Tracking the Evolution of Written Language Competence in L2 Spanish Learners

Alessio Miaschi⋆ ◇, Sam Davidson•, Dominique Brunato◇,
Felice Dell'Orletta◇, Kenji Sagae•, Claudia H. Sánchez-Gutiérrez•, Giulia Venturi◇
⋆Department of Computer Science, University of Pisa
◇ItaliaNLP Lab, Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa
•University of California Davis
alessio.miaschi@phd.unipi.it, {ssdavidson,sagae,chsanchez}@ucdavis.edu,
{dominique.brunato,felice.dellorletta,giulia.venturi}@ilc.cnr.it

Abstract

In this paper we present an NLP-based approach for tracking the evolution of written language competence in L2 Spanish learners using a wide range of linguistic features automatically extracted from students' written productions. Beyond reporting classification results for different scenarios, we explore the connection between the most predictive features and the teaching curriculum, finding that our set of linguistic features often reflects the explicit instruction that students receive during each course.

1 Introduction

In the last few years, research on language acquisition has benefited from the use of Natural Language Processing (NLP) technologies applied to large–scale corpora of authentic texts produced by learners, in both the first and second language context. The empirical evidence acquired from learner corpora, complemented with the increased reliability of linguistic features extracted by computational tools and machine learning approaches, has promoted a better understanding of learners' language properties and how they change across time and increasing proficiency level (Crossley, 2020). A first line of research has focused on providing automatic ways of operationalizing sophisticated metrics of language development to alleviate the laborious manual computation of these metrics by experts (Sagae et al., 2005; Lu, 2009). A second line of research has taken the more challenging step of implementing completely data-driven approaches, which use a variety of linguistic features extracted from texts to automatically assign a learner's language production to a given developmental level (Lubetich and Sagae, 2014).

A great amount of work has been carried out in the field of second language acquisition where the study of L2 writings is seen as a proxy of language ability development (Crossley, 2020). In this respect, much related work is devoted to predicting the degree of second language proficiency according to expert–based evaluation (Crossley and McNamara, 2012) or to modelling the evolution of grammatical structures' competence with respect to predefined grades, such as the Common European Framework of Reference for Languages (CEFRL) (Zilio et al., 2018). Given the difficulty of defining a unique indicator of linguistic complexity in the context of L2 language development, a great variety of features from all linguistic levels have been used as input for supervised classification systems trained on authentic learner data for different L2s. Such is the case e.g. of Hancke and Meurers (2013) and Vajjala and Lėo (2014), dealing with L2 German and L2 Estonian, respectively, and of Pilán and Volodina (2018), who also provided a features analysis focused on predictive features extracted from both receptive and productive texts in Swedish L2.

This paper adopts this framework and presents an innovative NLP-based stylometric approach to model writing development in learners of Spanish as a second and Heritage language. Our approach relies on a wide set of linguistically motivated features extracted from students' essays, which have already been shown relevant for a number of tasks related to modelling the 'form' of a text rather than the content. While the majority of previous studies on the evolution of language proficiency in L2 uses cross–sectional data, this study is the first, to our knowledge, using a longitudinal corpus of Spanish L2 essays to model writing development. Interestingly, a similar approach resulted in the successful prediction of the development of writing competence in a L1 acquisition scenario for the Italian language (Richter et al., 2015).

Contributions In this paper: (i) we present, to

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 92–101
July 10, 2020. ©2020 Association for Computational Linguistics

the best of our knowledge, the first data–driven study which uses linguistic features from student data to model the evolution of written language competence in Spanish as a Second Language (SSL); (ii) we show that it is possible to automatically predict the relative order of two essays written by the same student at different course levels using a wide spectrum of linguistic features; (iii) we investigate the importance of linguistic features in predicting language growth at different course levels and whether they reflect the explicit instruction that students receive during each course.

2 Motivation and Approach

Studies of L2 writing have focused on linguistic complexity as an indicator of writing development (Lu, 2011; Ortega, 2003). This construct, however, is still ill-defined, as evidenced by the divergent measures of complexity utilized in different studies. Typical measures of complexity have been the length of the T-unit (Hunt, 1965), the number of subordinate clauses in a text, or type to token ratios, among others. Instead of considering the construct as being multidimensional (Norris and Ortega, 2009; Bulté and Housen, 2012) and, thus, encompassing an array of different features, most studies have selected one or two of these measures and used them as single indicators of complexity (Bulté and Housen, 2014). This has prevented the development of much needed research that associates different steps of linguistic and written development with specific sets of characteristics. This situation has also prevented the formation of an in-depth picture of how those specific aspects develop in relation to the grammatical, lexical or stylistic content taught in classes at different language course levels. This second objective of characterizing writing at different proficiency levels may provide useful insights into how writing samples could be used for placement tests or other assessments to determine which language course is best suited to further develop a student's linguistic skills.

