IUSS-NeTS at LangLearn: The role of morphosyntactic features in language development assessment

Matilde Barbini[1], Emma Zanoli[1] and Cristiano Chesi[1]

[1]University School for Advanced Studies IUSS Pavia - NeTS Lab

Abstract

This report presents the system developed by the IUSS-NeTS team for the LangLearn evaluation task of the 2023 edition of EVALITA [1], which focuses on tracking the linguistic development of native Italian speakers (L1) and second language learners of Spanish (L2). Our approach focuses on the integration of specific linguistic features and surprisal-based metrics for both Italian and Spanish datasets (CItA and COWS-L2H, respectively). With our tests the system achieves a F-score of 0.67 for the Italian test set and 0.75 for the Spanish test set. These results demonstrate the effectiveness of our methodology in capturing and analyzing the linguistic development of L1 Italian and L2 Spanish learners.

Keywords

language development assessment, linguistic features, surprisal, NLP

1. Introduction

LangLearn [2] is the first shared task on automatic language development assessment. It is defined as a binary classification problem where participants are asked to predict the relative order of two essays written by the same student, for both Italian L1 and Spanish L2. According to the resources employed to train the models, the task was separated into two sub-tasks:

- prediction obtained using only the training data made available for the task;
- prediction acquired utilizing also extra resources (e.g. additional data for the training phase).

Building upon prior research in the field [3, 4, 5, 6], our study aims to investigate the linguistic evolution of native Italian speakers and second language learners of Spanish. To achieve this, we employ a comprehensive approach that combines the extraction of linguistic features from the examined texts, encompassing syntactic, morphosyntactic, and lexical dimensions, with the addition of surprisal-based metrics. This multifaceted methodology allows us to gain deeper insights (especially from a transparent formal linguistic perspective) into the language proficiency development of L1 Italian and L2 Spanish learners. This work is focused on the first sub-task: our approach does not use any additional task-specific resources, although our use of language models in the calculation of surprisal-based metrics must be noted.

2. Data

In this section we describe the training and the test datasets for both Italian L1 and Spanish L2.

2.1. Training set

The CItA corpus (Corpus italiano di apprendenti L1), an archive of written works composed by the same students between the first and second year of lower secondary school, was the first corpus made available by the task's organizers. There are 1352 essays in the original corpus, written by 156 students; 2394 pairs of essays were assigned to task participants for the initial training session (note that some texts were used twice to create the pairs). The structure of the file containing the pairs was as follow: *essay*$_1$, *essay*$_2$, *order*$_1$, *order*$_2$. The essays' chronological placement varied between the first and second year of lower secondary school, contingent upon when they were written; meanwhile, the essay number denoted its sequential position within the given year.

The second dataset provided in the task was the COWS-L2H corpus (Corpus of written Spanish of L2 and Heritage speakers); it comprises 3498 brief

Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)

matilde.barbini@iusspavia.it (M. Barbini); emma.zanoli@iusspavia.it (E. Zanoli); cristiano.chesi@iusspavia.it (C. Chesi)
0000-0002-0877-7063 (M. Barbini); 0000-0001-7116-9338 (E. Zanoli); 0000-0002-9421-8566 (C. Chesi)

CEUR Workshop Proceedings (CEUR-WS.org)

essays produced by second language learners attending a Spanish course at an American university, written in response to four different prompts; 1009 pairs were assigned to task participants for the initial training set. The structure of the file containing the pairs was the same as the Italian training set, however the columns labeled $order_1$ and $order_2$ contained chronological data to be interpreted as academic terms (quarters) and corresponding years. Each academic term spanned a specific time range: W denoted the time period from January to March, S from April to June, SU from July to September, and F from October to December[1].

2.2. Test set

The test set for the CItA corpus consisted of 278 texts organized into 307 pairs, whereas the test set for the COWS-L2H corpus comprised 471 texts organized into 320 pairs. It is important to note that, unlike in the train datasets, the essay identifying acronyms were presented in a random order.

