UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Dipartimento di Studi Linguistici e Letterari

Corso di Laurea Triennale Interclasse in
Lingue, Letterature e Mediazione culturale (LTLLM)
Classe LT-12

Tesina di Laurea

A study of connectors usage in English academic texts of Italian and German speakers

Relatrice

Prof.ssa Katherine Ackerley

Laureanda

Maria Elisabetta Floris

n° matr.1199749 / LTLLM

Anno Accademico 2021 / 2022

TABLE OF CONTENTS

Introduction

Corpus linguistics is the study of languages with an empirical approach that consists in the study of written or oral texts, produced by native or non-native speakers, which are collected depending on the purpose of the analysis.
The purpose of this study is to introduce the reader to this subject, demonstrating its importance in linguistic studies, and finally, to focus on a study that shows how different groups of non- native speakers of English vary in their use of sentence connectors.

The first chapter offers a historical outline of corpus linguistic studies, starting from the Corpus linguistics in 1800, when the first corpus-based studies appeared in the form of parental diaries. This study mentions the contrast with the approach of the American linguist Noam Chomsky, who claimed that linguistic studies should focus only on the speakers' personal knowledge of the language and not on the external use of it. Consequently, he believed that corpora were not useful for linguistic studies. The historical subsection concludes with studies in the late 90s, in particular with the appearance of computer-based corpora, which present an important evolution for this branch of linguistics. The first chapter focuses then on the classification of corpora, which depends on the characteristics of the texts and the aim of the study, and with the principal software tools used for the analysis.

The second chapter focuses on a specific type of corpora, namely Learner Corpora, which are collection of texts written by non-native speakers, learners of a foreign language. To understand how they are created, structured, and which are their main purposes, it is essential to introduce second language acquisition approaches. Second language acquisition refers to all the theories that have been formulated during the years concerning the process of learning a second language. The three main approaches that are analysed are Contrastive analysis, Error analysis, and Interlanguage.
Focusing on language for academic purposes, the study shows the importance of learner corpora analysis not only for research purpose, but also for the students, who have the possibility to improve their linguistic skills in the language that they are learning through the use of corpora, analysing their own texts or also those of other students.

The chapter then provides two examples of corpus-based studies on the use of connectors by academic students of English. One was carried out by Altenberg and Tapper (1998) on Swedish learners of English and the other one by Carrió-Pastor (2013) on Spanish students and native speakers.

The last chapter focuses on a study that I carried out of the usage of connectors by Italian and German academic students of English. The study is going to follow a contrastive approach on texts taken from the International Corpus of Learner English ICLE corpus. This corpus, created in the University of Louvain and directed by Sylviane Granger, includes today collection of texts written in English by students of 25 different mother tongue backgrounds. The study is going to analyse quantitative differences and similarities between the two groups in the use of specific connectors, with the purpose of proving or disclaiming an original hypothesis that Italian students tend to use conjuncts more than German students.

Chapter 1

Corpora

1.1. The history of corpora

The aim of this chapter is to present corpus linguistics, starting from a brief historical outline, that shows the most important years for corpus linguistic research.
Corpus linguistics could be defined as the study of languages based on examples of "real life" language use (McEnery, Wilson, 2001). A corpus is indeed a helpful tool consisting of collections of oral or written texts that can be analysed and compared in the different areas of grammar and linguistics. Usually, the texts are collected in different ways, depending on the register that is used; for example, Academic writing, by native or non-native speakers, The Corpus of Contemporary American English (COCA) (Davies, Mark, 2008) or The TV Corpus (Davies, Mark, 2019). Nowadays, with computers, it is easier to examine data from a semantic, phonetic, and syntactic point of view, whereas earlier studies were certainly more inclined to errors.
To recreate the history of corpora it is necessary to underline that, linguistic studies have evolved during the years and that technology had a fundamental impact on this field.

1.1.1. Early corpus linguistics

"Early corpus linguistics" is a term coined in Corpus Linguistics (McEnery-Wilson 2001: 3) to indicate linguistics before the criticism of Noah Chomsky toward the use of corpora as reliable sources in linguistic studies. Indeed, his thesis against the validity of corpora had an important impact on the opinion of many linguists.
Linguistic studies in these years are told to be predominantly corpus-based. (McEnery-Wilson 2001: 2). As Harris (McEnery-Wilson 2001: 3) explains "The approach began… with a large collection of recorded utterance from some language, a corpus. The corpus was subjected to a clear, stepwise, bottom-up strategy of analysis". One important study on child language carried out in the late 1800s was composed by parental diaries that were asked to report the child's locutions. Although this represents one of the earliest stages of linguistic studies, it is used as a source of normative data in language acquisition

research today (McEnery-Wilson 2001: 3). Other collection studies continued to be carried during the years, not only for language acquisition, but also for spelling conventions (Käding, 1897), language pedagogy (Traver 1940), and most of all, comparative linguistics, for which the use of corpora is essential. In 1940 Helen S. Eaton compared the frequency of word meanings in Dutch, French, German, and Italian in a study that appears to be very difficult even for today's standards (McEnery-Wilson 2001:4).

1.1.2. Noam Chomsky

It is believed that the criticism made by Noam Chomsky towards corpora as a source of information during the 1950s was the reason why for approximately twenty years after their use declined. Chomsky's approach reveals the historical contrast between rationalists and empiricists. Rationalists' main aim in linguistics is to develop a theory of language that focuses on the mental process of the creation of human language, while in an empiricist approach natural occurring data are fundamental to prove grammatical accuracy of sentences through evidence.

Chomsky puts the focus of linguistic enquiry not on abstract descriptions but on theories which reflected cognitively plausible models of language, therefore corpora are not considered as proper tools for linguistic studies, as its principal aim is language competence and not performance (McEnery-Wilson 2001: 6). Considering the rationalist's definitions, competence is our personal knowledge of the language, while performance is its external evidence often affected by factors not always strictly related to linguistic.

Chomsky's point of view can be summarised with his own words drawn from an interview conducted by József Andor in 2004:

> Corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, sciences don't do this (Andor, 2004: 97).

In addition, he argues that grammaticality cannot be concentrated in a finite collection of sentences, so corpora are said to be necessarily incomplete and, as a result, unreliable for the scientific purpose of linguistics: the explanation of linguistic competence rather than the enumeration of possible combinations of words.

1.1.3. The late 90s and modern corpus linguistics

All the argumentations raised by Chomsky became object of debates and studies. Linguists later advanced their theories refuting this criticism, conversely illustrating all the pros of corpus-based linguistic studies, even pointing out what could have been the problems in early corpus linguistics.

First of all, corpora are acknowledged to represent an essential component in analysing languages, namely observation. Another important argument emerged in the late 90s with Sampson, who observed that the type of sentences usually analysed by the introspective linguist was completely different from the evidence that usually occurs in corpus (McEnery-Wilson 2001: 14).

Crucial in modern corpus-based studies, and completely absent in Chomsky theories is frequency. The point is explained in a very clear and proper way in corpus linguistics (McEnery-Wilson 2001: 15), where it is demonstrated that corpora are more reliable tools of frequency-based data compared to introspection. To emphasise the importance of studying evidence, Fillmore (1992: 35) describes the "armchair linguist" and writes:

> He sits in a deep soft armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, 'Wow, what a neat fact!', grabs his pencil, and writes something down ... having come still no closer to knowing what language is really like.

Hence, as confirmed also by one of the leading figures in corpus research, Geoffrey Leech, (1992) the corpus methodology is considered more objective, and therefore more suitable for the scientific method than the introspective method is.

As previously mentioned, computers played a key role in corpus linguistics. McEnery reports that the first machine-aided linguistic tool was invented in 1967 by Father Roberto Busa, as he wanted to look for concordances in St Thomas Aquinas's texts. This new system was composed by 10,000 sentences divided into cards, that together with a manual

index could be read and compared (McEnery-Wilson 2001: 20). In the 1960s, at the Brown University, Nelson Francis, and Henry Kučera compiled the first machine-readable corpus that contained, together with the LOB Corpus (Lancaster-Oslo\Bergen) nearly 1 million words (Granger 1998: 6). Considering the today's size of corpora, like the Bank of English that counts 200 million words, the previous ones seem very limited. Taking advantage of this new system, Firth gave his contribution in the 1930s, producing a series of writings whose arguments were and still are shared by many linguists. He states that to better understand languages, it is necessary to pay attention to the *context* of communication. Consequently, Firth and his followers worked on corpora that contained complete texts rather than single sentences. The so-called neo-Firthians gave importance to the concept of collocation, which is the co-occurrence of two or more words together, and that is crucial in the creation of corpora tools, but this will be better explained in the next paragraphs (McEnery-Wilson 2001: 23-24).

