Linguistic features and proficiency classification in L2 Spanish and L2 Portuguese

Iria del Río

CLUL - University of Lisbon

Turku - 30th September 2019

Outline

1. Introduction
2. Datasets
3. Automatic Proficiency Classification
   - Theoretical Framework: CEFR
   - Methodological Framework
4. Experiments
   - Features and Algorithms
   - Results
5. Conclusions and Future Work

Outline

1. Introduction

2. Datasets

3. Automatic Proficiency Classification
    - Theoretical Framework: CEFR
    - Methodological Framework

4. Experiments
    - Features and Algorithms
    - Results

5. Conclusions and Future Work

NLP and Second Language Learning

Intersection of Corpus Linguistics and NLP techniques with the field of Second Language Learning

- Development of **language resources**: learner corpora - COPLE2 (around 300.000 tokens) > research in SLA, development of teaching materials, etc
- Development of **tools for language learning and teaching (CALL)**: automatic essay scoring, grammatical error detection and correction, exercise generator, selection of reading materials

Research Questions

This work focuses on **automatic proficiency classification in L2 Portuguese and L2 Spanish**. It tries to answer the following **research questions**:

- Which linguistic features capture better the proficiency of a L2 text in Spanish and Portuguese?
- Are those features similar between these two close languages?
- Is a cross-lingual approach possible for these two languages?
- When comparing L1 and L2 Spanish, which linguistic characteristics allow for predicting the level of linguistic development of a text?

Outline

1. Introduction

2. Datasets

3. Automatic Proficiency Classification
   - Theoretical Framework: CEFR
   - Methodological Framework

4. Experiments
   - Features and Algorithms
   - Results

5. Conclusions and Future Work

Data from L2 Learners

- Availability of data with linguistic annotations benefits different types of research, from theoretical analysis to statistical approaches like Machine Learning
- Learner data is particularly difficult to gather, because of the specific context where this data is produced
- L2 English: big collections of learner data available, like the Cambridge Learner Corpus (16 millions of words), but such type of collections are not common for other languages

NLI-PT Dataset

- **NLI-PT dataset** aims to solve this gap for European Portuguese
- Data from four learner corpora: COPLE2, CAL2, PEAPL2, Recolha de dados de Aprendizagem do Português Língua Estrangeira
- Originally compiled for NLI experiments
- Bigger and improved version: more texts, better annotations and a different and more intuitive organization of the data
- Student original text
- POS annotated: general POS class and fine-grained
- Syntactically annotated: constituencies and dependencies

CEDEL2 Corpus

- L2 Spanish corpus developed at the University of Granada by professor Cristóbal Lozano
- Data freely available
- 802,019 words coming from 2,578 participants; no annotations
- Two subcorpora:
    - L1 English: 512,873 words, 1,609 native speakers of English studying Spanish in different universities and schools all over the world
    - L1 Greek: 58,575 words coming from 173 native speakers of Greek who are learners of Spanish in Greece
    - +Control corpus of Spanish native speakers (230,571 words coming from 796 Spanish natives)

Outline

1. Introduction
2. Datasets
3. Automatic Proficiency Classification
   - Theoretical Framework: CEFR
   - Methodological Framework
4. Experiments
   - Features and Algorithms
   - Results
5. Conclusions and Future Work

Automatic Proficiency Classification and L2

- Proficiency classification is a common task in second language learning
- The development of the learner is usually defined in relation to a specific scale with different levels of linguistic complexity
- Common European Framework of Reference for Languages (CEFR): one of the most common scales used in Europe for measuring L2 proficiency

CEFR Levels

- 3 broad divisions: A, basic user; B, independent user; C, proficient user
- Subdivided into 6 development levels: A1 (beginner), A2 (elementary), B1 (intermediate), B2(upper intermediate), C1 (advanced) and C2 (proficient)
- Each level is related to specific linguistic features and skills
- Scale that shows a progression from a very rudimentary language to a performance close to a native production

Interest of Automatic Proficiency Classification

- Learners of a second language commonly perform placement tests that define their proficiency level
- Evident interest of an automatic system that can perform this task

Relation with SLA research

Several features used in Automatic Proficiency Classification have been identified as relevant in SLA research

- Lu (2012) for L2 English: relevance of features linked to lexical variation (like Type-Token ratio)
- Syntactic complexity
- Error patterns ("learner accuracy")
- Lexical-syntactic patterns – “phraseology”: good predictor for higher levels

Methodology: Models and Features

- Task modeled as classification or regression
- Features: complexity features usually identified in SLA research, BOW, POS n-grams, errors, morphological/syntactic/discursive features

Outline

1 Introduction

2 Datasets

3 Automatic Proficiency Classification
- Theoretical Framework: CEFR
- Methodological Framework

4 Experiments
- Features and Algorithms
- Results

5 Conclusions and Future Work

Types of experiments performed

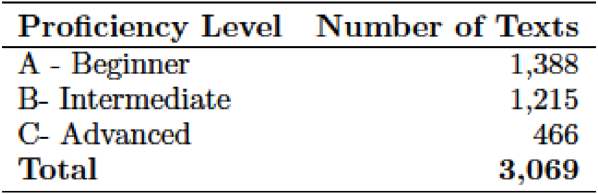
We performed three types of experiments:

