmann (1989) and Mel'čuk (1998) argued that the elements of a combination are distinct and have a specific relationship. According to Hausmann (1989), one of the two components, the one chosen first, is semantically independent (*Basis*) and the selection of the second component (*Kollokator*) depends on the former. Mel'čuk (1998) instead, proposed the concept of "keyword", which is a component freely chosen on the basis of its meaning, and "value" is the group of elements that the keyword has automatically selected in order to express a given meaning. For example, in *do a favour*, *do* is the value and *favour* is the keyword. Nesselhauf (2005, p. 15–18) pointed out that the main difference between Mel'čuk's "value" and Hausmann's *Kollokator* lies in that the former includes all the elements that can be combined with a keyword to express a given meaning, while the latter refers to one single element.

Mel'čuk (1998) recognised that collocations are a focal point that need further attention among phrasemes (p. 24): he inserted collocations within the framework of his Meaning-Text Theory (Zolkovskij & Mel'čuk, 1965) and described them in terms of their lexical functions. A lexical function is a unit which contains a general and abstract meaning and carries a deep syntactic role that can be expressed in a range of ways depending on the lexical unit, the keyword, to which it is applied (Cowie, 1998, p. 7). Mel'čuk identified a set of sixty Simple Standard lexical functions and from these a large number of collocations can be selected. Thus, Mel'čuk assumed a relational connection between the elements of a collocation: the choice of the collocate is subordinated to the base according to the meaning that the combination of the two aims to convey. To cite the earlier example, *do a favour*, *favour* was chosen for its meaning, while *do* was chosen because of its relation to *favour*, and certainly not for its meaning because other verbs, such as *make*, could have replaced it.

Mel'čuk's lexical functions' approach is an exhaustive and comprehensive scheme for the identification of collocations, which, according to the author, should always be mapped and transcribed for the language under investigation[4] (Mel'čuk, 1998, p. 24). Mel'čuk's approach and, more in general, the phraseological approach, imply a very articulate and complex structure. This is further complicated by the fact that, although many scholars share the same general view, they create and use a different terminology, thus increasing the number of definitions and classifications. However, this overview of the phraseological approach has shown that effort has been made to identify the underlying criteria and relations between collocations and their components. This is certainly a necessary perspective in disciplines such as pedagogy, lexicography, and language acquisition. For the present work,

[4]For a more in-depth overview of lexical functions, see Mel'čuk (1998).

the frequency-based approach was chosen over the phraseological approach for two simple reasons: the first is to facilitate the extraction of combinations from a corpus (by adopting the surface co-occurrence approach which will be detailed in the next Section), and the second is that the frequency-based approach provides more objective, easily identifiable, and comparable criteria for the identification of collocations.

3.4.1.2 The frequency-based approach

The frequency-based approach, or the statistically oriented approach[5], is slightly more recent than the phraseological one (as it can be traced back to Firth's 1957 definition of collocation) and adopts a bottom-up (or corpus-driven, as we will see) approach to identifying collocations, rather than the more deductive top-down phraseological approach. The frequency-based approach completely changed the language-based criteria conceptualised by the Russians and proposed a new categorisation of phraseological units based on lexical co-occurrences. This approach generates a wide range of new combinations (such as frames and colligations), some of which had not been identified by the previous approach and as such "pushed the boundary that roughly demarcates the 'phraseological' more and more into the zone formerly thought of as free" (Cowie, 1998, p. 20). Indeed, as said before in Section 3.1, phraseology is a multidisciplinary field of study and there is a wide spectrum of definitions and classifications which vary according to the perspective adopted. In this way, many combinations that had been regarded peripheral or not formulaic at all by the phraseological approach, have acquired a central role in the frequency-based approach on the basis of their frequency in the language. This approach considers collocations as the co-occurrence of two or more words in a specific span and distinguishes between frequent co-occurrences (or, more precisely, more frequent that expected if the words were combined randomly in the language (Gries, 2008; Hoey, 2005; Moon, 1998; Sinclair et al., 2004; Stubbs, 2001) and non-frequent co-occurrences.

The frequency-based approach dates back to J. R. Firth's (1957) formulations of collocations, which were later developed by the researchers of the Neo-Firthian school, such as Halliday (1966), Hoey (1991; 2005), Kjellmer (1987; 1994), Lewis (2000), Moon (1998), Sinclair (1991; 2004) and Stubbs (1996; 2001) among others. In particular, Sinclair (1991), the main representative of the Firthian approach, defined collocations as "the occurrence of two or more words within a short space of each other in a text" (p. 170). The space to which he referred is called the "span" and it has been set to 4 to the left and right of the central word (node) that affects the collocability of the other components. The span was ideally set

[5]Or the distributional approach as defined by Evert (2009).

to ± 4 after a study by Jones and Sinclair (1974) found that significant collocates (the other components of the phrase) usually occur within a span of 4, four words to the left and four words to the right of the node. The frequency of collocations is not based on the simple count of how many times they occur in a given language, but it is set at a frequency higher than what would be expected by random chance (Gries, 2008; Jones & Sinclair, 1974, p. 19).

Sinclair's notion of collocation was also used in determining word sense for lexicography purposes, such as for the *Collins COBUILD English Dictionary* (Sinclair, 1995), since the collocates can distinguish the meaning of the node according to the target message. Furthermore, this approach identified other abstract collocational entities such as colligations. Colligation was used by Firth (1957; 1968) to describe the syntactic relationship between grammatical categories (i.e., [verb of perception + object + bare infinitive/-ing form] – *I heard my neighbours fight*), as opposed to the syntactic relationship between lexical items of collocations. Sinclair (1998, p. 15) perfected this concept and provided the most common definition of colligation: the attraction between a lexical item and a grammatical category. For example, the verb *budge* is attracted to the grammatical construction [modal auxiliary verb + *budge*], such as in *will/won't budge* (Sinclair, 1998, p. 13). Hoey's work (2005) on lexical priming employs the concept of colligation to cover the distributional attraction patterns of linguistic items, so colligations can be used to explain the positive or negative relationship between lexical items and grammatical categories[6].

More recent developments in the Firthian approach have seen Evert (2005) propose another definition of collocation: "a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon" (p. 17). This definition includes the concept of prediction, or lack of it. This is a wide definition which can cover multiple types of collocations, some more restricted than others, but Evert suggested that, according to the research question, one can further restrict the field and choose to analyse a specific set of combinations. Perhaps in an attempt to address this issue, Evert (2009) later distinguished between empirical collocations and lexical collocations. Empirical collocations can be interpreted as "empirical statements about the predictability of word combinations: they quantify the 'mutual expectancy' (Firth, 1957, p. 181) between words and the statistical influence a word exerts on its neighbourhood" (Evert, 2009, p. 1214). Lexical collocations (also broadly called multiword expressions, see Sag et al., 2002) belong to the phraseological approach and are semi-compositional and lexically determined combinations such as *stiff drink*, *heavy*

[6]For more on colligations and lexical priming, see Hoey (2005).

smoker, *a school of fish*.

As regards empirical collocations, Evert (2009, p. 1212) proposes that they be operationalised according to three perspectives: as *surface co-occurrences*, *textual co-occurrences*, or *syntactic co-occurrences*. The surface approach is the broadest and perhaps most employed in the literature: the collocation is measured based on the span between its components. Since span size is set rather arbitrarily, this may cause some issues of identification when dealing with languages which have a relatively free word order, so that strong collocates may be found further away from the node (Evert, 2009). The textual approach regards word combinations that occur in the same textual unit, rather than the whole corpus as with the surface approach. The textual units are either sentences, utterances, or documents. This approach may be more appropriate than the surface co-occurrence as it eliminates the issue of span size and co-occurrence is measured in the same sentence. Lastly, the syntactic approach is the most inflexible of the three and it only accounts for word combinations which have a precise syntactic relationship, such as verb + object, verb + noun, or adverb + adjective.

Regardless of the approach taken, corpus linguistics allows researchers to automatically extract large quantities of data, which can then be analysed from the point of view of frequency or other collocational measures. The present work does not rely on a frequency measure, but it only emplyos additional measures that take into consideration the strength of attraction of the collocations. Some of these most common measures of attraction between collocates are lexical association measures, which assign a score to the word pair thus allowing the researcher to either set a score threshold (*threshold approach*) or classify the collocations on the basis of their association strength (*ranking approach*) (Evert, 2009). These will be presented in Section 3.5.2.

3.5 Collocations in language learning

As mentioned in Section 3.3, formulaic language plays an important role in SLA and learning as it is an essential component for the development of fluency (Ellis, 2002; 2003; Ellis et al., 2015; Howarth, 1998a; Lewis, 1997; 2000; Nattinger & De Carrico, 1992; Meunier & Granger, 2008; Nattinger & De Carrico, 1992; Wray, 2002) and effective comprehension. Lewis's (2000, p. 45) statement about collocations being "the most powerful force in the creation and comprehension of all naturally-occurring texts" could be extended to all formulaic language. There have been multiple studies on phraseological units, such as collocations, phrasal verbs, compounds, idioms, that have verified either the positive or neg-

ative influence on the linguistic competence and output of learners (Ellis, 2008, p. 5), but this Section will only focus on collocational studies in a language learning context. The Section will try and offer a bird's-eye view of the most relevant studies thus highlighting different approaches and different foci.

The reason why collocations have been and continue to be one of the most studied phraseological combinations could be attributable to three factors: firstly, a learner's knowledge of grammatical rules may lead to the production of grammatically correct utterances, but this knowledge does not necessarily result in the correct utterance of multiword units (Biber et al., 1999; Wray, 2002). Correct production of formulaic units is particularly relevant if the learners aim to achieve native-like fluency (Cowie, 1992, p. 10); secondly, formulaic language in general does not only aid productive knowledge, but also receptive knowledge, that is comprehension of texts (Hunston & Francis, 2000); thirdly, researchers agree about the important role of collocations in second language learning, but they have also demonstrated that learners tend to process them slower than native speakers (Siyanova & Schmitt, 2008; see Section 3.3.1), and this hinders the potential benefits of collocational knowledge.

As mentioned in the previous Section (3.4.1), collocations can be defined either based on the phraseological or the frequency-based approach. However, these are often combined with other methods depending on the research questions(s): for example collocations can be studied through the use of elicited data or natural language use (i.e., learner corpora). As regards elicited data, several studies employed the use of translation tasks (Bahns & Eldaw, 1993; Biskup, 1990; Farghal & Obiedat, 1995; Hasselgren, 1994), word-combination tests (Channell, 1981; Granger, 1998b; Gyllstad, 2005; Siyanova & Schmitt, 2008; Wolter & Gyllstad, 2011; Yamashita & Jiang, 2010), cloze tests (Al-Zahrani, 1998; Bahns & Eldaw, 1993; Schmitt et al., 2004c), and blank-filling tasks (Farghal & Obiedat, 1995; Hoffman & Lehman, 2000; Zhang, 1993) for the study of collocations. The use of elicitation methods for the study of collocations certainly has several benefits, such as the accurate observation and analysis of the learners' productive and receptive knowledge on a pre-selected group of collocations (which may be infrequent collocations), the possibility of collecting – through acceptability judgments – insightful data into the learners' intuition of collocations, and lastly the direct comparison of study results based on the same collection criteria.

For example, both Bahns and Eldaw (1993) and Ferghal and Obiedat (1995) selected a restricted set of combinations (15 English verb + noun and 22 common English collocations respectively) for their studies while employing both a translation task and cloze test the former, and both a translation task and blank-filling exercise the latter. These studies were able to apply a high degree of control on variables such as L1, participants, and tasks, which would be otherwise rather difficult in a natural language data context. As regards introspec-

tive data, studies by Channell (1981), Granger (1998b), Gyllstad (2005), and Siyanova and Schmitt (2008) administered slightly different acceptability judgements and were thus able to obtain data about the learners' knowledge of collocations. Elicitation-based data also allow for comparisons, such as the one performed by Bahns and Eldaw (1993) between the results of the translation task and the cloze test (which were not statistically significant), or the one found in Farghal and Obiedat (1995): the authors compared two groups of learners with different levels of proficiency performing similar tasks and concluded that both struggled with collocations. Another study by Biskup (1992) compared two groups of learners of different L1s, German and Polish. The learners were administered a translation task which tested their knowledge of restricted and unrestricted collocations. Furthermore, the collocations were divided between congruent (with a direct translation in the L1) and incongruent (no direct translation in the L1). The comparison between the two sets of learners found that German learners were more risk-takers and produced a wider range of collocations, whereas the Polish learners were more conservative and produced a higher number of restricted collocations.

Despite the interesting results obtained by studies based on elicited data, as mentioned in Section 2.2, this type of data also has some drawbacks: namely the degree of unnaturalness and the lack of generalisable results. As regards the latter, the researchers themselves are aware that small sets of collocations are not fair predictors of more generalised behaviours, as Bahns and Eldaw (1993, p. 108) indicated with reference to their 15-collocation sample. On the other hand, the issue of artificiality, which has already been thoroughly discussed in Section 2.2, "may lead learners to produce language which differs widely from the type of language they would use naturally" (Granger, 1998a, p. 5). In other words, "what we think we would say in a given situation is not necessarily the same as what we would actually say" (Gass & Selinker, 2008, p. 68). For example, if in a blank-filling task the learners are offered any kind of input, such as multiple choices, or other indications, this would influence their knowledge: they may be able to find the correct answer, but this would not indicate that they would be able to produce it in natural occurring circumstances. The difference between what may be known or considered acceptable and what is produced/used is highlighted both by Granger (1998b) and Gyllstad (2005). In her study, Granger noticed that the learners tended to match more adjectives with the adverb *highly* than they would actually produce in writing. On the other hand, Gyllstad showed that despite the native-like performance of advanced Swedish learners of English L2 in receptive tasks, their productive knowledge performance was less advanced.

In order to avoid the above-mentioned issues of elicited data, a growing number of scholars have started to conduct research on learner corpora. The main advantage of using learner

corpora for the study of collocations lies in the large quantity of data made available by such corpora and the possibility to observe naturally-occurring instances of phraseological units. Furthermore, learner corpora allow for the (semi-)automatic extraction and analysis of phraseological items. The majority of work on learners' phraseological competence has been done on written texts (Ädel & Erman, 2012; Durrant & Schmitt, 2009; Fan, 2009; Howarth, 1996; 1998a; Kaszubski, 2000; Laufer & Waldman, 2011; Li & Schmitt, 2009; Lorenz, 1999; Nesselhauf, 2005; Siyanova & Schmitt, 2008; among others), but an increasing body of research is also being carried out on spoken language (Crossley & Salsbury, 2011; De Cock, 1998; 2000; Ferraresi, 2019; Götz & Schilk, 2011; Marin Cervantes & Gablasova, 2017; Peng, 2016; Xu, 2018).

As regards LCR, the majority of studies have used synchronic corpora, that is, corpora collected at a specific point in time, such as ICLE. Gilquin (2007) investigated advanced French-speaking learners' knowledge of *make*-collocations in ICLE and combined three different approaches: the analysis of the learners' collocations, the comparison with native speakers' collocations, and elicitation together with acceptability tasks. The findings evidenced that learners do not produce many erroneous combinations, but tend to underuse *make*-collocations and produce a limited range of them compared to native speakers. Furthermore, when asked to rate the acceptability of the collocations, the learners displayed their collocational struggles (they were not able to fill the sentences with the appropriate verb, nor judge the acceptability of a given collocation and correct if wrong). This study managed to overcome one limitation of learner corpus data in collocational research, mainly that of only allowing the researchers to observe performance and not competence. However, Gilquin (2007), by combining learner corpus data with elicited data (blank-filling task and acceptability judgement), managed to obtain insight into the learners' competence. Other synchronic studies on collocations were conducted by Nesselhauf (2005), Durrant and Schmitt (2009), Bestgen and Granger (2014), and Wang (2016), among others. The higher number of studies on synchronic corpora is due to the large availability of synchronic learner corpora and the lack of longitudinal ones combined with the high degree of labour required for their compilation. However, as this gap is slowly being filled, more research is now adopting a longitudinal perspective, such as Li and Schmitt (2009), Siyanova-Chanturia (2015), and Yoon (2016).

Another feature of collocational LCR studies is that the L2 is mostly English, although this is changing and more studies are appearing with a focus on, for instance, French L2 (Forsberg, 2010), German L2 (Krummes & Ensslin, 2015), or Italian (Forti, 2019; Siyanova-Chanturia, 2015). The preferred levels of proficiency are usually intermediate/advanced. The choice of proficiency is often dictated by practical reasons, namely that it is easier

to collect highly natural texts with no elicitation from intermediate and advance language users. Low proficiency learners require a higher level of elicitation[7] and as collocations are seen as a component of fluency it is reasonable to expect that higher proficiency learners may be able to produce them in higher numbers. As mentioned above, research has investigated the written collocational production of learners more widely than spoken production due again to the larger availability of written learner corpora. The most widely used learner corpus is ICLE, although other corpora have been used, such as the *Uppsala Student English Corpus* (USE), or the *Corpus of Chinese Learner English* (CCLE); more work is now being developed on spoken learner corpora, such as LINDSEI and TLC.

Lastly, the majority of learner corpus studies have focused on a selected set of collocations, mainly verb + noun collocations (Laufer & Waldman, 2011; Nesselhauf, 2015; Wang, 2016) and adjective + noun (Durrant & Schmitt, 2009; Granger, 1998b; Siyanova & Schmitt, 2008). The rationale behind the choice of restricted focus may lie in the notion that collocations, as lexical combinations, frequently involve these parts of speech. Furthermore, mastery of frequent verbs and their collocations still represents a stumbling block for learners who aim to achieve native-like proficiency (Altenberg & Granger, 2001; Howarth, 1996; Nesselhauf, 2004). From the point of view of learner corpora, these combinations are also easy to extract from a POS-tagged corpus.

3.5.1 LCR findings on collocations

The vast pool of studies on collocations focus on different aspects, approaches, task types, and collocation properties, thus making it difficult to compare the results (Paquot & Granger, 2012, p. 131). However, some general trends can be identified, such as the learners' struggles (even at advanced levels) when it comes to collocations and the faster development of the receptive knowledge compared to the productive one (cf. Biskup, 1990; Gyllstad, 2005). Some of the issues encountered by learners lie in the frequency of collocations, namely the learners tend to either overuse or underuse specific sets of them; while other studies have demonstrated the influence of the L1 on collocational production and knowledge. In this Section, some relevant research findings will be presented.

[7] Learners with lower proficiency levels may not be able to produce free writing or engage in informal interviews in order to collect natural language data (the primary requirement for the compilation of a learner corpus). As such, collocational studies focusing on lower proficiency learners may need to adopt alternative collection methods to natural language, such as experimental (clinical or elicited) or introspective data (see Section 2.2).

3.5.1.1 Overuse/underuse

In LCR, collocations have been mainly investigated through the analysis of their frequency, which has highlighted some discrepancies among learners: some samples of learners overuse certain sets of collocations, while others underuse other sets of collocations. In this regard, Wang (2016) analysed a restricted set of six frequent delexical verbs and their noun-collocates and found that learners overused the combinations (tokens), but the overuse was due to the frequent repetition of the same collocation types (cf. Kaszubski, 2000). Another study by Laufer and Waldman (2011) investigated a broader range of verb + noun collocations. The authors showed that learners (even advanced ones) produced a significantly lower number of collocation types compared to native speakers. However, a study by Vincze et al. (2016) employed a phraseological approach to the analysis of all collocations (regardless of their syntactic pattern) produced by learners of Spanish L2. The authors found no statistically significant difference between the number of collocations extracted from the native speakers' and the non-native speakers' corpora. Furthermore, the researchers divided the collocations according to the syntactic pattern and found that verb + noun and noun + modifier were the two most frequent types of collocations, with statistically significant differences compared to native speakers. This means that, overall the learners tend to produce the same amount of collocations, but it is possible that they may either overuse or underuse specific sets of them. Interestingly, the study by Farghal and Obiedat (1995) showed that the underuse of collocations may be explained by lexical simplification strategies, such as synonymy, paraphrase, or avoidance. This means that the learners, when faced with the use of a collocation (either in natural occurring language or via an elicited task) may rely on some lexical simplification strategies, such as that of synonymy, whereby they may produce a collocation whose collocate or node is a synonym of the correct lexical item. For more on simplification strategies, see Farghal and Obiedat (1995).

As regards another type of collocation, namely adverb + adjective, a study by Granger (1998b) analysed both these collocations and formulae and observed that learners use significantly fewer *–ly* adverbs than native speakers in terms of types and tokens. This may result in a lack of diversification of linguistic items which contributes to the lack of native-like proficiency and fluency. Similarly, Lorenz (1999) found that learners underuse more restricted adverb + adjective collocations while they overuse less restricted ones. This also contributes to the *foreignsoundingness* of learners, even those at intermediate and advanced levels of proficiency. Granger's (1998b) study introduced the idea that there is a link between the overused collocations and the learners' L1 (cf. Kaszubski, 2000). This will be further discussed in the next Section (3.5.1.3).

Another perspective adopted for the investigation of collocations from the point of view of overuse/underuse has been that of association measures. Association measures have been used in a number of studies to delimit and classify collocations on the basis of their frequency, exclusivity (the restricted co-occurrence with specific lexical items), and dispersion. The general trend in these studies points towards the idea that learners are more sensitive to more frequent collocations (and thus will tend to produce them in similar or higher number compared to native speakers), rather than more strongly associated ones as identified by their MI scores (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Ellis et al., 2015). An in-depth study by Durrant (2014) with a test on phraseological competence of 19 collocations showed that frequency is moderately correlated to learner knowledge, whereas no correlation was found between learner knowledge of collocations and MI scores.

The explanation for these results may be found in the MI score itself, which, as it will be described in Section 3.5.2, tends to highlight infrequent and rare combinations. Frequency has been shown to positively relate to learners' processing and knowledge of collocations, thus infrequency of collocations could explain the lack of correlation between MI scores-defined collocations and learner knowledge. This is compatible with Li and Schmitt (2009) who suggested that less frequent and more strongly associated collocations identified by MI scores are "the type of item which is likely to be highly salient for native speakers" (p. 96), but perhaps not for non-native speakers, who may thus display underuse of these collocations.

In conclusion, learners use a more restricted set of collocations compared to native speakers (Fan, 2009) and this might be due to the fact that the learners overuse frequent collocations because they feel more confident with them: the so-called "collocational teddy bears" (Nesselhauf, 2005, p. 69). As regards the underuse of certain collocations, Farghal and Obiedat (1995) found that the learners tend to avoid or choose a different expression instead of the target collocation. Alternatively, the underuse could be attributable to the lower frequency and saliency of the given collocations or their incongruence with the L1. This latter factor will be analysed in Section 3.5.1.3.

3.5.1.2 Misuse

Some of these studies not only investigated the frequency of collocations, but they also took the erroneous use of collocations (misuse) into consideration. The present work also qualitatively analyses a small set of infrequent and unattested collocations for evidence of misuse (see Chapter 7). . For example, Laufer and Waldman (2011) demonstrated that one third of the collocations extracted from the learner corpus resulted in erroneous verb + noun

combinations, and Bestgen and Granger (2014) found that there was a negative correlation between the quality of essays and the lack of erroneous use of bigrams (e.g., university studies). Similarly, Nesselhauf (2005) found that approximately one third of the collocations she extracted from the learners' corpus were either unacceptable or questionable[8]. This could be an indication that learners had some difficulties in selecting the correct collocate verbs in the verb + noun collocations. Nesselhauf also noticed that the more restricted collocations were less prone to errors compared to medium-restricted collocations which allowed a wider range of collocate nouns (cf. Martelli, 2006; Nesselhauf, 2003). She suggested that this was due to the fact that learners were probably acquiring and storing more restricted collocations as a whole, whereas the less restricted ones were used more creatively, thus inducing a higher degree of errors (Nesselhauf, 2003, p. 233). Corroboration regarding the persistent errors at advanced levels of proficiency was also found outside LCR by Hoffman and Lehman (2000) who discovered, through a blank-filling task, that learners only achieved 34% accuracy in providing the correct adjective + noun and noun + noun collocations compared to native speakers. The misuse of collocations may be correlated to the learners' proficiency, as demonstrated by Siyanova and Schmitt (2008): a large percentage (75.3%) of the adjective + noun collocations produced by Russian advanced learners of English were acceptable and 45% of these were also frequent and strongly associated. The remaining collocations were not necessarily erroneous, but they were simply not taken into consideration because they did not occur in the BNC. Proficiency is a correlated factor of collocational knowledge as shown by Gitsaki (1999): the number of collocations increased as the learners' proficiency progressed and they became more proficient (cf. Zhang & Chen, 2006). Gitsaki's conducted a cross-sectional study and tested both receptive and productive knowledge among three groups of EFL learners with three different levels of proficiency. The most proficient learners performed better at adjective + noun collocations in translation tasks and acceptability judgements.

3.5.1.3 L1 influence

A variable which is strictly related to proficiency and has been widely investigated in research on collocations is L1 influence. Indeed, Granger (1998b) noticed that the learners' most frequent collocations had a translation equivalent in the L1, they were thus congruent

[8]Nesselhauf established her baseline of acceptable collocations according to a three-fold scheme: the collocation occurs in one of four dictionaries (the Oxford Advanced Learner's Dictionary, the Collins COBUILD English Dictionary, the BBI Dictionary of English Word Combinations, the Oxford Dictionary of English Idioms); the collocation occurs in identical form in at least five texts of the written component of BNC; the collocation which does not occur in the dictionaries or the BNC is submitted to native speakers for acceptability judgments.

collocations, while collocations which had no direct translation in the L1, non-congruent, were much fewer. She concluded that the learners tend to "'cling on' to certain fixed phrases and expressions which they feel confident in using" (Granger, 1998b, p. 156). This notion that learners may tend to use the same fixed phrases because they feel confident with them had also been recognised by Dechert in 1983 as "islands of reliability" (p. 184), by Hasselgren (1994) as "lexical teddy bears" (p. 237), and later updated to "collocational teddy bears" by Nesselhauf (2005, p. 69).

The influence of the L1 on the production of L2 collocations has been reported in multiple studies (Bahns, 1993; Granger, 1998b; Nesselhauf, 2005; Wang, 2016), with at least two perspectives on its two-fold role: an effect on congruent and non-congruent collocations and an explanation of learner errors. The issue of congruency has also been studied outside LCR by Wolter and Gyllstad (2013) who found that frequent congruent and incongruent collocations are processed faster than more infrequent congruent and incongruent ones by advanced learners. This study corroborates the idea that learners, at least advanced ones, are sensitive to the frequency of linguistic items in the L2. Frequency does not play such an important role for lower-proficiency learners, as shown instead by Siyanova-Chanturia et al. (2011), who found that their learners did not perform significantly faster on the more frequent binomials. Furthermore, Wolter and Gyllstad (2013) argued that the L1 also had a facilitating influence on the response times of learners to congruent collocations (as they performed significatly worse on incongruent collocations).

As regards congruency and non-congruency of collocations, it has been hypothesised that L1 congruency may help the learners' collocations, whereas non-congruency may hinder the production of L2 collocations (Bahns, 1993; Wolter & Gyllstad, 2011; Yamashita & Jiang, 2010). In particular as far as teaching practices are concerned, Bahns (1993), in her study on the feasibility of teaching L2 collocations, concluded that congruent collocations need not to be taught as the direct equivalent translation in the L1 and excluding this portion of formulaic language could simplify the learning process for collocations (which are present in tens of thousands in a given L2). Wolter and Gyllstad (2011) conducted research on collocation priming with a group of Swedish learners of English L2 and a control group of native speakers: the groups performed a primed lexical decision task on congruent, non-congruent, and unrelated items. The learners responded faster to congruent collocations than to non-congruent ones, although the authors warned that slower response rates to non-congruent collocations may be due to the lack of priming for this set of collocations. Yamashita and Jiang (2010) administered an online phrase-acceptability judgment task to a control group of native speakers and to two groups of Japanese learners of English ESL and EFL. The groups were tested for congruent and non-congruent collocations. Both the learner groups

made errors with non-congruent collocations, but the EFL learners had slower response rates to non-congruent than congruent collocations. Thus, the study concluded that the L1 has a greater influence on L2 collocation acquisition in the EFL context rather than the ESL and that L1 congruency combined with L2 exposure maximise the positive effect on the acquisition. However, these results were not corroborated by a subsequent study by Wolter and Yamashita (2015). The researchers tested non-native speakers on a lexical decision task and the results did not return significant signs of L1 activation. Although these studies adopted a psycholinguistic approach and only looked at the receptive knowledge of L2 collocations, similar findings were provided by Nesselhauf (2003; 2005). In both works, the author found similar percentages of errors for congruent and non-congruent collocations (11% and 42% in 2003 and 17% and 42% in 2005), thus suggesting that non-congruent collocations pose problems for learners, but so do congruent collocations. It is at the intersection of these results that a combined approach to the study of collocations is hoped for in order to collect introspective data on the learners' choices and cognitive processes underlying the production of congruent and non-congruent collocations.

As regards the second layer of L1 influence, namely L1 as explanation of learner errors, Kaszbuski (2000) and Granger (1998b) had already proposed that L1 influence was responsible for the overuse and/or overuse of certain sets of L2 collocations to the detriment of others. Other studies also found traces of L1 in erroneous collocations (Bahns, 1993; Biskup, 1992; Farghal & Obiedat, 1995; Martelli 2006; Nesselhauf, 2003; Wang, 2016).

3.5.1.4 Collocation learning lag

One of the many facets of collocational competence, which has already been mentioned in the previous sections, is that receptive knowledge develops faster than productive knowledge (Biskup, 1990; Gyllstad, 2005; Jaén, 2009; Koya, 2005). However, many other factors have been investigated as regards collocational competence. For example, knowledge of collocations has been shown to not develop concurrently with other competences (Biskup, 1992; Farghal & Obiedat, 1995; Laufer & Waldman, 2011; Schmitt & Carter, 2004; Wray, 2002). Schmitt and Carter (2004) pointed out that L2 formulaic language is a step behind the other learners' competences, while Wray (2002) concluded that "by the time the learner has achieved a reasonable command of the L2 lexicon and grammar, the formulaic sequences appear to lag behind" (p. 182). Indeed, the lexicon and the knowledge of collocations has been investigated by Bahns and Eldaw (1993) and it has been shown that, in translation tasks, the learners produce a higher number of collocation errors than lexical errors. This has led the authors to state that "knowledge of general vocabulary far outstrips their knowledge

of collocations" (Bahns & Eldaw, 1993: 108) (cf. Barfield, 2007).

Perhaps the phraseological competence lag of learners (even advanced ones) may be attributable to a fuzzy knowledge or awareness of collocational relationships. This has been demonstrated in studies that tested the learners' intuition about frequent collocations (Channel, 1981; Granger, 1998b; Siyanova & Schmitt, 2008). If the learners are not aware of the collocational relationship, their production is "subject to whatever IL rules the learner is operating under" (Yorio, 1989, p. 62). Lastly, another aspect of collocational knowledge is that it tends to develop in the late stages of acquisition. Laufer and Waldman's (2011) pseudo-longitudinal study found that a number of restricted collocations only increased at the advanced levels. This was also confirmed by Bestgen and Granger (2014) who employed the same approach of Durrant and Schmitt (2009) to the study of collocations. The authors compared two groups of learners, intermediate and advanced ones. They found that intermediate learners tend to use a larger number of high-frequency collocations and few infrequent and strongly associated ones, as opposed to advanced learners. The choice of collocations is extremely important in any study as they are not all good predictors of collocational knowledge, as indicated by Paquot (2017): adjective + noun collocations are weaker predictors of proficiency and collocational knowledge compared to verb + object collocations which are the best discriminators.

3.5.2 Association Measures for collocations

As mentioned in Section 3.4.1.2, frequency is not the sole predictor of collocations. There are additional measures which can inform on the strength of the attraction of the collocational items. In truth, there are many and diverse association measures, all of which may lead to different, if not contrasting results, and the choice of measure needs to be made carefully as it will impact the whole research work. A few of the most common association measures used in corpus linguistics for the analysis of collocations are *Mutual Information* (MI), *t*-score, Log Dice, Log-Likelihood, Delta *P.* Both MI and *t*-score will be analysed in the following Section as these are the association measures chosen for the present work.

As regards, Log Dice, Log-Likelihood and Delta *P,* these are statistical association measures based on contingency tables and present a few benefits over association measures such as MI and *t*-score. In particular, the Log- Dice is a statistical association measure that has not been explored much in LCR, but it represents an improvement over the MI as scores across different corpora can be compared and it provides a reference point for maximum values (for more on the Log Dice, see Gablasova et al., 2017). Log-Likelihood (Dunning, 1993) is a bidirectional measure which tests the significance of the hypothesis, thus how

strongly the collocations are attracted. It is the ideal measure for testing medium-frequency combinations which are not either particularly exclusive or non-exclusive. Lastly, Delta *P* was proposed by Gries (2013) as a unidirectional or asymmetric measure, which was first discussed by Ellis (2006) and then used in Ellis & Ferreira-Junior (2009). Even if a standard association measure scores is high, this does not tell the researchers any information regarding the direction of the attraction. Delta *P* takes into consideration the fact that perhaps one item is more attracted to the other and viceversa, and provides two scores which inform the researchers about the direction of the attraction.

3.5.2.1 T-score and Mutual Information

Two of the most common association measures (Evert, 2009) for the study of collocations are MI and *t*-score (Durrant, 2008; Durrant & Schmitt, 2009; Evert, 2009; Gries, 2013; Hunston, 2002; Spina, 2001; Siyanova & Schmitt, 2008). They interpret the observed co-occurrence frequency (i.e., how many times the combination occurs in a corpus), by comparing it with the expected frequency (i.e., how many times the combination is expected to occur based on the individual frequencies of its components and their chance to occur next to each other), and they thus calculate an association score which is a quantitative measure for the attraction between two words (Evert, 2009, p. 1225).

Mutual Information is an association measure first introduced by Church and Hanks (1990, p. 23) and it is based on a logarithmic scale and the ratio between the observed frequency of a collocation and its expected frequency[9]. MI exemplifies two general conventions for association scores that all association measures should adhere to: 1) higher scores suggest stronger attraction between words, i.e., a greater degree of collocativity; 2) an association measure should distinguish between positive association (the observed frequency is greater than the expected frequency) and negative association (the observed frequency is lower than the expected frequency), assigning positive and negative scores, respectively. MI satisfies both conventions: the more the observed frequency exceeds the expected frequency, the greater the association score will be; whereas if the observed frequency is equal to the expected frequency the result of the MI value will be 0. In practical applications, MI was found to have a tendency to inflate scores of low-frequency word pairs with an expected frequency equal or lower than 1, especially for data from large corpora. Thus, even a single co-occurrence of a two-word combination might result in a fairly inflated association score. In order to counterbalance this low-frequency bias of MI, in the present study an observed frequency threshold of 5 was set as the minimum requirement for the calculation of the

[9]For the formula, see Evert (2009, p. 1225).

score. As with all association measures, MI should be chosen according to the research question and should be employed especially when investigating exclusive and infrequent combinations.

The other association measure that has also been widely used in the literature combination is the *t*-score. It was first introduced as a measure in corpus linguistics by Church et al. (1991) and since then it has been subject to multiple critics (cf. Evert, 2005). The *t*-score measures "the strength of co-occurrence" (Wolter & Gyllstad, 2011, p. 436) and it is a derivation of the *t*-test.

The *t*-score is computed by subtracting the expected random frequency from the observed frequency and dividing it by the square root of the raw frequency[10]. This measure is not based on a standardised scale so it is not possible to compare scores of different corpora (Hunston, 2002) and an arbitrary cut-off point of *t*-score ≥ 2 has been suggested as a valid threshold for the classification of collocations (Hunston, 2002). The *t*-score highlights frequent combinations and it has been shown to strongly correlate to the raw frequency count (Gablasova et al., 2017). Indeed, Durrant and Schmitt (2009, p. 167) point out that the *t*-score rankings "are very similar to rankings based on raw frequency" (see also Hunston, 2002; Siyanova & Schmitt, 2008).

As regards the combined use of the *t*-score and MI to identify and extract collocations (on the opposite ends of a spectrum, i.e., frequent vs infrequent) from a corpus, Gablasova et al. (2017) warn the readers that despite much literature suggesting the *t*-score and MI as "co-extensional terms", this is not case as the *t*-score indeed highlights frequent combinations, but not all frequent combinations receive a high *t*-score. However, in Durrant and Schmitt (2009) and Granger and Bestgen (2014), MI and *t*-score have been successfully used concurrently to select and identify collocations by means of combining the two types of information these measures provide (Durrant & Schmitt, 2009). Furthermore, in a comparison among five different association measures (MI, MI3, *t*-score, Log-likelihood, chi-squared), Krenn (2000) has demonstrated that the *t*-score is the most reliable index for the extraction of collocations. The choice of the *t*-score, as with all other association measures, depends on the research question and on what the researcher would like to focus. For the present work, the *t*-score was employed in association with MI to extract collocations based on the similar approach of Durrant and Schmitt (2009).

[10]The formula is: $\frac{O-E}{\sqrt{O}}$. For more on the formula, see Evert (2009, p. 1225).

3.6 Summary

This chapter has reviewed the growing body of research that has been conducted in the past several decades on phraseology and, in particular, collocations. The overview of phraseology has highlighted the wide range of phraseological units that have been studied in the last few decades and the main concerns of the research community: what counts as a phraseological unit and how are they identified? Multiple approaches have been suggested, each with its own insights and limitations. Formulaic language has been shown to be pervasive to language, thus encouraging further research into this field. Moreover, although there are some contrasting results, it seems that non-native speakers tend to process formulaic sequences faster than non-formulaic language. Perhaps this is due to some variables at play, such as frequency of the formulaic sequences, or the congruency with the L1.

As regards collocations, the review has presented the two main approaches to the study of collocations, namely the phraseological and frequency-based approach. For the frequency-based approach, a brief summary of some of the most common association measures was provided in order to understand how collocations can be identified from the frequency-based perspective. Overall, the Sections on collocations in LCR have highlighted a wide range of data types, approaches, data collection methods, and foci. Previous research has covered many types of learners, including learners of different L1s or different proficiency levels. Different types of collocations have also been investigated, i.e., verb + noun collocations, adjective + noun, adverb + adjective, etc. Due to the vast pool of research studies on L2 collocations, direct comparisons of results are difficult, but it is possible to outline a general picture for the L2 learning of collocations, i.e., collocations pose special difficulty for L2 learners, as evidenced by deficiencies of collocation overuse, underuse and misuse even for learners at advanced levels. Furthermore, collocational competence tends to lag behind other linguistic competences such as vocabulary. All these factors point towards the need for more research into collocations.

Chapter 4

Methodology and research questions

THIS chapter introduces the research questions addressed in this thesis and describes the methods adopted to answer them. The first Section highlights the main points in the literature which led to the formulation of the three research questions; the next Section describes the rationale behind the compilation of the new corpus of Italian advanced learners of English (the ISLC), its design and collection criteria, tagging and transcription scheme, while the last subsection focuses on providing data regarding the corpus (e.g., size, balance, learners' metadata). The third Section presents the extraction of the collocations from the corpus; in particular, it outlines the approach to the identification of collocations, the span, syntactic pattern, and the chosen association measures. Section 4.4 discusses the methods used in both the quantitative and qualitative analyses. Specifically, this Section illustrates the procedures for the quantitative analysis of the extracted collocations and the scheme followed for the qualitative analysis.

4.1 Research questions

Section 3.3 argued that collocations and phraseology are pervasive to language (Altenberg, 1998; Becker, 1975; Bestgen & Granger, 2014; Biber et al., 1999; Bolinger, 1976; Cowie, 1991; 1998; Granger & Meunier, 2008; Howarth, 1998a; Kjellmer, 1994; Meunier & Granger, 2008; Nattinger & De Carrico, 1992; Pawley & Syder, 1983; Sinclair, 1991; Stubbs, 2007); play a central role in the acquisition of first and second languages and adult language production (Cowie, 1998; Pawley & Syder, 1983; Peters, 1983); and are an essential component for the development of fluency (Ellis, 2002; 2003; Ellis et al., 2015; Howarth, 1998a; Lewis, 1997; 2000; Meunier & Granger, 2008; Nattinger & De Carrico, 1992; Wray, 2002). However, despite the invaluable contribution of spoken learner corpora to the investigations of SLA (Myles, 2015), most research has been conducted on written

corpora (Ackerley, 2013; Bestgen & Granger, 2018; Bulon, 2020; Chen, 2019; De Haan & van der Haagen, 2013; Fan, 2009; Forti, 2019; Gonzalez & Alonso Ramos, 2013; Paquot, 2018; Schneider & Gilquin, 2016; Siyanova-Chanturia, 2015; Vincze et al., 2016; Wanner et al., 2013; among others). Furthermore and following a tradition in applied linguistics, the majority of studies on collocations and phraseology in LCR have focused on verb + noun collocations, in particular verb + noun collocations with high frequency delexical verbs such as *take* and *make* (Paquot & Granger, 2012). Other formulaic sequences that have proved to be error-prone and have thus attracted attention are phrasal verbs (De Cock, 2005) and speech formulae (Aijmer, 2009).

This thesis sets out to investigate a particular set of collocations in a spoken learner corpus, namely, adverb + adjective collocations. Indeed, although verb + noun collocations are very frequent structures (and thus are easily retrievable from a corpus), little work has been done on intensified adjective collocations (cf. Altenberg, 1991; Granger, 1998b; Granger & Bestgen, 2014; Lorenz, 1999) although intensification is an intricate part of foreign language learning (Lorenz, 1999, p. 26) and should need further attention. In the pioneering work by Lorenz (1999), an anecdote is reported regarding an advanced language learner asking whether she could substitute the adverb *extremely* with *immensely*. This, Lorenz argued, shows that intensification can be difficult even for advanced learners and learners themselves tend to be aware of the significance of intensification. Granger (1998b) also posited that intensifying adverbs:

> [...] constitute a particularly rich category of collocation, involving as they do a complex interplay of semantic, lexical and stylistic restrictions and covering the whole collocational spectrum, ranging from restricted collocability – as in bitterly cold – to wide collocability – as in completely different/new/free/etc. (p. 2)

This means that intensifiers can put great stress on the learners' collocational skills and, since adjectives are the most frequent focus of intensification (Lorenz, 1999), the adverb + adjective collocations could be expected to provide interesting insights into the language learning patterns of advanced learners of English.

In order to explore this set of collocations, this thesis addresses the following interrelated research questions and their sub-questions:

1. How do advanced Italian learners of English and native speakers of English compare in their production of different adverb + adjective collocations in spoken language?
    a) Are there any differences between the two groups in terms of quantity of combinations produced?

b) Are there any differences between the two groups in terms of quantity of infrequent/ unattested and grey area collocations produced?

c) Are there any differences between the two groups in terms of strength of association of the collocations produced?

2. What is the difference between the syntactic patterns and lexical meaning of the adverb + adjective collocations (see note 2 of Chapter 1) produced in spoken language by advanced Italian learners of English and native speakers?

3. Does L1 congruency have a transfer effect on the production of infrequent/unattested collocations generated in spoken language advanced by Italian learners of English?

The term *infrequent* or *unattested* collocation is used to refer those combination which, according to the study design that will be explained in the following Sections, have not been classified as collocations because they are either not present or occur fewer than five times in the BNC.

4.2 The Italian Spoken Learner Corpus

4.2.1 Rationale

The compilation of a corpus, either written or spoken, is a daunting task for any researcher, but in recent years many have succeeded in the feat and both the corpus linguistic and the LCR community can benefit from a large number of already available corpora (see Sections 2.3.1.1 and 2.3.1.4). However, in order to address the research questions of the present work, the compilation of a new learner corpus which could be comparable to a native-speaker reference corpus became necessary. Indeed, the research questions are aimed at the investigation of collocations in spoken language produced by advanced Italian learners of English within a CIA perspective. Two elements of this investigation posited the need for a new corpus: the poor availability of spoken learner corpora and the advanced proficiency level of the learners. At the moment of writing this thesis, the only widely available spoken learner corpus of Italian learners of English L2 is LINDSEI (Gilquin et al., 2010), which was also compiled according to strict design criteria. However, there is one drawback with the use of the Italian component of LINDSEI, mainly the proficiency level. At the time of its compilation, the team of compilers opted to rely on year of instruction as a proficiency level indicator. Further research has then indicated that reliance on external variables for

proficiency level is not an optimal choice and more text-centred or learner-centred variables should instead be preferred (see Section 2.3.1.3).

4.2.2 Corpus criteria

The compilation of the corpus for this work followed a precise objective, namely that of collecting data which was representative of the advanced IL spoken by Northern-Italian university students enrolled in a degree course of foreign languages and learning English as L2. Due to time constraints, resources needed for compiling the corpus, and the issues which arise when transcribing spoken language, the compiled corpus is a small learner corpus (66,898 words). However, the advantages of small corpora rich in metadata are a more qualitative and in-depth analysis of individual variation. As regards size, although it is true that size matters, Sinclair (1991, p. 18) suggested that a "corpus should be as large as possible, and should keep on growing". Hence, this corpus will continue to grow in the following years, as more recordings will be added and the degree of proficiency will expand to other CEFR levels.

In order to analyse adverb + adjective collocations, the corpus needed to be i) comparable, and ii) authentic. The criterion of comparability was essential since the analysis of the adverb + adjective collocations in this thesis adopts the CIA approach (see Section 2.4.1). Thus, the employment of the CIA approach required a corpus which would be as comparable as possible with the reference one(s). Comparable corpora are useful for contrastive and translation studies. As Aijmer et al. (1996) pointed out, comparable corpora can offer insight into the languages under investigation which a normal non-comparable corpus in a non-contrastive study would not yield. Furthermore, they allow the researcher to discover new language-specific or universal features that would otherwise not be easily accessible. Lastly, by comparing corpora, interesting teaching applications could also emerge. Although the term "comparable corpora" has a much stricter definition in the literature (cf. Hunston, 2002; McEnery & Xiao, 2007), in this thesis the term is used in the broadest sense, meaning a monolingual corpus compiled with a sampling frame similar to the reference corpus. The natural choice of reference corpus for CIA was LOCNEC, which was designed as the counterpart of LINDSEI (Gilquin et al., 2010). Therefore, the Italian Spoken Language Corpus was to be as similar as possible to LINDSEI and its "sister" reference corpus LOCNEC.

As regards the criterion of authenticity, it is widely known in the literature that learner corpora may not follow Sinclair's (1996) principle of authenticity, since learner corpora are not "gathered from the genuine communications of people going about their normal

business". Any intervention of the researcher qualifies the corpus as a "special corpus" and any other data obtained in "experimental conditions, or in artificial circumstances of various kinds" does not constitute a corpus (Sinclair, 1996). However, as already discussed in the literature review (see Section 2.3.1.5), although a certain degree of artificiality is inherent to learner language, ESL or EFL contexts are the appropriate environments for naturally occurring learner language (Tono, 2016, p. 124).

4.2.3 Corpus design and data collection

The decision to compile an ad-hoc corpus was made after careful consideration of the two elements that the research questions required from the research, namely advanced learners of English and spoken language. The new corpus needed to be controlled in terms of proficiency, transcription, task distinction, interviewer's output. The following paragraphs describe ISLC's design criteria, highlight differences and similarities with LINDSEI (see Appendix for LINDSEI's corpus design criteria), and explain the reasons why specific choices were taken as regards data collection. The study was subject to the approval of the Ethics Board of Università Cattolica del Sacro Cuore and, before the interview, students signed an informed consent in compliance with GDPR 2016/679. The data were collected between January and May 2019.

The data of the Italian learners of English L2 contained in the ISLC corpus were collected from three universities located in Northern Italy (Lombardy) and will be referred to as *University A*, *University B*, *University C*[1]. A minimum of 30 interviews was arbitrarily set for the corpus and, similarly to LINDSEI, the interview duration was established at approximately 15 minutes (five minutes per task). With regard to the interviewee's variables, the same questionnaire used in LINDSEI was filled in by the participants and Table 4.1 displays all the metadata collected.

[1]The university and participants' names were anonymised in compliance with the recent European General Data Protection Regulation (GDPR) 2016/679 and as approved by the researcher's ethics committee. Furthermore, anonymisation was also pursued based on the customary procedure in second and learner language research to often anonymise the participants and any data to which they may be associated in order to avoid any biases (Dörnyei & Taguchi, 2009).

Table 4.1: Learners' metadata collected during the interviews.

Learner-related	Education-related
age	MoI in primary school
gender	MoI in secondary school
country of origin	current studies
native language	current year of study
mother's native language	MoI at university
father's native language	years of English at school
languages spoken at home	years of English at university
	stay abroad experience (if yes, how long)
	other foreign languages

With reference to Table 4.1, *MoI* refers to *medium of instruction* and the stay abroad experience was only recorded if it occurred in an English-speaking country. The data were produced by students attending English courses at the three universities and attending either the undergraduate or postgraduate courses. The only requirements for the selection of volunteers was enrolment in a degree course of Foreign Languages of the aforementioned institutions and an advanced proficiency level ($\geq$ C1). No distinction was made between undergraduate or postgraduate, as opposed to LINDSEI (which includes only undergraduates), because it was thought that this had no influence on learner language since the only potential variable related to the degree level might be proficiency. On the other hand, in terms of proficiency, previous experience (Poli, 2014) and research (Carlsen, 2012; Leclercq et al., 2014; cf. Section 2.3.1.3), demonstrated that basing the proficiency variable on arbitrary parameters is highly unreliable. Individual variation is too unpredictable to allow for the adoption, for instance, of the course year as a predictor of proficiency level (see Section 2.3.1.3). Proficiency is a complex and multidimensional concept, so its assessment needs to be reliable and practical. There are standardised tests that the learners can take prior to the data collection process, or researchers can adopt a more text-centred approach (cf. Section 2.3.1.3). However, due to the high level of time and human resources needed to assess proficiency with this latter approach, in this thesis the assessment of the participants' proficiency level was carried out with a standardised test. Interviews were conducted with 58 participants, but only 34 were able to meet the requirements set out above and were thus included in the corpus.

The Oxford Online Placement Test[2] (OOPT) was used as a standardised assessment test for two main reasons: firstly, this test has already been widely used in University A for assessing the proficiency level of students applying for an Erasmus scheme or other stay-abroad programmes. Secondly, the administration of the test only requires a computer with an Internet connection and about 60 minutes of time, both of which are easy requirements to obtain. In addition, the test automatically marks the results and provides instant scoring. The test offers the following features: it can assess either British or American English since it is possible to select which variety (or a mix of both) to test; the test evaluates the person's understanding of meaning when communicating in English and not only the knowledge of grammar and vocabulary; the scores are aligned to CEFR; it is computer-adaptive thus it is short in length; and it is advertised as a valid and reliable assessment method having been already subject to rigorous test design, pretesting, and piloting. There are two main tasks in the test, the first is called *Use of English*, and the second is a listening task. The OOPT does not only test grammatical knowledge, but also the pragmatic one, for this reason the first part of the test is divided into four tasks: the first measures grammatical forms; the second measures semantic meaning; the third assess grammatical form and meaning; the fourth evaluates the learner's knowledge of the pragmatic meanings conveyed in situated interactions. The listening section is designed to present the learners with different listening passages from which they will have to identify the literal, intended, and implied meanings. There are three types of tasks in this section: the first comprises a series of short dialogues followed by a multiple-choice question; the second presents a longer dialogue followed by a multiple-choice question; and the third one is a monologue followed by a multiple-choice question. The recordings may be listened to twice. The following selections were operated regarding the administration of the test:

- Mix of accents (50% British English, 50% American English)
- Time limit 60 minutes
- Administration post-interview

It is worth commenting on the last element, the administration of the test post-interview, rather than prior. Evaluation is a great source of anxiety in learners, therefore if the participants believed that the interview was part of a test, they may be negatively influenced with higher levels of stress, anxiety, shyness, and lower levels of WTC. By administering the test after the interview, the learners were less biased, more relaxed, and confident that the interview was not going to be evaluated.

[2]https://elt.oup.com/feature/global/oxford-online-placement/?cc=it&selLanguage=it

4.2.3.1 Procedure

After having established the basic requirements, with the collaboration of the language teaching staff, 58 learners were identified and invited to participate in the project. Each learner was met in a private room for approximately two hours to conduct the interview and the proficiency test. All the interviews were conducted by the same interviewer (i.e., the researcher) who followed a transcript during the first part of the interview (see Appendix) in order to increase consistency and comparability of responses from the learners. Furthermore, the order of the interview tasks was followed rigidly, and, before starting the recording, the participants were informed about the general outline of the interview. The instructions and the interview were conducted in English. Only 34 participants met the requirements and were thus included in the study.

As already mentioned, ISLC followed the model chosen by the CECL team for the compilation of LINDSEI, that is, prompted production divided into three tasks. However, in the case of ISLC, the third task was modified into a story-telling task. The rationale behind this choice is two-fold: first, in LINDSEI the picture description task is the least productive in terms of number of words and the one which presents the higher degree of variability in terms of interviewer's prompts/responses. Secondly, story-telling leaves the learners "relatively free to select alternative constructions to express their meanings, either by preference or by avoidance" (Gass & Mackey, 2007, p. 81). As far as the former reason is concerned, as outlined in Gilquin et al. (2010), the average number of words produced in task three in LINDSEI is 126,860, which accounts for a mere 16% of the total interview learner language. In addition, Gráf (2017) pointed out that no guidance was given to the national sub-corpora teams in terms of task three elicitation/prompts, thus the interviewer did not know whether they were allowed to ask questions to keep the learner talking, or a more monological task was to be pursued (p. 30). Gráf also checked a random sample of sub-corpora and found that the interviewers varied in terms of finding solutions to very short picture descriptions. Thus, in the hope of increasing the talking time in this task, and in order to have clearer instructions for the interviewer, a prompted story-telling task was chosen. Not only does a story-telling task leave the learners freer to prefer or avoid certain structures, but it is also a cognitively more demanding task, which may be reflected in lower levels of linguistic accuracy (Foster & Skehan, 1996). For this reason, this more cognitively demanding task was believed to reveal further information regarding the learners' competence as regards adverb + adjective collocations.

Similarly to LINDSEI, the first task consisted in the learner choosing a topic from a selection of three set ones, and talking about it for approximately five minutes. The set

topics are the following:

- An experience you have had which has taught you an important lesson. You should describe the experience and say what you have learned from it.
- A country you have visited which has impressed you. Describe your visit and say why you found the country particularly impressive.
- A film/play you have seen which you thought was particularly good/bad. Describe the film/play and say why you thought it was good/bad.

After the first task, the interviewer went on to ask follow-up questions about the chosen topic, and then discussing other subjects, such as hobbies, travelling experiences, university life. In order to elicit more adverb + adjective collocations, the interviewer asked questions which required descriptions, such as: *how was the weather? How was the food? What was your favourite part of the trip? What was your least favourite part of the lessons?*. Lastly, the prompted story-telling was designed so that the learner would be instructed to read the prompt

> *A character is walking in a forest and suddenly they see a strange-looking box. They take it home and they open it...*

and look at a list of adverbs, which was designed as a circle (Figure 4.2.1). This was done to avoid a more traditional list design, since that may have biased the learners to use only the first adverbs, to the detriment of other adverbs further down the list.

While handing the prompt and the list of adverbs to the interviewee, a clear instruction was given:

> *Here is a prompt, here is a list of adverbs. I would like you to tell me a story starting from the prompt and use as many adverbs as you want.*

As regards the choice of adverbs, these were taken from Biber et al., (1999, pp. 561–565) from tables indicating the most common adverbs groups by semantic domain in British English and American English conversation and academic prose. Adverbs only occurring in the academic prose category were excluded. The adverbs were selected as follows:

- Degree adverbs: *very, really, too, quite, pretty, exactly*
- Stance adverbs: *probably, really, sort of, kind of*
- Amplifiers: *very, so, really, too, real, completely, absolutely, totally, rather*

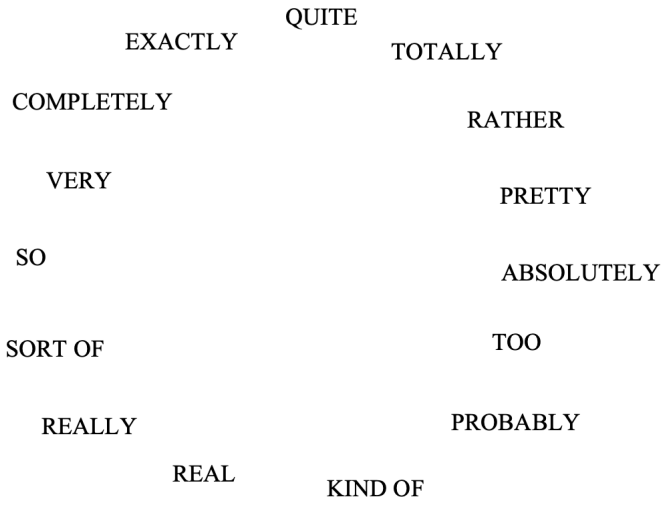

Figure 4.2.1: List of adverbs for prompted story-telling task.

The instruction given to the interviewee was carefully planned so as to give the learner the impression that they had no obligation to use all the adverbs or even some of them. They were free to include them in their narrative or not. If the learner asked further questions about this detail, the interviewer explained that it was up to the interviewee to choose whether to use any of the adverbs. This was done to avoid leading the learners to think that these adverbs were needed in the narrative. This clarification was believed to potentially avoid unnecessary learners' anxiety about inserting all adverbs in the narrative, thus leaving a certain amount of freedom to the learners.

4.2.3.2 The influence of the interviewer on learners' production

When embarking on the demanding task of compiling a corpus – either written or spoken – it is extremely important to establish strict design criteria prior to the collection of data. For the compilation of ISLC, the task was facilitated by the required high level of comparability between it and LINDSEI, hence most criteria were taken from the latter. For example, the informal interview genre was maintained although, as already described, a few adjustments were made to the data elicitation.

One of these aspects was the choice of interviewer, which in this case is believed to have positively influenced the learners' production. This belief is strictly related to the WTC construct, devised by McCroskey and Baer in 1985. The model stems from the idea that

"the more a person is willing to talk and to be nonverbally expressive, the more likely that person is to develop positive interpersonal relationships" (McCroskey & Baer, 1985, p. 96). People are different and some may be more prone to conversation than others, the WTC construct is a measurement scale of this willingness. An individual's WTC is not only personal, but it can be influenced by situational variables. For example, McCroskey and Baer (1985) demonstrated that the higher number of participants in a conversation and the more distant the relationship with the interlocutor, the less willing individuals are to speak. Over the years, other research tapped into the WTC model and its variables, such as shyness, motivation, language competence, learning context, and others (cf. Cao & Philp, 2006; de Saint Léger & Storch, 2009; Kang, 2005; Peng & Woodrow, 2010; Öz et al., 2015; Yashima, 2002; Yashima et al., 2004).

As regards the communicative context that was created for the collection of ISLC data, due to time constraints and limited economic resources, the data collector for the present work – hence, the interviewer – and the transcriber coincide with the author of this thesis. On the one hand, reducing the interviewer to one person consequently eliminates the interviewer variable (which is present instead in LINDSEI), but it also generates many benefits, chief among them, the relationship between interviewer and learners. The majority of the learners were already acquainted with the researcher (familiarity with the speaker may have a positive influence on WTC, see Pawlak & Mystkowska-Wiertelak, 2015) and, since the researcher was never perceived as a teacher or evaluator by the learners, and being closer to the learners' average age than the other teaching staff, it is plausible that the learners felt comfortable with the interviewer.

Moreover, since research has demonstrated that teacher immediacy behaviours reduce learners' shyness (Fallah, 2014, p. 144), the interviewer adopted multiple verbal and nonverbal immediacy behaviours: self-disclosure, humour, praise, smiling, proximity, eye contact, relaxed body position, speaking with learners outside the interview setting. In particular, before the interview would take place, the interviewer took some time to have an informal conversation with the learner (in their L1, Italian), asking about their day, telling them about their research, making jokes, offering chocolate to decrease the level of pressure. During the interview, the interviewer would sit next to the learner, rather than opposite (to decrease the distance), and she would use smiling, gestures, eye contact, nonverbal expressiveness to respond and provide positive feedback to the learner.

Furthermore, since research has shown that if students know what to expect, their anxiety level decreases and their WTC increases (Yashima et al., 2018, p. 126), the interviewer would explain the tasks clearly to the learners before starting the recording. In conclusion, on account of a significant relationship between immediacy behaviours and the learners'

L2WTC (Hsu, 2005; Zarrinabadi, 2014), all the interviewer's behaviours may have positively influenced the learners' WTC.