1.2. Corpus linguistics

In this section I will describe what corpus linguistics is, its role in the different fields of linguistics, and the use of computer software. Biber, Conrad and Reppen (1998: 4) refer to corpus linguistics by coining the term "The Corpus Approach" (Bennet 2010: 10) that has four main characteristics: firstly, it is empirical, because it is based on real and natural patterns of language use; secondly, it uses a large and reliable collection of natural texts; thirdly, it can rely on computer concordancing programs for the analysis; and lastly, it depends on quantitative and qualitative analytical techniques.

1.2.1. The aim of corpora

As mentioned before, the concept of corpus linguistics is today strictly related to computer-based analysis, which allows the researchers to collect unlimited quantities of texts and usually divide them by varieties of a single language, or by its specific uses, such academic language, spoken language, learner language, and more.

Before deciding the language area that should be analysed, scholars decide which is the principal aim of their research. So, they establish which is the field of linguistics to focus on. Some of them are phraseology, lexicogrammar, register, ESP (English for Specific Purposes), nuances of language, and appropriate syllabus design (Bennet 2010: 8).

Phraseology is the study of one word considered in a sequence of various words: the phrase. It embodies elements such as collocations, lexical bundles, and preferred sequences.

- Collocations gained importance within Neo-Firthians and still have a central role not only in phraseology, but in linguistics as a whole. Collocation is the statistical tendency of words to co-occur (Bennet 2010: 8). With the help of concordancing software that provide statistical data, it is possible to check a recurrent co-use of two specific words.

  For example, collocations of precise adverbs and adjectives. The online corpus Skell (Sketch Engine, 2014) gives statistic data of collocations: while "fully aware" counts 1.46 hits per million, "totally aware" counts 0.02 hits per million.

- The second important element of phraseology are lexical bundles. A description of them is given by Bibier, Johansson, Leech, Conrad, and Finnegan (1999: 990): a lexical bundle is a recurring sequence of three or more words, like "do you want me to". They are not complete phrases, and their meaning is not derivable from their parts (Bennet 2010: 9).

English for Specific Purposes refers to a methodology of language learning and teaching that focuses on specific areas of language use, with precise vocabulary, that one wants to investigate. There are specific corpora that contain only Academic language (e.g. British Academic Spoken English (BASE) corpus, 2001), or language used only in movies (e.g., Movie Corpus, 2019), or, even more specifically, language used in one single magazine (e.g. TIME Magazine Corpus, 2007). They aim to be representative of certain specific type of language, facilitating deeper studies.

Lastly, through corpora scholars can analyse languages from different perspectives, deciding whether with a more general or focused view. The above-mentioned fields are

some of the approaches researchers and teachers adopt to answer their questions or to prove or disprove their theories.

1.2.2. The classification of corpora

Nesselhauf (2011) describes the different types of corpora:

- General corpora, opposed to specialized corpora, aim to represent a language or its variety as a whole, they contain both spoken and written text or different text types (e.g., COCA Corpus). They are usually composed by more than 10 million words (Bennet, 2010: 13).
- Historical corpora aim to represent an earlier stage of a language (e.g., Helsinki Corpus). They are opposed to Present-day language corpora.
- Regional corpora, that represent a regional variety of a language (e.g., WCNZE = Wellington Corpus of Written New Zealand English).
- Learner corpora, in opposition to Native speaker corpora, that represent the language as produced by learners of one language (e.g., ICLE = International Corpus of Learner English).
- Multilingual corpora aim to represent several different languages, often with the same text types.
- Spoken corpora, contrarily to written or mixed corpora, are based on database of audios or transcriptions of oral conversations or speeches.

1.2.3. The importance of corpora in linguistics branches

Linguistic studies in nearly all the branches have obtained great advantages and improvements from the use of corpora. Indeed, it has allowed researchers to complement theoretical studies with research of natural use of language, whether proving or disclaiming them original thesis, or highlight something that cannot be seen only through theory.

To better understand corpora's importance, McEnery and Wilson (2001) offer detailed evidence of their application to the single fields of linguistics: speech research, lexical studies, grammar, language learning.

Speech research has gained benefits from corpora for two main reasons. Firstly, they are vast collections of data, that permit taking into consideration all the variables of context (such as speaker age, context of speech, and genre). Secondly, it is now possible to report real speech, that has not been elicited. Elicited speech usually results in artificial constructions that do not depict natural ones. Therefore, studies are more precise and reliable.

With the help of computer tools, lexicographers nowadays can enumerate the large variety of usage of words and phrases, and the time spent in lexical studies has lessen. Therefore, the writing of dictionaries became faster. Undoubtedly the studies have also improved from a qualitative point of view. Probably a clear example is the Longman Dictionary of Contemporary English, which provides not only the usual definition of lemmas, but also their frequency and some examples of common utilisation, taken from the Longman Corpus Network.

As for lexical studies, also for grammar ones, the creation of corpora has enabled new kinds of analysis. While in the past, subjective hypothesis about rules and frequency of usage were the only tools and resources one could rely on, nowadays it is possible to have objective data, with contexts of use, frequency of specific clauses, and grammatical constructions. So, it is possible to confirm or disprove the real application of grammatical rules. While in the past scholars could count mainly on qualitative data, now scholars have the possibility to analyse quantitative data (McEnery, Wilson 2001: 110).

Text-based studies are essential for historical linguistic studies. While in the past linguists were used to look at original texts and analyse them manually, it is now possible to put them together, compare and divide them by specific periods of time with the help of computer software. In doing so, it is possible to easily find estimates of frequency, repeated and preferred collocations. One example of an historical corpus is the Lampeter Corpus of Early Modern English Tracts (1999), which is a collection of English pamphlets of the 15th and 16th century (McEnery, Wilson 2001: 124).

Another differentiation is made also on regional varieties and sociolinguistics, as for example the Helsinki Corpus does, providing the diachronic evolution of British rural dialects (McEnery, Wilson 2001: 103-125).

Corpora are verifiable, authentic, and objective multifunctional tools with which researchers have reached high levels of specialisations and findings in linguistic studies. In other terms, linguists have the possibility to find out phases of learning and processing languages, both for native and foreign speakers, not only with theoretical basis but with innovated software for data analysis.

1.3. Corpora design

Creating corpora is a process that should follow precise steps. First of all, it is necessary that the texts are "authentic". This means that they are used for communicative purpose, and the unique intention is that of creating a corpus (Bennet 2010: 14-15).
The first aspect that should not be forgotten is obtaining permission to insert the texts in the corpora. Consent to use the sources is necessary to prevent legal issues because of copyright violation (Grigaliūnienė 2013: 27).

1.3.1. Concordance and collocations

Concordancing software is the prevalent tool that is used in corpus analysis. It consists of listing all the sentences that contain the search word, which is put at the centre of the page and highlighted. This type of display is called KWIC (Key Word In Context). Concordancer also allows the researcher to sort alphabetically to the right or to the left. (Grigaliūnienė 2013: 37).
Collocation, as mentioned in paragraph 1.1.3., refers to the co-occurrence of two or maximum four words.

Figure 1 illustrates a concordance for the word "ambivalent" in BNC (2004) sorted to one word to the right (Grigaliūnienė 2013: 39).

Figure 1.Concordance of "ambivalent" in the BNC (Grigaliūnienė 2013: 39).

1.3.2. Frequency

Frequency is probably the most common feature used in corpus analysis; it consists in the number of times that a word occurs in a corpus. It is important to specify that frequency is strictly related to the variety, genre, or register present in that corpus because, naturally, data could vary depending on these components. Frequency is useful evidence that helps scholars to find statistics about recurrent words in languages (Grigaliūnienė 2013: 43).

	BNC	Birmingham Corpus	Brown Corpus	LOB Corpus
the	1	1	1	1
of	2	2	2	2
and	3	3	3	3
a	4	5	5	5
in	5	6	6	6
to (Inf.)	6	4	4	4
it	7	9	12	10
is	8	11	8	8
to (Prep.)	9			
was	10	10	9	9
I	11	8	20	17
for	12	13	11	11
that (Conj.)	13	7	7	7
you	14	14	33	32
he	15	12	10	12
be	16	18	17	15
with	17	16	13	14
on	18	15	16	16
by	19	29	19	20
at	20	22	18	19

Table 1. Rank order of 20 most frequent word types in the BNC, the Birmingham Corpus, Brown Corpus and LOB Corpus (Grigaliūnienė 2013: 44).

Table 1 provides a comparison of the frequency of the most recurrent words in four different corpora. It reveals that the most frequent lemmas are function words (and not content ones), namely prepositions, articles, and particles (Grigaliūnienė 2013: 45). However, the words occurring in the four corpora are not the same, since they focus on different varieties of English.

In addition, Grigaliūnienė (2013: 45) claims that computer tools are used also to collect statistics about types, tokens, and type-tokens.

