
**DEPARTMENT OF PHILOSOPHY,**
**LINGUISTICS AND THEORY OF SCIENCE**

# WHO'S AFRAID OF COMPLEXITY?

## An Exploration of the Influence of Native Language Complexity on L2 Complexity

**Nadina Mariana Suditu**

| | |
|---|---|
| Master's Thesis: | 30 credits |
| Programme: | Master's Programme in Language Technology |
| Level: | Advanced level |
| Semester and year: | Spring, 2023 |
| Supervisor: | Aleksandrs Berdicevskis, Elena Volodina |
| Examiner: | Staffan Larsson |
| Keywords: | complexity, SLA, TTR, entropy, kolmogorov, linear regression |

# Abstract

The matter of linguistic complexity has been widely scrutinised in the last few decades, within theoretical linguistics, as well as in second language acquisition studies. A concept introduced in the last half of the previous century, it continues to be a matter of debate in the linguistic field, as it eludes a clear-cut definition and interpretation. In this thesis, the morphological and lexical complexity of learner essays in Swedish are calculated using type-token, entropy and Kolmogorov complexity measures in order to determine whether properties of the native language (L1) of the learner influence the complexity of the second language (L2). A supplementary measure is devised by aggregating morphological features pertaining to each language as taken from the World Atlas of Language Structures. The languages under analysis are: Arabic, Chinese, Dari, English, German, Persian, Spanish and Swedish. The measures are computed on both the L1s and L2s, which are afterwards analysed with the help of data visualisations and linear regression models. This study concludes that the complexity of the L1 is not a reliable predictor of L2 complexity, although it does act as a predictor regarding the instructional level of the learners.

# Preface

I would like to thank my supervisors, Sasha Berdicevskis and Elena Volodina, for their support throughout this process and for providing insight without which this study would probably not have taken shape as well as it has. Even though taking on a completely new (to me) area of research has been intimidating at times, the encouragement through good results and bad results has been one of the best lifelines I could have had these past few months.

I would also like to thank all of my friends alongside whom I have been working on the thesis, early mornings to late nights. Staying in Humanisten until 11 PM would not have been twice as fun without all the crossword breaks (although maybe it would have been more productive). To all my friends and family from far away, I'm thankful for all the patience and understanding as my response to almost every call and text in the last five months has been "sorry, I'm at the library".

Lastly, I'm of course grateful to the whole MLT program, for everything I have learned here, and for allowing me to confidently say now "Yes, I'm more of a humanities student, but I also know how to code".

# Contents

# 1 Introduction

Linguistic complexity has been an emerging field of study within second language acquisition (SLA) in the past decades, with a number of debates surrounding it. If for a long time the theoretical linguistics community has held the belief that all languages are equally complex (see Kusters, 2003; Sampson, 2009, as cited in Ehret et al., 2016), this consensus has been challenged increasingly in recent years. A big proponent for the theory that not all languages are equally complex was McWorther (2001), who argued that creole and younger languages are generally less complex than older language systems. This has led to an extensive interest in linguistic studies whose aim is to gauge the complexity of various languages, since the a significant number of researchers subscribe now to the notion that "human languages and dialects of the same language can and often do differ in their complexity" (Ehret et al. 2016:1). There is, though, a general lack of consensus revolving around complexity itself, regarding its definition and the way in which the concept is applied in linguistics, as well as regarding the measures that are used in order to assess it, all which will be discussed in the following chapter. Bulte and Housen (2012) argue that, "in order to have measures that tap into linguistic complexity in a meaningful and valid way" (p. 27) there is a need for a three dimensional approach: theoretical, observational and operational. According to them, the theoretical aspect is concerned with establishing clearly what complexity is, the observational with how complexity manifests itself "in actual language performance" (p. 27), and, finally, the operational aspect gauges the quantification of the before-mentioned manifestations. This thesis will focus on the notion of linguistic complexity, along with the external factors pertaining to the learners which can influence the complexity of emerging speech, such as age, gender, instructional level, and, most importantly, complexity of native language.

## 1.1 Research questions and scope

The present study focuses on an arguably minimally explored area of complexity in SLA. While there has been a considerable number of studies conducted on whether native language (L1) complexity or genetic distance affects the difficulty or pace at which learners acquire an additional language (such as Schepens et al., 2013, Schepens et al., 2017), there has been less interest directed toward how this affects the complexity of the second language (L2) produced by the learners. The choice of the L2 language submitted to analysis brings forth a somewhat novel angle as well, since Swedish has not been as extensively studied in this context as, for example, English or Dutch. As such, the main research question of this paper is centered on whether the complexity of the L1 influences the complexity of L2 Swedish as represented by learner essays, and to what degree this is an influential factor. A secondary aspect which will be taken into consideration as well is the influence of linguistic proximity on the matter of L2 complexity. The languages under investigation will be Arabic, Dari, English, German, Chinese, Persian, Spanish and Swedish, and their complexity will be assessed from the morphological and lexical viewpoint. The measures which will be used in this endeavor are of two natures: type-token ratio, word entropy and Kolmogorov complexity could be described as "distribution-based [approaches]" (Bentz et al. 2016:150), while an assessment of morphological complexity based on linguistic features as they are accounted in the World Atlas of Languages Structures (WALS) could be called a "paradigm-based approach" (p. 150).

## 1.2 Motivation and contributions

Part of the motivation behind this study stems from the ambition to provide insight within the field of theoretical and applied linguistics upon the ramifications of L2 complexity and, furthermore, to gain a better understanding on how these insights can prove beneficial for teachers and learners of second or additional languages. As purely theoretical and removed from real-life situations as this topic may seem, there have been studies investigating the impact that an L1 has on an immigrant's life and adaptation in another country. In a study by Helgertz (2010), it was examined whether the linguistic distance of an immigrant's L1 affects their position on the job market and social status in their transition to living in Sweden. It was found that immigrants fluent "in a language belonging to the Germanic language family at the time of migration" (p. 462) proved to be a significant advantage. This advantage is explained as consisting mainly of a "quicker pace in acquiring Swedish language skills" (p. 440). The complexity of the L1 was, although, not investigated, which adds further to the motivation of investigating this particular aspect.

# 2 Theoretical background and literature review

## 2.1 SLA and Interlanguage

It is essential to first introduce the notion of 'interlanguage' as the central focus of the current study revolves around this concept. In *The comparative fallacy in interlanguage studies: the case of systematicity*, Robert Bley-Vroman references the work of Selinker (1969, 1972), who in the late 1960s and early 70s consolidated the concept of an 'interlanguage' in second language acquisition (SLA) studies. Interlanguage, according to Selinker, would be "linguistically described using as data the observable output resulting from a speaker's attempt to produce foreign norm, i.e. both his errors and non-errors. It is assumed that such behavior is highly structured." (as cited in Bley-Vroman 1983:1). The way in which interlanguage and the L1 of a learner are interconnected is the belief held by specialists in the field that the complexity of the interlanguage depends not only on the level reached by the learner, but on the complexity of the target language as well (DeKeyser, 2016; Housen and Simoens, 2016; as cited in Brezina and Palotti, 2015). This statement forms the base of the current study, which aims to investigate how the complexity of an interlanguage, namely L2 Swedish at various preparatory levels, might be affected by the L1 of the learner.

## 2.2 Complexity, Accuracy, Fluency

In the last few decades, CAF (Complexity, Accuracy, Fluency) has been employed at the center of multiple second language and language acquisition studies (Pallotti, 2009). In their paper, *Complexity, Accuracy and Fluency,* Housen, Kuiken and Vedder (2012) produce a comprehensive overview of CAF, writing that generally, L2 proficiency is not "a unitary construct" (p. 1), but rather composed of multiple variables, namely complexity, accuracy and fluency. Originally, the concept of CAF emerged separately in the 1970s, as part of efforts in the field to accurately research proficiency levels in L2, and the three components were brought together for the first time in the middle 1990s, with a proficiency model introduced by Skehan (1996, 1998; as cited in Housen, Kuiken and Vedder, 2012). Generally, the three concepts are described in the following way:

*Complexity* is the ability to use a wide and varied range of sophisticated structures and vocabulary, as well as the extent to which a learner shows an inclination to be elaborate and adventurous in their use of the language (Bui & Skehan 2018).

*Accuracy* represents the ability to produce target-like and error-free language, arguably doubling in meaning as *correctness*, with the focus on how closely the learner can follow the rule system of the target language (Bui and Skehan, 2018).

*Fluency* represents the learner's proficiency and it focuses on the ability to produce the L2 with native-like rapidity, pausing, hesitation and reformulation (Housen, Kuiken and Vedder, 2012)

Ever since their introduction and consolidation in the speciality field, these three variables have been examined to study a wide range of aspects relating to second or additional language attainment , such as "the effects of age [...], the effects of instruction, of individual differences, the effects of learning context or of task design" (e.g. Bygate 1996, 1999; Collentine 2004; De Graaff 1997; Derwing & Rossiter 2003; Foster & Skehan 1996; Fotos 1993; Freed 1995; Mora 2006; Robinson 2011; Norris & Ortega 2000; Yuan & Ellis 2003, as cited in Housen, Kuiken and Vedder, 2012).

Several studies emphasize the importance of considering all three variables together to ensure the reliability and value of research findings (Norris & Ortega 2009; Ortega 1995; Skehan & Foster 1997, 2001 as cited in Housen, Kuiken and Vedder, 2012). This claim is backed by the hypothesis that the three components represent successive stages of a learner's developmental journey(Housen, Kuiken and Vedder, 2012). It is argued that the way in which complexity, accuracy and fluency are interconnected is cyclical, following the order of complexity as the initial stage, leading into accuracy, and ultimately to fluency. The thought process behind this is that at the initial stage, learners would be internalising new and more complex language structures, which then are modified according to base norms, leading to greater accuracy, followed by a consolidation of these acquired structures, and finally fluency (Housen, Kuiken and Vedder, 2012). However, the authors themselves admit that this theory is mainly speculative and rather simplistic, a sentiment shared by other researchers such as Pallotti.

Within the scope of this paper, the primary focus will be on examining complexity. As such, there is no need for anything more than a brief account on how accuracy and fluency have been viewed and addressed in the field. According to Housen, Kuiken and Vedder (2012), accuracy is often considered to be the most "straightforward" component of CAF, considering the fact that, conceptually, it strives to determine how far the learner's performance deviates from the norm of the language. Notwithstanding, this can prove problematic as well, since there is room for debate in the case of the criteria chosen for identifying deviations (Housen, Kuiken and Vedder, 2012), as well as pinpointing what an error is defined as, and whether certain elements of the produced speech are more erroneous than others (Housen, Kuiken and Vedder, 2012). On the other hand, fluency is less universally agreed upon, however, it can be generally described as the overall proficiency and 'smoothness' (Housen, Kuiken and Vedder, 2012) of a learner's speech. Furthermore, fluency is a phenomenon which manifests itself mainly at the phonological level (Housen, Kuiken and Vedder, 2012).

## 2.3 Complexity

The one aspect that is generally agreed upon about complexity is the fact that it might be the most 'controversial' element of CAF. (Housen Kuiken and Vedder, 2012; Pallotti, 2009; Bulte and Housen, 2012), with studies within the literature oftentimes are producing inconsistent or downright contradictory results (cf. Robinson 2007; Skehan 2009; Spada & Tomita 2010, as cited in Bulte and Housen, 2012), stemming from the inconsistencies in the definition of the term itself, which, according to Bulte and Housen, is frequently done in "general, vague or even circular terms" (p. 22). The meaning attributed to complexity varies widely not only from study to study, but sometimes even within singular studies (Housen, Kuiken and Vedder, 2012).

The definition of the term *complexity* is no less complex than the notion which it intends to depict. Inherently, the terms 'complexity' and 'complex' are multifaceted in themselves, with definitions ranging from "composed of two or more parts" (Pallotti, 2009:593), to the measure of the level of variety and "the existence of multiple alternatives" (p. 593), to "difficult, cognitively demanding" (p. 593) and, finally, simply holding the meaning of 'acquired late' (p. 593). As to whether the moment of acquisition plays a role when it comes to complexity, Pallotti believes it is a debatable issue, arguing that it might be the case of certain linguistic structures appearing in later speech because of infrequency or relevancy, not because of how complex they are (Hulstijn, 1995; as cited in Pallotti, 2009). Thus, it is crucial to make a clear separation between the notions of complexity and those of "progress" or "development" (Pallotti 2009). This perspective is especially relevant in the context of the current study, since, moving forward, learners' L2 complexity will be analysed in relation to some key background elements, one of them being the level at which the language had been produced. At surface level, it would seem sensible to deduce that, if the complexity appears to be greater at higher instructional levels, it means that complexity is monotonic with time. However, it might be the case that certain language structures which are more complex are not introduced to the learner via learning materials until higher instructional levels, consequently causing the learners to use them from the point of introduction forward.

The dilemma surrounding complexity stems to an extent from the fact that the term has been used to refer to two different concepts, which have been designated multiple terminology pairings, such as: linguistic complexity and cognitive complexity (DeKeyser, 1998; Housen & Kuiken, 2009; Housen, Van Daele & Pierrard, 2005; Williams & Evans, 1998 as cited in Housen, Kuiken and Vedder, 2012), or relative and absolute complexity, respectively. The confusion is further intensified by the fact that, sometimes, the two concepts are used interchangeably in the SLA literature (Housen, Kuiken and Vedder, 2012).

Cognitive, or relative complexity, is the terminology used to refer to the concept of linguistic complexity as affected by external factors pertaining to the learners or learning conditions. Included among these factors are: L1, age, aptitude, memory capacity, motivation, among others (Kuiken, 2022). Simply put, it "defines complexity in relation to language users" (Bulte and Housen, 2012:23).