In the concrete case of SSL, the literature indicates that one of the most difficult aspects to master for learners is the language's complex verbal morphology (Blake and Zyzik, 2016; Salaberry, 1999), given that verbal inflections express a complex cluster of person, number, tense, aspect and mood. Therefore, SSL courses tend to propose a step-by-step introduction to these different aspects of verbal morphology, generally following this order: (1) person and number in the present indicative, (2) past tenses (i.e., imperfect vs. preterite vs. pluperfect), and (3) mood (subjunctive vs. indicative). If this typical instructional sequence had to influence students' writing, it would be expected that learners show an increase in the variety of inflections that they are able to use over time. Nonetheless, several studies also indicate that a linguistic feature that has been learned in class may be mastered in exercises that focus on explicit knowledge but take additional time to unfold in tasks that require more implicit knowledge, such as free writing (Ellis and Shintani, 2013). This means that a simple classification of students' proficiency based on the presence or absence of features studied in a particular course may not be accurate, as some students may explicitly know the rules for a specific inflectional distinction but still be unable to use them accurately in writing. Taking lack of use in writing as evidence for lack of explicit knowledge could entail that students be mistakenly invited to enroll in courses where those features that do not show in their writing are unnecessarily explained to them again. A better approach would thus be to know what students are able to do when they are enrolled in different courses and, only then, compare those abilities to see which match, or mismatch, the contents seen in that particular class. By using a large set of linguistic features, it is possible to understand which phenomena change across proficiency levels and whether they are explicitly related to the teaching guidelines.

This study aims at tackling some of the still open methodological issues in the literature on Spanish acquisition by decomposing the problem into two main research questions: *(i)* verify if it is possible to predict the relative order of two essays written by the same student at different course levels using a wide set of linguistic predictors automatically extracted from Spanish L2 written productions; *(ii)* understand which typologies of language phenomena contribute more to the identification of writing skills' evolution and whether such properties reflect the teaching guidelines of the courses.

Following the approach devised in Richter et al. (2015) we addressed the first research question as a classification task: given a pair of essays written by the same student and ordered according to the course level (d_1, d_2), we classify whether $C(d_2) > C(d_1)$, where $C(d_1)$ and $C(d_2)$ correspond respectively to the course levels during

Course Level	Essays	Tokens	Students
Beginner (SPA 1-3)	2,058	485,435	1,130
Intermediate (SPA 21-22)	445	120,102	244
Composition (SPA 23-24)	536	151,197	287
Heritage (SPA 31-33)	459	130,684	244
Total	3,498	887,418	1,905[1]

Table 1: Summary of corpus composition.

Terms Enrolled	Students	Essays	Tokens
2	267	984	290,399
3	111	612	179,306
4	32	242	74,956
5	5	48	13,977

Table 2: Longitudinal data summary.

which the student wrote d_1 and d_2. Specifically, we model the problem as a binary classification task, training a Linear Support Vector Machine (LinearSVM) to predict the relative order of two essays written by the same student using a wide range of linguistic predictors automatically extracted from the POS tagged and dependency parsed essays. We rely on LinearSVM rather than more powerful learning algorithms, such as Neural Language Models, in order to obtain meaningful explanations when the classifier outputs its predictions to anchor the observed patterns of language development to explicit linguistic evidence.

We further extracted and ranked the feature weights assigned by the linear model in order to understand which typology of linguistic features contributes more to the classification task at different course levels. The assumption is that the higher the weight associated with a specific feature, the greater its importance in solving the classification task and, consequently, in modeling the student's written language evolution.

3 Corpus and Features

3.1 The COWS-L2H Corpus

We analyzed development of student writing from the *Corpus of Written Spanish of L2 and Heritage Speakers*, or COWS-L2H (Davidson et al., 2020). This corpus consists of 3,498 short essays written by students enrolled in one of ten lower-division Spanish courses at a single American university. Concretely, these courses are organized as follows: Spanish (SPA) 1, 2, and 3 are the introductory courses, which exposes students to the basic morphosyntax of Spanish; SPA 21 and 22 are the intermediate courses, focused on the development of reading and listening skills with a strong emphasis on lexical development; SPA 23 and 24 are two courses that specifically aim at improving writing skills with an emphasis on academic writing in Spanish; SPA 31, 32, and 33 are the Heritage speakers courses. These courses are grouped into four categories based on student proficiency and experience, as shown in Table 1.

Student compositions in the corpus are written in response to one of four writing prompts, which are changed periodically. During each period (an academic quarter, which consists of ten weeks of instruction) of data collection, students are asked to submit two compositions, approximately one month apart, in response to targeted writing prompts. These composition themes are designed to be relatively broad, to allow for a wide degree of creative liberty and open-ended interpretation by the writer. Prompts are intended to be accessible to writers at all levels of proficiency. Additionally, the use of broad themes invites the use of a variety of verb tenses and vocabulary. The use of specific writing prompts allows us to control for known topic effects on syntactic complexity among L2 learners (Yang et al., 2015).

The essays in the corpus were submitted by 1,370 unique student participants, with 415 student participants having submitted compositions in two or more academic terms (for a maximum of eight writing samples from each student). Thus, the corpus contains both cross-sectional and longitudinal data on the development of student writing in the context of a university language program. The distribution of the essays across the levels is uneven due to the distribution of student enrollment in Spanish courses. Because more students enroll in beginning Spanish courses than in advanced levels, a larger number of essays submitted to the corpus come from these beginner-level courses. The L2 Spanish learners are primarily L1 speakers of English, but due to the diverse student population of the source university, a large number are L1 speakers of other languages such as Mandarin. However, as English is the university's language of instruction, all students are either L1 or fluent L2 speakers of English. Those students enrolled in the Heritage courses (SPA 31 - 33) are, for the most part, L1 speakers of Spanish, having learned Spanish from

[1] This number differs from the 1,370 unique participants, as students who participated in more than one category are represented twice.

a young age in the home, and L2 speakers of English; these Heritage learners have had little-to-no academic instruction in Spanish.

