Eberhard Karls Universität Tübingen
Seminar für Sprachwissenschaft

Analyzing Linguistic Complexity of L2 Portuguese for Automatic Proficiency Classification

AUTHOR
Eric DeMattos

ADVISOR
Prof. Dr. Detmar Meurers

August 2020

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbstständig und nur mit den angegebenen Quellen und Hilfsmitteln einschließlich des WWW und anderer elektronischer Quellen angefertigt habe. Alle Stellen der Arbeit, die ich anderen Werken dem Wortlaut oder dem Sinne nach entnommen habe, sind kenntlich gemacht.

Eric DeMattos

Abstract

This thesis explores the interaction between different linguistic complexity features and their efficacy in automatically classifying proficiency levels in L2 Portuguese learners using supervised machine learning. We have found that even a small subset of complexity measures are useful, though supplementing them with other information may lead to further improvements. This analysis also corroborates previous findings in second language acquisition research that has posited various complexity features as good indicators of language growth.

Contents

1 Introduction

Automated essay scoring has received increased attention in recent years due to the mainstream adoption of digital education platforms and the proliferation of annotated corpora on which to evaluate such systems. One emerging application of this field is automatic proficiency classification: assessing language learner text and mapping their level on a scale, e.g. the Common European Framework of Reference for Languages (CEFR). For a time, research had mostly centered around English, though efforts have been made to expand the landscape to other languages such as German (Hancke and Meurers, 2013; Weiß and Meurers, 2019), Swedish (Östling et al., 2013; Pilán and Volodina, 2016), Estonian (Vajjala and Lõo, 2014), Norwegian (Berggren et al., 2019), Spanish (del Río, 2019b), and even cross-lingually for German, Czech, and Italian (Vajjala and Rama, 2018).

Following the introduction of a Portuguese learner corpus (del Río et al., 2018), the first automatic proficiency classification tests for the "language of Camões" have surfaced (del Río, 2019a,b). In these studies, del Río initially experimented with bag-of-words, POS and dependency *n*-grams, and a modest amount of linguistic complexity features, obtaining 72% accuracy with bag-of-words and POS *n*-grams being the best performing features. She later investigated cross-lingual Spanish-Portuguese classification in view of their morphosyntactic proximity, but did not surpass the results of her previous work.

Since linguistic complexity has been identified as one of the core dimensions characterizing language proficiency (Housen and Kuiken, 2009), one would expect it to be a well-suited metric for measuring growth. The lack of resources available in this area for Iberian languages, however, has heretofore impeded further research. In response, this thesis aims to expand support for extracting complexity features of European Portuguese and explore a larger set of feature interactions to determine whether it can indeed be a better indicator of language proficiency. To that end, it will provide a new platform on which to calculate such measures for Portuguese in a multilingual context.

2 Background

2.1 Portuguese

Portuguese is the mother tongue of approximately 215 million people globally, making it the sixth most spoken in the world by number of native speakers. It is the official language of seven countries and maintains co-official status in three others. The language is further divided into two main dialects that are generally mutually intelligible: European and Brazilian. The varieties of Lusophone countries in Africa and Asia often resemble their European counterpart more closely due to prolonged colonial rule, though the influence of indigenous languages and the transatlantic slave trade resulted in some shared phonological and prosodic divergence among certain African variants and Brazilian Portuguese in particular.

Similar to other Romance languages, Portuguese is characterized by moderate fusional inflection and, by consequence, a somewhat less rigid word order. While it is generally an SVO language, it also allows for SOV constructs when, for example, object pronouns are realized preverbally in what is known as proclisis. Object pronouns may otherwise be attached after the verb as an enclitic, with usage varying by dialect and context. There also exists a mesoclitic construction—restricted to verbs in the future or conditional tense—wherein a personal pronoun, object pronoun, or both may be inserted between a verb stem and its inflectional suffix.

(1) me chamo
1.SG=call.1.SG.PRES
"(I) call myself" (Galves et al., 2005)

(2) chamo-me
call.1.SG.PRES=1.SG
"(I) call myself" (Galves et al., 2005)

(3) dar-lhes-ão
give=DAT.3.PL=3.PL.FUT
"(they) will give them" (Quarezemin et al., 2018)

Contractions are also ubiquitous in both written and spoken Portuguese, irrespective of register. Many common function word pairs are realized exclusively in their merged forms, analogous to French *du* (de + le) and Spanish *al* (a + el). A few examples of mandatorily-contracted words are provided in Table 1.

	o	*a*	*os*	*as*
a	*ao*	*à*	*aos*	*às*
de	*do*	*da*	*dos*	*das*
em	*no*	*na*	*nos*	*nas*
por	*pelo*	*pela*	*pelos*	*pelas*

Table 1: Subset of prepositions merging with definite articles marked for gender and number. From top to bottom: to, from, in, for.

Portuguese, similar to other languages, contains multiple cleft constructions used to bring a constituent into focus. There are five main variants according to Lobo et al. (2019) that are acquired at different stages of linguistic development, including the atomic *é que* cleft illustrated in (4).

(4) Este ator é que a Academia escolheu.
this actor be.PRES that the Academy choose.PAST
"It was this actor that the Academy chose." (Lobo et al., 2019)

Another characteristic of Portuguese is the presence of an infinitive inflection, allowing for person and number agreement with an overt or implicit subject in embedded contexts without a subordinating conjunction (Rothman, 2007; Ambar et al., 2017). The distinction is exemplified in (5-6) with the verb *ir* "to go".

(5) Lamento que tu não vás à festa
(I) regret that you not go.2.SG.SUBJ to.the party
"I regret that you don't go to the party." (Ambar et al., 2017)

(6) Lamento tu não ires à festa
(I) regret you not go.2.SG.INF to.the party.
"I regret that you haven't gone to the party." (Ambar et al., 2017)

These features will be further discussed in Section 6 in relation to their efficacy in proficiency classification.

2.2 Automatic Proficiency Classification

Students learning a language are often categorized according to their skill level to determine their different abilities and needs. Various standards exist for assessing this, though one of the most popular is the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). Three coarse reference levels ranging from A (basic) to C (proficient) are used to delineate broad groups of students, each of which can be further divided into two sublevels: A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced), and C2 (mastery).

An empirical scale on which to measure proficiency is useful for identifying materials suitable for a student's level that will challenge them appropriately and foster growth. Comprehensible input is requisite for learners to build on concepts they already understand while exploring new aspects of the language ($i + 1$; Krashen, 1981). For this reason, it is crucial to place students at the appropriate level so that they are not under-stimulated, though not so high that they are pushed beyond their Zone of Proximal Development—the space in which they are able to perform tasks successfully, with or without assistance (Vygotski and Kozulin, 1986).

With the burgeoning availability of learner corpora being compiled for various languages in recent years, there is much interest for automating the task of proficiency classification. Yet there are still numerous obstacles to overcome, two of which being the processing of non-standard learner language in NLP contexts designed for fluent, native-like text, as well as clearly demarcating proficiency levels, especially since the assignment of levels to students is usually performed by humans—sometimes arbitrarily—and can therefore lead to ill-defined boundaries between classes.

2.3 Linguistic Complexity

Second language acquisition research has had a long-standing interest in isolating the factors most responsible for language growth. Three metrics in particular—complexity, accuracy, and fluency (CAF)—have been widely accepted for quantifying this progression (Housen and Kuiken, 2009). While *accuracy* relates to the errors students make and *fluency* to spontaneous, competent language production, *linguistic complexity* focuses on "the extent to which the language produced in performing a task is elaborate and varied" (Ellis, 2003) or whether structures generally considered to be "acquired late" appear in a learner's L2 system (Pallotti, 2009). This thesis is primarily focused on the third dimension of this triad: assessing the relationship between the increasingly rich constructions learners are able to produce vis-à-vis their current state of development.

Obtaining these measures has been greatly facilitated in recent years by the growing number of resources available in the realm of computational linguistics in addition to datasets born out of research in psychology, psycholinguistics, and cognitive science. Complexity can be measured objectively using a variety of ratios, frequencies, or formulas (Norris and Ortega, 2009), and the scope of the unit under observation will determine which approach to take.

2.3.1 Surface

Among the most basic measures to compute are length-based features, requiring minimal linguistic information. Surface length features include the length of a text or the average length of its subunits in various denominations ranging from word surface forms, characters, or syllables. In addition to the raw count, averages and normalizations can be obtained by dividing the cumulative sum by another value, e.g. mean word length in syllables. Tokenized words are calculated separately, the counts of which can also be leveraged in more complex lexical and sentential features described in the following sections.

Simple as they may be, length-based features have proven very powerful indicators of overarching complexity (Norris and Ortega, 2009; Housen and Bulté, 2012).

Absent everything else, using length of some sort should always be considered as a pillar in any complexity analysis.

2.3.2 Lexical

Word or token features encode lexical information. Lu (2012) refers to this notion as lexical richness and catalogues in detail three broad subcategories which are expanded on below: density, sophistication, and variation

Density is the ratio of words to the total number of words in a text. Lexical words, function words, and part of speech counts can be scaled by the total number of tokens, with or without the exclusion of function words. In the case of verbs, they can be normalized either by the number of verb tokens (Verb Variation 1, VV1) or the total number of lexical words (Verb Variation 2, VV2). Pallotti (2009) and Lu (2012) both found that generic density measures did not correlate sufficiently with their targets, though Lu (2012) indicated that modified versions of these ratios including Squared Verb Variation (SVV1) and Corrected Verb Variation (CVV1) performed well for assessing lexical richness.