The main objective when studying cognitive complexity would be, then, determining how the acquisition of language "elements" (Housen, Kuiken and Vedder, 2012) is influenced by these factors. For instance, certain structures have been proved to be more difficult to process than others, such as relative clauses or passives, in opposition to coordinate and active structures (Byrnes & Sinicrope 2008; Diessel 2004, as cited in Bulte and Housen 2012). These structures aren't independent of extraneous factors, and their level of difficulty greatly varies from learner to learner is susceptible to a variety of background variables (Kuiken 2022). Therefore, when referring to relative complexity, the more taxing a language is to its learners and users, and the more effort and resources are spent when trying to learn and acquire the specific linguistic structures, the more 'complex' a language becomes. In this context, complexity is interchangeable with *difficulty* (Pallotti 2009).

The absolute approach, on the other hand, defines language complexity in objective, quantitative terms as the number of discrete components that a language feature or a language system consists of, and as the number of connections between the different components (Bulte and Housen, 2012). Therefore, linguistic or absolute complexity could be described, as the half of the coin concerned with the purely structural

properties of the language, independent of the learner (Housen, Kuiken and Vedder), relating to the "intrinsic formal or semantic-functional properties of L2 elements (e.g. forms, meanings, and form-meaning mappings) or to properties of (sub-)systems of L2 elements." (p. 4). It is claimed that absolute complexity is 'real' complexity (Bulté and Housen, 2012; Palotti 2009). Pallotti prefers the term 'structural' as opposed to 'objective' or 'absolute' complexity, to try and steer clear of misinterpretations or further theoretical connotations and focus only on the structural aspect of the language.

Juola (2008) presents a distinctive view on the types of complexity, which is that "the information relevant to a text can be broken down into four major categories:

- the complexity of the idea(s) conveyed;
- the complexity of the author's style;
- the complexity mandated by the language in which the author writes;
- the shared information omitted between the author and her audience"

He claims the third type of complexity, which is attached to the linguistic constraints only, is what is meant by "linguistic complexity". There are further distinctions to be made regarding the concept of linguistic complexity in the literature, as can be observed in Figure 1. It has been depicted either as global/system complexity, or structural complexity. Global complexity can be described as a "dynamic property of the learner's L2 system at large" (Bulte and Housen, 2012:25), whereas local or structural complexity refers to "a more stable property of the individual linguistic items, structures or rules that make up the learner's L2 system" (p. 25).

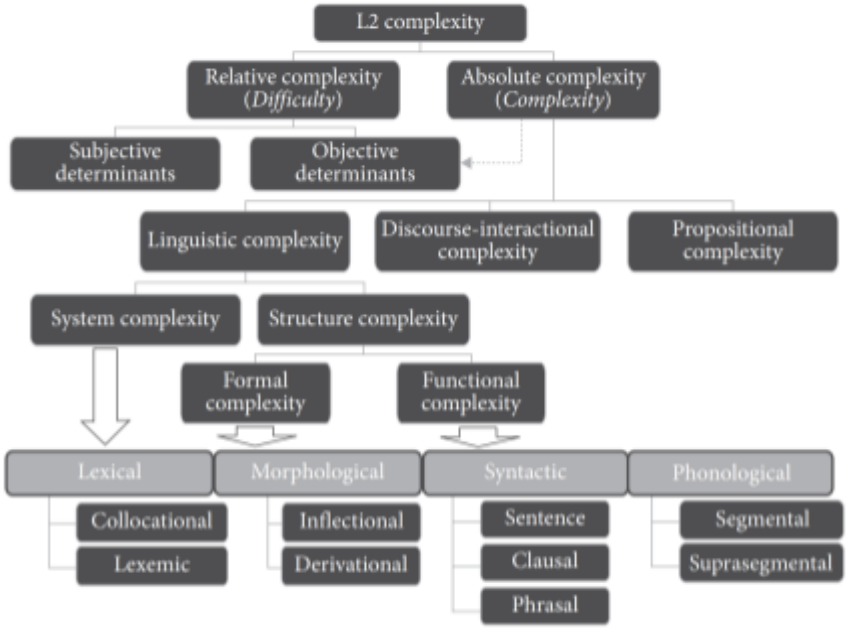


Figure 1: Taxonomy of linguistic complexity constructs, from Bulte and Housen, 2012

Broadly, two major components of 'linguistic complexity' can be identified: grammatical complexity and lexical complexity. The former is additionally split into two overarching categories: syntactic and morphological, "each further divisible into even smaller and finer-grained subcomponents." (p. 27). The latter is divided into three subordinating categories: density, diversity, and sophistication. Briefly, the three components can be described as follows: density represents the proportion of lexical words, diversity the number of different words, and sophistication the number of less frequent words (Kuiken, 2022).

## 2.3 Complexity Measures

The main forms of complexity being analysed for this study are morphological and lexical complexity, mainly lexical diversity and sophistication. According to Ortega (2012, as cited in Ehret et al. 2016), the quest for interlanguage complexity measures is guided by the following objectives: "(a) to gauge proficiency, (b) to describe performance, and (c) to benchmark development" (p.1).

It has been argued that it is in fact impossible to define the notion of overall complexity in an objective meaningful way (Deutscher, 2009), which is why many studies rely on measuring specific types of complexity individually, using dedicated measures, as the consensus seems to be that "measuring the complexity of subsystems of a language [...] seems less controversial" (Çöltekin and Rama, 2022:1). In line with the scope of this paper, measures which target syntactic complexity will not be mentioned and, instead, the focus will be directed towards those which evaluate lexical and morphological complexity, which often demonstrate interconnectedness or mutual influence.

It has often been remarked within the field that "the mere existence of morphology denotes complexity" (Anderson 2015; Carstairs-McCarthy 2010, as cited in Çöltekin and Rama, 2022:2). In Juola's view (1998), the endeavor of measuring morphological complexity has been motivated "by the fact that morphology is relatively straightforward and theory-independent, at least, in comparison to syntax" (Juola 1998).

In their paper, Çöltekin and Rama (2022) showcase two kinds of morphological complexity: *ennumerative* and *integrative*, notions first introduced by Ackerman and Malouf (2013). In their words, ennumerative complexity is "based on the number of morphosyntactic distinctions marked on words of a language" (p. 2), while integrative complexity refers to "the predictability of morphologically related words from each other" (p. 2). In the case of unannotated corpora, such as the ones which will be subjected to analysis in the current study, Çöltekin and Rama (2022) argue that a straightforward method of measuring complexity in such an instance can be "based on a measure of lexical diversity [...] since morphologically complex languages tend to include a larger number of word forms" (p. 4) and therefore exhibiting a higher degree of lexical diversity. Following this line of thought, another such common approach (p. 4) would be to measure the entropy of a text, since more complex morphologies lead to higher entropy scores, and vice-versa (Çöltekin and Rama, 2022).

Therefore, arguably two of the most prominently used measures of accounting for linguistic complexity, type-token ratio (TTR) and entropy make appearances in a considerable portion of the literature. These will be employed in the current study as well, along with Kolmogorov complexity.

### 2.3.1 Type-token Ratio

Type-token ratio is one of the most commonly used measures when it comes to linguistic complexity, namely diversity (Kuiken, 2022; Bulte and Housen, 2012), although it has been referred to as one of the simplest ones as well. It measures the ratio of total words to unique words in a text (Çöltekin and Rama, 2022) and, though it comes with its criticism (see Arnaud, 1986, Linnarud, 1986), it is generally deemed a straightforward measure (Çöltekin and Rama, 2022), with higher TTR indicating "rich or complex morphology" (p. 5). Even though by itself it doesn't lead to especially conclusive results, it has been found relevant alongside other measures, as well as correlating well with them (Bentz. et al. 2016).

Since this is a measure that highly depends and varies on text size, for this study it has been computed using a moving window average, as inspired by (Covington & McFall, 2010). In their paper, *Cutting the Gordian knot: The moving-average type-token ratio (MATTR)* the authors introduce a program which computes the TTR within a window size that can be chosen by the user, in the case of this study having the size of 100 tokens. The way in which this method works is that, depending on the window size, for example 500, the TTR is computed for words 1-500, then for words 2-501, 3-502 and so on, with the final measure being the mean average of all the previous TTRs.

### 2.3.2 Information theory

Before diving into more comprehensive overviews of entropy and Kolmogorov complexity, it is essential to first provide a backdrop of how these measures stem from information theory originally. In Juola's 2008 paper, *Assessing linguistic complexity*, he places linguistics within a framework highly based on information theory. With this as the foundation, the following scenario is used for demonstration: in a conversation there are two main parties, the speaker and the hearer, which exercise their 'economy' from distinct considerations. The speaker's economy "requires that he express messages in the most compact form possible" (p. 5), while the hearer's economy "requires that the speaker be easily understandable and thus that the amount of message reconstruction effort [...] be minimized" (p. 5). Therefore, both sides that partake in the communication process seek an optimization of the transmitted and received material, inasmuch as the speaker's message should consist only of the necessary information so as to be distinguishable without being redundant.

### 2.3.3 Entropy

With this as a base, we can transition to discussing Shannon entropy and Kolmogorov complexity as linguistic measures, starting with the former. Shannon, as Juola explains, has provided "a framework for mathematical analysis of the intuitive notion of language complexity" (p. 6).

Entropy is one of the more popular measures used for computing complexity, whether it be word, lemma, or character entropy. It can be described as "the average information content of a word in the text sample" (Çöltekin and Rama, 2022:6). In Shannon's definition, the entropy of a message source is the amount of information, typically measured in bits (yes/no questions), required to describe the successive messages emitted by that source to a recipient: "as the set of possible messages becomes larger, or the distribution of

messages becomes less predictable, the entropy of the source increases correspondingly." (Shannon, 1948 as cited in Juola, 1998). In Juola's words, the notion of predictability in a language "as well as the associated concepts of complexity, compressiveness, and randomness, can be mathematically remodelled using information entropy." (1998:142).

A complex morphology "creates many rare word forms, resulting in a longer tail of the word frequency distribution" (Çöltekin and Rama, 2022:6), hence less predictable words. On the other hand, a morphologically 'poor' language "typically uses more function words, resulting in more words with higher probabilities" (p. 6), which then leads to high predictability and lower entropy.

Entropy is, akin to TTR, sensitive to text size. For this study, once again, instead of truncating or splitting the sample texts in order to bring them to the same size, the choice has been made to compute using a moving window average, which will be described in greater detail in the next chapter.

It needs to be mentioned, however, that the above described measures have both had their fair share of criticism in the literature. TTR is accounted as being, by far, the simplest measure of this kind (Bentz et al., 2016), as it doesn't take into account "subtle differences in the distributions of word tokens over word types" (p. 149). Word entropy, although it performs better in this respect, does nonetheless not distinguish "between effects due to breadth of the base lexicon, on one hand, and word formation processes such as derivation, inflection or compounding" (p. 149). It is, additionally, not particularly reliable when it comes to reflecting "differences in regular and irregular morphological processes" (p. 149).

### 2.3.4 Kolmogorov Complexity

Kolmogorov complexity is closely linked to information theory, much as Shannon entropy (Ehret et al., 2016), with multiple studies in the literature having employed this measure (Ehret and Szmrecsanyi, 2016; Juola, 1998, 2008; Sadeniemi et al., 2008, as cited in Ehret et al., 2016). Fundamentally, this explores how the complexity of a system can be measured by the minimum length of a complete description of this system. In Juola's words (2008), "Shannon's entropy is an upper bound on (and asymptotically equal to) Kolmogorov complexity" (p. 8). However, Kolmogorov complexity measures "the information content of a string of symbols or text sample, not of a set of possible messages" (Li and Vitanyi 1997, as cited in Ehret et al. 2016:26). If the former measures the "informativeness" of a message source, the latter is applied in the case of a given string, "as the length of the algorithm required to describe/generate that string" (Juola, 2008:7).

Ehret et. al (2016) propose that this measure "bridges the gap between theoretical linguistics and applied linguistics" (p. 2) and, although it is said to be "formally uncomputable" (Kolmogorov, 1965; Li and Vitanyi, 1997, as cited in Juola 2008:8), it is feasible to compute under the premise that Kolmogorov complexity represents the ultimate possible file compression" (p.8), where "a good file compressor can be seen as an attempt to approximate this kind of complexity within a tractable formal framework." (p. 8). In a purely practical sense, the measure can define the complexity of a text as proportional to the length of the shortest algorithm that can generate that text (Brezina et al., 2016). Therefore, it is claimed to be less "reductionist" than other measures because it focuses on the predictability of future text based on "previously seen text" (Ehret et al, 2016:2). Kolmogorov complexity has also been described as being "a

text based, quantitative, holistic and global measure of structural surface redundancy" (p. 2), because it approximates complexity by measuring the amount of information and redundancy in a given text (Juola, 2008), using compression algorithms such as gzip (Ehret et al., 2016).

The final consensus over how it determines complexity is "related to the notion that more varied and/or diverse language should count as more complex (Bulté and Housen, 2014: 45, as cited in Ehret et al., 2016). Overall, "Comparatively larger scores indicate higher overall complexity of a text sample." (Ehret, 2016).

The current study will utilise the aforementioned measures by applying them on "non-parallel naturalistic texts", such as Ehret et al. (2016), as well as on parallel corpora, as in studies focused on bible translations, translations of *Alice in Wonderland* or the European Constitution (Ehret et al. 2016). While the main focus is on acquiring insight into L2 complexity by analysing corpora of collected learner essays, a secondary goal is comparing these results with the overall complexity of the learners' respective L1s. This feat will be achieved by applying the same measures on translations of the human rights declaration.