3. System description

In this section we present our approach for the automatic tracking of Italian L1 and Spanish L2 language development. Our method can be summarized in a binary classification based on the combination of linguistic complexity features and surprisal-based metrics.

3.1. Linguistic complexity features

The aim of our approach was to evaluate the relevance of specific morphosyntactic features in the assessment of the linguistic competence improvement both in L1 and L2. With this in mind, we selected a set of robust Natural Language Processing (NLP) tools which allowed us to extract linguistic features relevant to the classification task and which was compatible with both Italian and Spanish. After an extensive evaluation of various tools (Tint[2], Profiling UD[3], READ-IT[4] and CTAP[5]), we decided to rely on CTAP: Common Text Analysis Platform [7]. The CTAP system supports the analysis of language complexity in a easy to use, platform indipendent, flexible and extendable environment [7]. The system was firstly available for German [8] and English [9], then it was adapted also to other languages [10]. In particular as it is stated in the paper which presented the Italian adaptation of the tool [10] CTAP is the most comprehensive linguistic complexity measurement tool for Italian and the only one allowing the comparison of Italian texts to multiple other languages within one tool.

3.2. Surprisal

We chose to increase the chances of success of our classification also by using surprisal-based metrics extracted from the constituent texts of the two LangLearn corpora: it has in fact been observed that the surprisal value of a word is associated with behavioral measures of processing complexity [11]. Surprisal-based metrics have already been used in various contexts, such as the analysis of pathological speech [12] and Italian linguistic studies [13]. When referring to surprisal, we are talking about one of the many possible log-transformed probability measures expressing the likelihood that a token follows a sequence of tokens (or it is placed in a certain position into a given sequence of tokens). We performed an extensive evaluation of the best models to be used in this case, testing 13 models on different linguistic structures (such as relatives, passives, questions, various typologies of errors) for both Spanish and Italian. One important trait we considered was the low tolerance for errors: whenever a morphosyntactic error is present, certain tools tend to normalize the ill-formed token eventually neglecting the relevant features that, in fact, indicate true mistakes. Expanding on previous studies [14], the incorporation of surprisal-based metrics was a deliberate effort to preserve the meaningfulness inherent in error occurrences. We set out to evaluate which model demonstrated the least inclination to normalize errors, examining the surprisal values generated by different models. We gave preference to those models that yielded higher values of surprisal when faced with various types of errors, indicating a recognition of the mistakes, over models that produced lower values leading to normalization of errors. After the evaluation we decided to use two BERT models for both Italian[6] and Spanish[7], available on HugginFace. In particular we calculated the surprisal for every word in every text giving to the two BERT models the entirety of the text as context. Upon detecting errors in the tokenization

[1] For a detailed description of the two corpora see [6] and [5].

[2] https://dh.fbk.eu/research/tint/

[3] http://www.italianlp.it/demo/profiling-ud/

[4] http://www.italianlp.it/demo/read-it/

[5] http://sifnos.sfs.uni-tuebingen.de/ctap/

[6] https://huggingface.co/dbmdz/bert-base-italian-cased

[7] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

Surprisal-based metrics	
text surprisal KURT	kurtosis
text surprisal MAX	maximum value
text surprisal MEAN	mean
text surprisal MEDIAN	median
text surprisal MIN	minimum value
text surprisal OUT F	outliers (frequency)
text surprisal OUT N	outliers (number)
text surprisal Q1	first quartile
text surprisal Q3	third quartile
text surprisal SKEW	skewness
text surprisal STD	standard deviation
text surprisal IQR	interquartile range

Table 1
12 surprisal-based metrics; every metric represents a statistic computed for the token surprisal distribution of every token in every text

performed by both models, we made the decision to introduce an additional preprocessing step which involved utilizing Stanford's parser Stanza[8] for tokenization of both the Italian and Spanish texts, prior to further analysis. Stanza is a Python natural language processing toolkit that supports 66 different human languages. Stanza's text analysis pipeline is language agnostic (in other words, language indipendent): the parser has been trained on 112 datasets, and it has been found that the same neural architecture generalizes efficiently and performs well across all languages [15]. Following the discussed methodology, we were able to calculate on both CItA and COWS-L2H corpora 12 statistical measures (table 1) based on the extraction of word surprisal, which were then added to the linguistic features previously extracted through the CTAP tool.

