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-11

Tesina di Laurea

Errors in Second Language Acquisition: A study based on the German and Spanish components of COREFL

Relatore
Prof. Erik Castello

Laureanda
Jasmine Giamberardino
n° matr.1225567 / LTLLM

Anno Accademico 2021 / 2022

ABSTRACT

This dissertation analyses a set of errors concerning the omission of the third person *-s* made by Spanish and German L2 learners of English in the COREFL learner corpus. With a corpus-based approach, the errors were analysed from a grammatical point of view. The errors were also sorted according to some variables, including the learners' proficiency level and the texts' medium. The data analyses were carried out on Microsoft Excel, to perform calculations and visualize the results by putting the data into tables and charts. The results were also normalised in order to correctly compare them.

The findings suggest that the Spanish learners make more errors than the German learners, but both groups produce the errors mostly in the spoken medium. The results are also discussed in relation to some theories and hypotheses coming from the field of Second Language Acquisition.

The dissertation concludes by presenting some possible causes for this error type and some possible solutions to overcome it.

TABLE OF CONTENTS

INTRODUCTION

The aim of this dissertation is to investigate if and to what extent Spanish and German learners of English as a second language make errors concerning the omission of the third person *-s* (e.g. *she sing, he laugh)*. In particular, the focus of the study is on the amount of errors present in a dataset representing the production of Spanish and German learners and on the presence of notable patterns, if any.

The investigation of this type of errors is especially interesting from a Second Language Acquisition perspective: various studies in this field (see for instance Dulay and Burt, 1973, 1974; Bailey et al., 1974) report that the third person *-s* is one of the last morphemes acquired by learners of English as a second language, similarly to what was found in the acquisition of English as a first language (Brown, 1973).

The data on which the study is based were extracted from COREFL, a learner corpus containing spoken and written texts produced in English by Spanish and German learners. COREFL also contains two native control sub-corpora of English and Spanish, respectively. The corpus was investigated using the COREFL's online interface, which offers a variety of tools to browse the corpus.

The data were analysed with Microsoft Excel and the results were visualised thanks to tables and charts. More specifically, the frequency counts and the normalised frequencies of the errors were taken into account and commented.

The first chapter of this dissertation provides an overview of Second Language Acquisition, the field of study that investigates how the acquisition of a second language works. In order to make the presentation of this topic easier to follow, specific terminology pertinent to the field of Second Language Acquisition will be defined, such as "L2 acquisition", "second language" and "foreign language". The key goals of the discipline will then be covered, along with the main approaches to Second Language Acquisition which contribute to its progress. Furthermore, the role of the first language in the acquisition process of a second language will also be discussed because of its relevance. A comparison of the acquisition of a first language and of a second language will be useful to show the main differences and similarities between the two processes. Moreover, part of the chapter will be devoted to a diachronic outlook on Second Language Acquisition, which will touch upon the key theories that shaped the history of

the discipline. The focus will be on the theories that derive from the field of linguistics, because they are more pertinent to the study treated in this dissertation. In particular, the chapter will be concluded by an in-depth presentation of three approaches to Second Language Acquisition which constitute the theoretical background to the study: Error Analysis, Morpheme Order Studies, and Learner Corpus Research.

The second chapter is of a more technical nature. First of all, it features a detailed description of the COREFL corpus based on the information present on the COREFL's official website, available at the address "corefl.learnercorpora.com". The design of the corpus will be illustrated, providing information about the composition of the corpus, the methods used by the researchers to elicit data from the participants, and the metadata present in the corpus. The features that constitute the COREFL's web interface are also thoroughly explained using screen captures of the website as visual aid. The main purpose of the web interface is to grant users the opportunity to navigate and download the corpus. Moreover, a small section of the chapter is dedicated to the illustration of the main statistics of the corpus. Finally, the second chapter also contains a description of the methodologies employed to extract the data from the corpus and analyse them on Microsoft Excel.

The focus of the third chapter is on the presentation of the frequency counts and the normalised frequencies of the errors collected from the corpus. The errors will be categorised according to a set of variables, namely the learners' first language, the verb type, the medium of the task, the proficiency level of the learners, and the self-corrections. The frequency counts and the normalised frequencies of the errors are visually represented thanks to tables and charts, the contents of which are explained. Finally, the findings of the analysis will be summarised and discussed at the end of the chapter, with further comments present in the Conclusions section.

CHAPTER 1
AN INTRODUCTION TO SECOND LANGUAGE ACQUISITION

This first chapter is meant to be a general overview of the field of Second Language Acquisition. It is divided into four sections: the first section covers some specific terminology and briefly presents the main goals of the discipline; the second section compares and contrasts first language acquisition and second language acquisition; the third section provides a diachronic outlook on Second Language Acquisition by describing the key theories that contributed to the development of the field; the last section concludes the chapter with a more thorough presentation of the three approaches to Second Language Acquisition research which influenced significantly the core investigation of this dissertation.

1.1 Defining Second Language Acquisition

Second Language Acquisition (SLA) can be defined as the "systematic study of how people acquire a second language" (Ellis, 1997, p. 3). However, the term "second language acquisition" is ambiguous, as it is also commonly used to identify the object of study of SLA (Ellis and Barkhuizen, 2005, p. 3). To avoid confusion, the disambiguation proposed in Ellis and Barkhuizen (2005) will be employed in this dissertation: the term "second language acquisition" (SLA) will be used to describe the field of study itself, while its object of enquiry will be referred to as "L2 acquisition".

In order to obtain a more thorough definition of SLA, it may be useful to look at some specific terminology. The term "second language" has multiple meanings in the field of SLA. If considered in a broader sense, a second language can indicate any language which is learned after the mother tongue. In other words, "[t]he additional language is called a second language (L2), even though it may actually be the third, fourth, or tenth to be acquired" (Saville-Troike, 2012, p. 2).

In a narrower sense, a second language is "an official or societally dominant language needed for education, employment, and other basic purposes. It is often acquired by minority group members or immigrants who speak another language natively" (Saville-Troike, 2012, p. 4). A second language, considered in this sense, is in contrast with a foreign language. According to the definition given by Saville-Troike

(2012, p. 4), a foreign language is usually learned in a classroom, it is scarcely used by the learners in their "immediate social context[s]", and has "no immediate or necessary practical application." On the other hand, as Mitchell et al. (2012, p. 1) and Ellis (2015) explain, foreign languages commonly fall within the aforementioned broader definition of second languages. In fact, the authors stress that the difference between the two types of language acquisition is merely contextual, while "the underlying learning processes are essentially the same" (Mitchell et al., 2012, p. 1).

SLA also takes into account all types of learning (Mitchell et al., 2012, p. 2): formal learning, which typically takes place in a classroom; informal learning, which happens in naturalistic contexts, e.g. through social interactions with native speakers of the second language, and learning that combines both formal and informal learning circumstances (Saville-Troike, 2012, p. 2).

After this brief definition of SLA, the goals of the discipline will now be discussed. Saville-Troike (2012) uses three questions which effectively summarise the main goals of SLA: "What exactly does the L2 learner come to know? How does the learner acquire this knowledge? Why are some learners more successful than others?" (Saville-Troike, 2012, p. 2). As the author stresses, these questions unfortunately do not come with easy answers. One of the main reasons behind this difficulty could be the strong interdisciplinarity of SLA: this field of study originated from both linguistics and psychology, including their subfields of applied linguistics, psycholinguistics, sociolinguistics, and social psychology (Saville-Troike, 2012, pp. 2–3). To this list, Gass (2013, p. 1) adds sociology, discourse analysis, conversation analysis, and education, plus more could be mentioned. On the bright side, the interdisciplinarity also grants the opportunity to gain various valuable perspectives on L2 acquisition, since researchers from each field focus on different aspects of the discipline.

Towell and Hawkins (1994, pp. 4–5) group these different approaches "into three broad categories: linguistic approaches, sociolinguistic approaches, and psychological or cognitive approaches." The authors explain that linguistic approaches "are of a single broad type": these approaches all start from the assumption that everyone is born with a "language faculty", which allows children to acquire their native language. The differences between L1 and L2 acquisition are due to "structural changes" which modify the language faculty over time. The sociolinguistic approaches focus on at least two

issues: one concerns the attitudes of L2 learners "towards the L2, the people who speak it, or the culture with which the language is associated", since those attitudes could influence the learner's motivation and "the nature of SLA itself"; the second one focuses on how the process of acquisition is affected by "the context in which the learner encounters or uses the L2." Towell and Hawkins (1994, pp. 4–5) then describe the psychological or cognitive approaches: said approaches deal with the difference in general cognitive maturity between L1 learners and L2 learners. While L1 learners "acquire knowledge of language and knowledge of the world simultaneously, […] L2 learners already know quite a lot about the world when they come to the task of SLA." This difference could play a part in why the language acquisition process varies between L1 and L2 learners. The authors add that psychological or cognitive approaches also deal with the mental devices which enable the comprehension, the storage, and the production of language, and "how this might be related to the way that L1 and L2 learners acquire particular languages."

1.2 First Language Acquisition vs. Second Language Acquisition

The role played by the first language in L2 acquisition is a very relevant one in the field of SLA. It is "one of the oldest and most continuously studied areas of SLA" (Gass, 2013, p. 79). Ellis (2015) reinforces Gass's statement by saying that research on first language acquisition had a strong impact on early work in SLA. He also adds that "[t]he role of the learners' native language in the acquisition of a second language, arguably, has received more sustained attention than any other area of SLA." In light of its importance, a brief section will now be dedicated to the aforesaid topic.

The processes involved in L1 acquisition are very much distinct from those which constitute L2 acquisition (Lightbown and Spada, 2013). This does not mean, however, that the two do not interact. For instance, when learners start acquiring an L2 they have necessarily completed the acquisition of their first language, or even multiple ones in the case of simultaneous multilingualism (Saville-Troike, 2012, p. 4). Therefore, the acquisition of a second language will inevitably be influenced by the learner's first language. This phenomenon is known as language transfer. As Saville-Troike (2012, p. 19) explains, transfer can happen on all levels of language and it is positive if "an L1 structure or rule is used in an L2 utterance and that use is appropriate or "correct" in the

L2", while there is negative transfer, also called "interference", if the use of L1 elements in an L2 utterance results inappropriate and incorrect in the L2. To these two types of transfer, Ellis (1997, pp. 51–52) adds two more: avoidance and overuse. Avoidance takes place when L2 learners do not employ a particular L2 element or structure while using the L2 because there is no equivalent in their L1; overuse occurs when L2 learners often resort to an L1 form which is not as frequent, or is even absent, in the L2. From this short overview of language transfer, it would appear that L1 influence has a mostly negative impact on L2 acquisition.

The main differences between L1 and L2 acquisition will now be briefly covered. Ellis (2015) argues that L2 acquisition is of a more complex nature than L1 acquisition, because the former entails additional factors. One of these is the influence of the L1 on L2 acquisition, which has already been presented. Another one lies in the contexts of acquisition, which are "much more varied than those of first language acquisition" (Ellis, 2015). As already mentioned in the previous subchapter, L2 acquisition can take place in formal and informal contexts, also known as instructed and naturalistic contexts (Saville-Troike, 2012). This variety of contexts does not, however, provide L2 learners with a learning environment as rich as the one available to L1 learners: while babies are acquiring their native language, there is normally at least one adult always taking care of their needs, thus babies "can devote unlimited amounts of time to language learning" (Towell and Hawkins, 1994, p. 155). An important aspect considered by Ellis (2015) and Lightbown and Spada (2013, p. 38) is that many learners have the opportunity to learn an L2 solely in the formal context of a classroom. This means that not only L2 learners generally have far less time to dedicate to language learning compared to small children acquiring their L1, but they also "range from simple exposure to the language in the country with no explicit instruction, to a total lack of contact with the language in the country where it is spoken and a complete reliance on instruction in a classroom setting" (Towell and Hawkins, 1994, p. 155).

Another important factor is age: a second language can be learned at any age after the first language has been acquired, while children always acquire their L1 in their first years of life. Plus "[t]here is a cut-off age for L1 acquisition, beyond which it can never be complete" (Saville-Troike, 2012, p. 14). In other words, native-speaker competence cannot be achieved after the "cut-off age" has passed. The "critical period hypothesis"

investigates this phenomenon. This hypothesis has provided substantial evidence which reinforces a common claim in the field of SLA: if an individual starts learning an L2 during adulthood, they have very low chance of attaining the same linguistic competence as a native speaker "in either grammar or pronunciation" (Ellis, 1997, p. 68).

Learning strategies constitute another divergence between the two types of language acquisition. Learning strategies can be defined as "the behaviors and techniques [L2 learners] adopt in their efforts to learn a second language" (Saville-Troike, 2012, p. 97), and they "account for how learners acquire and automatize L2 knowledge" (Ellis, 2015). L2 learners can make use of learning strategies, while they are unavailable to L1 learners. This is due to the former group's higher cognitive maturity (Ellis, 2015).

1.3 History of the discipline "Second Language Acquisition"

The most important theories which contributed to the development of SLA will now be briefly presented. Special emphasis will be placed on theories stemming from the field of linguistics.

The field of SLA is quite recent: the first theories concerning the crucial questions of SLA were formulated in the 1960s. Before that period, L2 acquisition was investigated with the main purpose of assisting language teaching (Mitchell et al., 2012, p. 28; Saville-Troike, 2012, p. 25). During the 1950s, Structuralism was the predominant linguistic model. In particular, language pedagogy relied on a version of Structuralism created by Palmer in the 1920s, and further developed in the 1940s by Fries and his Michigan colleagues (Mitchell et al., 2012, p. 28). This model, as corroborated by the summary made by Howatt (2004, pp. 299–300, in Mitchell et al., 2012), confirms how Behaviourism was "[t]he most influential cognitive model of learning that was applied to language acquisition at that time" (Saville-Troike, 2012, p. 26). According to Behaviourism, language is a skill to be acquired, just like any other. More specifically, skills are acquired through habit formation, which is possible thanks to the repetition of S-R-R sequences (stimuli, responses, reinforcement): taking language learning as an example, linguistic input constitutes the stimuli, the individual responds to these stimuli and if the responses produce the expected results, these are reinforced (Saville-Troike, 2012, p. 26). If this theory is applied to L2 acquisition, a predicted outcome is that "the

habits of the L1 interfere with the development of L2 habits" (Ellis, 2015). In other words, if language is considered as habit, L2 learning would basically consist in acquiring more habits. This last statement, together with Structuralism and Behaviourism more in general, provided the basis for the Contrastive Analysis Hypothesis by Lado (1957, in Ellis, 2015).

The main goal of Contrastive Analysis (CA) is to predict which obstacles, and which facilitations, learners would encounter when trying to acquire a second language. This is achieved by comparing the learner's native language with the target language, i.e. the L2. This method thus attributes a major role to transfer phenomena, as Gass (2013, p. 86) also suggests: "[i]t is assumed that learners tend to transfer the habits of their native language structure to the foreign language and it is also assumed that this is the major source of difficulty or ease in learning the structure of a foreign language."

In the 1960s, some shortcomings of CA started to emerge. On several occasions, the analyses proposed by CA did not find a match in actual learner linguistic production: many predictions about errors or areas of positive transfer could not be proved by evidence (Gass, 2013; Saville-Troike, 2012; Towell and Hawkins, 1994). This would lead one to believe that L1 interference cannot be held responsible for all learner errors. This is not, however, the only reason why CA was criticized by researchers. In fact, the behaviourist theory, upon which CA was built, was also being challenged. The main criticism came from the American linguist Noam Chomsky. He brought about great change in the field of linguistics with his Transformational-Generative Grammar (1957, 1965, in Saville-Troike, 2012), in which he cogently argues that the behaviourist theory applied to language acquisition is inappropriate, because it does not account for the creative capacity which characterizes language. The creative capacity of language is at the core of the "logical problem of language acquisition": Saville-Troike (2012, p. 201) effectively defines it as "[t]he question of how children achieve the final state of L1 development with ease and success when the linguistic system is very complex and their cognitive ability is not fully developed." Plus, the language input they are exposed to is not perfect: it inevitably presents some ungrammatical forms since all language users happen to make mistakes. According to Chomsky, the logical problem of language acquisition can be explained if it is assumed that there are some innate principles which characterize all languages (Gass, 2013; Lightbown and Spada, 2013). These innate

principles were then reformulated by Chomsky under the theory of Universal Grammar (UG), which had a revolutionary impact on linguistics. First and second language acquisition studies also benefitted greatly from it.

The revolution triggered by Chomsky undoubtedly influenced a "shift in perspective on the part of researchers in the late 1960s and early 1970s from a primary interest in transfer […] to a primary interest in staged development and cross-learner systematicity" (Towell and Hawkins, 1994, p. 23). The papers by Corder (1967) and Selinker (1972) are probably some of the most notable instances of this shift in perspective.

The paper that Corder published in 1967 played a major part in the establishment of a new approach in SLA, Error Analysis, which will be covered in greater detail in section 1.4.1 of this dissertation. According to Corder (1967), learners' errors are a manifestation of the system of language, and the strategies, that the learner is using at any specific stage of his L2 acquisition process. As Saville-Troike (2012, p. 41) interprets it, "errors are windows into the language learner's mind." Thus, researchers started analysing learners' errors in order to gain some insight on the processes which take place during L2 acquisition.