- Types indicate the number of different words in the corpus,
- Tokens are the total amount of words running in the corpus,
- Type-tokens refer to the number of types divided by the number of tokens.

A widely used software that provides type-token data is WordSmith Tools, which calculates type-tokens for every 1,000 words (Grigaliūnienė 2013: 45-46).

1.3.3. Annotation

Annotation is a procedure that helps researchers to better understand the data they read. It adds linguistic information to the raw corpus, and it can be done at different levels: morphosyntactic, syntactic, semantic, and pragmatic (Grigaliūnienė 2013: 43). Usually, corpora are annotated with grammatical or part-of-speech tags, namely every word is accompanied by its grammatical class that is given automatically by the computer. This electronic operation not only accelerates the procedure of analysis, but also helps the reader to spot meanings of words that could be difficult to capture (Grigaliūnienė 2013: 43). Precisely for the function just mentioned, learner corpora are usually P-O-S annotated, given that there could be errors made by the students, this helps to find out specific words used for particular part of speech (Bennet 2010: 14).

Sylviane Granger (1998: 19) provides a detailed description of the different types of annotation, which she also refers to as "tagging". She claims that software can produce Part-of-speech, syntactic, semantic, discoursal, and error tagging.

<ICLE-FIN-JYV-0001.1>	IGN(left)
We	PRON(pers,plu)
in	PREP(ge)
Finland	N(prop,sing)
are	AUX(prog,pres)
just	ADV(excl)
getting	V(montr,ingp)
a	ART(indef)
law	N(com,sing)
that	CONJUNC(subord)
also	ADV(add)
women	N(com,plu)
can	AUX(modal,pres)
do	V(cxtr,infin)
military	ADJ
service	N(com,sing)
as	PREP(ge)
volunteers	N(com,plu)
.	PUNC(per)

Table 2. Sentence tagged with the TOSCA Tagger (Granger 1998: 20).

Table 2 shows an example of a sentence P-O-S tagged with the software TOSCA tagger. P-O-S tagging consists in attaching a word category tag to each word in a text (Granger 1998: 19).

In conclusion, these are the few principal operations that the invention of computer tools has enabled, it is clear that linguists and learners have nowadays the possibility to inquire about all the most detailed features of languages.

Chapter 2

Learner Corpora

2.1. Second language acquisition

Second language acquisition is a term that refers to all the theories that have been formulated during the years concerning the process of learning a second language. The main goal of SLA research is "to uncover the principles that govern the process of learning a foreign\ second language" (Granger, 1998:4). Gass et al. (1998: 409) identify the 1960s and 1970s as the first years of this field of linguistics. It was precisely in these years that the most influential approaches in this domain were born: contrastive analysis (CA), error analysis (EA), and interlanguage (IL). This section will briefly illustrate these main theories, to offer an introduction to the use of corpora together with the SLA method.

2.1.1. Contrastive Analysis

Contrastive analysis was developed in Europe and North America during the 1950's and 1960's, with scholars such as Fries, Lado and Weinreich. It has always been associated with structuralism. This current stated that language learning is based on an imitation of first language (L1) patterns, and it consequently becomes difficult when the patterns of the target language are different from the L1 (Murad, 2018: 96). CA is a process in which researchers compare languages to determine potential errors and finally to isolate what needs to be learned (Murad, 2018: 96). When it was developed, this method consisted in finding out the similarities and the differences between L1 and L2 at the communicative verbal and non-verbal aspects. The analyses were carried out by teachers in the classrooms, so they weren't corpus-based (Murad, 2018:97).

Therefore, it is possible to predict common errors and find the main difficulties of the learners. This naturally helped pedagogical and teaching studies to build specific learning materials for diversified groups of learners. Despite these advantages, researchers revealed some limits of CA. Firstly, its reliability is strictly related to the similarity or difference of L1 and L2 standards, secondly researchers state that attributing errors to the influence of L1 system only, is too selective (Murad, 2018: 97).

2.1.2. Error Analysis

The second theory is Error Analysis (EA). It was developed in the 1960s by the British linguist Corder and his colleagues (Murad 2018: 98). This method analyses the causes of systematic and frequent errors of L2 learners to later evaluate and correct them (Murad, 2018: 98).

Crystal (1980) states that:

> "In language teaching and learning, error analysis is a technique for identifying, classifying and systematically interpreting the unacceptable forms produced by someone learning a foreign language using any of the principles and procedures provided by linguistics".

The difference with contrastive analysis is that while CA focuses on the link between errors and L1 system, EA dissociates them from L1. This approach only focuses on the level of learner competence and performance in the second language. *Competence* and *performance* are two terms used by Chomsky: *competence* indicates the knowledge of the language, while *performance* refers to the capability to use it in a concrete situation (Chomsky, 1965).

This focus on errors led many scholars to consider this approach too restrictive, ignoring the correct uses by L2 learners (Murad, 2018: 98).

2.1.3. Interlanguage

The third approach to Second Language Acquisition is Interlanguage. It emerged in 1972 with Selinker. Inspired by the works of Weinreich, Selinker states that there's a latent psychological structure that is thought to have a role in SLA. Interlanguage theory's main aim, after having found the errors, is to account for the mental process responsible for specific features of the language of L2 learner (Murad, 2018: 99). Selinker explains that L2 learners develop their own methods for the L2, and they build a linguistic system that can be influenced both from their L1 and L2 but is separated from them. Selinker saw interlanguage as unnatural as it is not linked to language universally but is something subjective (Murad, 2018: 99).

In conclusion, one can say that all the different approaches to Second Language Acquisition research have provided and still are providing many new methods that are useful both for learning and for teaching. While for example IL helps to analyse the learning processes, CA supports learners with the comparison to native speakers or with other learners, helping them to improve their skills. A corpus-based approach to these studies, more precisely learner corpora, give important empirical data that help teachers, researchers, and students to find out the natural use of certain structures or categories of words in the target language.

2.2. Learner corpora

2.2.1. Learner corpora design

The aim of this section is to go through the steps of the creation of a learner corpus, analysing the principal types of learner corpora used in SLA.

Learner corpus research is a branch of corpus linguistics focused on non-native varieties. It emerged in the late 1980s. Granger (2015: 1) states that "like any corpus, the learner corpus is a collection of machine-readable authentic texts, which represent a particular language or language variety […]; it represents language as produced by foreign or second language learners".

The language contained in the corpus should be "natural" or rather, "near-natural" (Gilquin, 2015: 10). Indeed, learner collections of writings or speech, even if not elicited, are usually restricted to the classroom context, so related to a specific task. Therefore, while normal corpora language can be considered as natural and genuine, given that their only purpose is communication, the aim of texts produced for learner corpora is pedagogical, so the output cannot be seen as totally natural (Gilquin, 2015: 10).

When compiling a learner corpus, the researchers have the possibility to consider many variables that should be shared by all the data present in the corpus. They could be for example the proficiency level in the target language, or the communication environment, so the setting in which the learners use the language (they could use them only at school, or also outside the classroom). Another important feature is the distinction between oral and written data (Gilquin, 2015: 17). Among others, Granger (2002: 9) also mentions the

mother tongue of the learners, their use of reference tools and the audience or the interlocutor.

Granger (1998: 8) summarizes the principal criteria of the design of a learner corpus in Table 3.

Language	Learner
Medium	Age
Genre	Sex
Topic	Mother tongue
Technicality	Region
Task setting	Other foreign languages
	Level
	Learning context
	Practical experience

Table 3. Learner corpus design criteria (Granger 1998: 8).

Indeed, learner language is influenced by a wide quantity of variables. The compilers are free to decide which of the distinctive features should be considered as necessary in the design of the corpora.

2.2.2. Learner corpus typology

Granger (2002:10) reports that corpus typologies are usually classified in dichotomies, and four of them are particularly relevant for learner corpora. They will be illustrated in this section.

- *Monolingual and multilingual corpora*

  Monolingual corpora are the most common type of corpus. They contain one single language, and they can be used for analysis such as checking the different uses of a word or a phrase, observing the frequency with which they are used, or also identifying common errors.

  In contrast, multilingual corpora are composed of two or more monolingual corpora. They are usually similar in their features to allow the researcher to compare them and find similarities or differences (sketchengine.eu).

- *General and technical*

  As explained before (see 1.2.1) the different communicative context causes different language structures and requires particular lexis. In order to investigate the development of specific a feature or paradigm it is necessary that the corpus contains a sufficient collection of tokens of that particular feature for the results of the study to be reliable.

  Technical corpora, also known as "specialized" corpora, can answer very restrictive questions or also more open ones. They may be for example corpora containing Academic writing only.

  General corpora are instead composed by texts not related to specific features or themes but can offer a general and complete idea of the chosen language.