Sophistication is the number of advanced lexical words to the total number of lexical words. These measures refer to norm lists of words with an associated value compiled from a specific domain, such as age of acquisition (Carroll and White, 1973), concreteness (Paivio et al., 1968), familiarity (Gernsbacher, 1984), and imageability (Paivio et al., 1968)—all of which are concepts from psycholinguistics that have been of active interest in second language acquisition research. Additionally, word frequency is a popular element to consider as well, with modern resources such as the SUBTLEX project (New et al., 2007; Brysbaert and New, 2009) seeking to compile large-scale corpora of colloquial speech that is more representative than written text.

Variation describes the diversity of a learner's vocabulary. This can be figured using the type-token ratio, which relates the number of unique word types T to the total number of words in a text N (Templin, 1957). This naïve formula is often avoided, though, due to its sensitivity to text length: as the size of the text

increases, the number of unique words tapers off while the proportion dwindles due to the linearly-increasing denominator (McCarthy and Jarvis, 2007). Various alternatives seeking to remedy this issue have been proposed including Root TTR (RTTR; Guiraud, 1960), Corrected TTR (CTTR; Carroll, 1964), Bilogarithmic TTR (LogTTR; Herdan, 1964), and the Uber Index (Dugast, 1979). McCarthy and Jarvis (2010) have also introduced the Measure of Textual Lexical Diversity, which improves on previous approaches by calculating the TTR in forward and backward directions, dividing them into segments that are reset once a default threshold is attained.

2.3.3 Syntactic

At the syntactic-level, larger units are collected that measure phrasal, clausal, or sentential properties. The minimally terminable unit (T-Unit; Hunt, 1965) is another favored denomination on which to measure syntactic complexity due to its ability to capture stylistic variance. These constituents provide insight on coordination, subordination, and embedded structures (Ortega, 2003), with the T-Unit neutralizing the effect of coordination.

Housen and Bulté (2012) present a non-exhaustive list of measures that quantify a variety of syntactic features, length of units, sophistication of structural complexity, and the amount of coordination, subordination, and embedding. Global measures such as the mean length of T-Unit seem to be powerful indicators of overall complexity (Wolfe-Quintero et al., 1998; Ortega, 2003; Norris and Ortega, 2009) with subordination complexity, e.g. mean number of clauses per T-unit, and phrasal or subclausal elaboration, e.g. mean length of clause, not far behind (Norris and Ortega, 2009).

The occurrence of other morphosyntactic phenomena are also considered. Certain inflections or derivations are used to determine the structural sophistication, e.g. the total number of different verb tenses, classes, or moods. Infinitival sentences, conjoined clauses, and *wh*-clauses are also relevant for this purpose (Norris and Ortega, 2009; Pallotti, 2009).

Features also exist that are based on dependency relations in the domain of human language processing. The Dependency Locality Theory (Gibson, 2000) is one such example, measuring the distance between heads and their dependents. It takes into account both the storage of previous elements as well as their integration through sentence parsing and comprehension.

2.3.4 Discursive

According to Granger and Tyson (1996); Graesser et al. (2004), discourse markers in text are cohesion devices. These markers serve to link contexts between sentences and ultimately assist in resolving the overall theme. Commonly employed cohesives are connectives: explicit words or phrases that signal discourse progression and associate relations between ideas. Connectives can be additive (*also, moreover*), adversative (*however, in contrast*), causal (*because, consequently*), clarifying (*in other words, that is*), concessive (*although, regardless*), or temporal (*before, after*). Yang and Sun (2012); Crossley et al. (2016) have found that the presence of cohesives correlates positively with writing quality in L2 language production.

2.4 Common Text Analysis Platform

As with any other NLP task, linguistic complexity analyses require preprocessing from sentence segmentation and tokenization to dependency and constituency parsing. Pipelines developed specifically for Portuguese include LXService (Branco et al., 2008) and NLPPort (Rodrigues et al., 2018), but they were not designed for complexity feature extraction. Some tools built primarily for this purpose are CohMetrix (Graesser et al., 2004), L2SCA (Lu, 2010), and TAASSC (Kyle, 2016) though they only support English, or require installing software which may create a barrier to end users with inadequate system requirements, or who may otherwise not be technologically inclined. Pylinguistics (Castilhos et al., 2016) was initially created for Portuguese readability assessment and offers some complexity feature extraction but, like the previous systems, is limited in language support and requires

being familiar with computational methods.

Evaluating linguistic complexity is a multilingual problem: various metrics are language-independent while others require minor changes, such as swapping out norm lists compiled for different target languages. The Common Text Analysis Platform (CTAP; Chen and Meurers, 2016) addresses these concerns by providing a web-based, platform-independent resource on which to calculate lexical, syntactic, and discourse complexity measures for multiple languages. CTAP accepts unstructured corpora based on the Unstructured Information Management Architecture (UIMA; Ferrucci and Lally, 2004), and subsequently processes the text through a series of modularized annotators, visualized in Figure 1. Originally developed for English, it has since been extended to German (Weiß and Meurers, 2019) and Italian (Okinina et al., 2020), with support currently being added for Dutch, French, and Spanish.

CTAP has integrated Portuguese by way of Stanza, the Stanford NLP Group's newly introduced, fully neural, end-to-end NLP pipeline that is compatible with up

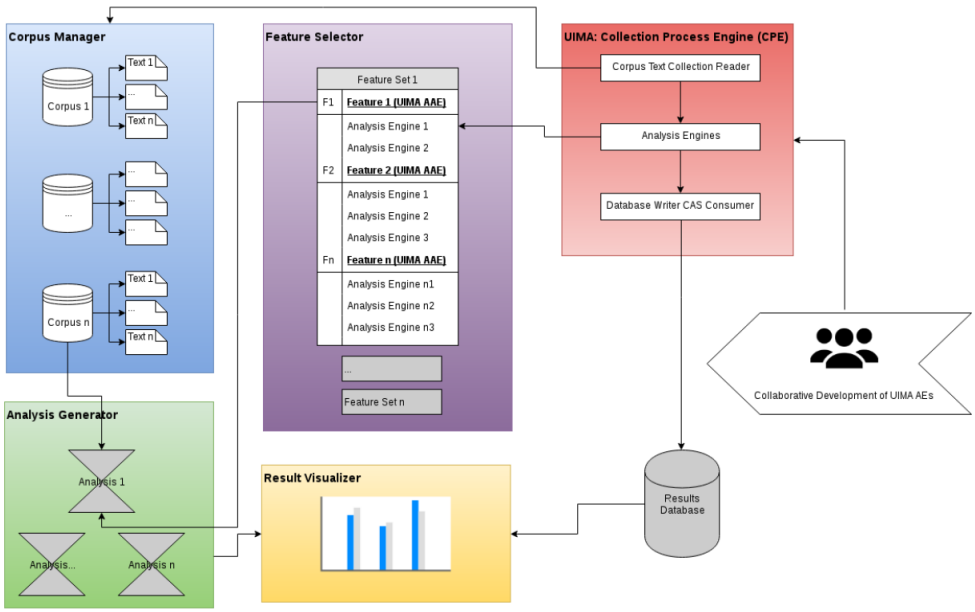


Figure 1: Overview of the CTAP architecture. Source: Chen and Meurers (2016).

to 66 languages (Qi et al., 2020). The LingMod Research Group has created a Java wrapper[1] around a containerized web service[2] hosting Stanza and its model binaries. This effectively lays the groundwork for supporting dozens of other languages out of the box and making CTAP a truly multilingual tool.

3 Data

The evaluation was conducted on the NLI-PT dataset, a compilation of 3,069 texts by European Portuguese learners originally collected for native language identification (del Río et al., 2018; del Río, 2019a). Entries are labeled according to their coarse CEFR proficiency level. The raw texts are accompanied by their tokenizations as well as part of speech, constituency, and dependency annotations—none of which were utilized for this analysis in order to make use of CTAP's end-to-end processing pipeline. Entries are labeled according to the corresponding student's native language and CEFR proficiency level group. The detailed breakdown is presented in Table 2.

	ARA	CHI	DUT	ENG	FRE	GER	ITA	JAP	KOR	POL	ROM	RUS	SPA	SWE	TET	TOTAL
A	6	133	15	174	42	231	319	34	33	39	45	27	257	16	17	1,388
B	6	204	20	173	59	149	182	26	9	22	33	59	266	3	4	1,215
C	2	108	8	68	4	58	64	11	26	16	1	8	91	0	1	466
	14	445	43	415	105	438	565	71	68	77	79	94	614	19	22	3,069

Table 2: Number of texts in NLI-PT, by CEFR level group and L1.

Training, development, and test splits were not defined in the original dataset, and can therefore not be used to directly compare with previous work. Thus, an arbitrary split was generated with `random_state=42`. 80% was used for training with the remaining 20% for evaluation. The division is further contingent upon CTAP's output, the sequence of which is inconsistent for each run. In addition to

[1] `https://github.com/lingmod-tue/stanza-java`
[2] `https://github.com/lingmod-tue/stanza-api`

the unequal representation of proficiency levels, the texts are based on 148 different tasks also suffering from a heavy class imbalance. Instances are not organized or labeled by their topic.

4 Methods

4.1 Preparation

CTAP accepts unstructured text as input, so virtually no preprocessing was performed on the corpus. Raw texts were fed into the Corpus Manager and segmented, tokenized, tagged, and dependency parsed with Stanza using models trained on the Universal Dependencies Bosque treebank (Rademaker et al., 2017).

Contractions and otherwise multi-word expressions such as a clitic pronouns as in (2, 3) needed to be teased apart for part-of-speech tagging and parsing. Expanding surface forms in this way offsets other token boundaries in the original text, wherein the original word's span no longer aligns with those of the newly expanded sentence. The CoNLL-U format is able to represent these multi-word expressions natively during tokenization, which can then be used to update span offsets, greatly simplifying the annotation process.

In order to preserve surface length features such as the number of characters or number of syllables in the original input text, surface forms need to be stored alongside their underlying tokens. A new surface form analysis engine was therefore added to CTAP to accommodate this discrepancy.