Corder's ideas, Error Analysis, and the innatist perspective on language formed the environment in which Selinker created the notion of "interlanguage" in 1972. It may be useful to mention interlanguage, because it is a term still used nowadays in the field of SLA (see for instance Mitchell et al., 2012; Saville-Troike, 2012). Interlanguage is defined by Selinker (1972) as "a separate linguistic system based on the observable output which results from a learner's attempted production of a TL [target language] norm." An interlanguage is shaped by the learner's L1 and by his target language, but is different from them, which makes it "a unique linguistic system" (Ellis, 1997, p. 33). Interlanguages are both systematic and dynamic: they are made of the learner's internal grammar with its own rules, but this grammar is subject to change over time as the learner is exposed to more linguistic input (Lightbown and Spada, 2013). Most learners' interlanguages stop developing before they achieve native-like proficiency in their target language, even though they keep receiving L2 linguistic input (Saville-Troike, 2012, p. 44). This phenomenon has been defined by Selinker as "fossilization". Mitchell et al. (2012, p. 36) stress how "[i]nterlanguage studies thus moved a major step beyond Error

Analysis, by focusing on the learner system as a whole, rather than only on its non-target-like features."

In those years, Chomsky's "revolution" and the approaches just described, led many researchers to investigate L1 acquisition in children. One of the most relevant studies is the one by Roger Brown (1973, in Mitchell et al., 2012). He conducted an important longitudinal study, known as the "morpheme study", on three children with different backgrounds. He focused on 14 English morphemes and compared their development: the findings showed that the children acquired the morphemes in approximately the same order, thus suggesting that there might be a common order of L1 acquisition. The results of this study, and others of similar nature, led SLA researchers to conduct analogous investigations on L2 learners, both children and adults. Of great importance are the Morpheme Order Studies, first conducted by Dulay and Burt in the early 1970s, which will be covered in detail in section 1.4.2 of this dissertation. Dulay and Burt (1973, 1974) took the results obtained from Brown's study in child L1 acquisition and applied them to child L2 acquisition. They compared the order of acquisition of a group of English morphemes on children who were L1 speakers of Spanish in the 1973 study. They expanded this investigation in the 1974 study by administering the same test to both L1 Spanish and Chinese children. The results showed a remarkably similar order of acquisition in both Spanish and Chinese children, which was also not too different from the L1 order obtained from Brown's study (Saville-Troike, 2012, p. 46). These findings led Dulay and Burt to hypothesize that "universal cognitive mechanisms are the basis for the child's organization of a target language, and that it is the L2 system, rather than the L1 system that guides the acquisition process" (Dulay and Burt, 1974, p. 52). In other words, according to Dulay and Burt's hypothesis, L1 transfer cannot be responsible for the L2 acquisition process. Instead, there may be some underlying "universal language processing strategies" (Dulay and Burt, 1974, p. 52) that govern L2 acquisition.

During the 1970s, the Morpheme Order Studies were followed by the Monitor Model (Krashen, 1978), a theory which has been criticized and mostly discredited, but nonetheless had a significant impact on SLA (Gass, 2013; Lightbown and Spada, 2013; Saville-Troike, 2012). Krashen's Monitor Model is composed of five hypotheses: the acquisition/learning hypothesis, according to which acquisition is unconscious while

learning is a conscious act; the monitor hypothesis, which states that "acquired" knowledge is used by L2 learners during spontaneous communication, while "learned" rules can only be used as a "monitor" to modify the production of the "acquired system"; the natural order hypothesis, which states that L2 rules are acquired in a predictable sequence, similarly to what happens in L1 acquisition; the input hypothesis, according to which comprehensible input makes language acquisition possible; the affective filter hypothesis, which supports the existence of a metaphorical barrier in learners' minds, the "affective filter", that when is active inhibits the processing of comprehensible input, meaning that language acquisition may not take place (Lightbown and Spada, 2013; Saville-Troike, 2012). Krashen's model is mainly remembered for its strong influence on language teaching, particularly in the USA in the 1980s and 1990s (Saville-Troike, 2012, p. 48).

Many other approaches followed the ones described in this section. The summary made by Saville-Troike (2012) will be used as a reference point to briefly present the other main models and theories applied to SLA.

In the field of linguistics, Chomsky kept working on and developing his Universal Grammar theory: in the 1980s, he implemented to UG the Principles and Parameters model, which defines principles as universal and common to all languages, while parameters are subjective to each language. The Minimalist Program was added to UG in the 1990s and gives the lexicon a prominent role. In the 2000s, it was followed by the Interface Hypothesis, which focused on linguistic interfaces and was not formulated by Chomsky. While all these approaches have an internal focus on SLA, the main frameworks that have instead an external focus belong to Functionalism. These frameworks are different from the others because they lay stress on the communicative function of language, thus looking at the various languages from an external perspective. Just like the approaches with an internal focus, those with an external focus have been very influential on SLA. To be more specific, the "[a]pproaches based on functional frameworks have dominated European study of SLA and are widely followed elsewhere in the world" (Saville-Troike, 2012, p. 27).

As mentioned before, also the fields of psychology and sociology contributed to the development of SLA. The summary by Saville-Troike (2012) will still be followed to briefly illustrate the main contributions of these two disciplines to SLA. Psychology's

main contributions came at first with neurolinguistics and Information Processing in the 1960s. The former focuses on the relationship between languages and the brain, while the latter focuses on learning processes. Processability also focuses on learning processes: it expands the concepts by Information Processing and applies them to L2 teaching. Another framework with the same focus is Connectionism, which was formulated in the 1980s and acquired importance with time. It does not start from the assumption that innate knowledge is involved in language acquisition, unlike most frameworks in SLA, but states that language learning arises from the strengthening of associations in the brain resulting from repeated exposure to determined inputs. Psychology is also interested in learner differences: the humanistic models within psychology which focus on this area started considerably influencing SLA research and second language teaching in the 1970s.

As for the social perspective, all of its frameworks "emphasize the importance of social context for language acquisition and use" (Saville-Troike, 2012, p. 28). The contributions to SLA research from a social perspective fall within the microsocial and macrosocial foci. Within the microsocial focus, one of the first approaches was Vygotsky's Sociocultural Theory, which is of great importance. It views "cognitive development, including language development, […] as a result of social interactions" (Lightbown and Spada, 2013, p. 118). In the 1990s, it also played a major role in the revival of the Interactionist approaches. Instead, in the 1960s-70s, the main approaches were those provided by the Variation Theory and the Accommodation Theory. In short, they focus on the different outcomes in linguistic learner production that are due to different contexts of use, and the possible variations in L2 acquisition which may arise even among learners of the same L2 (Saville-Troike, 2012, p. 29). The approach that flourished in the 2000s is the Computer Mediated Communication (CMC), and it is one of the first examples of computers used to improve and investigate L2 learning and teaching.

Research within the macrosocial focus mainly deals with L2 learning and use applied to different ecological contexts, namely educational, political, or cultural ones. One of the first frameworks to account for the importance of social and cultural knowledge in language use is the Ethnography of Communication framework. Formulated in the 1960s, this framework also considers language learners as part of a

community “with sociopolitical as well as linguistic bounds” (Saville-Troike, 2012, p. 29). In the 1970s and 1980s, it was followed by the frameworks within Acculturation Theory and Social Psychology. They are mostly concerned with how the results of the L2 acquisition process are altered by “such factors as identity, status, and values” (Saville-Troike, 2012, p. 29).

1.4 Some applications of Second Language Acquisition

In the following sections, three approaches to SLA research will be considered in more detail: Error Analysis, Morpheme Order Studies, and Learner Corpus Research. This choice was dictated by the influence that these approaches have on the research at the core of this dissertation.

1.4.1 Error Analysis

Interest in errors produced by language learners dates back way before the birth of Error Analysis (EA) in the late 1960s. For instance, Ellis and Barkhuizen (2005) report that native speakers’ errors were collected and studied already back in the 18th century by prescriptive grammarians, and this approach is still used nowadays. The goal was, and still is, to present what linguistic forms should be avoided and which ones should be used. As far as second language pedagogy is concerned, some dictionaries describing common learners’ errors started circulating in the 1930s and keep being produced to this day (Ellis and Barkhuizen, 2005, pp. 51–52).

Error Analysis is, however, the first approach that considers errors as more than “annoying, distracting, but inevitable by-products of the process of learning a language” (Corder, 1967). As already mentioned, EA was established mostly thanks to Corder’s famous paper “The Significance of Learners’ Errors”, published in 1967. In his paper, Corder bases his theses on the then-new Innatist hypothesis. Thus, he shares the idea that L2 learners, just like children acquiring an L1, must possess some kind of innate language mechanisms, whose functioning is however largely unknown, that make language acquisition possible. About errors, he states that they must be analysed, not discarded, since they provide evidence that the process of language acquisition is indeed taking place. Errors are also proof of “the system of the language that he [the learner] is using (i.e. has learned) at a particular point in the course” (Corder, 1967), which is, in

other words, what will be commonly known as an interlanguage. On the other hand, the production of a correct form on the part of the learner can be quite misleading according to Corder. This is because when the learner produces correct forms, it should not be taken for granted that said forms have actually been acquired: it may very well be that the learner is simply imitating some utterances they have previously heard.

Corder further stresses the relevance of errors by stating that they are important in three different ways: first, if they are systematically analysed, they can tell teachers how much progress the learner has made; second, researchers can get some insight from errors about learners' language learning strategies and the way language is being acquired; thirdly, errors are essential for learners, since errors can be considered as a tool that learners employ to test their "hypotheses about the nature of the language [they are] learning" (Corder, 1967). Corder also differentiates between "errors of performance" and "errors of competence", commonly known as "mistakes" and "errors". Mistakes are unsystematic, they are analogous to slips of the tongue. When an individual makes a mistake, they are generally able to correct it without much effort, which proves that mistakes do not indicate lack of knowledge. On the other hand, errors are systematic and usually point towards a gap in someone's linguistic competence.

When looking at learner language in the context of EA, only errors are taken into consideration, since mistakes do not offer any insight on the learner's interlanguage. In Corder's (1967) words, "[m]istakes are of no significance to the process of language learning", even though Ellis and Barkhuizen (2005) argue that mistakes are actually as important as errors both from a practical and a theoretical point of view.

In the late 1960s, EA thus became "the primary means of conducting research into L2 acquisition" (Ellis and Barkhuizen, 2005, p. 52). An Error Analysis is usually conducted according to the following steps (Corder, 1974, in Ellis and Barkhuizen, 2005):

1. Collection of a sample of learner language
2. Identification of errors
3. Description of errors
4. Explanation of errors
5. Error evaluation

Identifying errors can be quite challenging: first of all, distinguishing between errors and mistakes is not always a straightforward task, thus it often requires a deeper investigation (Corder, 1967; Ellis, 1997). Another difficulty can be found within the process of identification itself, which consists of a comparison between the collected learner's language sample and a reconstruction of what the sample would be like if a native speaker produced it (Ellis and Barkhuizen, 2005). The main difficulty lies precisely in the reconstruction of the sample, as the original sample may contain some ambiguous constructions difficult to interpret.

According to Ellis (1997), errors can then be described in various ways, for instance they could be classified into grammatical categories, or it may also be possible to describe how the errors differ from their reconstructed counterparts. Within these general error types there is "omission", which is the lack of an item necessary to make an utterance grammatical; "misinformation", which is the erroneous use of a form instead of the correct one, and "misordering", which refers to the wrong placement of words in an utterance.

Explaining errors can also prove to be a challenging task, as the goal is essentially to understand why errors occur. Many errors are systematic, even predictable. Many are also universal, meaning that they are found in the linguistic production of numerous learners, while others depend on the learner's L1 (Ellis, 1997). About the errors' sources, Saville-Troike (2012, p. 42) states that the main causes of L2 errors are "interlingual factors", such as negative transfer, and "intralingual factors", also called developmental errors. An example of intralingual/developmental errors is overgeneralization, of which Richards (1971) lists some examples: "he can sings, we are hope, it is occurs, he come from".

Finally, errors are evaluated by determining their gravity: errors that have little impact on the global comprehension of the utterance will be judged as less serious than errors which lead to misinterpretation (Saville-Troike, 2012, p. 42).

Nowadays, Error Analysis is still used to investigate L2 acquisition, but it is not the preferred approach as it presents some inadequacies. For example, Ellis and Barkhuizen (2005, p. 70) remark that EA does not provide a complete account of learner language, since errors are the only part taken into consideration. They also note the aforementioned methodological problems regarding the identification, description and

explanation of errors, which cannot be overlooked. Another important limitation of EA, mentioned by various researchers (see for instance Ellis and Barkhuizen, 2005; Gass, 2013; Saville-Troike, 2012), concerns learners' avoidance of some L2 structures: the lack of errors, especially regarding difficult constructions, may be caused by the avoidance of said constructions on the part of the learner. EA does not have the capacity to account for this possibility. This phenomenon is corroborated by the findings of a famous study by Shachter (1974, in Saville-Troike, 2012), which indicate that Chinese and Japanese L2 learners of English make few errors with relative clauses simply because they avoid using them.

Despite all its limitations, nowadays EA is the key component of Computer-aided Error Analysis (CEA), one of the most used methodologies in the subfield of Learner Corpus Research (e.g. Díez-Bedmar, 2021, p. 90). Very briefly, CEA is the error analysis of learner language carried out using learner corpora.

1.4.2 Morpheme Order Studies

The Morpheme Order Studies have been briefly presented in a previous section of this chapter. To recapitulate, they were first conducted by Dulay and Burt following Brown's footsteps in the early 1970s on children acquiring English as a second language. Regarding their objective, the Morpheme Order Studies "were directed at establishing whether there was a universal *order of acquisition* […] and, also, whether this order was the same or different from that found for L1 acquisition" (Ellis and Barkhuizen, 2005, pp. 73–74).

As previously mentioned, the main studies by Dulay and Burt were conducted in 1973 and 1974. In both instances, the test employed to carry out the investigations was the Bilingual Syntax Measure (BSM). The BSM consists of seven cartoon drawings and 33 related questions that the subjects have to answer. The questions are structured in such a way to ideally "elicit natural speech from children, not specific responses" (Dulay and Burt, 1973, p. 248). In particular, the questions were tailored to elicit some of the English morphemes used in Brown's study. In the 1973 study, the BSM was administered to three groups of Spanish-speaking children coming from three different areas of the United States. They were five to eight years old and were all learning English

as an L2. The results of the study showed a similar order of acquisition in all three groups of children, in spite of having different backgrounds.

Dulay and Burt (1974) decided to repeat the test on L1 Chinese and Spanish children who were L2 learners of English in order to confirm that the previous result was not "an artifact of the L1 Spanish language background of the subjects" (Towell and Hawkins, 1994, p. 25). The findings were encouraging: the sequences of acquisition in Chinese and Spanish children were, in Dulay and Burt's (1974) words, "virtually the same."

When comparing the L2 order of acquisition with the L1 order obtained by Brown in 1973, differences and similarities emerge. As noted by Ellis and Barkhuizen (2005, p. 75), the morpheme which is used more accurately in both sequences of acquisition is progressive *-ing*, while 3rd person *-s* is among the ones used with less accuracy. Regarding the differences, they include the articles, the copula, the auxiliary, and the irregular past tense: the first three are used more accurately by L2 learners, while the opposite stands for the irregular past tense. Either way, the findings of these studies reinforced the innatist hypothesis. Saville-Troike (2012, p. 47) reformulates Dulay and Burt's (1973) conclusions as follows:

> L2 learners are neither merely imitating what they hear nor necessarily transferring L1 structures to the new code, but (subconsciously) creating a mental grammar which allows them to interpret and produce utterances they have not heard before.

Similar studies were also conducted on adult L2 learners of English to check whether there was going to be a different order of acquisition. The most known is the one conducted by Bailey et al. (1974, in Towell and Hawkins, 1994) using the BSM on adult L2 learners of English from different L1 backgrounds. The results revealed once again a very similar order of acquisition. All these findings seemed to point towards the existence of a natural order of acquisition for grammatical morphology in both children and adults L2 learners of the same language, but with different L1 backgrounds and learning conditions (Towell and Hawkins, 1994, p. 25). However, other studies provided evidence that conflicted with the natural order of acquisition previously found. Ellis (2015) presents some of these studies, the findings of which generally show that the natural order of acquisition changes if there are grammatical morphemes that bear

meaning in the L2 but not in the L1. On the other hand, these same studies confirmed that the order does not vary significantly in the case of morphemes that do not carry meaning, like the English 3rd person singular, which are apparently equal in difficulty for L2 learners.

In spite of their promising results, the Morpheme Order Studies still received some criticism. Mitchell et al. (2012, p. 40) report that the criticisms mainly concern the Bilingual Syntax Measure, which allegedly biases the results, and the presupposed correlation between accuracy of production and acquisition order.