We focused our study on the longitudinal data in the COWS-L2H corpus. We were thus able to model the chronological development of L2 Spanish writing by monitoring how the writing quality of an individual student's compositions increase with time. Student participation is summarized in Table 2.

3.2 Linguistic Features

The set of linguistic features considered as predictors of L2 written competence evolution is based on those described in Brunato et al. (2020). It includes a wide range of text properties, from raw text features, to lexical, morpho-syntactic and syntactic properties, which were extracted from different levels of linguistic annotation. For this purpose, the COWS-L2H Corpus was automatically parsed using UDPipe (Straka et al., 2016) trained on the Spanish Universal Dependency Treebank (GSD section), version 2.5. We rely on these features since it has been shown that they have a high predictive power for several tasks all aimed at modelling the linguistic *form* of documents. This is the case for example of the automatic readability assessment task (Dell'Orletta et al., 2011a), of the automatic classification of the textual genre of documents (Cimino et al., 2017), or also of the automatic identification of the L1 of a writer based on his/her language production in a L2 (Cimino et al., 2018). Interestingly, for all mentioned tasks the set of linguistic features plays a very important role in the classification not only of a whole document but also of each single sentence. This is the reason why, as reported in the following sections, we modelled the prediction of the development of writing skills both as document and sentence classification tasks.

Although we used a state–of–the art pipeline, it is well-acknowledged that the accuracy of statistical parsers decreases when tested against texts of a different typology from that used in training (Gildea, 2001). In this respect, learners' data are particularly challenging for general–purpose text analysis tools since they can exhibit deviation from correct and standard language; for instance, missing or anomalous use of punctuation (especially in 1st grade prompts) already impacts on the coarsest levels of text processing, i.e. sentence splitting, and thus may affect all subsequent levels of annotation. Nevertheless, if we can expect that the predicted value of a given feature might be different from the real one (especially for features extracted from more complex levels of annotation such as syntax), we can also assume that the distributions of errors will be almost similar, at least when parsing texts of the same domain. Note also that the reliability of features checked against automatically annotated data was also empirically shown by Dell'Orletta et al. (2011b), who compared morpho-syntactic and syntactic features extracted from a gold (i.e. manually annotated) and an automatically annotated corpus of the same domain (i.e. biomedical language), showing that results are highly comparable.

As shown in Table 3, the considered features capture linguistic phenomena ranging from the average length of document, sentences and words, to morpho-syntactic information such as parts of speech (POS) distribution and fine–grained features about the inflectional properties of verbs. More complex phenomena are derived from syntactic annotation and model global and local properties of parsed tree structure, with a focus on subtrees of verbal heads, the order of subjects and objects with respect to the verb, the distribution of Universal Dependencies (UD) syntactic relations and features referring to the use of subordination.

Since it is acknowledged that lexical proficiency plays an important role in predicting L2 writing development (Crossley and McNamara, 2012), we also decided to add a small subset of features that model this property in terms of word frequency. Specifically, we considered the average class frequency of all word forms and lemmas in the essays (*Words Frequency Class*), where the class frequency for each word form/lemma was computed exploiting the Spanish Wikipedia (dump of March 2020) using the following measures: $C_{cw} = \lfloor \log_2 \frac{freq(MFW)}{freq(CW)} \rfloor$, $C_{cl} = \lfloor \log_2 \frac{freq(MFL)}{freq(CL)} \rfloor$, where *MFW* and *MFL* are the most frequent word form/lemma in the corpus and *CW* and *CL* are the considered ones.

A first overview of how and to what extent all these features vary across the documents of the COWS-L2H Corpus is provided in Table 4. Essays written by students in the first course levels are longer in terms of number of sentences but they contain shorter sentences compared with those written in the more advanced courses. As concerns the distribution of POS, essays written in the

Level of Annotation	Linguistic Feature	Label
Raw Text	Sentence Length	tokens_per_sent
	Word Length	char_per_tok
	Document Length	n_sentences
	Type/Token Ratio for words and lemmas	ttr_form, ttr_lemma
POS tagging	Distribution of UD and language–specific POS	upos_*, xpos_*
	Lexical density	lexical_density
	Inflectional morphology of lexical verbs and auxiliaries	verbs_*, aux_*
Dependency Parsing	Depth of the whole syntactic tree	parse_depth
	Average length of dependency links and of the longest link	links_len, max_links_len
	Average length of prepositional chains and distribution by depth	prepositional_chain_len, prep_dist_*
	Clause length (n. tokens/verbal heads)	token_per_clause
	Order of subject and object	subj_pre, subj_post, obj_pre, obj_post
	Verb arity and distribution of verbs by arity	verb_edges, verb_edges_*
	Distribution of verbal heads per sentence	verbal_head_sent
	Distribution of verbal roots	verbal_root_perc
	Distribution of dependency relations	dep_dist_*
	Distribution of subordinate and principal clauses	principal_proposition_dist, subord_dist
	Average length of subordination chains and distribution by depth	subord_chain_len, subord_*
	Relative order of subordinate clauses	subord_post, subord_prep

Table 3: Linguistic features according to different levels of annotation.