4.2.4 The transcription and tagging process

Once all the necessary data for the compilation of the corpus had been collected, in order to make the corpus viable and searchable, it was necessary to transcribe the interviews and perform a POS-tagging annotation. This Section describes the transcription scheme adopted for ISLC and the POS-tagging process.

Since there are numerous similarities between ISLC and LINDSEI, the choice of transcription scheme fell on an edited version of the same scheme used for LINDSEI. LINDSEI does not adopt a phonological transcription of the recordings, but is rather a collection of orthographic transcriptions. The transcription process was broken down into two steps: transcription of the data, post-transcription check of the transcribed data. The interviews were transcribed orthographically with little prosodic and phonetic information. The transcription scheme was adapted from Edwards (1992) and her minimalistic transcription standard. The transcription conventions used in LINDSEI are illustrated in the Appendix.

LINDSEI has undoubtedly provided learner corpus researchers with plentiful of data over the years, but a few observations were made regarding its transcription system. Gráf (2017), who was coordinator of the Czech sub-corpus, analysed some of the weaknesses of LINDSEI and, although he did not provide a solution for them, he suggested deciding which features are worth being transcribed in a spoken corpus as it may not be possible to include them all. He noted that it is extremely difficult for transcribers following this transcription scheme to identify unfilled pauses, filled pauses, and overlaps. The LINDSEI project started in 1995, which may explain why some of the design criteria may be considered weak, research on spoken corpora was not as advanced as it is today, and surely the team at CECL had to decide whether it was more important to devise a very detailed transcription scheme (which may not have been easily viable to all sub-corpora teams) or collect a large amount of learner data easily available to all researchers and non-researchers (e.g., teachers). Indeed, pauses generate several issues, as Gráf pointed out (2017, p. 31), for example the scheme indicates that one dot represents a short unfilled pause (less than a second), but no clear instructions are given as to how one can measure the length of the pause. There are no pause detection software available which are easy to use and with a low margin of error, so pause identification is largely subjective and impressionistic (Gráf, 2017, p. 31). The same can be said for filled pauses, which require the transcriber to measure the length and identify the type of nasalisation. Moreover, if long and short vowel sounds are to be transcribed, the

issue arises as to how a longer vowel sound can be identified as opposed to a standard length sound (i.e., *the* vs *the[i:]*). As regards other weaknesses not related to the transcription, proficiency may represent one of the primary ones. Carlsen (2012) has investigated proficiency in computer learner corpora and has concluded that proficiency is a "fuzzy variable". It is difficult to define it as it is multi-layered and multi-faceted; in the case of LINDSEI, the team opted for external parameters (i.e., institutional level).

Lastly, another problematic area identified by Gráf (2017) is the lack of clear instructions as to what the interviewer may say, may ask, may elicit. The set topic and the free discussion tasks may sometimes be difficult to distinguish: some learners are less talkative than others and if the interviewer is to prompt the learner, then the first task may turn into a dialogue, rather than a monologue. All these issues may be easily addressed by researchers compiling their own corpora: choices must be made prior to the compilation and clear instructions must be given to the collaborators. However, the very first choice that the researcher must make concerns the research question(s): the objective of the research inevitably influences the corpus design criteria.

Thus, Following Gráf's (2017) observation, the transcription scheme of LINDSEI was adapted to ISLC in accordance with the research questions. The corpus remains a collection of written transcriptions (the audio recordings are available upon request) and no prosodic or phonetic information are provided. LINDSEI's transcription scheme has been edited as illustrated by Table 4.2.

Table 4.2: ISLC's transcription scheme based on LINDSEI's transcription rules.

Text	Rule
Punctuation	No punctuation marks have been used to indicate sentence or clause boundaries.
Empty pauses	Empty pauses are not marked in the corpus.

Filled pauses and backchannelling	The standard set of filled pauses and back-channeling used to transcribe data in ISLC include 'uh', 'uhm', 'mhm', 'mm', and 'huh'. No distinction was made between long or short filled pauses. Filled pauses and backchannelling are enclosed in brackets to make it easier for researchers to exclude them from analysis if necessary (uh), (uhm), (mhm), (mm), (mhm), (huh).
Unclear passages or words	A three-tier system has been used to indicate the length of unclear passages or words: 1. <X> represents an unclear syllable or sound up to one word; 2. <XX> represents two unclear words; 3. <XXX> represents more than two unclear words. *e.g., <B> tons of films instead of <X> stories of course </B>* Unclear names of towns or titles of films for example have been indicated as *<name of town>* or *<title of film>*.
Truncated words	Incomplete words are immediately followed by an *equals* sign. e.g., *<B> when people speak abou= when people speak about uh New York <B/>*
Contracted forms	All standard contracted forms such as "I'm", "he'd" or "they'll" have been retained.

Non-standard forms	Non-standard forms that appear in the dictionary have been transcribed orthographically in their dictionary accepted way. The non-standard forms that can be found in ISLC transcripts are “cos”, “dunno”, “gonna”, “gotta”, “kinda”, “wanna” and “yeah”.
Foreign words	Italian or other foreign words are enclosed between angle brackets. e.g., *<B> we couldn’t uh <sorpassare> the cows <B/>*
Abbreviations	Abbreviations pronounced as sequences of letters have been transcribed as a series of upper-case letters separated by spaces. e.g., *<B> I play uh M M R P G which are massive multiplayer </B>* Abbreviations pronounced as words have been transcribed as a series of upper-case letters not separated by spaces. e.g., *<B> in NASA in the NASA museum . yes . I liked the different aircrafts spacecrafts <B/>* (example taken from LINDSEI’s transcription scheme. No instances of this type of abbreviation were found in the ISLC corpus).
Dates and numbers	All figures have been written out in words to represent the way they were pronounced by the speakers. e.g., *<B> sort of uhm five dimensional uhm construction </B>*
Anonymisation	To preserve anonymity the proper names mentioned in the interviews have been replaced with substitutes such as *<first name of interviewee>*, *<first name and full name of interviewer>* or *<name of professor>*. The names of famous people like singers or actors have however been kept.

As regards speaker turns, tasks, and interviews, the mark-up scheme in Table was adapted from LINDSEI.

Table 4.4: Mark-up scheme for speaker turns, tasks, and interviews adapted from LINDSEI.

Mark-up	Rule
Interview identification and delimitation	Each interview transcript is preceded by the code *IT* that indicates both the native language of the learner in the interview and the interview number. All interviews end with the symbol *</h>* on a separate line. e.g., *<h nt="IT" nr="IT001"> interview </h>* In the above example, the code *<h nt="IT" nr="IT001">* marks the beginning of the first interview from the ISLC corpus.
Task definition	The beginning and end of each of the three tasks making up the interview (i.e., set topic, free discussion and story-telling) have been marked as follows: • <S> set topic task </S> • <F> free discussion task </F> • <ST> story-telling task </ST> The markers used to delimit the three tasks always occupy a separate line.

Speaker turns	Speaker turns are displayed in vertical format, i.e., one below the other. While the letter "A" enclosed between angle brackets (<A>) marks the beginning of the interviewer's turn, the letter "B" between angle brackets (<B>) marks the beginning of the interview/learner's turn. The end of each turn is indicated by either </A> or </B>. e.g., *<A> ok very good so do you like travelling? </A>* *<B> yes </B>*
Overlapping speech	Speech overlaps are not marked in the corpus.

Lastly, in order to extract the collocations from the corpus and make the corpus available for further research, ISLC needed to be tagged. Generally, POS-tagging for learner corpora is performed with taggers originally created for native language[3]. Thus, it is fair to expect a poorer performance of the tagger on learner language compared to native language since the tagger is not being applied to the genre it was designed for.

The choice of the POS-tagger stemmed from the promising results that Van Rooy and Schäfer (2003) reported on CLAWS7, which is the most precise tagger according to their study (although the tests were carried out on written learner language). Indeed, after the initial automatic tagging, the incorrect tags are usually edited, either automatically or manually. Studies have shown that the main issues with tagging learner language are spelling, wrong lexical items, wrong inflections, omissions, non-standard syntax (Van Rooy & Schäfer, 2002; de Haan, 2000). Although the first, spelling, may not be relevant for spoken learner corpora (since the transcriber decides on the spelling during the transcription process), the others are still sources of issues for taggers. For example, if a speaker uses a wrong lexical item, or the wrong inflection, omits an article or builds the sentence according to non-standard syntax, the tagger (which is based on standard syntax and the assumption that the text was produced by native speakers) may erroneously tag one item with one grammatical label rather than another. However, CLAWS7 delivers a high level of accuracy and greatly reduces the amount of manual editing work needed.

Thus, CLAWS was chosen as the ISLC tagger because of its accuracy and the fact that it had already been used for spoken native language (Garside, 1995). Furthermore, the high

[3]This transversal use of POS-taggers is called "domain transfer" (Díaz-Negrillo et al., 2010, p. 142).

level of proficiency of the learners was thought to be an encouraging factor in the tagging accuracy. The chosen tagset was CLAWS5 since the aim of this thesis was to obtain a general tag for adverbs and adjectives and limit the amount of editing after the automatic tagging. After the corpus had been automatically tagged, a manual revision and correction of the errors made by the tagger was necessary. The main problems arising from the tagging were related to filled and unfilled pauses (i.e., wrong recognition of "mm", "uh", "uhm") and wrong verb inflection recognition due to non-standard syntax (i.e., *<B> so I didn't really enjoyed it </B>*). These were easily edited manually (no automatic editing was employed since the corpus is of limited size) and collocations checked for correct tagging.

Before drawing the conclusion of this Section, an observation about the tagging process should be made. Research on SLA and language learning aims to investigate IL on its own and comparisons with the TL are tolerated because the learners themselves aim to achieve a given target. The POS-tagging of a learner corpus is based on the grammatical categories of the TL, thus the description of the IL must take this into consideration when analysing the data. The grammatical categories may not be used in the same way as in the TL by the learners. This does not only lead to the manual correction of erroneous tags, but also to the consideration of the TL tags as guidelines for the IL, not as prescriptive rules. This observation may gain less attention in an advanced proficiency context as opposed to lower-proficiency level learner corpora, but nevertheless it should be remembered.

4.2.5 Description of the ISLC corpus

Table 4.6 presents basic data about the ISLC corpus (learners' turns only).

Table 4.6: Data on ISLC corpus (learners' turns only).

Type	Number
Number of texts	34
Number of speakers	34
Number of tokens	58,568
Number of types	15,072
Median text length	1,750
Interquartile range	529.25
Longest text (in tokens)	3,001
Shortest text (in tokens)	1,349

ISLC contains 34 interviews for a total of 66,898 words, this includes the interviewer and the learner's turns (filled pauses and backchannels are counted as words). The learners' turns account for 58,568 tokens. The median number of words per interview is 1,750 (IQR = 529.25); the longest text is 3,001 tokens, while the shortest is 1,349. Table 4.7 shows the distribution of tokens according to task:

Table 4.7: Distribution of words per task (learners' turns only).

	Number of words (%)		
	S	F	ST
ISLC	18,011 (30.88%)	26,653 (44.88%)	13,904 (24.23%)

As outlined in Table 4.7, the free discussion is the task which elicits the highest number of words, accounting for 45% of the interview on average. It is followed by the set topic task at 31% and the picture description at 24%. This is in contrast with LINDSEI_IT, where the set topic accounts for 52% of the total number of words. There is also a slight increase in the percentage of words of the third task, it would seem that changing the task from a picture description to a story-telling one had an effect, albeit a small one, on the number of words. The total duration of the interviews is 9:26:26 and the average duration is approximately 16 minutes and 20 seconds.

The data collection for ISLC involved interviews with 58 Italian learners of English L2. However, it was not possible to include them all in the corpus: 24 participants did not obtain a C1 level in the OOPT. Thus, the total number of interviews contained in the corpus is 34. The average age of the participants is 22 (median = 22, IQR = 2) and 76% are females (26 vs 8). The predominance of females is to be expected as with all humanities courses, thus, although the data may not appear well distributed, it is representative of the population found in a Foreign Languages degree course.

The participants are all Italian native-speakers with Italian native-speaker parents and they attended primary and secondary schools with Italian as the medium of instruction. The average number of years studying English at school is 13 (median = 13, IQR = 1), while the average number of years studying English at university is 4 (median = 4, IQR = 1). There is a 5-year gap between the ISLC's number of years studying English at school and the Italian sub-corpus of LINDSEI, whose learners' studied English for an average of 7.08 years. This is due to the following educational reforms that took place in Italy after the collection of LINDSEI_IT and which introduced the subject of English in elementary

school, thus substituting other languages such as French[4].

As far as university degree is concerned, the participants are well distributed between the undergraduate course (17) and the postgraduate course (17). Since there was no requirement for the attendance of a specific course year, the participants are distributed across the three course year: 12 learners attend the first year of the postgraduate and five attend the second and last year; in the undergraduate, no learners attend the first years, six are enrolled in the second year, and 11 in the third year (Figure 4.2.2).

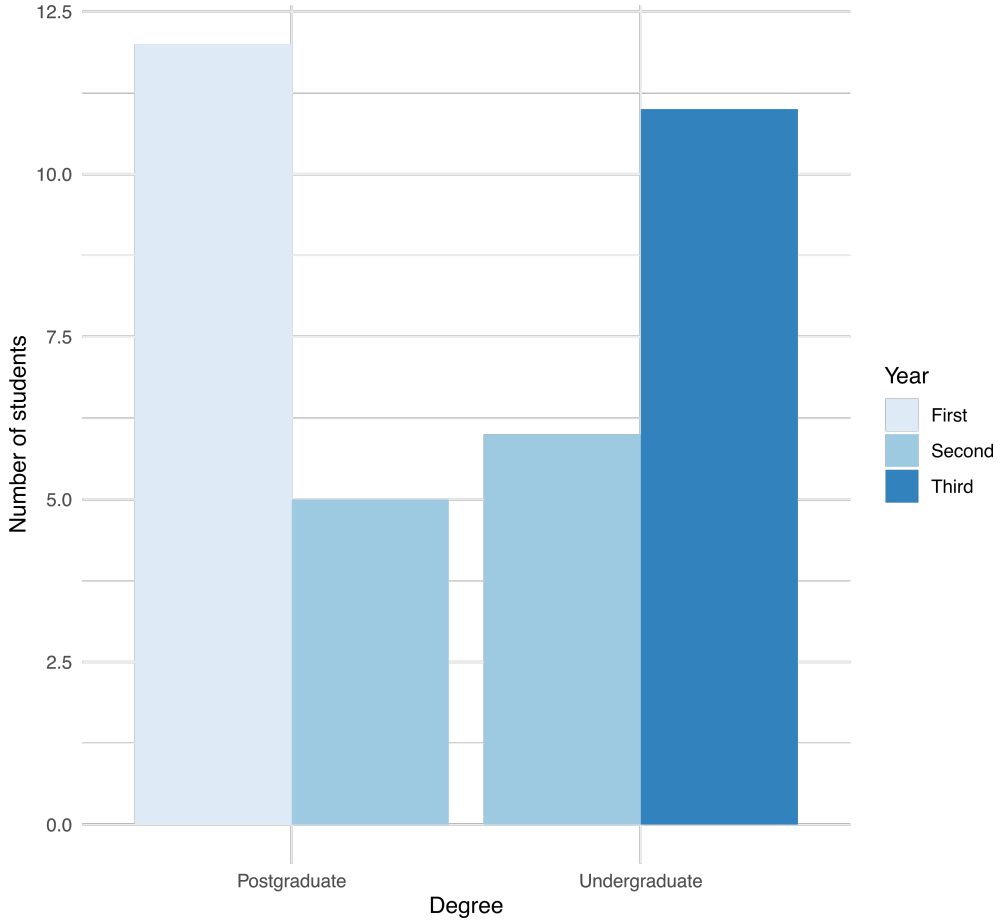

Figure 4.2.2: Distribution of learners attending the undergraduaate and postgraduate courses.

[4]In 1985, the President of the Republic (Sandro Pertini) issued a decree (D.P.R. 12th February 1985 n. 104) in which he stated that a foreign language should be taught in elementary school, giving free choice to the schools to choose what language to teach and for how many hours. Thus, Italian schools were free to manage the teaching of a foreign language until 2004, when English became compulsory in elementary school with a new reform (Moratti reform, 28th March 2003 n. 53). Since then, primary schools have been required to teach English one hour a week during the first year, two hours a week in the second year, and three hours a week for the remaining three years. However, teachers in elementary school are not language specialists and sometimes this is detrimental to the language learning process.

Another variable which was recorded in the learner questionnaire (see Section 4.2.3) is the type of course the learners are attending. Since the learners come from different universities and attend different degree courses, there is a wide range of courses available as variables: there are 11 courses. No official name of the courses is provided so as to guarantee the anonymisation of the universities. All courses are part of the Department of Foreign Languages. The majority of students attend three courses, two undergraduates and one postgraduate one.

As regards the other languages studied by the learners at university, German is the most studied (11 learners), followed by Russian (7) and French (5). There is an equal number of participants studying Chinese and Spanish (4), whereas only two learners attend Japanese courses.

The majority of learners come from University A (30) and were recruited across both campuses (in Lombardy). There are three participants from University B and one from University C. One interesting variable is whether the learners have stayed in an English-speaking country and if so, where and for how long. Figure 4.2.4 provides an overview of this variable.

Most learners (85%) have travelled abroad to an English-speaking country (29): the main destination is the United Kingdom (18), followed by a combination of the countries (UK, Ireland, USA, and Canada) (7). The majority of the learners went to an English-speaking country on holiday, thus for less than a week (18), while only few others spent more time in the country. Three participants spent between 1-4 weeks, one between 1-3 months, one between 3-6 months, and six spent more than 6 months. The numbers are particularly low since in Italy students of foreign languages are not required to spend time abroad for their studies, unlike in other countries, such as the United Kingdom. As regards proficiency (Figure 4.2.5), the majority of the learners obtained a score between 80 and 90 in the standardised test, while few others scored higher between 100 and 110 (all these scores are above the cut-off point of 80 corresponding to a C1 level).

4.3 Extraction of collocations from the corpus

In the selection of collocations, a number of decisions had to be made as to the parameters for the identification and subsequent extraction of collocations. Indeed, the choice of investigating adverb + adjective collocations is not sufficient to proceed with the extraction of these from the corpus as other essential parameters aside from the syntactic pattern need to be adjusted. These are reported in the Sections below. Once the criteria were set out, the

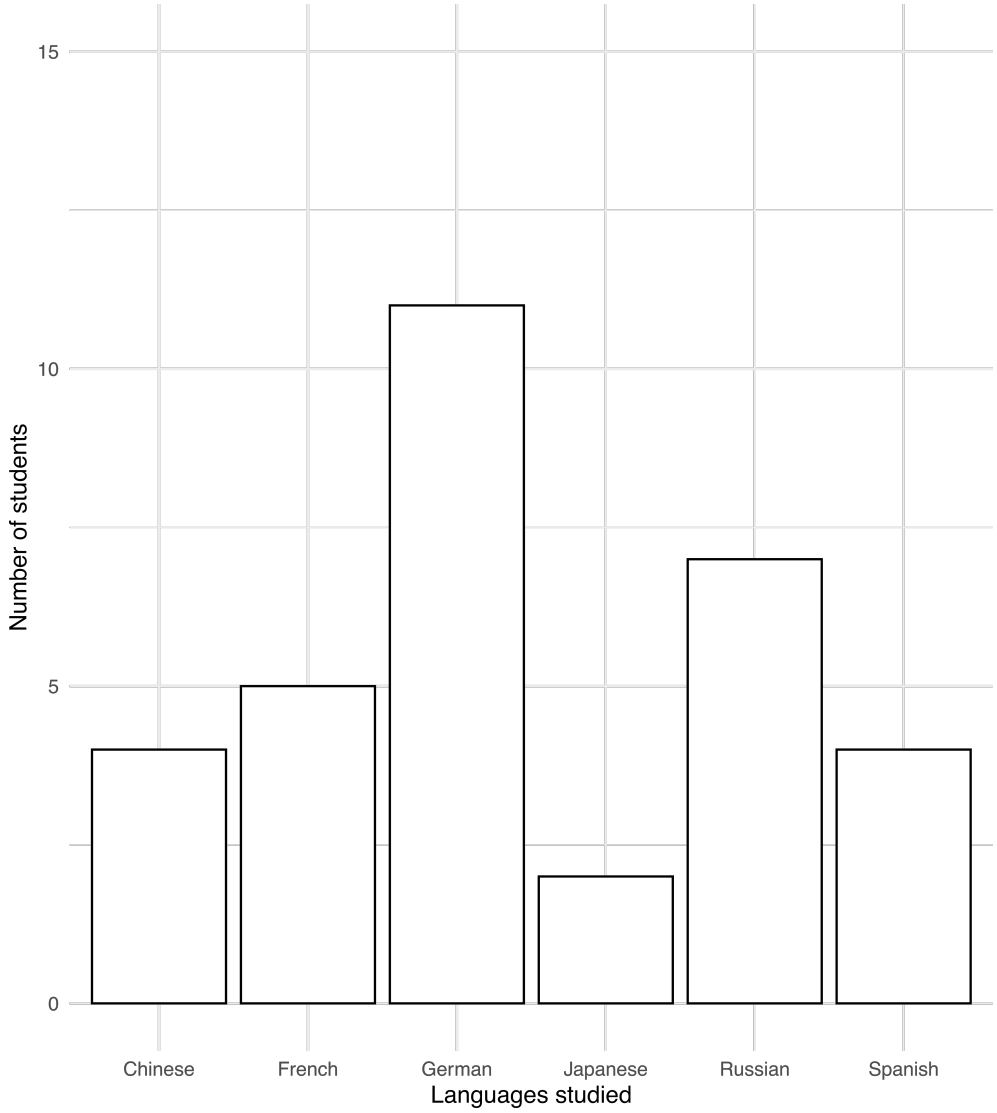

Figure 4.2.3: Other languages studied by the ISLC participants.

combinations were then extracted via means of AntConc (Anthony, 2019).

4.3.1 Approach

As described in Section 3.4.1, collocations can be identified according to two main perspectives or approaches: the phraseological approach (Cowie, 1981; Nesselhauf, 2005) and the frequency-based approach (Firth, 1957; Nesselhauf, 2005). The choice of approach not only impacts the identification of collocations, but it also influences the whole research

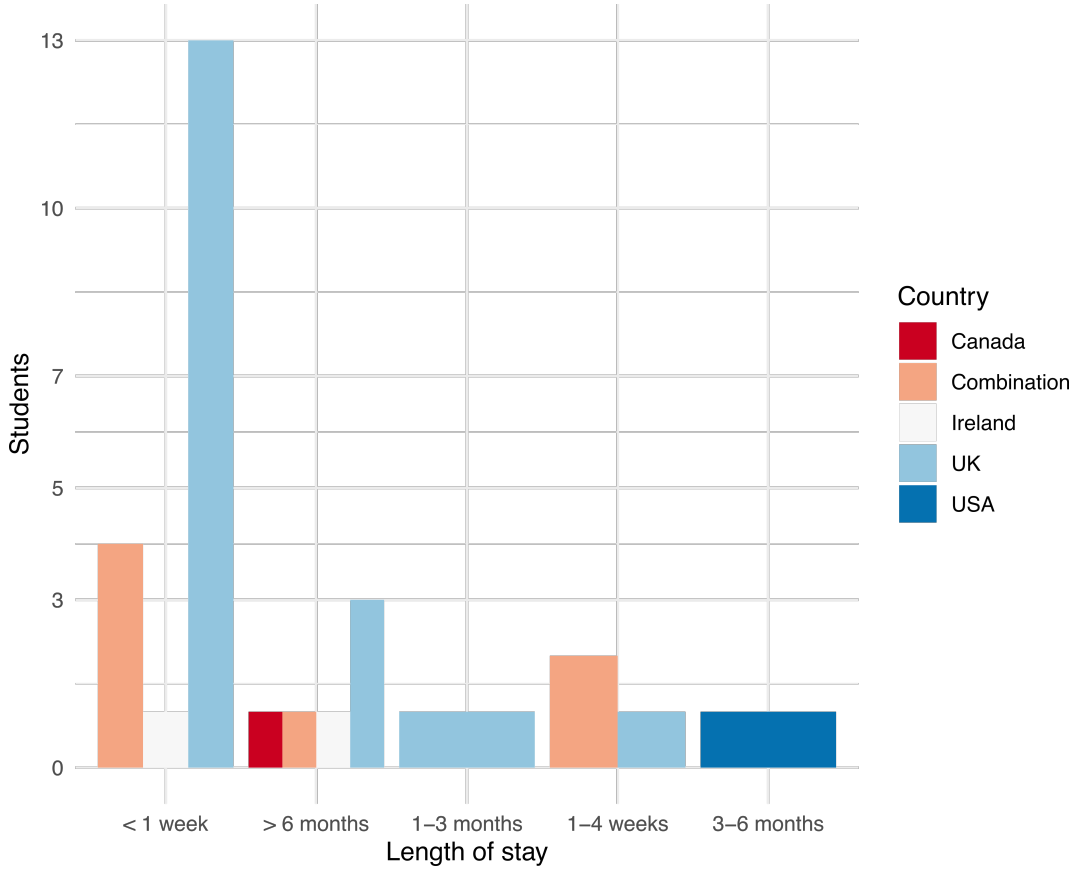

Figure 4.2.4: ISLC participants' amount of time and locations of stay-abroad experiences.

methodology.

In this thesis, the frequency-based approach was chosen for the analysis of collocations. The frequency-based approach does not only identify collocations on the basis of their raw frequency, but it also employs the concept of *more frequent than expected* (Gries, 2008, p. 4). Furthermore, association measures were employed to measure the strength of attraction of the collocations. Indeed, if the strength of attraction is measured, researchers can predict how likely two or more words are to co-occur together. This has also been called the statistical approach by Partington (1998, p. 15). In order to quantify frequency and strength of association, statistical measures that compare the individual frequencies of the collocation components and the joint frequencies of the collocation will be employed and described in Section 4.3.4.

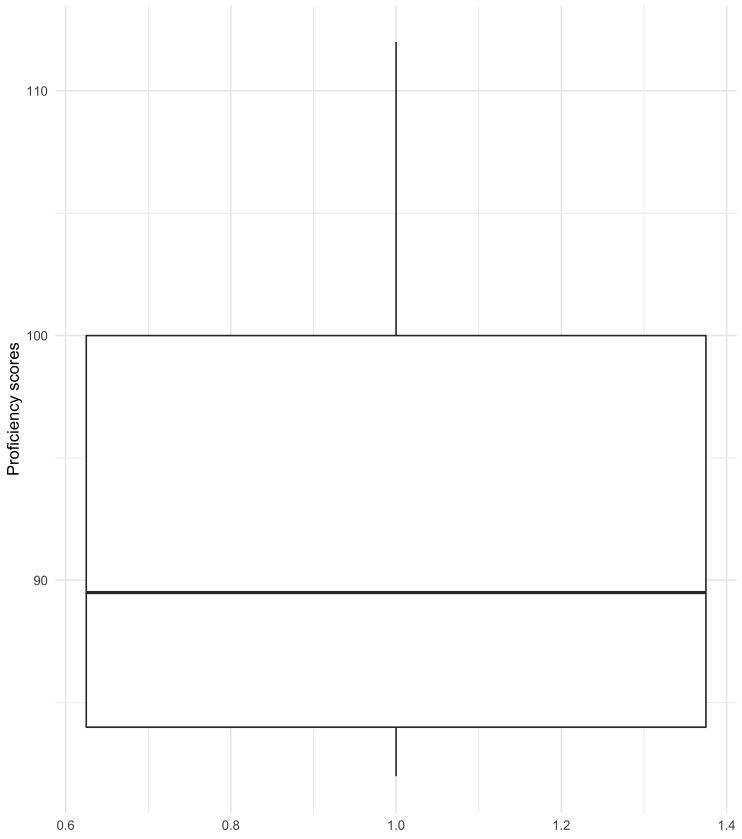

Figure 4.2.5: Boxplot of OOPT proficiency scores of ISLC participants.

4.3.2 Collocational span

Given a node, the number of collocates co-occurring either to the left or right of the node define the collocational span. Sinclair found that a collocational span of ± 4 is ideal for the investigation of collocations (Jones & Sinclair, 1974; Sinclair, 1991) (see Section 3.4.1.2); however, in the present study the span of the collocation was set to - 2, that is, two collocates to the left of the node, rather than - 1. The sorting option was set to two left in order to skip the word immediately preceding the tag *AJ0* which is the adjective itself tagged by CLAWS5. Furthermore, since in conversation attributive adjectives are more frequent and they typically occur in noun phrases before the head noun (Biber et al., 1999, p. 505-506), the choice of the - 2 span seems more appropriate than Sinclair's technical span which is more suitable for subject verb or verb-object collocations.

This allowed for the retrieval of bigrams, that is collocations made up of two words only. This choice was mainly driven by the structure of the collocation this work investigates,

adverb + adjective, for whose extraction it was necessary to limit the span to the left and to one collocate. This also improved performance of statistical measures when applied to adjacent collocational components (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Evert, 2005). During the extraction phase, as will be described in the following Section (4.3.3), the span was temporarily moved to the position -5 (five collocates to the left) in the concordance retrieval software in order to identify all those collocations which were interspersed with a pause, filled pause, or any other speech disfluency.

4.3.3 Syntactic pattern

As already mentioned at the beginning of this chapter, the current investigation examines adverb + adjective collocations according to the frequency-based approach. However, the selection of adverb + adjective collocations is not a straightforward and easy task since the extraction of tagged adverb and adjective items yields combinations which may not necessarily be intensified adjectives. In order to clean the results of the extraction, clear choices must be made regarding which structures will be included in the analysis and which ones will be discarded.

The first choice concerned the inclusion or exclusion of degree adverbs acting as comparatives and superlatives. Since the present study focuses on the learners' use of collocations according to the frequency-based approach, a collocation is identified as "the relationship a lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991, p. 7). This definition provided by Hoey is crucial because collocations are co-occurring "lexical" items, so that the co-occurrence of two items due to reasons other than frequency must be discarded from the study. This is also supported by Van Roey (1990, p. 46), who referred to collocations as "the linguistic phenomenon whereby a given vocabulary item prefers the company of another item rather than its 'synonyms' because of constraints which are not on the level of syntax or conceptual meaning but on that of usage". Thus, degree adverbs used to modify an adjective in the comparative or superlative form must be excluded (e.g., *[. . .] I would prefer to do something more practical more useful for the working environment [. . .]*). Lorenz (1999) included comparatives and superlatives in his analysis of intensification since he argued that they are a grammatical way of scaling upwards or downwards, that is, they are grammatical intensification. However, because collocations are lexical items that co-occur, comparatives and superlatives were excluded from this analysis on the basis of their more grammatical function. Indeed, *more* cannot be considered on par with an intensifier such as *very* due to its grammar-centred role. Furthermore, it is believed that the learners may perceive a combination such as *more*

comfortable not as the co-occurrence of two lexical items, one of which the learners are intensifying, but merely as the required grammatical rule for expressing a comparison.

The second step for the selection of collocations involved the exclusion of all combinations which included an adverb functioning as an non-intensifying adverbial plus an adjective, e.g., *[. . .] I had potatoes every every day there basically baked potatoes uh or smashed potatoes [. . .].* This was done to avoid combinations which were not instances of intensification. In addition, other instances of adverbs occurring before an adjective without direct modification of the adjective were also excluded. This meant the exclusion of adverbs of time (e.g., *[. . .] I also like enjoyable for a walk and listening to the music al= always English and Spanish songs [. . .]*) and additive/restrictive adverbs (e.g., *[. . .] a Japanese guy as well uh that was also interesting from a cultural point of view [. . .]*). Once more this was done to avoid the inclusion of combinations which were not occurrences of intensification.

As regards adverbs of manner, these were included in the extraction only if they took on the role of degree adverbs used as modifiers, such as *[. . .] and she's naturally good naturally beautiful so she can get into the this [. . .].* Indeed, Quirk et al. (1985, p. 448) signalled that subjunct and disjunct adverbs, such as *theoretically* or *technically*, can also function as modifiers: *theoretically sound* or *technically possible*.

Hedging fixed phrases such as *kind of* and *sort of* (and their slang forms *kinda*, *sorta*) were also included in the analysis as it was thought that learners may use these also as downtoners. Lastly, only directly adjacent combinations were extracted from the corpus (with one exception described below) for one main reason: admitting a wider range of distances could render the association measures scores not comparable between collocations (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger & Bestgen, 2014). However, since previous studies employing the approach based on the association measures scores used written learner corpora, the adjacent pairs were easy to identify and extract. In the case of a spoken learner corpus, otherwise adjacent pairs were sometimes interrupted by a filler, a pause, or a speech disfluency such as *[. . .] in Brescia at home I feel very like independent and as if I don't [. . .]*, or *[. . .] or maybe that could be a bomb so something very uhm uh dangerous uh but uhm [. . .].* In order to account for these variations, during the extraction process, the concordance lines were sorted to the left of the node up to 5 collocates in order to identify all those disfluencies or fillers which interrupted the otherwise adjacent collocation adverb + adjective.

One exception to these criteria was made as regards those combinations which included double intensification, such as *[. . .] and it's kind of very very different from ours but it's tasty [. . .].* Double intensification has already been noted in the literature (cf. Tagliamonte,

2008; Putri et al., 2020), although there is still ample ground for research on the topic. Quirk et al. (1985, p. 473) wrote that some intensifiers can be repeated for emphasis, but usually these are the ones pre-modifiying comparatives, rather than non-graded adjectives, although the authors warned that the repetition is generally permissible only if the repeated items come first or follow *so* (i.e., *very very good*, *so very very good*). However, the phenomenon also occurs in spoken Italian and is called *reduplicazione espressiva* – expressive reduplication – cf. De Santis, 2014). Thus, in light of the singularity of this phenomenon and its presence in the L1 (which could account for L1 congruency influence), the decision to include collocations with double intensification was made in order to further inspect the behaviour of learners.

4.3.4 Association measures

Frequency of co-occurrence in the identification of collocations has been discussed in Section 3.4.1.2, while presenting the frequency approach of the British school, and then in Section 3.5.2 in the description of statistical measures adopted in the literature for the investigation of collocations.

In order to measure the significance of collocations, measures of attraction between collocates called lexical association measures are adopted (see Section 3.5.2 for a description of the main association measures). In this thesis, two of the most common simple association measures (Evert, 2009) were chosen for the study of collocations: MI and *t*-score (Durrant, 2008; Durrant & Schmitt, 2009; Evert, 2009; Gries, 2013; Hunston, 2002; Schmitt & Siyanova, 2008; Spina, 2001). Both measures tend to highlight two sets of collocations: the former inflates the scores for less frequent but stronger collocations, while the latter assigns a higher score to frequent collocations. MI and *t*-score (computed on a large reference corpus) will be used concurrently to select and identify collocations by means of combining the two types of information these measures provide (Durrant & Schmitt, 2009). This is based on Durrant and Schmitt's (2009) work and subsequent research such as Bestgen and Granger (2014) and Granger and Bestgen (2014).

As in Durrant and Schmitt (2009), cut-off points were adopted for the association measures scores: since MI is not calculated on a specific scale with minimum and maximum values (the only notable value is 0, that is that observed and expected frequency coincide), a cut-off point of MI ≥ 3 has been indicated as an indicative value for collocations (Hunston, 2002). As regards the *t*-score, it has been suggested that *t*-score ≥ 2 is a valid threshold for the classification of collocations (Hunston, 2002). This approach proved to be successful in identifying sets of collocations in learner corpora because each measure highlights different

types of collocations and their combination allows for a balanced extraction of adverb + adjective pairs.

Furthermore, the association measures were calculated on the basis of a large reference corpus (i.e., the corpus needs to be large enough for the precise estimation of the association scores and it must be representative of the TL[5]), contrary to what Lorenz (1999) did in his work (he calculated the association measures on the basis of his learner corpus), which only highlighted salient collocations for the learners, rather than for native speakers of English. For this reason, the BNC, a 100-million-word collection of samples of written and spoken language representing British English from the second half of the 20th century, was chosen for the present work and a minimum frequency threshold per each combination was set to five.

In summary, each adverb + adjective combination was looked up in the BNC and assigned its MI and *t*-score according to the formulas reported in Evert (2009, p. 1225) and described in Section 3.5.2. These two association indices could only be computed if the combination occurred in the reference corpus (the frequency threshold was set to 5). If it failed to occur, the combination was included in the *infrequent/unattested collocations* category. If the association measures scores were below the cut-off points, the combination was included in the *grey area collocations* category. The categories were assigned these labels because of the following reasons: as regards "infrequent/unattested", since the association measures are computed on the BNC, it may be that some adverb + adjective collocations are chiefly American English, which would result in infrequent occurrence or no occurrence at all in a corpus of British English. Secondly, "grey area collocations", which have not been investigated in either Durrant and Schmitt (2009) nor Bestgen and Granger (2014), reflect the idea that these collocations received a score below the cut-off points and, because of the arbitrariness of these cut-off points, these might be considered collocations if analysed from a different perspective or approach (e.g., the phraseological approach). Thus, these are not collocations according to the identification criteria adopted in this study, but may be collocations if different parameters were to be applied.

4.4 Analyses

The following Sections describe the rationale and steps taken in the analysis of the data extracted from ISLC and LOCNEC, which was used to attempt to answer the research questions presented at the beginning of this chapter.

[5]See Bestgen and Granger (2014).

4.4.1 Quantitative analyses

4.4.1.1 Procedure

In order to address the first research question and its three sub-questions (see Section 4.1), which aim to provide a perspective on the quantity of collocations produced by the learners and native speakers, a quantitative approach was adopted.

First, the combinations extracted from ISLC and LOCNEC were counted based on types and tokens and compared in terms of average speaker production; this means that, in line with the need for more learner corpus studies accounting for individual variation (Callies, 2015) and following Durrant and Schmitt (2009), for each speaker the average number of combinations (types and tokens) was computed. Then, the averages were taken and compared with those of LOCNEC via means of visual inspection and an independent two-tailed *t*-test ran in R (R Core Team, 2017). The advantage of computing not only the overall average figure for each text, but also the average figure for each text, is that it is possible to take into account the degree of variation between texts (Durrant & Schmitt, 2009). This analysis was performed in order to assess whether the number of combinations extracted would be higher in ISLC or LOCNEC, thus answering the first sub-question, namely whether the number of collocations produced by learners would be similar or different to native speakers.

The second and third analyses, which address the second sub-question, stem from the desire to investigate a part of collocations that has often been discarded in other studies. Usually researchers examine the quantity and quality of collocations produced by the learners in comparison with those produced by native speakers, but infelicitous combinations often do not receive such attention, especially from a statistical approach perspective (cf. Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Nesselhauf, 2005). In order to explore this part of the learners' production, collocations which did not receive association measures scores (due to their infrequency in the BNC) or received a score lower than the cut-off points were divided into two categories: the infrequent/unattested collocations and the grey area collocations respectively. The choice of dividing infelicitous combinations into separate categories is two-fold: first of all, they are two different phenomena. Infrequent or unattested collocations are classified according to a frequency threshold in the reference corpus (which in this case is 5), which means that if they do not occur at least five times the BNC no association measures were calculated. This category has also been already investigated, albeit in a brief manner, in the literature: Durrant and Schmitt (2009), as well as Ferraresi and Milicevic (2019), conclude that non-native texts contain a higher number of low-frequency combinations. The second category instead involves combinations

which are not included in the study because the arbitrary cut-off points of *t*-score ≥ 2 and MI ≥ 3 are not met. This means that these combinations occur in the BNC and association measures scores were calculated for these combinations. Some of these combinations obtained negative association measures scores, which indicate rejection rather than attraction (Granger & Bestgen, 2014), while others simply did not qualify as collocations because their scores were short of a few decimals. Thus, this category includes low-collocational items. The second reason for the choice of keeping the two categories separate stems from Durrant and Schmitt's (2009) argument for their collocational bands. The authors posit that collocations should be viewed on a scale, rather than as a simple dichotomous collocation vs non-collocation concept. However, the authors limit their argument to collocations and treat non-collocations as a singular category, rather than distinguishing between different non-collocations. The present work thus aims to distinguish between low-frequency combinations and infelicitous or other variety-based (i.e., other English varieties such as American English) combinations in order to provide a more fine-grained view of non-collocations.

In order to analyse these two categories of collocations, a percentage of the number of infrequent/unattested or grey-area collocations over the total number of combinations produced was calculated for each of the 34 texts in ISLC and each of the 50 texts in LOCNEC (types and tokens) and for each category. The averages were taken and compared between the corpora via means of visual inspection and through either an independent two-tailed *t*-test (in case the data were normally distributed) or its non-parametric version, the Wilcoxon rank sum test (in case the data were not normally distributed). The tests were run in R. Again, the individual average and the overall average allow for better consideration of individual and text variation.

Finally, the last analysis addresses the third sub-question , that is whether there are any differences between the two groups in terms of strength of association of the collocations produced. In order to investigate any differences between the groups in terms of strength of association a new approach, which diverts from Durrant and Schmitt (2009), was adopted. Indeed, although this study has so far largely relied on Durrant and Schmitt's (2009) work on collocations, it could be expected that the next step employs the authors' second-level cut-off points for weak and strong collocations (namely, ≥ 10 for the *t*-score and ≥ 7 for MI). This would allow for a comparison between the groups in terms of strength of association below or over a new cut-off point for weak and strong collocations. However, Durrant and Schmitt's (2009) approach could not employed in the present work since no collocations with an MI score greater than 7 were produced in the two corpora, where 7 is the MI cut-off point for strong collocations. Thus, this method was discarded in favour of a more text-centred approach as found in Ferraresi (2019) and Bestgen and Granger (2014). The latter

computed the association measure scores of bigrams (*Collgram technique*) and compared their means for the longitudinal analysis. The former, in his analysis of collocations in constrained varieties of English, calculated the mean *t*-score and MI for each corpus text (both for types and tokens). Thus, two values correspond to each text, a mean *t*-score and a mean MI calculated on the basis of all the collocations contained in every single text. This mean score provides an indication of the strength of the collocation which is not based on an arbitrary cut-off point as in Durrant and Schmitt (2009), but rather it is an average that emerges from the texts of the corpus produced by the participants. Although the association measures scores on which the means are computed are based on the reference corpus, the averages thus computed give a better indication of the general trend of association measures scores without biases generated by arbitrary cut-off points giving rise to weak or strong collocations.

Following on from this, once the mean scores were computed, they were then compared between corpora both by visual inspection and by an independent two-tailed *t*-test for the average *t*-score and by a Wilcoxon rank sum test for the average MI (the frequency threshold was modified to 5 rather than Ferraresi's 3). Both statistical tests were run in R.

The advantages of this approach are three-fold: 1) the cut-off points suggested by Durrant and Schmitt (2009) are arbitrary values which, although they have been demonstrated to highlight differences between native and non-native texts, may not necessarily be appropriate in other texts/text types; 2) the mean score approach is better suited in the field of LCR, where individual variation has often been disregarded and instead requires more consideration (Callies, 2015): this approach does take it into account; 3) collocation scores are genre- and corpus-dependent (e.g., MI scores are usually higher in written texts, compared to spoken ones), thus the mean scores ensure that the analysis is carried out on the data inherent to the texts and corpora, without any generalisation regarding text type and genre.

4.4.1.2 T-test

The *t*-test is a parametric test which is used to verify whether there is statistical difference between two groups. It is a type of inferential statistic and, mathematically, it assumes that the means of the distributions of two populations are equal (null hypothesis). If the *t*-test rejects the null hypothesis, it indicates that the groups are indeed different (alternative hypothesis). The test can be *one-tailed* or *two-tailed*: this terminology refers to the alternative hypothesis, that is the difference between the two groups. In the case of the one-tailed version of the test, it is assumed that the variable observed in one group – for example, group A – is greater than the same variable in the other group, group B. This means that the statis-

tic will only test the difference in that direction and not in both ways. On the other hand, if the test is two-tailed, it means that the alternative hypothesis assumes that there is difference between the two groups, but there is no indication about the relationship between the groups and whether the variable is greater or lower than the other, so both directions will be tested.

Furthermore, the test can be *paired* (dependent) or *unpaired* (independent): this refers to whether the data of both groups come from the same participants or not. In the case of the paired test, two groups of data may have been collected in an experimental condition which consisted in collecting from the same subject data at different points in time (e.g., before/after some input). For example, the two sets of data may refer to participants' reaction times reading non-formulaic language vs reaction times reading formulaic language. In case of an unpaired test, the data comes from two sets of participants: native speakers' collocations vs learners' collocations.

The main assumptions of the *t*-test that should be met when performing the statistic in the field of corpus linguistics are:

- The observations are independent, both between the groups and within the groups
- The data in both samples are normally distributed and/or the sample sizes are greater than 30
- Homoscedasticity

As regards the first assumption, the independence of the observations, this refers to the fact that each observation (text or speech item) comes from a different speaker and the use of language by one speaker in the sample is not affected by the use of language by another speaker (such as in an interaction) (Brezina, 2018, p. 189). Secondly, the data should be normally distributed and/or the sample sizes greater than 30. In terms of distribution, this can be tested via the Shapiro-Wilk (Shapiro & Wilk, 1965) test for normality. The statistic tests the null hypothesis that the sample x comes from a normally distributed population. The test is used to derive the p value, which, if it is greater than 0.05, indicates that the null hypothesis cannot be rejected, and thus the sample is normally distributed. The test can be computed in R. As regards the last assumption, homoscedasticity, this refers to the equality of variances, i.e., the amount of variation in the two groups that are compared. However, in the literature (Brezina, 2018; Boneau, 1960; Levshina, 2015; Lumley et al., 2002) the *t*-test has been shown to be robust and can be used even with very skewed (not normally distributed) samples and, if the test is run in R (which is the case for all statistical tests in the present work), Levshina (2015) advised that there is no need to be concerned about

the variance since the standard implementation of the *t*-test includes Welch's adjustment, "which provides a correction for unequal variances" (p. 88).

Lastly, in addition to the statistical test, an effect size measure was calculated to evaluate in standardised terms the size of the difference between the two corpora (Brezina, 2018, p. 190). The most common effect size measures are Cohen's *d*, Pearson's correlation coefficient *r*, and odds ratio (Field et al., 2012, p. 262). For the present work, Pearson's correlations coefficient *r* was used as a measure of effect size due to its constrained nature to lie between 0 (no effect) and 1 (a perfect effect) and because the group sizes are discrepant (Field et al., 2012, pp. 263). Cohen (1988, p. 79–80) provided an indication of what constitutes a large or small effect:

- $r = 0.1$ small effect. In this case the effect explains 1% of the total variance;
- $r = 0.3$ medium effect. In this case the effect explains 9% of the total variance;
- $r = 0.5$ large effect. In this case the effect explains 25% of the variance.

The effect size measure was computed in R which requires first to retrieve the value of the *t* statistic and then to retrieve the degrees of freedom. The coefficient is then calculated by executing *sqrt(t^2/(t^2+df))*. This is the square root of the squared *t* statistic divided by the sum of the squared *t* and the degree of freedom (*df*) (Field et al., 2012, pp. 384-385).

4.4.1.3 Wilcoxon rank sum test

The Wilcoxon rank sum test (also called Mann-Whitney U test) is the non-parametric version of the *t*-test. Non-parametric refers to the fact that the test does not have any assumption regarding the normal distribution of the variable in the populations, nor about the homoscedasticity. Thus, the Wilcoxon rank sum test is the ideal test in case of extremely skewed (or not normally distributed) data. Contrary to the *t*-test which is based on actual scores of the populations, the Wilcoxon test uses the ranks of the scores in each group for the testing. Indeed, in the Wilcoxon test all observations from both groups are put together and ranked, then the test statistic *W* is computed as the sum of the ranks in the smaller group (Levshina, 2015, p. 109). Levshina (2015, p. 109) stated that the Wilcoxon statistic can be said to test the difference in medians between the groups; this is true in case the two groups share a similar distribution shape (e.g., positively skewed), while it would be more correct to say that the statistic tests mean ranks in case of different distributions.

The Wilcoxon rank sum test, just like the *t*-test, can be one-tailed or two-tailed and paired or unpaired, although in case the test is paired (or dependent) it acquires a different name:

Wilcoxon signed-rank test. In the present work, since the observations are independent, the unpaired version of the test was used. In addition, if the Wilcoxon rank sum test is run in R, like in the present work, there are two ways in which the software can compute the statistic: the exact test or normal approximation with continuity correction. In the first case, the software creates datasets that match the sample, but instead of putting participants into the correct groups, it puts them into a random group and repeats the process multiple times. In this way, the test looks at how often the difference that appears in the data when the null hypothesis is true is as large as the difference in the two groups (Field et al., 2012). The normal approximation instead does not assume that the data are normal (null hypothesis), but rather that the W statistic is normal, which means that a standard error can be computed and be used to calculate a z and hence a p value. This is normally used when the sample is larger than 40. When the test is run in R, the function is automatically set to default so R will decide based on the size of the sample whether to run the exact test or the normal approximation with continuity correction.

Lastly, an effect size measure was employed in order to quantify the difference tested by the Wilcoxon test. The coefficient r was calculated following the formula provided in Field et al. (2012) which involves retrieving the z-score from the computation by using the *qnorm()* function. Then, the z-score is converted into an effect size estimate following the formula (Rosenthal, 1991, p. 19):

$$r = \frac{z}{\sqrt{N}}$$

in which z is the z-score and N is the total number of observations on which z is based (Field et al., 2012, p. 1913). The same indication provided by Cohen (1988, pp. 79–80) regarding the large or small effect is valid:

- $r = 0.1$ small effect;
- $r = 0.3$ medium effect;
- $r = 0.5$ large effect.

4.4.2 Qualitative analysis

4.4.2.1 The importance of the qualitative approach

In order to address the second and third research questions (see Section 4.1), which require a further investigation of the collocation and their lexico-grammatical patterns, a qualitative

approach was adopted. Qualitative work is often discarded by the scientific community in favour of more quantitative analyses which are evidenced by experiments and statistical data (Tracy, 2010). In particular, the field of corpus linguistics has been largely dominated by quantitative approaches, especially after the advent of super-corpora and datasets such as the BNC (100 million words), or the COCA (450 million words) (Hasko, 2020, p. 2), and more recently for learner corpus studies, the TLC (4 million words). Indeed, although CIA had originally been conceived as both a quantitative and qualitative approach to learner language (cf. Granger, 2009, p. 18), LCR in general has tended to favour quantification over qualitative investigation (Ädel, 2014). However, Hasko (2020) posited that smaller genre-specific, learner-centred, context-dependent, and culture-bound learner corpora could be seen as a more ecologically valid[6] source of data for research into the dynamics of IL development of a specific learner population. Customised corpora can aid researchers in a more in-depth and fine-grained analysis of the details of learner language and relate these details to the specifics of the texts or the learner population. In addition, a purely quantitative approach to learner language can only provide notions of "overuse" and "underuse", that is, tell whether a linguistic phenomenon occurs more or less often in corpus X than in corpus Y, but it cannot evaluate the (over- or under-) use of a given phenomenon from the perspective of grammatical correctness (Ädel, 2014, p. 69). As regards advanced learners especially, the quantity may be similar to native speakers, but it is only through a qualitative investigation that researchers can uncover similar or different lexico-grammatical patterns, communicative competence, and the communicative goals of the learners.

Nevertheless, despite the now established importance of the qualitative approach, it is the combined approach between qualitative corpus analysis and quantitative studies that should be viewed as the ideal goal by researchers, as the merger between these can account for the contextual factors that situate quantitative findings in the ecology of human communication (Hasko, 2020, p. 4). Furthermore, qualitative analyses can often explain and illustrate the statistics of occurrence, thus providing a well-rounded and nuanced answer to the research questions (Timmis, 2015). Indeed, in the present work both approaches have been employed for the exploration of learner data in a comparative perspective.

4.4.2.2 Qualitative hallmarks

Thus, as explained in the previous Section, qualitative research can provide invaluable insights, provided that high quality qualitative methods are employed. For this reason, in the present work, an attempt was made to follow Tracy's (2010) eight universal hallmarks in

[6]That is, better predictors of real-life patterns.

order to provide the most rigorous qualitative analysis. The eight key markers of quality in a qualitative work provided by Tracy are the following:

1. worthy topic;
2. rich rigour;
3. sincerity;
4. credibility;
5. resonance;
6. significant contribution;
7. ethics;
8. meaningful coherence.

A worthy topic is necessary for qualitative research and it should be "relevant, timely, significant, interesting, or evocative" (Tracy, 2010, p. 840). The present work is concerned with analysing the production of collocations of advanced Italian learners of English with a frequency-based approach. The interesting slant of the research is the fact that collocations have seldom been investigated in spoken language. Despite spoken language being thought to be a more direct reflection of the learners' IL (Myles, 2015), there are few studies which have focused on collocations in spoken language. The results of the present work may add new insights to the knowledge of collocations the scientific community currently possess.

Secondly, rigour is required in qualitative research: this means that the researcher should provide sufficient and appropriate theoretical constructs, information regarding the collection of the data and the analytical processes. Sincerity is requested on behalf of the researcher to point out any subjective values or biases which may affect the research, as well as transparency about the methods and the limitations of the work. Credibility is achieved by showing the trustworthiness and plausibility of the research findings. This is done by providing "thick description(s)" (Tracy, 2010, p. 843) of the frameworks and details of the contexts and circumstances of the data. The goal is to show the readers so that they may come to their own conclusions. Furthermore, if a triangulation of methods is used, this could also increase the validity of the results if different and contrasting methods lead to identical findings.

Resonance, which is perhaps one of the most difficult markers to achieve, refers to the researchers' ability to reverberate their research to a wide audience. This can be attained by

either generating a beautiful and evocative text which moves the readers and lead them to enquire about the research itself, or by the research's potential to be valuable and insightful across a range of contexts and situations. This is "generalizability or transferability" and, despite these concepts being closely related to statistical generalisability, it is possible for qualitative research to produce generalisations and transferability.

Significant contribution is another hallmark which may be closely associated with the initial one, worthy topic. A significant contribution should complement the current state of the literature and answer questions which will contribute to the understanding of a specific topic.

The penultimate marker of quality has to do with ethical procedures, situations, relations, and exiting ethics. This means that researchers adopt the appropriate ethical procedures when involving participants by "securing all personal data" (Tracy, 2010, p. 847), thus avoiding any disclosures of their personal information. Furthermore, the results will be shared by taking into consideration the same ethics involved during the collection and analysis processes. In this regard, the present work was approved by the Ethics Committee of Università Cattolica del Sacro Cuore.

Lastly, meaningful coherence aims to encourage researchers to "a) achieve their stated purposes; b) accomplish what they espouse to be about; c) use methods and representation practices that partner well with espoused theories and paradigms; and d) attentively interconnect literature reviewed with research foci, methods, and findings" (Tracy, 2010, p. 848).

4.4.2.3 Procedure

In order to address the research questions from a qualitative perspective and by adopting a rigorous method of analysis, it was decided to systematically tackle each collocation according to a three-fold scheme. Furthermore, the qualitative analyses were split into two chapters, Chapter 6 focuses on the first set of collocations that were chosen for the analysis, while Chapter 7 deals with the second set of collocations analysed (more on this in the next paragraphs). The main hypotheses for these two investigations are the following: as regards the first set of collocations, it is expected that, since the collocations are frequent ones, the learners will display a similar behaviour to the native speakers, thus suggesting complete mastery of them. As regards the second set of collocations, which were infrequent or not present in the BNC, it is expected that these are examples of infelicitous combinations produced by the learners and L1 transfer is hypothesised to have had an influence on their production.

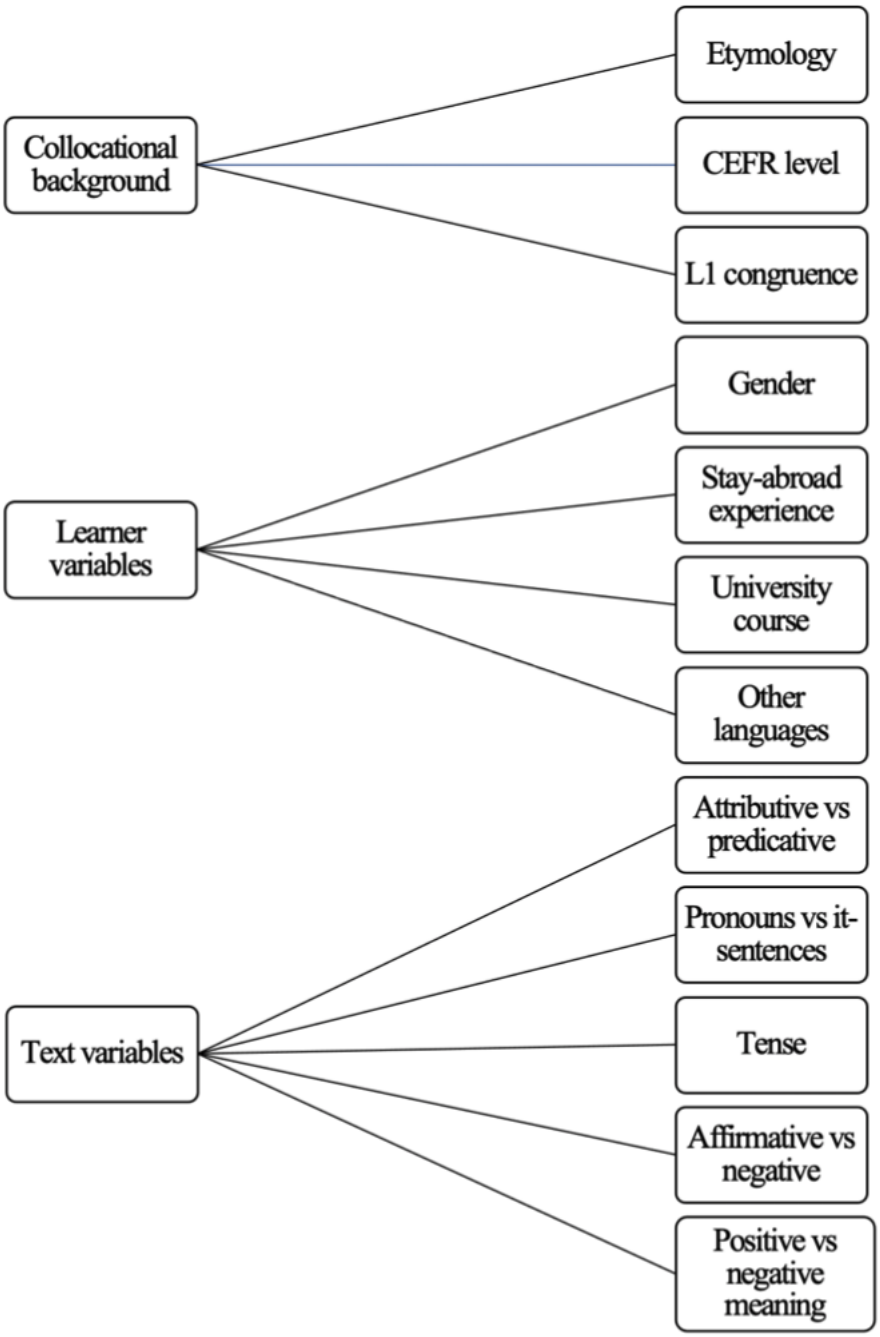

Figure 4.4.1: Qualitative analysis variables scheme.

The graph in Figure 4.4.1 shows three main variables according to which collocations are analysed. The first variable is the *collocational background* and this includes the etymology of the collocation (both the adverb and the adjective) and its L1 congruence or non-congruence. Despite the relatively infrequent use of etymology in learner corpus stud-

ies, the etymology of the individual components of the collocations may shed some light on the learner use and it is strictly connected to the L1 congruence. The Oxford English Dictionary (OED) has been used to check the etymology and use of all the collocational items analysed. If the etymology of part of a collocation can be traced back to the learner's mother tongue, then it could be hypothethised that the chances of the learner knowing the meaning, and potentially the use of the collocation, may be higher. Etymology is related to L1 congruence too since a word's origin may also dictate its similar use to the L1 (hence a better learner performance), as demonstrated by Granger (1998b). The collocations and their etymological definitions have been displayed in a table (Table 6.1) in an attempt to provide a first-glance overview of the collocation's etymology. Furthermore, three labels have been devised to simplify the reading of the table:

1. *same etymology;*
2. *mixed etymology;*
3. *different etymology*.