1.4.3 Learner Corpus Research

Learner Corpus Research (LCR) can be defined as a subfield of both SLA and Corpus Linguistics, as its main focus is on the investigation of learner language through the use of learner corpora. Meunier (2021, p. 23) quotes McEnery et al. (2006, p. 5) and provides a definition of a corpus. A corpus is a "collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety." By extension, a learner corpus shares the same characteristics, but contains samples of language produced by L2 learners. The corpus-based approach evidently has its advantages. According to Gass (2013, p. 41), two of the main advantages are tied to the opportunity of analysing the corpora by computer, which clearly automatizes and simplifies a good portion of the process, and to the size of learner corpora, which is way larger than other types of non-corpus samples, especially allowing for better statistical analyses. Since its genesis in the 1980s, LCR's studies results have proved to be particularly useful in informing SLA research and providing "useful input for applied projects", like the positive influence on teaching approaches, the creation of teaching materials, or the development of Natural Language Processing tools (Meunier, 2021, p. 23).

The ideal learner corpus would be based on samples taken from spontaneous L2 learners' linguistic production, what Ellis and Barkhuizen (2005, p. 23) refer to as "naturally occurring language use." Unfortunately, there are some major constraints which make this data collection task quite a challenge. To compensate, learner corpora are usually built upon the results of pedagogic tasks administered to the learners, as specified by Meunier (2021, p. 23). Meunier adds that the learner language samples that

constitute learner corpora are normally sorted according to several variables: learners' characteristics, such as L1, L2, proficiency level; task type, for instance if it is a written or oral production, formal or informal; task setting, like task interactivity. These variables are quite helpful because they are generally annotated on the samples as metadata, which are fundamental to organize and sort the data (Meunier, 2021, p. 23). These, however, are not the only annotations that can be added to the samples. For instance, through part-of-speech tagging all elements of the corpus are labelled with their corresponding part-of-speech tag. This task is automatically done thanks to existing processing tools. In addition to tools that allow automatic part-of-speech tagging, there are many others dedicated to corpus analysis. Some features they offer include the extraction of word lists or keywords, the comparison between sub-corpora, or the display of chosen elements such as words or tags (Meunier, 2021, p. 23).

A paramount asset of LCR is the possibility of comparing different language varieties. Granger (2015, p. 8, in Meunier, 2021, p. 26) states that two types of comparison have proved to be particularly prolific: a comparison of a learner language with a native language and a comparison between two interlanguages, especially coming from learners with a different native language. An example of the latter, presented in Meunier (2021, p. 27), could be the comparison between L1 French and L1 German speakers' production of L2 English, in acronym E2F vs E2G. In 1996, Granger theorised this approach, thus creating Contrastive Interlanguage Analysis (CIA). Callies (2015, p. 39) refers to CIA as "probably the most widely used methodological approach in LCR." However, he also presents a problem concerning this approach which has attracted much criticism. The issue is related to the comparison between an L2 corpus and a native corpus, also known as a control corpus: there are many native language varieties and choosing one over the other is interpreted by many as "the recognition of one idealized native speaker norm" (Meunier, 2021, p. 27). In 2015, Granger proposed CIA^2, a revised version of CIA which attempted to clear the issue by adding new terms.

There are many ways in which a corpus can be researched. According to Callies (2015, p. 35), "the choice of method(s) depends on the object(s) of study and the research question(s) being asked, and in turn, findings and results are highly dependent on the method(s) or database(s) chosen." Callies (2015) mentions three basic approaches: corpus-informed, corpus-based, and corpus-driven. Researchers employing the corpus-

informed approach use the learner corpus as a "general reference source for information" (p. 36), without working directly with the corpus data. In the corpus-based approach, the corpus is used as a data source to test existing hypotheses. In the corpus-driven approach, researchers instead formulate hypotheses based on the corpus data. The data can then be analysed in a quantitative or qualitative way: the former analysis is "primarily deductive, product-oriented and designed to test a specific hypothesis, which can then be confirmed or rejected, or refined and re-tested" (Callies, 2015, p. 36), plus quantitative data is generally the object of statistical analysis; the latter analysis is instead mainly based on generating hypotheses, it focuses on the observation and explanation of linguistic phenomena in naturalistic settings (i.e. "within naturally occurring social and cultural settings" (Callies, 2015, p. 37).

Furthermore, Granger (2012) explains that learner corpora can vary according to various dimensions: "time of collection, scope of collection, targeted language (L2), learner's mother tongue (L1), medium, and text type" (p. 11). The time of collection will now be briefly explained. A corpus is called "cross-sectional" if the language samples were collected from many learners "at a single point in time" (p. 11). It is called "longitudinal" when it contains samples taken from the same learners during a longer period of time, which can be challenging. Finally, some corpora are described as "quasi-longitudinal" as they are constituted by samples taken from learners at different proficiency levels, but at the same point in time.

There can be numerous kinds of learner corpora with different characteristics. However, as argued by Callies (2015, p. 38), there is a typical corpus configuration which dominates: "[a]rguably, the general methodology and procedure employed in LCR to date has mostly been corpus-based, quantitative, cross-sectional and comparative."

This chapter was intended to provide a general introduction to the field of Second Language Acquisition. Some theories and approaches have been covered in greater detail because they are relevant to the main study depicted in this dissertation. In chapter two, the learner corpus that was used as data source for this study, the COREFL corpus, will be thoroughly described. The chapter will also delineate the methodologies employed to perform both the data extraction and the data analysis.

CHAPTER 2
THE COREFL CORPUS: INVESTIGATION AND DATA ANALYSIS METHODOLOGIES

This second chapter is devoted to an in-depth description of the COREFL corpus and of the methodologies employed to investigate it. In the first section, a detailed account of COREFL will be given, i.e. the learner corpus used as the main data source for the investigation discussed in chapter three. The core parts of COREFL that will be covered are its design, its web interface, which allows for the navigation of COREFL's various sub-corpora, and its main statistics. The second section will present the methodologies used to isolate, and then extract, the corpus data on which the main investigation of this dissertation was performed. Finally, the last section will cover how the data were analysed using Microsoft Excel.

2.1 The COREFL corpus

The Corpus of English as a Foreign Language COREFL, is a learner corpus directed, designed and compiled by Cristóbal Lozano and Ana Díaz-Negrillo from Universidad de Granada, and Marcus Callies from Universität Bremen (Lozano et al., 2020). It is based on the same principles as the CEDEL2 corpus (Corpus Escrito del Español como L2) (Lozano, 2022), an L2 Spanish written corpus. The COREFL corpus can be consulted online at the following address: corefl.learnercorpora.com. This website not only grants free access to the corpus, but also provides an internal search engine that allows for a rather deep analysis of the corpus, without needing to resort to external software tools. However, there is also the possibility to download the corpus, or even just specific parts of it, thus allowing for the analysis of the data with other software. By navigating the website, users can also find various information about COREFL: the composition of the corpus, how to investigate it through the internal search engine, an in-depth presentation of its statistics, and much more. Currently, COREFL is available in its first version, but its creators are looking to expand it with a second version, which will include more data from both L2 learners and native speakers.

In the following sections, the characteristics of COREFL will be explained in detail. All of the information that will be presented can be found on COREFL's website.

2.1.1 Corpus design

COREFL is a quasi-longitudinal corpus mainly composed of samples of the language produced by Spanish and German L2 learners of English. However, COREFL also provides two native control sub-corpora: one includes data from different varieties of L1 English, mostly American and British; the other control corpus presents data collected from L1 Spanish, more specifically the Spanish and Latin American varieties. In the future, COREFL will be enriched with a third control sub-corpus containing language samples collected from native speakers of German.

COREFL contains both written and spoken texts, even though the latter constitute approximately only one third of the corpus. About the spoken texts, it is important to specify that each one can be paired with a written text. This is because some participants, at least 15 days after producing the written text, also produced a spoken text based on the same task. It is a significant feature because it allows researchers to conduct investigations focusing on the medium (i.e. spoken or written) without changing the participant and the task.

In order to collect the data, the participants were asked to carry out one task out of four. These four tasks were chosen out of the 14 used in the CEDEL2 corpus. The tasks used in the COREFL corpus are listed in the table below (Table 1). They correspond respectively to tasks number 2, 3, 13, and 14 from the list of CEDEL2 tasks.

Task title	**Task description**
Famous Person	Talk about a famous person. *Habla de una persona famosa.*
Film	Summarise a film you have seen recently. *Resume una película que has visto recientemente.*
Frog	Tell the story shown in the pictures. You can add new aspects to the story or ignore some aspects in the pictures. Your text should start "Un día / One day..." https://goo.gl/so3S6W
Chaplin	Watch the following Chaplin video clip (4 minutes). Summarise the story. You can watch the video clip more than once. https://www.youtube.com/watch?v=4QkTNJFhu-g

Table 1: COREFL tasks. Adapted from the COREFL website (corefl.learnercorpora.com/user_guide/corpus_design)

The researchers in charge of the data collection gathered information about a large number of variables from the participants. The variables are about the participant's linguistic background and the task itself. Also, the variables of learners are slightly different from those of native speakers and in a larger number. Some examples of linguistic background variables common to both learners and native speakers include the L1 of the participant, of the participant's father and mother, language(s) spoken at home, sex, age, educational institution, and a few others. Learners' linguistic background variables also include information about the age of exposure to L2 English, the years of study of English and details about stays in English-speaking countries. The task variables are less numerous. They are almost the same for both groups of participants: the learners have an extra variable regarding the place where they performed the task. The variables they have in common are the task title, the task text, the time in minutes spent to carry out the task, and the resources used to produce the task (e.g. dictionaries, grammar books and the like). The large number of variables available can be of great use to SLA researchers. For instance, all the texts in the corpus contain metadata based on the aforementioned variables. The files can be downloaded with or without the corresponding metadata attached. Besides having metadata, the corpus is also POS tagged: to each word of the corpus a tag was added (i.e. a label) that identifies its linguistic category, also known as part of speech (POS). POS tagging is quite advantageous, because it allows for advanced searches based on word category. However, it must be stressed that the tagging was done automatically by a POS tagger tool, meaning that any errors in learners' production are likely to be mislabelled.

Another interesting feature of COREFL is the assessment of the learners' proficiency level. In fact, COREFL used three proficiency-level measurements. The first is an objective measurement, which allowed for the classification of learners in the six levels proposed by the Common European Framework of Reference for Languages (CEFR), through the administration of a standardised placement test. The six levels are A1 (lower beginner), A2 (upper beginner), B1 (lower intermediate), B2 (upper intermediate), C1 (lower advanced), and C2 (upper advanced). The second is a subjective measurement, which consists in a self-assessment by the learners. The learners had to rate their own proficiency levels from A1 to C2 in the areas of writing, speaking, reading, and listening in English. The resulting four levels are then converted

into a one-to-six numeric scale and averaged. To put it in simpler terms, the A1 level corresponds to one, A2 to two and so forth until C2 to six. The average is thus a number from one to six, which creates a new variable. The third proficiency-level measurement is linked to the learners' possession of any English language certificate.

The data collection processes will now be briefly touched upon. The written data were collected using online forms. The forms were also written in the learners' native language, so as to make sure they would understand the task. However, some lower-level learners handwrote their written text, which then had to be reproduced in digital format. To distinguish the two types of data collection, the files in the corpus are assigned a specific label: "WRITING/AUDIO DETAILS: written_online" for the collection through online forms, and "WRITING/AUDIO DETAILS: written_offline_classroom" for the other type. Also spoken texts have undergone the same labelling procedure.

The spoken data were collected in four different ways. A limited number of audios were recorded by the participants themselves and then uploaded to the corpus database. These files are labelled as "WRITING/AUDIO DETAILS: spoken_online". The other three collection methods all took place at the Universidad de Granada. The majority of the audios were recorded with special equipment in a quiet room to ensure a high quality of the sound, thus allowing phoneticians to perform in-depth acoustic analyses. These files are marked in the corpus as "WRITING/AUDIO DETAILS: spoken_offline_lab". Another portion of spoken data were collected by a researcher, who used a laptop computer to record the participants. The label assigned to these texts is "WRITING/AUDIO DETAILS: spoken_offline_classroom". Finally, some participants' audios had to be collected online using the software Google Meet, which still allowed for a face-to-face format, due to the Covid-19 pandemic. These are marked in the corpus as "WRITING/AUDIO DETAILS: spoken_offline_googlemeet". All the audio files can be listened to and downloaded through the corpus web interface, which will be explored in section 2.1.2 of this dissertation. Also, the audio files have their own orthographic transcription, which is available in the corpus. On the COREFL website, the transcription conventions can be found in the "User Guide" > "Transcription conventions" tab.

2.1.2 The Web interface

As already mentioned, the corpus can be navigated using the web interface available on the COREFL website. From the home page, it suffices to click on the "Search / Download" tab (Figure 1) and select either the "Simple search" or the "Advanced search" option, which will now be touched upon.

Figure 1: Screen capture of the homepage of the COREFL website (corefl.learnercorpora.com), with focus on the tabs used to navigate the website. The "Simple search" and "Advanced search" options appear after clicking on the "Search / Download" tab

The two search interfaces (Figure 2; Figure 3) are both structured in an intuitive way: the main sections are easily distinguishable thanks to their different background colours, plus each search option presents a question mark symbol next to it. When clicking upon the symbol, it will display some helpful information about the search field in question.

The simple search interface (Figure 2) is ideal for searching a word or more in the corpus. The output of the search will be concordance lines containing the word or words searched for. In the "Corpus to search" area, users can select which sub-corpus they want to investigate ("Learners of L2 English" or "Natives") and the native language of the participants. The "Words (optional)" field is where words must be written in order to search for them in the corpus. If the search box is left blank, the sub-corpus which was selected in the previous section will be available for browsing and downloading. Finally, the "Result (Output)" field allows to choose how many concordance lines are to be showed per page.

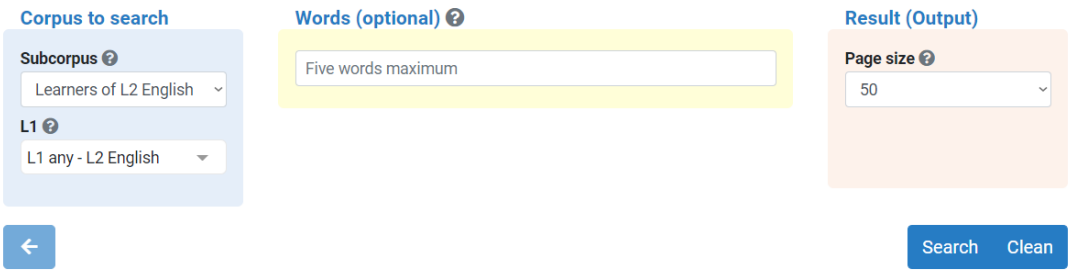


Figure 2: Screen capture of the "Simple search" interface on the COREFL website (corefl.learnercorpora.com/search_simple)

The advanced search interface (Figure 3) offers many more search variables.

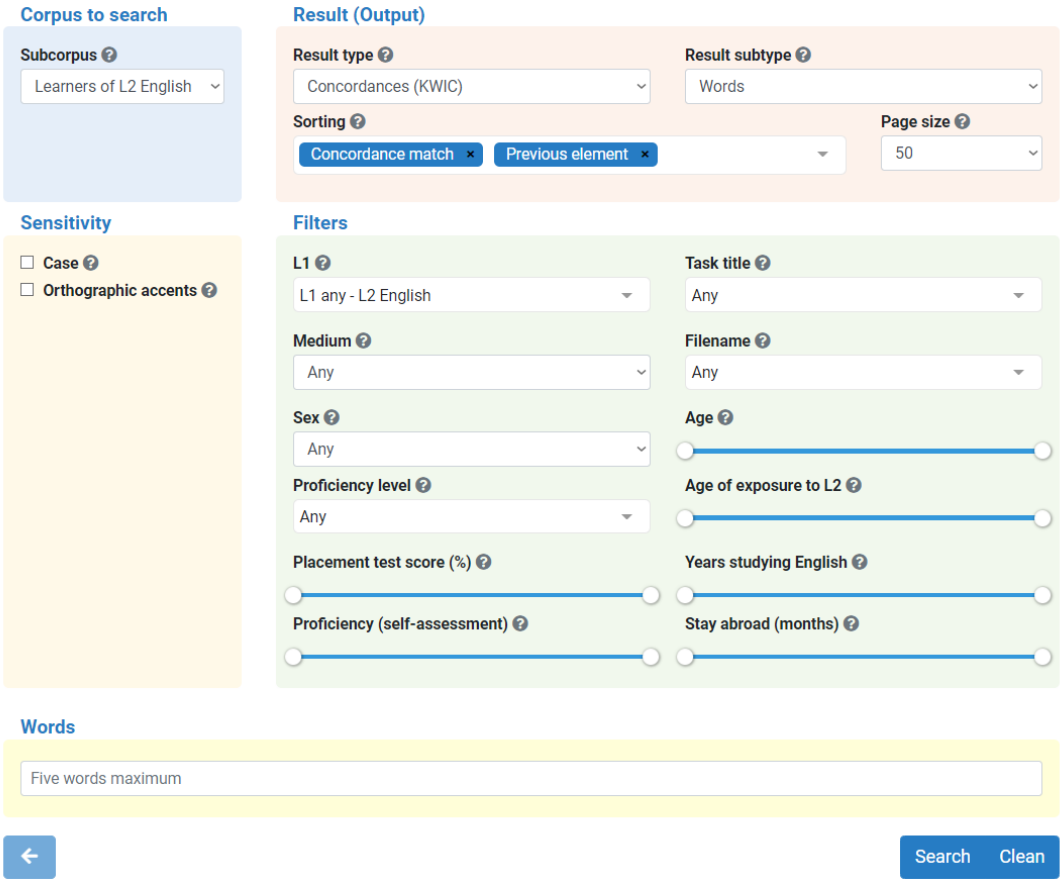


Figure 3: Screen capture of the "Advanced search" interface on the COREFL website (corefl.learnercorpora.com/search)

In the "Corpus to search" area, unlike the one in the simple search, it is possible to select only the preferred sub-corpus and not the participants' native language. On the other hand, the "Result (Output)" area is much richer. Here, users can choose the result type: concordances (KWIC, which stands for Key Word In Context), texts, simple frequency and full frequency. There is also the possibility to choose a result subtype between words, grammatical elements, proximal words and proximal grammatical elements. The order of appearance of results can also be customised by adding filters in

the “sorting” box. Furthermore, just like in the simple search, it is possible to select how many concordance lines should be showed per page. The “Sensitivity” area allows to choose whether the search will be case and accent sensitive. In the “Filters” section, users can accurately refine their search thanks to various variables: the participants’ L1, sex, age, proficiency level, age of exposure to English, years studying it, time spent abroad, placement test score, their self-rated proficiency, the medium of the text (spoken or written), the task title and the file name. Finally, the lower yellow area changes based on the chosen sub-result type: choosing “words” allows to search from one to five words and allows the use of wildcards, like “*” for any number of characters and “?” for one character; the function “grammatical elements” allows to search for both parts of speech and its subtypes (e.g. verb present), plus lemmas, which display all the forms of the searched lemma; the option “words proxim” allows to search for a word separated from another by a number of words which must be specified; “grammatical elements prox” works like “words proxim” but using grammatical words.