first years show a lower percentage of e.g. adpositions (*upos_ADP*) and subordinate conjunctions (*upos_SCONJ*) typically contained in longer and well-articulated sentences, while the use of main content words (e.g. *upos_NOUN*, *upos_VERB*) is almost comparable across years. The variation affecting morphosyntactic categories is reflected by the lexical density value, i.e. the ratio between content words over the total number of words, which is slightly higher in beginner essays. If we focus on differences concerning verbal morphology, a linguistic property particularly relevant in the development of Spanish curriculum, we can see how the use of more complex verb forms increases across course levels. Essays of the introductory courses contain a lower percentage of verbs in the past (*verbs_tense_Past*) and imperfect tenses (*verbs_tense_Imp*) (out of the total number of verb tenses) as well as a lower percentage of auxiliary verbs (*aux_**) typically used in more complex verb forms, such as copulative verbs or periphrastic moods and tenses. Interestingly, features related to verb inflectional morphology have the highest standard deviation, suggesting a quite wide variability among learners. A similar trend towards the acquisition of more complex verb structures can also be inferred by considering features extracted from the syntactic level of annotation: essays of the intermediate courses contain for example sentences with a higher average number of dependents of verbs (*verb_edges*) and in particular of verbs with a complex argument structures of 4 dependents (*verb_edges_4*).

As long as Spanish learners start mastering the second language, linguistic properties related to the construction of more complex sentences increase. This is for example the case of the depth of sentence tree (*parse_depth*) and of the length of syntactic relations (*max_links_len*) as well as of features concerning the use of subordination.

4 Experiments

We train a LinearSVM that takes as input pairs of essays written by the same students according to all the possible pairs of course levels (e.g. SPA 1 - SPA 2, SPA 2 - SPA 3, etc.). Specifically, we extract for each pair the linguistic features corresponding to the first and second essays and the difference between them. We standardize the input features by scaling each component in the range $[0, 1]$. To test the actual efficiency of the model, we perform the experiments with a 5-cross validation using different students during the training and testing phases. In order to provide our system with negative samples, we expand our datasets by adding reversed samples.

Since the students were asked to write essays responding to different prompts, we devise two set of experiments, pairing all the essays written by the same students that have: (i) the same prompt; (ii) both same and different prompts. Also, because of the small number of training samples for certain pairs of course levels we also decide to perform the experiments on a sentence-level, extracting the linguistic features for each sentence in the longitudinal subset of the COWS-L2H corpus and pairing them on the basis of the previously defined criteria. In order to obtain reliable results both on the document and sentence configurations, we consider only datasets at different pairs of course levels that contain at least 50 and 20 samples (including negative pairs) respectively. All the classification