## 2.4 Cross linguistic influences, Linguistic Distance, Typology

In this section, aspects of linguistic typology, cross linguistic influences and linguistic distance will be briefly considered. Even though the main focus of the paper is not determining whether linguistic distance influences L2 complexity, it is an angle worth taking into account, as it has been demonstrated in the literature that it does exert an influence.

Linguistic typology is the field concerned with how languages resemble or differ one another, based on classifications which have "strict empirical foundations" (O'Horan et al., 2016). Generally, modern typology branches out into three categories. The first one, qualitative, is concerned with developing variables for capturing similarities or differences in structures within and across languages (Bickel, 2007). The second category is that of quantitative typology, which explores "clusters and skewings" in the distribution of said variables, and, lastly, theoretical typology proposes theories that explain the clusters and skewings (Bickel, 2007).

In the book "Second Language Learning Theories", Mitchell et al. (2012) present the concept of "language transfer", a phenomenon which manifests itself in learners of an additional language who carry on structures or tendencies featured in their L1. Although, according to the authors, this concept has been claimed and rejected by researchers of the field along the years, theorists today "would generally accept once more that cross linguistic influences play an important role in L2 learning" (Mitchell et al. 2012, as cited in Ortega 2009). They mention, however, that attitudes toward this still vary, from researchers who claim that L1 influence is minimal (Klein and Perdue, 1992, as cited in Mitchell et al. 2012), to others, such as Ringböm (2007), who noted in a study that L1 speakers of Swedish acquire English at a faster rate than L1 speakers of Finnish, due to the fact that "Swedish and English are typologically close" (Ringböm, 2007).

An example of inconclusive findings regarding the possible effect that L1 might have on L2 learners is given in Ortega (2009): when Swedish researcher Kenneth Hyltenstam (1977) analyzed how beginner learners of L2 Swedish handle negation, the case of six L1 Turkish learners stood out in particular. Out of these, half of them found no difficulty in following Swedish negation patterns, while the other half was not as successful at this task initially. What is strange about this, Ortega (2012) highlights, is the fact that both Turkish and Swedish are post-verbal negation languages, leading to puzzlement over why the learners would retort to using pre-verbal negation in Swedish.

Other theories propose the idea that while L1 might not have an impact on the difficulty with which a learner will acquire the L2, it can have a noticeable influence on the learner's interaction with the stages of development in the process of learning (Ortega, 2009). This theory was introduced and expanded upon by Zobl (1982), who suggested that in this way, certain users might linger more in a certain development stage, while others might pass through that stage and onto the next one at a faster pace, depending on their L1 background.

Ortega highlights a study by Ellis (1985) that reports that between 23% and 36% of learner errors can be attributed to L1 transfer, which would be the "typical amount" (Ortega, 2008:51), but in reality, over the seven investigations that Ellis had performed, these results varied from 5% to 50%.

In another such study, *Linguistic Distance and the Language Fluency of Immigrants* (Otten and Isphording, 2011), the authors find that, in the case of first generation immigrants to Germany, higher linguistic distance results in higher acquisition costs and "decreases the probability of higher fluency in host country language". In the study "Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages" (Chiswick and Miller, 2005), the authors quote from Crystal's *Cambridge Encyclopedia of Language* (1987):

> *"The structural closeness of languages to each other has often been thought to be an important factor in FLL (foreign language learning). If The L2 is structurally similar to the L1, it is claimed, learning should be easier than in cases where the L2 is very different. However, it is not possible to correlate linguistic difference and learning difficulty in any straightforward way, and even the basic task of quantifying linguistic difference proves to be highly complex, because of the many variables involved."*

Therefore, although it seems to be intuitive and reasonable to affirm that the L1's similarity to the L2 is beneficial to the learner, this claim is not so easily quantifiable and can prove difficult in demonstrating.

# 3 Methods and Data

## 3.1 SweLL Corpus

The data for this project was extracted from the SweLL-pilot (Volodina et al., 2016) and SweLL-gold (Volodina et al., 2019) collections, both Språkbanken corpora. SweLL-pilot is a corpus of essays written by adult learners of Swedish, collected during 2007-2016, containing 502 essays. Subsequently, the SweLL-gold corpus is composed of 502 essays as well, collected during 2017-2021. The main difference between these corpora is that while SweLL-pilot is only anonymized and graded with CEFR labels (Volodina et al., 2016), SweLL-gold is pseudonymized, normalized and correction annotated (Volodina et al., 2019) and, instead of using CEFR labels, the essays are labeled by learner level using the following annotation: *Nybörjare* (beginner), *Fortsättning* (intermediate), and *Advancerad* (advanced).

The swell-pilot corpus is divided into three subcorpora:

- spIn - 256 essays collected from Language Introduction course for newly arrived refugees (v et al. 2016), most of them absolute beginners
- Sw1203 - 141 essays collected from university students in exam setting
- Tisus - 105 essays written as a part of a Test in Swedish for University Studies. All essays are argumentative, on the topic of "stress"

Volodina et al. (2019) state that the types of metadata that have been included in the swell corpus have been selected by importance and relevance to language learning criteria, and fall into two categories: personal information about the learner and information about the essay and the writing context, and the learner's performance on the essay if applicable.

Regarding the topics of the essays, they are mostly narrative or descriptive, with subjects relating to daily life, self introductions, relations to other people, free time, entertainment, arts, travel etc. (Volodina et al., 2016).

The metadata concerning personal details regarding the learners include, mainly: age, gender, information about native or any additional languages, education level, residence time in Sweden (Volodina et al. 2016, 2019). Due to privacy considerations, and to the personal nature of the essays contents, the swell data has been modified accordingly. In the case of swell-pilot, any name or address has been substituted with the placeholder "NN" or "NN-street" (Volodina et al., 2016), while in the case of swell-gold, a considerable effort has been made to pseudonymize any sensitive information, such as: country of origin or age, which is presented as a five year interval (Volodina et al., 2019).

For the current study, the extraction of data from the aforementioned corpora was done under two central criteria: L1 and length of the essays. Firstly, all essays shorter than 200 tokens were filtered out, since a part of the selected measures for complexity would be applied on the individual essays, hence incompatible with texts too small. Consequently, out of all the remaining essays of a suitable length, they

were categorised by the L1 to which the authors appertained, in the order of most prevalent L1, to least prevalent.

As such, the following languages have been chosen for this study: English, Arabic, Dari, Farsi, Spanish, German and Chinese. It is worth noting that a considerable number of the essays featured Swedish as one of the corresponding L1s. For both SweLL-Pilot and SweLL-Gold, the background information was provided by the learners via a metadata sheet (Volodina et al., 2019), where they had full freedom to note down either one, or multiple L1s, leading to somewhat challenging consequences for this study. Firstly, it creates the somewhat paradoxical situation of Swedish as an L1 for learners of Swedish. Since there is no section in the metadata with additional information on the provided L1s, this generates room for speculation. A plausible explanation could be that these learners are immigrants from households in which Swedish was spoken by one or more members, but not as a primary language, and subsequently immigrated to/back to Sweden. Consequently, this leads to learners who might be familiar with Swedish but hold a weak grasp on the language, which would require them learning it as an L2 in order to increase their proficiency.

| Language | Number of essays | Total number of tokens |
|---|---|---|
| Arabic | 73 | 34465 |
| English | 42 | 32216 |
| German | 34 | 17606 |
| Spanish | 33 | 15071 |
| Dari | 30 | 18390 |
| Chinese | 28 | 9362 |
| Persian | 26 | 16990 |
| Swedish | 8 | 10458 |

Table 1: Number of essays, by L1

Table 1 depicts the total number of essays and tokens for each of the selected languages. Although the number of essays which feature Swedish as an L1 is much lower than the rest of the L1s,, it was included in the study so as to explore how a learner who already possesses trace knowledge of the L2 might perform. It should be noted as well that, as seen from the number of tokens, these essays are quite lengthy.

The under-specification in the case of background information on L1s proves itself problematic in other instances as well. For example, in the swell metadata, “Chinese” is featured as an L1, but it is not specified which of the Chinese languages the learner is referring to, excluding a negligible number of essays in which Chinese is specified to refer to Mandarin. The same problem is posed by Arabic, where the specific dialect isn’t indicated. Albeit representing less of a concern, it does leave room for a certain level of ambiguity.

Having accounted for the selection of the L1 category, we can move on to the remaining background data selected as part of this study: CEFR level, age and gender. As mentioned above, Swell-Pilot categorises its essays utilising the classic CEFR labels, while SweLL-Pilot follows a different system of labeling. With the purpose of ensuring clarity, the terms present in SweLL-Gold have been substituted for the corresponding CEFR labels: level A for Nybörjare, level B for Fortsättning, and level C for Advancerad. Since in the SweLL-pilot collection the labels are further elaborated into A1/A2, B1/B2, C1/C2, they have been reduced to only A, B and C for the sake of simplicity.

Age appears in both corpora not as exact numbers, with the purpose of protecting author privacy, but as five year intervals. Originally, are eight groups of age in the SweLL corpora, which have been coded for this study with numbers from one to eight, correspondingly: 16-20: 1; 21-25: 2 ; 26-30: 3 ; 31-35: 4 ; 36-40: 5 ; 41-45: 6 ; 46-50: 7 ; 51-55: 8.

Gender appears in the metadata as male or female, with the options for the authors to not disclose it. As the number of instances in which gender isn't mentioned for the selected essays is relatively low, these have been excluded from the final study.

The SweLL data is organised into the XML format, where each essay represents a tree node, with subsequent nodes encompassing the sentences, and, respectively, the individual words. The texts are extracted using a Python script, with the tokens being stored in dictionaries. Each dictionary key is represented as the essay IDs, with values such as content and metadata information corresponding to them. The data appears in its final form as data frames saved in the comma separated values (CSV) format.

## 3.2 WALS and Universal declaration of human rights corpus

The grammatical information of the selected languages which will be used as part of the analysis was sourced from the World Atlas of Language Structures (WALS) (Haspelmath et al., 2005). WALS is a collection showcasing the structural properties (phonological, grammatical, lexical) of the languages of the world, published first as a book authored by more than 40 domain experts, and subsequently released in online form as well, as a searchable database (Cysouw, 2011). At the time of writing this paper, there are 2662 languages featured in WALS, and a total of 192 structural properties documented.

In the way that they are organised within the "Genealogy" section of WALS, the genealogical classification of the languages skips intermediary levels, including in the majority of cases only the family and the genus level. The authors explain that the highest level, the family category, is generally undisputed and widely accepted by specialists, whereas the genus notion is employed from Dryer (1989), where it is explained as "a level of classification which is comparable across the world, so that a genus in one family is intended to be comparable in time depth to genera in other parts of the world" (https://wals.info/languoid/genealogy). Otherwise, the classification that the WALS authors have adhered to is generally based on the 14th edition of Ethnologue given in Grimes (2000). Of the languages that have been chosen as subjects of this analysis, most of them are part of the Indo-European family (Swedish, German, English, Spanish, Persian), while Chinese is classified within the Sino-Tibetan family, and Arabic within the Afro-Asiatic.

Out of these aforementioned languages, some of them impose a certain level of complication, as mentioned earlier, namely Chinese, Arabic, and Persian. Since the metadata of the SWeLL corpora doesn't include additional background information such as the country of origin of the authors, this leads to ambiguities which concern this segment of the study especially. In WALS, there are 21 variants of Arabic featured, notwithstanding the fact that some of them are very sparsely documented. Out of these, the following contain the most entries: Egyptian (145), Gulf (62), Moroccan (58), Syrian (52), Iraqi (34), Kuwaiti (31), Modern Standard (30). Following the strategy employed by Coloma (2016), for the instances in which certain features were missing data points, they were "inferred by looking at other data points that belong to dialects of the same language". (e.g. Gulf Arabic instead of Egyptian Arabic). As for Chinese, in WALS, there are 7 languages under the genus 'Chinese', Mandarin being the most documented, with 153 feature entries, followed by Cantonese with 93, and then Hakka with 36 features. The remaining 4 languages hold less than 20 features each.

Moving forward, The SweLL metadata features both 'Dari' and 'Persiska' as an L1. In the book "Persian Studies in North America: Studies in Honor of Mohammad Ali Jazayery" (Marashi, 1994), the author describes Dari as holding the meaning of "the language of the court", stating furthermore that this name "had been available as an alternative to *farsi* as the name of the [Persian] language since the earliest times". It is claimed, therefore, that any emphasis on making a distinction between *Persian, Dari,* or another version of the language, such as *Tojiki,* as holding separate identities comes from a place of concern for "national cultural equality than for linguistic form". In WALS, Dari contains a very low number of documented features, specifically 5, with Persian at 147 entries. Upon a closer look at the features that are present for Dari, all five of them present the same values as in the case of Persian, which slightly simplifies the problematic situation. In light of these facts, a choice had to be made to either discard Dari altogether from the whole study, or exclude it only in the portion centered on WALS features, of which the latter was chosen. In the case of Chinese, only Mandarin will be considered, since it holds the largest number of entries out of itself and Cantonese.

| **Morphological** | **Tense, Possession, Aspect, Mood** |
|---|---|
| Inflectional Morphology (26) | Future Tense (67) |
| **Cases** | Past Tense (66) |
| Number of Cases (49) | Perfective/Imperfective (65) |
| **Possibility and evidentiality** | Morphological Imperative (70) |
| Situational Possibility (74) | Coding of Possessives (57) |
| Epistemic possibility (75) | Optative (73) |
| Overlap b/w Epistemic and Situational Possibility (76) | **Articles, Demonstratives, Pronouns** |
| Coding of Evidentiality (77) | Distance Distinctions in Demonstratives (41) |
| **Negation, Plurality, Interrogatives** | Expression of Pronominal Subjects (101) |
| Coding of Negation (112) | |
| Polar Question Coding (92) | |

Table 2: Selected WALS features from Lupyan and Dale (2010), WALS feature codes in parenthesis

In the paper *The role of morphological complexity in predicting the learnability of an additional language: The case of La (additional language) Dutch*, (Schepens et al., 2017) the authors utilise as a point of reference the 28 features extracted from WALS by Lupyan and Dale (2010), which have been incorporated in this paper as well, provided in Table 2. Originally 28 features were selected in total by Lupyan and Dale, out of which only 16 contain entries for all the languages under analysis in this paper.