2.1.3 Statistics

The most relevant statistical data will now be reported. The data are taken from the COREFL website, on the “Statistics” page (corefl.learnercorpora.com/statistics), where a wealth more can be found.

As of version one, COREFL contains data taken from 2,342 participants, who produced a total of 530,392 words, 583,513 grammatical elements and 2,447 documents. 1,810 documents were made by L2 learners of English, while the natives produced the remaining 637. About the learners’ documents, 1,361 were made by Spanish L2 learners of English, while 449 by the German learners. As briefly mentioned in section 2.1.1, written texts constitute a large majority of the corpus: in the learners’ sub-corpus, there are 1,459 written documents against 351 spoken ones. The age of the participants ranges from 12 to 78 years old. To be precise, learners are from 12 to 62 years old, while natives range from 16 to 78 years old. The number of documents for each proficiency level will now be presented: 231 documents were produced by learners with an A1 level, 295 by learners with an A2 level, 317 by learners with a B1 level, 398 by learners with a B2 level, 379 by learners with a C1 level, and 190 by learners with a C2 level.

2.2 Data extraction

The study at the core of this dissertation, which will be covered extensively in chapter three, is based on data taken from the "Learners of L2 English" sub-corpus. After navigating the sub-corpus, a particular type of error appeared to be quite common, thus it was decided to further enquire it. The error in question involves a lack of subject-verb concord (see Chapter 3.1): very briefly, it consists in the absence of the suffix *-s* which marks the third person singular of the present tense, thus resulting in wrong forms such as "she write".

The most efficient method that was found in order to isolate as many errors of that type as possible will now be presented. In the advanced search tab of the COREFL website, the "Grammatical elements" result subtype was selected. Then, the third person singular pronouns were written in the "Grammatical element" text box (Figure 4), separated from each other by the "|" symbol: "he|she|it". Thanks to the "|" symbol, it is possible to search for concordance lines containing either "he", "she" or "it". Afterwards, another series of search fields were added to the "Grammatical elements" search box by clicking on the plus symbol at the end of the first series of search fields. Finally, in the "Tag" search box on the second line, three tags were selected: "verb" for the category, "present" for tense, and "distinct from third person" for the person. All of the concordance lines that respect the aforementioned search conditions were displayed at the bottom of the page, after clicking on the "Search" button located on the right just below the search fields. The terms that were searched for can be easily found as they are highlighted in bold.

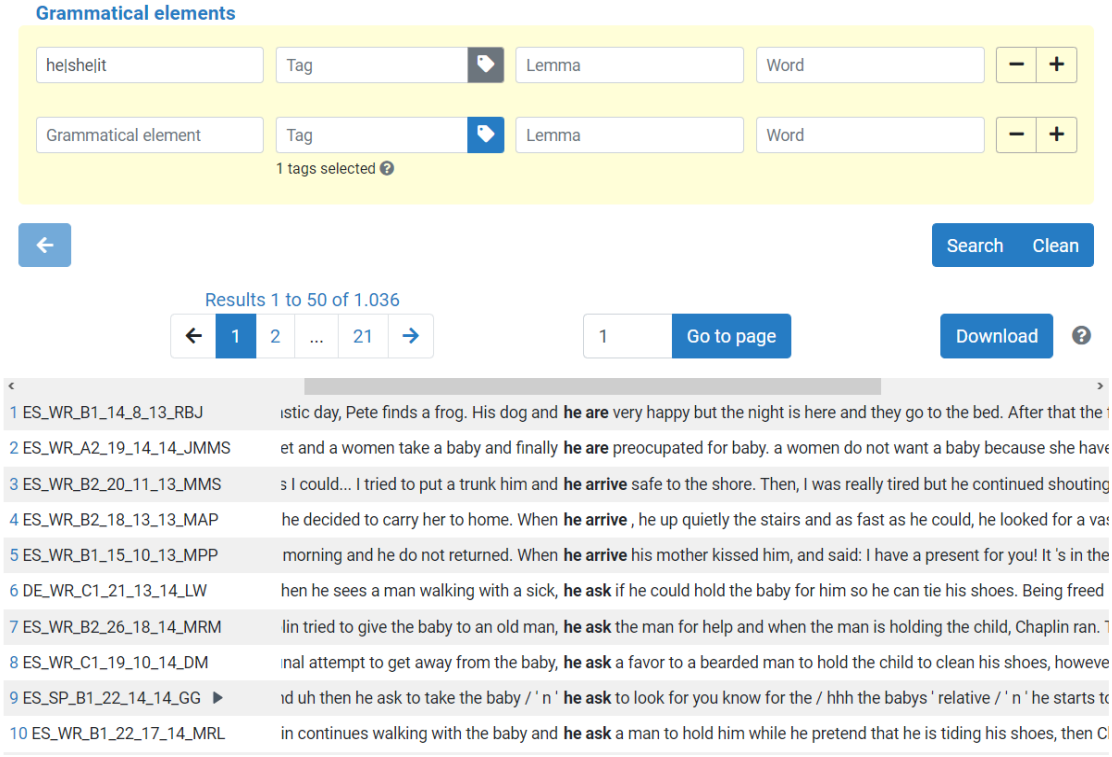


Figure 4: Screen capture of an advanced search with the "Grammatical elements" result subtype selected (corefl.learnercorpora.com/search)

Afterwards, the same search was carried out twice more in order to obtain the errors of the Spanish and the German learners separately. This was easily achieved by setting the "L1" filter to "L1 Spanish – L2 English" in the first search and to "L1 German – L2 English" in the second search. The two lists of concordances were then downloaded thanks to the "Download" button on the right, above the results. This option does not download the concordances alone: it also adds all the metadata attached to the files. For convenience, the data were transferred on two separate Excel worksheets.

2.3 Data analysis

As just mentioned, the data were uploaded on Microsoft Excel. In this software, the data is automatically arranged in rows and columns, allowing for more efficient analyses. Figure 5 shows the first nine columns and 24 rows of the worksheet, which holds the data from the Spanish learners of English.

	A	B	C	D	E	F	G	H	I
1	Filename	Task title	Proficiency level	Medium	Sex	Age	L1	Placement test score (%)	Age of exposure to English
2	ES_WR_A2_12_8_13_AGM	13. Frog	A2	Written	Female	12	Spanish	35	4
3	ES_WR_A2_12_8_13_AGM	13. Frog	A2	Written	Female	12	Spanish	35	4
4	ES_WR_A2_12_8_13_AGM	13. Frog	A2	Written	Female	12	Spanish	35	4
5	ES_WR_A1_12_8_13_EGF	13. Frog	A1	Written	Male	12	Spanish	25.8	4
6	ES_WR_A1_12_6_13_CAM	13. Frog	A1	Written	Female	12	Spanish	0	6
7	ES_WR_A2_12_8_13_AGM	13. Frog	A2	Written	Female	12	Spanish	35	4
8	ES_WR_A2_12_8_13_AGM	13. Frog	A2	Written	Female	12	Spanish	35	4
9	ES_WR_A1_12_9_13_ABG	13. Frog	A1	Written	Male	12	Spanish	29.2	3
10	ES_WR_A1_12_8_13_AMRF	13. Frog	A1	Written	Female	12	Spanish	25	4
11	ES_WR_A2_12_8_13_AGM	13. Frog	A2	Written	Female	12	Spanish	35	4
12	ES_WR_A2_12_8_13_AGM	13. Frog	A2	Written	Female	12	Spanish	35	4
13	ES_WR_A1_12_6_13_EOG	13. Frog	A1	Written	Female	12	Spanish	20.8	6
14	ES_WR_A1_12_7_13_DGG	13. Frog	A1	Written	Male	12	Spanish	22.5	5
15	ES_WR_A1_12_5_13_AGR	13. Frog	A1	Written	Female	12	Spanish	23.3	7
16	ES_WR_A1_12_8_13_ERP	13. Frog	A1	Written	Female	12	Spanish	25.8	4
17	ES_WR_A1_12_9_13_ABG	13. Frog	A1	Written	Male	12	Spanish	29.2	3
18	ES_WR_A1_12_9_13_ABG	13. Frog	A1	Written	Male	12	Spanish	29.2	3
19	ES_WR_A2_12_6_13_SLC	13. Frog	A2	Written	Female	12	Spanish	43.3	6
20	ES_WR_B1_12_9_13_AAE	13. Frog	B1	Written	Female	12	Spanish	46.7	3
21	ES_WR_A1_12_6_13_EAR	13. Frog	A1	Written	Male	12	Spanish	18.3	6
22	ES_WR_A1_12_6_13_UMV	13. Frog	A1	Written	Female	12	Spanish	18.3	6
23	ES_WR_A1_12_6_13_LRR	13. Frog	A1	Written	Female	12	Spanish	20.8	6
24	ES_WR_A1_12_7_13_JEC	13. Frog	A1	Written	Male	12	Spanish	20.8	5

Figure 5: Excel worksheet with data taken from the sub-corpus of Spanish learners of L2 English

The columns presented in Figure 5 contain some of the metadata which come together with the corpus files, the names of which are listed in column A ("Filename"). An exception is column C ("Proficiency level"): the metadata concerning the proficiency levels were missing, so they had to be added manually. It was not difficult to do so, because the proficiency levels can be found in the filenames. As a matter of fact, the filenames include some basic information about the participants. For example, the filename in row two (Figure 5) is "ES_WR_A2_12_8_13_AGM": the first two letters indicate the L1 of the participant, in this case Spanish; the second set of letters specify the medium of the task, which is written; A2 is the proficiency level of the participant; the following numbers are the age of the participant, the years studying English, and the task number respectively; the last letters are the initials of the participant's name. The other columns, which are not present in the picture, include more metadata, such as the number of years of study of English, the months of stays abroad if any, and the concordances themselves which are split into left context, word 1, word 2, and right context. Word 1 and 2 refer to the two search terms: word 1 will be either *he, she*, or *it*, while word 2 will thus be a word tagged as a verb, in the present tense, different from third person. Left and right context correspond to the groups of words which surround the two search terms.

Microsoft Excel also allows to filter data: the small arrow-shaped symbols present in every column's first row (Figure 5) provide many sorting and filtering options. For instance, data can be sorted alphabetically (from A to Z and vice versa) or by colour if

the cells are coloured, specific elements or categories can be manually excluded from view, and it is also possible to create sophisticated text filters. The options that were most useful during the data analysis phase were alphabetical sorting and the manual exclusion of given data categories. Another core feature of Microsoft Excel is the possibility to easily carry out mathematical operations. Specifically, the most used ones in the data analyses were sum, multiplication, division, and subtraction, even though Excel offers many mathematical functions of more complex nature.

After obtaining the results of various data analyses, which will be discussed in detail throughout Chapter 3, it was necessary to compare the different sets of results. However, direct data comparison was not always possible due to the variable sizes of the data sets. A good example of this problem is the difference in size between the Spanish and German data samples: hypothetically, if there were 40 errors in both data samples, it would not mean that Spanish and German learners produce the errors with the same frequency. Data comparison was thus made possible by normalising the frequencies. Normalised frequency is commonly achieved by dividing the raw frequency by the total number of words which compose the data section from where the raw frequency was extracted. Following the example presented above, 40 would be the raw frequency, while the totals would be the number of words in the Spanish and German data sets. The results of the aforementioned operations are usually multiplied by one million before comparing them, mainly because the division generally outputs very small numbers. The final result would thus be the frequency per million words.

In Excel, the data can also be visualised thanks to tables and charts. There are many types of charts and each one is tailored to highlight data in a different way. In the present analyses, two types of charts were used: pie charts and clustered columns charts. As the charts' labels in Excel indicate, pie charts are useful to display portions of a whole, while clustered columns charts are suitable for comparing values across a few categories.

In this second chapter, the main focus was on the COREFL corpus, which was explored in detail. The methodologies used to extract data from the corpus and analyse it were also accounted for. In chapter three, which is arguably the core of this dissertation, the analyses carried out on the data extracted from COREFL will be presented, and its results will be discussed.

CHAPTER 3
DATA ANALYSIS RESULTS AND DISCUSSION

This third chapter presents the results of the analysis carried out on the errors extracted from the COREFL corpus. The chapter is organised in five sections: the first introduces the data analysis and briefly covers the types of errors examined in the analysis; the second, third and fourth sections present statistical re-elaborations of the data in the form of tables and figures. These represent the frequency counts and the normalised frequencies of the errors across verb types, proficiency levels and the media, and the self-corrections; the last section provides a discussion of the results.

3.1 A study of the errors made by Spanish and German L2 learners of English

This study explores the cases of omission of the third person *-s* in the linguistic production of Spanish and German L2 learners of English attested in the COREFL corpus. The errors extracted from COREFL were carefully checked in order to remove the results which did not match the search parameters. These were likely due to the automatic tagging software which mislabelled some words. Before proceeding with the illustration of the results, it may be useful to provide more information about the aforementioned type of error.

The omission of the third person *-s* equals to a lack of concord between the subject and the verb. Biber et al. (2002, p. 232) define the subject-verb concord rule as follows:

> in finite clauses, the verb phrase in a clause agrees with the subject in terms of number (singular or plural) and person (first, second or third person). Except for the verb *be*, subject-verb concord is limited to the present tense, and to the choice between the base form (e.g. *walk*) and the *s*-form (e.g. *walks*) of the finite verb.

Based on this definition, it can be said that the learners made the wrong choice when choosing between the base form and the *s*-form of the finite verb. In other words, in erroneous forms such as “he play”, the learners used the base form of the verb, while they should have used the *s*-form in order to correctly follow the subject-verb concord rule.

The reason why this particular type of error was chosen as the core of this analysis is related to its relevance in the SLA theory and in particular with regard to the Morpheme Order Studies (see Chapter 1.4.2). In fact, the third person *-s* appears to be one of the morphemes which is used with less accuracy by learners of L2 English. This morpheme is found at the bottom of the "hierarchy of L2 acquisition", elaborated by Krashen (1977, in Ellis and Barkhuizen, 2005). The hierarchy is based on the results of the morpheme studies carried out in the 1970s, which proved that it was not necessarily conditioned by the learners' L1, age or setting.

The main goal of the analysis conducted in this dissertation was thus to investigate the extent to which errors concerning the omission of the third person *-s* were present in the linguistic production of Spanish and German L2 learners of English. More specifically, the research hypothesis explored whether approximately the same amount of errors is made by the two groups of learners or not.

Various aspects of the data were taken into account to carry out the analysis. The errors were analysed and categorised according to their verb type, the pronoun used (either *he, she,* or *it*), the learners' proficiency level, the medium of the task, and the self-corrections (i.e. whether the learners realised they made a mistake and corrected it). A combined study of the proficiency levels and the medium was also carried out. All of the above analyses were first performed separately on the Spanish and German sets of data, then the results were compared. However, in the following sections only the data related to some of the verb types, the proficiency levels, the media, and the self-corrections will be presented, because the remaining analyses did not provide interesting results.