Features	SPA 1	SPA 2	SPA 3	SPA 21	SPA 22	SPA 23	SPA 24	SPA 31	SPA 32	SPA 33
Raw Text Properties										
char_per_tok	4.3 ±.27	4.4 ±.27	4.42 ±.26	4.42 ±.26	4.43 ±.25	4.46 ±.23	4.41 ±.22	4.42 ±.25	4.42 ±.28	4.38 ±.3
n_sentences	20.0 ±7.0	24.01 ±7.15	23.57 ±6.87	20.8 ±5.99	20.17 ±5.15	19.54 ±6.33	17.92 ±5.44	16.06 ±4.05	16.31 ±3.78	15.46 ±3.63
tokens_per_sent	10.7 ±3.43	13.16 ±3.52	13.74 ±3.7	15.71 ±3.95	16.43 ±3.59	17.11 ±3.49	19.01 ±4.27	19.95 ±4.16	20.07 ±3.48	20.94 ±4.04
Morphosyntactic information										
lexical_density	.51 ±.05	.5 ±.04	.5 ±.04	.49 ±.03	.48 ±.04	.48 ±.03	.47 ±.03	.48 ±.04	.47 ±.04	.47 ±.04
upos_ADJ	.07 ±.03	.06 ±.02	.06 ±.02	.06 ±.02	.05 ±.02	.05 ±.02	.05 ±.02	.05 ±.02	.05 ±.02	.05 ±.02
upos_ADP	.09 ±.04	.1 ±.03	.11 ±.03	.11 ±.02	.11 ±.02	.12 ±.02	.12 ±.02	.13 ±.03	.12 ±.02	.13 ±.02
upos_NOUN	.16 ±.04	.16 ±.03	.16 ±.03	.16 ±.03	.16 ±.03	.17 ±.02	.17 ±.03	.17 ±.02	.16 ±.03	.16 ±.03
upos_PRON	.07 ±.04	.07 ±.03	.07 ±.03	.07 ±.03	.07 ±.03	.07 ±.03	.07 ±.03	.07 ±.03	.08 ±.04	.08 ±.04
upos_PUNCT	.14 ±.03	.13 ±.03	.12 ±.03	.12 ±.03	.11 ±.03	.11 ±.03	.11 ±.03	.09 ±.02	.09 ±.02	.09 ±.02
upos_SCONJ	.01 ±.01	.02 ±.01	.03 ±.01	.03 ±.02	.04 ±.02	.04 ±.01	.04 ±.02	.04 ±.02	.05 ±.02	.05 ±.02
upos_VERB	.12 ±.04	.12 ±.03	.12 ±.03	.12 ±.02	.12 ±.02	.12 ±.02	.12 ±.02	.13 ±.02	.13 ±.02	.13 ±.03
Inflectional morphology										
aux_mood_Cnd	.02 ±.09	.03 ±.09	.04 ±.12	.03 ±.07	.06 ±.11	.05 ±.11	.04 ±.08	.05 ±.09	.06 ±.12	.04 ±.11
aux_mood_Ind	.97 ±.14	.96 ±.12	.92 ±.15	.94 ±.14	.91 ±.13	.92 ±.14	.94 ±.1	.91 ±.16	.91 ±.12	.93 ±.12
aux_mood_Sub	.01 ±.04	.01 ±.04	.03 ±.07	.02 ±.05	.03 ±.05	.02 ±.08	.03 ±.06	.03 ±.06	.03 ±.06	.03 ±.05
aux_tense_Imp	.05 ±.16	.16 ±.25	.21 ±.26	.21 ±.25	.24 ±.25	.24 ±.26	.22 ±.24	.23 ±.28	.2 ±.27	.24 ±.29
aux_tense_Past	.02 ±.09	.1 ±.15	.09 ±.15	.12 ±.16	.12 ±.14	.11 ±.15	.12 ±.16	.11 ±.16	.12 ±.17	.11 ±.13
aux_tense_Pres	.92 ±.21	.73 ±.32	.69 ±.33	.65 ±.32	.63 ±.3	.65 ±.32	.66 ±.32	.63 ±.34	.66 ±.34	.63 ±.33
verbs_tense_Imp	.02 ±.06	.08 ±.12	.11 ±.13	.13 ±.13	.16 ±.14	.14 ±.15	.13 ±.13	.17 ±.15	.15 ±.15	.14 ±.14
verbs_tense_Past	.11 ±.19	.28 ±.23	.28 ±.22	.3 ±.2	.35 ±.22	.3 ±.22	.31 ±.19	.31 ±.21	.28 ±.18	.33 ±.19
Verbal Predicate Structure										
verb_edges	2.3 ±.36	2.5 ±.32	2.52 ±.3	2.62 ±.35	2.67 ±.28	2.63 ±.28	2.7 ±.32	2.71 ±.29	2.68 ±.26	2.76 ±.27
verb_edges_4	.09 ±.08	.13 ±.07	.13 ±.07	.16 ±.07	.16 ±.07	.15 ±.08	.16 ±.07	.16 ±.06	.16 ±.06	.16 ±.07
verbal_head_sent	1.52 ±.46	1.8 ±.53	1.92 ±.52	2.13 ±.54	2.26 ±.54	2.3 ±.51	2.54 ±.61	2.73 ±.58	2.86 ±.65	2.95 ±.66
Global and Local Parsed Tree Structures										
parse_depth	2.88 ±.65	3.27 ±.62	3.37 ±.61	3.6 ±.63	3.78 ±.55	3.94 ±.64	4.21 ±.69	4.49 ±.65	4.59 ±.67	4.56 ±.62
max_links_len	.65 ±.44	.7 ±.45	.72 ±.42	.96 ±.74	.92 ±.43	.99 ±.42	1.2 ±.68	1.24 ±.53	1.21 ±.42	1.39 ±.72
5rtoken_per_clause	7.17 ±1.56	7.49 ±1.58	7.28 ±1.39	7.52 ±1.51	7.41 ±1.26	7.55 ±1.26	7.62 ±1.24	7.42 ±1.3	7.16 ±1.09	7.26 ±1.32
Order of elements										
obj_post	.67 ±.18	.68 ±.15	.67 ±.15	.64 ±.16	.65 ±.15	.69 ±.13	.69 ±.14	.6 ±.17	.64 ±.17	.6 ±.16
obj_pre	.33 ±.18	.32 ±.15	.33 ±.15	.35 ±.15	.35 ±.15	.31 ±.13	.31 ±.14	.39 ±.16	.36 ±.17	.4 ±.16
subj_pre	.8 ±.19	.84 ±.15	.82 ±.15	.84 ±.15	.84 ±.13	.84 ±.13	.83 ±.13	.81 ±.12	.78 ±.13	.79 ±.14
Use of Subordination										
subord_chain_len	1.06 ±.25	1.15 ±.16	1.18 ±.14	1.21 ±.18	1.24 ±.15	1.24 ±.14	1.26 ±.16	1.29 ±.23	1.33 ±.16	1.32 ±.2
subord_2	.08 ±.14	.11 ±.11	.13 ±.1	.15 ±.11	.17 ±.1	.17 ±.11	.18 ±.11	.19 ±.11	.2 ±.1	.2 ±.1
subord_dist	.24 ±.14	.33 ±.13	.38 ±.12	.4 ±.12	.44 ±.12	.47 ±.12	.5 ±.12	.56 ±.12	.58 ±.08	.57 ±.1

Table 4: A subset of linguistic features extracted for each course level. For each feature it is reported the average value and the standard deviation.

experiments are performed using the majority class classifier as baseline and accuracy as the evaluation metric.

4.1 Tracking Writing Skills' Evolution

Table 5 reports the results obtained at both the document and sentence levels, pairing essays that have the same prompt (*Same* columns) and both the same and different prompts (*All* columns). As a general remark, we observe that best results are those obtained with the document-level experiments. This is quite expected, since sentence-level classification is a more complex task that often requires a higher number of features to gain comparable accuracy (Dell'Orletta et al., 2014). If we focus instead on the distinction between *Same* and *All* results, we notice that higher scores are mainly achieved considering pairs of essays that also have different prompts. Again, this result is not surprising because adding pairs of essays with different prompts within each datasets increases the number of training samples, thus leading to better scores. Despite this, the results obtained according to the *Same* and *All* configurations are quite similar and this allows us to confirm that classification accuracy is not significantly harmed if the two essay's prompts are the same, thus showing that our system is actually focusing on written language competence evolution properties rather than prompt-dependent characteristics.