Stanza does not include a constituency parser, which is required for extracting syntactic complexity features. Stanford CoreNLP's Shift-Reduce Parser is a suitable alternative already present in CTAP, but does not offer pre-trained models for Portuguese. LX-Center has released a publicly available model (Silva et al., 2010), but it is incompatible with newer versions of the Stanford Parser.

In response, a new model was trained on the CINTIL-Treebank (Branco et al., 2014), after aligning the part of speech tags with the existing UD tagset. Since

certain tags were not present in CINTIL, such as AUX for auxiliary verbs, these values had to be normalized as well. Phrase and bar-level categories were preserved even when their correlate head projections were modified to fit the UD scheme, as in PP for adpositions (ADP) and $\overline{\text{N}}$ for nouns (NOUN).

Customized, language-specific head-finding rules should normally be defined for training a new parsing model. However, a simple left-most head finder has been shown to perform sufficiently well for most tasks (Vadas and Curran, 2011). For this study, using the `LeftHeadFinder`[3] for training the constituency parser resulted in a surprisingly decent baseline

4.2 Feature Engineering

For the complexity analysis, most feature calculations had already been defined for other languages in CTAP and only slight modifications were necessary to include Portuguese. For instance, 21 density features were easily added by declaring the relevant Universal Dependencies tags: Adjective, Adverb, Article, Auxiliary Verb, Cardinal Number, Common Noun, Conjunction, Coordinating Conjunction, Determiner, Function Words, Interjection, Lexical Words, Noun, Particle, Preposition, Punctuation, Pronoun, Proper Noun, Subordinating Conjunction, Symbol, and Verb.

Lexical sophistication measures had also been integrated into CTAP, so it was therefore only necessary to add norm lists for Portuguese. Among them were: age of acquisition (Cameirão and Vicente, 2010), concreteness and imageability (Soares et al., 2017), familiarity (Marques et al., 2007), SUBTLEX frequency information (Soares et al., 2015), and discourse markers (Mendes et al., 2018). Each of these were calculated on a type or token level, further divided into three categories: lexical words (LW), function words (FW), or both (all words; AW).

Language-agnostic variation features previously discussed in Section 2.3.2 as well as the remaining lexical features are displayed in Table 3.

[3] `https://nlp.stanford.edu/nlp/javadoc/javanlp/edu/stanford/nlp/trees/LeftHeadFinder.html`

#	NAME	FORMULA
f_1	TTR	T/N
f_2	RTTR	$T/\sqrt{N}$
f_3	CTTR	$T/\sqrt{2N}$
f_4	LogTTR	$\log(T)/\log(N)$
f_5	UberTTR	$\log^2(N)/\log(N/T)$
f_6	MTLD	McCarthy and Jarvis (2010)
f_{7-12}	Adjective, Adverb, Lexical Type, Modifier, Noun, Verb	X_{type}/L
f_{13}	VV1	V_{type}/L
f_{14}	CVV1	$V_{type}/\sqrt{2V_{token}}$
f_{15}	SVV1	${V_{type}}^2/V_{token}$
f_{16-21}	Age of Acquisition	$X_{type\|token}/(N\|L\|F)$
f_{22-27}	Concreteness	
f_{28-33}	Familiarity	
f_{34-39}	Imageability	
f_{40-45}	SUBTLEX Word Frequency	
f_{46-51}	SUBTLEX Log Word Frequency	
f_{52-57}	SUBTLEX Contextual Diversity	
f_{58-63}	SUBTLEX Log Contextual Diversity	
f_{64-69}	SUBTLEX Top 1-5k; Below Top 5k	
f_{70-76}	SUBTLEX Zipfian Band 1-7	
f_{77-97}	POS Density	X/N

Table 3: Lexical features and their calculations. **Legend:** N = number of tokens, T = number of types, L = number of lexical tokens, F = number of function tokens.

There are 21 dependency-based features, 14 general constituency counts, and 42 syntactic complexity ratios available for Portuguese. The rules for extracting constituency phenomena were defined using Tregex (Levy and Andrew, 2006).

Some morphosyntactic features have been added specifically for Portuguese. Clitic usage (Flores and Barbosa, 2014; Costa et al., 2015), mood (Flores et al., 2017; de Jesus et al., 2019), and clefts (Lobo et al., 2019) are systems that have been demonstrated to follow certain development patterns and may correlate with L2 proficiency growth.

Table 4 catalogues all morphosyntactic features, with Portuguese clefts appearing separately in Table 5 to demonstrate a subset of Tregex rules.

#	NAME	FORMULA
f_{98-118}	Dependency Locality Theory	Gibson (2000)
$f_{119-139}$	Number of phrasal, clausal, or sentential constituents	X
$f_{140-137}$	Mean length of phrase, clause, or T-Unit	W/X
$f_{138-142}$	Clausal, sentential, and T-Unit complexity ratios	Variable
$f_{143-151}$	Complex Nominal, Complex T-Unit, Coordinate Phrase, Dependent Clause, NP, PP, Relative Clause, Verb Cluster, VP	X/S
$f_{152-158}$	Complex Nominals, Coordinate Phrase, NP, PP, Relative Clause, Verb Cluster, VP	X/CL
$f_{159-166}$	Complex Nominal, Coordindate Phrase, Dependent Clause, NP, PP, Relative Clause, Verb Cluster, VP	X/TU
$f_{167-168}$	Prenominal Modifer, Postnominal Modifier	X/NP
$f_{169-174}$	Conditional, Imperfect, Inflected Infinitive, Pluperfect, Preterite, Subjunctive	X/VP
$f_{175-178}$	Proclitic, Enclitic, Mesoclitic, All	
$f_{179-183}$	Clefts	

Table 4: Morphosyntactic features. **Legend:** W = number of words, S = number of sentences, CL = number of clauses, TU = number of T-Units.

#	NAME	TREGEX PATTERN
f_{184}	*é que* Clefts	`S|VP [< ((VP|VERB|AUX [<< /[Éé]/]) [$+ ((CP [< ((SCONJ [< /[Qq]ue/]))]) [< (SCONJ [$+ S])])])]`
f_{185}	It-Clefts	`S|VP [< ((VP|VERB|AUX [<< COPULA]) [$+ ((NP [!< (PRON [< /[Oo]/])]) [< (N' [< (CP\|NP [< ((NP < PRON) [$+ S|NP])])])])])]`
f_{186}	Pseudoclefts	`S|VP [ < ((NP [< ((NP < (PRON [< /[Oo]|[Qq]uem/])) [$+ S\|NP])]) [$+ (S|VP [<< (VP|VERB|AUX [<1 COPULA])])])]`
f_{187}	Inverted Pseudoclefts	`S|VP [< ((VP|VERB|AUX [<< COPULA]) [$+ (NP [<<, (PRON < /[Oo]|[Qq]uem/)] [< N'|CP|S])])]`
f_{188}	WH-Clefts	`(S|VP [< ((VP|VERB|AUX [<< COPULA]) [$+ NP])]) [$+ (NP [< ((NP < PRON) [$+ (S [< (NP [$+ VP])])])])])]`

Table 5: Tregex rules for syntactic complexity based on a constituency parse generated with UD-Bosque tags and left head-finding rules. COPULA refers to the regular expression pattern matching any relevant copular form: `/[Éé]|[Ff](o(i|r(am?|em)?|ssem?))|[Ee]ram?|[Ss](e(r(iam?|ão|á)|jam?)|ão|ido)/`

Cohesion measures are incorporated through both the raw tallies of different classes of connectives (see Section 2.3.4) as well as ratios indicating the usage of certain types in relation to the total number of connectives. Also taken into account are the proportion of single-word and multi-word phrasal connectives.

#	NAME	FORMULA
$f_{189-196}$	Number of connectives	X
$f_{197-206}$	Cohesive complexity	$X/(N\|CO)$

Table 6: Cohesion measures. **Legend:** N = number of tokens, CO = number of connectives.

The remaining 20 measures are global, require little to no linguistic insight, and are calculated straightforwardly. Finally, 15 native languages are represented in the corpus. These were included as categorical features, summing to a total of 241 features.

#	NAME
$f_{207-208}$	Standard deviation: token length (syllable, letter)
$f_{209-211}$	Standard deviation: sentence length (syllable, letter, token)
$f_{212-213}$	Number of word type, token with 2+ syllable
$f_{214-215}$	Percent word type, token with 2+ syllable
$f_{216-221}$	Number of letters, syllables, surface forms, tokens, token types, sentences
$f_{222-223}$	Mean token length: syllable, letter
$f_{224-226}$	Mean sentence length: syllable, letter, token
$f_{226-241}$	Native language (categorical)

Table 7: Shallow complexity features and L1.

4.3 Supervised Machine Learning

Traditional machine learning methods have been commonly used for several CEFR level classification tasks, and have generally performed quite well (Vajjala and Lõo, 2014; Pilán and Volodina, 2016; Vajjala, 2017; del Río, 2019a). Experiments for this linguistic complexity analysis of L2 Portuguese were performed with Random Forests, Support Vector Machines, and Logistic Regression using the scikit-learn library (Pedregosa et al., 2011).

For each system, hyperparameters were tuned using grid search and 10-fold cross-validation. Feature calculations were scaled to mitigate the effect of high cardinality values, which introduce bias against features with lower value ranges during training. Finally, ablation tests were performed by assessing the systems on different feature subsets to observe the performance of each combination.

4.4 Feature Selection

There are multiple calculations that measure similar if not the same metric. For example, the number of surface forms and the number of tokens are very highly correlated. Though usually not a bijective relation, these two features increase almost collinearly. This sort of redundancy should be avoided. Instead, distinct and complementary features have been argued to capture a more diverse, holistic picture of one's L2 system (Norris and Ortega, 2009). Care was therefore taken to prune undesirable features and reduce model variance prior to evaluation.