Due to the author's lack of knowledge of the Spanish and the German language, the errors could not be analysed with a contrastive approach, meaning that it was not possible to check whether L1 transfer influenced the production of the errors. Therefore, the errors were considered from a developmental perspective only.

Before presenting the data, a short premise must be made about the proficiency levels of the participants. In COREFL, the sub-corpus of German L2 learners of English does not contain data samples of the linguistic production of participants at the A1 and A2 proficiency levels. Because of this, in order to carry out a proper comparison between

the Spanish and the German data, the proficiency levels taken into account are only B1, B2, C1, and C2, unless otherwise stated.

Some general statistics about the data will now be presented. The combined total number of errors, which includes all proficiency levels, amounts to 882, of which 848 were made by the Spanish learners and the remaining 34 by the German learners. The number of errors found in the Spanish learners' samples drops to 421 without including the A1 and A2 proficiency levels. The frequency counts cannot be compared directly, so it was necessary to calculate the normalised frequencies. This was done by dividing the total number of instances of errors by the total number of tokens present in each sub-corpus. The normalised frequencies of the total numbers of errors for the Spanish and German components are 2.11 and 0.25, respectively. These figures confirm that the Spanish learners make more mistakes than the German learners.

In order to give more context to the figures mentioned above, Table 2 shows the numbers of learners at each proficiency level for both the Spanish and German corpora.

	B1	**B2**	**C1**	**C2**
Spanish	304	307	181	43
German	13	91	198	147

Table 2: Numbers of learners at each proficiency level for the Spanish and the German corpora

3.2 Errors and verb types

In this section, the errors will be presented by sorting them according to their verb type. Both the frequency counts and the normalised frequencies will be reported. Along with the frequency counts, some extracts from the corpus will be presented to exemplify the various verb types. The verbs were first identified as either auxiliary, stative, or dynamic, then they were further divided into the respective sub-categories. The nature of stative and dynamic verbs, and of their sub-categories, will now be briefly explained.

Stative verbs refer to *states* and are used prototypically with the simple aspect, while dynamic verbs "are verbs of *doing* and *happening*. They refer to actions, activities, events and processes, and are used with either simple and progressive aspect" (Falinski, 2008, p. 35). For the present analysis, stative verbs were further categorised as verbs of mental, emotional and cognitive states, verbs of perception and relational verbs. Dynamic verbs were further divided into telic and atelic verbs. Telic verbs differ from

atelic verbs because of their "completion point": telic verbs have an internal completion point, while atelic verbs do not have one and depict actions that "can be protracted indefinitely or broken off at any point, without attaining a final *object* or *result*" (Falinski, 2008, p. 94). Accomplishments can be seen as verbs lying in-between telic and atelic verbs: they depict atelic expressions that become telic when a "bounded" object is added. For instance, in the sentence "I am writing a letter", the object "a letter" sets a definite limit to the action of writing, thus making it telic (Huddleston and Pullum, 2002, p. 120). Finally, telic verbs were divided into achievement verbs, which entail actions having a definite result (e.g. *finish*), and semelfactive verbs, which describe actions that are repeated multiple times (e.g. *knock*).

Figure 6 and Figure 7 show the frequency count of the errors sorted according to their verb type: auxiliary, stative, and dynamic. Figure 6 presents the data related to the Spanish learners, while Figure 7 presents the data related to the German learners.

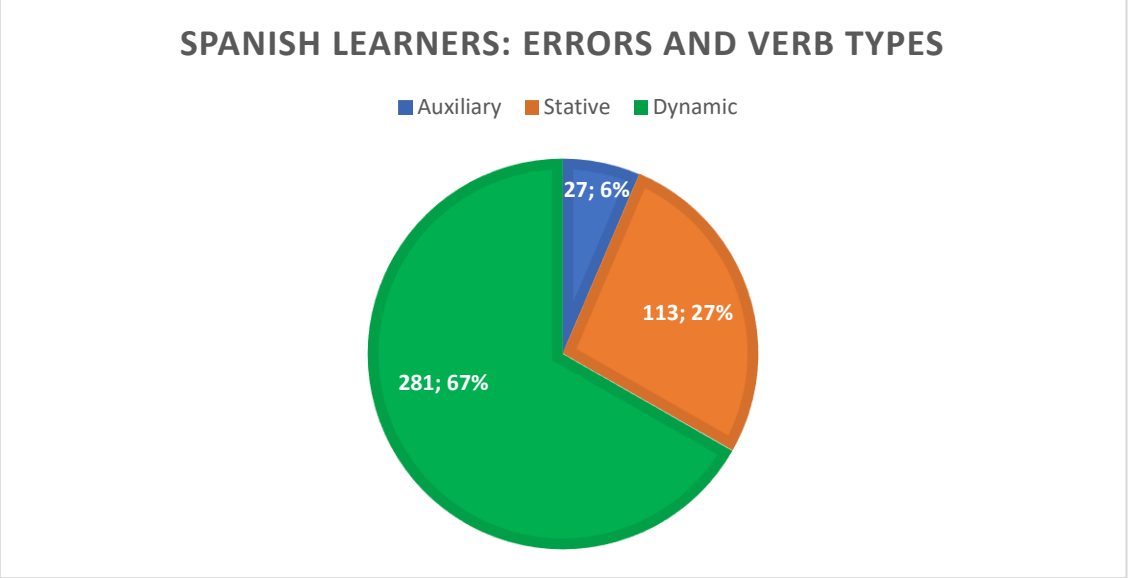


Figure 6: Frequency count of the errors produced by the Spanish L2 learners of English, sorted according to their verb type.

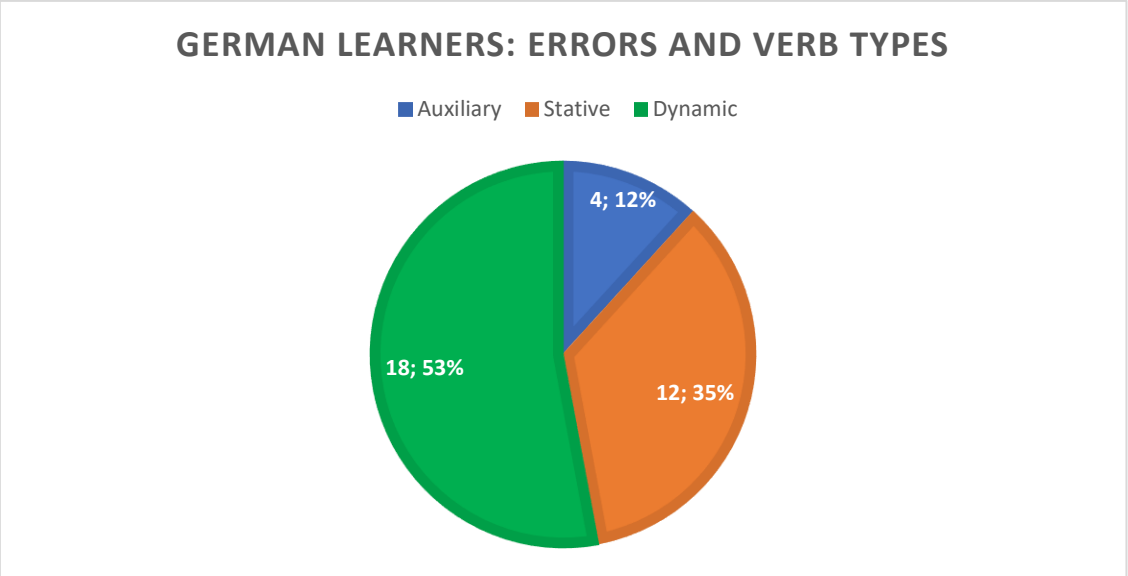


Figure 7: Frequency count of the errors produced by the German L2 learners of English, sorted according to their verb type.

Both charts suggest that the majority of errors take place when learners use dynamic verbs. In the following extracts, taken from the Spanish and the German learners' samples, the errors made with dynamic verbs are highlighted in italics:

(1) He wears a hat, a weird moustache and a stickwalk. Suddenly, *he find* a baby on the ground, next to a bin.
(2) He can not stand over the rock, so *he use* a little branch to stand himself.
(3) In the end of the video it looks as if Charlie is keeping the baby, because *he walk* away with it in his arms.

Auxiliary verbs appear to be used with more accuracy than stative and dynamic verbs. The examples from four to six contain the errors made with auxiliary verbs, while the examples from seven to nine contain the errors made with stative verbs:

(4) In the morning the boy is worried because *he don't* find her frog and he running […].
(5) […] she is the monster and in order to safe her family *she have* to kill herself.
(6) […] he decides to pick it up again and go away because *he don't* want to get in trouble with the policeman.

(7) He is tired and drops under the rock. He cries but *he see* the frog.
(8) Secondly the baby, when *he notice* that there is a baby behind him […].

(9) […] he's taking out the baby again and uh hhh *he seem* somewhat resignated and then he sits on on the curb […].

The main difference between the two charts is the higher frequency of errors concerning dynamic verbs in the Spanish learners' samples: they amount to 67% of the total of errors made by the Spanish learners, compared to 53% in the German learners' samples. German learners thus seem to produce more errors with stative and auxiliary verbs than their Spanish counterparts, namely 35% of the errors with stative verbs and 12% of errors with auxiliaries. By contrast, the Spanish learners made 27% and 6% of the two types of errors, respectively.

The frequency counts of stative and dynamic verbs' sub-categories are displayed in Figure 8 and Figure 9 respectively. The former shows the data related to the total of the sub-categories of stative verbs, while the latter shows the data related to the total of the sub-categories of dynamic verbs.

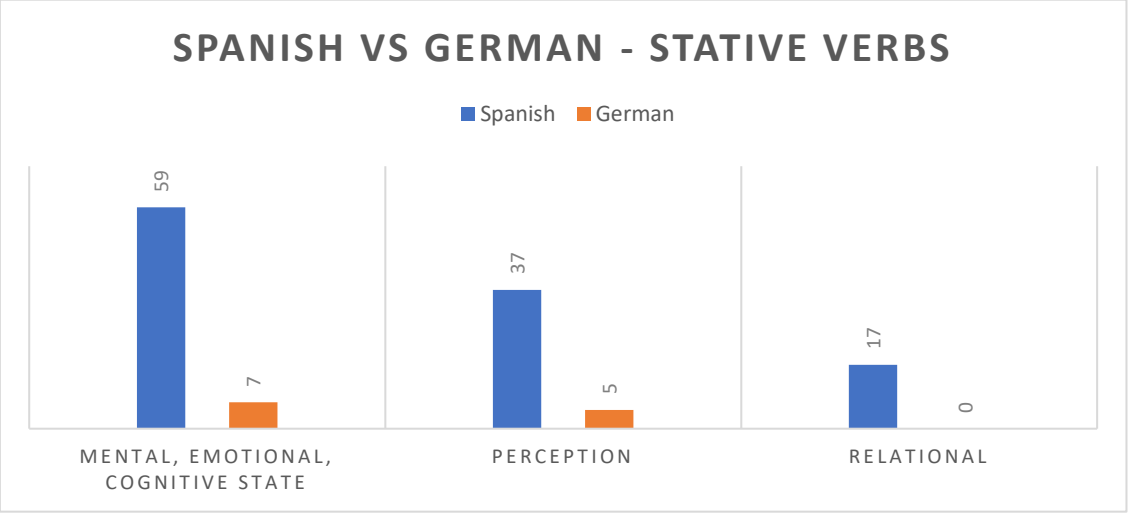


Figure 8: Frequency count of the errors made by Spanish and German L2 learners of English, sorted according to the sub-categories of stative verbs.

According to the first chart (Figure 8), in both the Spanish and German learners' samples, most of the errors concerning stative verbs are found when dealing with stative verbs of mental, emotional, and cognitive states (e.g. *decide*, *think*, *understand*), followed by perception verbs (e.g. *see*, *hear*, *seem*) and relational verbs (e.g. *get*, *become*). In particular, relational verbs are present in a significantly smaller amount than the other two categories: only 17 errors made by the Spanish learners and none by the

German learners. The extracts from ten to twelve contain the errors made with stative verbs of mental, emotional, and cognitive states; the extracts from thirteen to fifteen contain the errors made with perception verbs, and the extracts sixteen and seventeen contain the errors made with relational verbs:

(10) He is 9 years old. He has a dog and a frog. *He love* animals.
(11) The kid was sad because he could not recapture the frog, but *he understand* the message, the frog belongs to the forest.
(12) Chaplin was walking by a_second before and she 's seeing him so *she assume* that he 's responsible for it again.

(13) […] various bricks or something similar falling down on him. But, *it appear* that someone has taken those objects intentionally […].
(14) He lights up one of them, and *he see* a little baby on the ground […].
(15) […] so she gets really angry uh and *she see* uh what Chaplin is is doing […].

(16) He kills two men in a crisis, and *he become* the hero of Gotham because the citizens are upset with the rich people.
(17) When he insinuates her that she should have lost it *she get* angry and in a bad way tells him to get away.

From the second chart (Figure 9), it appears that both Spanish and German learners made more errors with telic achievement verbs (e.g. *receive*, *take*, *find*), with 165 and 15 errors respectively. In the following extracts, the errors made with telic achievement verbs are highlighted in italics:

(18) […] she even tries to hit him / and uh then *he ask* to take the baby […]
(19) They find a deer and *it push* he and drop to the river with his dog.
(20) Charlie sits down on the sidewalk, taking a closer look at the baby, when *he find* a letter in its clothes.

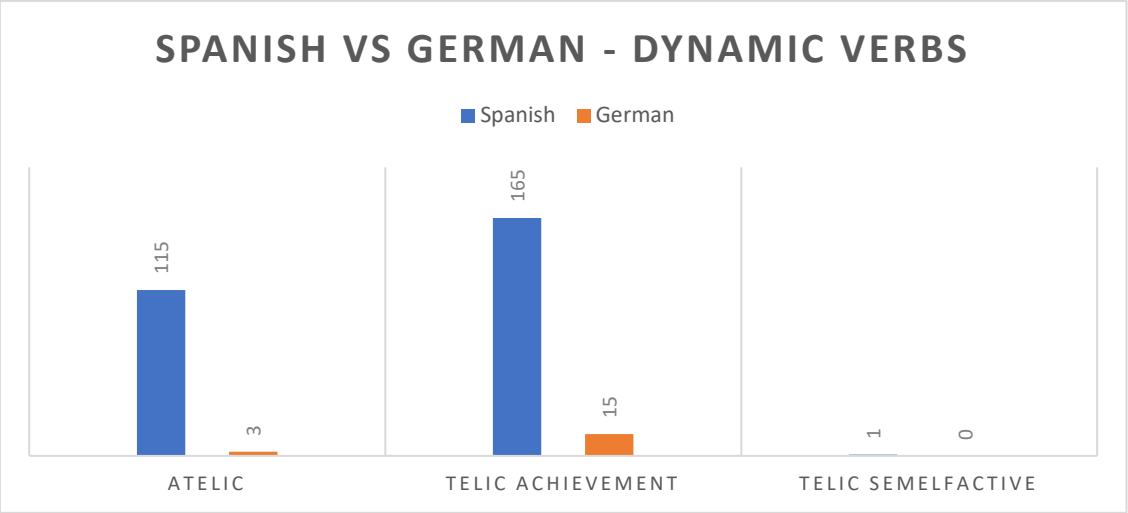


Figure 9: Frequency count of the errors made by Spanish and German L2 learners of English, sorted according to the sub-categories of dynamic verbs.

The other category with figures that are similar enough to the one just mentioned is the one concerning atelic verbs (e.g. *carry*, *sit*, *walk*), with 115 errors made by the Spanish learners and 3 by the German learners. The following extracts exemplify the errors made featuring atelic verbs:

(21) […] he can not see the frog, *he watch* a bee and the boy is look for […]

(22) Without knowing what to do, *he sit* on the floor and for an instant he thinks about letting the baby in sewers.

(23) […] it looks as if Charlie is keeping the baby, because *he walk* away with it in his arms.

On the other hand, only one error can be found associated with the telic semelfactive verb category referred to the Spanish learners' linguistic samples:

(24) […] to explain what he wants to do as a comedian hhh and at the end *he shoot* one man.

No errors of this type are found in the German learners' linguistic samples. Finally, there are no errors regarding accomplishment verbs in neither group of learners.

As previously mentioned, the frequency counts do not provide an accurate picture of the results. The normalised frequencies, however, enable us to compare the data directly. Figure 10, Figure 11 and Figure 12 contain the normalised frequencies per thousand words of the data presented thus far in this section. These normalised frequencies were calculated by dividing the number of errors found in each category by

the number of tokens and multiplying the result by 1000, first in the Spanish learners' data and then in the German learners' data.

The chart in Figure 10 shows the normalised frequency of the errors sorted according to their verb type. The chart confirms that both the Spanish and the German learners made more errors associated with dynamic verbs. It also confirms that auxiliary verbs were used with more accuracy by both the Spanish and the German learners, followed by the stative verbs. However, unlike the frequency counts in Figure 6 and Figure 7, it shows that the Spanish learners made more errors than the German learners in all of the three categories. The trend followed by the normalised frequencies of the German learners' data is also steadier than the one followed by the Spanish learners' data: the gaps between the normalised frequencies are smaller in the German learners' data.