- *Longitudinal and cross-sectional corpora*

  Being able to carry out a study that follows all the steps of the learning process of precise groups of people would be useful for the analysis of development of SLA, but it is nearly impossible. This is due to the impossibility for scholars to carry out studies that would cost time and finances. It is also difficult to collect data from the same group of people in a very long period of time. Myles (2015: 316) mentions that the longitudinal studies that have been carried out refer to a short period of time - months or years – and to small numbers of learners.

  Therefore, cross-sectional corpora are preferred by SLA researchers. They consist in data collected at a particular point in time, belonging to groups of learners of similar proficiency levels, that are considered to be representative of that particular level. In the process of collecting texts, what is essential to make this generalisation is to test the level of the members of the study, to better classify the different groups by proficiency level, and later compare them (Myles, 2015: 316).

- *Written and spoken*

  Written and spoken corpora are equally valid in the studies of SLA, but they are more suitable for certain types of analysis than the others could be.

  If the research aim concerns the spontaneous employment of a language, spoken

corpora are better, because oral production is influenced by the real-time pressure that "does not allow recourse to the metalinguistic system to the same extent, giving a better window to implicit knowledge" (Myles, 2015: 314).

Differently, written corpora are suitable for other types of investigations such as the development of lexis and discourse, the memorisation of fixed expressions, or the application of rules (Myles, 2015: 313). This is helped by the possibility to check errors after writing and reflect on the language before handing out the text.

2.2.3. Learner corpora and language for academic and specific purposes

In this subsection I will introduce the use of learner corpora in the field of Language for Specific Purposes (LSP) and in particular that of English for Academic Purposes (EAP). Gnutzmann (2009: 218) states that although "there seems to be no universally accepted definition of LSP", scientific and technical writing uses the same grammar as that of general English, but displays register-specific frequency patterns of both grammar and lexis (Flowerdew 2015: 465).

LSP studies are very wide in their purposes, and they split in two different approaches, one developed in Europe, the other Anglo-American. While the European approach focuses on terminology, lexicography, and translation of a specialised subject domain, the Anglo-American one gives importance to the social and cultural context in which the language is learned (Flowerdew 2015: 465).

Given its international domination in communication and research, English is the central language in LSP studies. Precisely for this relevance, linguists have coined the definition "English for specific purposes" (ESP), that can be divided into two approaches: English for Occupational Purposes (EOP), linked to the professional or workplace environment, and English for Academic Purpose (EAP), that concerns student's communication needs and skills (Flowerdew 2015: 466).

The studies carried out in ESP and EAP are reported in the *Journal of English for Academic Purposes* and *English for Specific purposes*, two of the leading journals in this field.

There are some procedures that are relevant in EAP studies. Firstly, a Contrastive Analysis approach is usually employed, indeed it helps to better understand not proper

forms in the learner writings (Flowerdew 2015: 469). Secondly, there are three main approaches that are usually chosen in EAP research, they are keywords, lexical-bundles and pattern grammar. The first refers to distinctive lexis of specific genres, the second to important fixed groups of words in the discourse, and the third to lexico-grammar patterns that tend to co-occur in precise genres (Flowerdew 2015: 470-474).

In conclusion, it is relevant to say that corpus application in ESP and EAP fields today does not only serve as empirical data source for preparation of reference materials, but it is also used by the students themselves in their classroom activities.

As a matter of fact, scholars state that this helps the process of learning owing to the possibility to find alternatives of some common words or phrases, or also find evidence of repeated structures in the given genre.

2.3. Connectors in learner corpora

In this section I will cite two main studies carried out on learner corpora, particularly on the use of English connectors by L2 students. To do so I will first introduce what connectors are, their categorisations and uses.

Adverbial connectors serve a connective function between two units of discourse, that can have different sizes. They may be sentences or also smaller than sentences, such as clauses (Biber, Conrad, Leech, 2002: 356). They express various relationships and Quirk et al. (1985: 634) divides connectors in their various semantical uses:

- Listing. These have a role of listing elements or adding information to the discourse, they are for example *firstly… secondly, in addition, furthermore.*
- Summative. These are used to clarify when a discourse is finished, and it is summarized. Some examples are *in sum, then, to conclude or all in all*.
- Appositional. These introduce the second unit as a restatement of the first. Some common examples are *i.e., for example, namely, in other words.*
- Resultive. These indicate that the second unit of discourse is a result or a consequence of the preceding unit. For instance, *consequently, thus, so, hence, therefore.*

- Inferential. These indicate a conclusion based on logic and supposition. They are for example, *otherwise, in that case, then*.
- Contrastive. These refer to some kind of conflict between the parts, they are namely *in contrast, alternatively, though, anyway, besides, on the contrary.*
- Transitional. These introduce a unit that does not directly follow the previous part, for example *meanwhile, by the by, incidentally.*

2.3.1. The use of adverbial connectors in advanced Swedish learners' written English

The first study I would like to refer to was carried out by Bengt Altenberg and Marie Tapper (Altenberg, Tapper, 1998: 80) and focuses on the use of connectors by advanced Swedish learners of English.

Starting from a brief description of what adverbial connectors are, they say that they are useful in the clarification of discourse clarity and coherence, even if they are not always required.

Moreover, the use of connectors is linked to register and discourse type, so learning how to use them goes together with the production of specific genres. Altenberg and Tapper add that connectors and their uses are different in the different languages, so for some learners it could be more difficult than it is for others (Granger, 1998: 81).

The research is a contrastive study of Swedish (SWICLE, referring to the Swedish subcomponent of the ICLE corpus) and native English (LOCNESS) writings, that it is not concerned with errors but with quantitative data of the under- and overuse of connectors by the two groups. The authors also provide a comparison with a corpus of L1 Swedish language, to analyse the possible link with the native language of students.

The analysed materials are argumentative or expository essays, with an average length of 500 words, all written by university students, most of them in their twenties. They were taken from the International Corpus of Learner English (ICLE).

The authors first report the overall frequency of conjuncts in L1 Swedish, Swedish writings of English and native speakers' writings. As previously theorized by the authors, it resulted that Swedish students tend to underuse conjuncts both in their native language and in English. While amongst native speakers the frequency is from 1 to 25, in Swedish

learner's essays is from 0 to 16, and it is nearly the same in their writings in their L1. Starting from the classification of connectors given by Granger and Tyson (1996) adding the "corroborative" subcategory, they also showed frequency in the use of conjuncts related to semantic types (Granger, 1998: 84-85).

Category	L1 Swedish		SWICLE		LOCNESS	
	n	*n* per 10,000	*n*	*n per* 10,000	*n*	*n* per 10,000
Listing	39	11.0	57	11.2	56	11.1
Summative	2	0.6	12	2.4	10	2.0
Appositive	69	19.5	70	13.8	53	10.5
Resultive	79	22.4	80	15.8	142	28.0
Contrastive	114	32.3	95	18.7	169	33.2
Transitional	0	0	2	0.4	0	0
Corroborative	26	7.4	50	9.9	52	10.3
Total	329	93.2	366	72.2	482	95.0

Table 4. Distribution of semantic types of conjuncts (Granger 1998: 85).

The study seems to show that Swedish learners overuse appositive conjuncts (i.e., *for example, for instance*), but this detail emerged only because of the preference of native speakers of English for the form *such as* that has the same function of appositive conjuncts but is not considered as one of them.

The researchers gave some data about the use of single connectors in the texts. The study revealed the 10 most used connectors by the two groups of students and showed that native speakers tend to exploit a limited set of connectors. In addition, there are some overused and underused conjuncts. For instance, *furthermore* is very common in SWICLE, while rare in LOCNESS, on the contrary, the resultive *thus* is underused in SWICLE and recurrent in LOCNESS (Granger, 1998: 86).

In order to find some features that are related to L1, the Swedish sub-corpus was compared with the ICLE sub-corpus of French learners of English, studied by Granger and Tyson (1996). The study revealed both similarities and differences between the two groups. Firstly, there is an evident underuse of conjuncts within Swedish learners that is not shown within the French ones. Because no evidence of a relation with the native

language emerged, a possible reason is the different educational background.
In addition, the similar items appeared to be over- and underused both by the Swedish and French learners, so another element precludes the transfer from L1 systems (Altenberg, Tapper, 1998: 88-91).

In conclusion, the study revealed in quantitative terms that there is no evident link between the use of conjunctions in English by the Swedish students and their native language. The only problem that emerged is that, equally to the French students, they tend to use fewer formal connectors instead of the formal ones used by English students (Granger, 1998: 92).