Coloma's study *Complexity trade-offs in the 100 language WALS sample* (2016) has served as additional inspiration on this section of the current study. Coloma extracts 60 WALS features which are then binarised into complexity variables, that is they hold either value 0 if the feature is considered to be 'less complex', or value 1, otherwise. The overall rule of thumb employed by the author was to follow the interpretation under which more complex means 'more overt distinctions and/or rules' (McWhorter, 2001).

As such, the selected features for this study were binarised following Coloma's method. Out of the aforementioned 16 selected features, 9 of them were showcased in Coloma's study as well, hence their values were adopted for the current study, while the remaining features were further binarised independent of Coloma's method. The values featured in the aforementioned study that coincide with the values used for the current paper, with explanation for the case in which they would get assigned value 1, are displayed in Table 3.

| WALS Feature | Complexity condition |
|---|---|
| Inflectional Morphology (26) | There is some inflectional morphology (prefixing, suffixing, or both) |
| Distance Distinctions in Demonstratives (41) | There are 3-way contrasts or more |
| Number of Cases (49) | Number of cases > 1 |
| Perfective/Imperfective (65) | There is grammatical marking of the perfective/imperfective aspect |
| Past Tense (66) | There is grammatical marking of the past tense |
| Future Tense (67) | There is an inflectional future tense |
| Optative (73) | There is an inflectional optative |
| Coding of Evidentiality (77) | There are some grammatical markers of evidentiality |
| Polar Question Coding (92) | There are polar question particles |

Table 3: WALS binarisation conditions, as in Coloma (2016)

In the case of the features selected for this paper that Coloma omitted on the basis of wanting to avoid arbitrariness, they were assigned values following the general guidelines from Lupyan and Dale (2010), as well as Schepens et al. (217), provided in Table 4.

| WALS Feature | Complexity condition |
|---|---|
| Situational Possibility (74) | Affixes on verbs > verbal constructions |
| Epistemic Possibility (75) | Affixes on verbs > verbal constructions |
| Overlap b/w Epistemic and Situational Possibility (76) | Separate markers > overlap for both possibility and necessity |
| Expression of Pronominal Subjects (101) | Affixes/clitics > other manners of expressing pronominal subjects |

Table 4: WALS binarisation conditions, as in Schepens et al. (2017)

The only features that ended up being omitted were 112 (Coding of negation) and 70 (coding of morphological imperative), as their potential coding of complexity posed the highest level of ambiguity and arbitrariness.

The concluding list of features was then aggregated into a table, where their binary values were thereafter averaged into a final complexity score.

The chosen parallel corpus on which to compute the selected complexity measures is the Universal Declaration of Human Rights corpus. Other corpora of larger sizes were considered as well, but ultimately they were discarded because of various reasons. One problem regarding this was that it proved difficult to find a corpus which included all the selected languages, mainly because a great number of corpora consist of only Indo-European languages. An inclusive corpus from this point of view is the multilingual parallel corpus created from translations of the bible[1].However, ironically, each language entry in this corpus is very large in size, making it difficult to compute the complexity measures on it, specifically the compression algorithm used for Kolmogorov complexity.

## 3.3 Applied Measures

The selected measures presented in the previous chapter (TTR, Entropy, Kolmogorov complexity) were applied on the learner essays, as well as on translations of the Declaration of Human Rights corpus. The first step was, naturally, preprocessing of the texts, which includes lowercasing and removing all punctuation and numbers. Both TTR and Entropy were measured on the individual essays, using a moving window of 100 tokens, via a dedicated script devised in Python. TTR is measured by calculating the ratio between the number of unique words (types) divided by the total number of words (tokens) in a given text.

Entropy, in this case, is calculated based on the frequency distribution of individual words in the text. First, the frequency distribution of each word is calculated, followed by determining the probability of each word, by dividing the frequency by the total number of words in the text. The entropy of each word is calculated using the formula -(probability of the symbol) * log2(probability of the symbol)[2]. Finally, the entropy of each word is summed in order to obtain the overall Shannon entropy of the text.

In the case of Kolmogorov complexity, since the measure is not compatible with such short texts, the same method as in Ehret (2016) was applied, namely: all essays of a particular language are aggregated into three different text files, based on their level, therefore resulting, for each language, in three text files (for example, Arabic A, Arabic B, Arabic C). This compression algorithm, applied in the programming language R, utilizes gzip to capture the recurrence of linguistic structures and regularities or irregularities (Ehret, 2016), at the lexical, morphological, and syntactic level. The complexity is determined "by taking two measurements for each text sample: the file size (in bytes) before compression and the file size after compression", (Ehret, 2016:27), these pairings being afterwards subjected to a regression analysis. The algorithm is based on text distortion in this process, for morphological and syntactic complexity: for

[1] https://github.com/christos-c/bible-corpus
[2] https://medium.com/@nishantnikhil/shannon-information-content-entropy-with-examples-20aca0c1fed6

syntactic distortion, 10% of all word tokens are deleted before compression), while for morphological complexity, 10% of all characters are deleted. (Ehret et al. 2016, Juola, 2008). The resulting scores (regression residuals) are interpreted thus: lower scores indicate lower complexity and higher scores correlate with higher complexity.

Having all the measures calculated on the L2 Swedish of the learner essays, the next step is to assess the complexity of the selected L1s as well. The first way in which this was done has been through the binarisation and averaging of WALS features, which has already been illustrated earlier in this chapter.

Moving forward, the measures implemented on the essays are applied to a parallel corpus which features all of the selected languages, the Universal Declaration of Human Rights. corpus has been chosen. The TTR, entropy, as well as the Kolmogorov complexity, are computed on the translation of each language.

The next step in the study involves conducting a statistical analysis in the R programming language, specifically via linear regression models and data visualisation through boxplots and scatterplots. Through these, the aim is to explore the relationship between the measures themselves, as well as, more importantly, the relationships between the complexity scores of the learner essays and the scores of their respective L1s. The first two linear regression models are computed as an investigation of how the complexity results of the essays, TTR and, respectively, entropy relate with the complexity results of each respective Human Rights declaration. With the next two models which are generated, the complexity scores derived from the Human Rights Declarations are substituted for the average aggregated WALS complexity of each language.

Afterwards, the correlation between the Kolmogorov complexity of all essays at each level and the Kolmogorov complexity score of each language as suggested via the declaration of human rights is visualised via boxplots.

Eight additional models are computed with the intention of determining whether the results of the statistical analysis are affected by errors in the essays or not, based only on the SweLL-Gold corpus, which is error annotated and normalised. The normalisation has been done manually, and it "takes care of the anomalies of learner language where different types of deviations and errors are re-written to represent a standard variant" (Volodina et al., 2019:68). The errors marked in the essays are of six major categories: Orthography (O), Lexis (L), Morphology (M), Syntax (S), Punctuation (P), and Other. For this endeavor, there have been excluded errors of the types O, L and M. As such, there are two additional subsets of the data: the text extracted from only the SweLL-Gold corpus, in two variants: with errors included, and with errors excluded, which are then subsequently used as dataframes for four additional linear regression models each. The main line of thought behind this part of the analysis was that it would be worth seeing whether the errors skew the results or not, and, subsequently, whether the data excluding these errors could be more reliable. However, this could arguably be seen as contradictory to the original aim of the study: analysing the L2 as it is produced by the learner, taking into consideration all aspects and particularities of the speech, including the errors, which ultimately play a role in determining the level of complexity as well. The interest for this kind of investigation was mainly of experimental nature, and as a way of accounting for all possible scenarios of analysis. Had the regression models delivered considerably different results compared with the main models, it could have meant, for example, that perhaps some of

the essays featured the same mistakes, or same kind of mistakes in a repetitive manner, which had altered the scores. However, as will be mentioned in the next section, all models, irrespective of errors being included or not, reported very similar results, which suggests that the above mentioned scenario, or other similar ones, were not the case.

# 4 Results and Discussion

## 4.1 L2 Complexity

As has been mentioned in the previous chapter, each of the selected measures have been computed on both the SweLL data and the declaration of human rights corpus.

Starting with examining TTR and entropy scores, data visualisation plots have been generated in order to gain a better understanding of what the results communicate. Firstly, we can take a look at scatter plots depicting the scores in relation to the instructional level, where each dot represents an individual essay:

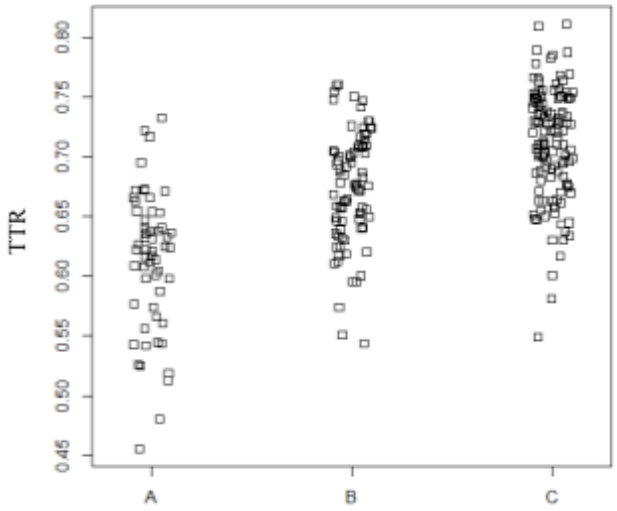


Figure 2: TTR by instructional level, scatterplot

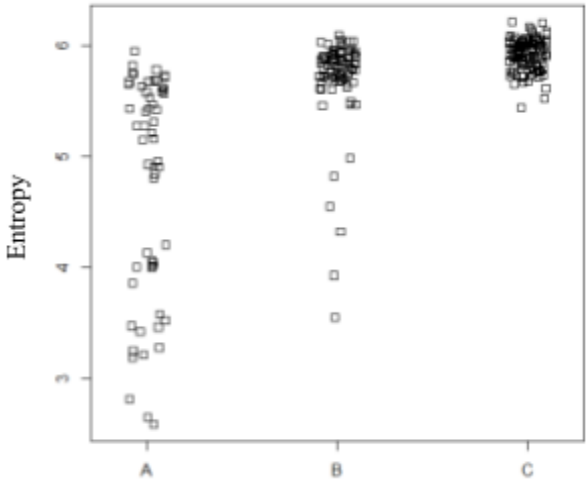


Figure 3: Entropy by instructional level, scatterplot

These plots seem to validate prior assumptions, with scores trending upwards as the instructional level increases, excluding the obvious outliers. In the case of TTR, as seen in Figure 2, it is evident however that the scores are more dispersed than in the case of entropy, but they appear to be spread in a similar manner for both measures at the lowest level. For TTR, this might be attributed to the fact that, considering TTR is calculated on the basis of frequency of unique tokens relative to total tokens, some learners might have an easier time producing albeit simple, grammatical language, while other learners' scores could be skewed by variation produced by lexical errors. In the case of entropy, presented in Figure

3, the same could be said regarding the learners' varying initial proficiency, which is later normalised at higher instructional levels, where the scores are more homogenous among themselves, clustering towards the top at levels B and C.

Having observed these patterns via the scatterplots, boxplots depicting these relationships were generated as well, with the intention of summarising the above-seen distributions and providing a clearer overall image.

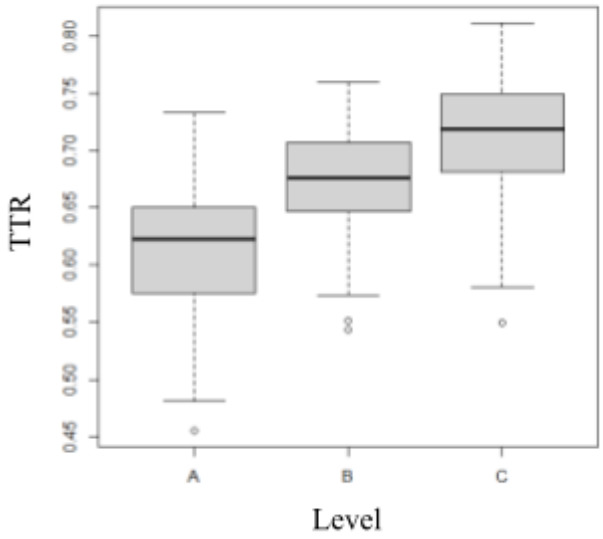


Figure 4: TTR by instructional level, boxplot

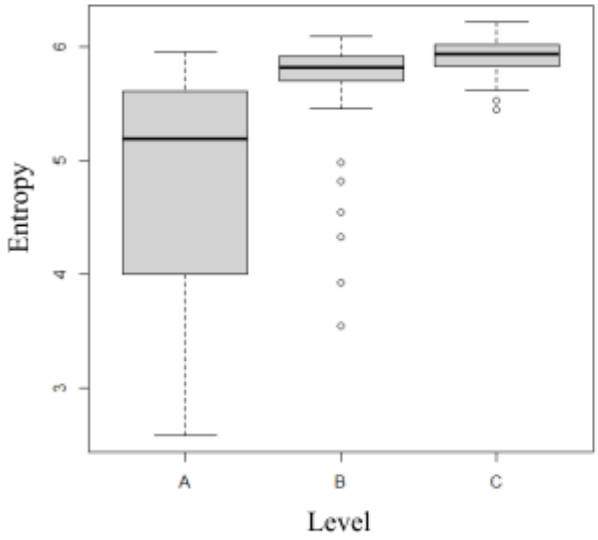


Figure 5: Entropy by instructional level, boxplot

As has been seen above as well, entropy for the higher levels, displayed in Figure 5, is concentrated homogeneously with higher scores with noticeably lower windows. The boxplots do reveal details that weren't apparent in the scatterplots, such as the fact that in the case of TTR, shown in Figure 4, there are not too many extreme outliers in reality. The medians, as well as the lower and upper quartiles, along with the whiskers of the plots seem to be relatively evenly distributed for TTR scores. While the median lines for entropy in relation to level follow are not unexpected, the box depicting entropy scores at level A is unusually tall, with the lower whisker stretching all the way to the bottom, suggesting a much greater diversity in scores compared to any other case in this context.