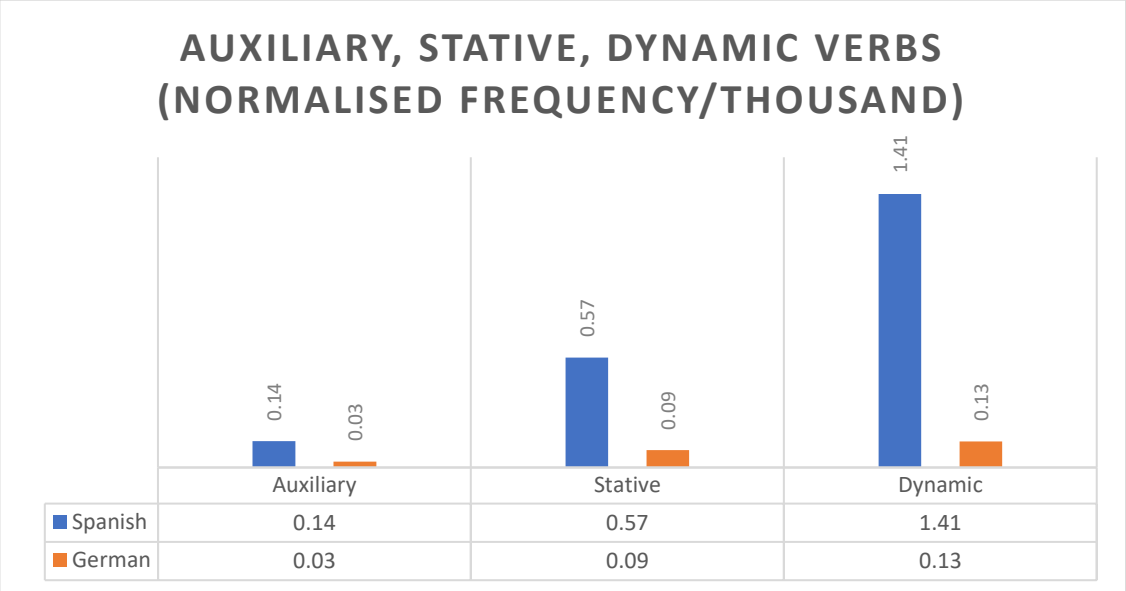


Figure 10: Normalised frequency per thousand words of the errors produced by Spanish and German L2 learners of English, sorted according to their verb type.

According to the data in Figure 11, the normalised frequency of the errors found associated with the sub-categories of stative verbs shows that the Spanish leaners made most of the errors in all categories. In both the Spanish and the German learners' samples, the majority of errors are found within the category of stative verbs of mental, emotional and cognitive states. The Spanish learners made fewer errors with perception verbs, and even fewer with relational verbs. On the other hand, the figures found in the German learners' samples are rather similar to each other: the normalised frequency is

0.05 for the category of stative verbs of mental, emotional and cognitive states and 0.04 for the category of perception verbs. Furthermore, there are no errors regarding relational verbs in the German learners' samples.

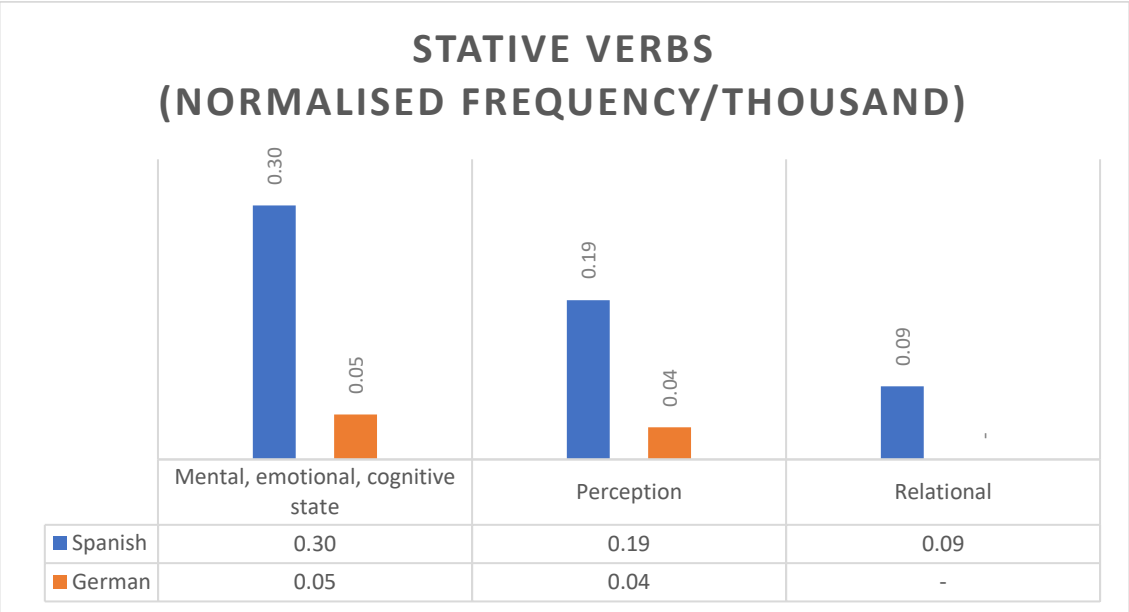
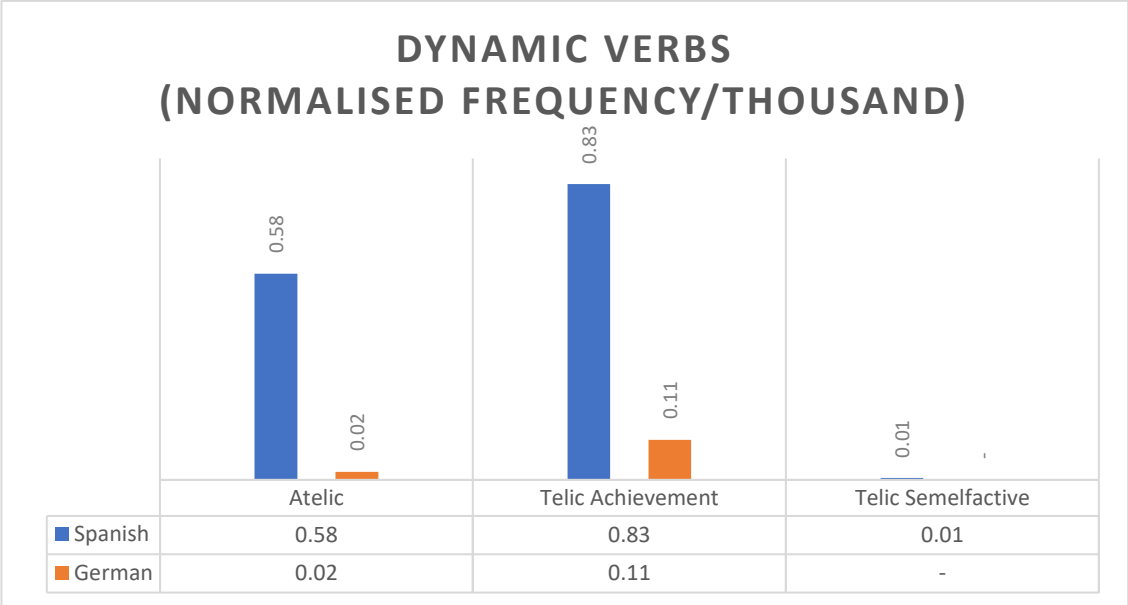


Figure 11: Normalised frequency per thousand words of the errors produced by Spanish and German L2 learners of English, sorted according to the sub-categories of stative verbs.

The chart in Figure 12 leads to very similar conclusions to the ones attested by the frequency count in Figure 9: both the Spanish and the German learners made more errors with telic achievement verbs, followed by atelic verbs. Figure 9 also showed that there is only one error associated with telic semelfactive verbs, and it is part of the Spanish learners' samples: this finding is clearly visible also in Figure 12, in which there is a normalised frequency of only 0.01 for the Spanish learners and nothing for the German learners. In this chart, just like in the others previously shown, most of the errors in all categories come from the Spanish learners, while the normalised frequencies of the errors made by the German learners keep being lower.



	Atelic	Telic Achievement	Telic Semelfactive
Spanish	0.58	0.83	0.01
German	0.02	0.11	-

Figure 12: Normalised frequency per thousand words of the errors produced by Spanish and German L2 learners of English, sorted according to the sub-categories of dynamic verbs.

3.3 Errors across proficiency levels and media

The results of the data analysis in relation to the proficiency levels and the media will now be presented. Firstly, the two sets of data will be presented separately, then the results of the combined analysis of both data sets will be provided.

Table 3 contains the frequency count of the errors sorted according to the learners' proficiency level. The data show that out of 848 errors made by the Spanish learners, the majority come from learners possessing the A2 and B1 proficiency levels, with 278 and 275 errors respectively. The Spanish learners with a C1 and a C2 proficiency level apparently made fewer errors, with 27 and 4 errors respectively. The errors made by the German learners appear to follow a different trend from the one just described. The German learners with a C1 and a C2 proficiency level seem to produce the most errors: out of 34, there are 15 and 11 errors respectively, while in the B1 and B2 proficiency levels the error count amounts to 3 and 5.

	A1	**A2**	**B1**	**B2**	**C1**	**C2**	**Total**
Spanish	149	278	275	115	27	4	**848**
German	0	0	3	5	15	11	**34**

Table 3: Frequency count of the errors produced by the German L2 learners of English, sorted according to the learners' proficiency level.

In Figure 13, the normalised frequency of the errors sorted by proficiency levels includes only the proficiency levels B1, B2, C1 and C2. The normalised frequency was obtained by dividing the number of errors in each proficiency level and group of learners by the number of tokens produced for each group of learners at the corresponding proficiency level.

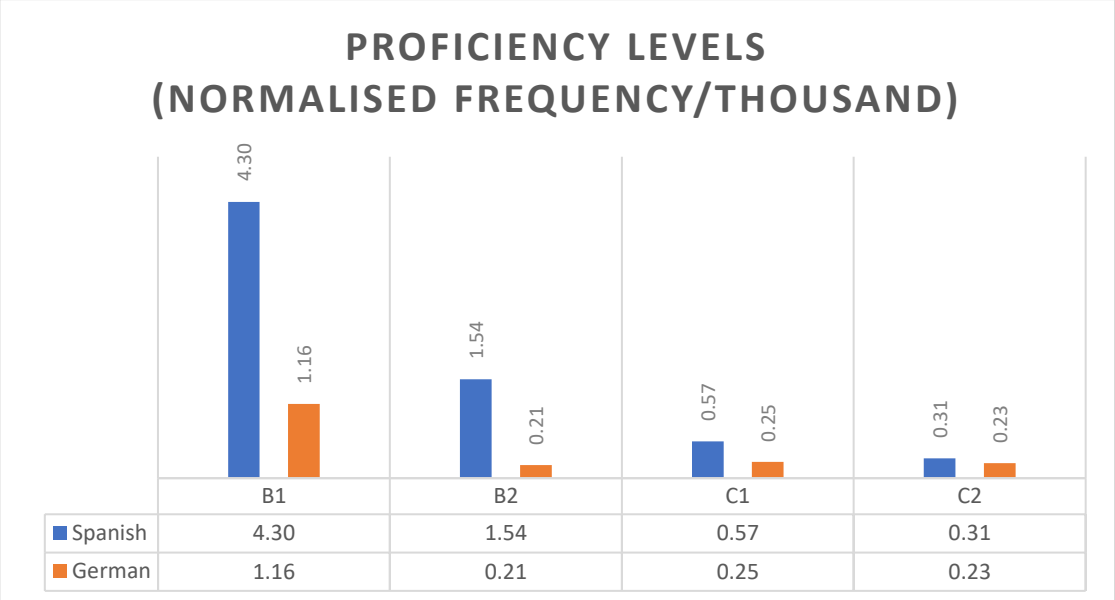


Figure 13: Normalised frequency per thousand words of the errors produced by Spanish and German L2 learners of English, sorted according to the proficiency levels B1, B2, C1 and C2.

The chart in Figure 13 suggests that the Spanish learners seem to gradually produce fewer errors as the proficiency level gets higher (i.e. from B1 to C2). The normalised frequency of the errors found at the B1 proficiency level decreases considerably at the B2 proficiency level, and it keeps decreasing, even if less prominently, at the C1 and C2 proficiency levels.

The normalised frequency of the German learners' errors significantly decreases from the B1 to the B2 proficiency level, similarly to the aforementioned Spanish learners' data. However, unlike the Spanish learners, the German learners apparently produce a steadier number of errors at proficiency levels B2, C1 and C2. There are some differences between the normalised frequencies found at the B2, C1 and C2 proficiency levels, namely that the frequency at the C1 proficiency level appears to be higher than the one found at the C2 and B2 proficiency levels, but these differences are rather negligible. It is also worth noting that the gap between the normalised frequencies of the

Spanish learners and the German learners' errors gets smaller as the proficiency levels get higher.

A comparison of the normalised frequencies of the errors produced by the Spanish and the German learners suggests that the former produce more errors than the latter, no matter the proficiency level.

The analysis of the errors throughout the media will now be presented. The data related to the media only contain the errors produced by learners with a proficiency level ranging from B1 to C2. Table 4 shows the frequency count of the errors sorted according to the tasks' medium. The data show that the Spanish learners produce more errors in written texts, while the opposite is true for the German learners. Some examples of spoken and written texts can be found in the extracts below: examples 25 and 26 exemplify the spoken medium, while examples 27 and 28 exemplify the written medium. Examples 25 and 27 were taken from the Spanish learners' samples, while examples 26 and 28 were taken from the German learners' samples. The errors are highlighted in italics, just like in the extracts previously provided:

(25) […] go to look to look uh hhh his fro = uh his frog and then uh he / uh hhh *he know* that the frog uh hhh was escape / was esc = / uh […]

(26) […] other things uh options where he could pick the bae uh put the baby and *he decide* to uh / look in the hole […]

(27) Charles starts to stare at him and smiling. In the final scene, *he stand* up from the floor and walks away with the child.

(28) In the end of the video it looks as if Charlie is keeping the baby, because *he walk* away with it in his arms.

	Written	**Spoken**
Spanish	231	190
German	8	26

Table 4: Frequency count of the errors produced by the Spanish and German L2 learners of English, sorted according to the medium of the task.

The normalised frequencies of the data presented in Table 4 are shown in Figure 14. Unlike the data in Table 4, the data in Figure 14 seem to indicate that both the Spanish and the German learners produce more errors in spoken texts. The data confirm

that the Spanish learners appear to generally produce more errors than the German learners, similarly to what was found in the previous data about proficiency levels and verb types. However, in both the Spanish and the German learners' data, the gap between the normalised frequencies of the errors found in the spoken and in the written medium is rather small. More specifically, the normalised frequency of the errors made by the Spanish learners in the written medium is 4.13, while it is 4.60 in the spoken medium. This rather small difference is found also in the normalised frequency of the errors made by the German learners: 0.10 is the normalised frequency in the written medium, while 0.46 is the normalised frequency in the spoken medium.

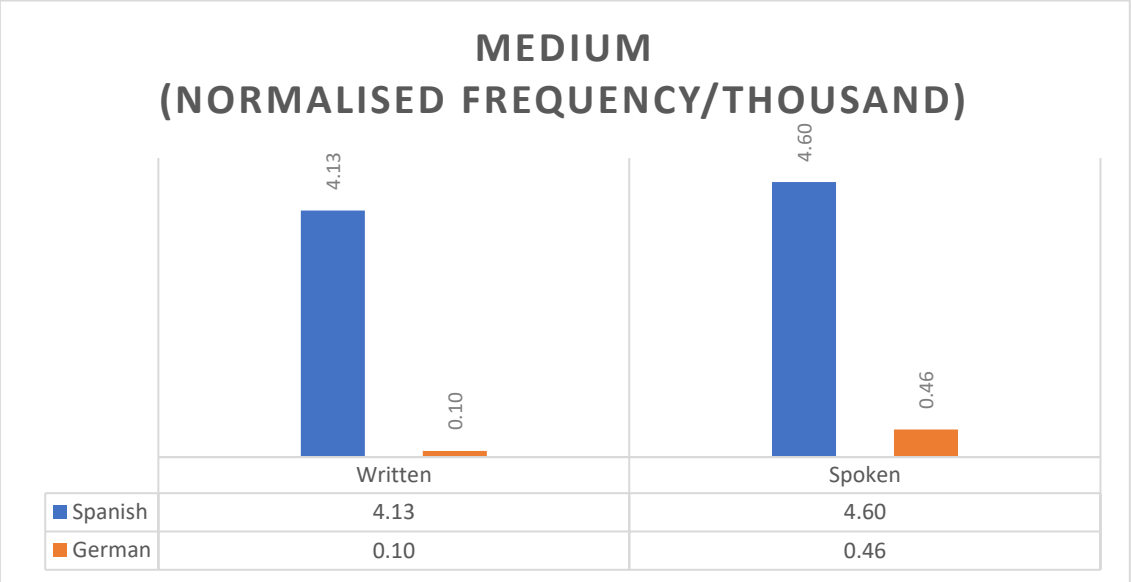


Figure 14: Normalised frequency per thousand words of the errors produced by the Spanish and German L2 learners of English, sorted according to the medium of the task.