2.3.2. A contrastive study of the variation of sentence connectors in academic English

This study carried out by María Luisa Carrió-Pastor in 2013 analyses the differences in the use of connectors in EAP, with the aim of finding whether native English speakers (NES) and non-native English speakers (NNES), in particular L1 Spanish speakers, make the same choices or not. More specifically, the analysis focuses on the influence of the linguistic and social background of the writers. To emphasis the concept of background, Carrió-Pastor takes the term "variation" used "to refer to the different manifestations of a given concept in a particular language which are neither mistakes nor errors and which appear in texts written by authors with different linguistic backgrounds" (Carrió-Pastor, 2013: 193). She states that the creation of a fluent discourse depends on various features, namely the influence of the mother tongue, social construct, and the writer's identity.
The study also analyses the use of connectors across the different sections of the research papers and aims to discern the possible causes of the variation in the use of sentence connectors (Carrió-Pastor, 2013: 193). Differently from the previous study, in which the subjects of the essays were various, here the chosen texts belong to the field of engineering.
The first results showed by Carrió-Pastor are the overall frequencies of connectors divided by Quirk (1985) categories (see section 2.3). NNES appear to use fewer listing, inferential, and contrastive connectors than NES. Secondly, the most used connectors appear to be the contrastive *however* and the appositional *for example.* What comes out

is that NNES writers use less variety of connectors in comparison with native speakers. In other words, Carrió-Pastor states that NNES do not underuse entire categories of connectors, but some of the connectors included in them, and for her, it is related to the personal style of the single writers (Carrió-Pastor, 2013: 195).

To analyse the use of connectors in the various sections of research papers, they were divided into abstract, introduction, methodology, result, discussion, and conclusion.
In the introduction, contrastive connectors are the most frequent because the writers tend to claim a niche for their research, creating a contrast between their and the other studies. In the methodology section, transitional, inferential, and appositional connectors were the most recurrent. Resultive and appositional connectors were predominant in the result section, this is linked to the necessity of showing data and results. In the discussion section, inferential and contrastive conjunctions dominate, while in the conclusion there is a majority of resultive and summative connectors.
Both NES and NNES are aware of the use of connects in the different sections of their papers, and their uses are generally equal, apart from some variation (Carrió-Pastor, 2013: 198-200).

In conclusion, the study confirmed the initial theories of the study: language is used in different ways depending on the linguistic background of the writers, both common and personal, for example "the Spanish writer may be more concerned with the ideas being expressed than with the rhetorical devices needed to guide the reader" (Carrió-Pastor, 2013: 200), so they organize the display of their thesis differently from native speakers.

The topic exposed in this chapter is introductory to the following chapter. More accurately, I decided to conduct a study on a learner multilingual corpus, namely the ICLE corpus, that has also been object of the above-mentioned studies. The study has been carried out with a contrastive approach, with the purpose of detecting whether academic students of English with different L1s use connectors in a similar way or not.

Chapter 3

Contrastive study on connectors in German and Italian students of English

3.1. Introduction

This is a contrastive study that analyses the differences in the use of conjuncts amongst German and Italian learners of English. Conjuncts have a "superordinate" role as compared with other clause elements. They have the role of conjoining independent units, so they should be considered as autonomous from the grammatical unit in which they appear (Quirk et al. 1985: 631). Adjuncts, on the other hand, are linked to the sentences in which they are located.

Quirk et al. (1985: 631) underline the difference between conjuncts and adjuncts. For example, taking the lemma *nonetheless:*

> She may be unable to attend the meeting. You should *nonetheless* send her the agenda.
>
> *It is *nonetheless* that you should send her the agenda.

Conjuncts, that here will also be called "connectors", are useful, but usually not mandatory, in the process of understanding a text. They contribute to the cohesion and coherence of the writings. Not being mandatory, a high frequency of conjuncts does not necessarily improve its cohesive quality, but on the other hand their underuse and misuse are likely to reduce the comprehensibility of the text (Crewe 1990).
Hence, the use of connectors is problematic for language learners, and a huge number of studies have been and still are being carried out to understand the process of learning. To further complicate learning, the variety of registers and genres requires groups of connectors that are more suitable than others.

Inevitably, connectors are different both in their categorisation and usage across all the languages, therefore students will have varying degrees of difficulty in learning them when they study foreign languages. It is worth mentioning the focus that Pastor (2013)

puts on the context of learning and practising the target language. She claims that connectors may be used in a different way by authors of different linguistic, social, and personal backgrounds, making different rhetorical choices and constructing a specific identity in their texts (Pastor 2013: 193). Consequently, all these elements play a role in the inevitable differences that will emerge in the usage of connectors between native speakers and L2 speakers. Many contrastive studies have been carried out regarding this topic, focusing for example on under- and overuse of connectors, such as Yang and Sun (2012) and the previously mentioned Altenberg and Tapper (1998), which compared native speaker and L2 use of connectors.

The aim of my study is to analyse the use of connectors amongst Italian and German students of English with a contrastive approach, hypothesising that Italian students use sentence connectors more than German students. I will focus on quantitative data, namely the under- or overuse of specific connectors in the two groups of writers. Secondly, I will compare them with the data resulted from other studies, more precisely Carrió-Pastor (2013) and Altenberg and Tapper (1998) to identify whether certain languages are more similar than others.

3.2. Material

This study has been carried out with texts written by German and Italian students of English, collected in The International Corpus of Learner English (ICLE). It is a corpus of essay writing by upper intermediate and advanced learners which was founded and coordinated by Sylviane Granger. When its first edition was released in 2002 (uclouvain.be), it contained 2.5 million words produced by learners of 11 mother tongue backgrounds. The second version, released in 2009, counts 3.7 million words and 16 mother tongue backgrounds. There is a new version, which still has to be completed, that includes 25 mother tongue backgrounds (uclouvain.be).

The two subcorpora that I have chosen for my analysis are the Italian and the German subcorpora, respectively from the University of Torino, and the University of Augsburg. There are 268 texts in the German sub-corpus and 180 in the Italian one. More specifically, the German corpus counts 131706 tokens and 11222 types; the Italian corpus counts 107592 tokens and 6228 types.

The texts were tagged with TagAnt (2015) with Part-of-Speech (POS) tagging (see 1.3.3), while the analysis was carried out with the software AntConc (2019), which allows the researcher to find Types and Tokens, concordance, collocates, frequency and the other main data necessary for corpora analysis. However, a manual stage was essential to distinguish the different uses of the chosen connectors, that often have not only the role of connectors but also that of adverbs (see 3.1). The following sentences are examples of two different uses of *therefore* in the German sub-corpus of the ICLE corpus: the first one has the role of adjunct, while the second one that of connector.

> "Tennis had developed to one of the most famous and *therefore* endlessly broadcasted sport in Germany but TV commercials destroy the joy of supporting the players feverishly."

> "Most cyclists who are run over by a car become badly injured - not to name their bike which in most cases cannot be repaired any more. *Therefore*, in my opinion, one has to consider both sides of cycling: The pleasures and the dangers."

3.3. Contrastive analysis on connectors usage

3.3.1. Specific connectors

I will focus on specific groups of conjuncts, find their frequencies in the two corpora, and compare the results. The aim of this analysis is to see if there are differences and similarities between the two groups of learners. I have chosen three groups of conjuncts of the semantical classification given by Quirk et al. (1985: 634), precisely resultive, listing and contrastive conjuncts.

Resultive conjuncts are used to signal that the second unit of discourse is a consequence of the preceding unit. As can be seen in Table 5, nearly all of them, except for *hence* are used more frequently by Italian students than by German ones. The connector *therefore* reveals a conspicuous difference between the two groups and a high use in the Italian corpus. The same occurs with the connectors *consequently* and *as a result.* By contrast, with *thus* and *hence* the data are different. Concerning *thus*, it appears that both German and Italian learners make a balanced use of it, even though Italian students seem to use it a bit more. Relevant is the different flexibility of the collocation of the connector *thus* in the sentences. If in the Italian sub-corpus students place it at sentence initial, German

texts have a less strict structure. The connector that appears to be almost absent from both corpora is *hence.* Indeed, only one Italian and two German students chose to use it.

Conjuncts		**German**		**Italian**	
		n	*n per 10,000*	*n*	*n per 10,000*
Resultive	*Therefore*	19	1.44	62	5.76
	Consequently	7	0.53	22	2.04
	Hence	2	0.15	1	0.09
	Thus	18	1.36	27	2.5
	As a result	1	0.07	14	1.3

Table 5. Four resultive connectors in German and Italian texts.

The second group of connectors is that of listing connectors (see Table 6), which introduce further information about the previous unit of text.

Equally to what emerged with resultive conjuncts, listing ones are used very often by Italian students, while German ones prefer not using this category of connectors. The most relevant frequency is that of *moreover*, which is considerably overused in the Italian corpus, and underused by German learners. A relevant element that stands out is the position of *moreover* in the Italian texts. As showed in Figure 2, this listing connector (62 tokens out of 90) is used as sentence initial conjunct and it is followed by a comma, which isolates it from the following sentence. German students not only use the lemma *moreover* very few times, but also, they use it both as conjunct and as adverb, as a result it has not a fixed position in the sentence (see Figure 3).