Correlation-based feature subset selection (CfsSubsetEval; Hall, 1998) from the WEKA library (Hall et al., 2009) was employed to incrementally purge highly correlated features, while retaining those with higher predictive ability of the class. Out of the 241 available features, including each student's native language, 53 were returned, all of which are catalogued in Table 8. Relevant features will be elaborated on in Section 6.

Table 9 ranks the best features by their information gain. Many of the features appearing in this list have been connected to a more complex language system. The

f_{1-2}	SUR	Mean Sentence Length in Letters, Tokens
f_3	SUR	Mean Token Length in Syllables
f_4	SUR	Number of Letters
f_5	SUR	Percentage of Tokens with More Than 2 Syllables
f_6	SUR	Percentage of Word Types with More Than 2 Syllables
f_7	SUR	SD Sentence Length in Letters
f_8	SUR	SD Token Length in Letters
f_{9-11}	LEX	Age of Acquisition (AW Token; AW, LW Type)
f_{12-14}	LEX	Concreteness (AW, FW Token; LW Type)
f_{15-17}	LEX	Familiarity (AW, LW Type; LW Token)
f_{18-20}	LEX	Imageability (AW Token; AW, LW Type)
f_{21}	LEX	SUBTLEX Contextual Diversity (FW Token)
f_{22-25}	LEX	SUBTLEX Frequency Band 2, 3, 4, 5
f_{26}	LEX	SUBTLEX Frequency Below Top 5000
f_{27-29}	LEX	SUBTLEX Log Word Frequency (AW Token; AW, FW Type)
f_{30-32}	LEX	SUBTLEX Word Frequency (AW, FW Type; FW Token)
f_{33}	LEX	Corrected Verb Variation 1
f_{34-38}	LEX	POS Density: Cardinal Number, Interjection, Punctuation, Subordinating Conjunction, Symbol
f_{39}	MSY	"é que" Cleft per VP
f_{40}	MSY	Dependent Clause per Clause
f_{41}	MSY	Dependent Clause per Sentence
f_{42}	MSY	Clause per T-unit
f_{43}	MSY	Mean Length of Noun Phrase
f_{44-46}	MSY	Number of Dependent Clauses, It-Clefts, Prenominal Noun Modifiers
f_{47}	MSY	Number of Proclitics per VP
f_{48-49}	MSY	Verb per VP: Inflected Infinitive, Subjunctive
f_{50-51}	COH	Connectives per Token: Additive, Concessive
f_{52}	COH	Number of Connectives
f_{53}	L1	Russian

Table 8: All features returned by CfsSubsetEval, unranked.

number of letters (#1) is a general length-based measure that has been shown to be a powerful indicator when paired with other finer-grained measures (Norris and Ortega, 2009). Ortega (2012) asserts that complex noun phrases such as (#3) are characteristic of advanced learners, and had previously found that the amount of subordination as in (#6) is relevant (Wolfe-Quintero et al., 1998; Ortega, 2003). Finally, Lu (2012) concluded that CVV1 (#9) is among the best features for assessing lexical variation.

The remaining features are based on lexical sophistication. A plethora of these measures are also present in the entire returned feature subset. These broadly measure the use of "advanced" words—often corresponding to lower frequency—that demonstrate higher lexical proficiency (Crossley and Skalicky, 2017). More than half of the top ten features by information gain are related to lexical sophistication: imageability (all and lexical word types), word frequencies (all word types), concreteness (all and lexical word types), and age of acquisition (lexical word types). Crossley et al. (2012) found that imageability was the strongest predictor of all which is in line with the results returned by CfsSubsetEval, being the second best

#	FEATURE NAME
1	Number of Letters
2	Imageability (AW Type)
3	Number of Prenomial Noun Modifier
4	Imageability (LW Type)
5	SUBTLEX Word Frequency (AW Type)
6	Number of Dependent Clauses
7	Concreteness (LW Type)
8	Age of Acquisition (LW Type)
9	Corrected Verb Variation 1
10	Concreteness (AW Token)

Table 9: Top 10 features ranked by information gain according to CfsSubsetEval.

feature after the general length measure. Chen and Meurers (2018) also assert that averaging word frequency information using frequency bands characterizes text readability, especially when normalized. Four different Zipfian bands appear, excluding the first band of highly frequent words as well as the lowest bands containing the most rare and obscure ones. Several other frequency features from the SUBTLEX corpus are relevant, further stressing the importance of lexical frequency in measuring growth.

There are many other language-agnostic features returned by CfsSubsetEval that have been established as anchors for complexity analyses in other languages. Notable are the Corrected Verb Variation 1 (Lu, 2012; Vajjala and Lõo, 2014), density of interjections and subordinating conjunctions (Vajjala and Lõo, 2014), dependent clauses per clause (Housen and Bulté, 2012), and clause per T-unit (Vyatkina, 2012).

5 Results

Average scores for 10-fold cross validation are reported along with the held-out test set. Following del Río (2019a), accuracy is used as the main evaluation metric with F1-score also provided due to the class imbalance. Since there is no test set provided by NLI-PT, direct comparisons are not possible.

For individual feature groups, lexical features are the clear winner. This is also the largest class with 97 different calculations. Morphosyntactic features perform only slightly better than surface ones, even with a 51-feature disparity between the two, showing that surface features alone can explain a large amount of variance between proficiency levels. Cohesive devices are also able to exceed naïve baselines and score similarly to surface features, but are generally the weakest set.

Using the feature subset provided by CfsSubsetEval, accuracy scores are on par with the entire feature set, though the F1-score reveals a noticeable dip when accounting for the class imbalance. Logistic Regression and SVMs are best when using all features, but Random Forests score better when using the subset.

	KIND	SIZE	10-FOLD CV		TEST	
			μ-F1	μ-ACC	F1	ACC
Random Baseline			0.33	0.33	0.33	0.33
Majority Baseline			0.20	0.45	0.20	0.45
Random Forest			0.50	0.62	0.49	0.61
Logistic Regression	SUR	20	0.46	0.63	0.47	0.63
Support Vector Machine			0.46	0.63	0.44	0.62
Random Forest			0.56	0.67	0.55	0.65
Logistic Regression	LEX	97	**0.57**	**0.66**	**0.57**	**0.68**
Support Vector Machine			**0.57**	**0.66**	**0.57**	0.64
Random Forest			0.50	0.63	0.53	0.65
Logistic Regression	MSY	71	0.48	0.63	0.47	0.62
Support Vector Machine			0.47	0.64	0.48	0.64
Random Forest			0.45	0.56	0.50	0.59
Logistic Regression	COH	18	0.44	0.61	0.44	0.61
Support Vector Machine			0.44	0.61	0.44	0.62

Table 10: Trials using surface, lexical, morphosyntactic, and cohesive feature sets.

	KIND	SIZE	10-FOLD CV		TEST	
			μ-F1	μ-ACC	F1	ACC
Random Baseline			0.33	0.33	0.33	0.33
Majority Baseline			0.20	0.45	0.20	0.45
Random Forest			0.56	0.67	0.58	0.68
Logistic Regression	CSE	53	0.58	0.67	0.56	0.68
Support Vector Machine			0.59	**0.68**	0.55	0.66
Random Forest			0.56	0.67	0.54	0.67
Logistic Regression	ALL	241	0.59	0.67	**0.61**	**0.69**
Support Vector Machine			**0.62**	**0.68**	**0.61**	0.67

Table 11: Scores achieved by systems using CfsSubsetEval (CSE) vs. all features.

The confusion matrix for the test set in Table 12 shows that CEFR-A is the easiest class to predict, with CEFR-B faring slightly worse. CEFR-C is by far the worst predicted class, with an outright majority of instances being classified as its lower neighbor, followed by CEFR-A.

Pred	A	B	C
True			
A	**225**	48	5
B	58	**177**	8
C	26	**53**	14

Table 12: Confusion matrix for Logistic Regression.

6 Discussion

6.1 Portuguese Features

The Portuguese clitic system is complex, with various phonological and syntactic constraints governing their form and position. Usually, object pronouns are realized as enclitics, immediately succeeding a verb attached by a hyphen as in (7). Proclitics, however, are required in situations affected by subordination, interrogation, quantification, negation, and aspect (Flores and Barbosa, 2014). In this case, the object pronoun precedes the verb, and is not attached by a hyphen as in (8).

(7) Ele viu-o
he see.3.SG.PAST=3.SG.MASC
"He saw him" (Flores and Barbosa, 2014)

(8) O João não a viu
the João not 3.SG.FEM=see.3.SG.PAST
"João never saw her" (Flores and Barbosa, 2014)

Due to these restrictions, it has been found that early learners generalize enclisis, and later acquire the contexts which call for proclisis (Flores and Barbosa, 2014). The reverse has not been observed, however: proclisis is never generalized in situations of enclisis.[4]

The results of the feature analysis for these two patterns are revealed in Figure 2. Enclitic distribution is similar across proficiency levels, whereas proclitics are clearly picked up later. Even with many outliers, proclisis in CEFR-A is very close to zero on average. The rate increases gradually in CEFR-B and even moreso in CEFR-C, with fewer outliers appearing in each class.

[4]It is worth noting that this is characteristic of European Portuguese in particular, on which the learner corporus is based, as the converse is true for Brazilian Portuguese: proclisis is generally preferred in most contexts with enclisis being considered formal, pedantic, or even archaic (Galves et al., 2005; Simoes, 2006). This also explains the acceptability of (1) in Section 2.1; strictly speaking, proclisis is canonically forbidden in the sentence-initial position, with Brazilian Portuguese having adopted looser conditions on such placement.