Table 5 and Table 6 contain the normalised frequencies of the errors found in both spoken and written English and across the proficiency levels.

Table 5 presents the data regarding the Spanish learners: the majority of the errors in both media appear to occur at the B1 proficiency level, and in both media the normalised frequencies decrease as the proficiency levels get higher. The normalised frequencies seem to stabilise at the C1 and C2 levels in the written medium. Also, the majority of errors is always found in the spoken medium, no matter the proficiency level.

Spanish	**B1**	**B2**	**C1**	**C2**
Spoken	6.93	2.93	1.29	0.46
Written	3.46	0.99	0.24	0.23

Table 5: Normalised frequency per thousand words of the errors produced by the Spanish L2 learners of English, sorted according to the medium of the task and the learners' proficiency level.

Table 6 shows the data related to the German learners. Analogously to the Spanish data, the normalised frequencies denote that most of the errors appear to be made at the B1 proficiency level in both media. However, the normalised frequencies do not regularly decrease as the proficiency levels get higher. At first, the normalised frequency follows the same pattern as the one found in the Spanish data (Table 5), that is, it is highest at the B1 proficiency level, then gets lower at the B2 level. But, unlike the Spanish data, the normalised frequency rises at the C1 level in the spoken medium, and in both media the normalised frequency is slightly higher at the C2 level than at the B2 level. Nonetheless, the difference between the normalised frequencies found at the B2 level and those found at the C2 level is negligible, similarly to the normalised frequencies of the errors made by the German learners at the B2, C1 and C2 proficiency levels (Figure 13).

German	**B1**	**B2**	**C1**	**C2**
Spoken	1.64	0.31	0.55	0.35
Written	0.73	0.14	0.03	0.15

Table 6: Normalised frequency per thousand words of the errors produced by the German L2 learners of English, sorted according to the medium of the task and the learners' proficiency level.

3.4 Self-corrections

In this section, the data related to the self-corrections found after the errors will be presented. An example of a self-correction, taken from the Spanish learners' samples, is highlighted in the following utterance:

(29) […] the woman comes off the establishment she was in and then she guess *she guesses* hhh this is Charles Chaplin trying to fool her again […].

The vast majority of the self-corrections, like the one just presented, were found in the spoken texts. This is very likely due to the fact that, in the written texts, it would be

rather strange to write a correction next to an error instead of just replacing the error with the correct version.

Table 7 shows the frequency count of all the self-corrections, sorted according to the learners' proficiency level and the learners' L1. The Spanish learners seem to add fewer self-corrections as the proficiency levels get higher, while the opposite situation can be found in the German learners' samples.

	B1	**B2**	**C1**	**C2**	**Total**
Spanish	11	5	3	0	19
German	0	2	3	5	10

Table 7: Frequency count of the self-corrections produced by the Spanish and German L2 learners of English, sorted according to the learners' proficiency level.

As previously mentioned, the normalisation of the frequency count is necessary to obtain data which is actually comparable. The normalised frequency of the self-corrections is shown in Figure 15.

The Spanish learners made the majority of the self-corrections at the B1 level, while no self-corrections were found at this proficiency level in the German learners' samples. The following extracts exemplify the self-corrections made by the Spanish learners at the B1 proficiency level:

(30) […] she wants to / hhh to / give the baby back to him / ' n ' she start / *starts* screaming at him […].

(31) […] then this old man sees a baby laying in the floor / 'n' he decide the *he decides* to take him […].

At the B2 proficiency level, on the other hand, the normalised frequencies of the self-corrections made by the Spanish and the German learners are almost identical: 0.07 is the normalised frequency of the Spanish learners' self-corrections, while 0.08 is the normalised frequency of the German learners' self-corrections. The self-corrections at the B2 level are highlighted in the following extracts, of which the first is taken from the Spanish learners' samples and the second from the German learners' samples:

(32) […] when she discovers again that the child is uh in the baby carriage / uh / she start to uh *she starts* to shout and to call Charles Chaplin again […].

(33) […] he discovers a piece of paper, unfolds it and after reading he know *knows* that this is an orphan child […].

The normalised frequencies at the C1 level are almost identical as well, but they are slightly lower than those found at the B2 level: the normalised frequency of the Spanish learners' self-corrections is 0.06, while the normalised frequency of the German learners' self-corrections is 0.05. The following extracts contain examples of self-corrections made at the C1 proficiency level:

(34) […] I think accidentally passed the woman and she comes out the store she see= *she sees* that there is another baby […].

(35) […] he has issues with walking so he give *gives* the baby to the man […].

Finally, the German learners made the majority of the self-corrections at the C2 proficiency level, while none were found in the Spanish learners' samples. In the following extracts, the self-corrections made by the German learners at the C2 proficiency level are highlighted:

(36) […] and forces him to take the child and he take *takes* the child and walks away and then sits down […].

(37) […] he tries to trick him by saying that he need = *needs* to tie his shoe laces and hands the baby over […].

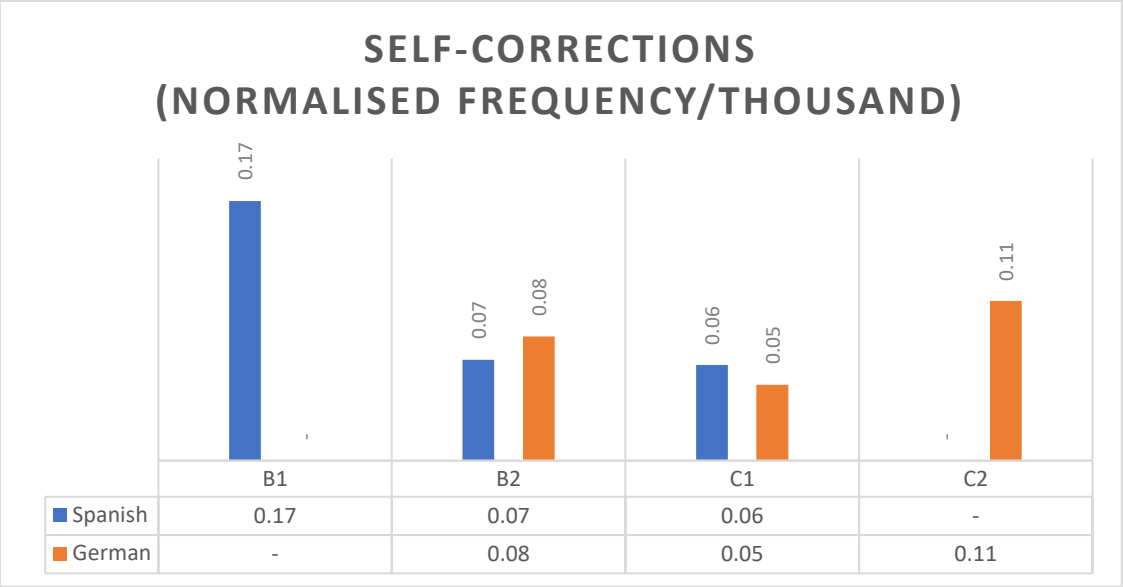


Figure 15: Normalised frequency per thousand words of the self-corrections produced by Spanish and German L2 learners of English, sorted according to the proficiency levels B1, B2, C1 and C2.

The normalised frequencies of the grand total of self-corrections found in the Spanish and the German learners' samples are rather close: 0.10 and 0.07, respectively. This means that, overall, the Spanish learners made slightly more self-corrections, but this does not stand true for all proficiency levels: at the B2 and C2 proficiency levels, the German learners made more self-corrections, even though the gap between the Spanish and the German learners at the B2 level is very small.

3.5 Discussion

The data presented in this chapter may be interpreted in the following way. Figures 6 and Figures 7 show that both the Spanish and German learners made more errors when using dynamic verbs. The presence of many dynamic verbs could be related to the nature of the tasks, which mostly require participants to produce a narrative text rather than a descriptive one. Dynamic verbs, being verbs that express actions and events, are more likely to be found in a narrative text compared to stative verbs. A similar claim can be made about atelic verbs (Figure 12): the learners seem to produce fewer errors when using atelic verbs, which may be explained by the fact that atelic verbs, unlike telic verbs, are often used in the progressive form. This would mean that they may be generally used less in their simple form, thus reducing the chances of finding errors related to the use of the third person *-s* with atelic verbs.

The normalised frequencies in Figure 10 about the errors found across auxiliary, stative and dynamic verbs show that the Spanish learners produce more errors than the German learners in all categories. However, it also shows that the German learners' errors follow a steadier trend. In other words, the normalised frequencies of the German learners' errors do not vary significantly, unlike the Spanish learners' data. This trend suggests that when the German learners have to apply the subject-verb concord rule, they do so with approximately the same difficulty in all three verb categories (i.e. auxiliary, stative and dynamic verbs). Furthermore, the same can be said about the normalised frequencies of the errors associated with the sub-categories of stative verbs (Figure 11): in this chart, the data concerning the German learners behave similarly.

The data regarding the presence of errors across the media (Figure 14) show that all learners, no matter their L1, seem to produce more errors in the spoken medium. What contributes to explaining the gap between written and spoken production is the psycholinguistic context, that is "whether learners have the opportunity to plan their production" (Ellis, 1997, p. 27). In this respect, Ellis's (1997) research can be quoted: he presents the case of Jean, a French learner of L2 English, who made a considerably higher number of errors in the production of an unplanned narrative compared to a planned one. This example would seem to confirm that the psycholinguistic context does play a role in the accuracy of learners' production.

The analysis of the errors throughout the proficiency levels (Figure 13) also provided some interesting findings. The trend followed by the errors in the German learners' samples turned out not to be linear: the errors decrease from the B1 to the B2 proficiency level, but they appear to remain steady at the B2, C1 and C2 proficiency levels, instead of gradually decreasing like in the Spanish learners' samples. In particular, the errors in the German learners' samples are slightly more numerous at the C1 proficiency level than at the B2 and C2 levels. The difference between the figures found at these proficiency levels is very small, but if the pattern they form is looked at more closely, it may be interpreted as an expression of the "U-shaped course of development". According to Ellis (1997, p. 23), "[a]cquisition follows a U-shaped course of development; that is, initially learners may display a high level of accuracy only to apparently regress later before finally once again performing in accordance with target-language forms." In other words, the small increase in number of errors found at

the C1 proficiency level does not necessarily indicate that learners having the C1 proficiency level are at an earlier stage of acquisition than learners at a lower proficiency level who produce fewer errors. It could rather indicate that the learners at higher proficiency levels, namely C1 and C2, are slightly less accurate than learners at a B2 level when dealing with verbs at the third person singular because they reflect less actively on the language: they perhaps take for granted their ability to correctly express verbs at the present tense, since it is something which is learnt at a relatively early stage of L2 acquisition.

In this regard, it is also important to mention that the German learners at the C2 proficiency level make more self-corrections than the Spanish learners (Figure 15). According to the distinction between "error" and "mistake" made in the field of Error Analysis, the errors that are self-corrected should actually be considered as mistakes. This would imply that the German learners may possess a better knowledge of the concord rule, because mistakes are generally the result of a slip of the tongue and should not be associated with a lack of competence.

In summary, some differences between the Spanish and the German learners' linguistic production have arisen. Of these, the major one is that the Spanish learners produce more errors than the German learners, thus implying that the German learners may encounter fewer difficulties when learning and using the subject-verb concord rule in English. One of the factors that might be responsible for this discrepancy could be the different L1 background of the learners. It may be useful to consider that Spanish is a Romance language, while German and English are both Germanic languages: the relationship between German and English is hence closer than the one between Spanish and English. Further observations regarding the omission of the third person *-s* will be explored in the Conclusions section of this dissertation.

CONCLUSIONS

The study presented in this dissertation has investigated the presence of errors regarding subject-verb concord in the linguistic samples of Spanish and German L2 learners of English, which were extracted from the COREFL learner corpus. The results discussed in the third chapter point towards a discrepancy between the two groups of learners: the Spanish learners seem to produce more errors than the German learners.

Some of the theories stemming from the field of Second Language Acquisition, as described in the first chapter of this dissertation, provided the theoretical framework for the study. These theories will now be used to further comment on the findings.

It is possible to look at the subject-verb concord error from an Error Analysis perspective. This type of error seems to be the result of misinformation, which is the use of the wrong grammatical form, and of overgeneralisation. About overgeneralisation, Richards (1971, in Richards, 1974, p. 174) states that "[w]ith the omission of the third person -*s*, over-generalization removes the necessity for concord, thus relieving the learner of considerable effort." Dušková (1969, in Richards, 1974, p. 174) goes into more detail and makes this point clearer:

> Since (in English) all grammatical persons take the same zero verbal ending except the third person singular in the present tense […] omissions of the -*s* in the third person singular may be accounted for by the heavy pressure of all other endingless forms. The endingless form is generalized for all persons, just as the form *was* is generalized for all persons and both numbers in the past tense.

In other words, the third person -*s* is an exception to the norm: the third person singular in the present tense does not take zero verbal ending like the other grammatical persons. It can be assumed that this exception may confuse the learners, therefore making it plausible that they may choose to avoid using the third person –*s* entirely, especially if the subject-verb concord rule is not very clear to them. Overgeneralisation of the "endingless" forms could thus be considered as one of the causes behind the production of this specific type of error.

The results obtained in this study also suggest that the hypothesis put forward by Dulay and Burt (1974) about the acquisition of an L2 may not be universally true. In particular, the findings of the Morpheme Order Studies (Dulay and Burt, 1973, 1974)

led the authors to hypothesise that the learners' L1 background does not drive the process of acquisition of the L2, because they found that the order of acquisition of a set of English morphemes does not change in L2 learners of English with different L1 backgrounds. However, the findings obtained in this study do not correspond to those obtained by Dulay and Burt: there is a clear discrepancy between the quantity of errors produced by the Spanish learners and those produced by the German learners, which would lead to believe that the L1 background is likely to play a part, at least in the acquisition of the *-s* morpheme of the third person singular in English. Nevertheless, the study carried out in this dissertation does not reflect an accurate picture of the situation because of its limitations, as discussed below. Furthermore, the third person *-s* was the only morpheme investigated in this study, so these results should not be compared too strictly to the hypothesis by Dulay and Burt (1974).

As regards the limitations of this study, it must first of all be noticed that the sample of errors extracted from the corpus does not contain all the instances of errors concerning the omission of the third person *–s*, but only the errors preceded by either *he, she* or *it.* The method adopted thus allowed me to retrieve only some of the errors of this type in the corpus. Another limitation is related to the difficulty in determining whether the learners, in some ambiguous cases, meant to express the verb in the past tense instead of the third person of the present tense. This means that some of the errors included in the sample may not be an expression of the omission of the third person *-s*, but might in fact represent a failed attempt on the part of the learners to express the past tense form of the verb. An example of this can be found in the following extract:

(38) […] he had been found it just the moment / in which / a policeman / appears / when *he give* the baby who was walking with their / and this man *leave* / the baby […].

In this extract, there are two errors (i.e. *he give* and *leave*). It is not very straightforward to determine whether the learner meant to express the past tense form of the verb or the third person of the present tense.

These two limitations inevitably affect the accuracy of the results obtained from this study. The first limitation could be overcome by elaborating a better (ideally automatized) way of identifying the errors in the corpus, which would enable the researcher to find all the instances of omission of the third person *-s*. With regard to the

second limitation, discriminating between cases of omission of the third person *-s* and cases of the present tense used instead of the past tense would be more challenging and would probably still require human intervention. A sophisticated tagging system would be ideal to distinguish between the two types of errors used, yet I believe that checking and interpreting errors with reference to the linguistic contexts in which they are used is of the paramount importance.

The impossibility of conducting a contrastive analysis on the errors is another limitation of this study. In order to examine the errors through a contrastive approach, it would be necessary to have some knowledge of all the languages involved in the study (i.e. Spanish, German and English). This approach would allow one to check for the presence of transfer phenomena in the production of the errors. The results of this analysis would arguably provide better insights into the possible causes behind the production of these errors, especially into the discrepancies between the findings obtained from the Spanish and the German learners' samples.

Regardless of the aforementioned weaknesses of this study, the presence of errors concerning the omission of the third person *-s* which emerged in this dissertation indicates that Spanish and German L2 learners of English have some difficulties in correctly using the subject-verb concord rule. The subject-verb concord rule is normally taught in the early stages of L2 English courses, and teachers probably assume that such an apparently basic concept is easily assimilated by the students. However, the findings of this study suggest otherwise, and teachers should not underestimate the challenges that students may face while actually producing texts. It may be useful for teachers to spend more time on this issue and encourage students to practice speaking and writing in English as much as possible.

By looking at the findings presented in the third chapter of this dissertation, it can be safely assumed that the L1 background of the learners does have some impact on the production of errors concerning the omission of the third person *-s*: the great majority of errors was found in the Spanish learners' samples. Because of this, it cannot be excluded that students with different L1 backgrounds might require different teaching and acquisitional approaches. Further studies in this direction may shed some light on this issue and hopefully provide more effective teaching methods tailored specifically for L2 learners of English with different L1 background.