The three labels are assigned according to one main parameter: the etymons of the lexical items of the collocation must stem from Latin. If both items stem from Latin, the collocation is considered to share the same etymology; if one of the two items stems from Latin, the collocation is labelled *mixed etymology*; and if none of the items stem from Latin, the collocation has a different etymology.

L1 congruence was verified following Wolter and Gyllstad (2013): in their work, Gyllstad, who is a native speaker of Swedish, determined whether the English collocations selected were congruent based on a direct translation into Swedish. The translation was made on the basis of the prototypical semantic value (Verspoor & Lowie, 1993), which means that each collocation had to have a direct word-by-word translation. A similar approach was employed by Bahns (1993) and Nesselhauf (2005). In the present work, an analogous approach was followed. However, due to high level of polysemy that the translations of some collocational items have, it was decided to involve seven native speakers of Italian, rather than only one like in Wolter and Gyllstad (2013). Indeed, the adjective *good* may be translated to Italian as *buono, bello, bravo, bene*, according to the context; similarly *nice* can have multiple translations depending on the context. Thus, rather than having the author choose the most frequent or immediate translation, it was decided to assign this task to seven different native speakers of Italian. The native speakers all have a background in linguistics, are either PhD students or members of the academic teaching staff of one of the universities involved in the research, and are proficient users of English L2. The native speakers were

asked to provide the direct word-by-word translation of the collocations object of the present analysis and the most frequent translation among the native speakers was chosen as the final one[7]. If the prototypical semantic translation was not possible or no agreement was found among the native speakers on one translation, the collocation was marked as *incongruent*. In addition, and following Wolter and Gyllstad (2013), frequency estimates about the Italian translations of the L1 congruent collocations were obtained from the Perugia Corpus[8] (PEC), a reference corpus for contemporary Italian which consists of 26 million words of spoken and written Italian divided in 10 different text genres. The frequency occurrences are reported normalised to one million words (pmw). Furthermore, whenever L1 congruence was hypothesised to be responsible for a particular usage of a collocation, occurrences from PEC were further analysed and compared in order to extract potential patterns of use in the learners' IL. This in particular was carried out in the analysis of Chapter 7 where the learners' collocations were thought to have been produced either by L1 influence or by exposure to another English variety (i.e., American English).

The CEFR level was added to the analysis as the proficiency level of the words composing the collocation was thought to be a good indication of whether the learners might have a good grasp of the individual components and perhaps the collocation itself. The CEFR was verified by consulting the Word Family Framework (WFF), a British Council free vocabulary resource which consists of 26,000 vocabulary items with their relative CEFR levels[9]. The interface allows the user to search for a word and the result indicates the grammar class and the CEFR level to which it belongs (as shown in Figure 4.4.2 and 4.4.3).

The second variable below collocational background is *learner variables* which includes gender, stay-abroad experience, university course, and other languages. This category includes some ethnographic data that may play a role in a learner's good or poor performance as regards collocations. Gender has been included since, despite the predominant number of females in language studies, it could be interesting to analyse whether there are any different collocational patterns between males and females. The second variable, stay-abroad experience, has been taken into consideration as other studies have shown that stay-abroad experiences may have a positive impact on learner language (Tracy-Ventura et al., 2016), although Bratankova (2015) has shown that 18 months of exposure to the TL is not sufficient to improve knowledge of Italian collocations. The third variable regards the university course attended by the learners. In Italian universities each department decides on a selection of courses to offer to the students who enrol. For example, the Department of Foreign

[7]The results are presented in the Appendix.
[8]https://www.unistrapg.it/perugiacorpus/index.html
[9]For more on the WFF, see West (2015).

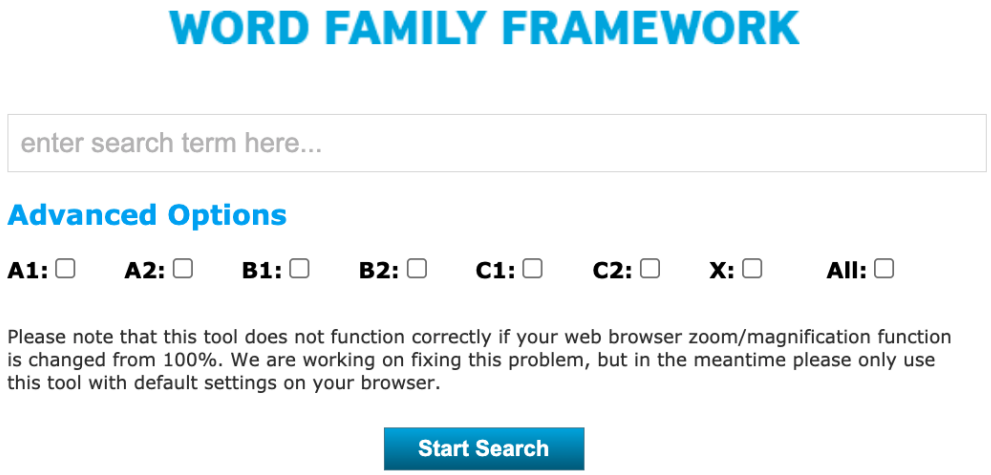

Figure 4.4.2: WFF search interface.

Languages may have a selection of three courses, each one with a different focus (e.g., languages and international relations, languages and literature, and so on). Each course then has another pre-selected collection of courses the students have to attend. The hypothesis here is that students attending different courses may show different patterns of collocational behaviour. Lastly, the knowledge of other foreign languages was thought to have either a positive or negative influence on the learners' knowledge of English collocations. This could be mainly due to either the L2-L3 congruence or non-congruence, such as in the case of a learner studying English and Chinese, where Chinese may have a negative effect on the production of English collocations due to the high level of non-congruence.

The third macro-category in Figure 4.4.1 regards *text variables* and its goal is to assess the effective use of the collocation in comparison with the reference corpus LOCNEC. The hypothesis is that in some cases, although the collocation may quantitatively appear identical to the native speakers' use, the learners could misuse it by either using it more frequently in an attributive position rather than predicative, or by generating a sentence with the *it*- construction rather than with a personal pronoun. For this reason, in this category the adjective function and the sentence construction have been included together with tense (e.g., *This was a very good film* vs *This has been a very good film*), the affirmative or negative valence of the sentence (e.g., *This is a very good film* vs *This isn't a very good film*), and the positive or negative meaning of the collocation (e.g., *very good* vs *very bad*). Furthermore, pre- and

Headword	A1	A2	B1	B2	C1	C2	X
house[1] home, dwelling	***house*** *nC*	***household*** *nC* ***housing*** *nU* ***move house*** see also @ **move**	***house*** *vT* put sb into a house ***housewife*** *nC* see also @ **wife** ***housework*** *nU* see also @ **work**[1] ***council house*** *nC Brit* see also @ **council**			***housekeeper*** *nC* ***housekeeping*** *nU* see also @ **keep**[2]**/ keeper/keeping**	***household*** *adj* ***househusband*** *nC* see also @ **husband** ***keep house*** ***housemate*** *nC* see also @ **mate**[2] ***housing estate*** *nC Brit* see also @ **estate**[1] ***terraced house*** *nC Brit* see also **terrace** ***doghouse*** *nC Am* see also @ **dog**
house[2] building		***guest house*** *nC* see also @ **guest**			***greenhouse*** *nC* see also @ **green** ***warehouse*** *nC* see also @ **ware / wares**	***in-house*** *adj / adv* ***farmhouse*** *nC* see also @ **farm** ***lighthouse*** *nC* see also @ **light**[1] ***firehouse*** *nC esp Am* see also @ **fire**[1]	***house*** *nC* building ***house*** *vT* store sth ***house*** *nC* parliament ***boarding house*** *nC* see also @ **board**[4] ***coffee house*** *nC* see also @ **coffee** ***courthouse*** *nC Am* see also @ **court**[1] ***curry house*** *nC* see also @ **curry** ***opera house*** *nC* see also @ **opera** ***public house*** *nC* see also @ **pub, public** ***slaughterhouse*** *nC*

Figure 4.4.3: WFF results page.

post-modification of the collocation has also been taken into consideration whenever this emerged as an interesting pattern of use compared to other learners or native speakers.

Overall, the systemic approach to the analysis showed in Figure 4.4.1 will allow for a thorough and well-rounded investigation of the collocations and the learners from a qualitative perspective. In addition, the collocation use will be compared to the LOCNEC patterns in order to understand how close the ISLC learners' use is to the native speakers'. LOCNEC and ISLC are comparable corpora due to their high level of similarity; however, in some cases the collocations analysed in ISLC are either not present in LOCNEC or there are too few occurrences. In these cases the analysis will be limited to the observable instances in ISLC or the few ones in LOCNEC and conclusions will be drawn, if possible, also on the basis of potential L1 transfer effects from Italian.

Following on from this, two sets of collocations were chosen for the qualitative analysis based on their frequency and association measures scores. The first set of 11 items (and making up Chapter 6) included collocations with t-score ≥ 2 and MI ≥ 3 and a frequency of 5 in the ISLC corpus:

- Completely different
- Very different
- Really good

- Very good
- Very important
- Really interesting
- Very interesting
- Really nice
- Very nice
- Totally different
- Very strange

The frequency parameter was chosen in order to evaluate the most frequent collocations in ISLC as these were thought to be sources of interesting data since they were used frequently and by a higher number of speakers. In addition, these collocations are a good sample of the learners' pattern of use compared to the native speakers since they represent similar use, overuse, and underuse. Indeed, the collocations were tested with a log-likelihood test[10] which indicated that five collocations were produced in similar amounts both in the learner and native-speaker corpora (*really interesting, very different, very interesting, very nice, very strange*), three were overused in the ISLC corpus (*completely different, totally different, very important*), and three were underused in the ISLC corpus (*really good, really nice, very good*) compared to LOCNEC[11].

The second set of 9 items (which make up Chapter 7) belonged to the category of infrequent/unattested collocations (see Section 4.4.1), that is, all the adverb + adjective combinations which did not occur (or occurred less than five times) in the BNC and for which no association measures were computed:

- Absolutely scared
- Really really amazing
- Kind of scared
- Really really really good

[10]The test was performed with the UCREL LL Wizard at http://ucrel.lancs.ac.uk/llwizard.html

[11]*P* values for overused collocations: $p < 0.01$ ($LL = 9.42$); $p < 0.05$ ($LL = 5.23$); $p < 0.01$ ($LL = 7.89$); $p < 0.001$ (11.61). *P* values for underused collocations: $p < 0.0001$ ($LL = 25.24$); $p < 0.05$ ($LL = 5.25$); $p < 0.01$ ($LL = 6.90$).

- Pretty curious
- Super strange
- Quite old-fashioned
- Very fun
- Really heart-breaking

These combinations were chosen because they are the most frequent ones in the ISLC corpus, occurring a number of times equal or greater than two. The frequency limit was set to a lower value compared to the first collocation set because these combinations do not occur more than 3 times in the ISLC corpus. The choice of analysing infrequent/unattested collocations was driven by the interest that these combinations may hold in terms of potential L1 transfer effects. These combinations are not present in LOCNEC, therefore the qualitative analysis may be limited to verifying whether there are any overt signs of L1 transfer and any other interesting patterns emerging from ISLC.

4.5 Summary

This Chapter began by presenting the research questions of the present work, which aim to tap into both the quantitative and qualitative use of collocations (and non-collocations) produced by learners. In order to address and investigate the research questions, it was necessary to compile a spoken learner corpus of advanced Italian learners of English L2. Since the chosen methodology for the analysis of the corpus is CIA, the corpus needed to be comparable to a reference corpus, which in this case was LOCNEC (sister corpus of LINDSEI), and contain natural learner language. Thus, the corpus was modelled after LINDSEI, although a few changes were operated as regards the tasks of the informal interview and the transcription scheme. The choice of changing the third task from a picture description to a story-telling task was dictated by the poor productivity of the picture description in terms of number of tokens, and by the lack of guidance for interviewer's input. Furthermore, the transcription scheme was simplified in order to quicken the transcription process and focus on the elements relevant to the analysis. Section 4.2.3 provided information on the corpus, its data, and the learners and their metadata, such as gender, age, university course, languages studied, and stay-abroad experiences. The following Sections focused on the rationale of the combined analysis approach to the collocations, quantitatively and qualitatively, and provided a description of how the analyses were performed. In particular, the

quantitative Section (4.4.1) described the three different sub-analyses carried out in order to address research question one and its three sub-questions, and offered a detailed account of the statistical tests used to compare the number of collocations and their strength of association. The qualitative Section (4.4.2) explained the reason why a qualitative approach should be used in combination with a quantitative approach in order to obtain ideal results and insights. Furthermore, it detailed the three-layered scheme adopted in the analysis of the collocations, the steps taken in the comparison between ISLC and LOCNEC, and the choice of collocations and infrequent/unattested ones to be investigated.

Chapter 5

Quantitative analysis

THIS chapter describes the quantitative analysis of the data collected in this study. First, the adverb + adjective combinations were extracted from the corpus via AntConc (Anthony, 2019). The concordances were sorted and filtered according to the parameters set out in Section 4.3.3, and then combinations were analysed. The first analysis compares the raw frequencies of the combinations between learners and native speakers; the second analysis comprises two sets of sub-analyses on infrequent/unattested and grey area collocations. These can provide insights into the learners' behaviour when it comes to infrequent or non-collocational combinations. Lastly, the third analysis, which adopted Ferraresi's (2019) method of inquiry into collocationality, compares the mean *t*-score and MI values of texts in ISLC and LOCNEC.

5.1 The extraction of the adverb + adjective combinations

The combinations constituted by adjectives preceded by adverbs were extracted from the ISLC corpus via AntConc. AntConc (Anthony, 2019) is a free concordancer software developed by Laurence Anthony and it is widely used for the retrieval of concordances, collocations, and keyword lists. The corpus was previously POS-tagged with CLAWS tagger (Garside & Smith, 1997 – see Section 4.2.4), and the extraction of the combinations was based on the CLAWS5 tagset. The first step consisted in the identification of all the adjectives (tag = *AJ0*) in the corpus by means of a simple tag search in AntConc. Then, manual sorting of the adjective concordances was carried out to highlight all the extracted adjectives preceded by adverbs (tag = *AV0*). This was achieved with the sorting option of AntConc which enables the user to visualise all the concordance lines in alphabetical order according to the desired selection. In the case of ISLC, the sorting option was set to two left, that is, the second word to the left of the node. The sorting option was set to two left

in order to skip the word immediately preceding the tag *AJ0* which is the adjective itself tagged by CLAWS5 (see Figure 5.1.1 and Section).

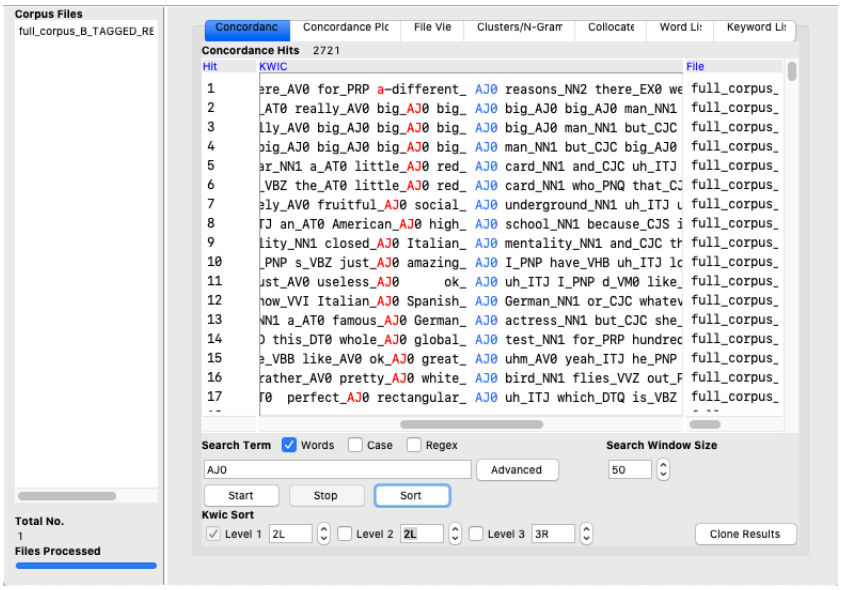

Figure 5.1.1: AntConc's sorting option set to - 2 (2L - bottom left) highlighting adjectives (*AJ0*) preceded by adverbs (*AV0*).

This retrieved all adjectives preceded by adjacent adverbs. A second extraction was performed to identify all those adverbs which did not directly precede the adjectives, as other parts of speech (such as filled pauses or disfluencies) occurred in between. This was done by increasingly moving the sorting filter to the left of the node until all instances were identified[1] (a maximum of five left was used to identify all concordances). The extraction performed in this way yielded 981 combinations; however, not all combinations were included in the analysis. Indeed, this initial set of combinations needed to follow the parameters of the syntactic pattern set out in Section 4.3.3. Thus, all the combinations which included an adverb functioning as an adverbial plus an adjective were excluded, as well as adverbs occurring before an adjective without direct modification of the adjective. Degree adverbs used to modify an adjective in the comparative or superlative form were also eliminated. On the other hand, cases which presented an adverb followed by *like* before the adjective were included in the analysis as *like* has been treated like a filler/speech disfluency the same as *uh*, or *uhm*. Adverbs of manner were also included in the extraction only if they took on the role of degree adverbs used as modifiers. In summary, the analysis considered degree adverbs and adverbs of manner acting as modifiers of adjectives, occurring either immediately after or immediately after up to three interjections/fillers/speech disfluencies. Therefore, time

[1] See Section 4.3.2.

and place adverbs, adverbials, and degree adverbs used in combination with an adjective in comparative or superlative forms were excluded. The hedging adverbs *kind of*, *sort of* and their slang forms (*kinda*, *sorta*) were included in the extraction.

After manually inspecting the 981 combinations identified in AntConc and discarding 285 of them, the remaining 696 adverb + adjective pairs were ranked according to their frequency. However, frequency of occurrence was not the only variable taken into consideration; indeed, additional measures were adopted in order to analyse the significance of collocations. Some of these measures of attraction between collocates, called lexical association measures, assign a score to the word pair thus allowing the researcher to either rank them on the basis of a score threshold (threshold approach) or classify them on the basis of their association strength (ranking approach) (Evert, 2009). As described in Section 3.5.2, there are several measures of strength of association that can be used in the frequency-based approach to identify and measure collocational strength according to how likely one word is to co-occur with the other.

In the present work, the strength of association of all 696 combinations occurring in the corpus at least once was calculated using two of the most widespread association measures in collocation studies (Durrant, 2008; Durrant and Schmitt, 2009; Evert, 2005, 2009; Gries, 2010; Hunston, 2002; Siyanova and Schmitt, 2008), MI and *t*-score. These two association measures are based on expected frequency (and random chance as regards *t*-score) and they tend to highlight different types of collocations. Indeed, MI tends to inflate the scores for word pairs which are less frequent but whose components often co-occur together and are rarely found on their own (Durrant & Schmitt, 2009). On the other hand, *t*-score highlights frequent combinations and it has been shown to strongly correlate with raw frequency counts (Gablasova et al., 2017), although it tends to lower the scores for combinations whose components co-occur with other lexical items. As regards these association measures, Stubbs (1995) and Hunston (2002) proposed that a *t*-score ≥ 2 and MI ≥ 3 can be taken as arbitrary cut-off points for the identification of collocations. This means that a combination which obtains a *t*-score equal or greater than 2 can be considered a collocation and, similarly, a combination with an MI score equal or greater than 3 can be considered a collocation (provided that the frequency threshold is set to a minimum of five). In this research and following Durrant and Schmitt's (2009) work, they are used in combination to single out potentially interesting collocations (or sift out irrelevant ones) extracted from a corpus. The association measures scores were calculated, as in Durrant and Schmitt (2009), on the basis of a large reference corpus, which in this case was the BNC. Occurrences of all 696 combinations extracted from ISLC were searched for in the BNC and their minimum frequency was set to at least five (this means that for combinations which occur in the BNC fewer than

five times no association measures scores were calculated). The association measure score for each of these was then calculated.

5.2 Frequency analysis of combinations: learners vs native speakers

In this Section, the combinations extracted from the ISLC corpus are compared to the combinations extracted from the native speaker corpus LOCNEC, the reference corpus for CIA. LOCNEC is the sister corpus of LINDSEI, the corpus on which the ISLC is based, and this guarantees that the level of comparability between ISLC and LOCNEC in terms of design criteria and genre is high. In addition, the native speakers are university students aged between 18 and 30, a similar background to the ISLC participants: university students aged between 20 and 28. This is a key point since, by ensuring that the two corpora and the speakers' backgrounds are as similar as possible, any differences between the two populations is more easily attributable to one variable only (native vs non-native speaker status). If more than one variable were involved in the comparison (such as different ages, different backgrounds, e.g., university students vs employed adults), identification of the possible cause of discrepancies would be impeded.

Table 5.1: Number of tokens in ISLC and LOCNEC corpora.

	ISLC	LOCNEC
Corpus size (learners' turns only)	58,568	117,417

The results of the extraction of all the combinations adverb + adjective from the two corpora are reported for combination types (Table 5.2) and tokens (Table 5.3), that is, for the number of different combinations and the number of occurrences of the combinations.

Table 5.2: Adverb + adjective combination types in ISLC and LOCNEC.

	ISLC	LOCNEC
Mean n. combinations per speaker	18.21 (*sd* = 7.02)	22.32 (*sd* = 8.30)
Normalised freq. per 1,000 words	7.48	5.82
Raw frequency	438	683

Table 5.2 shows that the native speakers produce on average a higher number of combi-

nation types, 22.32 (*sd* = 8.30) compared to the mean number of combination types of the learners which is 18.21 (*sd* = 7.02). Both sets of data are normally distributed as indicated by the Shapiro-Wilk test[2].

Table 5.3: Adverb + adjective combination tokens in ISLC and LOCNEC.

	ISLC	LOCNEC
Mean n. combinations per speaker	20.47 (*sd* = 8.32)	26.98 (*sd* = 10.95)
Normalised freq. per 1,000 words	11.88	11.49
Raw frequency	696	1,349

Table 5.3 shows that the mean number of combination tokens in ISLC is 20.47 (*sd* = 8.32), while the mean number of combination tokens in LOCNEC is 26.98 (*sd* = 10.95). Both sets of data are normally distributed as indicated by the Shapiro-Wilk test[3]. Figure 5.2.1 plots the distribution of combination types produced by individual speakers in ISLC and LOCNEC.

Boxplots represent the median values (the horizontal thick black line) and the distribution of the central data around them (the lower line of the box represents the first quartile, whereas the upper line is the third quartile). In Figure 5.2.1, the boxplot highlights an outlier in LOCNEC, which is speaker EN048 with 43 combination types. As regards the learners, the graph shows that the 1st and 3rd quartiles of combination types per speaker are between 12 and 25, as opposed to LOCNEC where the 1st and 3rd quartiles of combination types per speaker are found between 15 and 26. Visually, the difference between these two medians does not seem overly marked, but it is necessary to test this difference by means of a statistical test.

The choice of statistical test must take into consideration individual variation since, especially as regards second language learners, this is an aspect that is seldom computed (Callies, 2015). In order to account for individual variation, an independent two-tailed *t*-test was performed to test the difference between mean number of combination types produced by learners vs native speakers. The figures of mean number of combination types (and tokens) per speaker displayed in Table 5.2 (and plotted in Figure 5.2.1) met all the assumptions of the *t*-test (see footnote 2) so that it was possible to carry out the test and quantify whether there are significant differences between the mean number of combination types produced

[2]The combination types in ISLC are normally distributed ($W = 0.9519$, $p = 0.1399$) and the combination types in LOCNEC are also normally distributed ($W = 0.97917$, $p = 0.5173$).

[3]The combination tokens in ISLC are normally distributed ($W = 0.96289$, $p = 0.2945$) and the combination tokens in LOCNEC are also normally distributed ($W = 0.96953$, $p = 0.2318$).

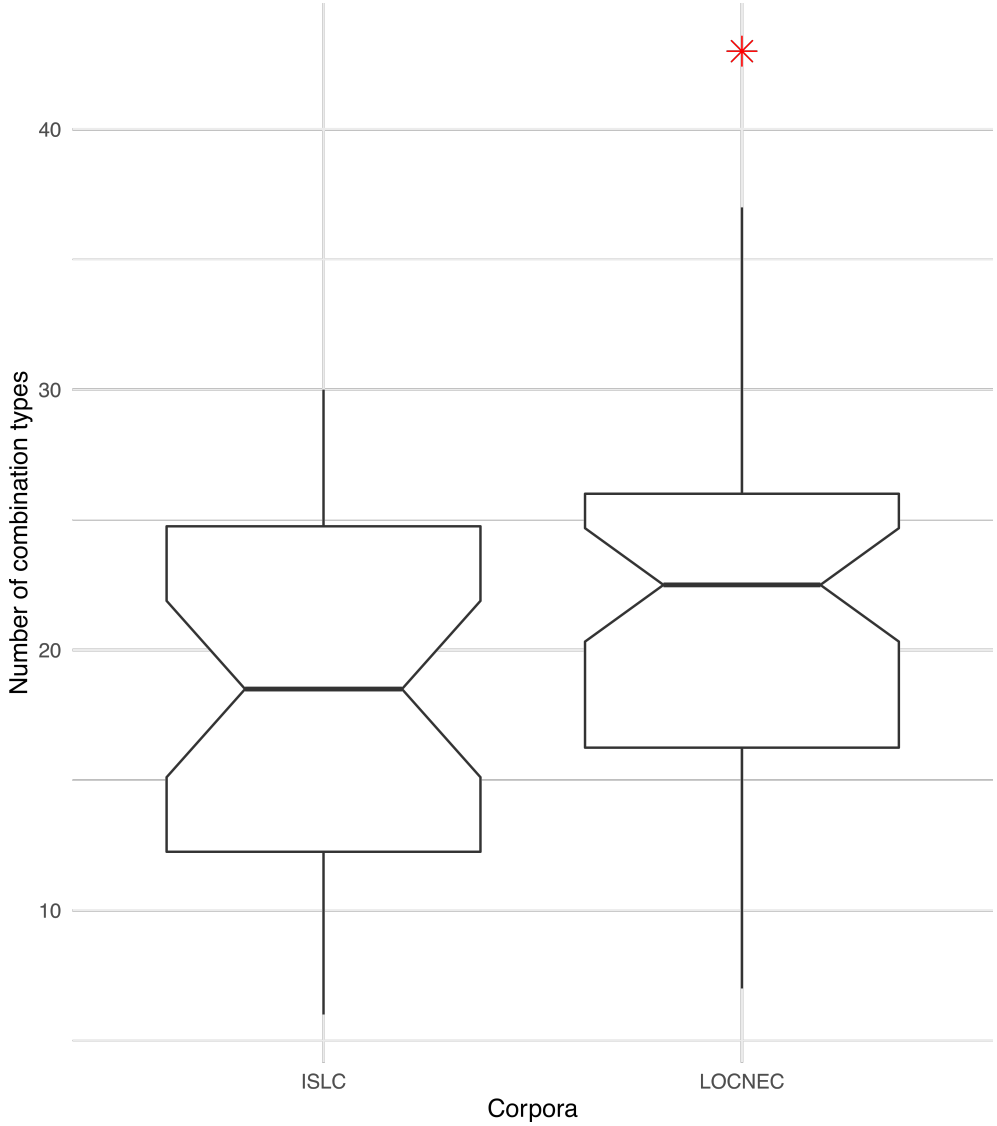

Figure 5.2.1: Boxplot illustrating the distribution of combination types produced by speakers in ISLC and LOCNEC.

by the learners and native speakers. The *t*-test indicates that there is a statistically significant difference between the two groups with a medium effect size (t = - 2.45, df = 78.03, p = 0.02, r = 0.27), thus suggesting that the native speakers produce, on average, a higher number of combination types. This already anticipates the learners' tendency to use fewer adverb + adjective collocations.

Figure 5.2.2 plots the distribution of combination tokens produced by individual speakers in ISLC and LOCNEC.

The boxplot in Figure 5.2.2 reveals that in LOCNEC there are two outliers, which have been identified as speakers EN015 and EN048.

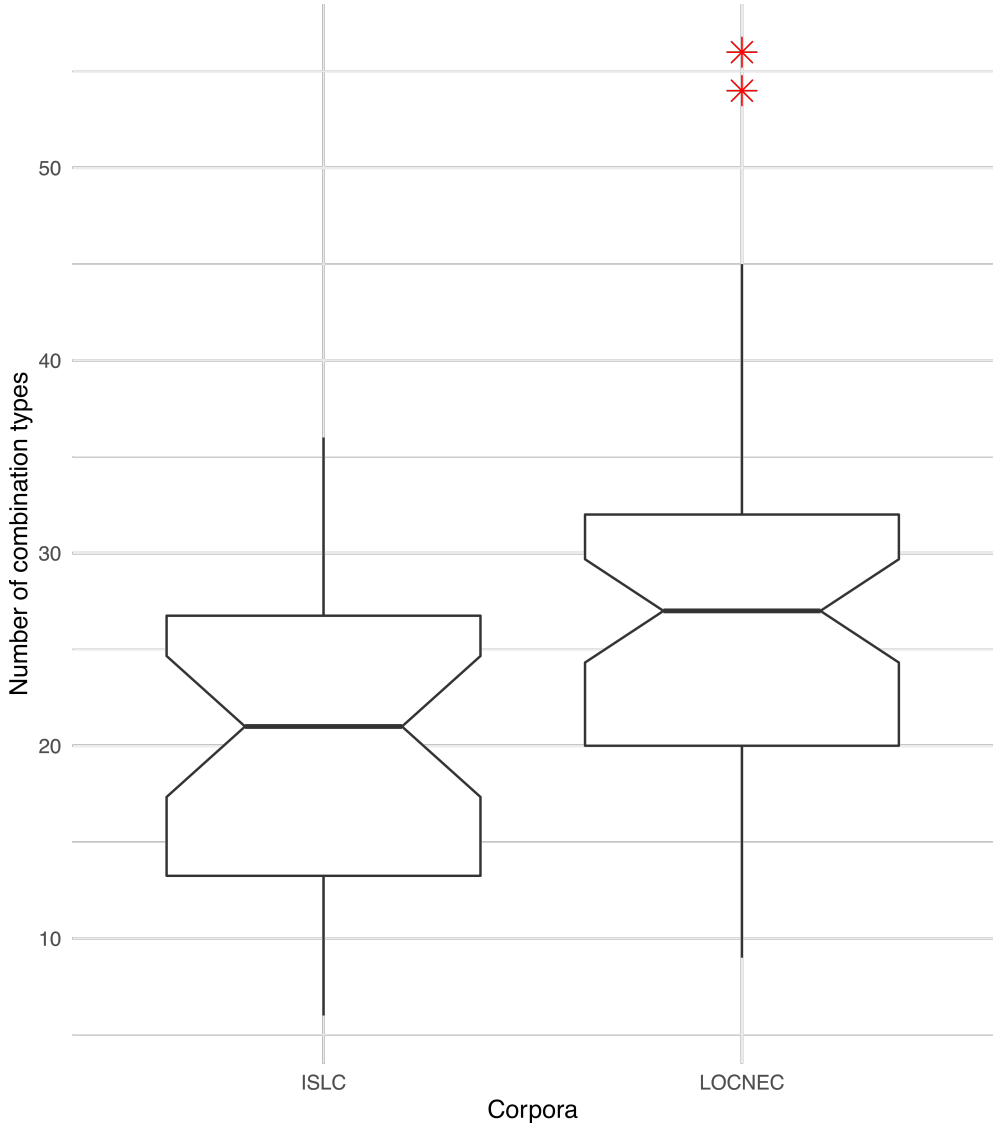

Figure 5.2.2: Boxplot illustrating the distribution of combination tokens produced by speakers in ISLC and LOCNEC.

These produce 54 and 56 combination tokens respectively. As regards the learners, the graph shows that the 1st and 3rd quartiles of combination tokens per speaker are distributed between 14 and 26, as opposed to LOCNEC where the 1st and 3rd quartiles are found between 20 and 32. The mathematical means of the two groups seem to differ, as the learners produce on average 20 combination tokens compared to the native speakers' 27 (see Table 5.3). As with the combination types, the difference was tested for significance with a two-tailed independent *t*-test since all assumptions were met (see footnote 3). The *t*-test assessed whether the mean number of combination tokens produced by LOCNEC speakers is indeed greater than that of the learners. The result confirm that the difference between the mean

number of combinations per speaker is also statistically significant with a medium effect size (t = - 3.22, df = 80.17, p = 0.001, r = 0.34), so that it is possible to state that speakers of the ISLC corpus produce, on average, a lower number of combination tokens compared to the native speakers of LOCNEC.

The analysis so far has examined the overall number of combinations extracted from the two corpora. The comparison has focused on the number of combination types and tokens produced by learners and native speakers. However, in the present study the main focus is on collocations as defined by the frequency-based approach (see Section 3.4.1.2), so that it is now necessary to investigate, among these combinations, the actual collocations and the non-collocations. As described in Section 4.4.1, the chosen approach to the identification of collocations is that of Durrant and Schmitt (2009) and Ferraresi (2019). In the following Section (5.2.1), the collocations are identified on the basis of association measure scores, in particular t-score and MI, following Durrant and Schmitt (2009). These are computed on the basis of a large reference corpus, the BNC, with a minimum frequency set to five occurrences. The collocations thus identified are divided between infrequent/unattested, grey area, and actual collocations. The first two categories are then analysed from the point of view of individual percentage production over the total number of combinations, while the third category is examined from a collocational strength perspective on the basis of Ferraresi's (2019) work which compared their mean t-score and MI values per text.

5.2.1 Frequency analysis of non-collocations: learners vs native speakers

In this Section, two analyses will be conducted on the non-collocations extracted from ISLC and LOCNEC. The first analysis focuses on infrequent/unattested collocations which are compared in terms of percentage between individual ISLC and LOCNEC speakers. The second analysis considers the grey area collocations, that is those collocations whose association measures scores are below the cut-off points[4]. This set of collocations is compared between learners and native speakers in terms of percentage.

5.2.1.1 Infrequent/unattested collocations

The present analysis of the collocations according to their association measures consists in examining the infrequent or unattested collocations. These collocations are either not present in the BNC (unattested) or occur fewer than five times, (infrequent), thus no t-

[4]For more on infrequent/unattested and grey area collocations, see Section 4.3.4.

score or MI were calculated (see Section 4.3.4). Examples of collocations included in this category are *absolutely scared*, *quite freezing*, *really creamy*, *totally impressed*, *pretty curious,* and others. Some of these combinations are plausible, such as *pretty curious* or *really creamy*, but they may either be chiefly American English or simply not frequent enough in the corpus.

Table 5.4: Infrequent/unattested collocation types in ISLC and LOCNEC.

	ISLC	LOCNEC
Median % of coll. per speaker	9.81	4.82
Interquartile range	10.44	10.83

Table 5.4 displays descriptive information regarding the infrequent/unattested collocation types. The median and the interquartile range are reported as the data is not normally distributed[5]. Individual variation was taken into account as regards this category of collocations too, so that the median percentage of infrequent/unattested collocation types was calculated by identifying, for each speaker, the proportion between the number of infrequent/unattested collocations produced over the total number of combinations. As far as the non-native speakers are concerned, the median percentage of infrequent/unattested collocations produced per speaker is 9.81%, while for native speakers this figure is 4.82% (the data is also visually represented in Figure 5.2.3).

This means that, on average almost 10% of the combinations produced by a single speaker in ISLC either does not occur in the BNC as a collocation, or it is too infrequent (< 5 times) to be categorised as a collocation according to the collocation parameters adopted in this study and, e.g., in Durrant and Schmitt (2009). On the other hand, only 5% of the combinations produced by the native speakers fall into this category. This difference was tested with the non-parametric Wilcoxon rank sum test, since the assumptions of the *t*-test were not met by the data: both sets of data are not normally distributed. The test was computed and the results confirm that indeed there is statistical difference between the groups with a medium effect size ($W = 10179$, $p = 0.002$, $r = -0.33$).

Table 5.5 instead displays descriptive information about infrequent/unattested collocation tokens. The median and the interquartile range are reported as the data is not normally distributed[6].

[5]ISLC percentage data is normally distributed ($W = 0.94$, $p = 0.06$), while LOCNEC is non-normally distributed ($W = 0.88$, $p = < 0.001$).

[6]ISLC data is non-normally distributed ($W = 0.93$, $p = 0.05$). LOCNEC is also non-normally distributed ($W = 0.84$, $p = < 0.0001$).)

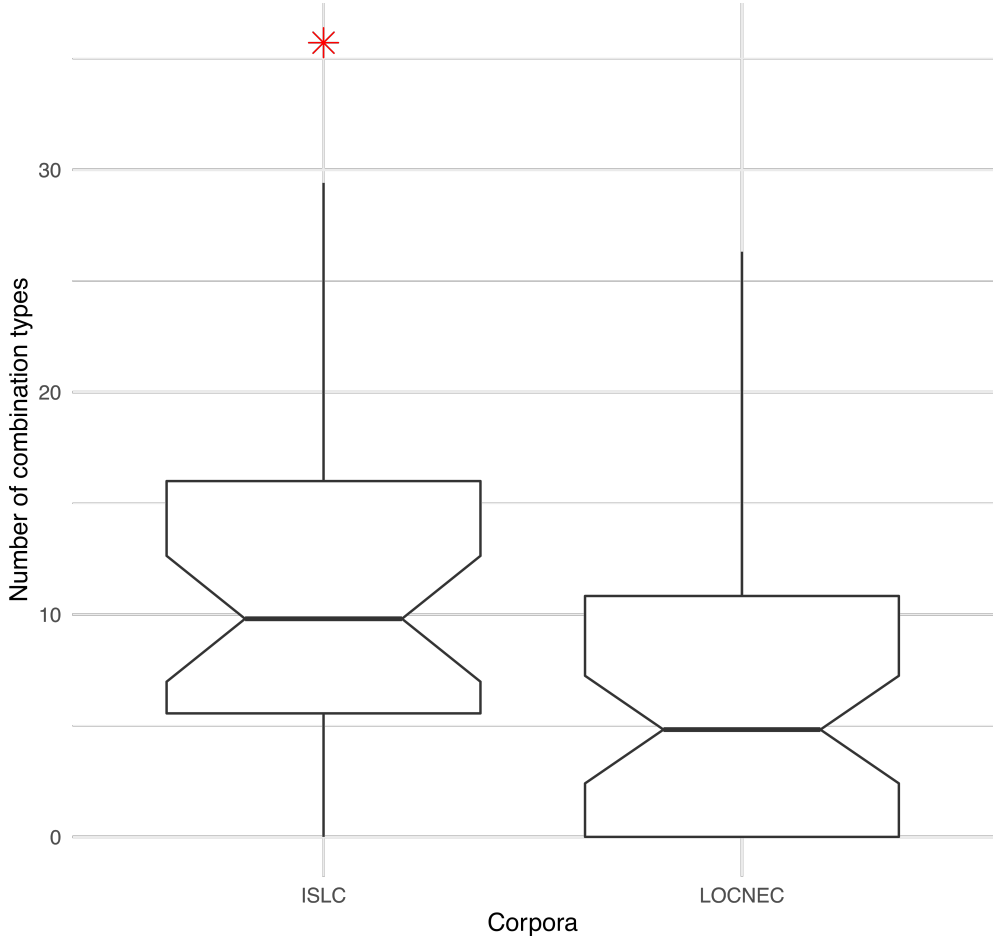

Figure 5.2.3: Boxplot illustrating percentage of infrequent/unattested collocation types produced by ISLC and LOCNEC speakers.

Table 5.5: Infrequent/unattested collocation tokens in ISLC and LOCNEC.

	ISLC	LOCNEC
Median % of coll. per speaker	9.31	4.26
Interquartile range	10.93	4.68

As done with the infrequent/unattested collocation types, individual variation was also taken into account as regards the tokens, so that the median percentage of infrequent/unattested collocation tokens was calculated by identifying, for each speaker, the proportion of infrequent/unattested collocations produced over the total number of combinations. As far as the non-native speakers are concerned, the median percentage of infrequent/unattested collocations produced per speaker is 9.31%, while the figure for the native speakers is 4.26%. These

percentages coincide with the infrequent/unattested collocation types (see Figure 5.2.3; the data was also visually represented in Figure 5.2.4).

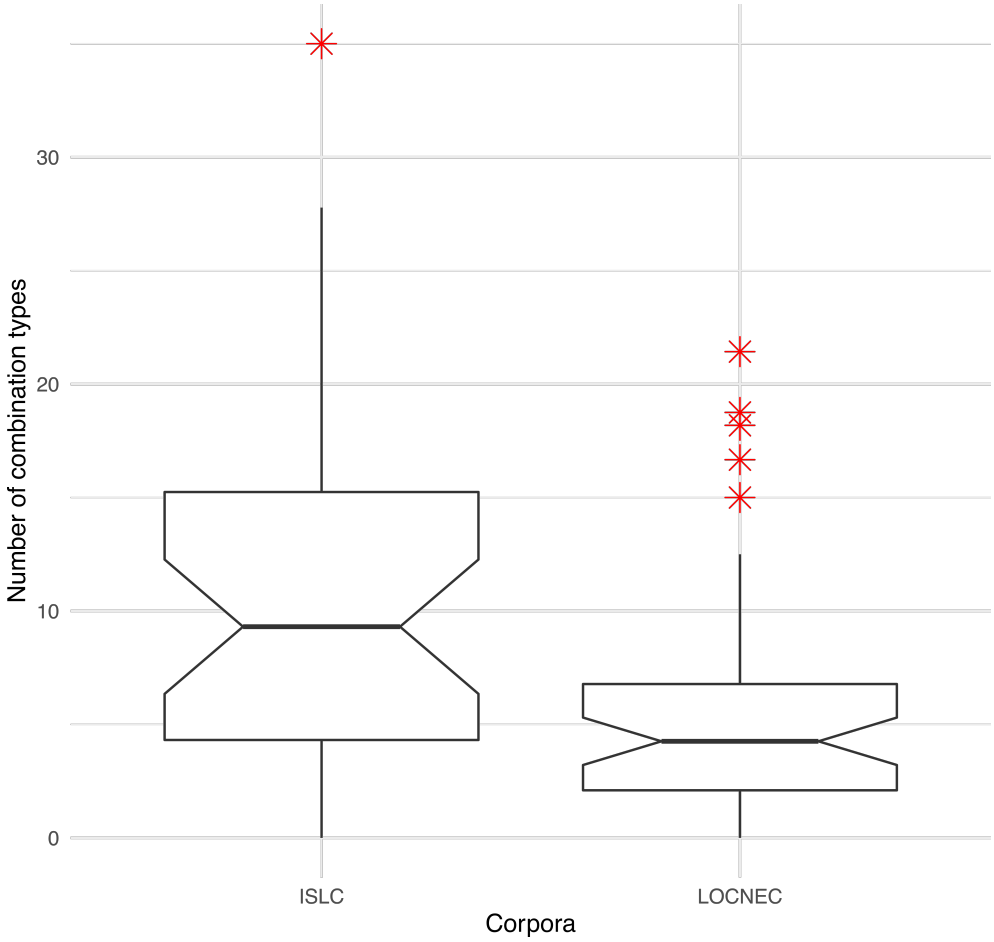

Figure 5.2.4: Boxplot illustrating percentage of infrequent/unattested collocation tokens produced by ISLC and LOCNEC speakers.

This means that, similarly to the infrequent/unattested collocation types, on average almost 10% of the combinations produced by a single speaker in ISLC either does not occur in the BNC as a collocation, or it is too infrequent (< 5 times) to be categorised as a collocation according to the collocation parameters adopted in this study. On the other hand, just over 4% of the combinations produced by the native speakers fall into this category. This difference was tested with the non-parametric Wilcoxon rank sum test, since the assumptions of the *t*-test were not met by the data: both sets of data are not normally distributed (see footnote 6). The test was computed and the results confirm that there is statistical difference between the groups with a medium effect size ($W = 1198.5$, $p = 0.001$, $r = -0.35$).

In conclusion, this Section has analysed the infrequent/unattested collocation types and tokens extracted from ISLC and LOCNEC. The visual comparisons (Table 5.4 and Table

5.5) show that the learners seem to produce a higher percentage of infrequent/unattested collocation types and tokens. A closer inspection of the data with a statistical test to verify the significance of the difference has revealed that, as regards the collocation types, the ISLC speakers indeed produce a higher percentage of infrequent/unattested collocation types with a medium effect size. Similarly, the Wilcoxon rank sum test on the infrequent/unattested collocation tokens has highlighted that there is significant difference with a medium effect size between the native and non-native speakers.

5.2.1.2 Grey area collocations

The analysis of the grey area collocations consists in investigating the percentage of collocations with a *t*-score lower than 2 and a MI score lower than 3. This includes all combinations that occur in the BNC at least five times and for which it was possible to calculate the association measures scores, but which did not rank as collocations (according to this study) due their scores being below the cut-off points. Since the association measures scores for all combinations was based on the BNC which is a corpus of British English, it may be that some of these combinations did not classify as collocations due to their chiefly American English use. Combinations included in this category are *very fascinating*, *really curious*, *pretty difficult*, *kind of dark*, *quite terrible*, and others.

Table 5.6: Grey area collocation types in ISLC and LOCNEC.

	ISLC	LOCNEC
Mean % of coll. per speaker	41.32	30.97
Standard deviation	15.72	9.90

Table 5.6 displays descriptive information regarding the grey area collocation types. The mean and the standard deviation are reported as the data is normally distributed[7]. Individual variation was taken into account as regards this category of collocations too, so that the mean percentage of grey area collocation types was calculated by identifying, for each speaker, the proportion of grey area collocations produced over the total number of combinations. As far as the non-native speakers are concerned, the mean percentage of grey area collocations produced per speaker is 41.32%, while the figure for the native speakers is 30.97% (the data was also visually represented in Figure 5.2.5).

[7]ISLC data is normally distributed ($W = 0.97$, $p = 0.37$). LOCNEC is also normally distributed ($W = 0.98$, $p = 0.54$).

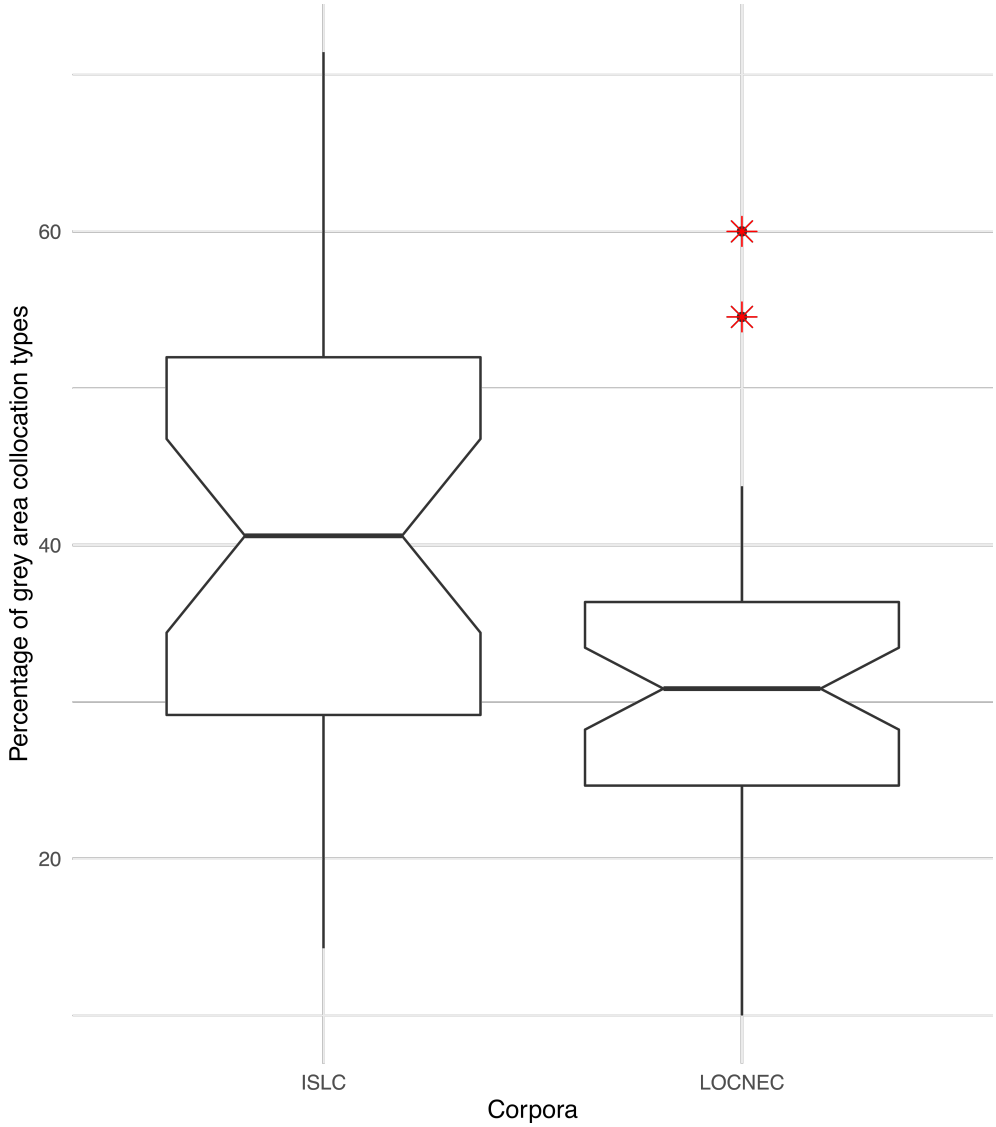

Figure 5.2.5: Boxplot illustrating percentage of grey area collocation types in ISLC and LOCNEC.

This means that, on average just over 40% of the combination types produced by a single speaker in ISLC has scored a *t*-score and an MI value lower than the cut-off points and, as such, cannot be regarded as a collocation according to the collocation parameters adopted in this study. On the other hand, about 30% of the combination types produced by the native speakers fall into this category. This difference was tested with an independent two-tailed *t*-test since the assumptions of the parametric test were met by the data: both sets of data are normally distributed (see footnote 7). The test was computed and the results indicate that there is statistical difference between the groups with a large effect size ($t = 3.4063$, $df = 50.725$, $p = 0.001$, $r = 0.43$).

Table 5.7: Grey area collocation tokens in ISLC and LOCNEC.

	ISLC	LOCNEC
Mean % of coll. per speaker	41.02	28.37
Standard deviation	17.24	9.74

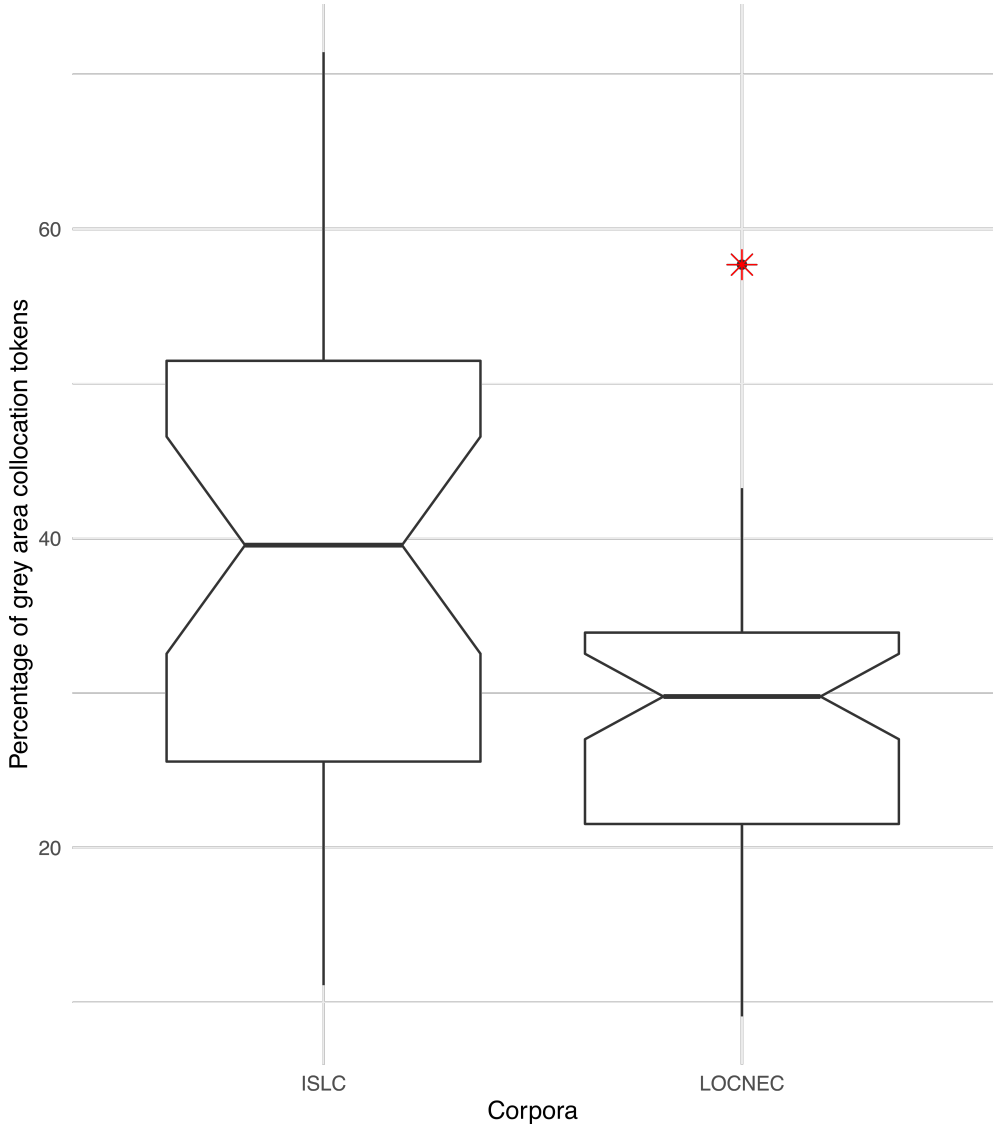

Figure 5.2.6: Boxplot illustrating percentage of grey area collocation tokens in ISLC and LOCNEC.

Table 5.7 instead displays descriptive information about grey area collocation tokens. The

mean and the standard deviation are reported as the data is normally distributed[8]. As done with the grey area collocation types, individual variation was also taken into account as regards the tokens, so that the mean percentage of grey area collocation tokens was calculated by identifying, for each speaker, the proportion of grey area collocations produced over the total number of combinations. As regards the non-native speakers, the mean percentage of grey area collocations produced per speaker is 41.02%, while the figure for the native speakers is 28.37%. These percentages coincide with the grey area collocation types (cf. Figure 5.2.5 – the data was also visually represented in Figure 5.2.6).

This means that, on average, just over 40% of the combinations produced by a single speaker in ISLC has scored a *t*-score and an MI value lower than the cut-off points and, as such, cannot be regarded as a collocation according to the collocation parameters adopted in this study. On the other hand, less than 30% of the combinations produced by the native speakers fall into this category. This difference was tested with an independent two-tailed *t*-test since the assumptions of the parametric test were met by the data: both sets of data are normally distributed (see footnote 8). The test was computed and the results indicate that there is statistical difference between the groups with a large effect size ($t = 3.8782$, $df = 47.38$, $p = 0.003$ $r = 0.49$).

In conclusion, the visual representations of the grey area collocation types and tokens indicated that there is difference between the mean percentage of this set of collocations produced by the learners and native speakers. An independent two-tailed *t*-test performed on both sets of data (types and tokens) verified that the difference is indeed statistically significant with a large effect size. Thus, learners produce a greater amount of grey area collocations compared to native speakers.

5.3 Analysis of t-score and MI values

This study has so far largely relied on Durrant and Schmitt's (2009) work on collocations, namely their identification as per the association measure scores. Thus, it could be expected that the present analysis also employs the authors' second-level cut-off points for weak and strong collocations (namely, 10 for the *t*-score and 7 for MI). However, this method was discarded in favour of a more text-centred approach as found in Ferraresi (2019, see Section 4.4.1). The analysis of the *t*-score and MI values for the collocations follows Ferraresi's method of calculating the mean *t*-score and the MI for each corpus text. Thus, two values

[8] ISLC data is normally distributed ($W = 0.96$, $p = 0.32$). LOCNEC is also normally distributed ($W = 0.98$, $p = 0.42$).

correspond to each text, a mean *t*-score and a mean MI calculated on the basis of all the collocations contained in every single text. This mean score provides an indication of the strength of the collocation based on the corpus itself and on the participants. The mean scores can then be compared between corpora both by visual inspection and by statistical testing. The *t*-score and MI values were calculated for all combinations and only the combinations with a score ≥ 2 and ≥ 3 respectively were taken into consideration. This was done so as to avoid including in this analysis the combinations previously investigated (infrequent/unattested and grey area collocations).

5.3.1 *T*-score

The mean *t*-score values were thus calculated for all collocation tokens with a *t*-score and MI values ≥ 2 and ≥ 3 respectively. A slight modification was performed as regards the frequency threshold, which was set to 5, rather than 3 as in Ferraresi (2019).

Figure 5.3.1 shows the distribution of mean *t*-score values for tokens in ISLC and LOCNEC and Table 5.8 displays the mean *t*-score values per text.

Table 5.8: Mean *t*-score values per text in ISLC and LOCNEC.

	ISLC	LOCNEC
Mean *t*-score value per text	13.9	16.6
Standard deviation	6.4	5.11

Table 5.8 shows that the mean *t*-score of ISLC is 13.9 (*sd* = 6.4), while that of LOCNEC is 16.6 (*sd* = 5.11). The width of the plot shown in Figure 5.3.1 represents the number of texts whose collocation values are situated on the *y* axis: the wider the plot, the more collocations display similar values. In the case of Figure 5.3.1, the LOCNEC texts display a greater degree of similar values between 10 and 20 compared to ISLC, whose shape is narrower between these values. In terms of median, despite the presence of an outlier in ISLC, LOCNEC displays a higher median value compared to ISLC and is overall "more collocational" (Ferraresi, 2019, p. 213).

In order to confirm the visual inspection, statistical testing was performed on the data after a Shapiro-Wilk test to verify distribution[9]. The *t*-test assessed whether there is significant difference between the mean scores of the two corpora and revealed that the mean

[9]Both the ISLC and LOCNEC data are normally distributed ($W = 0.96289$, $p > 0.05$ and $W = 0.97001$, $p > 0.05$ respectively).

collocation *t*-score in LOCNEC is statistically higher than ISLC with a medium size effect (t = - 2.0722, df = 60.238, p = 0.04254, r = 0.26).