More interestingly, we notice that considering all the possible course level pairs at the same time our system is able to achieve quite good results, especially at document level classification (0.68 and 0.70 of accuracy for *Same* and *All* configurations respectively), thus showing that it is possible to automatically predict the chronological order of two essays written by the same student by using a wide spectrum of linguistic properties.

In general, our best scores are obtained by considering all the experiments that include essays written by students in the Beginner category (SPA 1, 2 and 3). This is particularly evident for the experiments that compare essays written during SPA 1 as one of the two considered course levels, most likely because the evolution from knowing nothing at all of a specific L2 to knowing enough to start writing is actually bigger that the difference between knowing a little and then learning a little more. Additionally, students at this beginning stage of L2 acquisition tend to use markedly fewer words per sentence, and the words they user are shorter; these features are particularly salient for the classifier. Observing instead the results ob-

Course Levels	Documents				Sentences			
	Same		All		Same		All	
	Score	Samples	Score	Samples	Score	Samples	Score	Samples
All Levels	0.68	2,208	0.7	5,536	0.59	1,047,156	0.61	2,570,366
SPA 1 - SPA 2	0.88	280	0.9	624	0.7	143,660	0.71	316,264
SPA 1 - SPA 3	0.97	178	0.95	440	0.75	85,032	0.75	209,048
SPA 1 - SPA 21	#	#	0.91	116	0.61	14,298	0.7	46,738
SPA 2 - SPA 3	0.62	528	0.62	1,192	0.56	323,332	0.56	724,400
SPA 2 - SPA 21	0.61	62	0.61	188	0.57	35,754	0.58	104,442
SPA 2 - SPA 22	#	#	0.59	68	0.55	8,048	0.63	29,670
SPA 2 - SPA 23	#	#	0.77	52	#	#	0.58	27,420
SPA 3 - SPA 21	0.59	158	0.55	364	0.53	82,104	0.54	190,596
SPA 3 - SPA 22	0.61	64	0.58	186	0.54	31,886	0.6	93,486
SPA 3 - SPA 23	#	#	0.89	106	0.59	13,404	0.59	45,804
SPA 3 - SPA 24	#	#	#	#	#	#	0.68	11,276
SPA 21 - SPA 22	0.59	132	0.62	302	0.52	57,326	0.54	132,454
SPA 21 - SPA 23	0.52	58	0.74	154	0.54	27,038	0.57	67,634
SPA 21 - SPA 24	#	#	0.7	92	0.47	9,268	0.56	35,384
SPA 22 - SPA 23	0.71	76	0.69	186	0.55	35,272	0.56	79,168
SPA 22 - SPA 24	0.69	158	0.73	164	0.5	23,446	0.56	66,184
SPA 23 - SPA 24	0.45	168	0.49	386	0.48	61,654	0.49	137,786
SPA 31 - SPA 32	0.8	100	0.63	212	0.55	27,608	0.55	57,790
SPA 31 - SPA 33	0.52	100	0.53	198	0.51	24,830	0.48	48,990
SPA 32 - SPA 33	0.54	96	0.59	256	0.5	24,154	0.55	66,466

Table 5: Classification results in terms of accuracy obtained both at document and sentence levels along with number of samples for each dataset. **Same** and **All** columns report the results obtained by pairing essays that have same prompt and both same and different prompts respectively. Since the labels within each dataset has been balanced, baseline accuracy is 0.50.

tained pairing student essays belonging to the other three course level categories (Intermediate, Composition and Heritage), we notice a considerable drop in classifier performance. For instance, if we compare essays written by students in the Composition category (SPA 23 - SPA 24) we can see that all the classification results are below the majority class baseline classifier. A possible reason might be that these two courses are specifically aimed at improving learners' writing skills, with an emphasis on academic writing in Spanish, thus involving specific properties, such as discourse-level characteristics, which are possibly not covered by our set of features.

4.2 Understanding Linguistic Predictors

Beyond classification results, we were interested in understanding which typologies of linguistic phenomena are more important for solving the classification task and whether such properties correlate to the teaching curriculum. To better explore this second research question, we perform a feature ranking analysis along with the classification experiments, which allows us to establish a ranking of the most important features according to the different classification scenarios. That is, we evaluate the importance of each linguistic property by extracting and ranking the feature weights assigned by the LinearSVM. Table 6 reports the feature rankings obtained with sentence-level classification results, including pairs of essays that have the same prompt (*Same* configuration). We considered in particular six different course level pairs which are mostly representative of different stages of writing development. The focus on sentence-level results rather than document-level allows capturing more fine-grained linguistic phenomena.