REFERENCES

Bailey, N., Madden, C., & Krashen, S. (1974). Is there a "natural sequence" in adult second language learning? *Language Learning*, *21*, 235-243.

Biber D., Conrad S., & Leech G. (2002). *Longman Student Grammar of Spoken and Written English*. Harlow: Longman.

Brown, R. (1973). *A first language: the early stages*. Cambridge: Harvard University Press.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 35-55.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Corder, S. P. (1967). The Significance of Learners' Errors. *International Review of Applied Linguistics*, *5*(4), 161–170.

Corder, S. P. (1974). Error Analysis. In J. Allen & S. Corder (Eds.), *The Edinburgh Course in Applied Linguistics Volume 3: Techniques in Applied Linguistics*. Oxford: Oxford University Press, 122-154.

Díez-Bedmar, M. B. (2021). Error Analysis. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora*. London and New York: Routledge, 90-104.

Dulay, H. C., & Burt, M. K. (1973). Should we teach children syntax? *Language Learning*, *23*(2), 243–258.

Dulay, H. C., & Burt, M. K. (1974). Natural sequences in child second language acquisition. *Language Learning*, *24*, 37–53.

Dušková, L. (1969). On Sources of Errors in Foreign Language Learning. *International Review of Applied Linguistics*, *7*, 11–36.

Ellis, R. (1997). *Second Language Acquisition*. Oxford: Oxford University Press.

Ellis, R. (2015). *Understanding Second Language Acquisition* (2nd ed.). Oxford: Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford: Oxford University Press.

Falinski, J. (2008). *English Verbs: A Practical Grammar*. Padova: CLEUP.

Gass, S. M. (2013). *Second Language Acquisition: An Introductory Course* (4th ed.). New York: Routledge.

Granger, S. (2012). How to Use Foreign and Second Language Learner Corpora. In A. Mackey & S. M. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide*. West Sussex: John Wiley and Sons, 7-29.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7-24.

Howatt, A. P. R. (2004). *A history of English language teaching* (2nd ed.). Oxford: Oxford University Press.

Huddleston, R., & Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Krashen, S. (1977). The Monitor Model for adult second language performance. In M. Burt, H. Dulay, & M. Finocchiaro (Eds.), *Viewpoints on English as a Second Language*. New York: Regents Publishing, 152–161.

Krashen, S. (1978). The Monitor Model for second language acquisition. In R. C. Gingras (Ed.), *Second Language Acquisition and Foreign Language Teaching*. Arlington: Center for Applied Linguistics, 1-26.

Lado, R. (1957). *Linguistics Across Cultures: Applied Linguistics for Language Teachers*. Ann Arbor: University of Michigan.

Lightbown, P. M., & Spada, N. (2013). *How Languages Are Learned* (4th ed.). Oxford: Oxford University Press.

Lozano, C. (2022). CEDEL2: Design, compilation and web interface of an online corpus for L2 Spanish acquisition research. *Second Language Research*, *38*(4), 965–983.

Lozano, C., Díaz-Negrillo, A., & Callies, M. (2020). Designing and compiling a learner corpus of written and spoken narratives: COREFL. In C. Bongartz & J. Torregrossa (Eds.), *What's in a Narrative? Variation in Storytelling at the Interface Between Language and Literacy*. Berlin: Peter Lang, 21–46.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Meunier, F. (2021). Introduction to Learner Corpus Research. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora.* London and New York: Routledge, 23-36.

Mitchell, R., Myles, F., & Marsden, E. (2012). *Second Language Learning Theories* (3rd ed.). London and New York: Routledge.

Richards, J. C. (1971). A Non-Contrastive Approach to Error Analysis. *English Language Teaching*, *25*(3), 204-219.

Richards, J. C. (1974). *Error Analysis: Perspectives on Second Language Acquisition*. London: Longman.

Saville-Troike, M. (2012). *Introducing Second Language Acquisition* (2nd ed.). Cambridge: Cambridge University Press.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*, 219–231.

Shachter, J. (1974). An error in error analysis. *Language Learning*, *27*, 205-214.

Towell, R., & Hawkins, R. (1994). *Approaches to Second Language Acquisition*. Clevedon: Multilingual Matters.

SUMMARY IN ITALIAN

In questo studio si andranno ad investigare alcuni errori commessi da parlanti nativi spagnoli e tedeschi apprendenti d'inglese come seconda lingua. I dati utilizzati per svolgere lo studio sono stati estratti dal COREFL, un vasto learner corpus elettronico contenente testi prodotti in inglese come seconda lingua da apprendenti spagnoli e tedeschi.

Gli errori investigati riguardano l'omissione del morfema *-s* della terza persona singolare coniugata al tempo presente in inglese. Questi errori saranno esaminati prendendo in considerazione alcuni dei modelli teorici proposti nel campo della *Second Language Acquisition*, ovvero lo studio dell'apprendimento di una seconda lingua.

Di particolare ispirazione per lo svolgimento del presente studio sono stati i risultati riportati dai *Morpheme Order Studies* (Dulay e Burt, 1973, 1974), ovvero degli studi sull'ordine di acquisizione di alcuni morfemi grammaticali in apprendenti di inglese L2. Dai risultati è emerso che il morfema *-s* della terza persona singolare è uno degli ultimi ad essere appreso dagli studenti.

Lo scopo principale dello studio è quindi di comparare dal punto di vista quantitativo gli errori commessi dagli apprendenti spagnoli e tedeschi, oltre a verificare l'esistenza di eventuali pattern formati dagli errori.

Il primo capitolo della tesi fornisce un'introduzione generale sulla disciplina della *Second Language Acquisition* (SLA), soffermandosi in particolare sui concetti più pertinenti allo studio presentato in questa tesi.

Secondo la definizione data da Ellis (1997, p. 3), la SLA è lo studio sistematico di come le persone apprendono una seconda lingua. Con il termine seconda lingua (L2) non si intende letteralmente solo la seconda lingua appresa, bensì una qualunque lingua imparata dopo la madrelingua (L1) (Saville-Troike, 2012, p. 2).

La SLA è un campo di studio dal carattere fortemente interdisciplinare con contribuzioni provenienti da svariate discipline, tra cui la linguistica, la psicologia e materie affini a queste (Saville-Troike, 2012, pp. 2–3). Questa interdisciplinarità permette di guardare al problema dell'apprendimento della seconda lingua da molte prospettive diverse.

Una delle aree più investigate nel campo della SLA fin dai suoi inizi è il ruolo che la prima lingua (L1) assume nel processo di apprendimento di una L2 (Gass, 2013, p. 79). Tra i fenomeni più noti riguardo al ruolo della L1 nell'apprendimento di una L2 si trova il *transfer*. Questo fenomeno fa riferimento all'influenza che la L1 può esercitare sull'apprendimento di una L2. Nello specifico, un fenomeno di transfer viene considerato positivo se l'uso di una struttura tipica della L1 risulta appropriato nella L2, mentre il transfer è negativo se quest'uso genera frasi considerate come agrammaticali o inappropriate nella L2 (Saville-Troike, 2012, p. 19).

Per quanto riguarda la storia della SLA, è nata come disciplina intorno alla metà del Novecento. In quegli anni, le idee dello strutturalismo in linguistica e del comportamentismo in psicologia si sono poste alla base della nascita dell'analisi contrastiva (Lado, 1957, in Ellis, 2015). Questa ipotesi prevede il confronto della L1 del parlante con la L2 che intende apprendere, così da prevedere gli ostacoli e le agevolazioni che caratterizzeranno il percorso di apprendimento della L2. Da questa definizione emerge che l'analisi contrastiva attribuisce un'importanza centrale ai fenomeni di transfer linguistici (Gass, 2013, p. 86).

Tuttavia, negli anni Sessanta del Novecento sono stati trovati diversi difetti nelle proposte dell'analisi contrastiva. In particolare, le critiche mosse dal linguista americano Noam Chomsky hanno introdotto una nuova prospettiva sull'acquisizione del linguaggio, la cui influenza persiste tutt'ora nel campo della linguistica. Chomsky, infatti, ha elaborato la teoria della Grammatica Universale o Generativa (1957, 1965, in Saville-Troike, 2012), secondo cui il linguaggio è regolato da principi innati, comuni a tutte le lingue del mondo.

Su queste basi, varie teorie sono emerse nel campo della SLA. Un importante saggio pubblicato da Corder nel 1967 ha permesso la nascita dell'analisi degli errori (*Error Analysis*), ovvero un nuovo approccio allo studio dell'apprendimento di una L2. Questa teoria considera centrali gli errori commessi dai parlanti: gli errori vengono infatti interpretati come una manifestazione del sistema linguistico e delle strategie di cui il parlante si sta servendo in quel particolare momento del suo percorso di apprendimento della L2 (Corder, 1967). In altre parole, secondo Corder gli errori sono una dimostrazione dell'esistenza di un processo di apprendimento.

Nella sua teoria sull'analisi degli errori, Corder stabilisce una importante distinzione tra errori di esecuzione, chiamati *mistakes*, ed errori di competenza, chiamati *errors*. I primi vengono solitamente scartati dall'analisi, in quanto corrispondono a lapsus e non rispecchiano l'effettiva competenza del parlante. I secondi sono invece al fulcro dell'analisi: sono errori prodotti con sistematicità dal parlante, che quindi indicano la presenza di lacune nella sua conoscenza linguistica.

Sulle orme di Corder, Selinker (1972) ha sviluppato il concetto di interlingua, impiegato tutt'oggi nel campo della SLA. Un'interlingua corrisponde ad un sistema linguistico basato sulla produzione di un parlante in una determinata fase del suo percorso di apprendimento di una L2.

Negli anni Settanta, prendendo spunto dai vari studi sull'acquisizione della prima lingua nei bambini, sono emersi i primi studi sull'acquisizione della seconda lingua, ovvero gli studi sull'ordine dei morfemi (*Morpheme Order Studies*) condotti da Dulay e Burt (1973, 1974). In questi studi, sono stati somministrati dei test a dei bambini con L1 diverse apprendenti di inglese L2 con lo scopo di verificare l'ordine di acquisizione di una serie di morfemi inglesi. I risultati hanno riportato un ordine di acquisizione quasi identico per tutti i bambini, portando Dulay e Burt (1974, p. 52) a postulare l'esistenza di meccanismi cognitivi universali alla base del processo di acquisizione della seconda lingua, escludendo quindi un possibile ruolo centrale della L1.

Gli studi sull'apprendimento della seconda lingua hanno mosso un ulteriore passo in avanti negli anni Ottanta con la nascita della *Learner Corpus Research,* ovvero lo studio di corpora contenenti la produzione linguistica di apprendenti L2. Un corpus è una raccolta di testi autentici rappresentativi di una specifica lingua o varietà linguistica (McEnery et al., 2006, p. 5, in Meunier, 2021, p. 23). Questo approccio comporta vari benefici, come la possibilità di condurre analisi utilizzando un computer e l'opportunità di lavorare su campioni di grandi dimensioni.

Il secondo capitolo della tesi offre una descrizione dettagliata del corpus impiegato nello studio e dei metodi adottati per condurre l'analisi degli errori.

Il COREFL (*The Corpus of English as a Foreign Language*) è un learner corpus elettronico composto specialmente da testi, scritti e orali, prodotti da apprendenti spagnoli e tedeschi di inglese L2 (Lozano et al., 2020). Il corpus contiene inoltre due corpora di riferimento di spagnolo L1 ed inglese L1. I documenti presenti nel corpus

sono 2.447, di cui 1.810 prodotti da apprendenti di inglese L2. Di questi, 1.361 sono stati prodotti dagli spagnoli apprendenti di inglese L2, mentre 449 dai tedeschi. Il numero totale di token presenti nel corpus è 530.392. È possibile navigare liberamente il corpus al seguente indirizzo web: corefl.learnercorpora.com.

Per raccogliere i dati che compongono il corpus, ai partecipanti è stato chiesto di svolgere un'attività tra quattro proposte, che consistono nella produzione di un testo di carattere narrativo.

I ricercatori hanno inoltre raccolto una quantità considerevole di informazioni dai partecipanti, che costituiscono i metadati dei testi, accessibili tramite il sito. I testi sono anche stati sottoposti ad annotazione grammaticale automatica (*POS tagging*), ovvero un software ha analizzato ogni parola del corpus e le ha assegnato un'etichetta che ne descrive la categoria linguistica di appartenenza. Ovviamente questo metodo non è totalmente affidabile, specialmente nell'annotazione di errori prodotti dai partecipanti.

La presenza di dati annotati permette di svolgere ricerche anche avanzate nel corpus. L'interfaccia web di COREFL è dotata di due motori di ricerca interni, uno per ricerche semplici ed uno per ricerche avanzate. Il primo è ideale per verificare l'occorrenza di una o più parole nel corpus, o solo in parte di esso, mentre il secondo offre la possibilità di personalizzare molti più parametri di ricerca. Per esempio, gli errori impiegati in questo studio sono stati estratti cercando i pronomi personali della terza persona singolare, ossia *he, she* e *it*, insieme a parole etichettate come verbi al presente distinti dalla terza persona singolare, ottenendo quindi errori come *he sing*, *she eat*. I risultati della ricerca sono stati controllati manualmente per eliminare quelli non pertinenti ai parametri di ricerca. Una volta estratti, gli errori sono stati trasferiti sul programma Microsoft Excel. Questo ha permesso di semplificare le operazioni di calcolo e di filtrare gli errori in base ad alcune variabili, come il livello di competenza linguistica dell'apprendente, la sua L1 ed altre.

Nel terzo ed ultimo capitolo della tesi, i risultati delle analisi sugli errori vengono riportati e discussi. Gli errori raccolti sono stati classificati in base a diverse variabili: il tipo di verbo che contengono, ovvero verbi stativi, dinamici o ausiliari e le loro rispettive sottocategorie; il livello di competenza degli apprendenti dal B1 al C2; il mezzo scritto o orale della produzione e le autocorrezioni, ossia la presenza della forma corretta subito dopo l'errore. I conteggi degli errori raggruppati in ognuna delle categorie sono stati

divisi in base alla L1 dei partecipanti, quindi spagnolo e tedesco. Le frequenze sono poi state normalizzate in base al numero di token presenti nelle rispettive categorie in modo da ottenere dati comparabili. Tutti questi dati sono visualizzati tramite l'uso di tabelle e grafici commentati. Inoltre, ogni categoria di errori è accompagnata da una serie di estratti dal corpus per esemplificarne la natura.

Le frequenze normalizzate del totale degli errori mostrano che i partecipanti spagnoli hanno commesso molti più errori rispetto ai partecipanti tedeschi: le frequenze normalizzate ottenute sono rispettivamente 2,11 e 0,25. Questa tendenza è confermata in quasi tutte le categorie analizzate, ad eccezione di alcuni livelli di competenza e le autocorrezioni. Nel primo caso, le frequenze normalizzate sono comunque superiori nel campione spagnolo, ma ai livelli C1 e C2 queste diminuiscono, avvicinandosi alle frequenze calcolate per il campione tedesco. Nel secondo caso, ai livelli B2 e C1 le frequenze normalizzate delle autocorrezioni sono quasi identiche, mentre al livello C2 non sono state trovate autocorrezioni nel campione spagnolo e la frequenza normalizzata del campione tedesco è superiore a quella trovata ai livelli B2 e C1.

La principale conclusione che emerge dai risultati è sicuramente la discrepanza tra gli errori commessi da tedeschi e spagnoli. Una delle possibili cause di questo fenomeno potrebbe essere attribuita alla diversità delle L1: il tedesco e l'inglese sono lingue più vicine essendo entrambe lingue germaniche, mentre lo spagnolo è una lingua romanza. Dal punto di vista teorico, questi risultati sembrerebbero andare contro l'ipotesi avanzata da Dulay e Burt (1974) secondo cui l'ordine di acquisizione di una L2 non è influenzato particolarmente dalla L1 degli apprendenti. Tuttavia, il campione di dati utilizzato è troppo limitato per poter avanzare con certezza tale ipotesi. Senza considerare la differenza tra spagnoli e tedeschi, la produzione di errori riguardanti l'omissione del morfema *-s* dalla terza persona singolare potrebbe essere il risultato di un processo di sovrageneralizzazione. Questo termine deriva dal campo dell'analisi degli errori e in questo caso indica che l'omissione della *-s* avvenga perché in tal modo gli apprendenti evitano lo sforzo di accordare verbo e soggetto, operazione necessaria in inglese solamente per la terza persona singolare al presente (Richards, 1971, in Richards, 1974, p. 174).

In ogni caso, ulteriori studi di maggiore scala su questo fenomeno sarebbero indispensabili, soprattutto per proporre metodi di insegnamento mirati a facilitare

l’apprendimento da parte degli studenti del morfema -*s* della terza persona singolare. In particolare, metodi di insegnamento specifici in base alla L1 degli apprendenti potrebbero rivelarsi ancora più efficaci.