Conjuncts		German		Italian	
		n	*n per 10,000*	*n*	*n per 10,000*
Listing	*Furthermore*	9	0.68	30	2.78
	Moreover	13	0.98	90	8.36
	In addition	3	0.22	20	1.85
	Besides	12	0.91	19	1.76

Table 6. Four listing connectors in German and Italian texts.

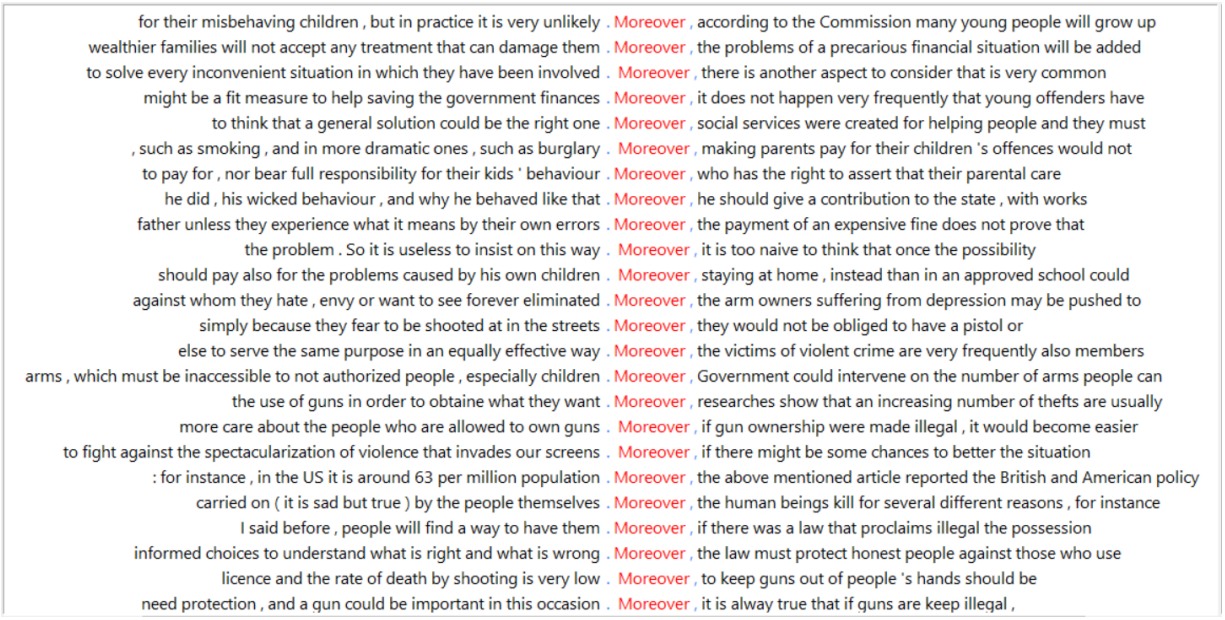

Figure 2. Collocates with KWIC of the connector "moreover" in the Italian sub-corpus of ICLE corpus.

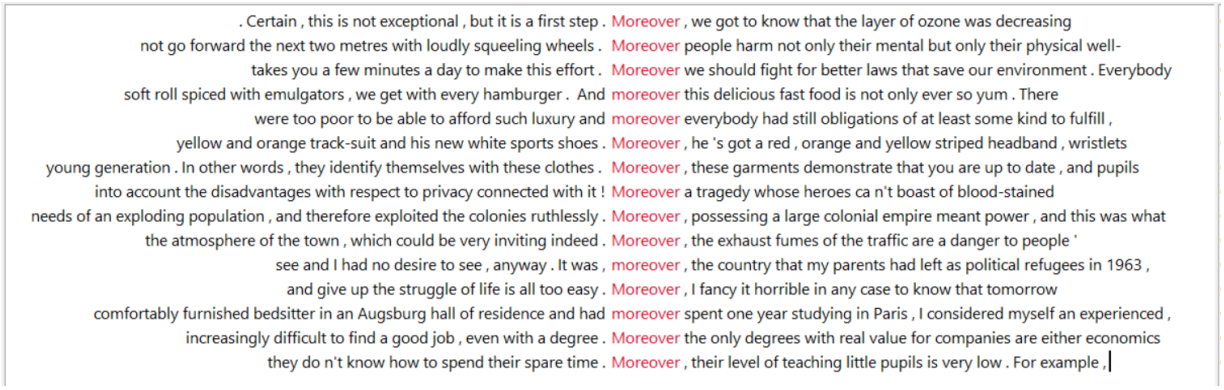

Figure 3. Collocates with KWIC of the connector "moreover" in the German sub-corpus of ICLE corpus.

The third group chosen for the analysis are contrastive connectors, which introduce a conflict between the two units of text. When compared to the other two categories of connectors, these ones are used less frequently. Taking into analysis some of them, the most underused connector both in the Italian and German corpus is *in contrast*. Considering the frequency of usage in the afore mentioned categories shown in Table 5 and 6, this table shows that the frequency of use between the two L1s does not differ so much. Indeed, except for *on the contrary*, which occurs 22 times in the Italian corpus and 5 times in the German one, the data seem to be balanced. The similarities are reflected also in the syntactic position of the connectors. More specifically, both *nevertheless* and *on the other hand* are used by the two groups at sentence initial.

Conjuncts		**German**		**Italian**	
		n	*n per 10,000*	*n*	*n per 10,000*
Contrastive	*On the contrary*	5	0.37	22	2.04
	In contrast	3	0.22	1	0.09
	Nevertheless	20	1.51	21	1.95
	On the other hand	20	1.51	23	2.13

Table 7. Four contrastive connectors in German and Italian texts.

Taking all this into account, it appears that Italian students are more inclined to use, and in some cases overuse, sentence connectors, when compared to German students. More specifically, the most used category in the Italian sub-corpus is that of listing connectors, with 90 hits only for the conjunct *moreover*, compared to the German sub-corpus that reveals very low frequency in this category, e.g., it counts only 13 hits for *moreover* (see Table 6). The contrastive category is the one that reveals more balanced frequency between the two groups, indeed, except for *on the contrary*, the other analysed conjuncts are used equally by the two groups of students.

Another difference that emerged in the study is that German students tend to place some connectors, namely *moreover* and *thus,* in varied positions in the sentence, while Italian students place them at sentence initial.

3.3.2. Comparison with other studies

Before ending the study, I will compare the results of this analysis with those of other researchers to find out if there are any similarities between Italian, German, Spanish, Swedish and French students. Altenberg and Tapper (1998: 87) show that Swedish advanced students underuse resultive connectors, numbers are similar to the German ones showed in Table 5, though German students use these conjuncts even less than the Swedish learners. For instance, the connector *therefore* appears 26 times in the SWICLE corpus, (the Swedish sub-corpus of the ICLE corpus) (Altenberg and Tapper, 1998: 87) while the German sub-corpus counts 19 hits. Meanwhile, the study carried out by Pastor (2013: 133) shows that Spanish academic writers use resultive connectors more than native speakers do, notably the conjunct *therefore* is the most frequently used. Nevertheless, Italian students appear to use them the most.

Observing the group of listing connectors, the present study revealed a conspicuous difference in frequency between Italian and German students. As a matter of fact, while the German corpus is low in this category, Italian learners are inclined to use them more, so much so that the corpus counts 90 hits for *moreover* as sentence conjunct, compared to the German corpus that counts only 13 hits. The comparison with the study of Altenberg and Tapper (1998: 90) seems to cast doubts about a hypothetical L1 influence for this specific connector: Swedish and French alike overuse it. Conversely, Pastor

(2013: 195) states that the category of listing connectors is the most used both in Spanish and English corpora, even if the connector *moreover* is not among the most used by these two groups. Alternatively, the second most frequently used connector (Pastor 2013: 196) is the listing *furthermore* both in NS and NNS lists. The element that brings together Spanish and Italian writers is *in addition,* which is very frequent in the two corpora.

The third category taken into account in my study is that of contrastive conjuncts. This group of connectors is highly used by all the groups, except for Italian and German students (see Table 7), which seem to underuse the connectors selected for my analysis. Nevertheless, there is a common tendency to use the contrastive *on the contrary* in Italian, Spanish (Pastor 2013: 196), Swedish and French Altenberg and Tapper (1998: 91) students, while the German corpus counts only 3 hits (see Table 7).

To conclude, this brief analysis reveals that German students seem to underuse connectors, when compared to the other groups of learners. However, it is essential to specify that this study has its limits. In fact, only some connectors have been analysed. Hence, it would be interesting to see if a wider and more complete study could disprove or confirm this underuse of conjuncts.

3.4. Conclusion

The leading conclusion confirms the original hypothesis: Italian academic learners of English use sentence connectors more than Germans do. All three analysed categories, resultive, listing, and contrastive connectors, show a higher usage in the Italian corpus, with also some cases of overuse. Secondly, it appeared that Italian students in some cases opt for fixed structures, namely they place the conjunct in sentence initial position (see Figure 2).