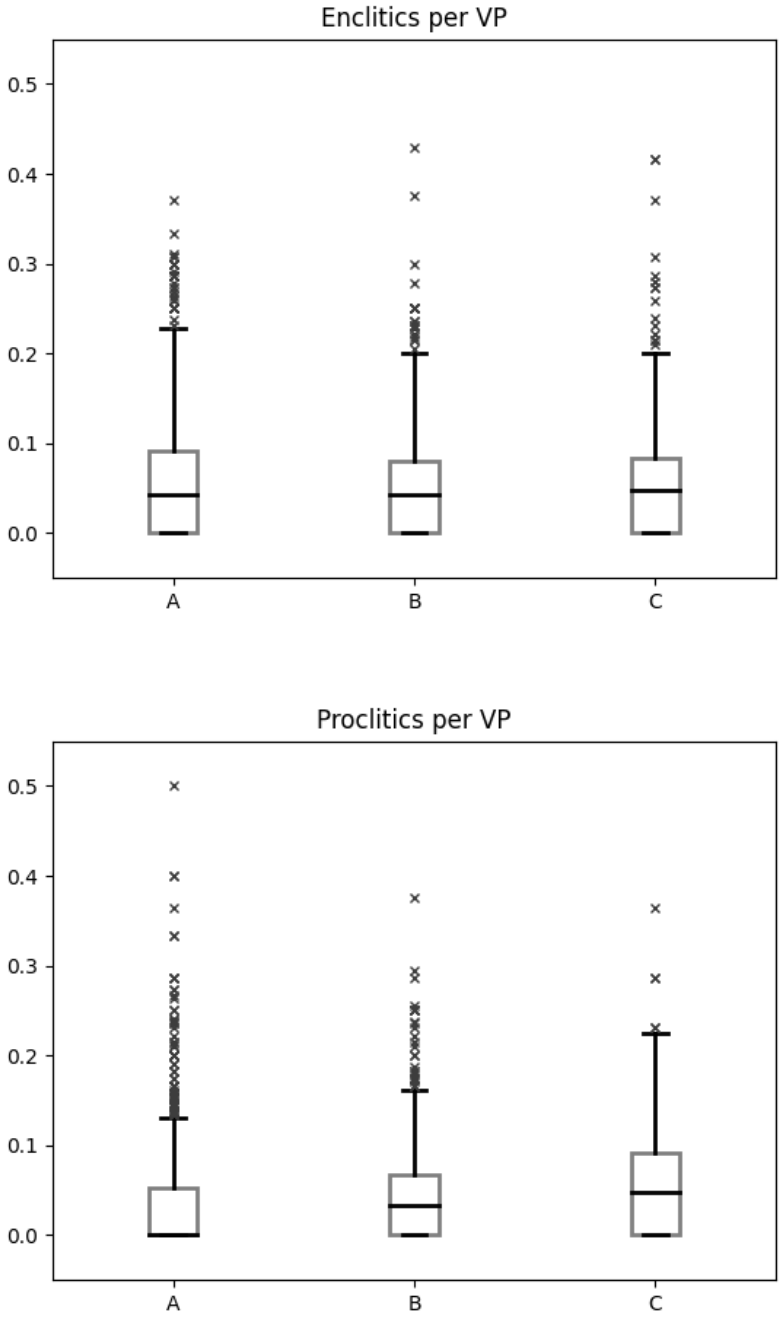


Figure 2: Use of clitic constructs in NLI-PT.

It follows that the number of proclitics employed is better correlated with proficiency level than enclitic use, as indicated by CfsSubsetEval. This corroborates the claim that proclitics are acquired in a later stage of development (Flores and Barbosa, 2014; Costa et al., 2015).

Curiously, the elusive mesoclitic is absent from the feature list. Being restricted to the future and conditional tenses, there are fewer opportunities for its use. The arcane manner in which it must split a verbal stem from its inflection may also deter students, preferring instead the more familiar clitic varieties. Indeed, average use rates for mesoclitics are near-zero for all levels, with only a handful of outliers almost equally distributed across classes, negating any potential effect.

Moving on to verbal morphology, two inflectional paradigms of interest make an appearance in CfsSubsetEval: the subjunctive mood, a pattern difficult for many students learning Romance languages, and the extremely rare inflected infinitive, present in few other languages.

The inflected infinitive is an unusual form that allows for person and number agreement with an untensed verb. There is always a grammatically-licensed finite alternative, meaning their use is never mandatory (Iverson and Rothman, 2008). Rothman et al. (2010, 2013) have suggested that these patterns are acquired by late childhood, but may still cause problems even for adults who may accept certain pragmatic cues and draw inferences that alter the intended meaning of the utterance. Unfortunately, most research on this phenomenon have focused on monolingual native speakers in L1 acquisition.

The usage distribution across proficiency levels in NLI-PT is seemingly random, with no general trend in increased usage for higher proficiency levels. It is possible that, since the first and third person singular forms are inflected with a null morpheme and are therefore homographs with each other as well as their uninflected form, this could be either unintended use by the student or even misidentified by the Stanza parser. These conjectures, however, still do not explain its predictive value according to CfsSubsetEval, other than the fact that it is not strongly correlated to other input variables.

Subjunctive, on the other hand, is much easier to interpret. Romance languages are well known for their extensive verbal morphology which can be overwhelming for students coming from an L1 without such variation. Furthermore, this mood is employed in non-epistemic situations that would otherwise make use of the indicative. While commonly triggered by certain verbs such as *querer* "to want", it may be difficult to identify when required by less common or more opaque predicates (Flores et al., 2017; de Jesus et al., 2019).

Beginner students are less likely to use the subjunctive, being introduced to it at an intermediate stage. This is clearly evident in the distribution in NLI-PI, illustrated in Figure 3. The vast majority of level A students do not use subjunctive at all, even though there are many outliers. However, the distribution between levels B and C are almost the same, though slightly less for the more advanced. This could be due to the fact that this mood is introduced and practiced more at the intermediate stage. The distribution indicates that it is a good proxy by which to discriminate between beginner texts and the higher levels.

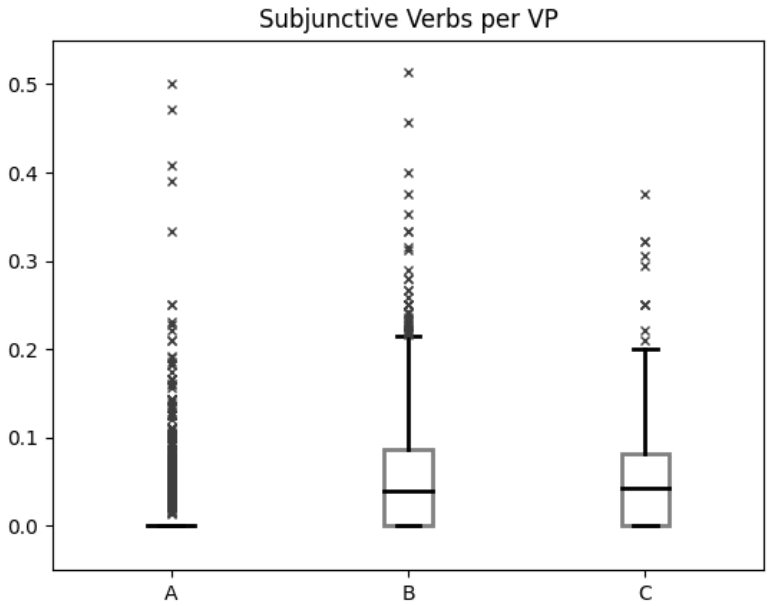


Figure 3: Use of subjunctive in NLI-PT.

The various clefting strategies available in Portuguese give rise to the potential for great syntactic variation. Of the five cleft constructions described in Section 2.1, only two were identified often enough in the corpus to make any impact: it-clefts and *é que* clefts. *WH*-clefts were not found at all, and the remaining pseudoclefts and inverted pseudoclefts were sparse. However, this may be the result of poor tagging and parsing of the learner texts using models trained on native text, and consequently the possibly unintuitive matching patterns described in Table 5, which were designed for the parsing model trained using simple left head-finding rules.

Lobo et al. (2015) found that both it-clefts and *é que* clefts are acquired earlier than the remaining varieties and are generally widespread in child speech. In NLI-PT, standard clefts are used less frequently overall than *é que* clefts, but the former is employed similarly across proficiency levels compared to the latter, which appears more often in higher proficiency levels. This also suggests that it is helpful in classifying higher levels compared to CEFR-A.

Interestingly, few features appear that target morphology, which would have been expected given the highly inflectional nature of Portuguese in addition to the results obtained by Vajjala and Lõo (2014) for Estonian, which conclude that morphological features can be highly predictive. Del Río (2019a)'s use of POS *n*-grams with fine-grained morphological information seems to fill this gap, with Vajjala and Rama (2018) also observing promising results using word and POS *n*-grams, though Vajjala and Lõo (2014) reported low scores using POS *n*-grams.

Encoding native language as a categorical feature has also been postulated to be a useful feature in proficiency classification (Vajjala, 2017), but no information gain was observed in this study with the exception of Russian being the only one to appear as a viable L1 feature. This is an interesting development, since certain vocabulary and syntactic choices may be characteristic of a particular L1 and may have potential transfer effects. For example, Spanish speakers may be expected to produce a higher rate of grammatical sentences even at lower levels given the close proximity between the two languages. This substantiates Crossley and McNamara

(2011), having shown homogeneity between learners of varying L1s.

6.2 Comparisons

Del Río (2019a) in her first experiments with Portuguese proficiency classification found complexity measures, which she refers to as "descriptive" features, among the weakest predictors, with the best scores being achieved using bag-of-words and POS *n*-grams. However, the set of approximately 39 complexity features she used is much smaller than the amount employed in this study.