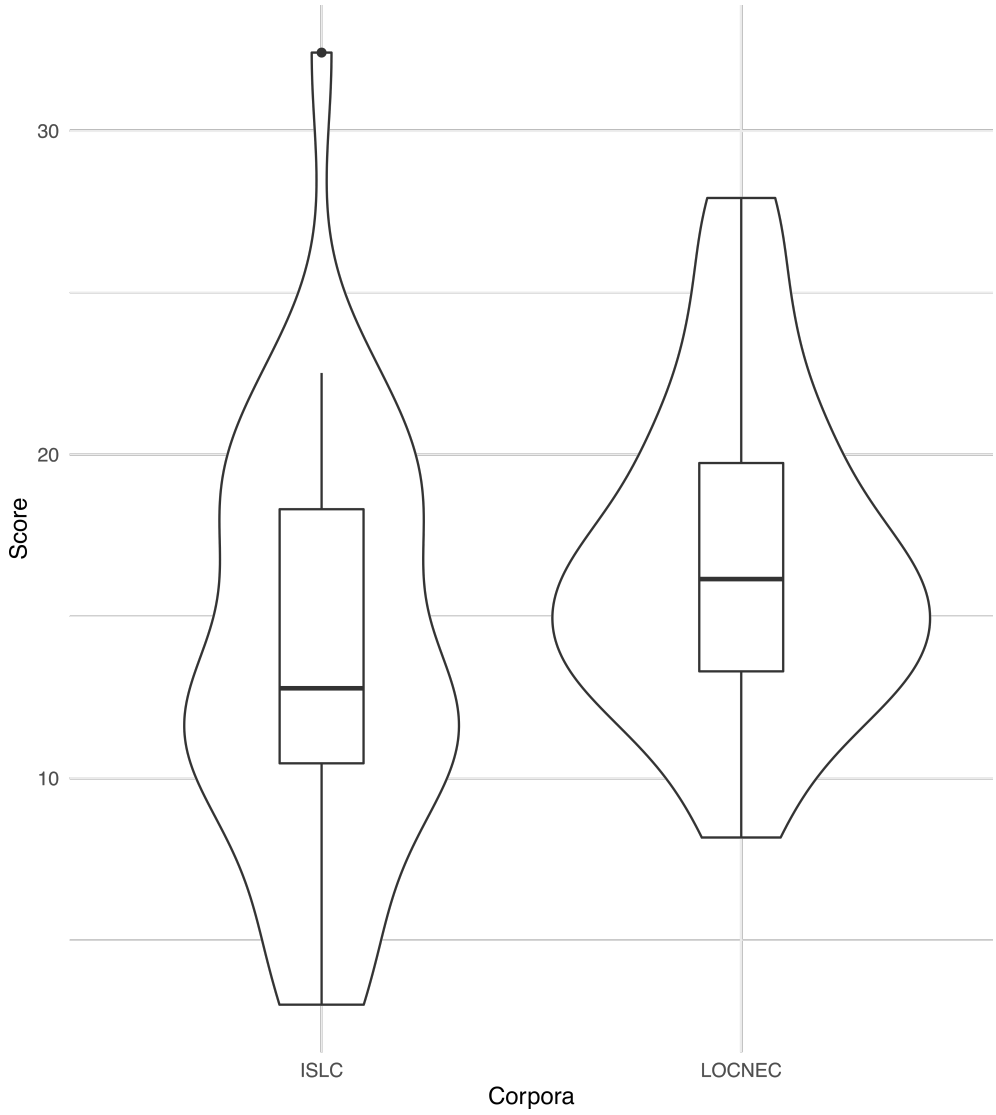

Figure 5.3.1: Mean t-score of collocations per text in ISLC and LOCNEC.

In order to increase the comparability with the literature suggesting that advanced learners are more sensitive to frequency (Ellis et al., 2008; 2015; among others) and thus use a larger number of high-frequency collocations (*t*-score collocations) (Bestgen & Granger, 2014; among others), a brief comparison of the percentages of *t*-score collocations was performed. The percentage of *t*-score collocations (tokens only) was computed for each speaker (both in ISLC and LOCNEC), over the total number of combinations. This resulted in a mean percentage of 64.83% of *t*-score collocations per speaker in ISLC (sd = 17.71) and a median percentage of 80.86% in LOCNEC (IQR = 9.96). Since one set of data was

not normally distributed (W = 0.94, p = 0.01 for LOCNEC; while W = 0.95, p = 0.11 for ISLC), a Wilcoxon rank sum test was run and the difference between the two groups resulted statistically significant with a large effect size (W = 389, $p < 0.0001$, r = - 0.46). Therefore, learners significantly underuse *t*-score collocations compared to native speakers. In conclusion, the native speakers produce, on average, higher *t*-score values per text compared to the learner.

5.3.2 Analysis of MI score

Similarly to the *t*-score, MI mean values were calculated for all collocation tokens with a *t*-score and MI values ≥ 2 and ≥ 3 respectively in each text in ISLC and LOCNEC. A slight modification was performed as regards the frequency threshold, which was set to 5, rather than 3 as in Ferraresi (2019). The data is displayed in Table 5.9 (median and interquartile range are shown since the data are not normally distributed[10]). The median MI value per text in ISLC is 4.72 compared to 5.03 for the native speakers.

Table 5.9: Median MI values per text in ISLC and LOCNEC.

	ISLC	LOCNEC
Median MI value per text	4.72	5.03
Interquartile range	0.46	0.41

In this case, the shape of the violin plots in Figure 5.3.2 are more alike than in the previous analysis (cf. Figure 5.3.1). The width of the plot has a similar shape in both corpora, showing that the majority of texts contain collocations whose MI values are situated between 4 and 6. However, in LOCNEC it is possible to observe that the plot is at its widest in correspondence with the *y* axis value of 5, indicating that the mean score of LOCNEC is slightly higher than ISLC. This is also confirmed by the median, represented by the thick black line in the boxplot within the violin plot. Indeed, in LOCNEC the line coincides with 5, whereas in ISLC the line is located slightly below it. In both cases, there are outliers: one in ISLC has a score above 6, whereas two in LOCNEC scored around 7 and 10.

In order to confirm the visual inspection, statistical testing was performed on the data. Since both sets of data are not normally distributed (see footnote 10), a Wilcoxon rank sum test was computed to assess whether there is significant difference between the mean scores

[10]Both the ISLC and LOCNEC data are not normally distributed (W = 0.90074, $p < 0.01$ and W = 0.59041, $p < 0.001$ respectively).

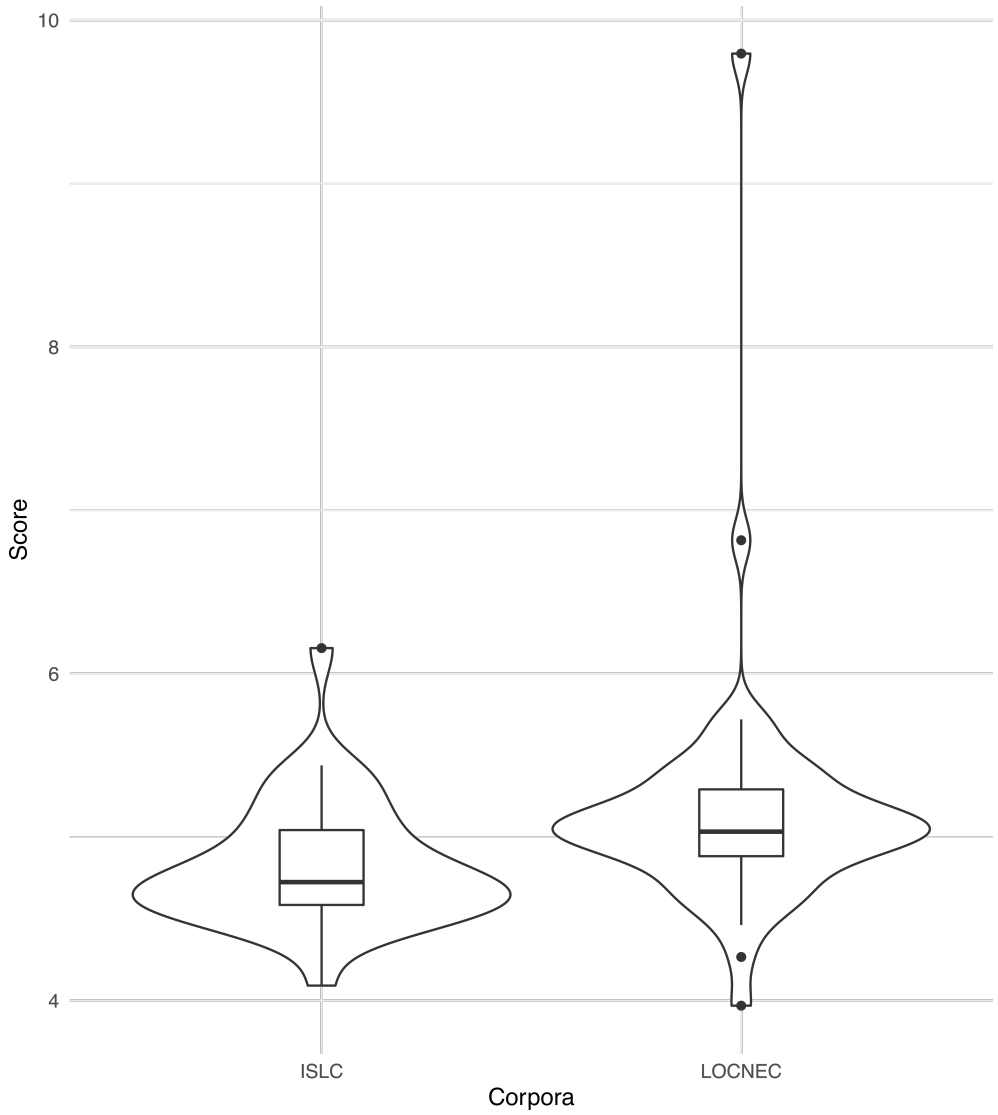

Figure 5.3.2: Mean MI score of collocations per text in ISLC and LOCNEC.

of the two corpora. The test revealed that the average collocation MI score in LOCNEC is statistically higher than ISLC with a medium effect size ($W = 491.5$, $p < 0.01$, $r = -0.35$).

In order to increase the comparability with the literature suggesting that even advanced learners struggle with infrequent and more strongly associated collocations (MI collocations) (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Ellis et al., 2008; Li & Schmitt, 2009), a brief comparison of the percentages of MI collocations was performed. The percentage of MI collocations (tokens only) was computed for each speaker (both in ISLC and LOCNEC), over the total number of combinations. An independent two-tailed *t*-test was

performed to test for significance since both sets of data were normally distributed (W = 0.97, p = 0.37 for ISLC; W = 0.97, p = 0.34 for LOCNEC). The test indicated that there is significant difference between ISLC and LOCNEC with a large effect size (t = - 5.6243, df = 49.16, $p < 0.0001$, r = 0.62). Therefore, learners significantly underuse MI collocations compared to native speakers. In conclusion, as regards mean MI values, native speakers produce collocations with a higher association measure score compared to learners.

5.4 Discussion

In this Chapter, we have presented a quantitative analysis of the data extracted from the ISLC corpus. Upon initial extraction, almost a thousand combinations were identified in ISLC and, upon manual sorting and selection, a total of 696 adverb + adjective combination tokens and 438 types were taken into consideration for the analysis. A similar extraction was performed on LOCNEC and led to 1,349 combination tokens and 683 types. The data was first compared in terms of mean number of combinations produced by speakers in ISLC and LOCNEC. An independent two-tailed t-test indicated that the difference between these two groups as regards the mean number of combination types produced by the speaker is statistically significant. The same test was run on the mean number of combination tokens produced by both groups of speakers and this also highlighted a significant difference. This initial analysis already anticipated that the learners produce a lower number of collocations compared to the native speakers. Indeed, if the mean number of combinations is higher in LOCNEC, it is plausible to expect that the native speakers also produce a greater mean number of collocations.

The second analysis, which comprised a sub-set of two different analyses, set out to verify the quantitative behaviour of learners vs native speakers as regards infrequent/unattested and grey area collocations. In the present work, collocations have been identified according to the frequency-based approach (see Section 3.4.1.2) and following Durrant and Schmitt's (2009) work. The first sub-analysis concerned the infrequent/unattested collocations, that is, those collocations that occur fewer than five times in the BNC or do not occur at all. This set of collocations was also divided between types and tokens and the median percentage of collocations produced per speaker was compared in the two corpora. The statistical test of significance on types indicated that the learners make a larger use of infrequent/unattested collocation types compared to the native speakers. The same test was run on the infrequent/unattested collocation tokens and significant difference was also found. The findings are not unexpected: as it has already been described in the literature (e.g., Laufer & Wald-

man, 2011; Nesselhauf, 2005), even advanced learners display signs of poor collocational production and competence.

The second sub-analysis took into consideration grey area collocations, in other words those collocations which occur more than five times in the BNC, but whose association measures scores are below the cut-off points. These were also divided between types and tokens and the mean percentage of collocations per speaker was compared between the two corpora. An independent two-tailed *t*-test on types indicated that there is significant difference between the mean percentage of grey area collocations produced by individual speakers in ISLC and LOCNEC. Although in a recent study by Leńko-Szymańska (2020) on performance-based assessment and the correlation between indices of lexical complexity and proficiency both the grey area and infrequent collocations were included in the analysis, in the parallel study of Durrant and Schmitt (2009) on which the present work is based, this type of collocations was not part of the analysis (nor were these considered in another similar study by Granger & Bestgen, 2014): only strong collocations and low frequency combinations were investigated. However, this finding adds another piece to the puzzle of collocational behaviour of language learners in spontaneous speech. Indeed, if previous research had already verified that even advanced learners of English produce a significantly greater amount of low-frequency collocations compared to native speakers, the present work informs of another behaviour, namely that non-native speakers also produce a greater number of grey area collocations. Thus, it is not only a matter of infrequent collocations (which the literature has already shown to be produced in lower numbers precisely due to their infrequency), but also of other frequent combinations whose collocationality can depend on the choice of reference corpus (and thus variety) or on the creativity of the learners. However, this finding may be strictly inherent to the speech mode, as Leńko-Szymańska's (2020) work on bigrams, which included the infrequent and grey area collocations in the analysis, found opposite results: her learners produced a similar amount of infrequent collocations and non-collocational bigrams to the native-speaking counterparts. This could indicate that perhaps her sample of learners possessed a higher level of proficiency, or that the production of collocations in spoken language is a distinct learner ability which may lag behind the written one, as demonstrated by the higher number of infrequent and grey area collocations.

The final analysis that was conducted on the extracted collocations focused on the mean *t*-score and MI values. This analysis followed Ferraresi's (2019) work on constrained varieties for which the author compared the mean *t*-score and MI values per text across the corpora. This approach has a few advantages (see Section 4.4.1) over the method employed by Durrant and Schmitt (2009) of dividing between strong and weak collocations according to a *t*-score and MI value based on their corpora. The scores calculated per text and

the mean obtained at the end provide an indication of the strength of the collocation based on the corpus itself and on the participants. As regards the *t*-score, the average *t*-score in ISLC is 13.9 and 16.6 in LOCNEC. This difference between the mean *t*-score value per text in ISLC and LOCNEC was indicated to be statistically significant by an independent two-tailed *t*-test. This means that, on average, native speakers produce higher mean *t*-score values per text compared to the learners. The second part of the comparison involved MI scores: the distribution of the scores is similar in both corpora, with a higher concentration around 5 for LOCNEC and between 4 and 6 for ISLC. The average MI score for LOCNEC is 5.2 compared to 4.8 in ISLC. The difference is statistically significant, which means that, on average, native speakers produce higher MI values compared to learners. Furthermore, since previous studies on collocations adopt different methods, in order to increase comparability with the present work and verify whether, as most research points out, learners are more sensitive to frequent collocations (*t*-score collocations) and struggle with less frequent and more strongly associated ones (MI collocations), a comparison was carried out between the percentages of *t*-score and MI collocations produced by individual speakers. The analysis revealed that the learners, as regards both the *t*-score and MI, produce a significantly lower number of collocations compared to the native speakers. Thus, not only are the collocations produced by the learners less collocational than those of native speakers, but they are also, predictably, present in lower numbers compared to native speakers.

Lastly, as regards the effect sizes computed for all the analyses, their values ranged from medium (0.26) to large (0.49) thus attesting to the robustness of the magnitude of the observations. This indicates that the differences identified by the statistical tests are not negligible and should be taken into account.

Underuse of collocations in language learners has been widely evidenced in the literature: for example, Laufer and Waldman (2011) found underuse of verb + noun collocations in language learners and, while Vincze et al. (2016) employed a phraseological approach and found that there was no statistically significant difference between the number of collocations extracted from the native speakers' and the non-native speakers' corpora, they concluded that it is possible that learners may either overuse or underuse specific sets of them. Another study by Farghal and Obiedat (1995) found underuse of collocations in language learners and the authors proposed that simplification strategies, such as synonymy, paraphrase, or avoidance, may explain the underuse in language learners. Furthermore, a study on adverb + adjective collocations by Granger (1998b) observed that learners use significantly fewer *–ly* adverbs than native speakers in terms of types and tokens. This may result in a lack of diversification of linguistic items which contributes to the lack of native-like proficiency and fluency. Since all the studies mentioned above analysed collocations

in written texts, the finding of this combined analysis expands the concept of underuse and less collocationality of intensified adjectives collocations to spoken language too.

These results partially coincide with the literature on collocations in language learning, at least as regards the MI values. Indeed, infrequency has been shown to negatively relate to learners' processing and knowledge of collocations, as found in Li and Schmitt (2009), who suggested that less frequent and more strongly associated collocations identified by MI scores are "the type of item which is likely to be highly salient for native speakers" (p. 96), but not for non-native speakers. Similarly, Lorenz (1999) found that learners "significantly overuse the high-frequency items", while "the reverse is true for native speakers, whose usage is characterised by a high number of low-frequency word-forms and phrases" (p. 185). Furthermore, he verified the "learners' predilection for recurrent, well-attested combinations" as opposed to more infrequent and strongly associated ones (p. 186). This was also found in Ferraresi (2019) whose native speakers' speech proved to be more collocational than the non-native speakers'. Indeed, impromptu non-native speeches show lower collocationality values compared to the native speakers (Ferraresi, 2019, p. 215) and the reason may be found in the low-frequency nature of the high-MI collocations. Research has shown that our brains store in the long-term memory sequences of language which are then retrieved automatically without the need to compose them online through the open-choice principle (cf. Conklin & Schmitt, 2012). Since these sequences are infrequent, the learners' decreased level of exposure to them consequently results in a lower number or in less strongly associated collocations (cf. Ellis et al., 2008). However, the evidence of phraseological processing is mixed for non-native speakers: it seems that, unsurprisingly, proficient learners of a second language process formulaic language similarly to native speakers, whereas lower proficiency users tend to process words individually and struggle with opaque sequences such as idioms (Conklin & Schmitt, 2012). In the case of high-MI collocations, although these are not opaque, they tend to be less frequent and thus less easily retrieved. Indeed, Ellis et al. (2008; 2015) found that, contrary to native speakers, advanced learners are more sensitive to frequency rather than high-MI collocations. This was also confirmed by Bestgen and Granger (2014) who employed the same approach of Durrant and Schmitt (2009) to the study of collocations. The authors compared two groups of learners, intermediate and advanced ones. They found that intermediate learners tend to use a larger number of high-frequency collocations (high *t*-score) and few infrequent and strongly associated ones (high MI score), as opposed to advanced learners. Durrant and Schmitt (2009) also found that non-native speakers significantly underuse high-MI collocations.

The significant difference between the mean MI values of native and non-native speakers (similar to those found in Ferraresi, 2019) is thus expected and has already been widely

accounted for in the literature. However, the significant difference between the *t*-score values found in native speakers' texts and in the learners' is in contrast with most research findings. Indeed, Ferraresi (2019) found no significant difference between *t*-score text varieties in his work, meaning that both non-native and native speakers show similar levels of collocationality in speeches. The same result emerged from Durrant and Schmitt (2009) who concluded that "non-native writers make at least as much use of collocations with very high t-scores as do natives" (p. 174). Furthermore, more recent studies in psycholinguistics tested the L1 and L2 processing of collocations and corroborated that both groups of speakers are sensitive (i.e., process items faster) to high-frequency collocations (Öksüz et al., 2020; Siyanova-Chanturia et al., 2011a; Wolter & Yamashita, 2018; Yi, 2018). However, since psycholinguistics studies have tested collocational knowledge and processing rather than production, it may be that collocational performance lags behind collocational reception. This, in turn, could account for the lower mean *t*-score values found in the present work since these refer to collocation production. Furthermore, multiple studies have shown that the productive knowledge of collocations lags behind their receptive knowledge (Biskup, 1990; 1992; Farghal & Obiedat, 1995; Gyllstad, 2005; Jaén, 2009; Koya, 2005; Laufer & Waldman, 2011; Schmitt & Carter, 2004; Wray, 2002), thus despite the high level of proficiency of the participants, collocational competence may be attested a lower level. Therefore, it is not as surprising to find that learners produce weaker and fewer *t*-score collocations compared to native speakers in their oral production.

In conclusion, although much research has focused on the processing of collocations and has found that advanced learners process frequent formulaic sequences faster than infrequent ones (as native speakers do) (cf. Siyanova & Schmitt, 2008; Wolter & Gyllstad, 2013), the present analysis has revealed that learner collocation production differs from processing, especially in spoken language. Indeed, not only learners produce fewer collocations, but their association measures scores are lower compared to native speakers'. This is in contrast with findings from studies on written texts which highlight similar patterns as regards frequent collocations between learners and native speakers and should call for the attention of teachers. Collocation production lags behind collocation competence and, although this line may be very thin and not significant in written texts, the gap widens in spoken language and needs to be addressed.

Chapter 6

Qualitative analysis: frequent collocations

In the previous Chapter, a first evaluation of the data extracted from the ISLC and LOCNEC corpora was conducted via quantitative analyses. The analyses employed frequency counts and statistical tests to examine the empirical collocations (those whose *t*-score and MI were equal or greater than 2 and 3 respectively), as well as other combinations which did not qualify as collocations, but were still widely used by the learners. In the present Chapter, a selection of collocations from the ISLC corpus are analysed from a qualitative perspective:

- Completely different
- Really good
- Really interesting
- Really nice
- Totally different
- Very different
- Very good
- Very important
- Very interesting
- Very nice
- Very strange

In order to address the research questions (see Section 4.1), it was decided to systematically tackle each collocation according to a three-fold scheme illustrated in Section 4.4.2.3.

As mentioned in Section 4.4.2.3, the lexico-grammatical pattern of the collocations will be examined in comparison to LOCNEC. It is worth remembering that the descriptions and comparisons provided in these Sections have a purely descriptive goal. There is no prescriptive intention in this analysis, but rather a mere investigation of how this group of learners behave compared to their native speakers' peers. The hypothesis is that, since the collocations are frequent ones, the learners will display a similar behaviour to the native speakers, thus suggesting a complete mastery of them. The first Section presents the etymology of all the collocations object of this analysis; the second Section displays the L1 congruency of the collocations to Italian; the third Section presents the learner variables; and the fourth Section is dedicated to the qualitative investigation of the collocations.

6.1 Etymology of collocations

Table 6.1 illustrates the etymology (according to the OED, see Section 4.4.2.3) of the collocational items constituting the lexical collocations chosen for this analysis. In order to operationalise the description of the etymology and to aid the readers in obtaining a clear and concise picture, the Table presents a third column which assigns a label to the collocation: *same etymology*, *mixed etymology*, *different etymology*.

The labels were chosen on the basis of the following parameters:

- same etymology: the etymology of both collocational items can be traced back to Latin;
- mixed etymology: the etymology of one of the collocational items can be traced back to Latin;
- different etymology: the etymology of none of the collocational items can be traced back to Latin.

Furthermore, the etymological definitions are repeated for the same lexical items present in multiple collocations so as to facilitate the readers whenever they consult the Table. Indeed, in the following paragraphs, the single collocations will be analysed and this Table will remain as reference for all the etymological elements that may aid the analysis.

Table 6.1: Etymology of collocations.

Collocation	Etymology	Shared etymology
Completely different	*completely*: derivation from the adjective *complete* and the suffix *-ly*. The adjective *complete* comes from Latin *complētus*, past participle of *complēre* to fill up, finish, fulfil. Compare French *complet*, *complète.*	same etymology
	different: of multiple origins. Partly a borrowing from French. Partly a borrowing from Latin. < (i) Anglo-Norman and Middle French *different* (French *différent*) (adjective) having divergent attributes, distinct, not of the same kind, (noun) difference, disagreement, quarrel, and its etymon (ii) classical Latin *different-*, *differēns* differing, different, in post-classical Latin also eminent, superior (Vulgate), use as adjective of present participle of *differre differ* v. Compare Spanish *diferente*, Portuguese *diferente*, Italian *differente*.	
Very different	*very*: < Anglo-Norman *verrai*, *verrey*, *verai*, *veray*, Old French *verai*, *varai*, *vrai* (modern French *vrai*, Provençal *verai*), < the stem of Latin *vērus* true.	same etymology

	different: of multiple origins. Partly a borrowing from French. Partly a borrowing from Latin. < (i) Anglo-Norman and Middle French diffe*rent* (French *différent*) (adjective) having divergent attributes, distinct, not of the same kind, (noun) difference, disagreement, quarrel, and its etymon (ii) classical Latin *different-*, *differēns* differing, different, in post-classical Latin also eminent, superior (Vulgate), use as adjective of present participle of *differre differ* v. Compare Spanish *diferente*, Portuguese *diferente*, Italian *differente*.	
Really good	*really*: Formed within English, by derivation. < *real* adj. + *-ly* suffix. Compare post-classical Latin *realiter* actually, in fact (frequently from 11th cent. in British sources; from 14th cent. in continental sources; 6th cent. as a variant reading, where the correct reading is probably *dealiter* divinely), Middle French *reaument* (1310 in Old French), *realment* (1353; French *réellement*).	mixed etymology
	good: A word inherited from Germanic. Cognate with Old Frisian *gōd*, Old Saxon *gōd*, Old Dutch *guot*, *guod*, Old High German *guot*, Old Icelandic *góðr*, Norn *go-*, *goug(h)-* , *gu-*, Old Swedish *goþer*, Old Danish *goth*, Gothic *gōþs*, probably < an ablaut variant (*ō* -grade) of the base seen also in *gather* v., with an original sense 'fitting', 'suitable'; compare further Lithuanian (archaic and regional) *guodas*, Latvian gods, both in sense 'honour'.	

Very good	*very*: < Anglo-Norman *verrai*, *verrey*, *verai*, *veray*, Old French *verai*, *varai*, *vrai* (modern French *vrai*, Provençal *verai*), < the stem of Latin *vērus* true.	mixed etymology
	good: A word inherited from Germanic. Cognate with Old Frisian *gōd*, Old Saxon *gōd*, Old Dutch *guot*, *guod*, Old High German *guot*, Old Icelandic *góðr*, Norn *go-*, *goug(h)-* , *gu-*, Old Swedish *goþer*, Old Danish *goth*, Gothic *gōþs*, probably < an ablaut variant (*ō* -grade) of the base seen also in *gather* v., with an original sense 'fitting', 'suitable'; compare further Lithuanian (archaic and regional) *guodas*, Latvian gods, both in sense 'honour'.	
Very important	*very*: < Anglo-Norman *verrai*, *verrey*, *verai*, *veray*, Old French *verai*, *varai*, *vrai* (modern French *vrai*, Provençal *verai*), < the stem of Latin *vērus* true.	same etymology
	important: Probably partly a borrowing from Latin. Probably partly a borrowing from French. Probably partly (i) < post-classical Latin *important-*, *importans* momentous, use as adjective of present participle of classical Latin *importāre* in its post-classical Latin sense 'to be of consequence, weight, or force', and partly (ii) (in later use) < Middle French *important* consequential, weighty, momentous (< Italian *importante*, use as adjective of present participle of *importare* import v.). Compare Catalan *important*, Spanish *importante*, Portuguese *importante*, Italian *importante*.	

Really interesting	*really*: Formed within English, by derivation. < *real* adj. + *-ly* suffix. Compare post-classical Latin *realiter* actually, in fact (frequently from 11th cent. in British sources; from 14th cent. in continental sources; 6th cent. as a variant reading, where the correct reading is probably *dealiter* divinely), Middle French *reaument* (1310 in Old French), *realment* (1353; French *réellement*).	same etymology
	interesting: < *interest* v. + *-ing* suffix. *Interest* is an alteration of the earlier *interess* v., after *interest* n. Specifically, an alteration of the earlier *interess* n., apparently after the cognate French *interest*, modern French *intérêt*, apparently a noun use of Latin *interest* it makes a difference, concerns, matters, is of importance, 3rd person singular present indicative (used impersonally) of the verb, of which the infinitive *interesse* was used as a noun in medieval Latin, and in the other Romanic languages and Middle English.	
Very interesting	*very*: < Anglo-Norman *verrai*, *verrey*, *verai*, *veray*, Old French *verai*, *varai*, *vrai* (modern French *vrai*, Provençal *verai*), < the stem of Latin *vērus* true.	same etymology

	interesting: < *interest* v. + *-ing* suffix2. *Interest* is an alteration of the earlier *interess* v., after *interest* n. Specifically, an alteration of the earlier *interess* n., apparently after the cognate French *interest*, modern French *intérêt*, apparently a noun use of Latin *interest* it makes a difference, concerns, matters, is of importance, 3rd person singular present indicative (used impersonally) of the verb, of which the infinitive *interesse* was used as a noun in medieval Latin, and in the other Romanic languages and Middle English.	
Really nice	*really*: Formed within English, by derivation. < *real* adj. + *-ly* suffix. Compare post-classical Latin *realiter* actually, in fact (frequently from 11th cent. in British sources; from 14th cent. in continental sources; 6th cent. as a variant reading, where the correct reading is probably *dealiter* divinely), Middle French *reaument* (1310 in Old French), *realment* (1353; French *réellement*).	mixed etymology
	nice: A borrowing from French. < Anglo-Norman *nice*, *nis*, *nise* and Old French *nice* < classical Latin *nescius* (see *nescious* adj.). Compare Old Occitan *nesci*, Spanish *necio*, Catalan *neci*, *nici*, Portuguese *necio*, Italian *nescio*, all in sense ‘foolish, simple, ignorant’. The semantic development of this word from ‘foolish, silly’ to ‘pleasing’ is not paralleled in Latin or in the Romance languages. The precise sense development in English is unclear.	

Very nice	*very*: < Anglo-Norman *verrai*, *verrey*, *verai*, *veray*, Old French *verai*, *varai*, *vrai* (modern French *vrai*, Provençal *verai*), < the stem of Latin *vērus* true.	mixed etymology
	nice: A borrowing from French. < Anglo-Norman *nice*, *nis*, *nise* and Old French *nice* < classical Latin *nescius* (see *nescious* adj.). Compare Old Occitan *nesci*, Spanish *necio*, Catalan *neci*, *nici*, Portuguese *necio* , Italian *nescio*, all in sense 'foolish, simple, ignorant'. The semantic development of this word from 'foolish, silly' to 'pleasing' is not paralleled in Latin or in the Romance languages. The precise sense development in English is unclear.	
Totally different	*totally*: < *total* adj. + *-ly* suffix: compare Schol. Latin *totāliter*, Old French *totalement*. *Total* < French *total* = Spanish *total*, Portuguese *total*, Italian *totale*, < Schol. Latin *tōtālis*, < Latin *tōtus* entire.	same etymology
	different: of multiple origins. Partly a borrowing from French. Partly a borrowing from Latin. < (i) Anglo-Norman and Middle French *different* (French *différent*) (adjective) having divergent attributes, distinct, not of the same kind, (noun) difference, disagreement, quarrel, and its etymon (ii) classical Latin *different-*, *differēns* differing, different, in post-classical Latin also eminent, superior (Vulgate), use as adjective of present participle of *differre differ* v. Compare Spanish *diferente*, Portuguese *diferente*, Italian *differente*.	

Very strange	*very*: < Anglo-Norman *verrai*, *verrey*, *verai*, *veray*, Old French *verai*, *varai*, *vrai* (modern French *vrai*, Provençal *verai*), < the stem of Latin *vērus* true.	same etymology
	strange: < Old French *estrange* (modern French *étrange*) = Provençal *estranh*, *estrang*, Spanish *estraño*, Portuguese *estranho*, Romanian *strâin*, Italian *strano* adjective, *stranio*, *strangio* noun < Latin *extrāneus* external, foreign (see extraneous adj.), < *extrā* adverb, outside, without.	

The readers will thus be able to consult the Table for each collocation and find all the etymological definitions pertaining to the collocation without needing to compose the definition and the etymological label themselves. Lastly, in Section 4.4.2.3 etymology has been argued to be strictly related to L1 congruency; the etymology label in Table 6.1 will also anticipate some of the findings in Table 6.2 regarding L1 congruency.

The first column of Table 6.1 corresponds to the collocations under investigation; the second column is split into two rows to provide the etymology of the two lexical items composing the collocation. The first row is occupied by the adverb, while the second row focuses on the adjective. The etymological definitions are extracted from the OED, with minor editing, mainly related to the elimination of the dates referring to the first use of the etymons. The last column provides the etymological label (as mentioned above) to facilitate the reading of the table. For example, in case of the first collocation, *completely different*, both the adverb and the adjective's etymons come from Latin (*complētus* and *different-*), thus the collocation can be deemed to share the same etymology with the corpus participants' L1.

The majority of the collocations indeed share the same etymology with Italian (and thus could be anticipated to be L1 congruent), but one note should be made regarding the collocations containing *very.* The adverb comes from Anglo-Norman and Old French, both of which stem from the Latin *vērus*, which means "true"; in this case the etymology is not directly related to Latin, but French influence is interposed between the contemporary form of the adverb and its root. However, this adverb has been considered to have the same etymology of the L1 because in contemporary Italian *vero* (adj.), *veramente* (adv.) are still used today and they share the same Latin root *vērus*. Furthermore, as regards collocations containing *nice*, although the adjective stems from Latin *nescius*, the shift in meaning between the Latin "foolish, simple, ignorant" and the contemporary form of the adjective is too big to account for any similarities between the two words of which the participants may be aware. Thus, the adjective *nice* has been deemed to have a different etymology.

6.2 L1 congruency of collocations

In a similar fashion to the previous Section and in an attempt to operationalise as many variables as possible (as per diagram in Figure 4.4.1), L1 congruency is displayed in Table 6.2. As explained in Section 4.4.2.3, L1 congruency determination is loosely based on Wolter and Gyllstad's (2013) approach, namely the prototypical semantic value translation of the collocations by a native speaker (who in this case are seven, rather than one) and the frequency estimates in a reference corpus.

Table 6.2: L1 congruency of collocations.

	Translation to L1	PEC occurrences (pmw)	L1 congruency
Completely different	completamente divers*	7.55	L1 congruent
Very different	molt* divers*	17.82	L1 congruent
Really good	davvero buon*	0.57	L1 congruent
Very good	molto buon*	3.51	L1 congruent
Very important	molt* important*	24.95	L1 congruent
Really interesting	davvero interessant*	0.98	L1 congruent
Very interesting	molt* interessant*	11.10	L1 congruent
Really nice	veramente carin*	0.26	L1 congruent
Very nice	molto carin*	4.4	L1 congruent
Totally different	totalmente divers*	1.93	L1 congruent
Very strange	molt* stran*	2.91	L1 congruent

As explained in Section 4.4.2.3, since Italian has multiple translations for some of the collocational items (e.g., *really, very, good, nice*), it was decided to employ seven native speakers who provided direct translations for the collocations. The most frequent translation among the native speakers was chosen as the final one[1]. The second column of Table 6.2 displays the final translation obtained by the group of native speakers. The asterisks have been placed where the adverb or adjective can be declined according to gender, singular or plural, in Italian; the translations have been searched in PEC using the asterisk as a wild card to obtain all declinations of the collocational items. The third column displays the frequency occurrences of the L1 translations in the PEC corpus normalised to one million words (pmw). The last column illustrates whether the collocation is L1 congruent or incongruent: all collocations have been indicated as L1 congruent.

[1] All the L1 native speakers' translations are available in the Appendix.

Table 6.3: Learner metadata variables with reference to the collocations.

	gender		stay-abroad experience			university courses										other languages				
	♀	♂	yes		no	1	2	3	4	5	6	7	8	9	10	11	CHI	FRE	GER	RUS	SPA
			< 6 m.	> 6 m.																	
Completely different	✓	✓	✓	✓		✓		✓	✓				✓		✓		✓	✓	✓	✓	✓
Very different	✓	✓	✓					✓		✓			✓	✓	✓		✓		✓	✓	✓
Really good	✓	✓	✓		✓	✓	✓	✓	✓	✓			✓		✓		✓	✓		✓	✓
Very good	✓	✓	✓	✓		✓		✓				✓	✓	✓			✓	✓	✓		✓
Very important	✓	✓	✓	✓				✓		✓				✓	✓			✓	✓	✓	✓
Really interesting	✓	✓	✓	✓				✓	✓					✓	✓			✓	✓	✓	✓
Very interesting	✓		✓											✓	✓				✓	✓	
Really nice	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓						✓	✓		✓
Very nice	✓		✓	✓				✓			✓	✓			✓			✓		✓	✓
Totally different	✓	✓	✓	✓	✓		✓	✓				✓		✓			✓	✓	✓		✓
Very strange	✓	✓	✓	✓		✓		✓							✓		✓	✓	✓		

6.3 Learner metadata

As reported in Section 4.4.2.3, the learners' ethnographic metadata (i.e., gender, stay-abroad experience, university course, other languages) have also been systematically gathered in a table (Table 6.3).

This was done to facilitate the interpretation of any correlation between the collocations and the learner variables. Table 6.3 is built around the collocations and the learner variables, which have been reported in the top row.

The first is gender (distinguished by the symbols ♀ and ♂); then there is stay-abroad experience (if yes, indicated if for less than six months (< 6 m.) or more than six months (> 6 m.); university courses. See Section 4.2.5. ; and lastly, other languages (in particular, the speakers' L3 ranked in terms of proficiency by the speakers themselves – see Section 4.2.3).

From the analysis of the Table a few facts emerge: both males and females make use of this set of collocations (the males in a smaller number due to the lower proportion of males in the degree courses of Foreign Languages); the stay-abroad experience does not seem to have an impact on the use of collocations, since all of these are used by learners who have either spent less than six months or more than six months indiscriminately. In addition, in four cases learners who have not had any experience abroad in an English-speaking country have also used these collocations together with the other groups of learners. As regards the university courses, it seems that courses n° 3 and n° 10 (corresponding to courses belonging to the Department of Foreign Languages of University A) are the most productive since n° 10 and nine speakers respectively attend these courses and have produced the majority of collocations. Course n° 3 is an undergraduate course, whereas course n° 10 is a postgraduate one. It must be noted that University A is the university which generated the highest number of participants and courses, thus the high number of courses n° 3 and n° 10 may be due to this fact. Lastly, the other languages which were factored in the analysis, namely Chinese, French, German, Russian and Spanish, do not show any explicit correlation, although French, German, and Spanish seem to be the most frequent in correlation with the collocations. However, since these are also the most frequent languages studied among the learners, it is difficult to claim whether learners who study these languages are the ones producing this set of collocations, or this is simply due to the limited sample size of learners.

6.4 Analysis of collocations

In this Section, 11 collocations will be analysed from a qualitative perspective, as anticipated in Section 4.4.2.3, following the diagram presented in Figure 4.4.1. References will be made

to the tables presented in the Sections above (i.e., Table 6.1, 6.2, and 6.3).

6.4.1 Completely different

Different is a common classifier adjective occurring 80 times per million words in the English language (Biber et al., 1999); it is one of the most common adjectives which functions as subject predicative of copular *be* occurring over 40 times per million words (Biber et al., 1999, p. 440). It is categorised as belonging to the A1 level of CEFR for learners of English by the WFF. Furthermore, the adjective shares the same etymology with Italian (see Table 6.1), thus it seems plausible that Italian learners of English may be positively influenced in their use of this adjective. The adjective can be used both in predicative and attributive position.

Completely is an adverb categorised as belonging to the A1 level of CEFR by the WFF; the adverb, similarly to the adjective, shares the same etymology with Italian. Furthermore, the collocation is L1 congruent (see Table 6.2). The overall picture of the average user of this collocation which emerges from Table 6.3 is quite varied, although it seems that the speakers who have not spent much time abroad did not use this collocation. Speakers of all other L3 languages produced this collocation. The collocation is L1 congruent.

In ISLC, *different* in conjunction with *completely* is used seven times with a predicative function, such as in:

(6.4.1) <B> [. . .] life in a big family the school was *completely different* from mine and I loved it [. . .] </B>

(6.4.2) <B> [. . .] I (uh) landed in Beijing it was *completely different* so (uhm) I've seen both the [. . .] </B>

The remaining four occurrences of *different* used with an attributive function, such as in:

(6.4.3) <B> [. . .] the main idea itself in a *completely different* world so I can imagine [. . .] </B>

(6.4.4) <B> [. . .] to consider that it is a *completely different* country it is another world [. . .] </B>

In terms of functions, both are used by the learners without particular preference for one over the other. All occurrences of the collocation are found in sentences with a positive meaning, no negatives were found to be associated with this collocation. Furthermore, in the majority of the cases (seven), the learners used the collocation in a sentence supported by a verb in

the present tense, whereas the remaining sentences (four) were held by a verb in the past tense, such as in (6.4.1) and (6.4.4). The verbs in the sentences containing the collocations are all affirmative and, as regards, the construction of the sentence, the preferred one seems to be the standard *subject + copular verb + collocation*, which occurs five times, such as in (6.4.1); there are two instances of an *it-* structure (*it + copular verb + collocation*), as shown in (6.4.2) and (6.4.4).

In terms of collocational post-modification, in practically all the instances except one, the learners produced sentences which explicitly mentioned the subject or the object (according to the adjectival function). The nouns associated with the collocation seem to indicate that the learners, even different ones, use *completely different* to refer to similar topics, e.g., school life, travelling, people.

During the analysis of these instances, the main communicative function of *completely different* emerged. It seems that the collocation is used to highlight a difference between one element of the clause and another. The function further divides into three some subtle sub-categories:

- comparison;
- expectation;
- opposites.

The comparison function of *completely different* refers to all those cases in which the collocation is used to simply compare two terms, either in a positive, negative, or neutral way. Examples of this below:

(6.4.5) <B> [. . .] so I stayed with four different families there (uhm) they had kids (uh) younger and older than me so I wasn't alone and it was also a different experience because I'm an only child so I me= I experienced (uh) the life in a big family the school was *completely different* from mine and I loved it because I didn't have (mm) much work to do at home because we worked a lot in school in class [. . .] </B>

In this case, the learner used the collocation to emphasise the difference between the school she attended in the foreign country where she did her experience abroad and her home school. The difference is positive for the subject of the sentence "the school" in Canada, so the predicative function of *completely different* is to convey a positive experience, i.e., even better than the school back home in Italy. A similar example of this positive comparison is the following:

(6.4.6) <B> [...] I was there to study German and (uh) it was my first time alone in a country for so long and there I went to university (uh) I (uhm) I studied a lot of new (uh) subjects that I don't study here (uh) and (uh) it was very interesting because (uh) the (uh) university there is *completely different* (uh) we had a lot of different projects also grouped group group projects and (uhm) we (uh) had to do lot of (uhm) different presentations [...] and (uhm) it was very interesting and [...] </B>

This is another example of a positive use of *completely different* which is found in predicative position and positively characterises the university the learner attended in the foreign country. The collocation is also used to express a negative feeling to the comparison, such as in the sentence below:

(6.4.7) <B> [...] but then there are some like spices which are different I mean we don't use here (mm) other differences well for sure the political situation the government is *completely different* because here we live and we've got everything from the government everything we want really and there they just don't have a real government [...] </B>

In this case, as opposed to the previous ones, the collocation is used to negatively connote the subject, the government, which is different from the Italian one, it is worse.

Lastly, *completely different* can also be used to compare in a neutral way, simply stating that there is difference between two terms, not necessarily positive or negative:

(6.4.8) <B> [...] they're very strange in fact because they're not very social as we are here in in the Western countries but we have to consider that it is a *completely different* country it is another world we can say because the first contact with these people began in the second half of the nineteenth century so I think this is normal they are very (uh) (uh) secretive we can say very secretive [...] </B>

The learner is describing his experience abroad in China and he is marking the fact that China and Italy are *completely different* countries, without assigning a particularly positive or negative connotation to either. In this case, the collocation is used in attributive position and has a neutral comparison function.

As regards the second sub-category function of *completely different*, the expectation function, this has been identified as a way of expressing how something turned out to be *completely different* from what was expected. The unexpected event or object can either be positive or negative. Excerpts (6.4.9) and (6.4.10) provide an example of this:

(6.4.9) <B> [...] I went to to China for an internship since I study (uh) Chinese and (uh) I expected China as my (uhm) as my professor <Chinese> (uh) (uh) told me (uhm) China is the future is the (uhm) the new (uh) vanguard country (uh) Beijing is like the new New York and I trust her but (uh) as I (uh) landed in Beijing it was *completely different* so (uhm) I've seen both the faces of (uh) China because [...] </B>

In this example, the learner had built an expectation of Beijing following his professor's descriptions and memories; however, on arriving in Beijing, his expectations were not met and he described the city as *completely different* from what his professor had told him. Thus, in this case it is possible to observe the function of the collocation which is used to highlight a difference that is ascribable to an expectation. This learner's expectation was not met, Beijing turned out not to be what his professor had told him, so the difference tends towards the negative. In the following example, the unexpected event was a positive one:

(6.4.10) <B> [...] (uhm) well (uhm) . (uhm) . a detective is working in a forest and he suddenly sees a (uhm) a pretty strange-looking box and he decides to take it home (uh) to analyse it and he was absolutely sure that it wasn't anything special so he decided to open it but what he found was *completely different* from (uh) what he was expecting (uhm) in fact he found (uhm) a sort of collier in it made of diamonds and it looked very expensive [...] </B>

The learner is performing the story-telling task, and she is starting to set the circumstances for the story. A detective finds a box in a forest and "he was absolutely sure that it wasn't anything special", so he decides to open the box without any expectations, but what he finds exceeds his expectations as the object is an expensive necklace. Once again, the collocation is used to qualify the object as something different from previous expectations, which in this case were exceeded, thus assigning a more positively marked meaning to *completely different*.

The last type of function identified within the broader comparison function of *completely different* is that of the opposites. This sub-category of comparison refers to all those instances in which the collocation is used to mark the difference between two terms which are in fact opposites:

(6.4.11) <B> [...] and people are generally very poor and they cannot afford just the the minimal (uh) goods for their basic life so food but very very the the conditions are very very low (uhm) while (uhm) in in the heart of Beijing is *completely different* there are many malls (uh) Armani Gucci (uh) whatever (mm) it is very particular because nobody nobody say anything about the outskirts so [...] </B>

This is another description of a learner's experience abroad in China: the learner is describing the difference between the countryside and the cities in China. The countryside is characterised by small villages where the sanitary conditions are very poor, whereas the cities, such as Beijing, are *completely different*, that is, the opposite. There are several shopping centres and the city is generally richer than the countryside villages. Thus, the difference that the collocation is qualifying is not simply a matter of one being better or worse than the other, but the two terms of comparison being quite the opposite. This also happens in the following example which incidentally includes two instances of the same collocation:

(6.4.12) <B> she's a girl and she's the wise girl that had to (uhm) turn into a snake to flee from Christians that who wanted to kill the pagans and her refugee is near (uhm) near water [...] so I had to tell this story [...] so it was really really hard for me because I'm a woman but it was in really beautiful see to see people with tears in their eyes (uh) while listening to me because I what= what I was saying was true and while the= in the previous show I had to do a *completely differen*t character because I was the mistress (uh) because my=my teacher wants us to ch=choose characters that are *completely different* from us so (uh) it's like (uh) a yeah a challenge for us [...]

This is a lengthy recollection of a theatre play in which the learner took part and played the role of a woman from the Celtic mythology who was persecuted by the Christians and had to flee her home. The narration explored the theme of domestic violence and the learner continued by explaining that the character she played was *completely different* from the character she had played in a previous play, a mistress. Although it may be debated whether a female Celtic heroine may represent the opposite of a modern-day mistress, the general understanding that these two characters do not share common values could be agreed upon. Thus, the collocation is used once again to mark a stark contrast between opposite characters, a female heroine fighting for women's rights and a mistress with different different priorities. This difference is stressed in the following line when the learner explains that the drama teacher encouraged his group to portray characters that are *completely different* from them, practically the opposite. In terms of L1 congruency, a random reproducible sample of 20 occurrences of *completamente divers** revealed that Italian native speakers use the collocation similarly as regards the functions. Indeed, the comparison and the opposites functions clearly emerged from the reference corpus, with a preference for the L1 speakers for the opposites (12 cases). No instances of expectation function were found.

As far as LOCNEC is concerned, *completely different* occurs four times and the native speakers use the collocation mostly in predicative position with one instance of attributive function. Similarly to ISLC, all the sentences which produced the collocation were positive,

and two contained a verb in the present tense, while the others presented the verb in the past tense. As regards the construction of the sentence, there were two instances of an *it*-structure, while in the other two cases the collocation was produced within an elliptical context[2]. In terms of function, the collocation is used mainly to compare two terms and resembles the comparison function identified in ISLC, with a positive, negative, or neutral connotation. In the case of LOCNEC, three sentences contain a neutral comparison, while one contains a positive comparison. There are no instances of other functions, as identified in ISLC, used in LOCNEC.

6.4.2 Really good

Good is an evaluative descriptor adjective occurring more than 20 times per million words (Biber et al., 1999, p. 509) and can be found in either attributive or predicative position. It is ranked by Biber et al. as one of the most common adjectives across registers occurring over 200 times per million words in at least one register. The etymology of *good* stems from Old English *gōd* (see Table 6.1) and thus there is no etymological link between *good* and an L1 equivalent in Italian such as *buono*. However, *good* is classified by the WFF as an A1 CEFR level word, which indicates that the advanced learners of ISLC should be more than aware of its meaning and use.

On the other hand, the etymology of *really* is Latin, so that the collocation shares a mixed etymology with Italian. The intensifier also belongs to the A1 CEFR level according to the British Council's WFF. Both items are classified as A1 words and as such good knowledge of the individual words could be expected from the learners. Furthermore, *really good* is a very frequent collocation as indicated by Biber et al. (1999, p. 543) who state that it occurs at least 20 times per million words. This further corroborates the hypothesis that the ISLC learners may use the collocation well and in a similar way to the native speakers. In terms of L1 congruency, the collocation is also L1 congruent.

As regards the ISLC speakers, eight participants use this collocation in their texts, three of them twice. Both males and females use it and no learner with no experience abroad has produced it. This collocation is used across the highest number of university courses (seven), although interestingly, despite the adjective's shared etymology with Germanic languages, no speakers studying German as L3 produced it.

The 11 instances of *really good* in ISLC show a predisposition for the collocation to be used in a predicative position, thus it usually post-modifies the subject:

[2]Ellipsis is a common phenomenon in conversation and it usually occurs when one or more elements of a sentence (which can easily be deducted from the context) are omitted.

(6.4.13) <B> [. . .] different from Jer=Jerusalem but I found it I found it *really good* (uh) this (uh) this city (uh) really impressed me [. . .] </B>

(6.4.14) <B> [. . .] and (uh) the meat they make (uhm) for example also cakes were *really good* there were really a lot of types and (uhm) also [. . .] </B>

In example (6.4.13) the collocation is post-modifying the pronoun *it* which refers to Jerusalem, the city the speaker has been to on holiday. In four cases the modified adjective *good* is found in attributive position, such as in the following examples:

(6.4.15) <B> [. . .] about London (uhm) probably that (uhm) if you don't have a *really good* (mm) sense of orientation you can get lost very [. . .] </B>

(6.4.16) <B> [. . .] Hard Rock Cafè so (uhm) we I think everyone can enjoy a *really good* hamburger [. . .] </B>

The collocation's post-modified nouns refer mainly to food (cakes, hamburger), entertainment (actors, episode), and travelling. In terms of tense, this group of speakers has a preference for the past tense since seven out of 11 occurrences are in the past simple. Only three speakers use the present tense, with IT017 producing two instances in the present. (6.4.17) and (6.4.18) show the use of the tenses:

(6.4.17) <B> [. . .] probably people yes people and food it was not *really good* in particular in the North . actually we tried [. . .] </B>

(6.4.18) <B> [. . .] them just a little bit but actually it it's still a *really good* episode one of the best I've ever seen the whole [. . .] </B>

Excerpt (6.4.17) is particularly interesting as it introduces one feature of the learners' use of *really good*, namely that of a negation. This occurs in two other cases beside excerpt (6.4.17):

(6.4.19) <B> [. . .] about London (uhm) probably that (uhm) if you don't have a really good (mm) sense of orientation you can get lost very [. . .] </B>

(6.4.20) <B> [. . .] how the (uh) the film would be but it wasn't (uh) *really good* because it wasn't like the books (uhm) for example [. . .] </B>

In these cases, it appears clear that the positive connotation of *really good* remains, but the speaker is trying to negate it by turning the sentence into a negative. In (6.4.19) a good sense of orientation is needed not to get lost in London, so that the more natural sentence for

speaker IT017 becomes "if you don't have a really good sense of orientation". She prefers the negation of the collocation as opposed to a different lexical choice, such as *poor sense of orientation*. The same happens in (6.4.20) where she was expecting a really good film, but this was not the case and the film becomes *not really good*. This form of negation may be considered as a linguistic hedge according to Fraser (2010), although not all negations should be regarded cases of hedging. Hedging is a pragmatic competence that is embedded in language and is used to communicate the intended message "with all its nuances in any socio-cultural context" (Fraser, 2010, p. 15). Furthermore, the sentence is preceded by *probably*, which reduces the speaker's commitment and says something about the reliability of the knowledge (Chafe, 1986, p. 263), or rather, it expresses uncertainty about the truth of the proposition (Hübler, 1983, p. 124).

A further element of interest is the construction of the sentence, namely whether it is built with a pronoun or an *it-* structure. It seems that the ISLC speakers prefer to use personal pronouns as the subjects of the main sentence (*pronoun + copular verb + collocation,* or *pronoun + modal verb + collocation*) containing the collocation since there are eight occurrences of this.

(6.4.21) <B> [. . .] how the (uh) the film would be but it wasn't (uh) *really good* because it wasn't like the books (uhm) for example [. . .] </B>

(6.4.22) <B> [. . .] probably people yes people and food . it was not *really good* in particular in the north . actually we tried [. . .] </B>

(6.4.23) <B> [. . .] Hard Rock Cafè so (uhm) we I think everyone can enjoy a *really good* hamburger [. . .] </B>

These three excerpts show the use of *really good* in combination with a pronoun. In excerpt (6.4.21) the collocation is post-modifying *it*, which is the deictic reference to film. In (6.4.22) the collocation is once again post-modifying *it* which refers to food; and lastly in (6.4.23) *everyone* is the subject pronoun of the sentence containing *really good*. There are only three instances of an *it-* structure:

(6.4.24) <B> [. . .] so home with his cat he can relax there it was a *really good* evening [. . .] </B>

(6.4.25) <B> [. . .] eat small portions of different kinds of foods and it was *really good* and (uhm) also (uhm) we ate we ate [. . .] </B>

(6.4.26) <B> them just a little bit but actually it it's still a *really good* episode one of the best I've ever seen the whole [. . .] </B>

From this short sample of occurrences and the scarce use of the impersonal structure, it is clear that the learners prefer to use the collocation with a more personal link to the subject or object of the sentence. In terms of L1 congruency, it has been noticed in a random reproducible sample of 20 occurrences in PEC, that the L1 equivalent *davvero buon** usually occurs in a positive sentence and is used to refer to the quality of food.

The main noticeable difference between ISLC and LOCNEC is the number of occurrences: there are 71 instances of *really good* in LOCNEC. There is a prevalence of predicative function in LOCNEC with 48 occurrences of predicative use and 23 of attributive use. Furthermore, due to high number of occurrences, there is a wide variety of nouns used in both the attributive and predicative function of the collocation. They mainly refer to entertainment and experiences. There are also three instances of sentences being used as the subject of the post-modifying collocations supported by the relative pronoun *which*:

(6.4.27) <B> [. . .] and they give out blankets which is re= re= *really good* w= when you walk in [. . .] </B>

(6.4.28) <B> [. . .] they could actually have the photo themselves which was *really good* cos they thought it was amazing .. but . [. . .] </B>

(6.4.29) <B> [. . .] railway and children come and ride on it which is *really good* <laughs> [. . .] </B>

Overall, the speakers in LOCNEC mainly use the collocation in association with entertaintment and fun, so much so that *really good fun* could be considered a collocation of its own given the restricted collocability of the elements. Entertainment seems to be the main semantic category of *really good*, with *film*, *fun*, *friends*, *music*, *acting*, *book*. There is also reference to the university world with *reputation*, *block*, *linguistics department*, *staff* and *corridor*. As regards the tense used in association with the collocation, there is an equivalent distribution of occurrences between past simple and present simple, 38 and 33 respectively.

Another important difference between ISLC and LOCNEC is the lack of negative sentences in LOCNEC. All the sentences containing *really good* in LOCNEC are positive, they all express positive evaluations about objects, people, or facts. Since the negative sentences found in ISLC were considered linguistic hedges, it is interesting to note that in LOCNEC hedging is present, but through other types of hedges, such as in:

(6.4.30) <B> [...] the way they did was well I thought was *really good* because they had brought it into: . context into [...] </B>

This excerpt highlights that the native speakers' may use other hedging devices, such as the use of *I thought* and the hesitant *well* (cf. Johansen, 2020).

Lastly, as regards the structures and the use of pronouns for the main clause containing the collocation, in LOCNEC we find a higher level of variety, probably due to the bigger sample. However, the main two types of occurrences are sentences built around personal pronouns (mainly *it*) and the impersonal *it*- structure, occurring 27 and 24 times respectively.

6.4.3 Really interesting

Interesting is an evaluative descriptor adjective occurring at least 10 times per million words (Biber et al., 1999, p. 530) and is a derivation of Latin *interest*. According to the WFF, the word belongs to the A1 CEFR level. As already described above, *really* is a derivation of Latin *realiter* (see Table 6.1) and the intensifier also belongs to the A1 CEFR level according to the British Council's WFF. Thus, in the case of this collocation, the components share similar etymological origins. Furthermore, both items are classified as A1 words and as such good knowledge of the individual words is expected from the learners, especially since the collocation is also L1 congruent. As regards the ISLC speakers, four participants use this collocation in their texts and all four have different experiences: either they have had a stay-abroad experience of less than six months, more than six months, or none at all. They belong to four different university courses (n° 3, 4, 9, 10) and speak four languages except Chinese.

There are five instances overall of *really interesting* in the ISLC corpus. The five instances of *really interesting* show a predisposition for the collocation to be used in a predicative position, thus it usually post-modifies the subject. The predicative function of the collocation is displayed in the two following extracts:

(6.4.31) <B> [...] is an inscript under this golden fish statue (mm) which is *really interesting* and (uh) and it makes the character [...] </B>

(6.4.32) <B> [...] and (uhm) also everything about (uh) tourism that I studied was *really interesting* and I'm really proud of myself because [...] </B>

In the first case, *really interesting* refers back to the *inscript* (sic) under the golden fish statue with the use of the pronoun *which*. In the second case, the collocation refers to the clause "everything about tourism I studied". All five occurrences of the collocations are used in predicative position. The learners do not show a preference for the tense of the clause containing the collocation, indeed there are two instances of past simple:

(6.4.33) <B> [. . .] and (uhm) also everything about (uh) tourism that I studied was *really interesting* and I'm really proud of myself because [. . .] </B>

(6.4.34) <B> [. . .] to the Hermitage (uhm) I did do different things and that was *really interesting* (uh) the best part (uh) was that I fell in [. . .] </B>

Two instances of present simple:

(6.4.35) <B> [. . .] is an inscript under this golden fish statue (mm) which is *really interesting* and (uh) and it makes the character [. . .] </B>

(6.4.36) <B> (uh) (uh) is a really (uh) really big and really (uhm) well *really interesting* also from a point of view of history (uhm) [. . .] </B>

And one sentence containing the conditional tense:

(6.4.37) <B> [. . .] that probably there's a message in it so it would be *really interesting* to open it and maybe find something about [. . .] </B>

As can be seen from these excerpts, all sentences are positive, there are no negatives and the collocation is used to express an evaluative judgement about an object, event, or experience. The collocation is built around pronouns (*pronoun + copular verb + collocation*) and one explicit subject (the Central Station in New York). The pronouns are *that*, *which* and a clause:

(6.4.38) <B> [. . .] one of the most beautiful things I saw was the Great Station that (uh) (uh) is a really (uh) really big and really (uhm) well (uh) (uh) is a really (uh) really big and really (uhm) well *really interesting* also from a point of view of history (uhm) [. . .] </B>

(6.4.39) <B> [. . .] to the Hermitage (uhm) I did do different things and that was r*eally interesting* </B> <B> (uh) the best part (uh) was that I fell in [. . .] </B>

(6.4.40) <B> [. . .] is an inscript under this golden fish statue (mm) which is *really interesting* and (uh) and it makes the character [. . .] </B>

(6.4.41) <B> [. . .] and (uhm) also everything about (uh) tourism that I studied was *really interesting* and I'm really proud of myself because [. . .] </B>

It appears that in the first two cases, (6.4.38) and (6.4.39), the learners are making a straightforward use of the collocation (evaluation function) without assigning it new functions.

However, it is interesting to note that two speakers (IT006 and IT008) produced the collocation while performing the story-telling task and in these cases the collocations are different in terms of pragmatic function.