The causes of these evident differences should be also searched in a comparison with texts written in their native languages, even though the comparison with other groups of learners, in section 3.3.2, highlighted some similarities between the Italian sub-corpus of ICLE and the others, which seems to disprove L1 effect. That is, Italian, Swedish, English and French do not share the majority of their grammatical features and structures, so a hypothetical effect of mother tongue structures is unlikely to be the cause. A huge number

of elements play important roles in the process of writing in a second language. The teaching methods, the aim of the essay, and the academic, personal, and social background of students all have an important impact in the choice of words, and in the overall structure of the texts. Schleef (2009: 1107) gives an important interpretation of German choices in academic writing. When comparing their strategies with American ones, he cites House (2006: 250–251) and claims that "speakers of German tend to prefer more content-oriented strategies, such as explicitly introducing a topic and expanding on it".

However, as previously said, this study is limited in data, so conclusions are not totally complete. Further studies could provide more precise answers, for example analysing the use of connectors with different approaches. Corpus-based studies of English for Academic Purposes both on native and non-native speakers contribute to provide complementary materials and give a more complete overview of the various processes.

Conclusion

The aim of the contrastive study was to find out if Italian students of English use sentence connectors more than German students. To do so it was necessary to introduce corpus-based studies in linguistics. Empirical data are essential materials in the study of languages, they help scholars to find out the natural uses of syntactical structures, lexical choices, or the application of grammatical rules of a given language. The possibility to categorise texts by written or oral production, by genre of the texts, and by age, linguistic background, mother tongue and learning context of the speaker allows researchers to conduct very targeted studies and find similarities and differences between them. This is helpful for the development of linguistic studies in all its branches: syntax, morphology, phonetics, and semantics. Indeed, if theoretical studies provide general and standard rules that define the single languages, natural data offer their real application, distinguishing results in the light of the previously mentioned features. These studies also allow scholars to understand how the use of languages vary during the centuries, in fact texts that are collected in corpora are not only present-day texts, but also historical texts.

A Corpus-based approach determines various advantages in the studies of Second language acquisition. When adopting an Error analysis, Interlanguage or Contrastive approach, scholars and students have the possibility to study target languages not only by standard theorical grammars. This method allows students to have a more dynamic approach to the process of language learning, and at the same time it provides the teachers the data necessary to find out the more recurrent errors in learning, so that to improve teaching methods.

Contrastive analysis approach plays a significant role in this context. Firstly, since these studies often aim to help teachers and students to understand learning processes of the target language, the process of comparison with native speakers becomes fundamental to understand the learners' skills. Secondly, scholars adopt this approach also to compare groups of non-native speakers to study possible shared features in the use of a common target language, and consequently prove or disclaim hypothetical L1 effects in this process.

In this context, the studies described above provide both NS\NNS and NNS\NNS contrastive analysis, revealing that the use of connectors does not generally depend directly on L1 influences in the use of the target language.

The use of connectors in written academic essays depends on a variety of elements, some of them shared by the group, some personal. Rhetorical choices of the writer, personal cultural and linguistic background play an important role in the building of the texts, and consequently in the use of sentence connectors that could help to better deliver the message. Teaching methods and learning contexts are two other main features that contribute to the choice of connectors. These ideas are exemplified by the results of the study of connectors used by Spanish learners conducted by Carrió-Pastor, she claims that "the Spanish writer may be more concerned with the ideas being expressed than with the rhetorical devices needed to guide the reader" (Carrió-Pastor, 2013: 200).

The comparison between the Italian and German sub-corpora of the ICLE corpus revealed that overall, German students use fewer conjuncts than Italian students, so it confirms the original hypothesis. The most conspicuous difference is shown by listing connectors, that are overused by the Italian group and underused by the German ones. The category that showed balanced frequency between the two groups is that of contrastive connectors. These are not highly used by either groups, for example *in contrast* appears only once in the Italian sub-corpus and three times in the German one.

The reasons why these results are not totally evident, because various features contribute to the process. Hence, to have a clearer overview, it would be necessary to conduct other studies. On the one hand, a study of connectors' usage in the mother tongues would clarify hypothetical similar choices both from a semantical and syntactical point of view. On the other hand, a wider study which includes a higher number of texts, and a wider range of L1s could reveal different semantic choices.

References

Andor, J., 2004. *The master and his performance: An interview with Noam Chomsky*. January 12, *Intercultural Pragmatics*, 1-1, pp. 93–111. https://www.degruyter.com/document/doi/10.1515/iprg.2004.009/html

Bennett, G. R., 2010. *Using Corpora in the Language Learning Classroom: corpus linguistics for Teachers*, Michigan ELT. https://www.press.umich.edu/pdf/9780472033850-part1.pdf

Biber, D., Conrad, S., & Reppen, R., 1998. *Corpus Linguistics: Investigating Language Structure and Use*, New York: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finnegan E., 1999. *Longman Grammar of Spoken and Written English*, Harlow: Pearson Education Limited.

Carrió-Pastor, M.L., 2013. *A contrastive study of the variation of sentence connectors in academic English,* in *Journal of English for Academic Purposes,* 12, pp. 192-202.

Corpus types: sketchengine.eu https://www.sketchengine.eu/corpora-and-languages/corpus-types/

Crewe, W.J., 1990. *The illogic of logical connectors*, ELT Journal 44, pp. 316-325.

Crystal, D., 1980. *A first Dictionary of Linguistics and Phonetics*, London: Andre Deutsch.

Davies, A., Elder, C., 2004. *Handbook of Applied Linguistics*, Blackwell Publishing.

Fillmore, C., 1992. *Directions in Corpus Linguistics*, Proceedings of Nobel Symposium 82.

Flowerdew, L., 2015. *Learner corpora ad language for academic and specific purposes* in Granger, S., Gilquin, G., Meunier, F., *The Cambridge Handbook of Learner Corpus Research*, United Kingdom: Cambridge University Press.

Gass, S., Fleck, C., Leder, N., Svetics, I., 1998. *Ahistoricity revisited. Does SLA have a history? Studies in second language acquisition.*

Gilquin, G., 2015. *Learner language*, in Bibier, D., Reppen, R., (eds.) *The Cambridge Handbook of Learner Corpus Research*, United Kingdom: Cambridge University Press.

Gnutzmann, C., 2009. *Language for specific purposes vs general language,* in Knapp, K., Seidlhofer, B., Widdowson, H., (eds) *Handbook of Foreign Language Communication and Learning*, Berlin: Walter de Gruyter, pp. 517-544.

Granger, S., 1998. *Learner English on computer*, Harlow: Addison Wesley Longman Limited.

Granger, S., Hung, J., Petch-Tyson, S., 2002. *Computer learner corpora, second language acquisition and foreign language teaching, Language learning and Language teaching*. Amsterdam: Benjamins.

Granger, S., Tyson, S., 1996. *Connector usage in the English essay writing of native and non-native EFL speakers of English*, World Englishes, 15, pp. 17–27.

Grigaliūnienė, J., 2013. *Corpora in language studies,* Vilnius: Vilnius University.

House, J., 2006. *Communicative styles in English and German*, in *European Journal of English Studies* 10, pp. 249-267.

McEnery, T., Wilson A., 2001. *Corpus Linguistics, an introduction*, Edinburgh: Edinburgh Textbooks in empirical linguistics.

Murad, A., 2018. *Approaching the Language of the Second Language Learner: Interlanguage and the Models Before*, English Language Teaching, 11, pp. 10: 95-108.

Myles, F., 2005. *Interlanguage corpora and second language acquisition research*, *Second Language Research*, SAGE Publications.

Nesselhauf, N., 2011. *Corpus Linguistics: A Practical Introduction*. https://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J., 1985. *A comprehensive grammar of the English language*, London: Longman.

Schleef, E., 2019. *A cross-cultural investigation of German and American academic style*, in Journal of Pragmatics, 41, pp. 1104–1124.

The International Corpus of Learner English (ICLE). uclouvain.be https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

Yang, W., Sun, Y., 2012. *The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency level,* Linguistics and Education, 23, pp. 31-48.

Corpora and software

AntConc 3.5.8 Laurence Anthony, 2019. Faculty of Science and Engineering, Wased University.

BAISA, Vít a Vít SUCHOMEL. SkELL – Web interface for English Language Learning, 2014. In Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, Brno: Tribun.

Davies, M., 2004. *British National Corpus*, Oxford University Press.

Davies, M., 2007. *TIME Magazine Corpus*.

Davies, M., 2019. *The Movie Corpus*.

Granger, S., Dagneaux, E., Meunier, F., Paquot, M., 2009. *The International Corpus of Learner English*, v2, Louvain la-Neuve: Presses universitaires de Louvain.