3.3. Classification

For the classification phase we decided to use Weka [16], since its rich selection of algorithms, ease of use, feature selection, evaluation capabilities and integration options make it a favorable choice for classification tasks[9]. Firstly we constructed ARFF files associating each text of the two corpora with the attributes consisting of the features extracted through CTAP and the 12 surprisal-based metrics. We set our classification in this way: give 1 if it is true that the $essay_1$ (i.e. the first essay of the pair) was written before $essay_2$ (i.e. the second essay of the pair), otherwise give 0. We pre-processed our files and we performed attribute selection on the extracted linguistic features and surprisal-based metrics: we calculated the correlation between each attribute and the output variable and selected only those attributes that had a moderate-to-high positive or negative correlation (close to -1 or 1), discarding the ones with a low correlation (value close to 0). This made it possible to eliminate attributes that were not significant in our analysis. For the Italian language we had initially 266 attributes which became 40 after the attribute selection phase, while for the Spanish language we had 400 attributes which became 75. To make a good use of the two training datasets, we performed the analysis with the cross-validation method (also known as k-fold validation), which is particularly useful to reduce the overfitting problem. To obtain the most comprehensive view we then evaluated several possible classification algorithms offered by Weka. We finally observed that the best performing algorithms were the Logistic regression, J48 and Random Forest, with the last one being the absolute best and thus our choice for the final analysis. Random Forest is a class of ensamble methods which aggregates the prediction of different classifiers to improve accuracy; Weka makes available the implementation of the algorithm described in [18].

4. Results

The final results yielded an F-score of 0.67 for the Italian test set and 0.75 for the Spanish test set. Weka facilitated the extraction of diverse informative measures regarding the classification performance (detailed performance metrics during the 10-fold cross-validation phase can be found in tables 2 and 3).

It is crucial to interpret all this information collectively to gain an accurate feedback on the classification. Among these measures, the ROC Area stands out as a highly informative index. It gauges

[8] https://stanfordnlp.github.io/stanza/

[9] For an extensive discussion of Weka's features see [17].

Detailed accuracy by class								
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,749	0,304	0,716	0,749	0,732	0,446	0,791	0,781	0.0
0,696	0,251	0,731	0,696	0,713	0,446	0,791	0,786	1.0
Weighted average								
0,723	0,278	0,723	0,723	0,723	0,446	0,791	0,784	

Table 2
Detailed metrics of the cross-validation phase on the Italian training dataset.

Detailed accuracy by class								
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0,734	0,263	0,735	0,734	0,734	0,471	0,818	0,825	0.0
0,737	0,266	0,736	0,737	0,736	0,471	0,818	0,824	1.0
Weighted average								
0,735	0,265	0,735	0,735	0,735	0,471	0,818	0,824	

Table 3
Detailed metrics of the cross-validation phase on the Spanish training dataset.

the model's overall performance by calculating the area under the curve in a graph plotting the TP (True Positive) rate (y-axis) against the FP (False Positive) rate (x-axis), as the classification decision threshold changes. An optimal classifier exhibits ROC area values close to 1, while 0.5 is akin to a random guess (similar to a Kappa statistic of 0).

5. Analysis and discussion of the results

The intuition behind our approach was that, in tracing the linguistic evolution, we should observe a change/maturation in the syntactic, morphosyntactic and lexical characteristics of the texts examined. Nevertheless, it was necessary to make further considerations and formulate precise hypotheses: if one could expect an increase in lexical sophistication and variety for both the Italian and Spanish corpus texts, it was however doubtful what phenomena we would have witnessed from the syntactic point of view. Even though we have evidence supporting the theory that, in normal acquisition, functionally higher features (such as those involved in complementation, question or relative clauses formation) are developed at later stages of acquisition [19], we wanted to assess as many linguistic features as possible in order to map a large number of phenomena and characteristics of the students' essays. Let us now examine in more detail which attributes were selected for classification.