Because the COWS-L2H corpus was collected from a single university with set curriculum, we are able to compare the features utilized by the LinearSVM with the course curriculum. We find that the feature rankings as obtained from the LinearSVM can in many cases be explained by differences in curriculum at each level. For example, from SPA 1 to SPA 2 the most important features used by the model are all related to verbal morphology, particularly morphology of auxiliary verbs. This can be explained by the fact that SPA 1 and 2 are the courses where students are introduced for the first time to the notions of verb tense and person. SPA 1 is focused on managing the idea of person and number in a tense that is not particularly difficult to understand for a speaker of English: the present tense. SPA 2, however, introduces the difficult difference between the three tenses in the past: imperfect, preterite and plus-perfect. This

SPA 1 - SPA 2	SPA 1 - SPA 3	SPA 2 - SPA 3	SPA 3 - SPA 21	SPA 22 - SPA 23	SPA 31 - SPA 32
aux_mood_Ind	lexical_density *	aux_tense_dist_Pres *	lexical_density	upos_PUNCT	upos_ADP *
aux_tense_Pres *	upos_ADP *	aux_mood_Ind	upos_DET	dep_punct	dep_case *
aux_tense_Imp *	upos_VERB *	aux_tense_Imp *	dep_punct	upos_ADV	verbal_head_sent
aux_tense_Past *	upos_NOUN *	aux_tense_Past	upos_VERB	dep_advmod	upos_PUNCT
upos_ADP *	upos_ADJ	dep_punct *	aux_tense_Pres	upos_CCONJ	upos_PRON
verbs_tense_Past *	upos_PRON	upos_PUNCT *	upos_ADJ	dep_cc *	dep_mark
upos_VERB *	dep_det	dep_nsubj *	upos_NOUN	upos_VERB	dep_punct
upos_INTJ *	upos_PUNCT *	dep_iobj	dep_nsubj *	dep_case	aux_tense_Imp
verbal_head_sent *	upos_PROPN	upos_PRON	upos_PRON	aux_form_Part	verbs_tense_Pres
verbs_tense_Imp *	dep_case *	verbal_head_sent *	upos_SCONJ	upos_ADP	subord_dist
upos_ADJ *	upos_SCONJ *	dep_cop	upos_ADV *	dep_mark	dep_cop
ttr_form	upos_AUX	subj_post *	upos_PUNCT	dep_compound	dep_cc
upos_PRON *	dep_punct *	aux_form_Fin	aux_form_Fin	upos_INTJ *	lexical_density
upos_PROPN *	subord_dist *	verbs_tense_Imp *	dep_cc *	dep_nsubj *	upos_AUX
upos_PUNCT *	upos_CCONJ *	upos_AUX	aux_tense_Imp	upos_AUX	upos_ADV

Table 6: Feature rankings obtained with sentence-level (*Same*) classification results for six different course level pairs. Features that vary in a statistically significant way with Wilcoxon Rank-Sum test are marked with *.

fact explains why distribution of past tense main verbs (*verbs_tense_Past*) differs between essays written during SPA 1 and SPA 2. Additionally, SPA 2 introduces composed verb tenses that require an auxiliary. Specifically, the auxiliary verbs “haber”, “estar”, and “ser” are introduced in SPA 2 as part of the past tense forms. Thus, it is not surprising that the top four features used by our classifier for distinguishing between essays written in SPA 1 and SPA 2 are related to the use of auxiliary verbs.

Classification of essays written by students while enrolled in SPA 2 and SPA 3 also relies largely on differences in verbal morphology. While the distribution of present tense auxiliary verbs is the most important distinguishing feature, other compound verb tenses play a role at these levels. For example, differences in the distribution of imperfect auxiliary verbs (*aux_tense_Imp*) may be explained by the use of the pluperfect tense.

Between SPA 1 and SPA 3, the most important discriminating feature is lexical density. While there is no specific focus on lexical density in the course curriculum, this feature is a natural extension of increasing sentence complexity. Davidson et al. (2019) shows that as students progress through the Spanish course sequence, lexical density tends to decrease due to the increased use of function words in more complex sentences. Additionally, one of the final items covered in the SPA 1 curriculum is the use of the prepositions “por” and “para”. Also, at all three beginning levels students are taught to use prepositions in constructing more complex sentence structures. This may explain why preposition usage (*upos_ADP*) is a key discriminating feature between essays written in SPA 1 and SPA2, as well as between SPA 1 and SPA 3. The prominence of this feature indicates that students are learning to more confidently use prepositions as their writing skills develop. The fact that (*upos_ADP*) is not a key discriminating feature between SPA 2 and SPA3 indicates that these changes are occurring primarily at the SPA 2 level, which accords with the course curriculum.

In spite of the still reasonable accuracy in discriminating more advanced levels, making a direct connection between the features used by the SVM and the course curriculum becomes more difficult. At these more advanced levels students have developed an individual writing style which results in a more complex relationship between the curriculum and the syntax used by students. At the SPA 3 - SPA 21 interval, the only three features which vary in a statistically significant way are the distributions of nominal subjects (*dep_nsubj*), adverbs (*upos_ADV*), and coordinating conjunctions (*dep_cc*). While the increased use of adverbs may be seen as a general sign of increased writing complexity, coordinating conjunctions are taught explicitly during SPA 3. Conjunctions are also practiced intensively during both SPA 21 and SPA 22 explaining their importance as a discriminating feature between these levels.

One of the clearest connections between curriculum and the features used by the LinearSVM occurs at the Heritage levels SPA 31 and SPA 32. Heritage learners of Spanish raised in an English-dominant country are known to use “English-like” prepositions in Spanish. For example, Pascual y Cabo and Soler (2015) report on preposition stranding (which is grammatical in English by ungrammatical in Spanish) among Heritage speakers of Spanish in the United States. We find that distributional differences in the use of prepositions, represented by the features *upos_ADP* and *dep_case*, is the key distinguishing feature between essays written by the same student during SPA 31 and SPA 32. This difference indicates that students are learning to use prepositions in a more "Spanish-like" manner, which is one of the major areas of feedback which instructors provide to Heritage students.