(6.4.42) <B> [...] Benny and Andy are walking through a forest and they are just talking about (uh) their lives and sharing thoughts and looking around and (uh) they suddenly see a very strange-looking box (uhm) Benny say= Benny says that (uhm) it would be totally crazy to take it home but Andy as a boy and as (uhm) sort of (uhm) crazy and curious (uh) boy decides (uh) to take the box home (uh) because he thinks that probably there's a message in it so it would be *really interesting* to open it and maybe find something about (uh) the world or something secret but of course Benny doesn't agree with this but so they start fighting abou= about this over this (uh) but in the end they (uh) agree [...] </B>

(6.4.43) <B> [...]A character is walking in a forest and suddenly they see a strange-looking box they take it home and they open it they take it home and they open it and they get really surprised by what's inside this box (uh) in this box there is (uh) this golden fish statue which makes the the character that foun= that found the box very surprised (uh) there is an inscript under this golden fish statue (mm) which is *really interesting* and (uh) and it makes the character completely interested in reading what's what's there so basically (uh) this inscript says that if you rub this statue this golden fish statue ten times [...] </B>

The longer extracts show how the speakers are performing the task online and struggling to give the text meaning. The collocation refers to the possibility of a message in (6.4.42) and to an inscription in (6.4.43), but no personal opinion is expressed. The collocation is attributed to a character and to an inscription, so that none of the collocations are uttered directly by the speaker. In (6.4.42) "Andy [...] think that probably there's a message in it so it would be *really interesting* to open it" is the character that thinks and in (6.4.43) the inscription on the golden fish statue is interesting and "makes the character completely interested in reading" it. On the contrary, in the other occurrences it is the speaker himself or herself who evaluates directly a place or an experience:

(6.4.44) <B> [...] one of the most beautiful things I saw was the great station that (uh) (uh) is a really (uh) really big and really (uhm) well (uh) (uh) is a really (uh) really big and really (uhm) well *really interesting* also from a point of view of history (uhm) [...] </B>

(6.4.45) <B> [. . .] I also (mm) studied Russian was which was something really different from what I used to study in high school and (uhm) it's also a satisfaction about it because (mm) it's something really also distant from our language so when I finally understand what it's written I'm really pleased and (uhm) also everything about (uh) tourism that I studied was *really interesting* and I'm really proud of myself because I also succeeded in passing exams of (mm) economy which wasn't something that I expected from myself [. . .] </B>

(6.4.46) <B> [. . .] I ate where Russian people usually eat so it's not like a restaurant (uh) it's more like (uhm) (uh) I don't know the word in English (uh) (uhm) <Italian> (uhm) so and I tried Russian food and it's kind of very very different from ours but it's tasty it's homemade and (uhm) then I went to different museums for example in Saint Petersburg I went to the Hermitage (uhm) I did do different things and that was *really interesting* </B>

In all these three cases, the speakers are describing experiences, in (6.4.41) a trip to New York, in (6.4.42) the experience of studying at university, and in (6.4.43) a study-abroad experience in Russia. Thus, they are expressing personal judgements on these experiences that they have lived, rather than imagining what a character would feel. In terms of L1 congruency, no other pattern was identified except that of the L1 equivalent to occur in positive sentences and expressing a quality on a variety of topics.

In LOCNEC there are nine occurrences of *really interesting* uttered by seven different speakers. Two speakers use the collocation twice (EN012 and EN016) and the preference for the position of the intensified collocation is in predicative position; while there is only one occurrence of attributive function (as expected in spoken language according to Biber at al., 1999, p. 518). As regards the tense used in the sentences containing the collocation, in the majority of the cases (five) the tense is in the past, whereas there are three instances of the present simple and one occurrence of future. All sentences are positive and contain no negatives. Furthermore, there is a variety of structures in terms of pronouns, subjects, and impersonal constructions (e.g., *pronoun + copular verb + collocation; noun + copular verb + collocation; it + copular verb + collocation*). There is one occurrence of an *it* pronoun referring to a page of a university prospectus and there is one occurrence of a *which* pronoun referring to cultural freedom. There is one case of explicit subject:

(6.4.47) <B> [. . .] (er) of that costume and that setting was (er) . was *really interesting* </B> <B> yeah </B> <B> it was at i= [. . .] </B>

And one instance of a *there + to be* sentence:

(6.4.48) <B> [. . .] but . they're nice </B> <B> but (erm) there are *really interesting* places to go and see as well .. we went [. . .] </B>

There are also three examples of an impersonal *it-* structure, such as in:

(6.4.49) <B> [. . .] <overlap /> started off in Kenya and then I travelled down to Zimbabwe through <overlap /> a few other countries </B> <B> so it was quite it was *really interesting* it was (em) it was <XX> land truck so twenty of us on a truck [. . .] </B>

Lastly, and more interestingly, there are two instances which are built around the pronoun *that* and the impersonal *it* and are preceded by the verb *find*, as in:

(6.4.50) <B> [. . .] there's a difference in . in attitude I find (erm) the way my dad described it to me was .. that they look at . somebody who can build a boat is as good as somebody who can be a a famous pop star so it's it's like the[i:] idea of the sort of value of on . that they put on what people d= </B> <B> do </B> <B> (erm) .. so I find that *really interesting* .. and there's there's the sort of .. I don't know cultural difference I mean that people sort of <X> Amsterdam as a sort of sex and drugs and <overlap /> rock and roll sort of thing </B>

The speaker chose to construct the predicative sentence and attenuate the expression with the verb *find* preceding the pronoun *that*. This also happens in one other case.

(6.4.51) <B> [. . .] <overlap /> so I took a year off and . I was working in a fish restaurant . and . it was . it was very interesting because although I've spent a lot of time at boarding school so away from my parents . I'd never been to America before . (erm) and so: I found it *really interesting* getting to know the people . and having to live in a completely foreign country on my own . <overlap /> (erm) [. . .] </B>

A similar construction is also found in another sentence, where between the pronoun *which* and the collocation there is a similar softening expression, preceded by a hedge (e.g., [. . .] a *sort of* cultural freedom which I find *really interesting* as well [...]).

6.4.4 Really nice

Nice is an evaluative descriptor adjective borrowed from Old French (see Table 6.1) which occurs more than 200 times per million words (Biber et al., 1999, p. 512). According to the WFF, the word belongs to the A1 CEFR level. As already described above, *really* is a derivation of Latin *realiter* (see Table 6.1), and belongs to the A1 CEFR level according to the British Council's WFF. Thus, in the case of this collocation the components share a

mixed etymological origin, as, although they both share Latin roots, *nice* has been subject to a meaning shift which has detached the contemporary word form from its original meaning. Thus, the semantic development of *nice* means that the learners cannot be aware that the contemporary meaning derives from quite the opposite ("fool", see Table 6.1). In this case, the etymology of the adjective should not be regarded as an element of help for L1 positive transfer, although, as displayed in Table 6.2, the collocation is L1 congruent.

This collocation is very frequent, occurring more than 20 times per million words in American and British English (Biber et al., 1999, p. 545) so that it is expected that the learners will have mastered its use. Only females use this collocation and both learners with an experience abroad and no experience at all have produced it. The majority of learners who have used this collocation are interestingly equally split between undergraduate courses (n° 1, 2, 3) and postgraduate ones (n° 5, 6, 7). However, the L3 languages are limited to French, German, and Spanish, contrary to previous cases.

There are eight instances of *really nice* in ISLC which are found mainly in predicative position (five):

(6.4.52) <B> [...] or I don't know but (uh) it (mm) (uh) it was *really nice* (uh) also what we did because we visited a lot [...] </B>

(6.4.53) <B> [...] city (uh) I don't know but then the people were (uh) *really nice* (uh) really friendly (uh) they helped you if they [...] </B>

There are only three instances of attributive function in ISLC for the collocation and in two cases *really nice* modifies the noun *people*:

(6.4.54) <B> [...] more relaxed relaxed from what I've seen (uhm) they're *really nice* people even though (uhm) stereotypes don't say [...] </B>

(6.4.55) <B> [...] knowledge that Thailand is a really welcoming country with *really nice* people (uhm) but I wasn't expecting it to be like [...] </B>

(6.4.56) <B> [...] out of place otherwise she can find (uh) (uh) so many flowers *really nice* flowers animals and she's absolutely she [...] </B>

As regards the tense, the learners employ a combination of present tense and past simple, such as in (6.4.52) with "it was really nice" or in (6.54) "they're really nice people". There are four instances of present simple and four instances of past tense. All sentences are positive and all instances of the collocations are used to express a positive judgement.

The most interesting aspect about the use of this collocation by the ISLC learners is the construction and the modifying function. In the five occurrences where the collocation is

used in predicative position, *really nice* is constructed around a noun in two cases (6.4.57), a personal pronoun in one (6.4.58), and the *it-* structure in two (6.4.59):

(6.4.57) <B> [...] city (uh) I don't know but then the people were (uh) *really nice* (uh) really friendly (uh) they helped you if they [...] </B>

(6.4.58) <B> [...] to study Spanish at Salamanca university and (uhm) they were *really nice* as well and (uhm) (uhm) (mm) . maybe the fact that [...] </B>

(6.4.59) <B> [...] their characters because they can see it so it's it's *really nice* I have fun </B> <B> yes </B> <B> ok [...] </B>

In (6.4.54) the explicit subject is *people*, in (6.4.55) the subject is the personal pronoun *they* and in the last extract the sentence is built around the impersonal *it-* structure. It is the use of these nouns and pronouns that lies the interesting fact about the use of the collocation. In the majority of the cases (five instances) the collocation is modifying, either in attributive or predicative position, the noun *people*, *colleagues*, or simply the personal pronoun *they* which refers to a category of people. This highlights two patterns: the collocation is mainly used in association with a noun in the plural form, unless it is making reference to an experience (6.4.52) (6.4.59); and the collocation is used by the learners to modify categories of people. Specifically, in the corpus, the learners refer to people (namely citizens of Canada and Thailand), colleagues, or family members. There is only one case in which the collocation is modifying flowers and this occurs during the story-telling task (6.4.56). In terms of L1 congruency, the L1 equivalent is used to describe a wide range of topics, but it shows a preference for qualifying people (especially *ragazzo*, *ragazza*); in addition it is almost always excusively used in the singular form.

In LOCNEC, this collocation is instanced 30 times by 16 different speakers, which means that almost all speakers use it on average twice. This initial statement is in stark contrast with the previous analysis on ISLC since only approximately 23% of the learners used it as opposed to 30% of the native speakers. The greater number of instances leads to the hypothesis that there will also be greater variation and flexibility in terms of usage.

The preferred function of the native speakers for the collocation is the predicative one with 24 instances of it compared to the attributive one which is instanced only six times (compatibly with spoken language). In terms of tense used in sentences containing the collocation, the native speakers make a similar use of the present and past tense. There are 16 occurrences of present tense sentences and 10 occurrences of past tense (four sentences were excluded since the collocation was part of an ellipsis). All occurrences of *really nice*

are instanced in positive sentences and the collocation is used to express a positive judgement. The preferred construction for the use of the collocation is *pronoun + copula verb + collocation*, such as in:

(6.4.60) <B> [. . .] her friends from her high school .. and they were all *really nice* </B> <B> and really made the[i:] effort [. . .] </B>

(6.4.61) <B> you know family seaside holidays </B> <B> but that's *really nice* too that's something <overlap /> <XX> [. . .] </B>

There are variations of this structure in *noun + copula verb + collocation*, like in:

(6.4.62) <B> [. . .] <X> </B> <B> <overlap /> the[i:] Opera House is *really nice* </B> <B> <overlap /> (erm) </B> <B> yeah [. . .] </B>

Or *pronoun + indeterminate article + collocation + noun* if the collocation is used in attributive position:

(6.4.63) <B> [. . .] to Stuttgart </B> <B> yeah (erm) . also that's a *really nice* city as well <overlap /> .. yeah [. . .] </B>

This structure and its variations are the preferred choice for the native speakers, since they instance it in 20 cases. Another type of structure used by the learners is the impersonal *it*-one:

(6.4.64) <B> [. . .] something like that but it was it *was really nice* </B> <B> yeah </B> <B> <overlap /> yeah [. . .] </B>

This happens in five cases only and it is used to refer to experiences. The remaining occurrences display different types of structures, such as the *make + object pronoun + verb* in (6.4.67):

(6.4.65) <B> [. . .] so he does it again . and make her look *really nice* so I presume this is in her house [. . .] </B>

Or the structure *end up with + collocation + noun* in (6.4.66):

(6.4.66) <B> [. . .] been round their houses some have ended up with *really nice* well decorated houses </B> <B> some have ended [. . .] </B>

Lastly, the collocation is used to express a positive judgement mainly regarding places, experiences, and people. Indeed, there are 12 cases in which *really nice* is making reference to a place which can be a city (6.4.63), a region:

(6.4.67) <B> [. . .] stay just to the north .. (er) which is it's *really nice* <overlap /> and </B> <B> there isn't much [. . .] </B>

or a country:

(6.4.68) <B> [. . .] England's not a nice country but Canada was like *really nice* </B> <B> you know it was <overlap /> really [. . .] </B>

This reference to places seems to be the most frequent instance of the collocation. The second most frequent reference (eight cases) is to experiences which can vary from volunteering to help stage manage concerts:

(6.4.69) <B> [. . .] I'm supposed to be helping out with the concerts I'm supposed to be stage managing them . I don't believe I volunteered for it actually <overlap /> but I did and it <X> it was *really nice* actually </B> <B> because I said to the[i:] [. . .] </B>

or sharing a house with friends:

(6.4.70) <B> [. . .] yeah but I mean I've I've got a house with a few friends this year which is *really nice* .. and it's quite cheap rent at the moment as well so <overlap /> that's good [. . .] </B>

or taking drinks or food inside a multiplex room. The third most frequent reference (seven cases) is made to people, so there are examples of a single subject:

(6.4.71) <B> so y= yeah . that was .. <X> so .. so (mm) that that that that . got me pretty annoyed </B> <B> because because<?> <X> I'd just met this *really nice* girl and . I wasn't able to [. . .] </B>

or multiple subjects, such as friends:

(6.4.72) <B> [. . .] yeah <laughs> but I mean most of the photos I've got are of people I have to admit cos like </B> <B> <laughs> all all my friends were there and then .. (erm) the girls who I was staying with I've got photos of them cos they were all *really nice* [. . .] </B>

The remaining references are made to a painting, food, and birds.

6.4.5 Totally different

Different is a common classifier adjective occurring more than 40 times per million words in spoken language (Biber et al., 1999, p. 440) and categorised as belonging to the A1 level of CEFR for learners of English by the WFF. Furthermore, the adjective derives from Latin *differentem*, thus it seems plausible that Italian learners of English may be positively influenced in their use of this adjective (see Table 6.1). *Totally* is an adverb stemming from Latin *totāliter* and Old French *totalement* and thus shares the same etymology with Italian. The word belongs to the A2 according to the WFF. The collocation is used throughout different groups of learners, both females and males, both learners with a stay-abroad experience (of either less or more than six months) and no experience at all, and a variety of university courses. The L3 languages spoken by the learners are Chinese, French, German and Spanish. The collocation is also L1 congruent since the Italian equivalent is *totalmente diverso.*

Totally is more frequent in American English (occurring more than 100 times per million words) so that the collocation may be more American English rather British English. There are six instances of *totally different* in ISLC and all of them see the collocation in predicative position, except for one in attributive position.

Excerpts (6.4.73) and (6.4.74) show an example of each function:

(6.4.73) <B> [. . .] Munchen and I realised that the southern part of Germany was *totally different* . they were really polite, they liked talking [. . .] </B>

(6.4.74) <B> [. . .] but there it was like it was another place it was a *totally different* universe nothing (uhm) particularly related [. . .] </B>

In terms of tense, the collocation is used both in the present tense (6.4.75) and in the past (6.4.76):

(6.4.75) <B> [. . .] Tel Aviv then it is the (uh) the modern a modern city there *totally different* from Jer= Jerusalem but I found it [. . .] </B>

(6.4.76) <B> [. . .] changed my life (uh) for the best (uh) I fe= I felt *totally different* when I came back and coming back is= [. . .] </B>

All sentences are positive and do not contain any negatives, however the collocation is not only used to express a mere difference, but rather it has other multiple functions. Due to the similarity of *totally different* with *completely different*, the same functions of the latter were reprised and identified in the former (see Section 6.4.1). The main three functions

are: comparison (comparison of two terms which can be positive, negative, or neutral), expectation (something turned out to be different from what was expected and it can either be positive or negative), and opposites (marking the difference between two terms which are in fact opposites). As far as *totally different* is concerned, it was possible to identify two of these functions, comparison and expectation:

(6.4.77) <B> [. . .] we stayed for three days in Jeru= in Jerusalem and we (uhm) we visited the the old cities (uhm) the old city and then we stayed for (uh) I think two days in Tel Aviv then it is the (uh) the modern a modern city there *totally different* from Jer= Jerusalem but I found it I found it really good (uh) this (uh) this city (uh) really impressed me [. . .] </B>

(6.4.78) <B> [. . .] we passed throughout all Germany because we were travelling using the car and we had to stop one night in a little town called Rothenberg it is in the same land of (mm) Munchen and I realised that the southern part of Germany was *totally different* . they were really polite . they liked talking to people from other countries (uh) the food was better tha= a lot better I ate a strudel and it was amazing really amazing [. . .] </B>

(6.4.79) <B> [. . .] I was very stressed for exams and all that stuff and I went there and it seemed like it was a paradise like there was no problem with like it was the period when Trump was about to be elected but there it was like it was another place it was a *totally different* universe nothing (uhm) particularly related to the rest of the world was happening it was just there living that moment and so I went there [. . .] </B>

These three extracts show instances of positive expectation since the learners first describe the location (interestingly they all refer to places) and compare it to another with a tone of unexpectedness. For example, in (6.4.77) the learner is comparing Jerusalem and Tel Aviv and she is saying that although Tel Aviv was *totally different*, she enjoyed it. The same happens in (6.78) where the learner is recounting a trip to Germany which she did not enjoy except for the last part in the Bavaria region where she stopped in Rothenberg. Southern Germany is described as unexpectedly different, people are polite, the food is better. Her expectations were clearly pointing towards a different direction and she was instead surprised. The last extract behaves similarly since the speaker is describing a summer holiday in Portugal and she was not expecting to find such a beautiful and relaxing place.

The second function that was identified in this sample is that of the comparison. There are two instances of positive comparison:

(6.4.80) <B> [. . .] I think that I did some things that I would have never thought I would do (uh) without going there so (uh) that really changed m= my life I always= I always say that that experience changed my life (uh) for the best (uh) I fe= I felt *totally different* when I came back and coming back is= is the hardest part a little bit because you're getting used to that kind of mentality [. . .] </B>

(6.4.81) <B> [. . .] so my friend was driving and when my friend was driving I didn't even notice she fell from the sled so I was sitting there and the dogs were running yeah but I didn't realise and she managed to catch the sled again so I was safe and that was for sure the best trip ever because it's (uh) a country which is *totally different* from what I've seen so far I have I've already had some journeys all around the world but never in such a place and it was great and I love snow so I was in the right place [. . .] </B>

and one instance of neutral comparison:

(6.4.82) <B> [. . .] obviously talking about China it's an amazing country which (uh) a wide story so (uh) thousands and thousands years of history so while (uh) wh= you can visit (uh) incredible (uh) landscapes and and works so (uh) talking about their (uh) culture historical (uh) background it's an incredible country but talking about their (uh) (uh) everyday culture it's *totally different* (uh) compared to the Italian one so (uh) I'm talking about politeness because (uh) people there have a very different behaviours (uh) in in every (uh) in every asp= aspect of (uh) everyday life [. . .] </B>

In (6.4.80) the speaker is describing her year abroad in the US and saying how this experience changed her life for the best. The key to the positive comparison is for the best since she came back *totally different* from the experience, a different person, a better person thanks to all the great experiences she had. In (6.4.81) the speaker is narrating a trip she took to Norway during the winter and how she went sledding with a friend. She describes the country as *totally different* from the other countries she visited and says that "I've already had some journeys all around the world but never in such a place", thus indicating that Norway is the best trip she had ever had. In (6.4.82) the comparison becomes purely neutral as the speaker is talking about his experience abroad in China. The collocation refers to the Chinese culture, which is *totally different* from the Italian one, but without any specific positive or negative judgement. Indeed, "people there have a very different behaviours in every aspect of everyday life" (sic), which does not point towards a positive or negative connotation, but merely a genuine difference between the two cultures.

The collocation only occurs once in LOCNEC:

(6.4.83) <B> and he paints her picture and he paints it . realistically </B> <B> <laughs> and he paints it as he sees it and she doesn't like it .. she doesn't like what she sees so he does it again and he makes her look . *totally different* makes her look very glamorous <overlap /> this isn't it's nothing like her [. . .] </B>

The collocation is used in the structure *make + object pronoun + verb + collocation* and its main function is that of marking the (neutral) difference between the woman and the painting. The lack of occurrences may be attributable to the fact that *totally different* is chiefly American and the speakers in LOCNEC are all British university students.

6.4.6 Very different

As reported above, *different* is a common classifier adjective, categorised as belonging to the A1 level of CEFR for learners of English by the WFF. *Very* is a common intensifier which shares the same root with Latin (see Table 6.1). The adverb is categorised as an A1 word by the WFF and it is L1 congruent (see Table 6.2). Both female and male learners use the collocation and only learners with a short stay-abroad experience produced it. There seems to be no correlation between the learners' university courses and this collocation, and all L3 languages match the production of the collocation except French. Among the most frequent collocations found in ISLC, this is the third one which includes the adjective *different* as its component. The collocation is L1 congruent.

The learners produce 11 occurrences of the collocation and this is used both in attributive and predicative position: there are six instances of attributive function and five instances of predicative. The excerpts below show an example of predicative and attributive respectively:

(6.4.84) <B> [. . .] I learnt a lot of (uh) things about their culture which are *very different* from Italy and I think this will help me also [. . .] </B>

(6.4.85) <B> [. . .] me lessons because for example (uhm) the fact that even two *very different* people can become friend friends and they (mm) [. . .] </B>

There is a preference for using the collocation in present tense sentences with only two instances of past simple, such as in:

(6.4.86) <B> [. . .] I have been in many places in China and (uh) and Beijing was *very different* different from the other places because I [. . .] </B>

One other aspect that has been noted in the analysis is the use of the preposition *from* in three cases. All the instances of the collocation are positive sentences and, as with the

other two collocations containing *different*, the collocation performs distinct functions. The functions of the collocations have been previously identified as comparison (comparison of two terms which can be positive, negative, or neutral), expectation (something turned out to be different from what was expected and it can either be positive or negative), and opposites (marking the difference between two terms which are in fact opposites). However, in the case of *very different*, a further function was identified, that of *listing*. The listing collocation performs the function of introducing a list of several different items that the narrator wants to describe. Excerpt (6.4.87) below shows the only case of listing found in the corpus:

(6.4.87) <B> [. . .] my favourite place is Central Park (uhm) because (uh) you can see (mm) *very different* things you ha= for example at seven in the morning when we visited the first time . yes because (uh) we had the jet lag we . so it's quite difficult and (uhm) at seven a.m. you see people doing yoga people running people just walking with their dogs or also people eating their breakfast at some (uhm) little bars or kiosks that were there and well it's it's something that you can't= we cannot imagine here in Italy [. . .] </B>

In this case, the speaker is narrating a holiday in New York where she and her boyfriend, very early in the morning, took a walk around Central Park and were able to see *very different* things. The things are: people doing yoga, people running, people walking their dog, people eating breakfast. This is a list of four different activities that the speaker was able to witness during her walk. The activity are indeed different from one another, but it appears that the function of the collocation is not so much that of stressing the difference among the activities, but rather introduce a list of multiple items which are different.

The other instances of the collocation perform the comparison, expectation, and opposites functions. The pragmatic function of comparison is found in the majority of the sentences (five) and in four cases the comparison is neutral, such as in (6.4.88):

(6.4.88) <B> [. . .] it's not just something I read on the books in the books but it's something I can actually experience so by going there by talking to people keeping in touch with them (uhm) watching watching movies in Russian it's still something very difficult but I can try (uhm) listening to Russian podcasts and it's like so yeah it's actually a culture *very different* from ours but very interesting [. . .] </B>

In this excerpt, the speaker is describing her stay-abroad experience in Russia where she got to live among Russians and experience life in the country. Her final comment describes the Russian culture as *very different* to the Italian one, but no other evaluative comment is made following this and the sentence is merely stating a fact, a difference between the two

cultures. Furthermore, the postmodification of the collocation is a clear L1 transfer ("una cultura molto diversa dalla nostra") and could be said to be rather unnatural.

There is one instance of negative comparison:

(6.4.89) <B> [...] the food (mm) yeah (uhm) also the food if I think about that (uhm) because it's *very different* from (mm) what I'm used to eat here in Italy (uhm) but it was not that bad because (uhm) for example once we went to the famous Hard Rock Cafe so (uhm) we I think everyone can enjoy a really good hamburger [...] </B>

This is a reply to the interviewer's question "What about the food?" which followed the previous question "What's the worst part about London?". The speaker is thus trying to find negative aspects about her holiday in London and as soon as food is mentioned her answer is "yeah also the food", meaning that food, together with the city's underground, is probably the worst element about London. Thus, the comparison between English food and Italian food takes a negative note as regards English food, which is *very different* from the Italian one. The other function performed by the collocation is the opposites one. There are four cases of this:

(6.4.90) <A> so how did Irish food compare with Spanish food? </A> <B> (uh) (uh) it's *very different* because well Ireland is a cold country so they have a lot of soup for example and especially fish soup with they usually serve it with black= black bread and butter and well they cook a lot of potatoes as well (uh) I had potatoes every every day there basically baked potatoes (uh) or smashed potatoes (uh) but in Spanish they cook well paella [...] </B>

(6.4.91) <A> which one did you prefer Irish or Spanish cuisine? </A> <B> (mm) I I think they're (mm) there's no winner I appreciate both of them because they are they are *very different* so (mm) there is no comparison actually for both of them </B>

Both extracts come from the same text and speaker, who is talking about a study holiday in Ireland. The interviewer asks the speaker to compare Irish and Spanish food and the speaker immediately says that the food is *very different* between the two countries, because one is a cold country (Ireland) and the other is (implicitly) a warmer country. The food reflects the climate with Ireland having traditional soups and potatoes as a staple food, while Spain paella and other fried dishes. When the speaker is further pressed by the interviewer to choose her favourite cuisine between the two, she replies with "there's no winner", because they are *very different*, the opposite in fact, and she stresses that "there is no comparison", thus effectively highlighting the fact that *very different* is pointing towards opposite cuisines, rather than just different. There is one instance of the expectation function:

(6.4.92) <B> [...] was (uh) a very (uhm) particular experience because I went to a *very different* country [...] </B>

Lastly, it is worth mentioning that the collocation mainly modifies and refers to nouns belonging to the semantic category of culture (four cases), food (three cases), and places (two cases). The remaining references are made to people and activities. In terms of L1 congruency, a random reproducible sample of 20 occurrences of the L1 equivalent has shown that the collocation primarily performs the comparison and the opposites functions. There were no instances of expectation or listing.

In LOCNEC, the collocation occurs 15 times and it is used almost exclusively in predicative position, while there are only three cases of the collocation in attributive position. The preferred tense is the present since 13 sentences are built in the present, whereas only two contain a structure in the past. As regards the prepositions used to the right of the collocation, we observe that, as opposed to the learner corpus, the preferred preposition is *to* (three cases). There is one occurrence of *from* and one occurrence of *for*, used to specify to whom the difference was relevant. The use of the preposition *to* is very frequent and unique to British English (Leech et al., 2002, p. 85), which is why it is found in LOCNEC, since it is a corpus of British native speakers. In terms of nouns associated with this collocation, these are the following: *style*, *experience*, and *attitude*. Experience and attitude may be grouped under the same semantic category of lifestyle/culture, whereas *style* refers to how a writing course is done and therefore may generally be categorised as other topics. However, if the nouns and/or topics post-modified by the collocation in predicative position are taken into consideration, the references expand to places and people too. In particular, places such as Belgium, Kent, and Durham appear in the sentences in question and are the terms of comparisons.

The only function of *very different* in LOCNEC is that of comparing, either implicitly or explicitly, two or more objects. For example:

(6.4.93) <A> what did you do </A> <B> I was an insurance broker <overlap /> <laughs> </B> <A> <overlap /> oh <laughs> </A> <B> yeah </B> <A> it's quite different </A> <B> yeah it's all *very different* I know and I just .. I quite like children [...] </B>

In this case, the interviewee is being asked what job she did before becoming a teacher, to which she replies that she was an insurance broker, which is *very different* from a primary school teacher. No positive or negative judgement follows the statement and therefore this collocation can be considered a neutral comparison. Another example of neutral comparison is the following:

(6.4.94) [. . .] yeah around exam time I think you get people <X> other people's tensions and stress as well <X> also just . I think I feel I'm a bit in a different position now being a postgrad student and I don't want to be surrounded by undergrads who are going through a *very different* experience [. . .] </B>

Here the speaker is comparing the experience of being a postgraduate student to that of being an undergraduate. The two are *very different*, but there is no positive or negative connotation to the comparison, the speaker is simply stating a fact. There is an abundance of *it* + *copula verb* + *collocation* structures (nine occurrences), five of which are impersonal *it*- structures.

6.4.7 Very good

Good is a descriptor adjective frequent in conversation, occurring more than 200 times per million words according to Biber et al. (1999, p. 512). It stems from Old English and thus has no shared etymology with Italian, although the adjective belongs to the A1 CEFR level according to the WFF. As mentioned above, *very* is a common intensifier categorised as an A1 word by the WFF and shares the same L1 etymology. The collocation is L1 congruent as per Table 6.2. The collocation is used by female and male learners, learners with stay-abroad experience (both less and more than six months), and attending a variety of university courses. All L3 languages are involved in the production of the collocation except Russian. The collocation is extremely frequent, occurring more than 100 times per million words in British English and more than 50 times per million words in American English (Biber et al., 1999, p. 545).

There are 14 occurrences of *very good* in ISLC, eight of which are found in predicative position, while six are in attributive position. Below is an example of a subject predicative, which is the preferred position of the predicative collocations:

(6.4.95) <B> [. . .] uhm s= I used to write short poems but they're not *very good* and (uhm) I obviously like going out with my friends [. . .] </B>

In this particular case, the collocation is complemented by a negative, which, as it has already been mentioned, could be considered a linguistic hedge. Below is an example of an attributive function:

(6.4.96) <B> [. . .] stressed I really detached from everything and I had a *very good* summer (uh) it was the second experience I was doing [. . .] </B>

In three cases the intensified predicative adjective complements a copular verb and occurs without a complement, as it is expected in conversation (Biber et al., 1999, p. 518), but in four cases it occurs with a phrasal complement, such as a prepositional phrase:

(6.4.97) <B> [. . .] had a cough and thr= throat= throat ache so it was not *very good* for us yes [. . .] </B>

(6.4.98) <B> [. . .] worst one meat because I don't like it I'm not *very good* at cooking it </B> <B> ok </B> here is a list [. . .] </B>

As regards the tense, there are eight occurrences of the collocation in the present tense – such as (6.4.98) – and four occurrences in the past tense – such as (6.4.97). One occurrence was excluded from the count as it is an infinitive sentence. Previous collocations showed a pattern of positive sentences with a lack of negatives; this is not the case for *very good* which occurs seven times with the negative:

(6.4.99) [. . .] whom I run with yeah I like cooking I'm not a *very good* but I'm improving and of course listening to [. . .] </B>

It is interesting to note that in all seven cases of negation, the speaker is talking about an ability or talent (6.4.99) and is employing the negative as a potential hedging strategy to attenuate the statement. The following excerpt is another example of negation:

(6.4.100) <B> [. . .] well my main hobby is drawing (uh) I'm not *very good* at it but I mean I only do it for myself so I don't really care (uhm) I have a graphic tablet I draw on my comp= with my computer and I'm trying to learn (mm) I've been drawing for four or five years now [. . .] </B>

The speaker is clearly trying to attenuate the statement about her drawing skills, which, undoubtedly are not *not very good*, since she has been drawing for four or five years and she uses a graphic tablet (which requires a certain amount of skill). The same phenomenon occurs in the other cases, except in (6.4.101):

(6.4.101) <B> [. . .] if they are not used to this kind of (uh) place full of (uhm) trash and very (uh) polluted it's not easy as for example me and my sist= my sister and I (uh) had an attack of allergy to (uh) to dust yeah we didn't have any type of allergy but then (uh) we had to go to the pharmacy because we had a cough and thr= throat= throat ache so it was not *very good* for us . yes [. . .] </B>

This is an extract from a conversation with speaker IT011 who is remembering a trip to India with her sister. They did not particularly enjoy the experience since they struggled with the level of hygiene and developed hayfever due to the pollution. In this case, the speaker is not trying to employ a hedging strategy to attenuate the statement about an ability or talent, but rather she is relying on a safe collocation to express a negative feeling. It could

be hypothesised that she did not use *very bad* since it may have sounded too strong for her and thus she preferred to hedge the negative feeling with a negation of *very good*, or that she relied on a safe collocation. A similar occurrence happens with speaker IT014 and her sentence (6.4.102):

(6.4.102) [. . .] uh I study I study and sometimes I work (uh) in a shop that sells (uh) wedding clothes (uh) a friend of mine owns this shop and so (uh) since sometimes there are clients that are foreign because it's a *very good* shop (uh) he asks me to come there and help him out with English or French (uh) the other day there was a group of (uh) Swiss women they were from Zurich [. . .] </B>

The speaker is telling an anecdote which took place in a wedding dress shop where she worked for a short period of time. In this extract, she is explaining the presence of foreign customers as a consequence of the fact that the shop is a *very good shop*. However, the shop is not simply *very good*, but what she means is that the shop is exclusive, luxurious, well-known, and other adjectives which may better express the elitist nature of the shop. Once again, when Italian speakers find themselves short of a better adjective or collocation, they tend to rely on those safe ones which are so frequent in their repertoire. Lastly, the remaining occurrences of *very good* are used to make a personal positive evaluation about food (two instances), experiences (two instances), and university life, as opposed to the other five sentences expressing (in)ability. In the PEC corpus, the L1 equivalent translation of *very good* was found to occur always in positive sentence and no hedging strategies were employed to mitigate the expressions.

In LOCNEC, there are 50 occurrences of *very good*, which is expected due to the high frequency of this collocation. Since the adjective good can function both as a predicative and attributive adjective, the sample presents 34 occurrences of *very good* in predicative position and 15 in attributive. One occurrence of *very good* is an exclamation. In the majority of *very good* concordances in LOCNEC, the modified adjective is predicative and complements the subject and a copular verb. Although attributive adjectives are rare in conversation (Biber et al., 1999, p. 518), there are 15 instances of them in LOCNEC. The nouns modified by the collocations in attributive position are: linguistics department, friends, film (three cases), actress, story-teller, technique (two cases), cover, career, Chinese, experience, question, thing, value. They represent both specific language (linguistics department, Chinese, friends) and vague language (thing, experience), so there is a combination of specific references (mainly to university life, job, entertainment) and vague ones which encompass a broader selection of topics. As regards tense, there is no overt preference for either the present or the past, there are 27 instances of the collocation complementing a verb in the present tense com-

pared to 20 complementing the past tense (the remaining occurrences were discarded due to the fact that the collocation was either a detached predicative or an exclamation).

As briefly mentioned above, two instances of the collocations were not included in the count between present and past tense since they are part of an elliptical construction. In the case of LOCNEC, there are two instances of this (both contained in this extract):

(6.4.103) <B> [. . .] <overlap /> <laughs> (erm) have you seen Seven .. <overlap /> the new </B> <A> <overlap /> no but it's not . on this term but </A> <B> no it will be on next term with any luck </B> <A> (mhm) </A> <B> very . *very good* </B> <A> (mhm) </A> <B> *very good* (erm) . I sort of just I was spell bound from beginning to end . <overlap /> and [. . .] </B>

In this instance the speaker is commenting on the film Seven, which according to the speaker is *very good.* However, the intensified adjective is standing alone as part of an elliptical construction which has eliminated the pronoun and the copular verb (e.g., it is very good*). The other case in which the collocation is found detached from a verb phrase is in (6.4.104):

(6.4.104) <B> [. . .] <A> <overlap /> and I'd like you to tell me that story </A> <B> that story </B> <A> yes so it goes like that one two three four </A> <B> right okay right I'm with you yeah ah *very good* . okay right there's a guy and he's painting the picture of a[ei] woman .. and he paints the picture .. exactly as he sees the woman </B>

In this case, the interviewee is struggling to understand the third task of the picture description, so the interviewer intervenes and explains the order of the pictures for the narration. The interviewee understands and confirms having understood the instructions. *Very good* acts as a discourse marker rather than a modifying collocation for a subject or object of the sentence. Similarly to ISLC, there are ten occurrences of the collocation in a negative sentence. In four cases the negation is found in a sentence where the collocation is evaluating a personal skill or ability:

(6.4.105) <B> [. . .] and not working </B> <B> <overlap /> and I'm not *very good* at things like <overlap /> that I have to work [. . .] </B>

(6.4.106) <B> [. . .] well yes I suppose it is really . I'm not *very good* at it and I'm more interested in [. . .] </B> i) <B> [. . .] <overlap /> I wa= I wasn't. I wasn't very good at getting to nine o'clock lectures when I [. . .] </B>

It can also be found with the adverb *never*, such as in the following:

(6.4.107) <B> [. . .] but . (er) I was I never v= <X> was *very good* at (eh) .. critical analysis [. . .] </B>

This sentence was counted as a negative due to the presence of the adverb *never* which negates the fact that the speaker was *very good* at critical analysis. The rest of the negative occurrences are mainly hedged statements through epistemic verbs, such as in:

(6.4.108) <B> [. . .] yeah </B> <B> I don't think the library's *very good* anyway they don't have many copies of .. [. . .] </B>

This is a case of hedging strategy since the speaker is preceding the statement with the verb *think* while concurrently using the negative to downplay the impact of the claim. All the positive occurrences of the collocation are used to express a personal positive evaluation.

6.4.8 Very important

Important is an evaluative adjective which is most common in academic writing (over 200 times per million words) and in news language (more than 80 times per million words), but it is not as frequent in conversation (Biber et al., 1999, p. 512). According to the OED and Table 6.1, the adjective is probably a borrowing from Latin and it belongs to the A1 CEFR level according to the WFF. *Very* has already been described above and can be referred to in Table 6.1. It also belongs to the A1 CEFR level according to the WFF. Both female and male learners produce this collocation, as well as learners with a short or long experience abroad. Both undergraduate and postgraduate learners use *very important* and in terms of L3, French, German, Russian, and Spanish are the other languages spoken by the learners. The collocation is L1 congruent as per Table 6.2.

There are eight occurrences of *very important* in ISLC compared to a single occurrence in LOCNEC. As *important* is an adjective rarely found in conversation, the sole occurrence of LOCNEC is more predictable than the multiple instances found in ISLC. *Important* is both an attributive and predicative adjective and in ISLC is used three times in attributive position and five times in predicative:

(6.4.109) <B> [. . .] of Liberty Island and (uh) Ellis Island too in= in that is a *very important* (uh) place in which immigrates were brought [. . .] </B>

(6.4.110) <B> [. . .] with my courses I study Russian here in Italy so it's *very important* for me to follow a lesson in a regular way in [. . .] </B>

The attributive position of the intensified adjective refers to place in (6.4.109), social topics, cases, "founding", "lesson". All these nouns belong to different semantic domains so that there is not a preferred reference for the use of *very important*. The three instances of the collocation in predicative position all complement the subject of the sentence which is an impersonal *it*- structure, and are followed by a clause complement:

(6.4.111) <B> [...] experience and what I learned wa= also is (uh) that it's *very important* to adapt (uh) to whatever situation you find [...] </B>

(6.4.112) <B> [...] a company which deals with the Germany so I think it's *very important* to know also about the culture of a country [...] </B>

(6.4.113) <B> [...] with my courses I study Russian here in Italy so it's *very important* for me to follow a lesson in a regular way in [...] </B>

In all these three cases it is possible to notice that the clause complement following the collocation is an infinitive clause. As regards the tense, almost all occurrences are in the present tense, with one exception in the past tense (6.4.114):

(6.4.114) <B> [...] uhm the discovery stayed a mystery but the book was (uhm) a *very important* (uhm) founding in the (uhm) (uhm) for the history [...] </B>

All sentences are positive and there are no negatives; let us look at (6.4.109) and (6.4.112): in excerpt (6.4.112) the speaker is stating that in the future she would like to work in a company which has contacts with Germany and therefore a knowledge of the country's culture is *very important*.

As mentioned at the beginning of this analysis, there is only one occurrence of *very important* in LOCNEC. This is not unexpected since the adjective *important* is not common in conversation. Below is the only instance of *very important* found in LOCNEC:

(6.4.115) <A> <overlap /> well behind <laughs> </A> <B> yes well there's a Spanish house <starts laughing> there's a Spanish streets <stops laughing> and things <overlap /> like that </B> <A> <overlap /> Spanish people </A> <B> yeah <laughs> *very important* [...] </B>

In this sentence, the collocation occurs as an elliptical construction in response to the interviewer's comment on Spanish people. The interviewee replies that they are indeed *very important*, which is why they appear in the pictures rather than Spanish landscapes. The restricted use of *very important* can only show a predicative function and an elliptical construction.

6.4.9 Very interesting

Interesting is a common evaluative adjective which occurs at least 10 times per million words (Biber et al., 1999, p. 530) and can be found both in attributive and predicative position. *Interesting* stems from Latin *interest* (see Table 6.1) and the adjective belongs to the A2 CEFR as indicated by the WFF. *Very* has already been defined as an adverb sharing the same etymology as Italian (see Table 6.1) and belonging to the A1 CEFR level. The collocation is L1 congruent (see Table 6.2). Only female learners produce this collocation and they have short stay-abroad experiences; they all attend postgraduate courses and they speak French and German as L3s. The hypothesis is that due to the shared etymology with the word *interessante* in Italian and the congruence of the collocation, the learners will perform fluently and use this collocation with a good degree of flexibility and variation.

The collocation occurs seven times in ISLC and it is used only in predicative position, thus reflecting the scarcity of attributive adjectives in spoken language (Biber et al., 1999, p. 511). Since the number of occurrences is limited, all excerpts will be reported below:

(6.4.116) <B> [...] it was my first time alone in a country for so long and there I went to university (uh) I (uhm) I studied a lot of new (uh) subjects that I don't study here (uh) and (uh) it was *very interesting* because (uh) the (uh) university there is completely different (uh) we had a lot of different projects [...] </B>

(6.4.117) <B> [...] in particular I had chosen the (uh) food topic and in particular the (uh) problems in poor countries with food and (uhm) it was *very interesting* and for me because I learnt how to (uh) do a presentation how to (uhm) talk in front of people [...] </B>

(6.4.118) <B> [...] they absolutely want to find out (uh) which city is shown (uh) also because (uh) there's an ex= an ex on the map and maybe (uh) they can find something *very interesting* there so they (uh) try to understand what city is represented there [...] </B>

(6.4.119) <B> [...] they start looking for something (uh) they don't know what they will find but (uhm) maybe they will find gold or something something very precious or something *very interesting* (uhm) they really want to (uh) find that thing and so they start looking into the (uhm) (uh) woods [...] </B>

(6.4.120) <B> [...] and in this sense (uh) the way in which they integrate (uhm) modernity and contemporaneity and ancient times <und> this type of things (uh)

well it's *very interesting* because you can s= you can see a completely (uhm) modern (uh) building and a skyscraper near one of the most ancient church [. . .] </B>

(6.4.121) <A> what are you watching right now? </A <B> suits the series with Meghan Markle and that is well loc= located in New York and <laughs> </B> <A> what's it like? </A> <B> (uhm) in which sense? </B> <A> how's the show? </A> <B> ah ok well it's *very interesting* is (uh) it talks about (uhm) (mm) lawyer (uh) lawyers and this law= lawyer (uhm) (uh) studium in New York and this is very famous as it deal with very important cases in the city so lots of money [. . .] </B>

(6.4.122) <B> [. . .] watching movies in Russian it's still something very difficult but I can try (uhm) listening to Russian podcasts and it's like so yeah it's actually a culture very different from ours but *very interesting* [. . .] </B>

Except for (6.4.116) and (6.4.120), the collocation is in predicative position and with no complement. In (6.4.116) and (6.4.120) instead, the collocation is followed by a clause complement which begins with *because*. In all cases the collocation complements a subject, except in (6.4.118), (6.4.119), and (6.4.122); the first two show a similar construction with the collocation preceded by the pronoun *something*, whereas in the last instance *very interesting* complements the object of the sentence which is *culture*. As regards tense, the preferred one is the present simple as it occurs in four sentences (6.4.118), (6.4.120), (6.4.121), (6.4.122); there are two instances of past simple (6.4.116), (6.4.117), and one occurrence of future (6.4.119). All sentences are positive and contain no negatives. It is worth noting that the first four sentences were produced by the same learner and (6.4.118) and (6.4.119) were instanced during the third task. Incidentally, the sentences produced during the third task show a similar structure: *they* + *find* + *something* + *collocation*. The same can be said for (6.4.116) and (6.4.117) where the collocation is preceded by *it was* and the impersonal construction is preceded in turn by a filled pause (uh) or (uhm). Another speaker, speaker IT014, produces two sentences which present similar features, (6.4.120) and (6.4.121). Both sentences are in the present tense and the collocation complements an *it-* structure, which is in turn preceded by the discourse marker *well*. The collocation mainly refers to experiences of university life (6.4.116, 6.4.117) and travelling (6.4.120), as well as entertainment (6.4.121) and culture (6.4.122). In terms of L1 congruency, the L1 equivalent translation of the collocation has been found to be practically in all cases in mid-sentence position, although it is mainly used to express an evaluation regarding opinions, project, and ideas.

In LOCNEC, there are 12 instances of *very interesting*, all found in predicative position. There are three cases in which the collocation is followed by a clause complement (*because*

+ *reason clause*):

(6.4.123) <B> [. . .] <stops laughing> but (erm) I thought it was *very interesting* because it (erm) .. highlighted the sort of [. . .] </B>

And by a relative clause complement:

(6.4.124) <B> us all thinking for well over a week and also *very interesting* that someone else we went with we went the [. . .] </B>

In (6.4.124) it is also possible to notice an elliptical construction of the collocation which is found without a copular verb (i.e., *it was*). As regards construction, the collocation always complements a subject and the preferred structure is the *it + copular verb* one. The preferred tense is the past, with eight occurrences, compared to the present which was only instanced once. In the same sentence where the present tense is instanced, the verbal expression *I thought* is present and functions as a hedging stategy. The verbal expression is also used in (6.4.127) and in:

(6.4.125) <B> [. . .] through their children and (erm) .. I just thought it was *very interesting* because of that it was (erm) .. it's [. . .] </B>

In the case of (6.4.124) and (6.4.125), the sentences are contained in the text produced by the same speaker (EN001). All sentences are positive except for:

(6.4.126) <B> [. . .] just destroying everything </B> <B> <overlap /> it's not *very interesting* </B> <B> gardening <laughs> <overlap /> [. . .] </B>

In (6.4.126), the collocation is complemented by a negative copular verb. This is used, as it has been described in previous analyses, as a hedging strategy to mitigate the strength of the statement. In this case, the speaker is telling the interviewer about a work experience in a greenhouse. The speaker was meant to keep the greenhouse clean from weeds and cutting the grass, a rather boring job, which the speaker defined as *not very interesting*.

One last pattern that should be pointed out is the fact that the collocation, in five cases, is found in end-sentence position, so that the sentence ends with the collocation. This is not the case in ISLC where in all occurrences (except one) the collocation is in mid-sentence position. This is perhaps the only distinctive feature between the two groups of learners; no other major differences were noted in the comparison.

6.4.10 Very nice

Nice is an evaluative descriptor adjective occurring more than 200 times per million words (Biber et al., 1999, p. 512) which stems from Latin (see Table 6.1). According to the WFF, the word belongs to the A1 CEFR level. *Very* has been amply described in the previous analyses as a common intensifier sharing the same etymology with Italian and belonging to the A1 CEFR level. The collocation is L1 congruent (see Table 6.2). Only female learners use this collocation and learners with either a long or short experience abroad. Both postgraduate and undergraduate learners produced the collocation and they are speakers of French, Russian and Spanish L3. Both the components of the collocation and the collocation itself are very frequent (*very nice* occurs more than 100 times per million in British English conversation and more than 20 times per million words in American English conversation (Biber et al., 1999, p. 545)) so that it is expected that the learners have become accustomed to it and have mastered its use.

All sentences extracted from ISLC are positive and there are seven occurrences of *very nice* in ISLC, six of which are found in predicative position:

(6.4.127) <B> [...] (uh) vendor to like pick up the clothes and yeah this is *very nice* this really suits you this really compliments your [...] </B>

and one in attributive position (with a negative functioning as a linguistic hedge):

(6.4.128) <B> [...] and nobody had lived there for like six months so not a *very nice* smell and the house was a mess because the flatmates [...] </B>

As regards the predicative function of the collocation, this is not followed by any phrasal or clause complement and it always complements a subject. In (6.4.129) the collocation complements a subject, but the latter is omitted and the collocation occurs as an elliptical construction:

(6.4.129) <B> [...] suggested by the <name of Italian newspaper> and I enjoyed it yeah *very nice* [...] </B>

In this case, the subject of the collocation is the same *it* of *I enjoyed it*, but the pronoun and the copular verb were omitted. As regards the tense, the learners employ both the present tense (four times) and the past tense (three times) showing no particular preference for one or the other. Below is an example of the collocation in the past tense:

(6.4.130) <B> [...] so five in total yeah with me (uh) but the university was *very nice* very big very modern (uhm) I met some incredible [...] </B>

The variety of structures used in the sentences containing the collocations is limited to *pronoun + copular verb + collocation* (a noun (*university*) is used instead of a pronoun in (6.4.130)) where the pronouns varies from *this* (6.4.127) to *they* (6.4.131) or *it* (6.4.132):

(6.4.131) <B> [. . .] but if you (uhm) get acquainted with them (uhm) then they are *very nice* they are helpful they try to get in touch with you [. . .] </B>

(6.4.132) <B> [. . .] her what should I I I found this this key it's *very nice* actually you can see it's probably really really [. . .] </B>

The references of the collocation are varied and include a dress, Russian people, work colleagues:

(6.4.133) <B> [. . .] I'm fine with it and also I like the environment (uh) my colleagues are great yeah all young people and they're *very nice* (uh) I always bother them with a lot of questions [. . .] </B>

a book and a key. As regards the last one, the reference to a key, it is interesting to notice that this collocation was produced during the third task and the association between *very nice* and *key* appears slightly unnatural. Perhaps the unnatural-sounding *very nice* reference to a key was generated by the cognitive demanding task and the pressure the speaker felt to produce a collocation starting from the adverb *very*. In PEC, the L1 equivalent of *very nice* has been found to occur mainly in relation to people, as an evaluation of the physical and personal traits of people, especially women.

In LOCNEC, there are 23 instances of *very nice* suggesting a more widespread use of the collocation in the native speakers' corpus. In 20 cases, the collocation is used in predicative position, while there are instead only three instances of the collocation in attributive position. The scarcity of *intensifier + adjective* collocations in attributive position in the corpus is due to the nature of the corpus itself, which is a spoken corpus. Predicative adjectives are mainly used in conversation (Biber et al., 1999). The native speakers show a preference for the present tense when it comes to the sentences containing the collocations, thus there are 15 occurrences of sentences in the present tense. On the other hand, the past tense occurs in seven sentences and there is also one instance of conditional. All sentences are positive, except for (6.4.134) which contains a negative:

(6.4.134) <B> [. . .] Liverpool or Manchester derby that's not *very nice* there .. but normally when it is just [. . .] </B>

As it has been previously indicated (see *Very interesting*), the use of the negative combined with the collocation is a hedging strategy to attenuate the strength of the statement. In this

case, the speaker is describing to the interviewer his main hobby, which is football. He goes to the stadium for every football match and the interviewer wonders whether it might be dangerous at times. He replies that it can be dangerous, especially at Leeds, Liverpool or Manchester stadium. Thus, going to a football match there "that's not very nice" and the use of the negative combined with an intrinsically positive collocation provides an attenuated statement about the potential dangers of being caught up in a hooligan fight. In terms of hedging strategy, the use of the negative is not the only strategy identified in this sample. There are three other examples of speakers hedging the collocation statement:

(6.4.135) <B> [. . .] Schwaebisch Alps which are sort of . yeah .. yeah *very nice* as well (eh) beautiful [. . .] </B>

(6.4.136) <B> [. . .] my corridor </B> <B> (er) seems to be *very nice* so (er) .. yeah I I quite [. . .] </B>

In (6.4.135) the collocation is preceded by a hesitating moment and by the hedging approximator *sort of* (Prince et al., 1982), which attenuates the strength of the statement about the Alps. Similarly, in (6.4.136) the speaker is describing their student accommodation and stating that the students on his or her corridor *seem*(*s*) to be *very nice*. Once again, the hedging device employed with the use of the epistemic verb *seem* diminishes the strength of the collocation in reference to the people living in the same student accommodation. The preferred sentence structure for the native speakers is *pronoun + copular verb + collocation*, indeed 14 sentences are built thusly. The pronouns range from *that*, *she*, *he*, and *everybody*.

As far as *it* is concerned, there are no occurrences of impersonal *it-* structures in this sample of *very nice* extracted from LOCNEC. Another structure is represented in (6.4.137):

(6.4.137) <B> [. . .] to buy it cos it doesn't look *very nice* . so they have all these their fruit [. . .] </B>

This is produced by the same speaker who generated three similar sentences all containing the collocation *very nice*. All the occurrences of *very nice* refer to a variety of topics ranging from clothes to football and university life. The primary reference is represented by places, such as Manchester stadium or the Schwaebisch Alps (seven references), followed by people (five references). The remaining references are food (three times) and individual subjects such as a painting, a letter, Christmas presents and going to a conference.

6.4.11 Very strange

The descriptive adjective *strange* originates from Latin and is frequent in British spoken language, occurring more than 40 times per million words (according to the BNC). *Strange*

is classified as an A1 level word according to the WFF. For *very* see the analyses above (see Table 6.1). The collocation is L1 congruent. The collocation is used by a small set of learners, both females and males, and both with short and long experience abroad. The majority of the learners are undergraduates, but there is also a postgraduate learner; the L3 languages are Chinese, French, and German.

There are five instances of *very strange* in ISLC, which are reported below (two instances occur in the same sentence):

(6.4.138) <B> [...] maybe the box (uh) helped the owner (uh) realising with (mm) realising that (uhm) (mm) he (uhm) he had to give back what he finds (uh) (mm) and (uhm) instead of kee= keeping things (uh) to himself </B> <A> ok and what if the original owner of the box had left the box on purpose in the forest? </A> <B> well that would be *very strange* but I think the (uh) the character will eventually realise what the (mm) the goal of the original owner is [...] </B>

(6.4.139) <B> [...] they see that the shape of the map of the streets is like a star and they are pretty sure that (uhm) it's not a city in their zo= area (uh) because (uhm) they (uhm) they find it *very strange* a *very strange* shape for a city and (uh) actually they haven't they have never seen something like this [...] </B>

(6.4.140) <B> [...] but generally (uhm) the people that (mm) earn more more money than the others work less hours than the the common worker we can say and that's *very strange* because here in Italy yeah there are many (uh) disequalities but (uh) at least the work (mm) is proportional to the (mm) (uhm) to the (mm) to the to the hours [...] </B>

(6.4.141) <B> [...] I didn't like the (uhm) the internship but I like (uhm) I liked China (uhm) the way they live (uh) how do they approach to foreigners they're *very strange* in fact because they're not very social [...] </B>

In four cases the collocation is found in predicative position, such as in (6.4.138), whereas in (6.4.139) one occurrence of *very strange* is found in attributive position. As regards the predicative function of the collocation, *very strange* complements the subject of the sentence in (6.4.138), (6.4.140), (6.4.141), whereas in (6.4.139), the collocation complements an object (i.e., they find *it* very strange). In terms of complements, in (6.4.138) and (6.4.140) the collocation is followed by a clause complement, whereas in (6.4.139) the attributive collocation is followed by a phrase complement. The preferred tense is the present simple, with one instance of conditional in (6.4.138); the sentences are all positive.

The structure of the sentence is *pronoun + copular verb + collocation*, such as in *that's very strange* (6.4.140) or *they're very strange* (6.4.141). In (6.4.139), the first occurrence of *very strange* is preceded by *they find it*, thus the collocation is used to express someone's else judgement regarding the shape of a city.

In two cases the collocation is used to express an evaluation regarding an event: in (6.4.139), the speaker is replying to the interviewer's question about the possibility of someone leaving a box in the forest. The speaker finds this *very strange* and this collocation was produced during the third task. In (6.4.140), the speaker is comparing life in China to life in Italy and is expressing a judgement about the fact that people who work less in China earn more than those who work more hours. This is defined as *very strange* by the speaker as normally in Italy salary is proportional to number of hours. The other three occurrences of the collocation refer to the shape of a city (6.4.139) and to Chinese people. In terms of L1 congruency, no similar pattern of use was identified in the PEC corpus. A random reproducible sample of the L1 equivalent collocation indicated that the lexical items are used to express an evaluation about a wide range of topics, while the sentences are always positive.

The analysis on the five instances of *very strange* in LOCNEC revealed that the collocation is used in predicative position in three cases and in attributive position in two. The predicative collocations complement the subject in two instances, whilst one complements an object. Specifically, the latter is a sentence similar to (6.4.139), where the structure is *personal pronoun + find + it + collocation*. However, in LOCNEC the speakers enriched the collocation with a hedging strategy in two cases (*I think that's a very strange idea* and *it sounds very strange*). The collocations in LOCNEC make reference to a place (Liverpool), to a film, to the fact that a Spanish university does not have a campus, to a plane, and lastly to the ban on American dollars in Cuba.

6.5 Discussion

In this Chapter, we have presented a qualitative analysis of the first set of collocations extracted from the ISLC corpus. These collocations were selected on the basis of their *t*-score and MI ($\geq$ 2 and $\geq$ 3 respectively) and their frequency in ISLC ($\geq$ 5). A scheme was employed for the analysis following the diagram set out in Figure 4.4.1 which required the investigation of three macro-variables: collocational background, learner metadata, and text variables. The collocational background required the definition of the collocation's etymology based on the OED. This was reported in Table 6.1 and a label was assigned to each collocation on the basis of its etymology: shared, mixed, different. The majority of the

collocations (seven) share the same etymology with Italian, while the remaining three have a mixed etymology. The second collocational variable, L1 congruency, was verified on the basis of a similar approach to Wolter and Gyllstad (2013), with the addition of seven native speakers of Italian for the task of L1 congruency translation and evaluation. All the translations provided for the collocations by the seven native speakers confirmed that the lexical combinations are all L1 congruent (see Table 6.2). Lastly, as regards the CEFR level, which was verified through the WFF, all collocations (lexical items composing the collocation) have been categorised as either A1 or A2. These results are not unexpected: the most frequent collocations used by the learners are also frequent combinations in English (according to Biber et al., 1999), they belong to the beginners' level of proficiency (thus it is plausible to hypothesise that they have been acquired in the early stages of language learning), and they are all L1 congruent.

L1 influence has been amply reported in the literature (Bahns, 1993; Granger, 1998b; Nesselhauf, 2005; Wang, 2016) as regards either the effect of the L1 on congruent or non-congruent collocations, and as an explanation for learner errors. The fact that all the collocations of this set are L1 congruent is in line with research into learner collocations, which posits that non-congruency may hinder the production of L2 collocations (Bahns, 1993; Wolter & Gyllstad, 2011; 2013; Yamashita & Jiang, 2010). Although the majority of collocations produced both by ISLC and LOCNEC speakers are L1 congruent[3], some collocations in LOCNEC can be found to be in fact non-congruent (such as *a bit itchy, bitterly cold, bloody knackering, terribly modern*), while all the collocations produced in ISLC are all L1 congruent[4]. Thus, in the case of these collocations, L1 congruency has undoubtedly played a major role in the learners' production, together with frequency.

The second set of variables which were taken into consideration for the analysis of these 11 collocations were learner metadata: these included gender, stay-abroad experience, university courses, and other foreign languages. The variables were also included in a Table (6.3) in order to systematise the analysis as much as possible. No overt patterns of correlation were identified between these variables and the collocations produced. Interestingly, the stay-abroad experience was hypothesised to show an effect on language production (e.g., the learners who had stayed abroad the longest would produce more collocations and use them appropriately) following Tracy-Ventura et al. (2016); however, this was not the case. The authors had warned that not all areas of language may improve following an experience

[3]This is mainly due to the fact that Italian structures intensification similarly to English, thus adverb + adjective combinations are also very frequent.