Later, del Río (2019b) obtained generally poor results in her experiments training systems on Spanish for cross-lingual classification, though there are a few interesting takeaways. While bag-of-words, POS, and dependency linguistic features outperform complexity features, their interclass results are quite inconsistent with lows ranging from 0.19 F-1 macro for classifying CEFR-B with dependency n-grams, to 0.68 F-1 macro in CEFR-A using POS n-grams. CEFR-C was not able to be classified with these features at all, scoring 0.0 F1-macro across the board. Complexity features, however, are more stable: general measures scored a high of 0.60 F1-macro for A, with other subsets scoring between 0.40 and 0.48. Among levels B and C, F1 scores are almost the same for all subsets: 0.48-0.49 for B and 0.25-0.30 for C. This suggests that, while the small set of complexity features on their own may not have been the best, they are still insightful as predictive variables.

Combining these complexity measures with POS *n*-grams resulted in further accuracy gains for classes A and B, though F1 scores for level C surprisingly plummeted back to 0.0. Given that the best results she observed for Spanish monolingual classification, however, were achieved through a combination of POS *n*-grams and a larger amount of complexity features, it follows that using these in tandem for monolingual Portuguese classification could lead to improved scores.

6.3 Detractors

Recall that linguistic complexity and accuracy are separate entities in the CAF triad for measuring second language acquisition (Housen and Kuiken, 2009). While complexity focuses on elaborateness and variation in learner expression, it is not concerned with errors made by students. However, the veracity of these feature calculations is susceptible to grammar and mechanical errors. Incorrect orthography or word choice obfuscates the student's intent, leading to incorrect tagging and parses and a contaminated analysis.

For example, one CEFR-A student achieved an impressive 50% subjunctive verb usage per verb phrase, clearly visible in Figure 3. The subjunctive mood is often acquired later and it is therefore unlikely that a student would be capable of producing this construct so early, much less as often as was indicated. Upon closer inspection, it appears that this student—whose native language is Spanish—erroneously used the subjunctive present forms *tome* and *volte* instead of the indicative past *tomei* and *voltei*, which is clearly negative transfer from Spanish simple past, e.g. *tomé*.

In the previous example, the verbal inflection was correctly identified, tagged, and parsed even though it was used incorrectly. Other situations result in a bad parse that propagates throughout the pipeline, culminating with an incorrect parse tree and, accordingly, inaccurate feature calculations. Sentence (9) illustrates one such instance. The verb *tendem* takes the optional prepositions *a* or *para* depending on context. This student has introduced an inappropriate preposition *de* that is labeled by a confused tagger as a subordinate conjunction in what is really a simple sentence with no subordination.

(9) * Os actores tendem de ser mais dramáticos também
The actors tend to to be more dramatic too

While Stanza otherwise provides a fairly competent parse of input text and relatively high accuracy is generally achieved for tokenization, tagging, lemmatization, and dependency parsing, there are still some shortcomings due to its data-driven

approach and comparably smaller training set. For example, it seems that mesoclitics are not well-represented in the UD training corpus, because they are often poorly tokenized. This can be minor, as is the case with *telefonar-lhes-ia* "I would call them", which is split into *telefonará* + *lhes*, changing the tense from conditional to future. In other situations, it can result in very obscure results, such as merging an object pronoun with the inflected verb, introducing a new pronoun that was not present in the original mesoclitic, or even creating a brand new word not licensed by the language. This was resolved by creating hard-coded rules, searching for up to two object pronouns between hyphens and an inflectional suffix from one of the two appropriate tenses.

Another example would be the adverbial modifier "how" and the first person singular conjugation of "to eat", which are homonymous: *como*. Only the former is represented in the UD training set which naturally leads to a questionable tag in a situation where the verb is clearly required. The parse tree for a sentence affected by this is depicted in Figure 4.

Incorporating learner errors could provide a further dimension in which to better predict proficiency level. Vajjala (2017); Weiß and Meurers (2019) both found that errors in conjunction with complexity features were useful in their proficiency classification experiments. Further, del Río (2019a) hypothesized that bag-of-words performed strongest for Portuguese proficiency classification because it can help identify orthographic mistakes. POS *n*-grams, which were the next best feature in isolation, are also able to capture, for example, agreement errors in gendered nouns and adjectives. Both of these performed better in her experiments over the linguistic complexity features, so combining their effects with the augmented complexity feature set introduced in this work could be beneficial, as would incorporating other error patterns such as the incorrect expansion of mandatory contractions (see Section 2.1) and allomorphic variation of clitics.

Task variation may also negatively impact the classifier due to varied learner language across different prompts. As mentioned in Section 3, texts are not labeled according to their associated topic and resolving topics from the texts themselves

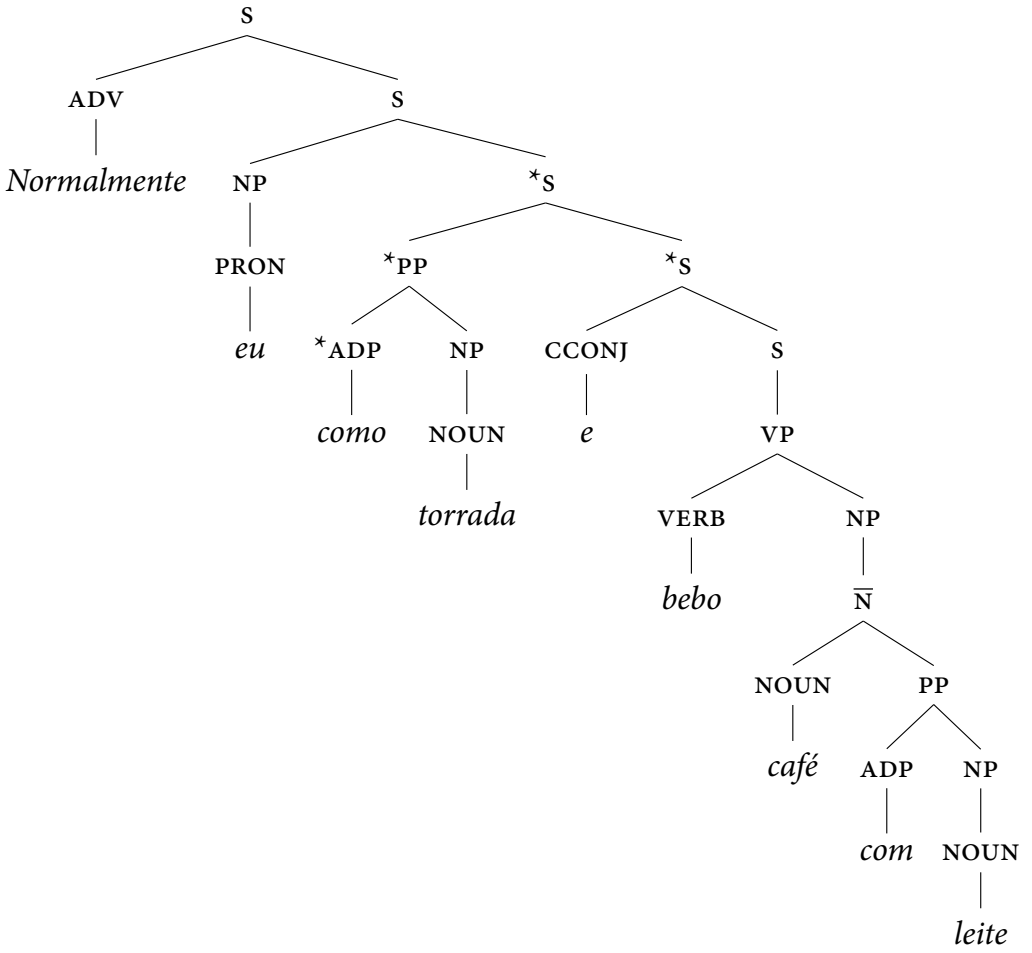


Figure 4: Constituency tree for "Normally I eat toast and drink coffee with milk". The verb *como* "eat" is mistaken for its prepositional homonym, projecting to a prepositional phrase instead of a verbal phrase. The error percolates to the sentence conjunct, which should be a complementizer phrase.

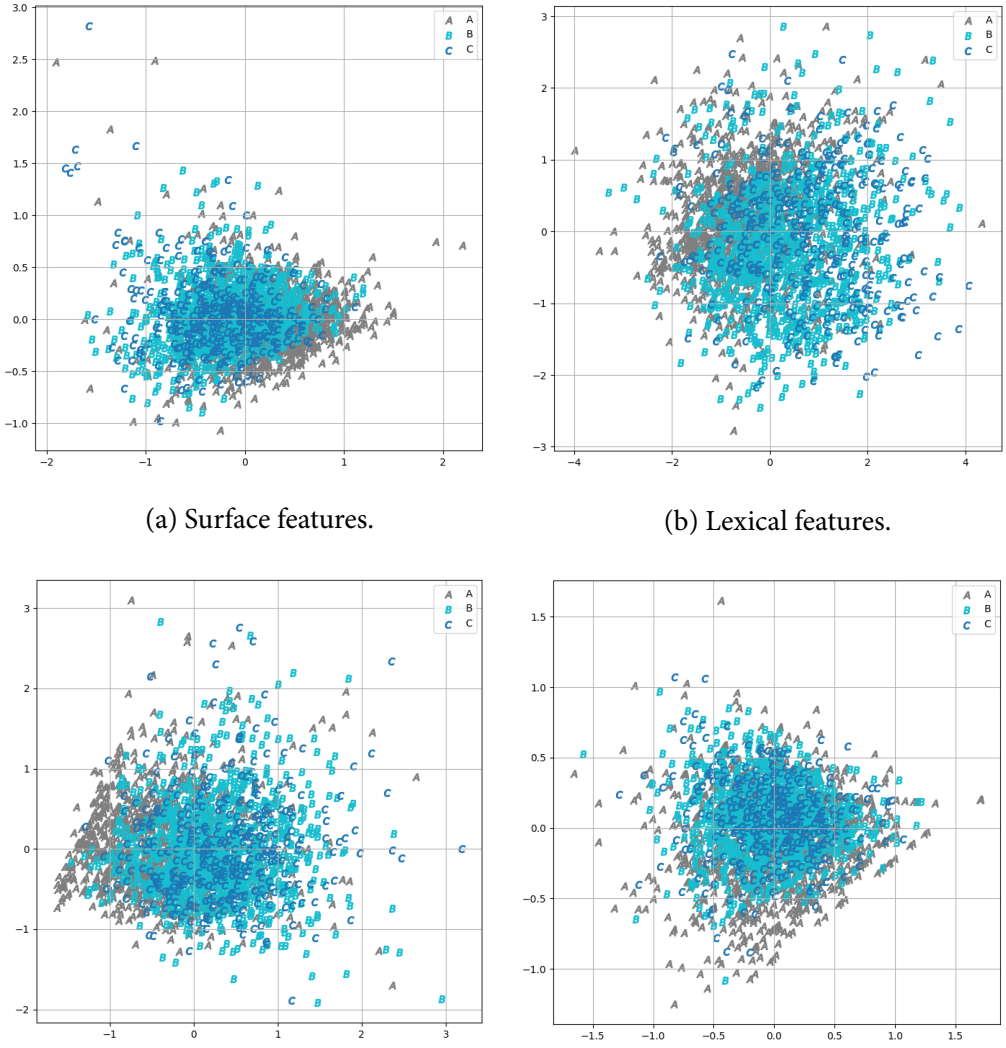
is non-trivial. Since linguistic complexity analyses are sensitive to the topic around which each text is based, as varied writing prompts can elicit different language patterns (Tracy-Ventura and Myles, 2015; Yang et al., 2015), this has the potential to adversely influence the analysis because similar tasks cannot be grouped together during evaluation.