Moving forward, the relationship between scores and L1 is investigated, starting once again with scatter plots.

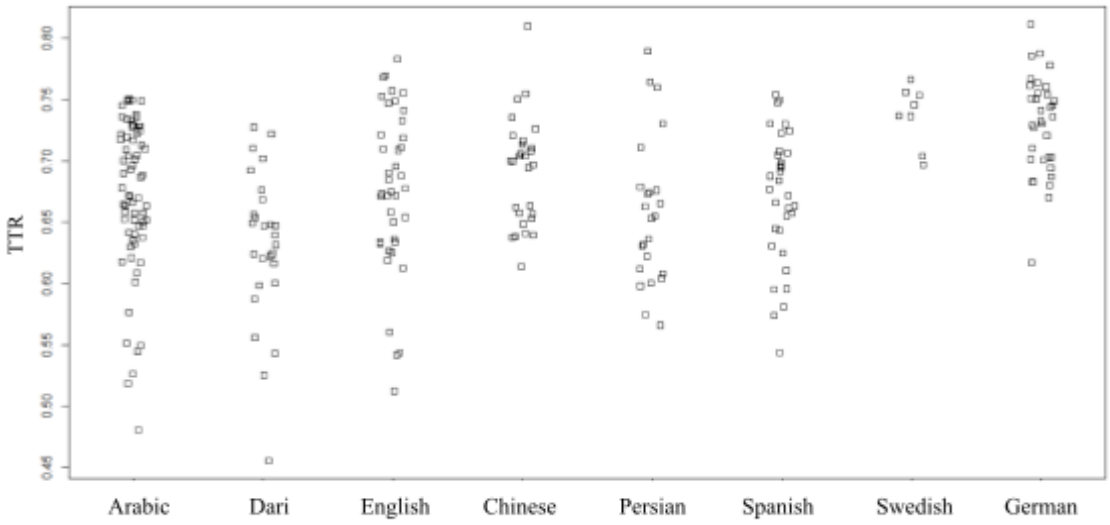


Figure 5: TTR by L1, scatterplot

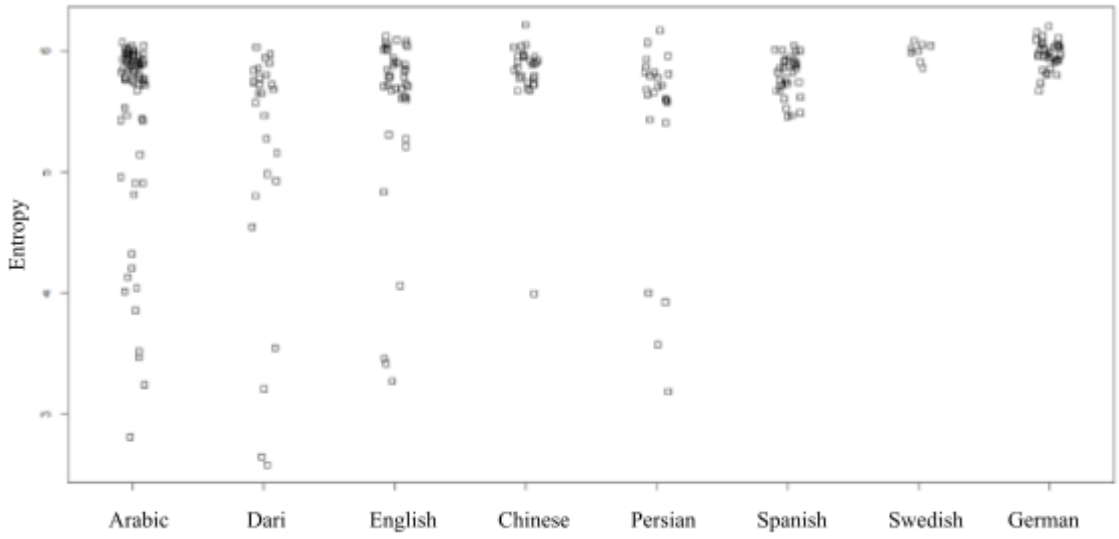


Figure 6: Entropy by L1, scatterplot

It can be seen in Figure 6 that, as has been previously noted in the case of instructional level, entropy scores tend to form clusters, barring the data points which are clear outliers, in contrast to TTR scores shown in Figure 5, where it is not as immediately evident which dots represent outliers, and which do not. As has been mentioned in the previous chapter, there are considerably fewer essays with Swedish as an L1, and their authors having previously been exposed to the language, it is unsurprising that the scores are relatively high and present no outliers. Arabic, on the other hand, holds the highest concentration of data

points, which leads to amplified variation. From these scatterplots, it is difficult to determine whether L1 influences the distribution of the results, even more so as part of the inconclusiveness stems from the difference in data points numbers across L1s. In the case of entropy, most data points are once again clustered toward the top of the scatterplot, making it perhaps easier to draw conclusions when looking at the lower halves of the plots. While Dari and Arabic seem to draw nearer to the bottom, this can be misleading, since they are the two L1s with the highest number of corresponding essays, and it is not exactly clear which of them are outliers. While other L1s such as English and Persian show a low amount of outliers, Spanish, Swedish, German and Chinese (with the exception of one essay) present no outliers. On the other hand, the scatterplot depicting TTR in relation to L1 displays more widely distributed data points, with Arabic and Dari once again descending closer to the bottom of the plot, while German and English seem to hold the overall highest scores.

As was previously the case, considering only the scatterplots does not provide a comprehensive outlook on the distribution of the scores, which is why additional boxplots were generated as well.

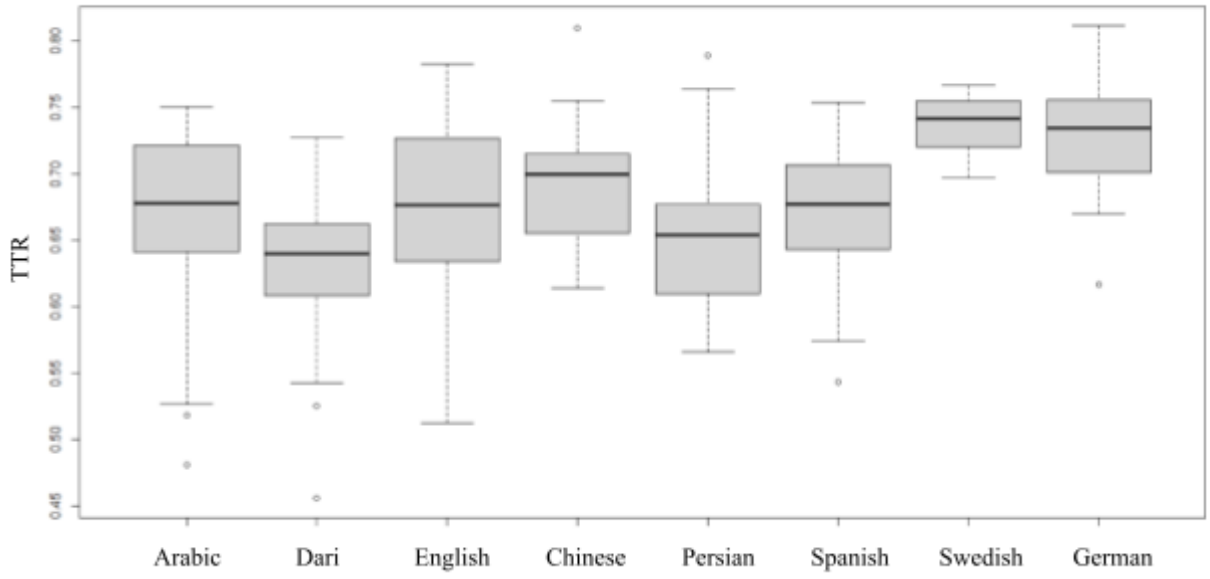


Figure 7: TTR by L1, boxplot

In the box plot depicting the relationship between L1 and TTR, it is now more clear that the highest medians are held by Swedish and German. English L1, although having a lower median than Chinese, Arabic and even Spanish, trends higher in the upper quartile. It is also the L1 that trends the lowest in the bottom quartile, followed by Arabic and Dari.

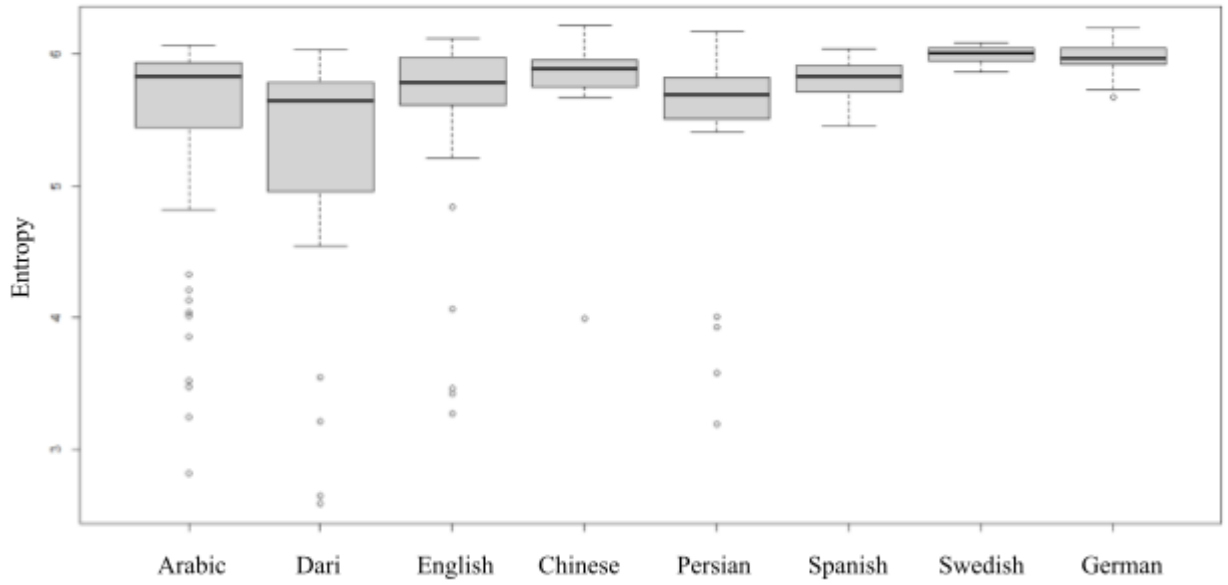


Figure 8: TTR by L1, boxplot

Examining the boxplot depicting entropy score distributions, Arabic presents the most outliers, followed by Dari, English and Persian, all with a similar amount. Chinese, once again, has only one outlying essay, while Spanish, Swedish and German have none. Here, it is even less clear which L1 holds the highest scores, as they are all gathered towards the top of the plot, with mainly much shorter boxes in comparison to the TTR distributions. However, Chinese seems to trend slightly higher than the other L1s in the upper quartile, while Dari hangs the lowest.

| Level C | Level B | Level A |
|---|---|---|
| English | English | Dari |
| Dari | Arabic | English |
| German | German, Spanish | Arabic |
| Chinese | Chinese | Chinese |
| Spanish | Persian | Spanish, Persian |
| Arabic, Swedish | Dari | |
| Persian | Swedish | |

Table 5: L1 Kolmogorov complexity ranked by instructional level

Moving on to the Kolmogorov complexity scores, the results don't necessarily seem to show a substantial correlation to TTR and entropy in the current paper, as it seems that if considering only this measure, complexity rankings as seen in Table 5 are quite different, despite other studies in the literature reporting quite high correlations with other measures (e.g. Ehret et al., 2016). The Kolmogorov scores don't suggest

a clear trend on which L1 determines a higher complexity either. According to the literature “Kolmogorov complexity scores are inherently relative” (Ehret et al, 2016:41), meaning that, while they do not hold inherent meanings and “are hard to interpret in absolute terms” (p. 41), they are meaningful when considered in the context of “the rankings in which they are presented” (p. 41). Overall, Persian does seem to linger at the bottom, with English at the top.

## 4.2 Brief manual analysis

In order to investigate whether the scores, as depicted in these visualisations, were affected by factors such as a high number of lexical errors, which would mean a higher number of ‘different’ words, thus skewing the results, a brief manual analysis was conducted as well.