Schmied, J., Claridge, C., Siemund, R., 1999. *The Lampeter Corpus of Early Modern English Tracts*.

TagAnt version 1.2.0, 2014, Laurence Anthony.

Riassunto

La linguistica dei corpora è un indirizzo sempre più adottato nello studio delle lingue. Essa consiste nell'analisi di testi scritti o parlati, che vengono collezionati in *corpus* in base agli obiettivi della ricerca. L'obiettivo di questo studio è quello di introdurre il lettore all'argomento, dimostrarne l'importanza per l'evoluzione degli studi linguistici, a cornice dello studio oggetto della tesi: una analisi dell'uso di connettivi frasali utilizzati in testi accademici da studenti non madrelingua inglese.
La storia della linguistica dei corpora vede le sue origini negli ultimi decenni del 1800, quando, con l'impiego dei "diari parentali", si riportavano le espressioni utilizzate dai bambini, che venivano poi analizzate dai linguisti. Ad essi seguirono i primi studi normativi e comparativi, non solo sull'apprendimento del linguaggio, ma focalizzati anche su aspetti pedagogici e strutture grammaticali.

L'approccio empiristico, che vede nel concreto utilizzo della lingua il suo oggetto fondamentale, negli anni 50 del Novecento vede l'opposizione delle teorie del razionalista Noam Chomsky. Alla base della teoria razionalista c'è la distinzione tra *competenza* ed *esecuzione.* Il linguista americano e i suoi sostenitori affermano che la linguistica studi la *competenza,* ovvero le regole innate che un parlante possiede, mentre ritengono che l'*esecuzione*, ovvero l'effettivo utilizzo di quelle regole, non possa concorrere nelle ricerche linguistiche, in quanto influenzata da numerosi fattori, e di conseguenza non oggettiva. Al contrario, per gli empiristi l'esecuzione, e di conseguenza la creazione dei *corpora*, rappresentano una parte fondamentale dell'analisi linguistica: l'osservazione. I corpora, infatti, rendono possibile l'analisi della frequenza effettiva dell'uso di determinate strutture o lessemi.
L'invenzione dei computer diede un contributo sostanziale agli studi dei corpora. Nel 1967, padre Roberto Busa sviluppa il primo sistema di analisi linguistica assistito dall'utilizzo di un computer, avente oggetto una collezione di testi di San Tommaso d'Aquino. Ad esso si susseguirono software innovativi che tutt'oggi permettono analisi sempre più approfondite.

Il primo passo nell'analisi di un corpus è l'identificazione della disciplina linguistica che si vuole prendere in oggetto, tra esse le più ricorrenti sono: fraseologia, la lessico-grammatica, il registro, le varianti di una lingua, e la Lingua per scopi speciali (LSP).

Dopo aver ottenuto il consenso all'utilizzo dei testi, i corpora vengono creati. Essi possono essere classificati a seconda delle caratteristiche dei testi che vengono scelti per la loro composizione. Nesselhauf (2011) li classifica in:

- Corpora specialistici o generali
- Corpora sincronici o diacronici
- Corpora regionali
- Corpora di apprendenti o di madrelingua
- Corpora multilingue o monolingue
- Corpora di lingua scritta o parlata.

I corpora sono quindi strumenti oggettivi, autentici, verificabili e multifunzionali che hanno permesso, e tutt'oggi permettono, agli studiosi di svolgere analisi approfondite e specialistiche delle lingue, del loro uso e apprendimento. Gli strumenti utilizzati per le analisi permettono di individuare alcuni principali elementi della frase: la *collocazione* di parole o gruppi di parole, la *frequenza* del loro uso in determinati generi di testo o registri, e i *token* che aiutano ad individuarne la ricchezza lessicale. Attraverso i software informatici è inoltre possibile annotare i corpora da un punto di vista sintattico, morfologico, semantico e pragmatico, quindi individuare tramite una procedura computerizzata la categoria di appartenenza di ogni singolo lemma presente nei testi. Questo procedimento risulta fondamentale nell'analisi di testi di apprendenti, che spesso utilizzano elementi della frase in modo errato o non preciso.

Gli studi dell'Apprendimento della seconda lingua (in inglese SLA: Second Language Acquisition) hanno infatti a loro volta sviluppato metodi di analisi di corpora dettagliati, in modo da "comprendere i principi che governano il processo di apprendimento di una seconda lingua" (Granger:1998:4). Questi studi si svilupparono negli anni 60 e 70 del Novecento con il distinguersi di tre approcci: analisi contrastiva, analisi degli errori e teoria dell'interlingua.

L'analisi contrastiva consiste nell'analisi comparata di due o più lingue e ha l'obiettivo di individuare una possibile influenza della prima lingua (L1) sull'apprendimento della seconda (L2). L'analisi degli errori cerca le cause di errori frequenti negli apprendenti e, di conseguenza, i metodi per correggerli. Infine, l'interlingua è la teoria secondo la quale

l'apprendente che non padroneggia ancora a pieno la lingua che sta imparando adotta strutture soggettive affiancandole a strutture della sua lingua nativa.

Dopo aver scelto quale approccio utilizzare, è possibile quindi affrontare uno studio dei corpora di apprendenti. Essi si classificano in: corpora monolingue o multilingue, generici o specialistici\tecnici, paralleli o comparabili, scritti o parlati. Come già accennato, i corpora analizzano anche mirati campi della Lingua per Scopi Specifici (LSP), ed in questo caso dell'Inglese per Scopi Accademici (in inglese EAP: English for Academic Purposes). I testi collezionati sono testi scritti in ambito accademico da studenti di lingua inglese a diversi livelli, su svariati argomenti. L'analisi di questi corpora, generalmente contrastiva, non è importante solo per l'evolversi degli studi linguistici in sé, ma crea materiale di apprendimento per gli studenti stessi. Essi, infatti, hanno la possibilità di eseguire analisi su propri testi, o scritti da colleghi, e di conseguenza prendere consapevolezza di strutture alternative a quelle utilizzate, errori frequenti e possibili soluzioni.

Esemplari di questo approccio d'analisi sono due studi contrastivi sull'uso dei connettivi linguistici da parte di studenti di lingua inglese. I connettivi sono elementi grammaticali, avverbi, congiunzioni o altre parti del discorso, che hanno la funzione di unire due parti del discorso creando coesione e coerenza nel testo, essi possono essere categorizzati sulla base della funzione semantica che svolgono. Il loro utilizzo e apprendimento è influenzato da innumerevoli fattori, sia personali che di gruppo, come la lingua nativa, il contesto di apprendimento, le conoscenze pregresse di ogni studente, o le scelte retoriche nella composizione del testo.

Il primo studio, condotto da Altenberg e Tapper nel 1998, è uno studio contrastivo sull'uso dei connettivi da parte di studenti svedesi di inglese e madrelingua inglesi, che mette a confronto i dati dei due gruppi scelti dal punto di vista quantitativo, ad esempio la frequenza dell'uso di singoli connettivi in specifiche parti del testo. I testi analizzati appartengono al corpus ICLE (The International corpus of Learner English), che contiene sotto-corpora di testi scritti da apprendenti di inglese di 25 lingue native diverse. Il secondo studio, condotto da Carrió-Pastor nel 2013, analizza invece testi di studenti spagnoli e madrelingua inglese scelti dal corpus ICLE appena citato.

Infine, lo studio oggetto della tesi analizza l'uso dei connettivi in inglese in testi di studenti italiani e tedeschi collezionati nel corpus ICLE, con l'obiettivo di confermare o smentire l'ipotesi iniziale secondo la quale gli studenti italiani facciano un uso maggiore dei connettivi rispetto a quelli tedeschi. In seguito ad una analisi contrastiva di specifici connettori appartenenti a tre categorie (connettivi consecutivi, aggiuntivi e di opposizione), è emerso che il corpus tedesco mostra nella maggior parte dei casi un uso altamente inferiore dei connettivi rispetto a quello italiano. La categoria in cui si riscontra maggiore differenza è quella dei connettivi consecutivi, e tra essi il connettivo che dimostra maggiore divario è *therefore (perciò\ dunque)*, utilizzato 62 volte nel corpus italiano e 19 volte in quello tedesco. Lo stesso accade con i connettivi aggiuntivi, mentre i connettori contrastivi vengono utilizzati in modo più bilanciato.

Le ragioni di questo divario non emergono chiaramente dallo studio. Sarebbe necessario infatti condurre ulteriori ricerche, prendendo in oggetto una quantità superiore di connettivi e categorie per poter verificare le strutture e i lemmi preferiti dai due gruppi di studenti. Inoltre, sarebbe utile condurre un'analisi comparativa con l'uso dei connettivi nella lingua nativa per verificare eventuali similitudini semantiche e sintattiche.