6.4 Proficiency Levels

Most other CEFR proficiency evaluations have determined that levels A and C classes are easier to classify, while this analysis observes C as being the hardest with A and B achieving similarly good results. NLI-PT has a large class imbalance, containing a similar number of entries for classes A and B, but roughly one third the amount of each for level C. Del Río (2019a) accounts for this by performing trials on both the full dataset as well as a balanced subset, equally representing all classes. While predictions for C improve dramatically compared to other levels with the balanced set, it remains the class most often misclassified.

For a three-way classification problem with a random baseline of 33%, complexity measures performed somewhat decently with consistent scores in the high 60% accuracy range, rivaling the existing work performed by del Río (2019a). However, since proficiency level is measured on a spectrum, it follows that there is much overlap between adjacent levels, to say nothing of the variance caused by human error or bias when manually assigning levels. As such, it is of interest to visualize how closely these levels are at an empirical level.

Dimensionality reduction was conducted using principal component analysis with scikit-learn to inspect the trends between two automatically-extracted features that were deemed to be predictive of the target variable. Each feature group was performed separately, presented in Figure 5, in addition to the combined feature set, shown in Figure 6. For surface, lexical, and morphosyntactic features, CEFR-A is the only group that appears to cluster somewhat distinctly. There is much overlap between CEFR-B and CEFR-C across all feature sets, explaining some of the difficulty in distinguishing them.



(a) Surface features.

(b) Lexical features.

(c) Morphosyntactic features.

(d) Cohesion features.

Figure 5: Principal component analysis for different feature subsets.

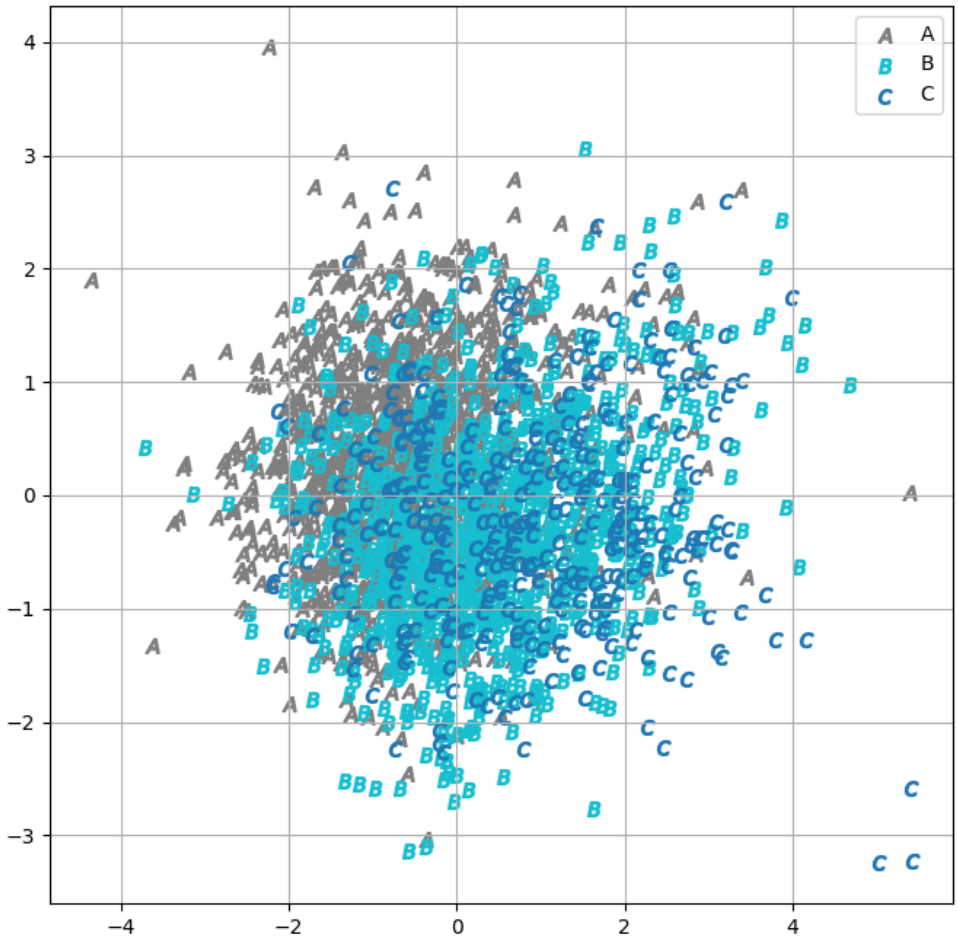


Figure 6: Principal component analysis for combined features.

7 Conclusion

This thesis has presented the first purely linguistic complexity analysis for the automatic proficiency classification of L2 Portuguese. Until recently, Portuguese had been largely neglected in this space, though its prominence as a global language calls for its inclusion. A large and diverse set of complexity features were compiled, with several global, lexical, morphosyntactic, and discursive features demonstrating good predictive ability, corroborating the assertions of similar analyses for other languages. This thesis has shown that complexity features can indeed be leveraged to classify proficiency, though other metrics including part-of-speech *n*-grams, errors, and task information may be combined for better results. In addition, resources fine-tuned for Portuguese learner text will likely improve accuracy even further.

References

Manuela Ambar et al. Overtly/Non-Overtly Inflected Infinitives in Romance. 2017.

Johan Stig Berggren, Taraka Rama, and Lilja Øvrelid. Regression or classification? Automated Essay Scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102, Florence, Italy, August 2019. Association for Computational Linguistics.

António Branco, Francisco Costa, Pedro Martins, Filipe Nunes, João Silva, and Sara Silveira. LX-Service: Web Services of Language Technology for Portuguese. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

António Branco, Catarina Carvalheiro, Francisco Costa, Sérgio Castro, João Silva, Cláudia Martins, and Joana Ramos. DeepBankPT and Companion Portuguese Treebanks in a Multilingual Collection of Treebanks Aligned with the Penn Treebank. pages 207–213, 10 2014. ISBN 978-3-319-09760-2.

Marc Brysbaert and Boris New. Moving beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. *Behavior research methods*, 41:977–90, 11 2009.

Manuela L. Cameirão and Selene G. Vicente. Age-of-acquisition norms for a set of 1,749 Portuguese words. *Behavior Research Methods*, 42:474–480, 2010.

John B. Carroll. Language and Thought. 1964.

John B. Carroll and Margaret N. White. Word Frequency and Age of Acquisition as Determiners of Picture-Naming Latency. *Quarterly Journal of Experimental Psychology*, 25(1):85–95, 1973.

S. Castilhos, V. Woloszyn, D. Barno, and L. K. Wives. Pylinguistics: an open source library for readability assessment of texts written in Portuguese. *Revista de Sistemas de Informação da FSMA*, 18, 2016. ISSN 1983-5604.

Xiaobin Chen and Detmar Meurers. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading*, 41:486–510, 2018.

Xiaobin Chen and Walt Detmar Meurers. CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *CL4LC@COLING 2016*, 2016.

João H. C. Costa, Alexandra Fiéis, and Maria Raika Guimarães Lobo. Input variability and late acquisition: clitic misplacement in European Portuguese. *Lingua*, 161:10–26, 2015.

S. Crossley and D. McNamara. Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing*, 20:271–285, 2011.

S. Crossley, Tom Salsbury, and D. McNamara. Predicting the proficiency level of language learners using lexical indices. *Language Testing*, 29:243 – 263, 2012.

S. Crossley, K. Kyle, and D. McNamara. The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32:1–16, 2016.

Scott A. Crossley and Stephen Cameron Skalicky. Examining Lexical Development in Second Language Learners: An Approximate Replication of Salsbury, Crossley McNamara. *Language Teaching*, 52:385–405, 2017.

Alice Pereira de Jesus, Rui Cunha Marques, and Ana Lúcia Santos. Semantic features in the acquisition of mood in European Portuguese. *Language Acquisition*, 26:302 – 338, 2019.

Iria del Río. Automatic proficiency classification in L2 Portuguese. *Procesamiento del Lenguaje Natural*, 63:67–74, 2019a.