One such instance of an outlier is an essay situated on the lowest position in the scatterplot depicting the relationship between L2 and L1 TTR, corresponding to Arabic. This essay has the TTR score of 0.48, which is relatively low compared not only to the 0.67 average that Arabic holds overall, but to the average TTR for L1 Arabic essays at level A as well, which is 0.67. Taking a closer look at its contents, it seems that the author does indeed use predominantly verbs in the infinitive form in instances where this is ungrammatical, e.g: “han prata så mycket”, “jag träffa honom”, bringing the TTR score down.

Another interesting outlier case is an essay situated on the opposite end of the spectrum, with a higher TTR score than is normal for the trend displayed in the case of Chinese L1. This level C essay stands out with a score of 0.80, whereas the average TTR for Chinese L1 is 0.68, and the average TTR for Chinese level C is 0.69. This argumentative essay with the topic of discussing a selection of articles from the declaration of human rights holds such a high score because it displays a higher vocabulary variety. This


Figure 9: Word map of L1 Chinese essay with highest TTR score

can be seen with the help of a word map depicting the frequency of the used words in this particular essay, displayed in Figure 9, compared to a word map depicting the most used words overall in all L1 Chinese essays at this level, shown in Figure 10.


Figure 9: Word map for all Chinese essays at level C

On the other hand, in the case of entropy, a good example would be a clear outlier, once again pertaining to Chinese L1, and level A. This essay displays a significantly lower entropy score, its dot on the scatterplot being completely isolated from the other essays in its group, which is made even more evident by the fact that its entropy score is 3.99, compared to the overall average of 5.80. The only other essay with Chinese as an L1 at this level (A) has an entropy score of 5.67. When comparing these two essays, it is clear that the lower entropy score is determined by a much simpler sentence structure overall. It should be noted as well that the essay with the exceptionally low entropy score has, on the other hand, an ordinary TTR score relative to the average in its group. This can be explained by the fact that, while a text can exhibit a variety of distinct words, thus a generally high diversity, if they appear in a predictable, or repetitive manner in the text, without much variation in word order or sentence structure, the entropy score will naturally be much lower.

## 4.3 Correlation between L1 and L2 complexity

| Language | TTR | Entropy |
|---|---|---|
| Arabic | 0.7939516703 | 6.089768253 |
| German | 0.6929558477 | 5.837983775 |
| Swedish | 0.6914891274 | 5.830756869 |
| Chinese | 0.676716486 | 5.824581803 |
| Persian | 0.6707336957 | 5.719889311 |
| Dari | 0.6651319261 | 5.713529513 |
| Spanish | 0.6243049327 | 5.586298334 |
| English | 0.6188474995 | 5.568799518 |

Table 6: TTR and entropy scores for the universal declaration of human rights

Computed on the declaration of human rights corpus, the two measures show very high correlation and agreement, as can be seen in Table 6. In this case, it is perhaps most intuitive to consider the score rankings alongside the aggregated WALS complexity scores, which, as mentioned in the previous chapter, were binarised and averaged.

| Language | WALS Complexity Average |
|---|---|
| Arabic | 0.57 |
| Persian | 0.57 |
| Spanish | 0.42 |
| German | 0.28 |
| Swedish | 0.28 |
| Chinese | 0.21 |
| English | 0.21 |

Table 7: WALS complexity averages for each language

The results shown by the WALS scores, displayed in Table 7, are somewhat simplistic, but they are in line with the general consensus in the field over the documented complexities of these languages as relative to one another, with Arabic and Persian on the higher end of the spectrum and English and Chinese on the lower end. In a study conducted by Bentz et al. (2016), in which languages' complexities were calculated using a WALS-based equation, the following complexity measures were reported, among

other languages: Arabic (0.563), Spanish (0.440), English (0.329). This relative ranking seems to support the findings of the current study, with Arabic demonstrating the highest complexity followed by Spanish, and then by English. It should be noted as well that, in multiple studies on this topic, it has been indicated that the complexity of English is usually lower than that of other languages. (ehret 2016, joula 2008, ehret and szmrecsanyi 2016). The TTR and entropy scores computed on the declarations of human rights can be then investigated with this as a backdrop. In both cases, Arabic scores the highest, and English the lowest. However, in the case of the remaining L1s, the two tables present discrepancies, and don't seem to suggest a clear agreement.

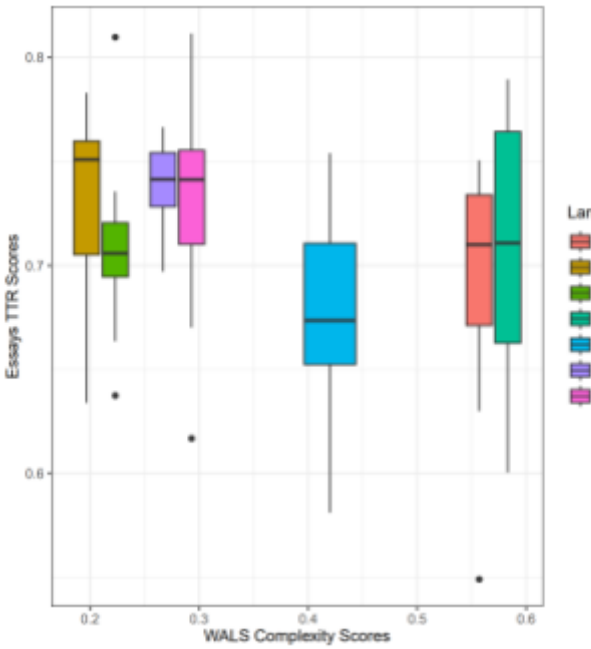


Figure 11: L1 TTR in relation to WALS

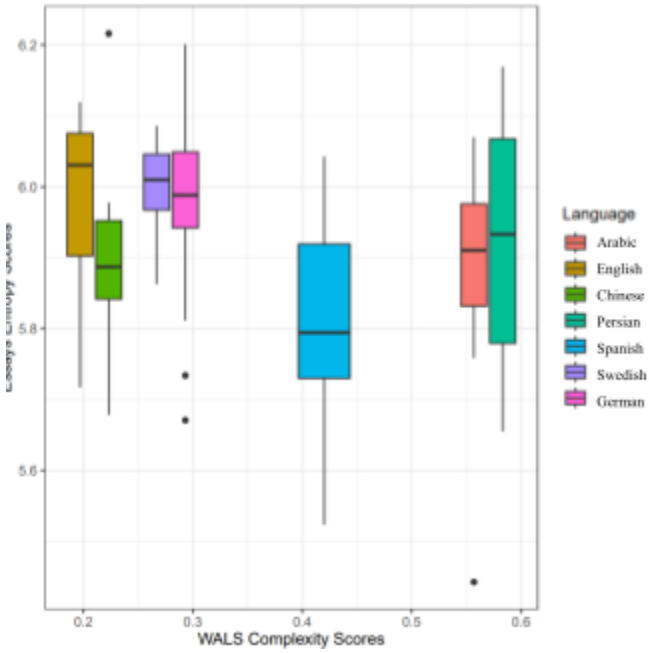


Figure12: L1 entropy in relation to WALS

The boxplots above displayed in Figures 11 and 12 represent the L1 complexity as suggested by the TTR and entropy scores in relation to the WALS complexity scores for each language, both extremely similar to one another, although differences can be seen for example in the differing medians for Swedish and German, as they are lower in the case of entropy.

In the case of how Kolmogorov complexity of the L1 accounts for L2 complexity, the results are mainly inconclusive as well. Scatterplots have been generated in order to visualise the L1 Kolmogorov complexity in relation to the L2 Kolmogorov complexity. The essays extracted at the highest instructional level, level C, are not only the longest in size, which the compression algorithm favours, this level is also the only one at which all L1 languages are represented. Hence, it was considered that the data extracted at this level would be the most representative, as opposed to the lower instructional levels, and it is presented in Figure 13. An additional data point was included, representing all the essays at that level,

regardless of L1, to be regarded as a standard base of complexity, although it seems to place unusually low on the plot. Regarding L2 complexity, with the exception of English and Persian located at the highest and, respectively, lowest extremes, the data points representing the remaining languages are not situated at particularly suggestive positions. On the other hand, considering the L1 complexity scores, it can be observed from the visualisation of the data that Chinese holds unusually high scores. These unreliable results can be largely attributed, in this instance, as the Declaration of Human Rights is in the Chinese script, to the fact that the Chinese writing system is logographic, as opposed to the Arabic or Latin alphabets, which are phonetic or syllabical, an observation reported in other similar studies as well (Ehret, 2018). As the Chinese alphabet uses symbols to represent meanings other than sounds (https://www.berlitz.com/blog/chinese-alphabet), the compression algorithm cannot work properly on these symbols. At the other end of the spectrum are German, Swedish, and, leftmost, English.

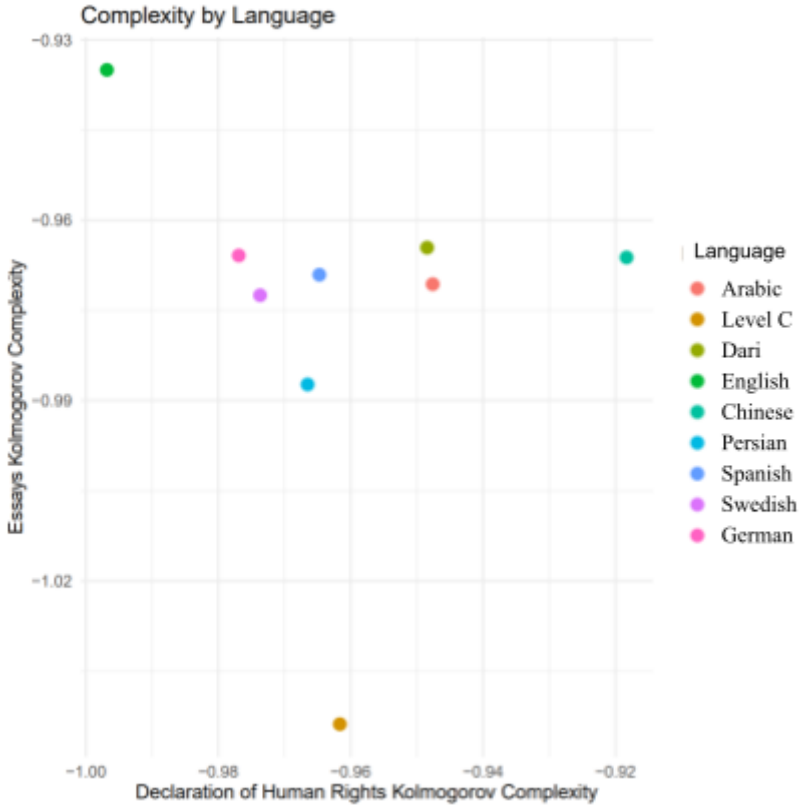


Figure 13: Correlation between L1 and L2 Kolmogorov complexity, level C

## 4.4 Statistical linear regression analysis

As stated in the previous chapter, linear regression analysis was performed under the hypothesis that L1 complexity, as suggested from the scores yielded from the Declarations of Human Rights, as well as the WALS features aggregation, would influence the complexity of essays written in the L2. Hence, the dependent variable of the model was L2 complexity, with the independent variables being L1 complexity and level, as well as the additional age and gender. The models included a supplementary interaction term between L2 complexity and level to account for potential interaction effects. By including this interaction,

the model investigates into greater detail whether the relationship between L1 and L2 complexity differs depending on the Level variable.

Model formula, in R notation: l*2 complexity ~ l1 complexity * Level + Age + Gender*

| | Dependent variable | | |
|---|---|---|---|
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | 0.62 | 0.49 – 0.74 | **<0.001** |
| TTR | -0.02 | -0.20 – 0.16 | 0.790 |
| Level B | 0.04 | -0.14 – 0.22 | 0.649 |
| Level C | 0.08 | -0.08 – 0.24 | 0.302 |
| Age | 0.01 | 0.00 – 0.01 | **0.002** |
| Gender:Man | -0.00 | -0.02 – 0.01 | 0.437 |
| TTR x Level B | 0.02 | -0.24 – 0.28 | 0.877 |
| TTR x Level C | 0.02 | -0.21 – 0.24 | 0.881 |
| Observations | 263 | | |
| $R^2$ / $R^2$ adjusted | 0.399 / 0.382 | | |

Figure 14: Linear regression model summary, for L1 TTR as an L2 complexity predictor

The first model followed the above formula, with the dependent variable being the TTR score of the L2 essays. The reference levels are Gender: Female and Level A. In the resulting summary of the model, displayed in Figure 14, the following findings were reported:

- The overall p-value of the model yields a score of <2.2e-16, which, along with an F-statistic score of 17.7, would suggest that the independent variables as a group hold an overall statistically significant relationship to the dependent variable.
- The R-squared value of 0.3521 indicates that the independent variables account for approximately 35% of the variance of the model.
- The coefficient for the L1 complexity is negative, with a value of -0.051918, which indicates that, as the L1 complexity increases, the L2 complexity decreases. Overall, this coefficient is not statistically significant.

- Although the coefficients relating to the instructional level are not particularly significant either, the trend does seem to suggest that as level increases, L2 complexity increases as well.
- The coefficient for age is statistically significant, indicating that, as age increases, L2 complexity increases as well.
- The gender coefficient, although not statistically significant, is positive, suggesting that male learners yield higher scores overall.