[4]This has been verified by the author only, not through the translation task assigned to seven native speakers as in the case of Chapter 6 and Chapter 7.

abroad and their findings showed that spoken language did not show signs of significant improvement, especially once the learners had returned home. Nevertheless, it was interesting to notice that all learners who had a short experience abroad (from one week to up to six months) were responsible for the production of these collocations, while the opposite was not true, that is, learners who had no experience abroad were not the sole users of a specific collocation. As far as the university courses are concerned, no clear pattern is discernible, except that course n° 3 and course n° 10 collect the higher number of collocations (10 and nine respectively). Course n° 3 belongs to the Department of Foreign Languages of University A and is an undergraduate course; while n° 10 still belongs to University A, but is a postgraduate course. However, any disparity in numbers among the university courses cannot be solely addressed as evidence that particular groups of learners have a tendency to produce more collocations, but simply has to be attributed to the small (and skewed) sample size of learners. In a similar fashion, the other L3 languages involved in the analysis do not show any specific signs of correlation. It was expected that perhaps the collocations which shared the same etymology (i.e., French and Spanish) would be the most productive, but German was also equally productive. Interestingly, the only collocation which was used across all L3 languages is *completely different*. Regrettably there is not enough research on the influence of L3 on L2 since most of the studies concentrate on the opposite direction, L2 on L3 (see Section 2.3.1.2), but it would be interesting to investigate whether L3s play any role on the production of L2 collocations.

As regards the third macro-category of variables, namely textual ones, these were taken into consideration during the analysis of the 11 collocations. The results point towards a general misuse of the frequent collocations, although the misuse is often not identified on a lexico-grammatical level, but mainly on a pragmatic function level. The following paragraphs analyse the results of the analysis, first collocation by collocation, and then concurrently.

The first collocation analysed is *completely different*, which has a shared etymology with Italian and shows a similar lexico-grammatical pattern to LOCNEC. The main difference arises in terms of communicative function, which, in the ISLC sample, has been identified and divided into three sub-functions: comparison, expectation, opposites (see Section 6.4.1). However, in LOCNEC the only function found in the sample is that of the comparison. Following the literature on L1 congruency (see Section 3.5.1.3) which, in some cases, can account for errors in collocations, a reproducible sample of 20 occurrences was analysed in PEC. The occurrences revealed that *completament divers** behaves similarly to the ISLC *completely different*, with the comparison and the opposites function. In particular, in Italian *completamente divers** is mainly used to express a comparison between opposites, which

make up the majority of the sample occurrences, followed by the comparison. However, no instances of expectation were found. This points towards the fact that the learners appear to be more creative than their native-speaker peers as regards the functions of this collocation, but this creativity is not a case of L1 transfer. On the contrary, it may be hypothesised that the use of the collocation to express the pragmatic function of comparing expectations is a simplification strategy as suggested by Farghal and Obiedat (1995). The authors posited that the underuse of certain collocations may be attributable to simplification strategies, such as synonymy, paraphrase, or avoidance. In this case, it could be argued that the learners rely on a safe collocation in order to express a concept for which they struggle to find better words or expressions.

The second collocation analysed is *really good*, which has a shared etymology with Italian, but shows a dissimilar pattern to LOCNEC both as regards the lexico-grammar and the pragmatic function. Despite the purely qualitative perspective offered in this chapter, it cannot go unnoticed that this collocation only occurs 11 times in ISLC and 71 times in LOCNEC. This generates a wider variety of lexico-grammatical patterns in LOCNEC compared to ISLC. Indeed, in ISLC the preferred construction of the sentence containing the collocation is *pronoun + copular verb + collocation*, whereas LOCNEC shows a much greater variety (see Section 6.4.2). The disparity in terms of structures may be attributable to a lack of the learners' mastery of the collocation (Römer et al., 2020), although the small corpus size may also play a significant role in this and there may be not enough data to understand the reason. Furthermore, the analysis has revealed that the ISLC learners employ the collocation with a negative as a form of hedging strategy when it comes to the expression of abilities or qualities. The hedging role of the collocation has been identified as a way to mitigate the strength of the argument, but it could also point towards a simplification strategy as suggested by Farghal and Obiedat (1995). Indeed, if the learners wish to express a negative judgement as regards an ability or a quality, they may be more prone to rely on a safe collocation (a lexical teddy-bear), which is easier to retrieve from the mental lexicon, rather than take more time to compose a different collocation or expression. This is also corroborated by the fact that there are no instances of the L1 equivalent *davvero buon** in PEC used as a hedging strategy, so that L1 transfer cannot be accounted for the use of the learners of *really good* as a hedging device. Thus, despite the collocation being very frequent and L1 congruent, the learners show an underuse of it in their texts, which is reflected in the limited range of structures accompanying the collocation, and a creative use when it comes to its pragmatic function (hedging strategy). However, it is not clear whether the real intention of the learners was to utilise the collocation as a hedging device, or the production of *really good* is a consequence of a simplification strategy during a more cognitive demanding task

such as spoken L2 IL.

The third collocation analysed is *really interesting*, which has a shared etymology with Italian and similar pragmatic and lexico-grammar patterns, albeit with a few differences compared to native speakers. Indeed, in the ISLC sample the collocation is mainly found in predicative position and the sentences are all positive with the collocation used to express an evaluative judgement about an object, event, or experience. This also occurs in LOCNEC; however, there is also a greater variety of structures in the native-speaker corpus, which abounds in *pronoun + copular verb + collocation, noun + copular verb + collocation, it-* structures, *it + look + collocation* and many others. In particular, some of these structures can be considered linguistic hedges (e.g., the use of the epistemic verbs *I think/I find*), so that the native speakers appear to be more pragmatically competent in their communication than the non-native speakers. This is further corroborated by two instances of *really interesting* found in ISLC: in these sentences the learners, who were performing the third task, attributed the personal evaluation *really interesting* not to an experience or object defined by themselves, the speakers, but rather to a third character or inanimate object. Indeed, if in the other occurrences the collocation appears as the speakers' personal evaluation (e.g., "everything about tourism that I studied was really interesting"), in these two cases the evaluation is spoken by a third character (e.g., "he thinks there's probably a message in it so it would be really interesting to open it"). This could be an effect of the higher cognitive demand that the story-telling task exercises on the learners, who are clearly aware of the main function of the collocation (personal evaluation), but fail to embed it correctly in their online text. Furthermore, there is no indication of L1 transfer, since the L1 equivalent translation is only used to express positive personal evaluation uttered by the speakers themselves.

The fourth collocation analysed is *really nice*, which has a shared etymology with Italian and a similar, although more restricted, lexico-grammatical and pragmatic pattern compared to the native speakers. Indeed, in the ISLC sample the collocation is mainly found in predicative position, in either the present or past tense, and in *noun + copular verb + collocation*, *pronoun + copular verb + collocation* or *it-* structures. Interestingly, the collocation is used to express a positive evaluation only regarding people or categories of people (e.g., colleagues); in two instances only it is used in reference to another object (i.e., flowers) and an experience. On the other hand, in LOCNEC the collocation is found mainly in predicative position and with no preference for the tense, similarly to ISLC, and in *noun + copular verb + collocation* and *it-* structures and variations. The main difference between the two corpora which emerged from the analysis is that in LOCNEC the collocation is used to express a positive evaluation regarding a wider variety of objects, such as places, experiences, and people, while in ISLC *really nice* is practically only associated with people. Fur-

thermore, the collocation in ISLC always refers to plural objects (e.g., people, colleagues, they), whereas in LOCNEC instances of singular objects are also found. Interestingly, the L1 equivalent translations found in PEC are all instances in association with a singular object and reference is made to a plurality of topics and subjects, thus not restricted to people (on the contrary, reference to people is scarce). Thus, it seems that the learners have a good grasp of the lexico-grammatical pattern of this collocation and its pragmatic function. There seems to be no L1 influence on the production of this collocation, despite its L1 congruency and shared etymology, so that the learners have clearly developed their own concept and pattern for this collocation, although mastery has still not been achieved (due to the limited flexibility shown in the use of the collocation – see Römer et al., 2020).

The fifth collocation analysed is *totally different*, which has a shared etymology with Italian and only occurs once in LOCNEC. This collocation has been identified as chiefly American English due to the presence of *totally*, which is more frequent in American English (occurring more than 100 times per million words according to Biber et al., 1999, p. 440), so that it has been hypothesised that its scarce occurrence in LOCNEC is due to the fact that the corpus participants are British university students. Nevertheless, despite not being able to compare the two corpora in terms of lexico-grammatical patterns, the occurrences of *totally different* in ISLC showed a similar pragmatic pattern to *completely different*. Indeed, two of the three functions identified in the analysis of *completely different* emerged in combination with *totally different*: comparison and expectation. In particular, in ISLC the learners employed the collocation in order to compare two terms (either positively, negatively, or neutrally) and to mark an unexpected result compared to a previous expectation. The only occurrence identified in LOCNEC showed a comparison function. Similarly to *completely different* the function of expectation was hypothesised to be attributable to the L1 equivalent *totalmente divers**, thus pointing towards L1 transfer. This was verified by analysing a random (reproducible) sample of 20 occurrences of *totalmente divers** in PEC. The findings reveal that the L1 equivalent of the collocation is used to perform comparisons and express the difference between two terms as the opposites function. No evidence of expectation was found. This leads us to conclude that, as regards *totally different*, the collocation behaves similarly to *completely different* in terms of pragmatic functions; however, due to the scarceness of occurrences in LOCNEC and to the chiefly British English nature of the reference corpus, it is difficult to state whether in this case the learners have been creative with the communicative functions of the collocation. A comparison with an American English corpus could shed some light on this, although it is clear that there are is no evidence of L1 transfer on the learners' use of the collocation.

The sixth collocation analysed is *very different*, which has a shared etymology with Italian

and shows a similar lexico-grammatical pattern, but different pragmatic functions compared to LOCNEC. Indeed, as regards the collocation's positioning in the learners' texts and its textual variables, the collocation is found both in predicative and attributive position, with a preference for the present tense, and is often followed by the preposition *from*. On the other hand, the LOCNEC native speakers use the collocation mostly in predicative position (as is expected in spoken language, see Biber et al., 1999, p. 511), in the present tense, and followed by the preposition *to* (which is chiefly British English, see Leech et al., 2002, p. 85). Despite minor differences between the lexico-grammatical behaviour of the collocation in the learners' and native speakers' texts (which may be attributable to individual variation or the small sample size), the major difference which emerges from the analysis once again regards the pragmatic function. As with the previous collocations containing the adjective *different* (namely *completely different* and *totally different*), the learners have used *very different* with four different functions, three of which had been previously identified (comparison, expectation, opposites), and the fourth one was instanced for the first time in association with *very different*: listing. The listing function was used once by one ISLC speaker to start "listing" a series of activities that she witnessed while walking around Central Park in New York. The activities were all expressed in the same way (*-ing* form) and shared the same anaphoric subject (people): *people doing yoga, people running, people just walking their dogs, people eating their breakfast*. This was identified as a listing function of the collocation which preceded this list, as it appeared that the learner's intention was that of marking the beginning of a list of activities rather than stressing the difference between the activities. In LOCNEC, the only function that was successfully identified was the comparison one. Similarly to the previous collocations containing *different*, the only function found in LOCNEC was that of the comparison. Once again the greater creativity of the learners as regards the communicative function of the collocation was attributed to L1 transfer, but a random (reproducible) sample of the L1 equivalent (*molt* divers**) revealed that the L1 collocation is only used as a means of comparison and opposites. There was no evidence of expectation or listing functions. It was not possible to argue in the discussion of the previous collocation that the learners' creativity in terms of pragmatic function was due to a specific factor due to the lack of sufficient occurrences in LOCNEC and the chiefly American English nature of the collocation. However, in this case, a sufficient number of occurrences was found in the two corpora to allow for a comparison and it can now be argued that the extra functions that the learners assign to the collocation may stem from an avoidance strategy (see Farghal & Obiedat, 1995) whereby when the learners encounter a communicative situation for which they lack appropriate vocabulary, they may fall back on their "islands of reliability" (Dechert, 1983, p. 184).

The seventh collocation analysed is *very good*, which has a mixed etymology (*good* stems from Germanic) and shows slightly different patterns of lexico-grammatical and pragmatic behaviour, but overall is a good example of the learners' almost native-like mastery of the collocation. The collocation is found both in predicative and attributive position in ISLC, mainly in the past tense, and often occurs in a negative sentence. The collocation's main references are to abilities or skills except in two cases where *very good* is used in relation to an health issue and to define the quality of a shop. In LOCNEC, *very good* is found both in predicative and attributive position, the collocation does not show signs of preference for one tense or the other, and it is used to refer to a variety of subjects, both specific ("linguistics department") and vague ("thing"). The key element that emerged from the analysis on the two groups of speakers is the use of collocations preceded by a negation as a hedging strategy. Indeed, both in ISLC and LOCNEC, when the collocation is found in a negative sentence, this is generally because the speaker is attempting to attenuate the strength of the statement in relation to (their) personal abilities or skills: someone may be "not very good" at drawing in ISLC or "not very good at things like working" in LOCNEC. This hedging function of a negated *very good* is also strictly L2 related since there were no instances of it in the PEC corpus when searching for the collocation's L1 equivalent. Thus, despite two instances of infelicitous lexical choice in ISLC, it seems that the learners have grasped the main functions of *very good*, personal evaluation and hedging, and are aware of its preferred lexico-grammatical patterns.

The eight collocation analysed is *very important*, which shares the same etymology of Italian *molt* important** and is an unexpected collocation to find in learner spoken language, especially since there is only one occurrence in LOCNEC. Indeed, the adjective *important* is not frequent in conversation (Biber et al., 1999, p. 512) and is found mostly in attributive position (which is also infrequent for an adjective in spoken language; see Biber et al., 1999, p. 518). In ISLC, the collocation is found eight times, most often in predicative position and it refers to a variety of subjects. The preferred tense is the present and all sentences are positive. On the other hand, the collocation is only found once in LOCNEC, in a positive sentence and part of an elliptical construction with reference to Spanish people. Despite the lack of sufficient occurrences in the reference corpus for a comparison and a statement about the lexico-grammatical behaviour of the collocation, two elements have emerged from the short analysis: L1 transfer and written influence. Table 6.2 shows that the L1 equivalent of *very important* is the most frequent among all L1 collocations, occurring almost 25 times per million words. This is fairly frequent and may have influenced the learners' in their L2 production: since the Italian collocation is so frequent, they may have been more inclined to use the L2 translation in their spoken language. The second aspect is written influence:

research has shown the presence of spoken features in learner writing (cf. Gilquin & Paquot, 2008; Šimčikaitė, 2012), but in this case it appears the opposite may also be true. In an EFL context, it may be that some features of written language, such as the collocation *very important*, may be transferred to spoken language[5]. This may be especially true in an Italian EFL context where main and sole focus remains on written language. Spoken grammar is not taught or taken into consideration and the result if often transfer or written grammar to spoken language (cf. Forchini, 2012).

The ninth collocation analysed is *very interesting*, which shares the same etymology with Italian and shows similar lexico-grammar and pragmatic patterns, albeit with a few differences. The collocation occurs in ISLC only in predicative position, mostly in the present tense and with no negative sentences. Interestingly, one learner produced four instances of the collocation while performing the third task with the construction *they + find + something + collocation*. References is made to experience as regards university life, travelling, entertainment and culture. On the other hand, in LOCNEC the collocation is found only in predicative position, but mostly in the past tense, and the preferred construction is the *it + copular verb + collocation*. There are two cases in which the speakers employed a linguistic hedge by preceding the collocation with *I thought*, which is not present in ISLC. Furthermore, *very interesting* is practically only found in end-sentence position in LOCNEC, as opposed to mid-sentence position in ISLC. The different positioning may be due to L1 transfer since the L1 equivalent *molt* interessant** is only found in mid-sentence position in a random (reproducible) sample analysed. Thus, the analysis highlighted a few micro differences between the two groups of learners. These may be attributable to individual variation (or L1 transfer) or the small sample size; nevertheless, it seems that the learners have a good understanding of the pragmatic function of the collocation (personal evaluation), despite not employing the same linguistic hedges or strategies.

The tenth collocation analysed is *very nice*, which shares the same etymology as Italian and, similarly to the previous collocation *very interesting*, shows similar patterns to LOCNEC both as regards the lexico-grammar and the pragmatic functions to LOCNEC. However, a few minor differences also emerged from the analysis. In ISLC, the collocation is mainly found in predicative position, which is similar to LOCNEC, although in the native speakers *very nice* is also found in attributive position. In ISLC the speakers show no preference for the tense, so that the collocation can occur either in the present or past, while in LOCNEC *very nice* preferably occurs in the present tense. In both corpora the collocation is found in positive sentences, except in one case in LOCNEC which is an instance

[5]This also falls under the wider concept of "register awareness". For more on register awareness, see Altenberg (1998) and Granger and Rayson (1998).

of hedging strategy. The most frequent sentence construction in both corpora is *pronoun + copular verb + collocation*; however, and here is the main difference between the two corpora, in LOCNEC the speakers employ two more hedging devices in addition to a negative. In particular, the collocation is preceded by the approximator *sort of* and the epistemic verb *seems* in LOCNEC. This does not occur in ISLC, which shows no signs of hedging when the collocation is instanced (except in one case with a negative). The learners' use of the collocation without employing any hedging strategies may be attributable to the L1 pattern of *molt* carin**, which is usually found in PEC as a direct personal evaluation of a subject (especially people). Overall, the collocation seems to be used correctly by the learners and similarly to the native speakers, although the native speakers tend to use more hedging devices (which may be due to a wider range of speakers and thus invididual variation). This, in turn, shows better pragmatic competence.

The last collocation analysed is *very strange*, which shares the same etymology with Italian, and, similarly to the previous collocations *very interesting* and *very nice*, shows similar patterns to LOCNEC both as regards the lexico-grammar and the pragmatic function, albeit with a few differences. First, the collocation is mainly found in predicative position in ISLC and the preferred tense is the present. The most frequent construction is *pronoun + copular verb + collocation* and reference is made to a variety of topics, such as people, cities, and others. The main function of the collocation is to express the speaker's personal evaluation regarding a specific subject. In LOCNEC, similar patterns are found: the collocation is found primarily in predicative position, the preferred tense is the present, and a range of topics are the subjects of the collocational modification. However, the main difference between the two groups of learners is found in the use of the epistemic verbs *I think* and *it sounds* found in LOCNEC. These mark the speakers' attenuated statement, a phenomenon which is not found in ISLC. Furthermore, a similar use of the collocation to what was found in *really interesting* occurred in ISLC: in one case the collocation's evaluative function was not attributed to the speaker uttering it, but to a third person, a character of the story-telling task. No evidence of L1 influence was found after searching for the L1 equivalent translations in the PEC corpus. It may be argued that, similarly to *really interesting*, when the learners are subject to a higher cognitive load, such as a story-telling task, they may fail to retrieve the correct and wanted sequences from their mental lexicon and eventually fall back on safe collocations. In a way, these two instances of creative use of the collocations may be grouped under the more general term of "simplification strategy". If a learner is feeling pressured to produce a text online, they may struggle (even in the L1) to compose a logic and fluent text in a short time. Thus, the logical consequence could be that of simplifying the mental effort by relying on those lexical teddy-bears that can be retrieved much more

easily from the lexicon due to their L1 congruence and frequency.

The majority of research on collocations in learner language has focused on investigating the overuse/underuse, the misuse, and the L1 influence on learner collocations (see Section 3.5.1). While Chapter 5 of the present work was dedicated to the first of these three variables, this chapter enquired whether the learners are aware of the correct usage of the collocations from the lexico-grammatical point of view. However, during the analysis a further element emerged and drew our attention, namely pragmatic competence. Indeed, pragmatic competence, which is the ability to "communicate your intended message with all its nuances in any socio-cultural context and to interpret the message of your interlocutor as it was intended", is often disregarded in language teaching and learners may never be able to master it, thus leading them to compose grammatically correct speech, but fail at their communicative aims (Fraser, 2010, p. 15). The analysis has certainly highlighted this pattern of poor pragmatic competence in practically all the collocations analysed: in eight cases (*completely different, really good, really interesting, totally different, very different, very interesting, very nice, very strange*) the learners showed that they had acquired a good grasp of the lexico-grammatical pattern of the collocations (albeit with a few differences which may be attributable to individual variation or small sample size), but frequently displayed inconsistencies with their pragmatic competence. In the case of the three collocations containing the adjective *different*, the learners showed a pattern of creativity in assigning the collocations communicative functions which were not identified in LOCNEC. Although learners' creativity in language production should not always be considered a negative sign of acquisition, it is undebatable that, if the attribution of new communicative functions to specific collocations were to be verified elsewhere in larger reference corpora (or American English corpora in the case of *totally different*), this would contribute to the foreignsoundingness of the learners.

As regards the remaning five collocations, these showed two main patterns of poor pragmatic competence: the learners tend to employ the strategy of negation in their hedged expressions and they are prone to the use of fewer or more limited linguistic hedges compared to their native speaker counterparts. The first strategy, negation, is also used by the native speakers (especially with the collocation *very good*), but the learners show a consistent pattern of negation which could be interpreted as an actual strategy of simplification, rather than hedging. In other words, it seems that when the learners produce a negative sentence in association with the collocation, they are often doing so not so much as to attenuate their statement, but rather to compensate for the lack of a better expression (e.g., (6.4.17) and (6.4.20)). Farghal and Obiedat (1995) argued that the underuse of certain collocations may actually signal a simplification strategy whereby the learners obviate to the (temporary)

lack of a better expression with synonymy, paraphrase, or avoidance. It could be argued that certain hedged statements or certain linguistic hedges employed by the learners are also simplification strategies inasmuch they bridge a gap which the learner is struggling to fill. In terms of the second element, namely the limited range of linguistic hedges, this is a less unexpected finding as Johansen (2020) argued that previous research on learners of English L2 showed that they tend to use fewer hedging strategies compared to the native speakers. In addition, they employ a more limited range of hedging strategies, although further research is needed in order to obtain the full picture (p. 113).

Another element which was taken into consideration was L1 congruency: Bahns (1993) posited that congruent collocations do not need to be taught to language learners since the learners may already be aware of these due to their L1 equivalents. Thus, by excluding a portion of the L2 formulaic language, the learning process for collocations could be simplified. This is also supported by Yamashita and Jiang (2010) who found that L1 congruency combined with L2 exposure maximise the effect on the acquisition of collocations. However, this was not corroborated by Wolter and Gyllstad (2015) in their psycholinguistic study: they found no significant signs of L1 activation in a lexical decision task. Similarly, Nesselhauf (2003; 2005) found similar percentages of errors in congruent and incongruent collocations. The present work further corroborates these findings since practically no L1 influence was identified in the lexico-grammatical and pragmatic use of the collocations. Only in one case, *really ni*ce, was it found that the learners tended to use the collocation as an evaluation of people (e.g., *they're very nice*) as a potential L1 transfer from Italian where the L1 equivalent is used to describe people (especially girls). Furthermore, all learner metadata were taken into consideration when analysing the collocations and no significant qualitative correlation was found between gender, stay-abroad exprience, university course, or L3 languages and the collocations.

In conclusion, the analysis set out to verify whether advanced learners of English were able to use their frequent collocations in a similar way to native speakers as regards the lexico-grammatical pattern. No major differences were found as regards the syntactic patterns of the collocations compared to the native speaker corpus (apart from a lower degree of flexibility in sentence construction). However, an interesting and unexpected element emerged: pragmatic competence. All the collocations involved in the analysis can express personal evaluations regarding specific subjects and, as such, they may also be subject to hedging strategies. A description of the literature on hedging would be beyond the scope of this work, thus for an in-depth review of the literature see Johansen (2020). However, it did not go unnoticed that the learners struggle with the pragmatic competence associated with the production of the spoken collocations. They showed signs of pragmatic creativity

as regards three collocations, whereas a more general restricted use of hedging strategies was noticed in the others. In a few cases, the preferred linguistic hedge of negation was noticed and hypothesised to be a simplification strategy, rather than a hedging one. Despite the difficulty in mastering the syntactic patterns of language structures and formulaic language, these can be reasonably expected to have been grasped by advanced learners, especially in an EFL context which mainly focuses on the grammar aspects of the L2, rather than the communicative ones (cf. Murphy, 2003). However, it is not only a lack of lexico-grammatical proficiency that contributes to the foreignsoundingness of learner IL (cf. Granger, 1998b), but it is also a lack of pragmatic competence, or instances of creative pragmatic functions, or limited range and use of pragmatic devices (i.e., hedges). This calls for the attention of teachers who should focus on the pragmatic aspects of spoken formulaic language in an attempt to improve the learners' communicative skills.

Chapter 7

Qualitative analysis of infrequent/unattested collocations

In the previous Chapter, a set of 11 frequent collocations extracted from ISLC were analysed from a qualitative perspective and with a CIA approach. The analysis was based on a scheme of three main variables which were investigated in relation to the collocations produced by the learners. The findings revealed that, in line with the initial hypothesis that the learners would have acquired mastery of the lexico-grammatical patterns of the collocations, the learners indeed have reached an advanced level of use (in some cases) of certain collocations, but their pragmatic competence is attested at a lower level compared to their native speaker peers. Furthermore, L1 congruency was shown to have practically no influence on the lexico-grammatical or pragmatic patterns of the learners' collocations.

In the present Chapter, a second set of collocations extracted from the ISLC corpus is analysed from a qualitative perspective:

- Absolutely scared
- Kind of scared
- Pretty curious
- Quite old-fashioned
- Really heartbreaking
- Really really amazing
- Really really really good
- Super strange

- Very fun

These collocations were selected on the basis of their infrequency or non-occurrence in the BNC (which means they did not receive any association measures scores) and their frequency ≥ 2 in the ISLC corpus (see Section 4.4.2.3). Furthermore, in order to address the third research question (see Section 4.1) which enquires about the lexico-grammatical patterns of infrequent/unattested collocations, it was decided to systematically tackle each collocation according to a three-fold scheme – illustrated in Section 4.4.2.3 – similarly to the first set of collocations analysed in Chapter 6.

Since there are no occurrences of these collocations in LOCNEC, the lexico-grammatical pattern of the lexical combinations will be examined on its own, rather than within a CIA perspective. More emphasis will be attributed to L1 congruency (or incongruency) in this case, as the hypothesis is that, although some of these collocations may be infrequent or not present in the BNC, due to their chiefly American English nature, some are probably reflections of L1 transfer onto the IL. The hypothesis is that these combinations are examples of infelicitous collocations produced by the learners as a consequence of a negative L1 transfer effect.

The first Section presents the etymology of all the collocations in this analysis; the second Section displays the L1 congruency of the collocations to Italian; the third Section presents the learner variables; and the fourth Section is dedicated to the qualitative investigation of the collocations.

7.1 Etymology of collocations

Table 7.1 illustrates the etymology (according to the OED, see Section 4.4.2.3) of the lexical items constituting the combinations chosen for this analysis. In order to operationalise the description of the etymology and to aid the readers in obtaining a clear and concise picture, the Table presents a similar layout to Table 6.1 in Section 6.1 with a third column assigning a label to the collocation: *same etymology, mixed etymology, different etymology*.

The labels were chosen on the same basis as in the previous analysis, namely according to the following parameters:

- same etymology: the etymology of both collocational items can be traced back to Latin;
- mixed etymology: the etymology of one of the collocational items can be traced back to Latin;

- different etymology: the etymology of none of the collocational items can be traced back to Latin.

Table 7.1: Etymology of infrequent/unattested collocations.

Collocation	Etymology	Shared etymology
Absolutely scared	*absolutely:* Formed within English by derivation. < *absolute* adj. + *-ly* suffix. Compare classical Latin *absolutē*, Anglo-Norman and Middle French, French *absolument*, Middle French *absolutement*, Spanish *absolutamente*, Italian *assolutamente*.	mixed etymology
	scared: < *scare* v. + *-ed* suffix. *Scare* is a borrowing from early Scandinavian. Middle English *skerre* , < Old Norse *skirra* (Icelandic only in phrase *skirra vandræðum* to avoid strife, and reflexive *skirra-sk* with accusative to shrink from; but compare Norwegian *skjerra*, Swedish dialect skj*arra* to scare), < *skiarr* (< **skerro-*) shy, timid, startled.	
Kind of scared	*kind of*: intended as the adverb *kind of* which limits or expresses a degree of reservation about the description or designation that follows it, derives from *kind* probably partly (i) < the same Germanic base as Old English *-cund* (in e.g. *godcund* adj.), Old Saxon *-kund*, Old High German *-chund* , -ch*unt*, Old Icelandic *-kundr*, Gothic *-kunds*, suffix forming adjectives with the sense 'of the nature of' < a suffixed form of the Germanic base of *kin.*	different etymology

	scared: < *scare* v. + *-ed* suffix. *Scare* is a borrowing from early Scandinavian. Middle English *skerre* , < Old Norse *skirra* (Icelandic only in phrase *skirra vandræðum* to avoid strife, and reflexive *skirra-sk* with accusative to shrink from; but compare Norwegian *skjerra*, Swedish dialect *skjarra* to scare), < *skiarr* (< **skerro-*) shy, timid, startled.	
Pretty curious	*pretty*: formed within English, by conversion. The adjective *pretty* < *prat* n. + *-y* suffix. Compare Dutch *prettig* pleasant, nice, agreeable, comfortable (also Belgian Dutch †*pruttig*), German regional (Low German: East Friesland) *prettig* sportive, funny, humorous, Old Norwegian *prettugr* tricky, deceitful (compare Old Icelandic *prettóttr*), also (with metathesis) Middle Dutch *pertich* cunning, quick, lively (Dutch regional (southern) *pertig*, Belgian Dutch *pertig*, also in sense 'whimsical, capricious').	mixed etymology
	curious: < Old French *curius* = Provençal *curios*, Spanish *curioso*, Italian *curioso* < Latin *cūriōsus* used only subjectively 'full of care or pains, careful, assiduous, inquisitive'; French has also the objective sense in 14th cent.	

Quite old-fashioned	*quite*: of multiple origins. Apparently partly a borrowing from French. Apparently partly a variant or alteration of another lexical item. Apparently partly < Anglo-Norman *quite* without opposition, compare also Middle French *quittes* totally; < *quite*, adjective, and partly < qu*it* adj. (although this is apparently not attested in the most closely corresponding sense until considerably later); in sense probably influenced by Anglo-Norman and Middle French *quittement* freely, unconditionally, completely.	different etymology
	old-fashioned: formed within English, by compounding. The adjective *old* is a word inherited from Germanic. The word *fashioned* < *fashion* n. + *-ed* suffix. *Fashion* < Old French *façon*, *fazon*, Old Northern French *fachon* (modern French *façon*) = Provençal *faisso*, Italian *fazione* (the Spanish *faccion* is of learned origin) < Latin *factiōn-em*, noun of action < *facere* to make.	
Really heartbreaking	*really*: Formed within English, by derivation. < *real* adj. + *-ly* suffix. Compare post-classical Latin *realiter* actually, in fact (frequently from 11th cent. in British sources; from 14th cent. in continental sources; 6th cent. as a variant reading, where the correct reading is probably *dealiter* divinely), Middle French *reaument* (1310 in Old French), *realment* (1353; French *réellement*).	mixed etymology

	heartbreaking: formed within English, by compounding. *Heart* is a word inherited from Germanic. *Breaking* < *break* v., Old English *brecan* (*bricþ*, past tense *bræc*, *brǽcon*, past participle *brocen*), corresponding to Old Frisian *breka*, Old Saxon *brekan*, (Middle Dutch, Dutch *breken*), Old High German *brehhan* (Middle High German, modern German *brechen*), Gothic *brikan* (past tense *brak*, *brêkum*, participle *brukans*) < Old Germanic stem *brek-*, corresponding to Latin *frag-* (*frang-o*, *frēgi*, *frac-tum*), Aryan **bhreg-*.	
Really really amazing	*really*: Formed within English, by derivation. < *real* adj. + *-ly* suffix. Compare post-classical Latin *realiter* actually, in fact (frequently from 11th cent. in British sources; from 14th cent. in continental sources; 6th cent. as a variant reading, where the correct reading is probably *dealiter* divinely), Middle French *reaument* (1310 in Old French), *realment* (1353; French *réellement*).	mixed etymology
	amazing: formed as *amazing* n. + *-ing* suffix. The noun *amazing* is a derivation of the verb *amaze* < *a-* prefix ? intensive + *maze* n. *Maze* is probably the reflex of an unattested Old English noun **mæs* or **mase*, of which *amasian* amaze v. is a verbal derivative (it is possible that an unprefixed verb **masian* also existed); further etymology uncertain.	

Really really really good	*really*: Formed within English, by derivation. < *real* adj. + *-ly* suffix. Compare post-classical Latin *realiter* actually, in fact (frequently from 11th cent. in British sources; from 14th cent. in continental sources; 6th cent. as a variant reading, where the correct reading is probably *dealiter* divinely), Middle French *reaument* (1310 in Old French), *realment* (1353; French *réellement*).	mixed etymology
	good: A word inherited from Germanic. Cognate with Old Frisian *gōd*, Old Saxon *gōd*, Old Dutch *guot*, *guod*, Old High German *guot*, Old Icelandic *góðr*, Norn *go-*, *goug(h)-* , *gu-*, Old Swedish *goþer*, Old Danish *goth*, Gothic *gōþs*, probably < an ablaut variant (*ō* -grade) of the base seen also in *gather* v., with an original sense 'fitting', 'suitable'; compare further Lithuanian (archaic and regional) *guodas*, Latvian gods, both in sense 'honour'.	
Super strange	*super*: a borrowing from Latin. < classical Latin *super-*, use as prefix of *super* (adverb and preposition) above, on the top (of), beyond, besides, in addition < the same Indo-European base as Sanskrit *upari-* , ancient Greek ὑπερ- *hyper-* prefix, Old English *ofer- over-* prefix.	same etymology
	strange: < Old French *estrange* (modern French *étrange*) = Provençal *estranh*, *estrang*, Spanish *estraño*, Portuguese *estranho*, Romanian *strâin*, Italian *strano* adjective, *stranio*, *strangio* noun < Latin *extrāneus* external, foreign (see extraneous adj.), < *extrā* adverb, outside, without.	

Very fun	*very*: < Anglo-Norman *verrai*, *verrey*, *verai*, *veray*, Old French *verai*, *varai*, *vrai* (modern French *vrai*, Provençal *verai*), < the stem of Latin *vērus* true.	mixed etymology
	fun: probably formed within English, by conversion. The verb *fun* < origin uncertain. Perhaps a regional variant of *fon* v. (although this is apparently otherwise obsolete after the Middle English period in its specific sense 'to deceive, befool (a person)'.	

Furthermore, the etymological definitions are repeated for the same lexical items present in multiple collocations so as to facilitate the readers whenever they consult the Table. The readers will thus be able to consult the Table for each collocation and find all the etymological definitions pertaining to the collocation at once, without needing to compose the definition and the etymological label themselves.

As done in the previous Chapter, the first column of Table 7.1 corresponds to the collocations under investigation; the second column is split into two rows to provide the etymology of the two lexical items composing the collocation. The first row is occupied by the adverb, while the second row focuses on the adjective. The etymological definitions are extracted from the OED, with minor editing, mainly related to the elimination of the dates referring to the first use of the etymons. The last column provides the etymological label (as mentioned above) to facilitate the reading of the table. For example, in case of the first collocation, *absolutely scared*, the adverb stems from Latin (*absolutē*), while the adjective is a borrowing from old Scandinavian. Thus, the collocation is labelled as having a mixed etymology.

The majority of all combinations have a mixed etymology (six), while two of them display different etymologies, and only one shares the same etymology as Italian. The mixed or different etymology could anticipate some L1 incongruency in the case of these collocations. One note should be made regarding the adverb *very*: as mentioned in the previous Section (6.1), the adverb originates from Anglo-Norman and Old French, which both have their root in Latin *vērus*, thus the adverb has been considered to have the same etymology of the L1 because in contemporary Italian *vero* (adj.), *veramente* (adv.) are still used today and they share the same Latin root. As regards the collocation *quite old-fashioned*, this has been labelled as *mixed etymology*. This is due to the adjective *old-fashioned*, which has mixed compound roots and, despite its Latin origin as regards *fashion*, cannot be regarded as a facilitating word for the learners, since the L1 equivalent *moda* does not share the same root.

7.2 L1 congruency of infrequent/unattested collocations

In a similar fashion to the previous Chapter and in an attempt to operationalise as many variables as possible (as per diagram in Figure 4.4.1), L1 congruency is displayed in Table 7.2.

As explained in Section 4.4.2.3, L1 congruency determination is loosely based on Wolter and Gyllstad's (2013) approach, namely the prototypical semantic value translation of the collocations by a native speaker (who in this case are seven, rather than one) and the fre-

Table 7.2: L1 congruency of infrequent/unattested collocations.

	Translation to L1	PEC occurrences (pmw)	L1 congruency
Absolutely scared	no agreement	n/a	L1 incongruent
Kind of scared	no translation	n/a	L1 incongruent
Pretty curious	abbastanza curios*	0.23	L1 congruent
Quite old-fashioned	abbastanza vecchio stile	no occurrences	L1 congruent
Really heartbreaking	no agreement	n/a	L1 incongruent
Really really amazing	no translation	n/a	L1 incongruent
Really really really good	no agreement	n/a	L1 incongruent
Super strange	super stran*	no occurrences	L1 congruent
Very fun	molto divertent*	2.49	L1 congruent

quency estimates in an L1 reference corpus.

It has already been mentioned in Section 4.4.2.3 that since Italian has multiple translations for some of the collocational items (e.g., *really amazing, really good*), it was decided to request the participation of seven Italian native speakers who provided direct translations for the collocations. The most frequent translation among the native speakers was chosen as the final one. If no direct translation (no word-by-word translation) was provided by none of the native speakers, or if no agreement was found on a univocal translation, the collocation was labelled *incongruent*. The second column of Table 7.2 displays the final translation obtained by the group of native speakers. The asterisks have been placed where the adverb or adjective can be declined according to gender, singular or plural, in Italian; the translations have been searched in PEC using the asterisk as a wild card to obtain all declinations of the collocational items. The third column displays the frequency occurrences of the L1 translations in PEC normalised to one million words (pmw). The last column illustrates whether the collocation is L1 congruent or incongruent: the collocation is L1 congruent in four cases (*pretty curious, quite old-fashioned, super strange, very fun*) and L1 incongruent

in five cases (*absolutely scared, kind of scared, really heartbreaking, really really amazing, really really really good*).

7.3 Learner metadata

As reported in Section 4.4.2.3 and similarly to what was done in Section 6.3, the learners' ethnographic metadata (i.e., gender, stay-abroad experience, university courses, other languages) have also been systematically gathered in a table (Table 6.3). This was done to facilitate the interpretation of any correlation between the collocations and the learner variables. Table 6.3 displays the infrequent/unattested collocations and the metadata of the learners who produced them.

The first variable reported in the table is gender: all the collocations were produced by women and one, *really really really good*, was also generated by a man. The stay-abroad experience variable indicates that practically all learners using these collocations have had a short experience in an English-speaking country, while four of them spent more than six months. Only three learners did not have any experience abroad. As regards the university courses, it seems that the majority of the learners attend the postgraduate course n° 10 (corresponding to course IM belonging to the Department of Foreign Languages of University A); in addition, the majority of the learners attend postgraduate courses (11 speakers) rather than undergraduate ones (4 speakers). Lastly, in terms of the other languages factored in the analysis, namely Chinese, French, German, Russian and Spanish, French seems to be the most common among the learners using these collocations.

Overall, the Table shows that the most likely speaker who could produce these collocations is a female postgraduate student with a short experience abroad in an English-speaking country, whose L3 is French.

7.4 Analysis of infrequent/unattested collocations

In this Section the nine collocations will be analysed from a qualitative perspective, as anticipated in Section 4.4.2.3, following the diagram presented in Figure 4.4.1. References will be made to the tables presented in the Sections above (i.e., Table 7.1, 7.2, 6.3).

Table 7.3: Learner metadata variables with reference to the infrequent/unattested collocations.

	gender		stay-abroad experience			university courses										other languages				
	♀	♂	yes		no	1	2	3	4	5	6	7	8	9	10	11	CHI	FRE	GER	RUS	SPA
			< 6 m.	> 6 m.																	
Absolutely scared	✓		✓		✓		✓								✓		✓	✓			
Kind of scared	✓		✓	✓								✓			✓			✓			✓
Pretty curious	✓		✓												✓				✓	✓	
Quite old-fashioned	✓		✓	✓				✓							✓					✓	✓
Really heartbreaking	✓		✓							✓								✓			
Really really amazing	✓			✓				✓										✓			
Really really really good	✓	✓	✓		✓								✓		✓		✓	✓			
Super strange	✓		✓												✓						✓
Very fun	✓		✓	✓								✓		✓	✓			✓	✓		

7.4.1 Absolutely scared

The evaluative adjective *scared* is fairly frequent in English occurring approximately 14 times per million words (according to the BNC). According to the OED, *scared* is a borrowed word from early Scandinavian (see Table 7.1) and thus does not share the same etymology as Italian. The adjective is classified as belonging to the B1 CEFR level by the WFF.

Absolutely is a frequent amplifier in British spoken language occurring more than 100 times per million words (Biber et al., 1999, p. 565) and stemming from Latin (see Table 7.1). The adverb belongs to the A2 CEFR level according to the WFF. The collocation has a mixed etymology and it is not L1 congruent. As regards the ISLC speakers, two participants used this collocation, they are both female, and they either spent a short time abroad or no time at all. Furthermore, one attended an undergraduate course, while the other a postgraduate course. The languages spoken are French and Chinese.

There are two instances of *absolutely scared* in ISLC and both have been instanced during the third task.

(7.4.1) <B> [...] they find some (uhm) some money in it . quite a large sum (uhm) absolutely and being *absolutely scared* of losing the money (uh) they decide not to tell anyone [...] </B>

(7.4.2) <B> [...] they decided to call the police to solve the mystery and (uh) they were *absolutely scared* and pretty sure that they were (uh) controlled so uhm [...] </B>

The combination in (7.4.1) is part of a verb phrase and, as it is indicated by the preceding adverb *absolutely*, it appears as if the learner were trying to use one of the adverbs given in the prompt and after a short hesitation combined the adverb with *scared*. A similar event occurs in (7.4.2) where the learner is trying to narrate a story and after a short hesitation produces the combination above. Furthermore, both combinations are instanced in reference to the pronoun *they*, thus they are used to describe a feeling perceived by others. The adverb *absolutely* is chiefly British English and is frequent in conversation, occurring approximately 100 times per million words; it is less frequent in American English with 50 occurrences per million words (Biber et al., 1999, p. 565). Its main function is that of intensifying the adjective immediately following, which in conversation should generally be an adjective expressing a positive evaluation (Biber et al., 1999, p. 564). The two occurrences of ISLC above indicate that the learners applied the intensifying function of the adverb to an adjective which expresses a negative feeling, namely fear, thus overturning the hearer's expectations.

7.4.2 Kind of scared

Kind of is an adverb which is especially frequent in American English conversation, occurring more than 1,000 times per million words (Biber et al., 1999, p. 561). It belongs to the hedges category according to Biber et al. (1999, p. 557) and Fraser (2010) since its use marks the imprecision of the word choice immediately following. It is common for the adverb to modify a noun phrase or an adjective. Etymologically, it was probably partly inherited from Germanic and thus shares no etymology as Italian (see Table 7.1). It is categorised as a B1 level word according to the WFF. The evaluative adjective *scared*, as explained in the previous description, derives from Scandinavian and is classified as belonging to the B1 CEFR level by the WFF. In terms of L1 congruence, the native speakers who were judging the congruency agreed that there is no equivalent translation in Italian, thus the collocation is incongruent. As regards the speakers, two participants produced this collocation: both are females and they had either a short experience or long experience abroad, they both attended postgraduate courses, and they speak French and Spanish.

In ISLC, there are two instances of *kind of scared*, both of which are evaluations of personal feelings in reference to other people, not the narrator himself or herself:

(7.4.3) <B> [...] and of course they hoped it wouldn't be an animal (uh) they were *kind of scared* about having to deal with a strange animal [...] </B>

(7.4.4) <B> [...] someone knew something about it so he went to the market it was (uh) *kind of scared* to ask people because (mm) it finding (uh) an egg [...] </B>

In both cases the combination is found in predicative position and complements the subject of the sentence. In addition, both combinations are followed by a clause complement. The tense of the sentences is in the past simple and the combinations were produced during the third task. Similarly to the previous collocation, *absolutely scared*, *kind of scared* have been uttered by the speakers not in relation to a feeling they experienced, but they have been attributed to a character of their story-telling task. Indeed, in (7.4.3) the characters protagonist of the story are *kind of scared*, and in (7.4) it is once again the character of the story who is *kind of scared*.

7.4.3 Pretty curious

Pretty is a common amplifier (immediately preceding adjectives) present in American English conversation and occurring at least 400 times per million words (Biber et al., 1999, p. 561). The adverb is of Germanic origin and thus has no similar etymology with Italian; it is

categorised as A2 by the WFF. *Curious* is a frequent adjective occurring more than 25 times per million words (Biber et al., 1999, p. 532) which derives from Latin. The adjective shares the Italian etymology and belongs to the B1 CEFR level according to the WFF. As far as L1 congruence is concerned, the L1 equivalent of *pretty curious* is *abbastanza curios**, which occurs only six times in PEC (0.23 times per million words). The six instances are found in an academic article, a book, a newspaper article, two TV programmes, and a wikipedia web page. Interestingly, during the L1 translation task, all the native speakers suggested translations for *curious* which, in combination with an intensifier adverb, produce the meaning of *peculiar* in Italian, rather than simply the adjective describing someone with an eager interest in something else. For this reason, in PEC the six occurrences of *abbastanza curios** convey the meaning of something rather peculiar, which can be a piece of information, being hired by someone, events, someone's behaviour, a software feature. In terms of learners, only two ISLC participants produce this collocation: they are both females and with a short experience abroad; they both attend the same postgraduate course (n° 10) and speak German and Russian as their L3s.

In ISLC, there are two instances of *pretty curious*, both of which are evaluations of personal feelings in reference to other people, not the narrator himself or herself:

(7.4.5) <B> [...] it's a city (uh) near another one which is which has a round shape so they think it could be there so they (uhm) they're *pretty curious* about it and (uh) they decide to go there in the round city [...] </B>

(7.4.6) <B> [...] maybe that could be a bomb so something very (uhm) (uh) dangerous (uh) but (uhm) they were (uh) *pretty curious* (uh) so they absolutely wanted to open it [...] </B>

In both cases the combination is found in predicative position and complements the subject of the sentence. In addition, one combination is followed by a phrase complement (7.4.5). The tense of the sentences is the present simple for (7.4.5) and the past tense for (7.4.6). Both combinations were produced during the third task. The L1 equivalent combination is used either in the past tense or in the present. As mentioned above, in Italian the combination is used with the meaning of *rather peculiar* to describe certain events or facts; in the case of ISLC, both combinations are used to describe someone's attitude towards a city (7.4.5) and the content of a strange-looking box (7.4.6).

7.4.4 Quite old-fashioned

Quite is a common amplifier in British English conversation occurring more than 350 times per million words (Biber et al., 1999, p. 567). It is a borrowing from French, but shares no Italian etymology (see Table 7.1). The WFF classifies the adverb as an A1 level word. As explained in Biber et al. (1999), *quite* can occur with both gradable and non-gradable adjectives, but often this results in different meanings. With gradable adjectives whose meaning does not imply an absolute end-point of the scale, the adverb usually means "to some extent", but with non-gradable adjectives the meaning becomes that of "completely" (p. 556). The issue is that many adjectives can occur with *quite* and be used in both senses, such as *confident*. In spoken language, Biber et al. suggest paying attention to intonation, as that may be an indicator of one meaning or the other. However, they also warn that it is practically impossible to distinguish between the two different meanings. *Old-fashioned* is an adjective which was formed by compounding the adjectives *old* and *fashioned*. The two components of the adjective have different etymologies, *old* is a cognate from Old Frisian, while *fashioned* derives from Old French, thus the adjective shares no etymology (see Table 7.1). The adjective belongs to the C1 CEFR level according to the WFF. In terms of L1 congruence, the L1 equivalent has been indicated by the native speakers as *abbastanza vecchio stile*. Since the native speakers have provided a translation, the combination has been deemed L1 congruent; however, this translation is not without issues, since it is actually three words, rather two, so it is not technically a word-by-word translation. Furthermore, there are no occurrences of *abbastanza vecchio stile* in PEC. In terms of speakers, two ISLC participants produce this collocation: they are both females with either a short or long experience abroad, one attended an undergraduate course, while the other a postgraduate course; their L3s languages are Russian and Spanish.

In ISLC, there are two instances of *quite old-fashioned*, but in both cases the intonation was not of any help in distinguishing between the use of the diminisher or the intensifier. Both combinations were produced during the second task, that is the dialogue, and in both cases they refer to the same domain, entertainment:

(7.4.7) <B> [...] a good film (uh) well (uh) it's one of my favourite know it's= I know it's *quite old-fashioned* it's (uh) Gone with the Wind I really like it and (uhm) I don't know [...] </B>

(7.4.8) <B> [...] no no no last summer and also (mm) Lucio Battisti which is *quite old-fashioned* but I love him [...] </B>

In the first case (7.4.7), reference is made to the film *Gone with the Wind*, which the speaker

describes as her favourite film; in the second case (7.4.8), the speaker is referring to her favourite singer, the late Italian singer Lucio Battisti. Thus, in both cases reference is made to music and films and the combination is found in predicative position and complements the subject. In addition, the tense is the present simple in both cases and the sentences are positive.

7.4.5 Really heartbreaking

Really is a derivation of *real* + *-ly* and stems from Latin *realiter* (see Table 7.1). It is extremely common both in British English and American English conversation, occurring more than 350 and 600 times per million words respectively (Biber et al., 1999, p. 565). The intensifier belongs to the A1 CEFR level according to the WFF and shares the same etymology with Italian. *Heartbreaking* is an adjective formed within English by compounding the noun *heart* and the adjective *breaking*. Both the noun and the adjective originally derive from Old Frisian, so the adjective does not share the same etymology with Italian (see Table 7.1). The adjective is not categorised by the WFF according to a CEFR level because there is insufficient data. However, it is reasonable to expect the adjective to belong to either the upper intermediate level (B2) or the advanced level (C1). As regards L1 congruence, the native speakers found no agreement regarding a possible L1 translation of *really heartbreaking* so that the collocation has been labelled incongruent. As regards the speakers, only one ISLC participant uses this collocation twice: the learner is a female with a short experience abroad attending an undergraduate course and studying English and French.

In ISLC there are two occurrences of *really heartbreaking*, both produced by the same speaker while she is describing a film which impressed her. In particular, she is recounting a specific scene in which a dog, which has lost its owner, patiently waits in the same place in the hope that the owner will come back. She describes this scene as *really heartbreaking* and she repeats the combination in the same sentence:

(7.4.9) <B> [...] and wait for his owner day after day and you can see the scenes when the seasons pass and his dog gets older and more and more tired (uh) it's *really heartbreaking* but (uh) and knowing that it's a true story makes it even more *really* (uhm) *heartbreaking* yeah and I think [...] </B>

In the first instance the combination is in predicative position and complements the subject, while in the second instance the combination is in predicative position but complements the object. In both cases the tense used is the present simple and the sentences are positive. There is some hesitation before the second combination instanced by *even more*, which may

indicate the speaker was about to choose a different lexical item. Indeed, it sounds as if the speaker had planned on a different lexical item, but could not find it promptly and thus quickly replaced it with a repetition of the combination instanced earlier. This could be further confirmed by the filled pause (*uhm*) between the adverb and the adjective.

7.4.6 Really really amazing

As mentioned above, *really* stems from Latin and shares the same Italian etymology (see Table 7.1). It is extremely common both in British English and American English conversation and belongs to the A1 CEFR level according to the WFF. *Amazing* is an adjective probably deriving from Old Frisian, but etymology is uncertain and as such the collocation's etymology is mixed (see Table 7.1). *Amazing* belongs to the B1 CEFR level according to the WFF. The native speakers agreed that there is no translation for *really really amazing* so that the collocation is L1 incongruent. In terms of speakers, only one ISLC participant produced this collocation: it is a female student with a long experience abroad (in the USA) attending an undergraduate course and speaking French as L3.

The combination contains a repetition of the intensifying adverb *really*. Double intensification has already been noted in the literature (cf. Tagliamonte, 2008; Putri et al., 2020), although there is still ample ground for research on the topic. Quirk et al. (1985, p. 473) wrote that some intensifiers can be repeated for emphasis, but usually these are the ones pre-modifiying comparatives, rather than non-graded adjectives. In addition, the authors warned that the repetition is generally permissible only if the repeated items come first or follow so (i.e., *very very good*, *so very very good*). Although there is no L1 equivalent, double intensification also exists in Italian, especially in spoken language (this phenomenon is called *reduplicazione espressiva* – expressive reduplication – cf. De Santis, 2014).

In ISLC, there are two instances of *really really amazing*, both produced by the same speaker in the context of a description of the Niagara Falls and the Grand Canyon:

(7.4.10) <B> [...] Niagara Falls (uh) those waterfalls were really powerful and gigantic and they were *really really amazing* and also there= (uh) there= there's the (mm) border [...] </B>

(7.4.11) <B> [...] but still from the natural point of view that's really beautiful even though the Grand Canyon is *really really amazing* too but I like water [...] </B>

In both cases the combination is found in predicative position and it complements the subject. In addition, they are both followed by a clause complement. In the first case the tense is in the past, while in the second case the tense is in the present. The recording was listened

to and it revealed that the duplication of the adverb was not a mere hesitation, but it was purposefully produced to emphasise the descriptions. Indeed, throughout the whole interview the speaker generated multiple instances of double intensification. Furthermore, the same speaker produced several instances of *really* as a modifier of a clause, such as in:

(7.4.12) <B> [. . .] I made many friends (uh) that *really* had an= different vision of the world [. . .] </B>

(7.4.13) <B> [. . .] and (uhm) it *really* helped me to open my mind even more [. . .] </B>

(7.4.14) <B> [. . .] I *really* felt lucky cos I= ever= every time we think we have so many problems [. . .] </B>

An in-depth study by Paradis (2012) suggested that, in addition to its intensifying property when found in pre-adjective position, *really* is also a marker of epistemic stance. Indeed, in (7.4.12) the speaker is employing the adverb to emphasise her subjective view of the situation and the same happens in (7.4.13) and (7.4.14). Thus, it seems that this speaker's communication style includes the frequent use of emphasis, as regards adjectives but also clauses from the epistemic stance perspective. There is no baseline or text produced by the same speaker in Italian to verify whether this speaker indeed tends to overuse emphatic devices such as double adverbs.

7.4.7 Really really really good

As mentioned above, *really* stems from Latin and shares the same Italian etymology (see Table 7.1). It is extremely common both in British English and American English conversation and belongs to the A1 CEFR level according to the WFF. *Good* is a very common evaluative adjective in conversation occurring more than 200 times per million words according to Biber et al. (1999, p. 512). Its origin can be traced back to Old Frisian so that the adjective does not share the same Italian etymology. The adjective belongs to the A1 CEFR level according to the WFF. In terms of L1 congruency, due to the highly polysemous translation of *good* in Italian, no agreement was reached among the native speakers as regards a word-by-word L1 translation: the collocation is L1 incongruent. In terms of speakers, two speakers produced this collocation: one is a female learner and the other is a male learner, one had a short experience abroad, while the other no experience at all; they both attend a postgraduate course (n° 8 and n° 10) and they are learners of French L3.

There are two instances of *really really really good* in ISLC:

(7.4.15) <B> [...] so (uh) the characters (uh) weren't really m= well described and you couldn't (uh) sympathise with for example the main character the bad character or the fallen angel for example and well I was really disappointed because I expected *really really really good* and because this film took I think four or five years to be made so I really expected something (uh) well extraordinary [...] </B>

(7.4.16) <B> [...] so it was amazing few mistakes maybe I think they haven't read the other four so it's fine but I think it made a *really really really good* episode it's an hour and twenty most of it are like fifty minutes [...] </B>

In the first case (7.4.15), it is difficult to understand whether the combination was used in attributive or predicative position since the subject (and maybe the object too) is missing. It could be reasonable to think that the learner may have wanted to say "I expected it to be really really really good" or "I expected something really really really good", both of which would contemplate the combination in predicative position. In the second case instead, the combination is found in attributive position in reference with *episode*. In both positive sentences the tense is in the past and the general topic of the conversation is about a film and a TV series respectively (both were produced during the first task). The interpretation of the sentences and the audio recordings of the conversation seems to point towards the learners' communicative goal of providing a description of a film and TV series with an adjective that marks the high end-point of the scale. In other words, the film and the episode could have been *exceptional*, *amazing*, *remarkable*, *wonderful*, *impressive* and many other non-gradable adjectives, but perhaps these synonyms of *really good* could not be easily retrieved from their repertoire, so the learners settled for the intensification of the gradable adjective *good* via double intensification (or *reduplicazione espressiva* in Italian). However, knowledge of French is a feature that the speakers share and, intuitively, it is known that double intensification of the adjective *très* is extremely common in French, so that it is possible that in this case the L3 had an influence on the L2. Regrettably little research has been done on the influence of L3 over the L2, but this certainly seems like a fertile ground.

7.4.8 Super strange

Super is an intensifier, especially common among teenagers (Tagliamonte, 2008), which originally was a prefix stemming from from classical Latin *super*, meaning above, on the top (of), beyond (see Table 7.1). The intensifier shares the same etymology with Italian, but as regards the CEFR level, since its principal grammatical function is that of a prefix, the WFF does not provide a categorisation. The evaluative adjective *strange* originates from

Latin so that the adjective shares the same collocation with Italian. *Strange* is classified as an A1 level word according to the WFF. As regards the L1, intensification of adjectives with prefixes such as *super-* is attested in Italian (Grandi, 2017; Malloggi, 2017; Paradis, 2008; Dressler & Barbaresi, 1994), especially among teenagers who make a widespread use of intensification (Grandi, 2017). Furthermore, the native speakers provided the same translation of *super stran**, so that the combination is L1 congruent. *Super stran** though is not present in PEC, but this is predictable since the combination makes use of an infrequent type of intensification and one which is typically common among a specific group of speakers. In terms of speakers, *super strange* is used only by one speaker, speaker IT012, who is a female student with a short experience abroad attending a postgraduate course at University A and studying Spanish as L3.

In ISLC, there are three instances of the combination:

(7.4.17) <B> [. . .] at the end of my internship I was asked to stay and I I was so afraid of having this job and I was like no ok I'm leaving and it was *super strange* because when I decided to take this gap year I was like I wanted to work so much [. . .] </B>

(7.4.18) <B> [. . .] (mm) I thi= I I'm super it's *super strange* to say but I have like a more positive outlook on life [. . .] </B>

(7.4.19) <B> [. . .] it's *super strange* to be here now at university and I see my friends that happened to work in their life and they're I feel there is a difference [. . .] </B>

All three combinations were instanced by speaker IT012 and in all three cases the combinations are in predicative position, complement the subject (*it*), and are followed by a clause complement. In one case the tense is in the past, while in the other two cases the tense is the present simple. All sentences are positive. The use of the adjective *strange* was thought to have occurred in the context of the third task since the presence of the adjective *strange-looking* in the prompt may prime synonyms or near synonyms such as *strange*. However, the inspection of the combinations reveals that all three sentences were instanced in the context of the first task, where the speaker is describing her job experience as an intern in a company. It is interesting to notice that in all cases the combination is referring to a feeling: in (7.4.17) the speaker is saying that it was *super strange* to accept the internship because she felt torn between a new job experience and the more comfortable life of university; in (7.4.18) the speaker states that it is *super strange* to say that the job experience in the end gifted her with a new perspective on life, a more positive outlook; and in (7.4.19) the speaker affirms that now being back at university is *super strange*. There is no reference to objects or people but only to a personal feeling or attitude towards certain events. Another aspect which is

worth mentioning is the fact that speaker IT012 alone produced approximately 75% of all ISLC instances of *super*. In her text there are 11 *super* used in conjunction with adjectives (*shocking*, *stressful*, *serious*, etc.). The original hypothesis was that the use of *super strange* may derive from either the learner's exposure to a high degree of prefix intensification or the learner's personal communicative style. The fact that all three instances of *super strange* were found in the text of speaker IT012 and the high number of *super* present in conjunction with other adjectives suggest that it is the speaker's communication style which is dense with intensification. This intensification, which probably is present in Italian as well, has been transferred into the TL.