Iria del Río. Linguistic features and proficiency classification in L2 Spanish and L2 Portuguese. In *Proceedings of the 8th Workshop on NLP for Computer Assisted Language Learning*, pages 31–40, Turku, Finland, September 2019b. LiU Electronic Press.

Iria del Río, Marcos Zampieri, and Shervin Malmasi. A Portuguese Native Language Identification Dataset. In *BEA@NAACL-HLT*, 2018.

Daniel Dugast. *Théâtre et dialogue: études de lexicométrie organisationnelle sur les théâtre de Corneille, Racine et Giraudoux, sur des pièces de Corneille, Racine, Molière et Beaumarchais, sur un entretien entre Maurice Clavel et Philippe Sollers, précédées d'un historique des méthodes quantitatives en lexicologie et des fondements d'une explication nouvelle: UBER*. Travaux de linguistique quantitative. Verlag nicht ermittelbar, 1979. ISBN 9782051000598.

Rod Ellis. Task-based Language Learning and Teaching. 2003.

David Ferrucci and Adam Lally. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3-4):327–348, Sep 2004.

Cristina Flores and Pilar Barbosa. When reduced input leads to delayed acquisition: A study on the acquisition of clitic placement by Portuguese heritage speakers. *International Journal of Bilingualism*, 18:304 – 325, 2014.

Cristina Flores, Ana Lúcia Santos, Alice Pereira de Jesus, and Rui Marques. Age and input effects in the acquisition of mood in Heritage Portuguese. *Journal of child language*, 44 4:795–828, 2017.

Charlotte Galves, M. A. Moraes, and Ilza Ribeiro. Syntax and Morphology in the Placement of Clitics in European and Brazilian Portuguese. *Journal of Portuguese Linguistics*, 4:143–177, 12 2005.

Morton Ann Gernsbacher. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of experimental psychology. General*, 113 2:256–81, 1984.

Edward Gibson. The dependency locality theory: A distance-based theory of linguistic complexity. 2000.

Arthur C. Graesser, Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, Computers*, 36:193–202, 2004.

Sylviane Granger and S. C. Tyson. Connector usage in the English essay writing of native and non-native EFL speakers of English. 1996.

Pierre Guiraud. Problémes et méthodes de la statistique linguistique. 1960.

M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18, 2009.

Julia Hancke and Detmar Meurers. Exploring CEFR classification for German based on rich linguistic modeling. pages 54–56, 01 2013.

Gustav Herdan. Quantatative Linguistics or Generative Grammar? 1964.

Alex Housen and Bram Bulté. *Defining and operationalising L2 complexity*, pages 21–46. 10 2012. ISBN 9789027213068.

Alex Housen and F. Kuiken. Complexity, Accuracy and Fluency in Second Language Acquisition. *Applied Linguistics*, 30, 12 2009.

Kellogg W. Hunt. Grammatical structures written at three grade levels. 1965.

Michael Iverson and Jason A. Rothman. The Syntax-semantics interface in L2 acquisition: genericity and inflected infinitive complements in non-native Portuguese. 2008.

Stephen D. Krashen. The "Fundamental Pedagogical Principle" in Second Language Teaching. 1981.

Kristopher Kyle. Measuring Syntactic Development in L2 Writing: Fine Grained Indices of Syntactic Complexity and Usage-Based Indices of Syntactic Sophistication. 2016.

Roger Levy and Galen Andrew. Tregex and Tsurgeon: tools for querying and manipulating tree data structures. In *LREC*, 2006.

Maria Lobo, Ana Lúcia Santos, and Carla Soares-Jesel. Syntactic Structure and Information Structure: The Acquisition of Portuguese Clefts and Be-Fragments. *Language Acquisition*, 23, 03 2015.

Maria Lobo, Ana Lúcia Santos, Carla Soares-Jesel, and Stéphanie Vaz. Effects of syntactic structure on the comprehension of clefts. *Glossa*, 4:1–23, 2019.

Xiaofei Lu. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15:474–496, 2010.

Xiaofei Lu. The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, 96:190–208, 06 2012.

J. Frederico Marques, Francisca L Fonseca, Sofia Morais, and Inês Pinto. Estimated age of acquisition norms for 834 Portuguese nouns and their relation with other psycholinguistic variables. *Behavior Research Methods*, 39:439–444, 2007.

Philip McCarthy and Scott Jarvis. Vocd: A theoretical and empirical evaluation. Language Testing, 24, 459-488. *Language Testing - LANG TEST*, 24:459–488, 10 2007.

Philip McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42:381–92, 05 2010.

Amália Mendes, Iria del Río, Manfred Stede, and Felix Dombek. A Lexicon of Discourse Markers for Portuguese – LDM-PT. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

Boris New, Marc Brysbaert, Jean Veronis, and Christophe Pallier. The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28(4):661–677, 2007.

John Norris and Lourdes Ortega. Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. *Applied Linguistics - APPL LINGUIST*, 30:555–578, 12 2009.

Council of Europe and Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. European year of languages. Cambridge University Press, 2001. ISBN 9780521803137.

Nadezda Okinina, Jennifer-Carmen Frey, and Zarah Weiß. CTAP for Italian: Integrating Components for the Analysis of Italian into a Multilingual Linguistic Complexity Analysis Tool. In *LREC*, 2020.

Lourdes Ortega. Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College Level L2 Writing. *Applied Linguistics*, 24(4):492–518, 12 2003. ISSN 0142-6001.

Lourdes Ortega. *Interlanguage complexity: A construct in search of theoretical renewal.* 2012.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76 1: Suppl:1–25, 1968.

Gabriele Pallotti. CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics - APPL LINGUIST*, 30:590–601, 12 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Ildikó Pilán and Elena Volodina. Classification of Language Proficiency Levels in Swedish Learners' Texts. 2016.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *ArXiv*, abs/2003.07082, 2020.

Sandra Quarezemin, Núbia Saraiva Ferreira, Ana Lívia Agostinho, Giuseppe Varaschin, Karina Zendron da Cunha, and Luciano Denardin de Oliveira. The Handbook of Portuguese Linguistics, editado por W. Leo Wetzels, João Costa e Sérgio Menuzzi. 2018.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)*, pages 197–206, Pisa, Italy, September 2017.

Ricardo Rodrigues, Hugo Gonçalo Oliveira, and Paulo Gomes. NLPPort: A Pipeline for Portuguese NLP (Short Paper). In *SLATE*, 2018.

Jason Rothman. Heritage speaker competence differences, language change, and input type: Inflected infinitives in Heritage Brazilian Portuguese. *International Journal of Bilingualism - INT J BILING*, 11, 01 2007.

Jason Rothman, Acrisio Pires, and Ana Santos. Acquisition of inflected and uninflected infinitives child L1 European Portuguese. 01 2010.

Jason Rothman, Inês Duarte, Acrisio Pires, and Ana Lúcia Santos. How early after all? Inflected infinitives in European and Brazilian Portuguese L1 production. 2013.

João Ricardo Silva, António Branco, Sérgio Castro, and Ruben Reis. Out-of-the-Box Robust Parsing of Portuguese. In *PROPOR*, 2010.

Antonio Roberto Simoes. Clitic Attachment in Brazilian Portuguese. *Hispania*, 89: 380, 05 2006.

Ana Paula Soares, João Junior da Silva Machado, Ana Elizabeth Santos Costa, Álvaro Iriarte, Alberto Simões, José João de Almeida, Montserrat Comesaña, and Manuel Perea. On the Advantages of Word Frequency and Contextual Diversity Measures Extracted from Subtitles: The Case of Portuguese. *Quarterly Journal of Experimental Psychology*, 68:680 – 696, 2015.

Ana Paula Soares, Ana Elizabeth Santos Costa, João Junior da Silva Machado, Montserrat Comesaña, and Helena Oliveira. The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. *Behavior Research Methods*, 49:1065–1081, 2017.

Mildred C. Templin. Certain language skills in children. 1957.

Nicole Tracy-Ventura and Florence Myles. The importance of task variability in the design of learner corpora for SLA research. 2015.

David Vadas and James R. Curran. Parsing Noun Phrases in the Penn Treebank. *Computational Linguistics*, 37(4):753–809, 2011.

Sowmya Vajjala. Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features. *International Journal of Artificial Intelligence in Education*, 28:79–105, 2017.

Sowmya Vajjala and Kaidi Lõo. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden, November 2014. LiU Electronic Press.

Sowmya Vajjala and Taraka Rama. Experiments with universal cefr classification. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2018.

N. A. Vyatkina. The Development of Second Language Writing Complexity in Groups and Individuals: A Longitudinal Learner Corpus Study. *The Modern Language Journal*, 96:576–598, 2012.

L.S. Vygotski and A. Kozulin. *Thought and Language*. MIT paperback series. MIT Press, 1986. ISBN 9780262720106.

Zarah Weiß and Detmar Meurers. Broad Linguistic Modeling is Beneficial for German L2 Proficiency Assessment. *Widening the Scope of Learner Corpus Research: Selected Papers from the 4th Learner Corpus Research Conference*, pages 419–435, 2019.

K. Wolfe-Quintero, S. Inagaki, H.Y. Kim, and University of Hawaii at Manoa. Second Language Teaching & Curriculum Center. *Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity*. National Foreign Language Center Technical Reports. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, 1998. ISBN 9780824820695.

Weiwei Yang, Xiaofei Lu, and Sara Weigle. Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28, 06 2015.

Wenxing Yang and Ying Sun. The use of cohesive devices in argumentative writing by Chinese EFL learners at different proficiency levels. *Linguistics and Education*, 23(1):31 – 48, 2012. ISSN 0898-5898.

Robert Östling, Andre Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. Automated essay scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47, Atlanta, Georgia, June 2013. Association for Computational Linguistics.