The second model followed the same formula, only this time investigating entropy scores. Generally, it seems to follow the trend presented by the previous model. The most notable differences between the two are the following:

- Age does not seem to be a significant factor in this instance.
- The coefficient between L1 and L2 entropy is marginally statistically significant, indicating a potential linear relationship between the two.

| | **Dependent variable** | | |
|---|---|---|---|
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | 0.62 | 0.49 – 0.74 | **<0.001** |
| TTR | -0.02 | -0.20 – 0.16 | 0.790 |
| Level B | 0.04 | -0.14 – 0.22 | 0.649 |
| Level C | 0.08 | -0.08 – 0.24 | 0.302 |
| Age | 0.01 | 0.00 – 0.01 | **0.002** |
| Gender: Man | -0.00 | -0.02 – 0.01 | 0.437 |
| TTR x Level B | 0.02 | -0.24 – 0.28 | 0.877 |
| TTR x Level C | 0.02 | -0.21 – 0.24 | 0.881 |
| Observations | 263 | | |
| $R^2$ / $R^2$ adjusted | 0.399 / 0.382 | | |

Figure 15: Linear regression model summary, for WALS complexity as an L2 complexity predictor, represented by TTR score

The next two models replace L1 computed complexity scores with the WALS complexity averages. Hence, the first of the two, as shown in Figure 15, explores the relationship between L2 TTR and WALS complexity. Broadly, it reports similar findings as the previous two models. Notable are the reports that, once again, the coefficient for age is statistically significant, as well as the coefficient for level. The

coefficient for gender is, however, not significant. Finally, the model which examines the interactions between L2 entropy and WALS complexity reports overall similar findings. The models computing L1 entropy in relation to WALS complexity, as well as L1 entropy in relation to L2 entropy, can be seen in the Appendix, depicted in figures 16 and, respectively, 17

The models computed on the SweLL-Gold datasets, with and without errors, generally outline the same findings. However, they do suggest that the effect of L1 complexity on L2 complexity may be stronger when errors are present.

As an explanatory note, it should be mentioned why the linear regression analysis was performed only in the case of L1 TTR and entropy, and not in the case of Kolmogorov complexity. This was mainly due to the nature of the data, and of the tools used to calculate the complexity scores. For TTR and entropy, the scores were computed via python scripts for each individual essay, as the moving window technique explained earlier in the paper allowed for disregarding the size of the texts. In the case of Kolmogorov complexity, the compression algorithm called for longer texts, which is why the essays were aggregated into categories, by language and instructional level. This led to an incompatibility in matters of regression analysis, which is why, ultimately, the final decision was to depict Kolmogorov complexity only through data visualisation. Two additional such diagrams can be found in the Appendix, with L2 Kolmogorov complexity in relation to its L1 counterpart, for levels A and B, in figures 18 and, respectively, 19.

Overall, regarding the linguistic distance aspect, studies in the field have demonstrated that the proximity of the L1 and L2 might make a difference when it comes to L2 complexity. In a study by Ehret (2016) it was reported that “German learners of English tend to produce more Kolmogorov-complex essays than French, Italian, or Spanish learners of English”. This seems to hold true, tangentially: speakers of L1 English and German seem to score higher L2 Swedish complexity. However, pursuing this particular aspect of analysis was beyond the scope of this paper, due to the complex nature of the subject, and to a lack of clear definite consensus in the field on how linguistic genetic distance should be measured.

Regarding the initial hypothesis, whether L1 complexity indicates L2 complexity, the results are generally inconclusive, although this was to be somewhat expected when accounting for similar studies. In the paper *CAF: Defining, Refining and Differentiating Constructs*, Pallotti prefaces his study by addressing two broader issues, one of which he calls the “necessary variation fallacy”. He argues that, even though researchers generally look for measures of language performance which show clear variance, the measures which highlight “constants and similarities” (Palloti, 2009) are equally as valuable. He goes on to defend this claim, stating that “if after an experimental treatment two groups of subjects do not show any difference, then this is not a nonresult, but a result just as interesting as their being different” (Pallotti 2009:590).

# 5 Ethical considerations

The main ethical concern regarding this thesis stems from the data being used, specifically relating to privacy and anonymity, as the subject of analysis consisted of learner essays collected from assignments or exams in which the topics at hand could include personal and sensitive information. However, this issue was addressed in the first place by the team who conducted the collection of data originally, as part of Språkbanken corpora SweLL-Pilot and SweLL-Gold corpora. In order to comply with the EU General Data Protection Regulation (GDPR), the team took multiple precautions in order to steer clear of any violation of privacy (Volodina et al., 2019). This included anonymisation or pseudonymisation of sensitive information such as age, names, locations, or any facts pertaining to the authors of the essays or people they referenced, in order to conceal their identity. As such, with all the data used in this study having been previously anonymised, it cannot be used to violate the authors' privacy.

Another ethical concern stems from the consideration that certain findings might hold negative implications toward learners. If, for example, speakers of a certain L1 which ranks lower in terms of complexity are found to be at a disadvantage regarding the L2, this may lead to biases against this group of learners as immigrants in the corresponding country. Since the results of the current study have not indicated that certain groups of learners might be at an advantage or disadvantage based on their L1, there is no need in expanding upon this particular aspect if it is not a realistic scenario.

# 6 Critiques and limitations, future work

Since the main question around which the thesis revolved did ultimately not have a clear answer, there is still work that could be done on the subject, in order to come as close as possible to determining whether a learner's L2 complexity would be higher depending on their L1 complexity. First of all, a larger dataset would be needed in order to assure that it is as representative as possible of the L1, respectively L2. As has been mentioned in the Methods section, the parallel corpus chosen for this study is not very large in size, as opposed to other available corpora, whose size would have proven problematic when computing some of the measures. Thus, under the conditions of more time and more efficient equipment, the measures could be calculated on this corpus, which might provide more representative and reliable results.

Another aspect which might have had an impact on the elucidation of the research question is the type of complexity that was investigated. As has been previously stated, the complexity measured for this study was of the lexical and morphological nature, albeit with a stronger concentration on the lexical aspect. A more comprehensive study could include assessing syntactic complexity, as well as morphological complexity in a more detailed manner. Out of the two generally agreed upon types of morphological complexity, inflectional and derivational, which are overall a relatively new field of study within SLA (Brezina and Pallotti, 2016), derivational complexity has been the least researched, hence investigating it would result, most probably, in valuable insight. Moreover, part of the measures which have been used in this study have been described as rather simplistic (e.g. Çöltekin and Rama 2022). According to Bulte and Housen (2012) as well, the focus on lexical and syntactic measures of complexity has led to "a rather narrow, reductionist, perhaps even simplistic view on and approach to what constitutes L2 complexity"

(p. 34). This leads to the conclusion that expanding the scope and employing additional measures would prove beneficial. Additionally, in order to compute measures which rely on an annotated corpus, for a more detailed study, the data would have to be POS-tagged. This would be advantageous for a more detailed exploration of what type of errors influence the results to a greater extent.

# 7 Conclusions

In this thesis, the subject of investigation was the effect of L1 complexity on the complexity of L2 essays written by learners of Swedish L2. The L2 corresponding data was extracted from the SweLL-Pilot and SweLL-Gold corpora, which consist of essays collected from learners of Swedish, each with varying L1 backgrounds. The essays were selected on the basis of length and most frequent L1s present in the corpora: Arabic, Chinese, Dari, English, German, Persian, Spanish, and Swedish. The measures used to calculate the complexity of these essays were: type-token ratio, Shannon entropy and Kolmogorov complexity. The complexity of the L1s was calculated by applying the same measures on translations of the Declaration of Human Rights parallel corpus. An additional assessment of L1 complexity was devised by extracting grammatical features from the World Atlas of Language Structures and approximating a complexity score based on the features corresponding to each language. Linear regression statistical models were generated in order to determine whether a higher L1 complexity would result in a higher L2 complexity, as well as to investigate how other factors such as instructional level, gender and age would influence the model. Final results did not confirm the hypothesis at the base of the study, as the linear regression models did not show a clear correlation between L1 and L2 complexity. However, from the data visualisations, it was observed that the L2 complexity measures did exhibit a clear tendency over levels, which is an interesting predictor of level that could be studied further. Additionally, the choice of utilising unannotated data was motivated by the ambition to provide simple and cheap language-agnostic methods of working with multiple languages without the need of prior annotation.

# References


Arnaud, P. J. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. Vocabulary and Applied Linguistics, 133–145.

Ackerman, Farrell, and Robert Malouf. 2013. “Morphological Organization: The Low Conditional Entropy Conjecture.” Language 89 (September): 429–64.

Anderson, Stephen R. 2015. “Dimensions of Morphological Complexity.” In Understanding and Measuring Morphological Complexity, edited by Matthew Baerman, Dunstan Brown, and Greville G. Corbett, 0. Oxford University Press.

Bentz, Christian, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. 2016. “A Comparison Between Morphological Complexity Measures: Typological Data Vs. Language Corpora.” In Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), 142–53. Osaka, Japan: The COLING 2016 Organizing Committee.

Bickel, Balthasar. 2007. “Typology in the 21st Century: Major Current Developments.” Linguistic Typology 11 (January): 239–51.

Bley-Vroman, Robert. 1983. “The Comparative Fallacy in Interlanguage Studies: The Case of Systematicity1.” Language Learning 33 (1): 1–17.

Brezina, Vaclav, and Gabriele Pallotti. 2019. “Morphological Complexity in Written L2 Texts.” Second Language Research 35 (1): 99–119.

Bui, Gavin, and Peter Skehan. 2018. “Complexity, Accuracy, and Fluency.”

Bulté, Bram, and Alex Housen. 2012. “Defining and Operationalising L2 Complexity.” In, 21– 46.

Bygate, M. (1996). E.ects of Task Repetition: Appraising the Developing Language of Learners. In J. Willis, & D. Willis (Eds.). Challenge and Change in Language Teaching.

Bygate, Martin. 1999. “Quality of Language and Purpose of Task: Patterns of Learners’ Language on Two Oral Communication Tasks.” Language Teaching Research 3 (3): 185– 214.

Carstairs‐McCarthy, Andrew. 2011. “The Evolution of Morphology.” In The Oxford Handbook of Language Evolution, edited by Kathleen R. Gibson and Maggie Tallerman, 0. Oxford University Press.

Chiswick, Barry R, and Paul W Miller. n.d. “Linguistic Distance: A Quantitative Measure of the Distance Between English and Other Languages.”

Collentine, Joseph. 2004. “THE EFFECTS OF LEARNING CONTEXTS ON MORPHOSYNTACTIC AND LEXICAL DEVELOPMENT.” Studies in Second Language Acquisition 26 (2): 227–48.

Coloma, Germán. 2017. “Complexity Trade-Offs in the 100-Language WALS Sample.” Language Sciences 59 (January): 148–58.

Çöltekin, Çağrı, and Taraka Rama. 2022. “What Do Complexity Measures Measure? Correlating and Validating Corpus-Based Measures of Morphological Complexity.” Linguistics Vanguard, September.

Covington, Michael A., and Joe D. McFall. 2010. “Cutting the Gordian Knot: The Moving- Average Type-Token Ratio (MATTR).” Journal of Quantitative Linguistics 17: 94–100.

DeKeyser, Robert. 2016. “OF MOVING TARGETS AND CHAMELEONS: Why the Concept of Difficulty Is So Hard to Pin Down.” Studies in Second Language Acquisition 38 (2): 353–63.

Derwing, Tracey, and Marian Rossiter. 2003. “The Effects of Pronunciation Instruction on the Accuracy, Fluency, and Complexity of L2 Accented Speech.” Applied Language Learning 13 (January).

Deutscher, G. (2009). " Overall complexity": a wild goose chase?.

Diessel, Holger. 2004. The Acquisition of Complex Sentences. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

Dryer, Matthew S. 1989. “Large Linguistic Areas and Language Sampling:” Studies in Language 13 (2): 257–92.

Ellis, R. (1985). Understanding second language acquisition. Oxford University Press.

Ehret, Katharina. 2016. “An Information-Theoretic Approach to Language Complexity: Variation in Naturalistic Corpora.” PhD thesis.

Ehret, K. (2018). KOLMOGOROV COMPLEXITY AS A UNIVERSAL MEASURE OF LANGUAGE COMPLEXITY.

Ehret, Katharina, and Benedikt Szmrecsanyi. 2019. “Compressing Learner Language: An Information-Theoretic Measure of Complexity in SLA Production Data.” Second Language Research 35 (1): 23–45.

Foster, Pauline, and Peter Skehan. 1996. “The Influence of Planning and Task Type on Second Language Performance.” Studies in Second Language Acquisition 18 (3): 299–323.

Fotos, Sandra. 1993. “Consciousness Raising and Noticing Through Focus on Form: Grammar Task Performance Versus Formal Instruction.” Applied Linguistics 14 (4): 385– 407.

Freed, B. F. (1995). What makes us think that students who study abroad become fluent? Studies in Bilingualism, 123.

Graaff, Rick. 1997. “THE EXPERANTO EXPERIMENT.” Studies in Second Language Acquisition 19 (June): 249–76.

Grimes, Barbara F. (ed.) 2000. Ethnologue. Dallas: SIL International. (2 vols. 14th edition).

Gass, S. M., & Selinker, L. (1994). Language transfer in language learning. John Benjamins Pub.

Haspelmath, Martin, Hans-Jörg Bibiko, Jung Hagen, and Claudia Schmidt. 2005. The World Atlas of Language Structures.

Helgertz, Jonas. 2013. “Pre- to Post-Migration Occupational Mobility of First Generation Immigrants to Sweden from 1970-1990: Examining the Influence of Linguistic Distance.” Population Research and Policy Review 32 (3): 437–67.

Housen, Alex. n.d. “HousenPierrard 2005 - Introducing Investigations in Instructed Second Language Acquisition.” Accessed May 19, 2023.

Housen, Alex, and Folkert Kuiken. 2009. “Complexity, Accuracy, and Fluency in Second Language Acquisition.” Applied Linguistics 30 (4): 461–73.