7.4.9 Very fun

Very is a common intensifier which stems from Latin *vērus*, thus it shares the same etymology as Italian and is categorised as an A1 word by the WFF. *Fun* is a conversion within English from the verb *fun*, which is of uncertain origin, so the etymology is not shared. The word is categorised as belonging to the A2 CEFR level by the WFF. *Fun* can be described as an uncountable and abstract noun or a qualitative evaluative adjective (Quirk et al., 1985, p. 412). Indeed, Quirk et al. note that "*Fun* is primarily a regular uncount noun: *a great deal of fun*, *great fun*, *some fun*. However, in informal usage (esp AmE), *fun* has been fully converted into an adjective: *The party was fun*, *a fun person*; *a very fun party*" (p. 412). Since *fun* is a conversion from the verb *fun*, the adjective (and the noun) can be considered a homomorph, a term that Quirk et al. (1985, p. 70) use to describe words with the same morphological form, but different syntactic functions. In this case, *fun* is a noun, but also an adjective. In terms of intensification, a study by Enstad (2010) on *fun* and *funny* found that in the BNC *fun* is only modified by adverbs such as *quite*, *really*, *rather* and *kinda*. On the other hand, in COCA *fun* is pre-modified by a wider range of adverbs including *really*, *so*, *very*, *as*, *pretty*, *quite*, *especially*, *kinda*, *downright*. This confirms Quirk et al (1985)'s description of the adjective *fun* as predominantly informal (spoken) American English. In terms of L1 congruence, the native speakers identified *molto divertent** as the L1 equivalent of *very* fun, so that the combination is L1 congruent. The combination is also present in PEC, although rather infrequently, occurring 2.49 times per million words. In terms of speakers, three learners produce this combination: they are all females, either with a short or long experience abroad, all attending postgraduate courses at University A and University B, studying French and German as L3s.

In ISLC, there are three instances of *very fun*:

(7.4.20) <B> [...] two very different people can become friend friends and they (mm) can

learn things one from another (uh) from one another and but it was also *very fun* like the Italian man was (uh) really fun [. . .] </B>

(7.4.21) <B> [. . .] I listen to music (uh) yesterday I discovered (uh) an American singer called <name of singer> who did a song called <name of the song> she's very new (uhm) but she's *very fun* it was like she's my new spirit animal [. . .] </B>

(7.4.22) <B> [. . .] we (uhm) had a programme to (uh) in the night so (uh) it was snowing also in the night and it was not that good but it was *very fun* I was (uh) driving my own sled yeah [. . .] </B>

All three combinations occur in predicative position, they complement the subject of the sentence (*it*, *she*, *it*) and are not followed by a complement. In (7.4.20) and (7.4.22) the tense is in the past, while (7.4.21) is in the present simple. Both in (7.4.20) and (7.4.22) the combination is used in an *it + copular verb* structure with reference to an event: in (7.4.20) the speaker is talking about her experience abroad in Germany and the way school is organised differently there; in (7.4-22) the speaker is recounting a trip to Norway where she experienced dog-sledding at night. Excerpt (7.4.21) instead contains the combination used in reference to a female singer, so that the combination is a personal evaluation of a person. It is interesting to note that in PEC there are only two explicit references to people, while the remaining occurrences describe events, facts, objects (i.e., *scrittura*, *video*, *lettera*, videogames, theatre shows, skiing, school, and many others). As regards the two other cases, (7.4.20) and (7.4.22), both sentences show similar features to the Italian L1 equivalent in terms of references, present or past tense, use of personal or impersonal pronouns.

7.5 Discussion

In this Chapter, we have presented a qualitative analysis of the second set of collocations extracted from the ISLC corpus. These collocations were selected because of their infrequency or non-occurrence in the BNC and their frequency in ISLC (≥ 2). A scheme was employed for the analysis following the diagram set out in Figure 4.4.1 which required the investigation of three macro-variables: collocational background, learner metadata, and text variables. The collocational background required the definition of the combination's etymology based on the OED. This was reported in Table 6.1 and a label was assigned to each collocation on the basis of its etymology: shared, mixed, different. As opposed to the first set of frequent collocations in ISLC (see Chapter 6), the majority of combinations analysed in this Chapter have different or mixed etymology; in one case only is the etymology shared

(i.e., *super strange*). This already anticipates that probably the learners did not rely on their L1 knowledge for the production of these combinations. As it was mentioned in Section 4.4.2.3, a similar etymology was hypothesised to be a good predictor of L1 congruency; this is not always the case, but as it has been shown in the previous Chapter, a mixed or shared etymology usually points towards L1 congruency, while different or mixed etymology is usually an indication of L1 incongruency. The second collocational variable, L1 congruency, was verified on the basis of a similar approach to Wolter and Gyllstad (2013), with the addition of seven native speakers of Italian for the task of L1 congruency translation and evaluation. In the case of this set of collocations, not all the translations provided for the collocations by the seven native speakers indicated that the combinations were L1 congruent: indeed, five collocations were incongruent. The incongruent combinations were labelled as such due to the lack of translations or because the translators did not reach an agreement on the suitable word-by-word translation. The remaining four combinations were deemed L1 congruent following the native speakers' evaluation, but instances of the L1 equivalents occurred in PEC only in two cases: *pretty curious* and *very fun*. The findings regards L1 congruency further corroborate the initial result from the etymology variable, namely that the learners probably did not rely on their L1 equivalents for the production of these collocations. Lastly, as regards the CEFR level, which was verified through WFF, this set of collocations seems to have an increased CEFR level as opposed to the previous set where the collocational items all belonged to the A1 or A2 level. In this case, three adjectives belong to the B1 level, one belongs to the C1 level, while as regards the adverbs, the majority are still categorised as A1 or A2, but there is also B1 in the case of *kind of*. These results offer an interesting and perhaps unexpected perspective: the previous set of frequent and attested combinations was composed by CEFR beginner level collocations and lexical items, whereas this set of infrequent or unattested combinations is composed by upper-beginner/intermediate level lexical items.

As discussed in the previous Section (6.5), the L1 has been reported in the literature to often have an influence on formulaic language (Bahns, 1993; Granger, 1998b; Nesselhauf, 2005; Wang, 2016; among others) and L1 incongruency has been demonstrated to hinder the production of L1 collocations (Bahns, 1993; Wolter & Gyllstad, 2011; 2013; Yamashita & Jiang, 2010; among others). Following on from this, it was hypothesised that L1 congruency may thus aid the production of collocations, even infelicitous ones, such as those who are particularly infrequent or not present in a large reference corpus, such as the BNC. This was mainly based on the works of Bahns (1993), Bahns and Eldaw (1993), Biskup (1992), Ellis (2008), Nesselhauf (2003) who found that negative L1 transfer may be the source of deviations in the production of collocations. Furthermore, Koya (2003) and Huang (2011)

found that it is the most proficient learners of L2 that rely far more greatly rely on their L1 for the production of collocations. However, the analysis has so far indicated that the learners' production of this set of combinations (which occurred fewer than five times or not at all in the BNC) did not rely on L1 congruency, as only four of these combinations are L1 congruent and only two of these are frequent enough to occur in a 26-million L1 reference corpus (PEC).

The second set of variables which were taken into consideration for the analysis of these nine combinations were learner metadata: these included gender, stay-abroad experience, university courses, and other foreign languages. The variables were also included in a Table (7.3) in order to systematise the analysis as much as possible. Although it is difficult to state whether there is a significant correlation between these variables, it seems that the most probable user of one of the combinations analysed in this Chapter is a female learner, with a short experience abroad, attending a postgraduate course, and studying French as L3. In the previous Chapter, a note was made regarding learners with an experience abroad of over six months: Tracy-Ventura et al. (2016) had demonstrated that the learners who had spent the longest amount of time abroad produce a higher number of collocations and show better performance. Following on from this, the opposite may be true, that is, learners with a short experience abroad may be more prone to the production of creative collocations, lower production of attested collocations, more erroneous use of them. However, although the majority of learners who produced this set of combinations had a short experience abroad, in four cases out of nine, the learners also had a long experience abroad, so that it is difficult to identify a specific correlation. As far as the university courses are concerned, four learners attended an undergraduate course, while 11 a postgraduate course, specifically n° 10 (seven learners). The reason for the higher number of postgraduate students may be that this group of learners are more aware of the need to produce a greater range of varied collocations; this could be a response to widespread teaching practices in Italy of encouraging students to employ a wider range of vocabulary as they become more advanced. Lastly, as regards the L3, little research has been done on the influence of the L3 on the L2 so that the high number of learners studying French L3 does not have a clear correlation to the production of collocations. However, it has been noticed that French L3 may have been responsible for the production of one combination, *really really really good*, due to the presence of double (triple) intensification of the adjective *good*, which resembles the often doubled French *très*.

As regards the third macro-category of variables, namely textual ones, these were taken into consideration during the analysis of these nine combinations. The results, as already anticipated by the analysis of the other variables so far, indicate that there seems to be no direct link between the production of these combinations and L1 transfer. The following

paragraphs analyse the results of the analysis, first collocation by collocation, and then concurrently. A note should be made as regards the lexico-grammatical pattern: it has been practically impossible to compare the lexico-grammatical pattern of the collocations to the L2 and L1 native speakers' due to the lack of occurrences in LOCNEC and PEC (for most collocations). However, it may be reasonable to expect, based on the previous analyses of Chapter 6, that the learners have somewhat mastered the lexico-grammatical patterns of collocations, although these are not as frequent as those analysed in the previous Chapter.

The first collocation analysed is *absolutely scared*, which has a mixed etymology, but is L1 incongruent. This collocation was produced during the third task, thus during a highly cognitive-demanding and prompted task. It has been noticed that, as regards pragmatic competence (which has been previously shown to pose a few problems for the learners), the learners may seem to struggle. In this case, the learners showed difficulties with the correct choice of adjective for the intensifier *absolutely*. Indeed, Biber et al. (1999) indicate that *absolutely* is generally used to intensify an adjective expressing a positive evaluation, whereas in this case the intensifier was used in combination with an adjective expressing fear, a negative feeling. Furthermore, the lack of L1 equivalent and L1 congruency show that the production of this collocation was not an effect of L1 transfer.

The second collocation analysed is *kind of scared*, which has a different etymology as well as being L1 incongruent. This was also produced during the third task as *absolutely scared*, thus during a demanding and prompted task. In this case, *kind of* has been identified as a chiefly American English intensifier. This may explain the infrequency of the combination in a reference corpus of British English, such as the BNC. In Biber et al. (1999) there were no indications as regards what type of adjective can be modified by *kind of*, so that a similar comparison in terms of pragmatic competence with *absolutely scared* was not possible. Lastly, the lack of L1 equivalent and L1 congruency show that the production of this collocation too was not an effect of L1 transfer.

The third collocation analysed is *pretty curious*, which has a mixed etymology, but is L1 congruent. The collocation was also produced during the third task, as with the previous two combinations. The L1 equivalent translation *abbastanza curios** was found only six times in PEC (0.23 times per million words), thus indicating a rather infrequent use of the combination in Italian. However, despite the infrequency of the combination, a few elements were identified from the analysis: the combination occurs mainly in written texts (an academic article, a book, a newspaper, a wikipedia page) and only twice in spoken language (TV show); it is used to define a multiplicity of subjects (events, someone's behaviour, a software release); and it is mainly constructed in predicative position complementing the subject of the sentence (i.e., *Questo dato è abbastanza curioso, anche se facilmente spiega-*

bile). However, it must be pointed out that the L1 equivalent word-by-word translation of *pretty curious* acquires a different meaning in Italian when translated to *abbastanza curios**, which could be approximated to *rather peculiar.* Despite the only two instances of *pretty curious* in ISLC, it has been noted that the combination is used to describe the attitude of the characters of the story-telling task, following a pattern which already emerged from the previous analyses, namely that of attributing the description of a feeling or object to a character rather than the speaker uttering the sentence (i.e., *they're pretty curious about it*). In this case, L1 congruency suggested that there may have been an L1 transfer effect, but this was discarded as the L1 equivalent has been shown to acquire a different meaning, namely that of *rather peculiar*, so that, although there could be another L1 equivalent translation, there is no direct L1 effect link between *pretty curious* and *abbastanza curios**.

The fourth collocation analysed is *quite old-fashioned*, which does not share the same etymology, but has been regarded as L1 congruent. However, as regards the congruency, the translation provided by the native speakers comprises three words rather than two, so that it cannot be considered a pure word-by-word translation. Furthermore, *abbastanza vecchio stile* does not occur in PEC, so that even if the translation were acceptable from the point of view of equivalent congruency, the combination is too infrequent to occur in the reference corpus. In ISLC, the combination was produced to describe a film and a singer, but no other elements were identified in the analysis. The fact that there are no L1 occurrences in PEC suggests that the learners did not rely on their L1 to produce this combination.

The fifth collocation analysed is *really heartbreaking*, which has a mixed etymology, but is L1 incongruent. Indeed, no agreement was reached among the native speakers as regards a univocal translation equivalent. The combination was produced twice by one learner only as a repetition of the first instance. Furthermore, the learner showed clear signs of hesitation before producing the second instance of *really heartbreaking*, so that it is plausible to hypothesise that the learner was trying to retrieve another lexical item, but as she failed to do, she fell back on the previously instanced combination. As regards potential L1 transfer effects, no such link was identified in the analysis.

The sixth collocation analysed is *really really amazing*, which has a mixed etymology, but is L1 incongruent. It was produced twice by the same learner and the most interesting aspect about this collocation is the presence of double intensification, which is also present in Italian (called *reduplicazione espressiva*): *really really.* A further inspection of the learners' text reveals that this speaker heavily relies on the intensifier *really.* Although there is not enough evidence to expand the research and the learner's baseline speech is not available, this use of the combination may lead to the following two conclusions: the learner is either displaying a simplification strategy behaviour by heavily relying on the intensifier *really*

(which leads to double intensification), or her personal communication style comprises the use of double intensification. In either case, no L1 transfer effect has been identified in the analysis.

The seventh collocation analysed is *really really really good*, which has a mixed etymology and it is not L1 congruent as the native speakers did not reach agreement about its L1 equivalent translation. Similarly to the previous collocation, *really really really good* presents double (or in this case triple) intensification, which is a common phenomenon in spoken language and in Italian too (viz. *reduplicazione espressiva*). However, despite the shared intensification phenomenon, the L1 native speakers were not able to reach an agreement about the word-by-word translation. In both cases in which the combination was uttered, the speakers were providing a description, an evaluation of a film or a TV show episode, so that it is possible to hypothesise that they were perhaps attempting to mark the adjective *good* on the high-end point of the scale. As a replacement of a potentially better lexical item, they relied on a teddy-bear intensifier such as *really* to increasingly mark the adjective. As regards L1 transfer effects, no direct link between the L1 and the IL was identified in the analysis.

The eighth collocation analysed is *super strange,* whose etymology concides with Italian and it is thus L1 congruent. The collocation is instanced three times by the same speaker, IT012, and its L1 equivalent does not occur in PEC. All three cases of *super strange* occur in predicative position and they are used to express a feeling (i.e., *it's super strange to be here now at university*). Similarly to *really really amazing,* the learner's text was further inspected and other instances of *super* as an intensifier were found. Indeed, IT012 with 11 occurrences of *super* produced 75% of all instances of the intensifier in ISLC. As with the other combinations, it may be possible that the combination arose in the learner's production as a consequence of her personal communication style, which includes the use of *super* as an intensifier. This could have been picked up by exposure to American English, in which the intensifier has been identified to be very frequent[1]. An L1 transfer effect has been discarded as *super* it is indeed used as an intensifier in Italian, but it is not frequent (only occurring 2.87 times per million words).

The last collocation analysed is *very fun*, whose etymology is mixed, and it is L1 congruent. The particular aspect of this collocation is the employment of *fun* as an adjective, rather than a noun, which is its traditional form. Indeed, in informal American English *fun* becomes an adjective (Biber et al., 1999, p. 412) which makes it modifiable by an adverb,

[1]In the BNC, the adverb *super* occurs 1.95 times per million words. In a reference corpus for American English, the Contemporary Corpus of American English (COCA), the adverb *super* occurs 40.81 times per million words.

such as *very* (Enstad, 2010). In PEC there are 66 instances of *molto divertent** (2.49 times per million words) indicating that the combination is not particularly frequent, but it is still used in Italian. This leads to two possible conclusions: the speakers were either predominantly exposed to American English which led them to the production of *very fun*, or the speakers were not aware of the peculiar behaviour of *fun* in informal American English, but they produced the collocation because it is the L2 congruent translation of *molto divertent**. More data is needed to reach a solid conclusion, but in this case L1 could have played a role.

As mentioned in the Discussion section (6.5) of Chapter 6, the majority of research on collocations in learner language has focused on investigating the overuse/underuse, misuse, and L1 influence on learner collocations (see Section 3.5.1). Chapter 5 was dedicated to the investigation of the first variable (overuse/underuse), while Chapter 6 focused on whether the lexico-grammatical patterns of a set of frequent ISLC collocations presented any differences or similarities with those produced by the native speakers, thus practically addressing the second variable of misuse. This Chapter instead was concerned with investigating whether the production of infrequent/unattested collocations (as identified in Chapter 5) is a consequence of L1 (negative) transfer effects. The data available for the analysis was limited to two or three instances of these combinations in ISLC, since no occurrences were found in LOCNEC. Furthermore, L1 congruency was verified and only in four cases the combinations had L1 equivalent translations; however, only two of these four translations occurred in the L1 reference corpus PEC. The data was not enough to carry out a complete lexico-grammatical study from a Contrastive Interlanguage Perspective, but the L1 incongruency and infrequency of the L1 equivalents indicate that it is likely that there was no L1 transfer effect on the production of this set of collocations. Furthermore, although no relevant research was found in this regard, it should be noted that the collocations which contained double intensification (*really really amazing, really really really good*) were produced in both cases by learners who study French as L3. Double intensification is a common phenomenon in informal spoken French (S. Cigada, personal communication, 23rdNovember 2020), so that more research into the effects of the L3 on the L2 is needed in order to verify whether there is indeed a correlation between these phenomena.

This finding is in contrast with most of the literature on L1 congruency and collocations since the majority of studies agree that L1 congruency may aid the learners both in terms of collocation processing (Wolter & Gyllstad, 2013; Wolter & Yamshita, 2018; among others) and collocation production (Biskup, 1992; Ellis, 2008; Granger, 1998b; Koya, 2003; among others), so much so that Bahns (1993) suggested that congruent collocations do not need to be taught to language learners since the learners may already be aware of these due to their L1 equivalents. This was not verified by subsequent studies by Wolter and Gyllstad (2015)

and Nesselhauf (2003; 2005) who found no significant evidence of L1 activation in a lexical decision task and similar percentages of errors in congruent and incongruent collocations respectively. In particular, these results clash with Koya's (2003) finding that advanced proficiency learners of L2 rely heavily on the L1 for their collocation production following the U-shaped type of transfer behaviour posited by Kellerman (1979). If both low and advanced proficient learners of L2 rely on L1 for the production of collocations, then it follows that when the learners are not relying on the L1 they may be more prone to produce creative or incorrect collocations (cf. Biskup, 1992). Studies which have concentrated on the effect of L1 congruency and L1 transfer have mainly employed either a psycholinguistic approach (thus with the use of experimental elicited data) or have used written learner corpora.

The present work has extracted infrequent/unattested collocations from a spoken learner corpus of advanced learners of English L2 and found no clear evidence of L1 negative transfer effects. On the other hand, evidence was found of L1 positive transfer effect on the production of frequent collocations (see Chapter 6), thus confirming the general agreement in the literature about L1 congruency facilitating the production of collocations. However, in some cases L1 congruency aided the production, but not the correct pragmatic function of the collocation in the L2. The combination of the results suggest that collocation production in spoken language diverts from the written competence and performance, although some elements are retained: L1 congruency can help learners produce a high number of frequent and congruent collocations, although they may also transfer their L1 pragmatic function to the L2 (which in some cases may be different); L1 congruency does not directly affect the production of infrequent/unattested collocations as these seem to be more a product of personal communicative style or exposure to other English varieties (e.g., American English).

Chapter 8

Conclusions

THE aim of this chapter is to offer a summary of the research findings and describe the contributions that these results have for both the LCR and SLA communities. After identifying the limitations of the study, future directions and applications are presented with t he hope that they could potentially aid teachers and scholars in improving learner performance of spoken collocations and phraseology.

8.1 Summary of research findings

This study investigated the use of adverb + adjective collocations in a spoken learner corpus of advanced Italian learners of English L2 on the basis of three research questions:

1. How do advanced Italian learners of English and native speakers of English compare in their production of different adverb + adjective collocations in spoken language?

2. What is the difference between the syntactic patterns and lexical meaning of the adverb + adjective collocations produced in spoken language by advanced Italian learners of English and native speakers?

3. Does L1 congruency have a transfer effect on the production of infrequent/unattested adverb + adjective collocations generated in spoken language by advanced Italian learners of English?

The analysis stemming from these research questions provided the following findings:

1. Italian learners of English produce a fewer number of collocations; a higher number of non-collocations; and their collocations are less collocational than native speakers'

2. Not all collocations produced by Italian learners of English show different syntactic patterns and lexical meaning from those of native speakers; when learners deviate from the native speakers' baseline, they do so mainly for pragmatic reasons

3. No evidence of L1 (negative) transfer was found in relation to the learners' production of infrequent/unattested collocations

These findings can contribute to the LCR and SLA communities in a number of ways, as will be listed below. They also have some implications for Italian teaching practices of English L2.

8.2 Contribution to LCR and SLA

The first finding corroborates the previous LCR results on learners' collocations, namely that learners, even at advanced levels of proficiency, struggle with collocations. This can result in a lower number of collocations, or a higher number of a specific set of them ("teddy bears", Hasselgren, 1994). If the collocations' association measures are taken into consideration, it is often the case that learners produce a higher number of frequent collocations (as identified by their *t*-score) and a lower number of strongly associated collocations (as identified by their MI score). The present study has confirmed this general trend, but has added a further piece to the puzzle since most of the research on learners' collocations has been carried out either on written or elicited data; in this study the collocational lag has been found in spoken data and it has been demonstrated that, contrary to most findings, learners struggle with the production of *t*-score collocations and their collocationality.

The fact that the learners' collocations are less collocational than native speakers' can also inform SLA researchers that when it comes to spoken language, EFL learners are perhaps exposed to an insufficient amount of spoken input and they are, most probably, more used to written input/output. This could potentially challenge some of the tenets of SLA, namely that L2 second language acquisition displays similar patterns to L1 acquisition and that formulaic sequences are normally acquired as strings of words by (L1) learners. Indeed, in an EFL context such as the Italian one, the L2 acquisition process could be hindered by the lack of appropriate input. This lack is due to the several factors, which can vary from a tendency to dub foreign films in the native language to teaching practices which mainly focus on syntax and the written form. However, regardless of the factors affecting the amount and quantity of appropriate input, SLA theories should take into account the lessened acquisition processes taking place in the EFL classroom as these could affect learner performance,

such as in this case. Furthermore, and as regards collocations specifically, the assumption that formulaic language is acquired in chunks can be challenged by the EFL environment in which the lack of exposure to the target language greatly reduces the amount of formulaic language and thus the ability to acquire it in chunks. In any case, even if learners are able to acquire some formulaic sequences as a whole, these are often sequences found in the written language, rather than speech. Thus, EFL learners are frequently found to be producing formulaic sequences typical of written texts in their speech (cf. Forchini, 2012; Murphy, 2003).

The second finding will be of further interest to the LCR community by providing evidence that advanced learners tend to use collocations correctly from the point of view of syntactic patterns, although these are less varied compared to the native speakers. This means that, for example, they use a more restricted set of prepositions or use the collocations with a predominant negative or positive connotation. This could indicate that the learners have acquired the basic knowledge of given collocations but lack the full mastery of their syntactic use, thus prompting LCR researchers to investigate the syntactic patterns of spoken collocations.

However, the learners have been shown as regards the pragmatic functions they assign to collocations and this could be representative of several SLA phenomena: for example, by assigning a different pragmatic function to a collocation, the learners could be showing avoidance, for lack of knowledge or other reasons. In addition, the creative pragmatic functions could also be instances of fossilisation whereby the learners adapt one collocation's pragmatic function to a set of different contexts, without ever acquiring the appropriate collocation. These and other hypotheses regarding the creativity shown by the learners are fertile ground for SLA theories and studies which may further investigate this topic, either through learner corpora or with clinical and/or experimental data.

The third finding is a significant addition to the LCR panorama, since L1 congruency has been shown to play a key role in learners' production of collocations, but this does not seem the case in the present study. Indeed, no evidence of L1 (positive or negative) transfer was found either for L1 congruent or for L1 incongruent collocations. This goes against mainstream studies (cf. Bahns, 1993; Granger, 1998b; Nesselhauf, 2005; Wang, 2016; among others) showing an effect as regards L1 congruency; however, since the majority of these studies were based on either written or elicited data, the present work adds another tile to the picture and urges further investigations into other potential variables at play.

Similarly, the SLA community could benefit from this result since L1 negative transfer has long been associated with the production of wrong formulaic sequences, but this study has demonstrated that other processes must be affecting the production of collocations for

advanced learners.

It is also worth mentioning some of the contributions that the study design has also produced for the LCR and SLA communities. Among them, I identified four significant points of reflection:

1. ISLC is a new spoken learner corpus which is available to the LCR and SLA scholars and will be further expanded to include other proficiency levels
2. The participants' proficiency level was assessed before including their texts in the corpus
3. The original LINDSEI's third task was replaced by a newly adapted story-telling task
4. The quantitative analysis replaced Durrant and Schmitt's (2009) cut-off points with a more inclusive association measure score approach

As regards the first point, the compilation of ISLC was not solely aimed at setting up an ad-hoc corpus for the present research, but it was also designed with future research perspectives in mind. The corpus is an improved and newer version of the Italian component of LINDSEI and as such it is highly comparable with all other LINDSEI components; in addition, it is a further addition to the learner corpora panorama which is still largely dominated by written corpora.

The second design choice is a further improvement of the LINDSEI design criteria. The learners' proficiency was assessed with a standardised language test, which the literature has shown to be a more robust proficiency measure. Although standardised language tests are not the perfect tool for proficiency assessment (and this will be discussed in the section on limitations), the present research has shown that with minimal effort it is possible to obtain a more solid and homogeneous sample of a given proficiency level.

The third point directly built on Gràf's (2017) suggestions to further improve the already ground-breaking feat of LINDSEI. One of the main issues that was pointed out by the scholar during the compilation of the Polish component of LINDSEI was the picture description task. This task was suggested to be an unproductive source of language and, when the participants produced short texts, no clear indications had been set out to help proceed with the data collection. This study opted to replace this task with a prompted story-telling, which was shown to increase the number of words produced by the learners.

Lastly, the fourth point refers to the choice of not adopting Durrant and Schmitt's (2009) cut-off points for the analysis of collocations; this was done because the cut-off points were not applicable on spoken data since neither the learners nor the native speakers produced

collocations with an MI score greater than 7. The different approach adopted suggests that there are significant differences between written and spoken language in terms of collocations. This needs to be taken into consideration when relying on a solid and tested methodology such as Durrant and Schmitt's, which requires adjustments for different types of data, such as spoken language.

8.3 Limitations

The study was characterised by a number of limitations. First of all, ISLC was compiled with limited resources and as such, for the moment, it is a small corpus. The corpus is small both as regards the number of tokens (approximately 60,000 learner turns only) and the number of participants which amount to 34. Indeed, involving students in the project proved to be extremely difficult for two main reasons: the first is strictly related to the nature of the research, spoken language. Students were willing to participate but as soon as they were told they were meant to hold a conversation in English and it was going to be recorded, the majority did not feel particularly comfortable.

Secondly, in order to obtain a large and diverse number of participants, involvement of multiple universities was required as well as the necessary resources. This resulted in a limited diversification of participants who come from only three universities. Secondly, although a step was taken towards operationalising proficiency from a textual perspective through a standardised test, this method still has its limitations. First and foremost, the test assesses comprehension and grammar/lexical knowledge, but it does not verify speaking skills and due to the limited resources available to the researcher it was not possible to rely on the evaluations of trained raters. I will attempt to address this issue in the future by training a number of raters who will rate the texts based on a specific rubric.

Furthermore, another aspect which has to be taken into consideration and which is strictly related to the limited size of the corpus is the fact that the findings, especially the quantitative ones, cannot be generalised to the whole Italian advanced learners of English L2 population. This is mainly due to the fact that there are not enough speakers and texts to allow for a generalisation. It is true that there is no optimal size for a corpus, but to allow for generalisations, the corpus has to be representative of a given population. In this case, the corpus does not offer a complete picture of the learner situation across all regions of Italy. This means that different teaching practices were not taken into consideration and it may be that different universities employ different approaches to language learning. This in turn could be reflected in better collocational performance or even opposite displays of

behaviour (overuse of collocations). However, small corpora also have their benefits, as they "provide tailor-made solutions to their problems" (Gilquin, 2015, p. 15). As already mentioned above, the corpus expansion is under way, and it will attempt to include a larger number of participants from different universities (located in different regions of Italy).

Despite the qualitative approach to the analysis of two sets of collocations not requiring a large sample of data, it was difficult to draw some solid conclusions about a few collocations, due to the lack of data. This was especially true in the second part of the qualitative analysis where no instances of the collocations were available in the reference corpus LOCNEC and only few occurred in the BNC and PEC corpora. This limited the analysis and did not allow for an in-depth analysis of the lexico-grammatical features of the collocations. In addition, in cases of L1 congruency, often the collocations did not occur in the Italian reference corpus PEC, which again hindered the analytical work. The only way to tackle this is to have a larger amount of data, but at the moment this is not possible as both LOCNEC and PEC are limited corpora.

8.4 Desirable applications and future perspectives

The study has several implications for future research and applications, chief among them for Italian teaching practices of L2. Indeed, the collocations produced by the learners lead one to question some of the teaching practices adopted in the universities attended by the subjects. First of all, the three universities have not included phraseology in their syllabi and this is reflected in the learners' collocational gap even at advanced level. The courses seem to be concerned mainly with the teaching of morphology and syntax, text genres, discourse analysis. Although these are certainly essential topics for the solid development of a second language, the research findings suggest that an implementation of phraseology into the course syllabi would be a desirable integration which may also boost the learners' collocational knowledge.

Secondly, the present results and mainstream research have shown that there is difference between the learners' written and spoken performance; however, this does not seem to be addressed by the Italian universities. The lack of a clear distinction between the teaching of written and spoken grammar, as it is also suggested by reference grammars such as Biber et al.'s (1999) and Carter and McCarthy (2006), could be attributed to the collocational lag of the Italian learners. It would be beneficial to distinguish between the two modes in the L2 teaching practices so that learners could become aware of differences and associate specific formulaic sequences to one mode or the other (cf. Forchini, 2012; Murphy, 2003).

In addition, more phraseology-targeted activities could also be implemented in ordinary teaching so as to lessen the foreignsoundingness of the learners.

A way to educate students about the differences between spoken and written collocations and, more in general, language is introducing corpora in the classroom. Indeed, the use of spoken (or written) corpora in the EFL classroom can achieve two desirable objectives: first, it introduces corpora in the (university) classroom through the main door, rather than the back one (Bernardini, 2004, p. 15) with all the benefits that this reaps, and secondly it provides learners with a greater amount of input. This promotes phraseology, but also other categories such as grammar, lexis, and pragmatics. In this regard, research projects such as the one under development by Forchini (Forchini, 2021) consisting in the compilation of a reference corpus of movie language for research and pedagogical purposes are particularly commendable. In addition, the ISLC corpus will be further expanded and extended by collecting texts of students from other universities (and proficiency levels) so that a larger spoken learner corpus of Italian learners of English will be available to scholars wishing to analyse or use the data for pedagogical purposes.

As regards introducing corpora in the classroom, two main advantages are brought to the learners: first of all, as the literature has amply documented (cf. Sinclair, 2004), corpora can either be sources of insights into the second language (both from the teaching or learning perspective; cf. Conrad, 2004; Tsui, 2004) and they can directly influence learning and teaching processes. Although a review of the literature on the topic is beyond the scope of this Chapter, it is worth acknowledging that many studies are now placing corpora and data-driven learning practices at their core. The most recent ones are Forti's (2019) exploration of L1 incongruent collocations and how data-driven learning may positively impact L2 learning and Gablasova et al.'s (2019) contribution to the volume on learner corpora and language teaching with a case study based on the TLC and its helpful corpus-based addition to the teaching of expressions of disagreement, language adjustment, and engaged listenership.

Since the research findings have highlighted some pitfalls in L2 teaching practices in Italy and corpora in the EFL classroom could address some of these issues, it is worth mentioning that the use of corpora could also break a vicious circle that has been noted in the literature. Indeed, despite the emphasis that has been placed on integrating corpus literacy into teacher training programmes and the popularisation of (learner) corpora, the practice of English language teaching is little affected by corpus research (Callies, 2019; Römer, 2011). Indeed, both Mukherjee's (2004) and Callies' (2019) investigation into corpus literacy among German teachers of English have revealed that the use of corpora have not been sufficiently addressed in language teacher education. This is regrettable, at least from the perspective of

phraseology, since as it has been already noted, EFL learners are exposed to very little natural and spoken language, and the use of spoken corpora (especially a movie corpus which is complemented by audio-visual material too) may bridge the gap by offering linguistic insights and supplement teaching practices. Most importantly though, if university learners who intend to become language teachers learn how to adopt and use corpora, they may be more willing to use them in their teaching practices once they leave university and become qualified language teachers. This could mean introducing corpora and data-driven learning at lower levels of proficiency, thus impacting language development in a more significant manner.

In conclusion, the results have sparked a series of reflections mainly on Italian teaching practices of L2 and potential applications of corpora in the classroom. The fact that advanced learners of English show persistent difficulties with collocations, although not unexpected, is not ideal and reveals that the English L2 Italian teaching environment has not yet caught up with research advocating for more spoken language input and phraseology-targeted activities. It also points towards a lack of distinction between written and spoken grammar, to the detriment of naturalsoundingness and a weak acquisition of indicators of fluency, such as phraseology. Furthermore, it has been proposed that the use of spoken corpora in the EFL classroom, aside from the standard benefits from the use of corpora, could be a rich source of natural linguistic insights to the learners (and teachers) while encouraging corpus literacy for the learners who will then become L2 teachers. It is my hope that the universities involved in this research project will be able to benefit from the findings and review their course contents and pedagogical applications.

References

Ackerley, K. (2013). A comparison of learner and native speaker writing in online self-presentations: Pedagogical applications. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of learner corpus research. Looking back, moving ahead*, (pp. 1–10). Presses Universitaires de Louvain.

Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp, & M. B. Díez-Bedmar, (Eds.), *Linking up contrastive and learner corpus research,* (pp. 35–53). Rodopi.

Ädel, A. (2014). Selecting quantitative data for qualitative analysis: A case study connecting a lexicogrammatical pattern to rhetorical moves. J*ournal of English for Academic Purposes*, *16*, 68–80. https://doi.org/10.1016/j.jeap.2014.09.001

Ädel, A. & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

Aijmer, K. (2009). "So er I just sort of I dunno I think it's just becase. . . ": A corpus study of 'I don't know' and 'dunno' in learner spoken English. In A. H. Jucker, D. Schreier, & M. Hundt (Eds.), *Corpora: Pragmatics and Discourse*, (pp. 151–166). Rodopi. https://doi.org/10.1163/9789042029101_009

Aijmer, K., Altenberg, B. & Johansson, M. (Eds.)(1996). Languages in contrast. Papers from a symposium on text-based cross-linguistic studies. Lung: Studies in English 88. Lund University Press.

Aisenstadt, E. (1979). Collocability restrictions in dictionaries. *International Journal of Applied Linguistics*, *45*(1), 71–74. https://doi.org/10.1075/itl.45-46.10ais

Aisenstadt, E. (1981). Restricted collocations in English lexicology and lexicography. *International Journal of Applied Linguistics*, *53*(1), 53–61. https://doi.org/10.1075/itl.53.04ais

Al-Zahrani, M. S. (1998). *Knowledge of English lexical collocations among male Saudi college students majoring in English at a Saudi University.* [Unpublished doctoral dissertation]. Indiana University of Pennsylvania.

Alexopoulou, T., Michel, M., Murakami, A. & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*(1), 180–208. https://doi.org/10.1111/lang.12232

Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS Learner Corpus. *Komaba Journal of English Education*, *1*, 105–27.

Allerton, D. J., Nesselhauf, N. & Skandera, P. (2004). *Phraseological units: Basic concepts and their application*. Schwabe.

Altenberg, B. (1991). Amplifier collocations in spoken English. In S. Johansson, & A. Stenström (Eds.), *English computer corpora: Selected papers and research guide*, (pp. 127–148). Walter de Gruyter.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*, (pp. 101–122). Oxford University Press.

Altenberg, B. & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173–195. https://doi.org/10.1093/applin/22.2.173

Amosova, N. N. (1963). *Osnovui angliiskoy frazeologii* [The foundations of English phraseology]. University Press.

Andersen, R. (1977, February). *The improved state of cross-sectional morpheme acquisition/accuracy methodology* [Paper presentation]. Los Angeles Second Language Research Forum, Los Angeles, California.

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In L. Anthony, S. Fujita, Y. Harada, & W. Daigaku (Eds.), *Proceedings of IWLeL: An interactive workshop on language e-learning,* (pp. 7–13). Tokyo: Waseda University.

Anthony, L. (2019). *AntConc* (Version 3.5.8) [Computer Software]. https://www.laurenceanthony.net/software

Arnon, I. & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82. https://doi.org/10.1016/j.jml.2009.09.005

Aston, G. (2011). Applied corpus linguistics and the learning experience. In V. Viana, S. Zingier, & G. Barnbrook (Eds.), *Perspectives on corpus linguistics*, (pp. 1–16). John Benjamins. https://doi.org/10.1075/scl.48.01ast

Atkins, S., Clear, J. & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, *7*(1), 1–16. https://doi.org/10.1093/llc/7.1.1

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L.F. & Cohen, A.D. (1998). Language testing–SLA interfaces: An update. In L.F. Bachman, A.D. Cohen, & M.H. Long (Eds.), *Interfaces between second language acquisition and language testing research*, (pp. 1–31). Cambridge University Press.

Bachman, L.F. & Palmer, A.S. (1996). *Language testing in practice*. Oxford University Press.

Badecker, W. (2001). Lexical composition and the production of compounds: Evidence from errors in naming. *Language and Cognitive Processes*, *16*(4), 337–366. https://doi.org/10.1080/01690960042000120

Badecker, W. & Allen, M. (2002). Morphological parsing and the perception of lexical identity: A masked priming study of stem homographs. *Journal of Memory and Language*, *47*(1), 125–144. https://doi.org/10.1006/jmla.2001.2838

Bahns, J. (1993). Lexical collocations: A contrastive view. *ELT Journal*, *47*(1), 56–63. https://doi.org/10.1093/elt/47.1.56

Bahns, J. & Eldaw, M. (1993). Should we teach EFL students collocations? *System*, *21*(1), 101–114. https://doi.org/10.1016/0346-251X(93)90010-E

Baker, P. & Egbert, J. (2018). *Triangulating methodological approaches in corpus-linguistic research*. Routledge. https://doi.org/10.4324/9781315724812

Bardel, C. (2015). Lexical cross-linguistic influence in third language development. In H. Peukert (Ed.), *Transfer effects in multilingual language development*, (pp. 117–128). John Benjamins. https://doi.org/10.1075/hsld.4.05bar

Barfield, A. (2007). *An exploration of second language collocation knowledge and development* [Unpublished doctoral dissertation]. University of Swansea.

Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag.

Becker, J. D. (1975). The phrasal lexicon. In R. Schank, & B. Nash-Webber (Eds.), *Proceedings of Interdisciplinary Workshop on Theoretical Issues in Natural Language Processing*, (pp. 70–73). Yale University.

Benson, M., Benson, E., & Ilson, R. (1986). *The BBI combinatory dictionary of English*. John Benjamins.

Benson, M., Benson, E. & Ilson, R. (1997). *The BBI dictionary of English word combinations* (2nd ed.). John Benjamins. https://doi.org/10.1075/z.bbis

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. https://doi.org/10.1016/j.jslw.2014.09.004

Bestgen, Y. & Granger, S. (2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In S. Hoffmann, A. Sand, S. Arndt-Lappe, & L. M. Dillmann (Eds.), *Corpora and lexis*, (pp. 277--301). Brill; Rodopi. https://doi.org/10.1163/9789004361133

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257. http://otipl.philol.msu.ru/media/biber930.pdf

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamins. https://doi.org/10.1075/scl.23

Biber, D. & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D. & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers*, *26*, 181–190.

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. & Quirk, R. (1999). *Longman grammar of spoken and written English*. Longman.

Biskup, D. (1990). Some remarks on combinability: Lexical collocations. In J. Arabski (Ed.), *Foreign language acquisition papers*, (pp. 31–44). Uniwersytet Slaski.

Biskup, D. (1992). L1 Influence on learners' renderings of English collocations: A Polish/German empirical study. In P. Arnaud, & H. Bejoint (Eds.), *Vocabulary and applied linguistics*, (pp. 85–93). Macmillan. https://doi.org/10.1007/978-1-349-12396-4_8

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning*, *33*(1), 1–17. https://doi.org/10.1111/j.1467-1770.1983.tb00983.x

Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum*, *1*, 1–14.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t-test. *Psychological Bulletin*, *57*(1), 49–64. https://doi.org/10.1037/h0041412

Borin, L. (2004). New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and Language Learners*, (pp. 17–67). John Benjamins. https://doi.org/10.1075/scl.17.06bor

Bouton, L. F. (1988). A cross-cultural study of ability to interpret implicatures in English. *World Englishes*, *7*(2), 183–196. https://doi.org/10.1111/j.1467-971X.1988.tb00230.x

Bouwer, R., Béguin, A., Sanders, T. & van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83–100. https://doi.org/10.1177/0265532214542994

Bratankova, L. (2015). *Le collocazioni verbo+nome in apprendenti di italiano L2*. [Unpublished doctoral dissertation]. Univerzita Karlova v Praze.

Brezina, V. (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge University Press. https://doi.org/10.1017/9781316410899

Brezina, V., McEnery, T. & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173. http://dx.doi.org/10.1075/ijcl.20.2.01bre

Brindley, G. (1998). Outcomes-based assessment and reporting in language learning programmes: A review of the issues. *Language Testing*, *15*(1), 45–85. https://doi.org/10.1177/026553229801500103

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Bulon, A. (2020). Comparing the 'phrasicon' of teenagers in immersive and non-immersive settings. Does input quantity impact range and accuracy? *Journal of Immersion and Content-Based Language Education*, *8*(1), 107–136. https://doi.org/10.1075/jicb.18010.bul

Butler, C. S. (1997). Repeated word combinations in spoken and written text: Some implications for functional grammar. In R. A. Gatward, R. M. Vismans, C. S. Butler, & J. H. Connolly (Eds.), *A fund of ideas: Recent developments in functional grammar*, (pp. 60–77). Institute for Functional Research into Language and Language Use.

Caines, A. & Buttery, P. (2018). The effect of task and topic on opportunity of use in learner corpora. In V. Brezina, & L. Flowerdew (Eds.), *Learner corpus research: New perspectives and applications*, (pp. 5–27). Bloomsbury. http://dx.doi.org/10.5040/9781474272919.0007

Callies, M. (2009). *Information highlighting in advanced learner English: The syntax–pragmatics interface in second language acquisition*. John Benjamins. https://doi.org/10.1075/pbns.186

Callies, M. (2013a). Agentivity as a determinant of lexico-syntactic variation in L2 academic writing. *International Journal of Corpus Linguistics*, *18*(3), 357–390. https://doi.org/10.1075/ijcl.18.3.05cal

Callies, M. (2013b). Advancing the research agenda of interlanguage pragmatics: The role of learner corpora. In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2014: New domains and methodologies*, (pp. 9–36). Springer. https://doi.org/10.1007/978-94-007-6250-3_2

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), T*he Cambridge handbook of learner corpus research*, (pp. 35–55). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.003

Callies, M., Díez-Bedmar, M. & Zaytseva, E. (2014). Using learner corpora for testing and assessing L2 proficiency. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency. Perspectives from SLA*, (pp. 71–90). Multilingual Matters. https://doi.org/10.21832/9781783092291-007

Canale, M. & Swain, M. (1980). The theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1–47. https://doi.org/10.1093/applin/I.1.1

Cao, Y. & Philp, J. (2006). Interactional context and willingness to communicate: A comparison of behavior in whole class, group and dyadic interaction. *System*, *34*(4), 480–493. https://doi.org/10.1016/j.system.2006.05.002

Carlsen, C. (2009, March). *Proficiency levels in learner corpora – a source of error or an asset in SLA-research* [Paper presentation]. GURT 2009 Georgetown University Round Table, Georgetown, Washington.

Carlsen, C. (2012). Proficiency level – A fuzzy variable in computer learner corpora. *Applied Linguistics*, *33*(2), 161–183. https://doi.org/10.1093/applin/amr047

Carrol, G. & Conklin, K. (2014). Getting your wires crossed: Evidence for fast processing of L1 idioms in an L2. *Bilingualism: Language and Cognition*, *17*(4), 784–797. https://doi.org/10.1017/S1366728913000795

Carrol, G. & Conklin, K. (2017). Cross language lexical priming extends to formulaic units: Evidence from eye-tracking suggests that this ideas 'has legs'. *Bilingualism: Language and Cognition*, *20*(S2), 299–317. https://doi.org/10.1017/S1366728915000103

Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In Testing the English proficiency of foreign students.

Center for Applied Linguistics. [Reprinted in H.B. Allen, & R.N. Campbell, (Eds.) (1972) Teaching English as a second language: A book of readings (pp. 313-320). McGraw-Hill]

Cervantes, I. M. & Gablasova, D. (2017). Phrasal verbs in spoken L2 English: The effect of L2 proficiency and L1 background. In V. Brezina, & L. Flowerdew (Eds.), *Learner corpus research: New perspectives and applications*, (pp. 28–46). Bloomsbury. http://dx.doi.org/10.5040/9781474272919.0008

Chafe, W. (1986). Evidentiality in English conversation and academic writing. *Evidentiality: The linguistic coding of epistemology*, *20*, 261–272.

Channell, J. (1981). Applying semantic theory to vocabulary teaching. *ELT Journal*, *35*(2), 115–122. https://doi.org/10.1093/elt/XXXV.2.115

Chen, W. (2019). Profiling collocations in EFL writing of Chinese tertiary learners. *RELC Journal*, *50*(1), 53–70. https://doi.org/10.1177/0033688217716507

Chen, Y. H. & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology*, *14*(2), 30–49. http://dx.doi.org/10125/44213

Cheung, A., Matthews, S. & Lan Tsang, W. (2011). Transfer from L3 German to L2 English in the domain of tense/aspect. In G. De Angelis, & J. Dewaele (Eds.), *New trends in crosslinguistic influence and multilingualism research*, (pp. 53–73). Multilingual Matters. https://doi.org/10.21832/9781847694430-005

Choueka, Y. (1988). Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. *Proceedings of RIAO '88: User-oriented content-based text and image handling*, (pp. 609–623). Le centre de hautes etudes internationales d'informatique documentaire.

Christophel, D. M. & Gorham, J. (1995). A test-retest analysis of student motivation, teacher immediacy, and perceived sources of motivation and demotivation in college classes. *Communication Education*, *44*(4), 292–306. https://doi.org/10.1080/03634529509379020

Church, K. W. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29. http://dx.doi.org/10.5555/89086.89095

Church, K., Gale, W. A., Hanks, P. & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical acquisition: Using on-line resources to build a lexicon*, (pp. 115–164). Lawrence Erlbaum.

Cieslicka, A. (2006). Literal salience in on-line processing of idiomatic expressions by second language learners. *Second Language Research*, *22*(2), 115–144. https://doi.org/10.1191/0267658306sr263oa

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Academic Press. https://doi.org/10.4324/9780203771587

Conklin, K. & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics*, *29*(1), 72–89. https://doi.org/10.1093/applin/amm022

Conklin, K. & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074

Conrad, S. & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, *20*(2005), 56–71. https://doi.org/10.1515/9783484604674.56

Corder, P. (1967). The significance of learners' errors. *International Review of Applied Linguistics in Language Teaching*, *5*(1/4), 161–170. https://doi.org/10.1515/iral.1967.5.1-4.161

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423. https://doi.org/10.1016/j.esp.2003.12.001

Cortes, V., Jones, J. & Stoller, F. (2002). *Lexical bundles in ESP reading and writing* [Paper presentation]. TESOL Conference. Salt Lake City, UT.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics*, *2*(3), 223–235. https://doi.org/10.1093/applin/II.3.223

Cowie, A. P. (1991). Multiword units in newspaper language in perspectives on the English lexicon. A tribute to Jacques Van Roey. *Cahiers de l'Institut de Linguistique de Louvain*, *17*(1/3), 101–116.

Cowie, A. P. (1992). Multiword lexical units and communicative language teaching. In P. Arnaud, & H. Bejoint (Eds.), *Vocabulary and applied linguistics*, (pp. 1–12). Macmillan. https://doi.org/10.1007/978-1-349-12396-4_1

Cowie, A. P. (1994). Phraseology. In R. Asher (Ed.), *Encyclopedia of language and linguistics*, (pp. 3168–3171). Pergamon.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford University Press.

Crossley, S. & Salsbury, T. L. (2011). The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics in Language Teaching*, *49*(1), 1–26. https://doi.org/10.1515/iral.2011.001

Crosthwaite, P. (2016). L2 English article use by L1 speakers of article-less languages. *International Journal of Learner Corpus Research*, *2*(1), 68–100. https://doi.org/10.1075/ijlcr.2.1.03cro

Dagneaux, E., Denness, S. & Granger, S. (1998). Computer-aided error analysis. *System*, *26*(2), 163–174. https://doi.org/10.1016/S0346-251X(98)00001-3

De Cock, S. (1998): A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics*, *3*(1), 59–80. https://doi.org/10.1075/ijcl.3.1.04dec

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair, & M. Hundt (Eds.), *Corpus linguistics and linguistic theory. Papers from the Twentieth International Conference on English language research on computerized corpora (ICAME 20)*, (pp. 51–68). Rodopi. http://hdl.handle.net/2078.1/75985

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, *2*(1), 225–246.

De Cock, S. (2005). Learners and phrasal verbs. In M. Rundell (Ed.), *Macmillan phrasal verbs plus*. Macmillan. http://hdl.handle.net/2078.1/75983

De Cock, S., Granger, S., Leech, G. & T. McEnery. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer*, (pp. 67–79). Addison Wesley Longman.

De Felice, R. 2008. *Automatic error detection in non-native English*. [Unpublished doctoral dissertation]. University of Oxford.

De Felice, R. & Pulman, S. (2009). Automatic detection of preposition errors in learner writing. *Calico Journal*, *26*(3), 512–528. https://doi.org/10.1558/cj.v26i3.512-528

De Haan, P. (1992). The optimum corpus sample size? In G. Leitner, (Ed.), *New directions in English language corpora*, (pp. 3–19). Mouton de Gruyter.

De Haan, P. (2000). Tagging non-native English with the TOSCA-ICLE tagger. In C. Mair, & M. Hundt (Eds.), *Corpus linguistics and linguistic theory*. *Papers from ICAME 20 1999*, (pp. 69–79). Amsterdam: Rodopi.

De Haan, P. & van der Haagen, M. (2013). Assessing the use of sophisticated EFL writing. In B. Weltens, J. Hornikx, W. Lowie, P. Poelmans, & R. L. Present-Thomas (Eds.), *Writing assessment in higher education*, (pp. 16–27). John Benjamins. https://doi.org/10.1075/dujal.2.1.04deh

de Saint Léger, D. & Storch, N. (2009). Learners' perceptions and attitudes: Implications for willingness to communicate in an L2 classroom. *System*, *37*(2), 269–285. https://doi.org/10.1016/j.system.2009.01.001

De Santis, C. (2014). "Cresci, cresci, cresci…". La reduplicazione espressiva come strumento di espressione di relazioni transfrastiche. In C. De Santis, A. Ferrari, G. Frenguelli, F. Gatta, L. Lala, M. Mazzoleni, & M. Prandi (Eds.), *Le relazioni logico-sintattiche. Teoria, sincronia, diacronia*, (pp. 185–211). Aracne Editrice. http://dx.doi.org/10.4399/97888548780519

Dechert, H. (1983). How a story is done in a second language. In C. Faerch, & G. Kasper (Eds.), *Strategies in interlanguage communication*, (pp. 175–195). Longman.

Díaz-Negrillo, A., Meurers, D., Valera, S. & Wunsch, H. (2010). Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, *36*(1/2), 139–154.

Dörnyei, Z. & Taguchi, T. (2009). *Questionnaires in second language research. Construction, administration, and processing*. Routledge.

Dressler, W. U. & Barbaresi, L. M. (1994). *Morphopragmatics: Diminutives and intensifiers in Italian, German, and other languages*. Mouton de Gruyter. https://doi.org/10.1515/9783110877052

Dulay, H. C. & Burt, M. K. (1974). Errors and strategies in child second language acquisition. *TESOL Quarterly*, *8*(2), 126–136. http://dx.doi.org/10.2307/3585536

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*(1), 61–74.

Durrant, P. L. (2008). *High frequency collocations and second language learning.* [Unpublished doctoral dissertation]. University of Nottingham.

Durrant, P. L. (2014). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics*, *19*(4), 443–477. https://doi.org/10.1075/ijcl.19.4.01dur

Durrant, P. L. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics*, *38*(2), 165–193. https://doi.org/10.1093/applin/amv011

Durrant, P. L. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177. https://doi.org/10.1515/iral.2009.007

Edmonds, A. (2014). Conventional expressions: Investigating pragmatics and processing. *Studies in Second Language Acquisition*, *36*(1), 69–99. *https://doi.org/10.1017/S0272263113000557*

Edwards, J. (1992). Transcription of discourse. In W. Bright (Ed.), *International encyclopedia of linguistics*, vol. 1, (pp. 367–371). Oxford University Press.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, *18*(1), 91–126. http://dx.doi.org/10.1017/S0272263100014698

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. J. Doughty, & M. H. Long

(Eds.), *The handbook of second language acquisition*, (pp. 63–103). Blackwell. https://doi.org/10.1002/9780470756492.ch4

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24. https://doi.org/10.1093/applin/ami038

Ellis, N. C. & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*(1), 188–221. https://doi.org/10.1075/arcl.7.08ell

Ellis, N. C., Simpson-Vlach, R., Römer, U., O'Donnell, M. & Wulff, S. (2015). Learner corpora and formulaic language in SLA. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*, (pp. 357–378). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.016

Ellis, N. C., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Ellis, R. (1994). *The study of second language acquisition*. Oxford University Press.

Ellis, R. (1997). *Second language acquisition*. Oxford University Press.

Ellis, R., & Barkhuizen, G. P. (2005). *Analysing learner language*. Oxford University Press.

Enstad, R. I. (2010). *The phraseological patterns of fun and funny: A corpus-based investigation* (Unpublished MA thesis). University of Oslo.

Evert, S. (2005). *The statistics of word co-occurrences: Word pairs and collocations.* [Doctoral dissertation, University of Stuttgart]. OPUS – Publication Server of the University of Stuttgart. http://dx.doi.org/10.18419/opus-2556

Evert, S. (2007). *Room for improvement? Upper limits on collocation extraction with statistical association measures* [Poster presentation]. Annual Meeting of the German Linguistics Association (DGfS 2007), Siegen, Germany. http://www.dgfs2007.uni-siegen.de

Evert, S. (2009). Corpora and collocations. In A. Lüdeling, & M. Kytö (Eds.), *Corpus linguistics: An international handbook*, (pp. 1212–1248). Walter de Gruyter.

Fallah, N. (2014). Willingness to communicate in English, communication self-confidence, motivation, shyness and teacher immediacy among Iranian English-major undergraduates: A structural equation modeling approach. *Learning and Individual Differences*, *30*, 140–147. https://doi.org/10.1016/j.lindif.2013.12.006

Fan, M. (2009). An exploratory study of collocational use by ESL students. A task based approach. *System*, *37*(1), 110–123. https://doi.org/10.1016/j.system.2008.06.004

Farghal, M. & Obiedat, H. (1995). Collocations: A neglected variable in EFL. *International Review of Applied Linguistics in Language Teaching*, *33*(4), 315–331. https://doi.org/10.1515/iral.1995.33.4.315

Ferraresi, A. (2019): Collocations in contact: Exploring constrained varieties of English through corpora. *Textus*, *32*(1), 203–222. http://dx.doi.org/10.7370/93190

Field, A., Miles, J. & Field, Z. (2012). *Discoversing statistics using R*. Sage.

Firth, J. R. (1951). Modes of meaning. In J. R. Firth (Ed.), *Essays and studies*, (pp. 118–149). Oxford University Press.

Firth, J. R. (1957). *Papers in linguistics 1934-1951*. Oxford University Press.

Firth, J. R. (1968). *Selected papers of J.R. Firth 1952-59*. Longman.

Forchini, P. (2012). Teaching spoken language through movie corpora., Nuova Secondaria, *6*, 76–81.

Forchini, P. & Murphy, A. C. (2010). 4-grams in comparable specialised corpora: Perspectives on phraseology, translation, and pedagogy. In U. Römer, & R. Schulze (Eds.), *Patterns, meaningful units and specialised discourse*, (pp. 87-103). John Benjamins. https://doi.org/10.1075/bct.22.05for

Forsberg, F. (2010). Using conventional sequences in L2 French. *International Review of Applied Linguistics in Language Teaching*, *48*(1), 25–51. https://doi.org/10.1515/iral.2010.002

Forti, L. (2019). *Developing phraseological competence in Italian L2: A study on the effects of data-driven learning*. [Unpublished doctoral dissertation]. Università per Stranieri di Perugia.

Foster, P. & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*(3), 299–323. http://dx.doi.org/10.1017/S0272263100015047

Fraser, B. (2010). Pragmatic competence: The case of hedging. In G. Kaltenbück, W. Mihatsch, & S. Schneider (Eds.), *New approaches to hedging*, (pp. 15–34). Brill. https://doi.org/10.1163/9789004253247_003

Gablasova, D., Brezina, V. & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(S1), 155–179. https://doi.org/10.1111/lang.12225

Gablasova, D., Brezina, V. & McEnery, T. (2019). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*, *5*(2), 126–158. https://doi.org/10.1075/ijlcr.19001.gab

Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. (2015). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*, *38*(5), 613–637. https://doi.org/10.1093/applin/amv055

Gaies, S. J. (1983a). The investigation of language classroom processes. *TESOL Quarterly*, *17*(2), 205–217. http://dx.doi.org/10.2307/3586650

Gaies, S. J. (1983b). Learner feedback: An exploratory study of its role in the second language classroom. In H. Seliger, & M. H. Long, (Eds.), *Classroom oriented research in second language acquisition*, (pp. 190–213). Newbury House.

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech, & G. Sampson (Eds.), *The computational analysis of English: A corpus-based approach*. Longman.

Garside, R. (1995). Using CLAWS to annotate the British National Corpus. URL: http://info.ox.ac.uk/bnc/garside_allc. html

Garside, R. & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech, & T. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora*, (pp. 102–121). Longman.

Gass S. M. & Selinker, L. (2008). *Second language acquisition: An introductory course*. Routledge.

Gass, S. M. & Mackey, A. (2007). Input, interactions, and output in second language acquisition. In B. Vanpatten, & J. Williams (Eds.), *Theories in second language acquisition*, (pp. 175-200). LEA.

Gass, S. M., & Selinker, L. (2001). *Second language acquisition: An introductory course*. Lawrence Erlbaum.