5 Conclusion

We present a first study aimed at modeling the evolution of written language competence in Spanish as a Second and Heritage Language, using data from the COWS-L2H Corpus. We have described a rich set of linguistic features automatically extracted from student writing, and have demonstrated that it is possible to automatically predict the relative order of two essays written by the same student at different course levels using these features, especially when considering students enrolled in beginner-level Spanish courses. Finally, we have shown that the linguistic features most important in predicting essay order often reflect the explicit instruction that students receive during each course.

This work can help instructors and language researchers better understand the specific linguistic factors which contribute to improved writing proficiency. Additionally, the appearance of features in the LinearSVM ranking helps clarify the effect of instruction on writing performance, specifically on effects such as the known delay between students being taught a concept and that concept appearing in the students' writing. We also believe that this work may contribute to the development of better language assessment and placement tools.

In future work we intend to explore the influence of student L1 on feature rankings, as L1 (and L2) transfer and interference effects may influence the rate at which students acquire specific linguistic features. Additionally we plan to conduct a cross-lingual analysis, investigating how the feature rankings we see in Spanish writing development relate to those seen in the acquisition of other languages.

Acknowledgments

Reported research started in the framework of a Short Term Mobility program of international exchanges funded by CNR. This work was also funded in part by the UC Davis Grant for Collaborative Interdisciplinary Research and by a gift from Microsoft Azure.

References

Robert J Blake and Eve C Zyzik. 2016. *El espanol y la linguistica aplicada*. Georgetown University Press.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7147–7153, Marseille, France. European Language Resources Association.

Bram Bulté and Alex Housen. 2012. Defining and operationalising L2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, pages 23–46.

Bram Bulté and Alex Housen. 2014. Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of second language writing*, 26:42–65.

Diego Pascual y Cabo and Inmaculada Soler. 2015. Preposition Stranding in Spanish as a Heritage Language. *Heritage Language Journal*, 12:186–209.

A. Cimino, F. Dell'Orletta, D. Brunato, and G. Venturi. 2018. Sentences and Documents in Native Language Identification. In *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Turin.

A. Cimino, M. Wieling, F. Dell'Orletta, S. Montemagni, and G. Venturi. 2017. Identifying Predictive Features for Textual Genre Classification: the Key Role of Syntax. In *Proceedings of 4th Italian Conference on Computational Linguistics (CLiC-it)*, pages 1–6, Rome.

Scott Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*.

Scott A Crossley and Danielle S McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2):115–135.

Sam Davidson, Aaron Yamada, Agustina Carando, Kenji Sagae, and Claudia Sánchez Gutiérrez. 2019. Word use and lexical diversity in second language learners and heritage speakers of spanish: A corpus study. In *American Association for Applied Linguistics*.

Sam Davidson, Aaron Yamada, Paloma Fernandez Mira, Agustina Carando, Claudia H. Sanchez Gutierrez, and Kenji Sagae. 2020. Developing nlp tools with a new corpus of learner spanish.

In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7240–7245, Marseille, France. European Language Resources Association.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2011a. READ-IT: assessing readability of Italian texts with a view to text simplification. In *Proceedings of Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, UK.

Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2011b. Ulisse: an unsupervised algorithm for detecting reliable dependency parses. In *CoNLL*.

Felice Dell'Orletta, Martijn Wieling, Giulia Venturi, Andrea Cimino, and Simonetta Montemagni. 2014. Assessing the readability of sentences: which corpora and features? In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173.

Rod Ellis and Natsuko Shintani. 2013. *Exploring language pedagogy through second language acquisition research*. Routledge.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the Learner Corpus Research (LCR) conference*.

K. W. Hunt. 1965. Grammatical structures written at three grade levels. *(NCTE Research Report Number 3). Urbana, IL: National Council of Teachers of English*.

Xiaofei Lu. 2009. Automatic measurement of syntactic complexity in child language acquisition. In *International Journal of Corpus Linguistics, 14(1):3–28*.

Xiaofei Lu. 2011. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL quarterly*, 45(1):36–62.

Shannon Lubetich and Kenji Sagae. 2014. Data-driven measurement of child language development with simple syntactic templates. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2151–2160.

John Norris and Lourdes Ortega. 2009. Measurement for understanding: An organic approach to investigating complexity, accuracy, and fluency in SLA. *Applied Linguistics*, 30(4):555–578.

Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to l2 proficiency: A research synthesis of college-level l2 writing. *Applied linguistics*, 24(4):492–518.

Ildikó Pilán and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico. Association for Computational Linguistics.

Stefan Richter, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2015. Tracking the Evolution of Written Language Competence: an NLP–based Approach. *CLiC it*, page 236.

Kenji Sagae, Alon Lavie, and Brian MacWhinney. 2005. Automatic measurement of syntactic development in child language. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 197–204. Association for Computational Linguistics.

M Rafael Salaberry. 1999. The development of past tense verbal morphology in classroom L2 spanish. *Applied Linguistics*, 20(2):151–178.

M. Straka, J. Hajic, and J. Strakova. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Sowmya Vajjala and Kaidi Lėo. 2014. Automatic CEFR Level Prediction for Estonian Learner Text. In *NEALT Proceedings Series Vol. 22, pages 113–127*.

Weiwei Yang, Xiaofei Lu, and Sara Cushing Weigle. 2015. Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28:53–67.

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon. 2018. An SLA corpus annotated with pedagogically relevant grammatical structures. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).