Housen, Alex, F. Kuiken, and Ineke Vedder. 2012. “Complexity, Accuracy and Fluency.” In, 1–20.

Housen, Alex, and Hannelore Simoens. 2016. “INTRODUCTION: COGNITIVE PERSPECTIVES ON DIFFICULTY AND COMPLEXITY IN L2 ACQUISITION.” Studies in Second Language Acquisition 38 (2): 163–75.

Hyltenstam, Kenneth. 1977. “Implicational Patterns in Interlanguage Syntax Variation.” Language Learning 27 (2): 383–411.

Isphording, Ingo E., and Sebastian Otten. 2011. “Linguistic Distance and the Language Fluency of Immigrants.”.

Juola, Patrick. 1998. “Cross-Entropy and Linguistic Typology.” In New Methods in Language Processing and Computational Natural Language Learning.

Juola, Patrick. 2008. “Assessing Linguistic Complexity.” Language Complexity: Typology, Contact, Change, January.

Klein, W., & Perdue, C. (1997). The Basic Variety (or: Couldn’t natural languages be much simpler?). Second Language Research, 13(4), 301–347.

Kortmann, B. & Szmrecsanyi, B. (2012). Linguistic Complexity: Second Language Acquisition, Indigenization, Contact. Berlin, Boston: De Gruyter.

Kuiken, Folkert. 2022. “Linguistic Complexity in Second Language Acquisition.” Linguistics Vanguard, October.

Kusters, C. W. (2003). Linguistic complexity: The Influence of Social Change on verbal inflection.

Li, M., & Vitányi, P. (1997). An Introduction to Kolmogorov Complexity and Its Applications.

Linnarud, M. (1986). Lexis in Composition. Lund: Lund Studies in English.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. PloS one, 5(1), e8559.

Marashi, M., & Jazayery, M. A. (1994). Persian studies in North America Studies in honor of Mohammad Ali Jazayery. Iranbooks.

McWhorter, J. H. (2001). The World’s Simplest Grammars Are Creole Grammars. Linguistic Typology, 5, 125-166. - References - Scientific Research Publishing.

Menezes, Vera. 2013. “Second Language Acquisition: Reconciling Theories.” Open Journal of Applied Sciences 03 (07): 404–12.

Mitchell, R., Myles, F., Marsden, E. (2019). Second Language Learning Theories (4th ed.). Taylor and Francis.

Morita, Emi. 2000. “A Cognitive Approach to Language Learning by Peter Skehan. Oxford: Oxford University Press, 1998, 324 Pp.” Issues in Applied Linguistics 11.

Muñoz, Carmen. 2006. Age and the Rate of Foreign Language Learning. Multilingual Matters.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. Language Learning, 50(3), 417–528.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Applied Linguistics, 30(4).

O'Horan, Helen, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. "Survey on the Use of Typological Information in Natural Language Processing." In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 1297–1308. Osaka, Japan: The COLING 2016 Organizing Committee.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. Studies in Second Language Acquisition, 21(1), 109–148.

Ortega, L. (2009). Understanding second language acquisition. Routledge.

Pallotti, Gabriele. 2009. "CAF: Defining, Refining and Differentiating Constructs." Applied Linguistics.

Ringbom, H. (2006). Cross-Linguistic Similarity in Foreign Language Learning.

Robinson, Peter. 2001. Cognition and Second Language Instruction. Cambridge University Press.

Robinson, Peter. 2011. "Second Language Task Complexity, the Cognition Hypothesis, Language Learning, and Performance." In, 3–38.

Sadeniemi, M., Kettunen, K., Lindh-Knuutila, T., & Honkela, T. (2008). Complexity of European Union Languages: A comparative approach*. Journal of Quantitative Linguistics, 15(2), 185-211.

Selinker, Larry. 1972. "INTERLANGUAGE." IRAL - International Review of Applied Linguistics in Language Teaching 10 (1-4).

Shannon, C. E. 1948. "A Mathematical Theory of Communication." Bell System Technical Journal 27 (3): 379–423.

SINICROPE, HEIDI BYRNES, CASTLE. 2008. "Advancedness and the Development of Relativization in L2 German: A Curriculum-Based Longitudinal Study." In The Longitudinal Study of Advanced L2 Capacities. Routledge.

Skehan, P. (1996). Second Language Acquisition Research and Task Based Instruction. In J. Willis, & D. Willis (Eds.), Challenge and Change in Language Teaching (Pp. 17-30). Oxford Heinemann. - References - Scientific Research Publishing.

Skehan, Peter. 2009. "Modelling Second Language Performance: Integrating Complexity, Accuracy, Fluency, and Lexis." Applied Linguistics 30 (4): 510–32.

Slik, Frans van der, Roeland van Hout, and Job Schepens. 2019. "The Role of Morphological Complexity in Predicting the Learnability of an Additional Language: The Case of La (Additional Language) Dutch." Second Language Research 35 (1): 47–70.

Volodina, Elena, Lena Granstedt, Arild Matsson, Beáta Megyesi, Ildikó Pilán, Julia Prentice [Grosse], Dan Rosén, et al. 2019. "The SweLL Language Learner Corpus: From Design to Annotation." The Northern European Journal of Language Technology 6 (December): 67– 104.

Volodina, Elena, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. “SweLL on the Rise: Swedish Learner Language Corpus for European Reference Level Studies.” In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16), 206–12. Portorož, Slovenia: European Language Resources Association (ELRA).

Williams, J., & Evans, J. (1998). What Kind of Focus and on Which Forms? In C. Doughty, & J. Williams (Eds). Focus on Form in Classroom Second Language Acquisition (Pp. 139–151). Cambridge: Cambridge University Press.

Yuan, Fangyuan, and Rod Ellis. 2003. “The Effects of Pre‐Task Planning and On‐Line Planning on Fluency, Complexity and Accuracy in L2 Monologic Oral Production.” Applied Linguistics 24 (1): 1–27.

Zobl, Helmut. 1982. “A Direction for Contrastive Analysis: The Comparative Study of Developmental Sequences.” TESOL Quarterly 16 (2): 169.

# A Resources (Links)

- Ethnologue. Volume 1: Languages of the World; Volume 2: Maps and Indexes [14th Ed.]. 2013. SIL International. https://www.sil.org/resources/archives/6229.
- https://github.com/katehret/measuring-language-complexity
- http://research.ics.aalto.fi/cog/data/udhr/
- https://wals.info/

# B Linear Regression Models

| | Dependent variable | | |
|---|---|---|---|
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | 5.01 | 4.60 – 5.42 | **<0.001** |
| WALS | -0.44 | -1.25 – 0.37 | 0.283 |
| Level B | 1.01 | 0.53 – 1.49 | **<0.001** |
| Level C | 1.03 | 0.57 – 1.48 | **<0.001** |
| Age | -0.01 | -0.04 – 0.03 | 0.708 |
| Gender: Man | -0.00 | -0.13 – 0.13 | 0.984 |
| WALS x Level B | -0.20 | -1.27 – 0.87 | 0.714 |
| WALS x Level C | 0.21 | -0.79 – 1.20 | 0.684 |
| Observations | 236 | | |
| $R^2$ / $R^2$ adjusted | 0.473 / 0.456 | | |

Figure 16: Linear regression model summary, for WALS complexity as an L2 complexity predictor, represented by entropy score

| | Dependent variable | | |
|---|---|---|---|
| *Predictors* | *Estimates* | *CI* | *p* |
| (Intercept) | 7.96 | 4.14 – 11.78 | **<0.001** |
| Entropy | -0.55 | -1.21 – 0.11 | 0.101 |
| Level B | -1.15 | -6.59 – 4.28 | 0.677 |
| Level C | -2.32 | -7.15 – 2.50 | 0.344 |
| Age | 0.02 | -0.02 – 0.06 | 0.423 |
| Gender: Man | -0.08 | -0.22 – 0.05 | 0.226 |
| Entropy x Level B | 0.35 | -0.58 – 1.29 | 0.458 |
| Entropy x Level C | 0.59 | -0.23 – 1.42 | 0.158 |
| Observations | 263 | | |
| $R^2$ / $R^2$ adjusted | 0.435 / 0.420 | | |

Figure 17: Linear regression model summary, for L1 entropy as an L2 complexity predictor

# C Additional Data Visualisations

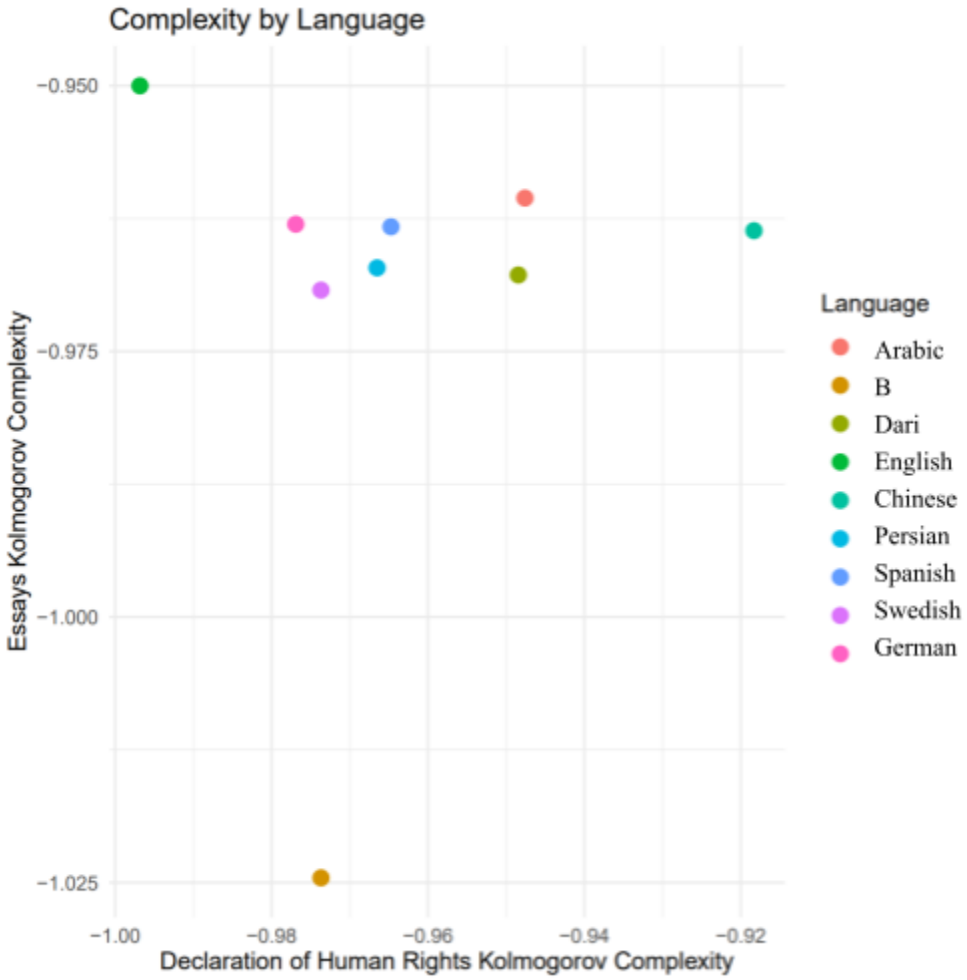


Figure 18: Correlation between L1 and L2 Kolmogorov complexity, level B

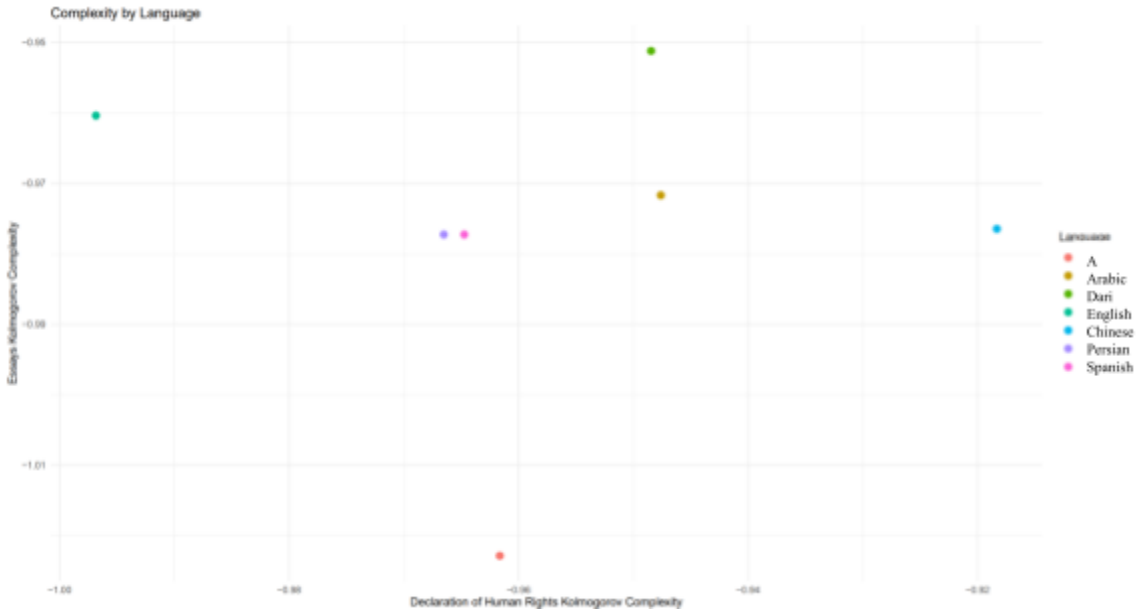


Figure 19: Correlation between L1 and L2 Kolmogorov complexity, level A