Gibbs, R. W. (1980). Spilling the beans on understanding and memory for idioms in conversation. *Memory & Cognition*, *8*, 149–156. https://doi.org/10.3758/BF03213418

Gilquin, G. (2007). To err is not all. What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik*, *55*(3), 273–91. https://doi.org/10.1515/zaa.2007.55.3.273

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*, (pp. 9–34). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.002

Gilquin, G. (2019). Light verb construction in spoken L2 English. *International Journal of Learner Corpus Research*, *5*(2), 181–206. https://doi.org/10.1075/ijlcr.18003.gil

Gilquin, G., De Cock, S. & Granger, S. (2010). *LINDSEI. Louvain International Database of Spoken English Interlanguage*. Presses universitaires de Louvain.

Gilquin, G. & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction*, *1*(1), 41–61. https://doi.org/10.1075/etc.1.1.05gil

Gitsaki, C. (1999). *Second language lexical acquisition: A study of the development of collocational knowledge*. International Scholars.

Götz, S. & Schilk, M. (2011). Formulaic sequences in spoken ENL, ESL and EFL. In J. Mukherjee, & M. Hundt (Eds.), *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*, (pp. 79–100). John Benjamins. https://doi.org/10.1075/scl.44.05sch

Gráf, T. (2017). The story of the Learner Corpus LINDSEI_CZ. *Studie Z Aplikované Lingvistiky*, *2*, 22–35.

Grandi, N. (2017). Intensification processes in Italian. In M. Napoli, & M. Raveto (Eds.), *Exploring intensification: synchronic, diachronic and cross-linguistic perspectives*, (pp. 55–77). John Benjamins. https://doi.org/10.1075/slcs.189.04gra

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast. Text-based cross-linguistic studies*, (pp. 37–51). Lund University Press.

Granger, S. (1998). *Learner English on computer*. Addison Wesley Longman.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer corpora, second language acquisition and foreign language teaching*, (pp. 3–33). John Benjamins. https://doi.org/10.1075/lllt.6.04gra

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, *37*(3), 538–546. http://dx.doi.org/10.2307/3588404

Granger, S. (2008). Learner corpora. In A. Lüdeling, & M. Kytö (Ed.), *Corpus linguistics. An international handbook*, (pp. 259–275). Walter de Gruyter.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. A critical evalution. In K. Aijmer (Ed.), *Corpora and language teaching*, (pp. 13–32). John Benjamins. https://doi.org/10.1075/scl.33.04gra

Granger, S. (2012a). How to use foreign and second language learner corpora. In A. Mackey, & S. G. Gass (Eds.), *A guide to research methods in second language acquisition*, (pp. 7–29). Basil Blackwell. https://doi.org/10.1002/9781444347340.ch2

Granger, S. (2012b). Learner corpora. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*, (pp. 3235–3242). Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0669

Granger, S. (2017). Learner corpora in foreign language education. In S. Thorne, & S. May (Eds.), *Language and technology. Encyclopedia of language and education* (3rd ed.), (pp. 1–14). Springer International. https://doi.org/10.1007/978-3-319-02237-6_33

Granger, S. & Bestgen, Y. (2014). The use of collocations by intermediate vs advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229–252. https://doi.org/10.1515/iral-2014-0011

Granger, S. & Meunier, F. (2008). *Phraseology: An interdisciplinary perspective*. John Benjamins. https://doi.org/10.1075/z.139

Granger, S. & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger, & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective*, (pp. 27–49). John Benjamins. https://doi.org/10.1075/z.139.07gra

Granger, S. & Rayson, P. (1998). Automatic profiling of learner texts. In S. Granger (Ed.), *Learner English on computer*, (pp. 119–131). Routledge.

Granger, S. & Thewissen, J. (2005, June). *Towards a reconciliation of a 'can do' and 'can't do' approach to language assessment* [Paper presentation]. 2nd Annual Conference of EALTA, Voss, Norway.

Granger, S., Gilquin, G. & Meunier, F. (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger, & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective*, (pp. 3–26). John Benjamins. https://doi.org/10.1075/z.139.06gri

Gries, S. T. (2010). Useful statistics for corpus linguistics. In A. Sánchez Pérez, & M. Almela Sánchez (Eds.), *A mosaic of corpus linguistics: Selected approaches*, (pp. 269–291). Peter Lang.

Gries, S. T. (2013a). 50-something years of work on collocations: What is or should be next... *International Journal of Corpus Linguistics*, *18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, S. T. (2013b). *Statistics for linguistics with R*. Walter de Gruyter.

Gries, S. T. (2015). Statistics for learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*, (pp. 160–181). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.008

Guo, X. (2006). *Verbs in the written English of Chinese learners: A corpus-based comparison between non-native speakers and native speakers.* [Unpublished doctoral dissertation]. University of Birmingham.

Gut, U. (2010). Cross-linguistic influence in L3 phonological acquisition. *International Journal of Multilingualism*, *7*(1), 19–38. https://doi.org/10.1080/14790710902972248

Gyllstad, H. (2005). Words that go together well: Developing test formats for measuring learner knowledge of English collocations. (The department of English in Lund: Working papers in linguistics, vol. 5). http://www.englund.lu.se/images/stories/pdf-files/workingspapers/vol05/gyllstad-wp-05.pdf

Gyllstad, H. & Wolter, B. (2016). Collocational processing in light of the phraseological continuum model: Does semantic transparency matter? *Language Learning*, *66*(2), 296–323. https://doi.org/10.1111/lang.12143

Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robins (Eds.), *In Memory of J. R. Firth,* (pp. 148–162). Longman.

Han, Z. (2004). Fossilization: Five central issues. *International Journal of Applied Linguistics*, *14*(2), 212–241. https://doi.org/10.1111/j.1473-4192.2004.00060.x

Handl, S. (2008). Essential collocations for learners of English: The role of collocational direction and weight. In F. Meunier, & S. Granger (Eds.), *Phraseology in foreign language learning and teaching*, (pp. 43–66). John Benjamins. https://doi.org/10.1075/z.138.06han

Hasko, V. (2020). Qualitative corpus analysis. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0974.pub2

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237–258. https://doi.org/10.1111/j.1473-4192.1994.tb00065.x

Hasselgren, A. (1997). The EVA corpus of Norwegian school English. *ICAME Journal*, *21*, 123–124. http://icame.uib.no/ij21/eva-corp.pdf

Hausmann, F. (1989). Le dictionnaire de collocations. In F. Hausmann, O. Reichmann, H. Wiegand, & L. Zgusta (Eds.), *Wörterbücher: ein internationales Handbuch zur Lexicographie. Dictionaries. Dictionnaires*, (pp. 1010–1019). Walter de Gruyter.

He, L. & Shi, L. (2012). Topic knowledge and ESL writing. *Language Testing*, *29*(3), 443–464. https://doi.org/10.1177/0265532212436659

Herbst, T. (1996). What are collocations: Sandy beaches or false teeth? *English Studies*, *77*(4), 379–393. https://doi.org/10.1080/00138389608599038

Hernández, M., Costa, A. & Arnon, I. (2016). More than words: Multiword frequency effects in non-native speakers. *Language, Cognition and Neuroscience*, *31*(6), 785–800. https://doi.org/10.1080/23273798.2016.1152389

Hilton, H. (2014). Oral fluency and spoken proficiency: Considerations for research and testing. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency. Perspectives from SLA*, (pp. 27–53). Multilingual Matters. https://doi.org/10.21832/9781783092291-005

Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics*, *41*(4), 667–683. https://doi.org/10.1016/j.pragma.2008.09.029

Hoey, M. (1991). *Patterns of lexis in text*. Oxford University Press. https://doi.org/10.1017/S0272263100013024

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. Routledge. https://doi.org/10.4324/9780203327630

Hoffman, S. & Lehmann, H. (2000). Collocational evidence from the British National Corpus. In J. Kirk (Ed.), *Corpora galore: Analyses and techniques in describing English*, (pp. 17–32). Rodopi.

Housen, A. (2002). A corpus-based study of the L2 acquisition of the English verb system. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer corpora, second language acquisition and foreign language teaching*, (pp. 77–116). John Benjamins. https://doi.org/10.1075/lllt.6.08hou

Howarth, P. (1996). *Phraseology in English academic writing. Some implications for language learning and dictionary making*. Max Niemeyer.

Howarth, P. (1998a). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications*, (pp. 161–186). Oxford University Press.

Hsu, Y. L. (2005). *Teacher's immediacy and students' willingness to communicate (WTC): Perceived teacher's verbal immediacy behaviors in English class*. [Unpublished MA thesis]. Ming Chuan University.

Huang, L. (2001). Knowledge of English collocations: An analysis of Taiwanese EFL learners. In C. Luke, & B. Rubrecht (Eds.), *Texas Papers in Foreign Language Education: Selected Proceedings from the Texas Foreign Language Education Conference, 2001*. Volume 6, n. 1, Fall 2001.

Hübler, Axel. (1983). *Understatements and hedges in English*. John Benjamins.

Hulstijn, J. (2010). Measuring second language proficiency. In E. Blom, & S. Unsworth (Eds.), *Experimental methods in language acquisition research*, (pp. 185–200). John Benjamins. https://doi.org/10.1075/lllt.27.11hul

Hulstijn, J. (2011). Language proficiency in native and non-native speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229–249. https://doi.org/10.1080/15434303.2011.565844

Hulstijn, J., Alderson, C. & Schroonen, R. (2010). Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them? In I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, (pp. 11-20). European Second Language Association. https://hdl.handle.net/11245/1.327493

Hundt, M., Nesselhauf, N. & Biewer, C. (2007). Corpus linguistics and the web. *Language and Computers*, *59*, 1–5. https://doi.org/10.1163/9789401203791

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Hunston, S. & Francis, G. (2002). *Pattern Grammar. A corpus-driven approach to the lexical grammar of English*. John Benjamins. https://doi.org/10.1075/scl.4

Hutchinson, J. (1996). *UCL Error Editor*. Centre for English Corpus Linguistics, Université Catholique de Louvain.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hymes, D. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.), *Sociolinguistics*, (pp. 269–293). Penguin Books.

Izumi, E., Saiga, T., Supnithi, T., Uchimoto, K. & Isahara, H. (2003). The development of the spoken corpus of Japanese learner English and the applications in collaboration with NLP techniques. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference*, (pp. 359–366). UCREL, Lancaster University. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.513.8250&rep=rep1&type=pdf

Izumi, E., Uchimoto, K. & Isahara, H. (2004). The NICT JLE corpus: Exploiting the language learners' speech database for research and education. *International Journal of the Computer, the Internet and Management*, *12*(2), 119–125. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.563.7156&rep=rep1&type=pdf

Jaén, M. M. (2009). A corpus-driven design of a test for assessing the ESL collocational competence of university students. *International Journal of English Studies*, *7*(2), 127–148.

Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon. *Language Learning*, *50*(2), 245–309. https://doi.org/10.1111/0023-8333.00118

Jarvis, S. & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. Routledge.

Jiang, N. A. & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal*, *91*(3), 433–445. https://doi.org/10.1111/j.1540-4781.2007.00589.x

Jones, M. & Sinclair, J. (1974). English lexical collocations: A study in computational linguistics. *Cahiers de Lexicologie*, *24*(1), 15–61.

Jones, N. (2011, April). *Linking classroom activities and the CEFR levels: How different kinds of assessment can support learning* [Paper presentation], CEF conference for teachers, Brussels, Belgium.

Juhasz, B. J. (2007). The influence of semantic transparency on eye movements during English compound word recognition. In R. P. G. Van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain*, (pp. 373–389). Elsevier Science. https://doi.org/10.1016/B978-008044980-7/50018-5

Kang, S. J. (2005). Dynamic emergence of situational willingness to communicate in a second language. *System*, *33*(2), 277–292. https://doi.org/10.1016/j.system.2004.10.004

Kang, S., Kim, Y. & Roemer, U. (2020, March). *Examining the longitudinal development of English verb-argument constructions: Comparing spoken and written learner production data*. In 2020 conference of the American Association for Applied Linguistics (AAAL). AAAL.

Kapatsinski, V. & Radicke, J. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. In R. L. Corrigan, E. A. Moravcsik, H. Ouali, & K. M. Wheatley (Eds.), *Formulaic language: Volume 2. Acquisition, loss, psychological reality, and functional explanations*, (pp. 499–522). John Benjamins. https://doi.org/10.1075/tsl.83.14kap

Kaszubski, P. (2000). *Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: A contrastive, corpus-based approach* [Unpublished doctoral dissertation]. Adam Mickiewicz University.

Kellerman, E. (1979). Transfer and non-transfer: where are we now? *Studies in Second Language Acquisition*, *2*, 37–57.

Kennedy, G. (1998). *An introduction to corpus linguistics*. Longman.

Kjellmer, G. (1987). Aspects of English collocations. In W. Meijs (Ed.), *Corpus linguistics and beyond: Proceedings of the Seventh International Conference on English language research on computerized corpora*, (pp. 133-140). Amsterdam: Rodopi.

Kjellmer, G. (1991). A mint of phrases. In K. Aijmer, & B. Altenberg (Eds.), *English corpus linguistics. Studies in honour of Jan Svartvik*, (pp. 111-127). Longman.

Kjellmer, G. (1994). *A dictionary of English collocations: Based on the Brown Corpus*. Oxford University Press.

Klyachko, E., Arkhangelskiy, T., Kisselev, O. & Rakhilina, E. (2013). Automatic error detection in Russian learner language. In *Proceedings of the First Workshop Corpus Analysis with Noise in the Signal (CANS 2013)*, UCREL, Lancaster University. http://ucrel.lancs.ac.uk/cans2013/

Koya, T. (2005). *The acquisition of basic collocations by Japanese learners of English.* [Unpublished doctoral dissertation]. Waseda University.

Krenn, B. (2000). *The usual suspects: Data-oriented models for identification and representation of lexical collocations*. DFKI & Universität des Saarlandes.

Krummes, C. & Ensslin, A. (2015). Formulaic language and collocations in German essays: From corpus-driven data to corpus-based materials. *The Language Learning Journal*, *43*(1), 110–127. https://doi.org/10.1080/09571736.2012.694900

Lado, R. (1961). *Language testing*. McGraw-Hill.

Lardiere, D. (2003). Revisiting the comparative fallacy: A reply to Lakshmanan and Selinker, 2001. *Second Language Research*, *19*(2), 129–143. https://doi.org/10.1191/0267658303sr216oa

Larsen-Freeman D. & Long, M. (1991). *An introduction to second language acquisition research*. Longman. https://doi.org/10.4324/9781315835891

Larsen-Freeman, D. (2014). Another step to be taken. Rethinking the end point of the interlanguage continuum. In Z. Han, & E. Tarone (Eds.), *Interlanguage. Forty years later*, (pp. 203–220). John Benjamins. https://doi.org/10.1075/lllt.39.11ch9

Laufer, B. & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Leclercq, P., Edmonds, A. & Hilton, H. (2014). *Measuring L2 proficiency. Perspectives from SLA*. Multilingual Matters. https://doi.org/10.21832/9781783092291

Lee, D. Y. W. & Chen, S. (2009). Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing*, *18*(4), 149–165. https://doi.org/10.1016/j.jslw.2009.05.004

Leech, G. (1993). Corpus annotation schemes. *Literary and Linguistic Computing*, *8*(4), 275– 281. https://doi.org/10.1093/llc/8.4.275

Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer*, (pp. xiv-xx). Longman.

Leech, G., Smith, N. & Rayson, P. (2012). English style on the move: Variation and change in stylistic norms in the twentieth century. In M. Kytö (Ed.), *English corpus linguistics: Crossing paths,* (pp. 69-98). Brill Rodopi.

Lessard, G. (1999). [Review of the book Learner English on computer, by S. Granger (Ed.)]. *Computational Linguistics*, *25*(2), 302–303.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins. https://doi.org/10.1075/z.195

Lewis, M. (1997). *Implementing the lexical approach. Putting theory into practice*. Language Teaching Publishings.

Lewis, M. (2000). *Teaching collocations: Further developments in the lexical approach*. Language Teaching Publishings.

Li, J. & Schmitt, N. (2009). The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication*, (pp. 23–46). Continuum.

Libben, G. (1998). Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. *Brain and Language*, *61*(1), 30–44. https://doi.org/10.1006/brln.1997.1876

Lindqvist, C. (2009). The use of the L1 and the L2 in French L3: Examining cross-linguistic lexemes in multilingual learners' oral production. *International Journal of Multilingualism*, *6*(3), 281–297. https://doi.org/10.1080/14790710902812022

Lindqvist, C. (2010). Inter-and intralingual lexical influences in advanced learners' French L3 oral production. *International Review of Applied Linguistics in Language Teaching*, *48*(2/3), 131–157. https://doi.org/10.1515/iral.2010.007

Lorenz, G. (1999). *Adjective intensification - Learners versus native speakers: A corpus study of argumentative writing*. Rodopi.

Lozano, C. & Mendikoetxea, A. (2013). Corpus and experimental data: Subjects in second language research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty*

years of learner corpus research: Looking back, moving ahead, (pp. 313–323). Presses Universitaires de Louvain.

Lumley, T., Diehr, P., Emerson, S. & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual review of Public Health*, *23*(1), 151–169. http://dx.doi.org/10.1146/annurev.publhealth.23.100901.140546

Malloggi, P. (2017). Intensifying structures of adjectives across German and Italian. In M. Napoli, & M. Ravetto (Eds.), *Exploring Intensification: Synchronic, diachronic and cross-linguistic perspectives*, (pp. 251–264). https://doi.org/10.1075/slcs.189.13ma

Martelli, A. (2006). A corpus based description of English lexical collocations used by Italian advanced learners. In E. Corino, C. Marello, & C. Onesti (Eds.), *Proceedings of the 12th EURALEX International Congress*, (pp. 1005–1011). Edizioni dell'Orso.

Martins, C. & Pinharanda, N.M. (2013). L2 to L3 transfer: Learner corpora analyses. *Learner Corpus Studies in Asia and the World*, *1*, 271–281.

McCroskey, J. C. & Baer, J. E. (1985). Willingness to communicate: The construct and its measurement. *Communication Reports*, *2*(2), 96–104.

McEnery, T. (2006). *Swearing in English: Bad language, purity and power from 1586 to the present*. Routledge.

McEnery, T. & Xiao, R. (2007). Parallel and comparable corpora. The state of play. In Y. Kawaguchi, T. Takagaki, N. Tomimori, & Y. Tsuruga (Eds.), *Corpus-based perspectives in linguistics*, (pp. 131–145). John Benjamins. https://doi.org/10.1075/ubli.6.11mce

McEnery, T. & Wilson, A. (1996). *Corpus linguistics*. Edinburgh University Press.

McEnery, T., Brezina, V., Gablasova, D. & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyze language use. *Annual Review of Applied Linguistics*, *39*, 74–92. https://doi.org/10.1017/S0267190519000096

Mel'čuk, I. (1988). Semantic description of lexical units in an explanatory combinatory dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, *1*(3), 165–188. https://doi.org/10.1093/ijl/1.3.165

Mel'čuk, I. (1998). Collocations and lexical functions. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*, (pp. 23–53). Oxford University Press.

Mendikotxea, A. & Lozano, C. (2018). From corpora to experiments: Methodological triangulation in the study of word order at the interfaces in adult late bilinguals (L2 learners). J*ournal of Psycholinguistics Research*, *47*, 871–898. https://doi.org/10.1007/s10936-018-9560-0

Meunier, F. & Granger, S. (2008). *Phraseology in foreign language learning and teaching*. John Benjamins. https://doi.org/10.1075/z.138

Meunier, F. & Littré, D. (2013). Tracking learners' progress: Adopting a dual "corpus cum experimental data" approach. *The Modern Language Journal*, *97*(S1), 61–76. https://doi.org/10.1111/j.1540-4781.2012.01424.x

Michel, M., Murakami, A., Alexopoulou, T. & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency. *Instructed Second Language Acquisition*, *3*(2), 124-152. https://doi.org/10.1558/isla.38248

Moe, E. (2007). SLA and LT research – friends or foes? In C. Carlsen, & E. Moe (Eds.), *A human touch to language testing. A collection of essays in jonour of Reidun Oanæs Andersen on the occasion of her retirement*, (pp. 172-183). Novus Press.

Moon, R. (1998a). *Fixed expressions and idioms in English*. Clarendon Press.

Moon, R. (1998b). Frequencies and forms of phrasal lexemes in English. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications*, (pp. 79–100). Oxford University Press.

Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In U. Connor, & T.A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective*, (pp. 239–250). Brill Rodopi. https://doi.org/10.1163/9789004333772_014

Mukherjee, J. (2009). The grammar of conversation in advanced spoken learner English: Learner corpus data and language-pedagogical implications. In K. Aijmer (Ed.), *Corpora and language teaching*, (pp. 203–230). John Benjamins. https://doi.org/10.1075/scl.33.17muk

Murakami, A. & Alexopoulou, T. (2016). L1 influence on the acquisition order of English grammatical morphemes: A learner corpus study. *Studies in Second Language Acquisition*. *38*(3), 365–401. https://doi.org/10.1017/S0272263115000352

Murphy, A. (2003). 'Naturalness is to text what grammatical correctness is to sentences': A corpus-driven perspective, Rassegna italiana di linguistica applicata, *1*(2), 207–223.

Murphy, A. & Poli, F. (2018). Reflections on a small word: The position of "also" in learner language. *Nuova Secondaria*, *10*, 80–84. http://hdl.handle.net/10807/122372

Mutta, M. (2014). Cross-linguistic influence in an oral translation task by L3 French learners. *Language, Interaction and Acquisition*, *5*(2), 279–313. https://doi.org/10.1075/lia.5.2.05mut

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research*, (pp. 309–332). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.014

Nagata, R., Mizumoto, T., Kikuchi, Y., Kawasaki, Y. & Funakoshi, K. (2018). A POS tagging model adapted to learner English. In W. Xu, A. Ritter, T. Baldwin, & A. Rahimi (Eds.), *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop in Noisy User-Generated Text*, (pp. 39–48). Brussels: Association for Computational Linguistics.

Nattinger, J. R. & De Carrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press.

Nesi, H. & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. In J. Flowerdew, & M. Mahlberg (Eds.), Lexical cohesion and corpus linguistics, (pp. 23–43). John Benjamins. https://doi.org/10.1075/bct.17.03nes

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24*(2), 223–242. https://doi.org/10.1093/applin/24.2.223

Nesselhauf, N. (2004). What are collocations? In D. Allerton, N. Nesselhauf, & P. Skandera (Eds.), *Phraseological units: Basic concepts and their application*, (pp. 1–21). Schwabe.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins. https://doi.org/10.1075/scl.14

Nicholls, D. (2003). The Cambridge learner corpus. Error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference. Technical Papers, Vol. 16,* (pp. 572–581). Lancaster: UCREL, Lancaster University.

Norris, J. M. & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition*, (pp. 717–761). Blackwell. https://doi.org/10.1002/9780470756492.ch21

Öksüz, D., Brezina, V. & Rebuschat, P. (2020). Collocational Processing in L1 and L2: The Effects of Word Frequency, Collocational Frequency, and Association. *Language Learning*. https://doi.org/10.1111/lang.12427

Orol González, A. & Alonso Ramos, M. (2013). A comparative study of collocations in a native corpus and a learner corpus of Spanish. *Procedia - Social and Behavioral Sciences*, *95*, 563–570. https://doi.org/10.1016/j.sbspro.2013.10.683

Ortega, L. & Byrnes, H. (2008). Theorizing advancedness, setting up the longitudinal research agenda. In L. Ortega, & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities*, (pp. 281–300). Routledge.

Osborne, J. (2008). Phraseology effects as a trigger for errors in L2 English: The case of more advanced learners. In F. Meunier, & S. Granger (Eds.), *Phraseology in foreign language learning and teaching*, (pp. 67–84). John Benjamins. https://doi.org/10.1075/z.138.07osb

Öz, H., Demirezen, M. & Pourfeiz, J. (2015). Willingness to communicate of EFL learners in Turkish context. *Learning and Individual Differences*, *37*, 269–275. https://doi.org/10.1016/j.lindif.2014.12.009

Palmer, H. E. (1933). *Second interim report on English collocations: Submitted to the Tenth Annual Conference of English teachers*. Institute for Research in English Teaching.

Paquot, M. (2013). Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics*, *18*(3), 391–417. https://doi.org/10.1075/ijcl.18.3.06paq

Paquot, M. (2017). L1 frequency in foreign language acquisition: Recurrent word combinations in French and Spanish EFL learner writing. *Second Language Research*, *33*(1), 13–32. https://doi.org/10.1177/0267658315620265

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, *15*(1), 29–43. https://doi.org/10.1080/15434303.2017.1405421

Paquot, M. & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. https://doi.org/10.1017/S0267190512000098

Paradis, C. (2008). Configurations, construals and change: Expressions of degree. *Journal of English Language and Linguistics*, *12*(2), 317–343. https://doi.org/10.1017/S1360674308002645

Paradis, C. (2012). *Constructing credibility in discourse Modes of knowing, temporality and epistemic control*. Abstract from Modality, corpus, discourse: Usage-based methods, epistemic stance, subjectivity, evidentiality, perspectivisation, Sweden.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. John Benjamins. https://doi.org/10.1075/scl.2

Pawlak, M. & Mystkowska-Wiertelak, A. (2015). Investigating the dynamic nature of L2 willingness to communicate. *System*, *50*, 1-9. https://doi.org/10.1016/j.system.2015.02.001

Pawley, A. (1986). Lexicalization. In D. Tannen, & J. E. Alatis (Eds.), *Language and linguistics: The interdependence of theory, data and application. Georgetown University round table on languages and linguistics*, (pp. 98-120). Georgetown University Press.

Pawley, A. (1991). How to talk cricket: On linguistic competence in a subject matter. In R. Blust (Ed.), *Pacific linguistics: Papers on Austronesian languages and ethnolinguistics in honour of George Grace*, (pp. 339–368). Pacific Linguistics.

Pawley, A. (2001). Phraseology, linguistics and the dictionary. *International Journal of Lexicography*, *14*(2), 122–134. https://doi.org/10.1093/ijl/14.2.122

Pawley, A. (2007). Developments in the study of formulaic language since 1970. In P. Skandera (Ed.), *Phraseology and culture in English*, (pp. 3–45). Mouton de Gruyter. https://doi.org/10.1515/9783110197860.3

Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards, & R. W. Schmidt (Eds.), *Language and communication*, (pp. 191–226). Longman.

Pecina, P. (2009). Lexical association measures and collocation extraction. *Language Resources and Evaluation*, *44*, 137–158. https://doi.org/10.1007/s10579-009-9101-4

Peng, J. E. & Woodrow, L. (2010). Willingness to communicate in English: A model in the Chinese EFL classroom context. *Language Learning*, *60*(4), 834–876. https://doi.org/10.1111/j.1467-9922.2010.00576.x

Peng, X. (2016). *Use of verb-noun collocations by advanced learners of Chinese*. [Unpublished doctoral dissertation]. University of Pennsylvania.

Pérez-Paredes, P. & Díez-Bedmar, M.B. (2019). Certainty adverbs in spoken learner language. *International Journal of Learner Corpus Research*, *5*(2), 253–279. https://doi.org/10.1075/ijlcr.17019.per

Pérez-Vidal, C. (2014). *Language acquisition in study abroad and formal instruction context*. John Benjamins. https://doi.org/10.1075/aals.13

Peters, A. M. (1983). *Units of language acquisition*. Cambridge University Press.

Plonsky, L. & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, *36*, 73– 97. https://doi.org/10.1017/S0267190516000015

Poli, F. (2014). *Fossilization in EFL: An investigation through learner corpora.* [Unpublished MA thesis]. Università Cattolica del Sacro Cuore.

Prince, E., Frader, J. & Bosk, C. (1982). On hedging in physician-physician discourse. In J. R. Di Pietro (Ed.), *Linguistics and the professions: Proceedings of the second annual Delaware Symposium on language studies*, (pp. 1-29). Norwood-New Jersey: Ablex.

Putri, C., Salikin, H. & Khazanah, D. (2020). She's Really Kind and Hella Weird! − The Use of Intensifiers among Teens: A Sociolinguistic Analysis. *k@ta*, *22*(1), 36–45.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of contemporary English*. Longman.

Rayson, P. & Baron, A. (2011). Automatic error tagging of spelling mistakes in learner corpora. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds.), *A taste for corpora. In honour of Sylviane Granger*, (pp. 109–126). John Benjamins.

Richards, J. (1971). Error Analysis and Second Language Strategies. http://eric.ed.gov/?id=ED048579 [last accessed 16th March 2020]

Römer, U., Skalicky, S. C., & Ellis, N. C. (2020). Verb-argument constructions in advanced L2 English learner production: Insights from corpora and verbal fluency tasks. *Corpus Linguistics and Linguistic Theory*, *16*(2), 303-331. https://doi.org/10.1515/cllt-2016-0055

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, *31*, 205–225. https://doi.org/10.1017/S0267190511000055

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Sage. http://nrs.harvard.edu/urn-3:HUL.InstRepos:38784661

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing*, (pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45715-1_1

Sánchez, L. (2015). L2 activation and blending in third language acquisition: Evidence of crosslinguistic influence from the L2 in a longitudinal study on the acquisition of L3 English. *Bilingualism: Language and Cognition*, *18*(2), 252–269. https://doi.org/10.1017/S1366728914000091

Schachter J. & Celce-Murcia, M. (1977). Some reservations concerning error analysis. *TESOL Quarterly*, *11*(4), 441–451. http://dx.doi.org/10.2307/3585740

Schachter, J. (1974). An error in error analysis. *Language Learning*, *24*(2), 205–214. https://doi.org/10.1111/j.1467-1770.1974.tb00502.x

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing and use*. John Benjamins. https://doi.org/10.1075/lllt.9

Schmitt, N. & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use*, (pp. 1–22). John Benjamins. https://doi.org/10.1075/lllt.9.02sch

Schmitt, N., Dörnyei, Z., Adolphs, S. & Durow, V. (2004): Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use*, (pp. 55–86). John Benjamins. https://doi.org/10.1075/lllt.9.05sch

Schneider, G. & Gilquin, G. (2016). Detecting innovations in a parsed corpus of learner English. In S. C. Deshors, S. Götz, & S. Laporte (Eds.), *Linguistic innovations: Rethinking linguistic creativity in non-native Englishes*, (pp. 177–204). John Benjamins. https://doi.org/10.1075/ijlcr.2.2.03sch

Schone, P. & Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In L. Lee, & D. Harman (Eds.) *Proceedings of the 2001 conference on empirical methods in natural language processing*, (pp. 100–108). https://www.cs.cornell.edu/home/llee/emnlp/papers/schone.pdf

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, *24*(1), 97–123.

Scott M. (2004). *WordSmith Tools* (Version 4.0). Oxford University Press.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(3), p. 209–232. https://doi.org/10.1515/iral.1972.10.1-4.209

Selinker, L. (2014). Interlanguage 40 years on. Three themes from here. In Z. Han, & E. Tarone (Eds.), *Interlanguage. Forty years later*, (pp. 221–246). John Benjamins. https://doi.org/10.1075/lllt.39.12ch1

Shapiro, S. S. & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591–611. http://dx.doi.org/10.2307/2333709

Shatz, I. (2016). Native language influence during second language acquisition: A large-scale learner corpus analysis. In M. Hirakawa, J. Matthews, K. Otaki, N. Snape, & M. Umeda (Eds.), *Proceedings of the Pacific Second Language Research Forum (PacSLRF 2016)*, (pp. 175–180). Hiroshima, Japan: Japan Second Language Association.

Šimčikaitė, A. (2012). Spoken discourse markers in learner academic writing. *Kalbų studijos*, *20*, 27–33.

Sinclair, J. (Ed.) 1987. *Looking up: An account of the COBUILD Project in Lexical Computing*. Collins.

Sinclair J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. (1995). Corpus typology. A framework for classification. In G. Melchers, & B. Warren (Eds.), *Studies in Anglistics*, (pp. 17–33). Almqvist and Wiksell International.

Sinclair, J. (1996). *EAGLES. Preliminary recommendations on corpus typology* (date of access 13/03/2020) http://www.ilc.cnr.it/EAGLES96/corpustyp/node12.html#SECTION00452000000000000000

Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive lexical semantics*, (pp. 1–24). John Benjamins. https://doi.org/10.1075/cilt.171.02sin

Sinclair, J. (2004). *Trust the text*. Routledge.

Sinclair, J., Jones, S. & Daley, R. (2004). *English collocation studies: The OSTI report*. Bloomsbury.

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, *11*(2), 285–301. https://doi.org/10.1515/cllt-2014-0016

Siyanova-Chanturia, A. & Pellicer-Sánchez, A. (2018). *Understanding formulaic language: A second language acquisition perspective*. Routledge. https://doi.org/10.4324/9781315206615

Siyanova-Chanturia, A. & Van Lancker-Sidtis, D. (2018). What online processing tells us about formulaic language. In A. Siyanova-Chanturia, & A. Pellicer-Sanchéz

(Eds.), *Understanding formulaic language. A second language acquisition perspective*, (pp. 38–61). Routledge. https://dx.doi.org/10.4324/9781315206615-3

Siyanova-Chanturia, A., Conklin, K. & Van Heuven, W. J. (2011a). Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences. *J Exp Psychol Learn Mem Cogn*, *37*(3), 776–784. https://dx.doi.org/10.1037/a0022531

Siyanova-Chanturia, A., Conklin, K. & Schmitt, N. (2011b). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*(2), 1–22. https://doi.org/10.1177/0267658310382068

Siyanova, A. & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, *64*(3), 429–458. http://dx.doi.org/10.3138/cmlr.64.3.429

Sosa, A. & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word 'of'. *Brain and Language*, *83*(2), 227–236. https://doi.org/10.1016/S0093-934X(02)00032-9

Spina, S. (2001). *Fare i conti con le parole. Introduzione alla linguistica dei corpora*. Guerra Edizioni.

Spina, S. & Siyanova-Chanturia, A. (2018). *The Longitudinal Corpus of Chinese Learners of Italian (LOCCLI)* [Poster presentation], 13th Teaching and Language Corpora Conference, Cambridge, UK.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, *2*(1), 23–55. https://doi.org/10.1075/fol.2.1.03stu

Stubbs, M. (1996). *Text and corpus analysis: Computer-assisted studies of language and culture*. Blackwell.

Stubbs, M. (2001). *Words and phrases. Corpus studies of lexical semantics*. Blackwell.

Stubbs, M. (2007). An example of frequent English phraseology: Distributions, structures and functions. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on*, (pp. 87–105). Brill; Rodopi. https://doi.org/10.1163/9789401204347_007

Sung Park, E. (2004). The comparative fallacy in UG studies. Working Papers in TESOL and Applied Linguistics 4(1). http://www.tc.columbia.edu/academic/tesol/Webjournal/forum2004.htm

Swinney, D. A. & Cutler, A. (1979). The access and processing of idiomatic expression. *Journal of Verbal Learning and Verbal Behavior*, *18*(5), 523–534. https://doi.org/10.1016/S0022-5371(79)90284-6

Tagliamonte, S. A. (2008). So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language and Linguistics*, *12*(2), 361–394. https://doi.org/10.1017/S1360674308002669

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, *9*, 123–143. https://doi.org/10.1016/0889-4906(90)90003-U

Thewissen, J. (2015). *Accuracy across proficiency levels*. Presses Universitaires de Louvain.

Thomas, M. (1994). [Review of the article Assessment of L2 proficiency in second language acquisition research]. *Language Learning*, *44*(2), 307–36.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. Norris, & L. Ortega (Eds.), *Synthesizing research on language learning and teaching*. John Benjamins. https://doi.org/10.1075/lllt.13.13tho

Timmis, I. (2015). *Corpus linguistics for ELT: Research and practice*. Routledge. https://doi.org/10.4324/9781315715537

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tono, Y. (2002). *The role of learner corpora in SLA research and foreign language teaching: The multiple comparison approach.* [Unpublished doctoral dissertation). Lancaster University.

Tono, Y. (2003). Learner corpora: Design, development and applications. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference*, (pp. 341).

UCREL Technical Paper Number 13. Lancaster: Lancaster University. http://ucrel.lancs.ac.uk/publications/cl2003/papers/tono.pdf

Tono, Y. (2016). What is missing in learner corpus design? In M. Alonso-Ramos (Ed.), *Spanish learner corpus research. Current trends and future perspectives*, (pp. 115–202). John Benjamins. https://doi.org/10.1075/scl.78.02ton

Tracy-Ventura, N., McManus, K., Norris, J. & Ortega, L. (2014). 'Repeat as much as you can': Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency. Perspectives from SLA*, (pp. 143–166). Multilingual Matters. https://doi.org/10.21832/9781783092291-011

Tracy-Ventura, N., Mitchell, R., & McManus, K. (2016). The LANGSNAP longitudinal learner corpus. Design and use. In M. Alonso-Ramos, (Ed.), *Spanish learner corpus research. Current trends and future perspectives*, (pp. 117–142). John Benjamins. https://doi.org/10.1075/scl.78.05tra

Tracy-Ventura, N., & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, *1*(1), 58– 95. https://doi.org/10.1075/ijlcr.1.1.03tra

Tracy, S. J. (2010). Qualitative quality: Eight "big-tent" criteria for excellent qualitative research. *Qualitative Inquiry*, *16*(10), 837–851. https://doi.org/10.1177/1077800410383121

Tremblay, A., Derwing, B., Libben, G. & Westbury, C. (2011). Processing advantages of lexical bundles: Evidence from self-paced reading and sentence recall tasks. *Language Learning*, *61*(2), 569–613. https://doi.org/10.1111/j.1467-9922.2010.00622.x

Tsui, A. B. (2004). What teachers have always wanted to know – and how corpora can help. In J. Sinclair (Ed.), *How to use corpora in language teaching*, pp. 39-61. John Benjamins.

Underwood, G., Schmitt, N. & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use*, (pp. 155–172). John Benjamins. https://doi.org/10.1075/lllt.9.09und

Van Lancker-Sidtis, D. & Postman, W. A. (2006). Formulaic expressions in spontaneous speech of left-and right-hemisphere-damaged subjects. *Aphasiology*, *20*(5), 411–426. https://doi.org/10.1080/02687030500538148

Van Lancker, D., Canter, G. J. & Terbeek, D. (1981). Disambiguation of ditropic sentences: Acoustic and phonetic cues. *Journal of Speech, Language, and Hearing Research*, *24*(3), 330–335. http://dx.doi.org/10.1044/jshr.2403.330

Van Lancker, D. & Kempler, D. (1987). Comprehension of familiar phrases by left- but not by right-hemisphere damaged patients. *Brain and Language*, *32*(2), 265–277. http://dx.doi.org/10.1016/0093-934x(87)90128-3

Van Roey, J. (1990). *French-English contrastive lexicology: An introduction*. Peeters.

Van Rooy, B. & Schäfer, L. (2002). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies*, *20*(4), 325–335. https://doi.org/10.2989/16073610209486319

Van Rooy, B. & Schäfer, L. (2003). Automatic POS tagging of a learner corpus: the influence of learner error on tagger accuracy. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference*, (pp. 835-844). UCREL, Lancaster University.

Verspoor, M. & Lowie, W. (1993). Making sense of polysemous words. *Language Learning*, *53*, 547-586.

Vincze, O., García-Salido, M., Orol, A. & Alonso-Ramos, M. (2016). A corpus study of Spanish as a foreign language learners' collocation production. In M. Alonso-Ramos (Ed.), *Spanish learner corpus research: Current trends and future perspectives*, (pp. 299–331). John Benjamins. https://doi.org/10.1075/scl.78.11vin

Vinogradov, V. V. (1947). Ob osnovnuikh tipakh frazeologicheskikh edinits v russom yazike [About the basic types of phraseological units in English]. In A. A. Shakhmatov (Ed.), *Sbornik statei i materialov* [The collection of articles and materials], (pp. 339–364). Nauka.

Vinogradova, O., Spina, S., Forti, L., Torubarov, I. & Login, N. (2018). Error annotation in learner corpora: Tools and applications in English and Italian. Pre-conference workshop 13th Teaching and Language Corpora Conference, University of Cambridge.

Wang, Y. (2016). *The idiom principle and L1 influence: A contrastive learner-corpus study of delexical verb+noun collocations*. John Benjamins. https://doi.org/10.1075/scl.77

Wang, Y. (2017). Lexical bundles in spoken academic ELF: Genre and disciplinary variation. *International Journal of Corpus Linguistics*, *22*(2), 187–211. https://doi.org/10.1075/ijcl.22.2.02wan

Wanner, L., Alonso Ramos, M., Vincze, O., Nazar, R., Ferraro, G., Mosqueira Suárez, E. & González, S. P. (2013). Annotation of collocations in a learner corpus for building a learning environment. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Twenty years of learner corpus research. Looking back, moving ahead*, (pp. 493-503). Presses Universitaires de Louvain.

West, R. (2015). Keeping it in the family. *English Teaching Professional*, 97.

White, L. (2003). On the nature of interlanguage representation: Universal Grammar in the second language. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition*, (pp. 19–42). Blackwell. https://doi.org/10.1002/9780470756492.ch2

Wolter, B. & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics*, *32*(4), 430–449. https://doi.org/10.1093/applin/amr011

Wolter, B. & Gyllstad, H. (2013). Frequency of input and L2 collocational processing: A comparison of congruent and incongruent collocations. *Studies in Second Language Acquisition*, *35*(3), 451–482. https://doi.org/10.1017/S0272263113000107

Wolter, B. & Yamashita, J. (2015). Processing collocations in a second language: A case of first language activation? *Applied Psycholinguistics*, *36*(5), 1193–1221. https://doi.org/10.1017/S0142716414000113

Wolter, B., & Yamashita, J. (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing: What accounts for L2 performance?. *Studies in Second Language Acquisition*, *40*(2), 395–416. https://doi.org/10.1017/S0272263117000237

Wray, A. (2002). Formulaic language in computer-supported communication: Theory meets reality. *Language Awareness*, *11*(2), 114–131. https://doi.org/10.1080/09658410208667050

Wulff, S. & Gries, S. (2011). Corpus-driven methods for assessing accuracy in learner production. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance*, (pp. 61–87). John Benjamins. https://doi.org/10.1075/tblt.2.07ch3

Xu, J. (2018). Measuring "spoken collocational competence" in communicative speaking assessment. *Language Assessment Quarterly*, *15*(3), 255–272. https://doi.org/10.1080/15434303.2018.1482900

Yamashita J. & Jiang, N. (2010). L1 Influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*(4), 647–668. https://doi.org/10.5054/tq.2010.235998

Yang, W. (2014). *Mapping the relationships among the cognitive complexity of independent writing tasks, L2 writing quality, and complexity, accuracy and fluency of L2 writing.* [Unpublished doctoral dissertation]. Georgia State University.

Yang, W., Lu, X. & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgements of writing quality. *Journal of Second Language Writing*, *28*, 53–67. https://doi.org/10.1016/j.jslw.2015.02.002

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *The Modern Language Journal*, *86*(1), 54–66. https://doi.org/10.1111/1540-4781.00136

Yashima, T., MacIntyre, P. D., & Ikeda, M. (2018). Situated willingness to communicate in an L2: Interplay of individual characteristics and context. *Language Teaching Research*, *22*(1), 115–137. https://doi.org/10.1177/1362168816657851

Yashima, T., Zenuk-Nishide, L. & Shimizu, K. (2004). The influence of attitudes and affect on willingness to communicate and second language communication. *Language Learning*, *54*(1), 119–152. https://doi.org/10.1111/j.1467-9922.2004.00250.x

Yi, W. (2018). Statistical sensitivity, cognitive aptitudes, and processing of collocations. *Studies in Second Language Acquisition*, *40*(4), 831–856. https://doi.org/10.1017/S0272263118000141

Yoon, H. J. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System*, *66*, 130–141. https://doi.org/10.1016/j.system.2017.03.007

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. https://doi.org/10.1016/j.jslw.2016.11.001

Yoon, H. J. & Polio, C. (2016). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, *51*(2), 275–301. https://doi.org/10.1002/tesq.296

Yorio, C. A. (1989). Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam, & L. K. Obler (Eds.), *Bilingualism across the lifespan*, (pp. 55–72). Cambridge University Press. https://doi.org/10.1017/CBO9780511611780.005

Zarrinabadi, N. (2014). Communicating in a second language: Investigating the effect of teacher on learners' willingness to communicate. *System*, *42*, 288–295. https://doi.org/10.1016/j.system.2013.12.014

Zhang, W. Z. & Chen, S. C. (2006). EFL learners' acquisition of English adjective-noun collocations - A quantitative study. *Foreign Language Teaching and Research*, *38*(4), 251–258.

Zhang, X. (1993). *English collocations and their effect on the writing of native and non-native college fresh-men.* [Unpublished doctoral dissertation]. Indiana University of Pennsylvania.

Zolkovskij A. & Mel'čuk, I. (1965). O vozmoˇznom metode i instrumen- tax semantiˇceskogo sinteza [On a possible method and instruments for semantic synthesis (of texts)]. *Nauˇcno-texniˇceskaja informacija* [Scientific and Technological Information], *6*, 23–28.

Appendix

LINDSEI corpus design criteria

Since LINDSEI was an international project involving several teams, for the researchers it was important to obtain a high level of comparability as regards data collection and transcription. This led them to the choice of text type: the informal interview. The informal interview was chosen because it does not impose any constraints or limits to the learners and, as such, it is the best choice for obtaining spontaneous speech.

The procedure consisted in three stages: set topic, free discussion, picture description. The first task, which is to all effects a monologue, consists in the learner choosing a topic from a selection of three set ones, and talking about it for approximately five minutes. The set topics are the following:

1. An experience you have had which has taught you an important lesson. You should describe the experience and say what you have learned from it.

2. A country you have visited which has impressed you. Describe your visit and say why you found the country particularly impressive.

3. A film/play you have seen which you thought was particularly good/bad. Describe the film/play and say why you thought it was good/bad.

After this initial activity, the interviewer moved on to the second task, the free discussion. Initially this consists in asking follow-up questions on the chosen topic, and then discussing other subjects, such as university life, hobbies, travelling. This task was conceived to resemble a dialogue. Lastly, the description of a picture was designed so that the learner was asked to look at four pictures making up a story and reconstruct the story according to their own interpretation.

Other important design criteria for LINDSEI: the EFL learning context, the advanced proficiency level, and the attendance of an undergraduate course. As regards, proficiency, the level was initially set at advanced, but since the researchers relied on arbitrary criteria (i.e., course year), there are many differences between texts and sub-corpora. It is best to define the proficiency of LINDSEI as ranging from higher intermediate to advanced. EFL context, advanced proficiency, and undergraduate course were set as the requirements for the compilation of the corpus, but other variables concerning the interview, interviewer, interviewee were also recorded. The additional variables were thought to be useful for future research and can be used as a search criteria in the digital interface of LINDSEI. Figure 8.0.1 shows the other recorded variables.

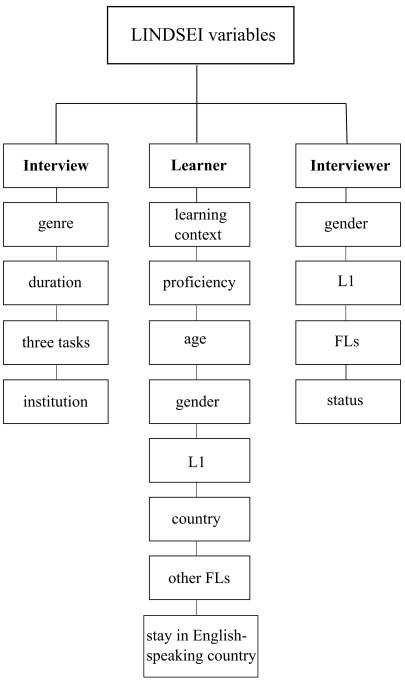

Figure 8.0.1: LINDSEI variables (Gilquin et al., 2010, p. 7).

The duration of the interview was set at about 15 minutes (thus resulting in approximately 2,000 words per interview) and the majority of sub-corpora data were collected in a single institution, except for the Italian, Japanese, and Spanish sub-corpora, whose data collection took place in two different universities.

The LINDSEI learners come from 11 L1s: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish, and Swedish.

Description of LINDSEI

The average age of LINDSEI learners is 22.38 and the number is slightly higher in the Italian sub-corpus, 23.54. Most interviewees are females (79% of the whole corpus), and this is no exception for the Italian sub-corpus with 80% female learners. Other interesting variables regarding the Italian sub-corpus are the average number of years studying English at school (7.08), the average number of years studying English at university (3.46) and lastly, the average number of months spent in an English-speaking country (2.13). These numbers are

within the standard averages of LINDSEI, as the overall average number of years studying English at school is 7.33, the overall average number of years studying English at university is 2.99, and the overall average number of months spent in an English-speaking country is 3.73. LINDSEI contains 554 interviews for a total of 1,079.681 tokens and 11 national sub-corpora. The total duration of the interviews is 130:39:28 hours, while the average duration is approximately 14 minutes. The Italian sub-corpus has an average interview duration of 13 minutes and 40 seconds.

Table 8.1: Distribution of tokens in LINDSEI.

NATIONAL SUBCORPUS	Number of interviews	Number of words
BULGARIAN	50	96,064
CHINESE	53	82,536
DUTCH	50	100,454
FRENCH	50	143,887
GERMAN	50	114,916
GREEK	50	103,706
ITALIAN	50	80,047
JAPANESE	51	56,239
POLISH	50	115,722
SPANISH	50	84,749
SWEDISH	50	102,081
LINDSEI	554	1,079,681

LINDSEI transcription convention scheme

The transcription scheme was adapted from Edwards (1992) and her minimalistic transcription standard. The transcription conventions used in LINDSEI are as illustrated in Table XX.

Table 8.3: Transcription convention for LINDSEI.

Text	Rule

Punctuation	No punctuation marks have been used to indicate sentence or clause boundaries. Some punctuation marks have been used to indicate other types of information (cf. empty pauses below).
Empty pauses	Empty pauses are defined as a blank on the tape, i.e., no sound or when someone is just breathing. The following three-tier system has been used: • one dot for a "short" pause (less than 1 second); • two dots for a "medium" pause (between 1 and 3 seconds); • three dots for a "long" pause (more than 3 seconds). <A> (uhu) . why do you like prefer modern art </A> <B> (erm) . . . cos it's strange </B>
Filled pauses and backchannelling	The standard set of filled pauses and back-channeling used to transcribe the LINDSEI data include 'eh' (for a brief filled pause), 'er', 'em', 'erm', 'mm', 'uhu', and 'mhm'. Filled pauses and backchannelling are enclosed in brackets to make it easier for researchers to exclude them from analysis if necessary. (eh), (er), (em), (erm), (mm), (uhu), (mhm)

Unclear passages	A three-tier system has been used to indicate the length of unclear passages: • <X> represents an unclear syllable or sound up to one word; • <XX> represents two unclear words; • <XXX> represents more than two unclear words. <B> that's probably also why I understood and . <X> okay </B> Unclear names of towns or titles of films for example have been indicated as <name of town> or <title of film>.
Uncertain words or word ending	The symbol <?> directly following a word indicates that the transcribers were not entirely sure of the word or word ending they transcribed. <B> yeah and you have some special . way of choices <overlap /> <X> shouldn't<?> <X> <B/>
Truncated words	Incomplete words are immediately followed by an equals sign. <B> it was . in France in Italy well not not in Italy but Britain and Ger= and Germany and <B/>
Contracted forms	All standard contracted forms such as 'I'm', 'he'd' or 'they'll' have been retained.

Non-standard forms	Non-standard forms that appear in the dictionary have been transcribed orthographically in their dictionary accepted way. The non-standard forms that can be found in the LINDSEI transcripts are ‘cos’, ‘dunno’, ‘gonna’, ‘gotta’, ‘kinda’, ‘wanna’ and ‘yeah’.
Foreign words	Learners’ use of words from a language other than English is indicated by <foreign> (before the word or expression) and </foreign> (after the word or expression). <B> I don’t know the the . the name in English but (eh) here we call it <foreign> Traducción e Interpretación </foreign> <B/>
Abbreviations	Abbreviations pronounced as sequences of letters have been transcribed as a series of upper-case letters separated by spaces. <B> (er) but that was enjoyable what I what I what I plan to do after . after I finish my . M A course </B> Abbreviations pronounced as words have been transcribed as a series of upper-case letters not separated by spaces. <B> in NASA in the NASA museum . yes . I liked the different aircrafts spacecrafts <B/>
Date and numbers	All figures have been written out in words to represent the way they were pronounced by the speakers. <B> it’s an old film and . actually I think (em) produced in nineteen eighty-one </B>

Phonetic features:

1. syllable lengthening	1. A colon is added at the end of a word to indicate that the last syllable is lengthened. It is typically used with words like to, so, or or. Colons have not been inserted within words. <B> we went to: Santiago first and then . to the south </B>
2. pronunciation of articles	2. When pronounced as /ei/, the definite article a is transcribed as a[ei]. <B> it was (erm) . (erm) . a[ei] (er) parody . of me . <overlap /> so </B> When pronounced as /ði:/, the definite article the is transcribed as the[i:]. <B> (eh) . I can see the[i:] the sketches of of a lady's portrait </B>
3. foreign pronunciation	3. As a rule, foreign pronunciation is not noted, except in the case where the foreign word and the English word are identical. If in this case the word is pronounced as a foreign word, this is marked using the <foreign> tag (see above). <B> cos I I thought . well I didn't have the[i:] (erm) .. distinc= <foreign> distinction </foreign> so I said well </B>

Anonymisation	To preserve anonymity the proper names mentioned in the interviews have been replaced with substitutes such as <first name of interviewee>, <first name and full name of interviewer> or <name of professor>. The names of famous people like singers or actors have however been kept.

In addition to the transcription scheme, LINDSEI was also marked up in order to signal the beginning and ending of speaker turns, tasks, and interviews. Other information and non-verbal sounds were also indicated with the mark-up scheme, which is as follows:

Table 8.4: Mark-up convention scheme for LINDSEI.

Feature	Rule
Interview identification and delimitation	Each interview transcript is preceded by a code that indicates both the native language of the learner in the interview (two letters of the name of the language, e.g. DU for Dutch, PL for Polish) and the interview number. All interviews end with the symbol </h> on a separate line. <h nt="FR" nr="FR001"> interview </h> In the above example, the code <h nt="FR" nr="FR001"> marks the beginning of the first interview from the French mother tongue component of LINDSEI.
Task delimitation	The beginning and end of each of the three tasks making up the interview (i.e. set topic, free discussion and picture description) have been marked as follows: <S> set topic task </S> <F> free discussion task </F> <P> picture description task </P> The markers used to delimit the three tasks always occupy a separate line.

Speaker turns	Speaker turns are displayed in vertical format, i.e. one below the other. While the letter 'A' enclosed between angle brackets (<A>) marks the beginning of the interviewer's turn, the letter 'B' between angle brackets (<B>) marks the beginning of the interviewew/learner's turn. The end of each turn is indicated by either </A> or </B>. <A> so which topic have you chosen </A> <B> I have chosen topic topic (er) number two </B>
Overlapping speech	The marker <overlap /> is used to indicate the beginning of overlapping speech in the two turns where it occurs. The end of overlapping speech is not indicated. <B> oh I don't remember <overlap /> it </B> <A> <overlap /> you don't remember no </A>
Voice quality	If a particular stretch of text is said laughing or whispering for instance, this is marked by inserting <starts laughing> or <starts whispering> immediately before the specific stretch of speech and <stops laughing> or <stops whispering> at the end of it. <B> (er) nothing sorry <starts laughing> I'm very nervous <stops laughing> </B>
Non-verbal vocal sounds	Non-verbal vocal sounds are enclosed in angle brackets. <B> <coughs> even <coughs> even if you don't tell her .. she's it you know it's something noticeable </B>

Contextual comments	Non-linguistic events are indicated between angle brackets only if they are deemed relevant to the interaction (if one of the participants reacts to it, for example). <B> and (erm) . at the[i:] end . . . <someone enters the room and says I'm sorry> </B> <A> <X> no problem just can you give us ten minutes <laughs> go on </A>

Transcription of interviewer's instructions

At the beginning of the interview, the interviewer would repeat the same prompt to the learners:

"Good morning, I have now started the recording and I will now ask you choose a topic that you would like to talk about for around five minutes. These are the three topics. Please take a minute to make your choice and think about what to say."

L1 congruency translation of collocations by native speakers

First set of collocations analysed in Chapter 6 in Table 8.5.

Second set of collocations analysed in Chapter 7 in Table 8.6.

Table 8.2: Distribution of interviews/words per subcorpus (adapted from LINDSEI, Gilquin et al., 2010, p. 23).

NATIONAL SUBCORPUS	Number of words (%)		
	S	F	P
BULGARIAN	27,709 (40%)	22,633 (33%)	18,827 (27%)
CHINESE	28,261 (44%)	21,277 (33%)	14,004 (22%)
DUTCH	24,460 (31%)	46,870 (59%)	8,322 (10%)
FRENCH	46,076 (50%)	38,836 (42%)	6,490 (7%)
GERMAN	31,721 (37%)	38,494 (45%)	15,737 (18%)
GREEK	30,124 (40%)	32,822 (43%)	13,052 (17%)
ITALIAN	31,234 (52%)	16,977 (28%)	11,362 (19%)
JAPANESE	14,408 (39%)	15,873 (43%)	6,485 (18%)
POLISH	50,747 (54%)	32,706 (35%)	9,668 (10%)
SPANISH	26,588 (41%)	22,666 (35%)	15,550 (24%)
SWEDISH	39,066 (54%)	25,733 (36%)	7,005 (10%)
LINDSEI	350,393 (44%)	314,887 (40%)	126,860 (16%)

Table 8.5: L1 congruency translations of frequent collocations.

Combination	NS 1	NS 2	NS 3	NS 4	NS 5	NS 6	NS 7
Completely different	completamente different*	completamente divers*	totalmente divers*	completamente divers*	completamente divers*	completamente divers*	totalmente different*
Really good	molto buon*	veramente brav*	veramente bene	davvero buon*	davvero buon*	proprio bell*	davvero buon*
Really interesting	davvero interessant*	davvero interessant*	veramente interessant*	davvero interessant*	davvero interessant*	davvero interessant*	davvero interessant*
Really nice	molto carin*	veramente carin*	veramente carin*	davvero bell*	davvero bell*	proprio gentil*	veramente carin*
Totally different	totalmente different*	totalmente divers*	completamente divers*	totalmente divers*	totalmente divers*	del tutto divers*	totalmente different*
Very different	molto divers*	molto different*	molto divers*	molto divers*	molto divers*	molto divers*	molto divers*
Very good	molto buon*	molto brav*	molto bene	molto buon*	molto buon*	molto bell*	molto buon*
Very important	molto important*	molto important*	molto important*	molto important*	molto important*	molto important*	molto important*
Very interesting	molto interessant*	molto interessant*	molto interessant*	molto interessant*	molto interessant*	molto interessant*	molto interessant*
Very nice	molto carin*	molto carin*	molto carin*	molto bell*	molto bell*	molto gentil*	molto carin*
Very strange	molto stran*	molto stran*	molto stran*	molto stran*	molto stran*	molto stran*	molto stran*

Table 8.6: L1 congruency translations of infrequent/unattested collocations.

Combination	NS 1	NS 2	NS 3	NS 4	NS 5	NS 6	NS 7
Absolutely scared	completamente spaventat*	assolutamente spaventat*	decisamente spaventat*	mi ha fatto/ho avuto/avevo stra paura	terrorizzat*	molto spaventat*	no translation
Kind of scared	abbastanza spaventat*	un po' spaventat*	no translation	mi ha fatto/ho avuto/avevo stra paura	no translation	un po' spaventat*	un po' spaventat*
Pretty curious	abbastanza stran*	piuttosto curios*	abbastanza curios*	piuttosto curios*	abbastanza curios*	molto stran*	abbastanza curios*
Quite old-fashioned	abbastanza fuori moda	piuttosto vecchio stile	abbastanza vecchio stile	abbastanza vecchio stile	abbastanza datat*	molto antiquat*	un po' vecchio stile
Really heartbreaking	molto trist*	davvero straziant*	veramente straziant*	troppo trist*	incredibilmente straziant*	davvero straziant*	davvero commovent*
Really really amazing	veramente vearmente stupend*	veramente veramente incredibil*	no translation	davvero davvero splendid*	no translation	assolutamente fantastic*	veramente veramente bell*
Really really really good	molto molto molto bene	molto molto molto bene	no translation	davvero davvero davvero buon*	incredibilmente buon*	proprio ma proprio bell*	super bell*
Super strange	molto stran*	super stran*	molto stran*	super stran*	super stran*	super stran*	davvero tanto stran*