5.1. Features selected for the Italian dataset

Out of the 40 attributes chosen for the Italian dataset, 11 were specifically associated with the density of different part-of-speech categories. These included the density of conjunctions, subjunctive mode verbs, pronouns of various types, and auxiliary verbs. Additionally, 2 quantitative features were considered, namely the count of sentences and the average sentence length measured in tokens. 15 attributes pertained to the quantity of various types of syntactic constituents identified within the texts. Furthermore, 12 attributes were focused on assessing the sophistication, refinement, and variety of the lexicon employed. Lastly, only 1 surprisal-based metric was selected, which involved the identification of outliers (essentially the count of anomalous values that significantly deviated from the remaining observations)[10].

5.2. Features selected for the Spanish dataset

Out of the 75 attributes chosen for the Spanish dataset, 6 were specifically associated with the density of various part-of-speech categories. These included the density of verbs, adverbs, adjectives, and others. Additionally, 9 attributes were quantitative features, encompassing measurements such as the mean sentence length in letters, the number of tokens, the standard deviation of sentence length in syllables and letters, and the count of tokens with more than two syllables, among others. A total of 29 features were utilized to track morphological and syntactic complexity. These features encompassed the count of various syntactic constituents, as well as specific indicators such as the number of relative clauses, prepositional phrases per clause, and the

[10] For a list of the selected features for Italian see https://drive.google.com/file/d/1FCIn1DsawUR2YYg3L685q09SZ4NaEtnX/view?usp=sharing

occurrence of passive structures. Moreover, 24 features were dedicated to assessing the sophistication, refinement, and variety of the lexicon employed. 2 additional indices focused on textual cohesion, particularly examining lemma overlaps. Lastly 5 metrics based on surprisal were selected. These metrics included the interquartile range (i.e., the spread of the middle half of the data), the median, the minimum value, the third quartile (i.e., the middle value between the dataset's median and the highest value), and the standard deviation [11].

5.3. Concluding remarks

As a general observation, it is evident that the performance achieved on the Spanish dataset surpasses that of the Italian dataset. This discrepancy can potentially be attributed to multiple factors. One plausible explanation is the larger number of available attributes for the Spanish language compared to Italian. Additionally, it is noteworthy that the acquisition of a second language (L2) exhibits more significant variation and development within a shorter temporal period compared to a first language (L1). The elevated significance of surprisal-based metrics in relation to the texts from the Spanish L2 corpus can also be rationalized by a higher occurrence of errors and inconsistency/incoherence within the texts during the initial stages of L2 language learning. It is our belief that the outcomes of this study should be interpreted as motivation to further explore the linguistic characteristics of texts, with a particular emphasis on the development of additional indices pertaining to linguistic complexity and information theory. This pursuit aims to gain a more comprehensive understanding of the similarities and differences in the progression of L1 and L2 languages.

Acknowledgments

We are thankful to the colleagues of NeTS lab at IUSS Pavia who discussed with us the preliminary results of these experiments. This research has been partially founded by PON Governance 2014-2020: Next Generation UPP Project - CUP D19J22000240006.

[11] For a list of the selected features for Spanish see https://drive.google.com/file/d/16ugOhuL6P8eW1v-C7YaKuVhL8RmifRRv/view?usp=sharing

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[2] C. Alzetta, D. Brunato, F. Delll'Orletta, A. Miaschi, K. Sagae, C. H. Sánchez-Gutiérrez, G. Venturi, Langlearn at evalita 2023: Overview of the language learning development task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[3] S. Vajjala, Automated assessment of non-native learner essays: Investigating the role of linguistic features, International Journal of Artificial Intelligence in Education 28 (2016) 79–105.

[4] Z. Weiss, D. Meurers, Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school, in: Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Florence, Italy, 2019, pp. 380–393. URL: https://aclanthology.org/W19-4440. doi:10.18653/v1/W19-4440.

[5] A. Miaschi, S. Davidson, D. Brunato, F. Dell'Orletta, K. Sagae, C. H. Sanchez-Gutierrez, G. Venturi, Tracking the evolution of written language competence in L2 Spanish learners, in: Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Seattle, WA, USA → Online, 2020, pp. 92–101. URL: https://aclanthology.org/2020.bea-1.9. doi:10.18653/v1/2020.bea-1.9.

[6] A. Miaschi, D. Brunato, F. Dell'Orletta, A nlp-based stylometric approach for tracking the evolution of l1 written language competence, Journal of Writing Research 13 (2021) 71–105. URL: https://www.jowr.org/index.php/jowr/article/view/778. doi:10.17239/jowr-2021.13.01.03.

[7] X. Chen, D. Meurers, CTAP: A web-based

tool supporting automatic complexity analysis, in: Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC), The COLING 2016 Organizing Committee, Osaka, Japan, 2016, pp. 113–119. URL: https://aclanthology.org/W16-4113.

[8] J. Hancke, S. Vajjala, D. Meurers, Readability classification for German using lexical, syntactic, and morphological features, in: Proceedings of COLING 2012, The COLING 2012 Organizing Committee, Mumbai, India, 2012, pp. 1063–1080. URL: https://aclanthology.org/C12-1065.

[9] S. Vajjala, D. Meurers, Assessing the relative reading level of sentence pairs for text simplification, in: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Gothenburg, Sweden, 2014, pp. 288–297. URL: https://aclanthology.org/E14-1031. doi:10.3115/v1/E14-1031.

[10] N. Okinina, J.-C. Frey, Z. Weiss, CTAP for Italian: Integrating components for the analysis of Italian into a multilingual linguistic complexity analysis tool, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7123–7131. URL: https://aclanthology.org/2020.lrec-1.880.

[11] N. J. Smith, R. Levy, The effect of word predictability on reading time is logarithmic, Cognition 128 (2013) 302–319. URL: https://www.sciencedirect.com/science/article/pii/S0010027713000413. doi:https://doi.org/10.1016/j.cognition.2013.02.013.

[12] N. Rezaii, J. Michaelov, S. Josephy-Hernandez, B. Ren, D. Hochberg, M. Quimby, B. Dickerson, A computational approach for measuring sentence information via surprisal: theoretical implications in nonfluent primary progressive aphasia, medRxiv (2022). URL: https://www.medrxiv.org/content/early/2022/11/29/2022.11.25.22282630. doi:10.1101/2022.11.25.22282630. arXiv:https://www.medrxiv.org/content/early/2022/11/29/2022.11.25.22282630.full.pdf.

[13] J. A. Michaelov, B. K. Bergen, Do language models make human-like predictions about the coreferents of italian anaphoric zero pronouns?, 2022. arXiv:2208.14554.

[14] G. Kharkwal, S. Muresan, Surprisal as a predictor of essay quality, in: Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics, Baltimore, Maryland, 2014, pp. 54–60. URL: https://aclanthology.org/W14-1807. doi:10.3115/v1/W14-1807.

[15] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, C. D. Manning, Stanza: A python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 101–108. URL: https://aclanthology.org/2020.acl-demos.14. doi:10.18653/v1/2020.acl-demos.14.

[16] E. Frank, M. A. Hall, G. Holmes, R. Kirkby, B. Pfahringer, I. H. Witten, Weka: A machine learning workbench for data mining., Springer, Berlin, 2005, pp. 1305–1314. URL: http://researchcommons.waikato.ac.nz/handle/10289/1497.

[17] D. Merlini, M. Rossini, Text categorization with weka: A survey, Machine Learning with Applications 4 (2021) 100033. URL: https://www.sciencedirect.com/science/article/pii/S2666827021000141. doi:https://doi.org/10.1016/j.mlwa.2021.100033.

[18] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.

[19] N. Friedmann, A. Belletti, L. Rizzi, Growing trees: The acquisition of the left periphery, Glossa: a journal of general linguistics 6 (2021). URL: https://www.glossa-journal.org/article/id/5877/. doi:10.16995/glossa.5877.