- Proficiency classification in L2 Spanish
- Classification of texts considering proficiency levels + native texts in Spanish
- Cross-lingual proficiency classification Spanish>Portuguese and vice versa

Datasets

- We used data from NLI-PT and CEDEL2
- Since CEDEL2 is not annotated, we annotated the corpus at the same levels as NLI-PT: POS (general and fine-grained) and syntactical (dependencies)
- We also extracted descriptive and complexity metrics from CEDEL2

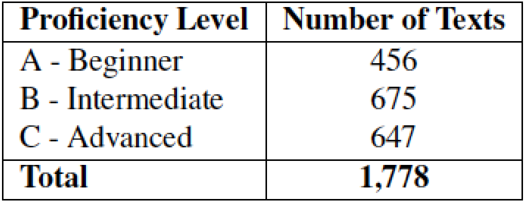
NLI-PT Distribution by Proficiency Level



Proficiency Level	Number of Texts
A - Beginner	1,388
B- Intermediate	1,215
C- Advanced	466
Total	**3,069**

Figure 1: Distribution of texts per class in dataset

CEDEL1 Distribution by Proficiency Level



Proficiency Level	Number of Texts
A - Beginner	456
B - Intermediate	675
C - Advanced	647
Total	**1,778**

Figure 2: Distribution of texts per class in CEDEL2

Classes considered

- In NLI-PT data, the CEFR levels were different in the original corpora: two consider five levels (A1-C1) while the other two consider only the three major levels (A, B, C)

- Therefore we consider only the three major levels in our experiments: A, B, and C

Feature Set

- We were interested in investigating the impact of different linguistic features in the classification task
- Two main types of features:
  - Representation of linguistic levels: lexical (BOW), morphological (POS) and syntactic
  - Complexity metrics: general descriptive and lexical metrics

Features: Representation of Linguistic Levels

Bag of words using the original word form

- In preliminary experiments we tested the impact of different representations: word form, tokenized form and lemmatized form and word form got the best results

POS n-grams

- Fine-grained representation from NLI-PT (it could potentially show agreement errors)
- We experimented with n-grams of different sizes

Dependency triplets n-grams:

- Dependency triplets with the form head, relation, dependent
- They may show different syntactic proficiency

Features: Descriptive and Complexity Metrics

- **Set of 20 features** linked to proficiency by SLA studies

- Those features are not present in CEDEL2; we extracted them using our own scripts

- Different types of metrics:
  - **Morphological features**: number of nouns, number of verbs, number of adverbs, number of connectives, ...
  - **Lexical features**: lexical diversity, content diversity, ...
  - **Descriptive measures**: average syllables per word, syllable count, word count, etc.
  - We also used the Portuguese adaptation of the Flesch reading index

Methods

- We model the task as a classification problem
- We split the datasets into training (80%) and test (20%) sets
- Metrics: general accuracy and F1-Score (general and by class)
- Baseline: text length

Algorithms

- Algorithms: 10-fold cross-validation experiments with the training set + different sets of features for algorithm selection

- We tested the best algorithm for each set of features against the test set

Experiment 1: L2 Spanish

Features	Accuracy	F1-Score
Baseline_RF	0.60	0.58
BOW_LB	0.70	0.70
POS_RF	**0.73**	**0.72**
Dep_LB	0.70	0.70
LING_LR	0.72	0.71
CoLex_LR	0.63	0.61
CoMor_NB	0.49	0.47
CoDesc_LR	0.70	0.70
COMP_LDA	**0.70**	**0.70**
POS+Co_RF	**0.74**	**0.74**
POS+Dep+Co_LR	0.74	0.73
ALL_LR	0.72	0.72

Table 1: General results for L2 Spanish.

Results per Class

Features	**A-F1**	**B-F1**	**C-F1**
Baseline_RF	0.68	0.33	0.69
BOW_LB	0.71	0.59	0.79
POS_RF	**0.76**	**0.60**	**0.82**
Dep_LB	0.72	0.61	0.78
LING_LR	0.77	0.59	0.80
CoLex_LR	0.71	0.43	0.73
CoMor_NB	0.44	0.34	0.61
CoDesc_LR	0.74	0.60	0.77
COMP_LDA	**0.73**	**0.61**	**0.77**
POS+Co_RF	**0.77**	**0.62**	**0.83**
POS+Dep+Co_LR	0.76	0.60	0.80
ALL_LR	0.77	0.62	0.80

Table 2: Results per class for L2 Spanish.

Experiment 2: L1 vs L2 in Spanish

Features	**Accuracy**	**F1-Score**
Baseline_LR	0.50	0.43
BOW_RF	**0.73**	**0.73**
POS_NB	0.39	0.33
Dep_LR	0.37	0.30
LING_LR	**0.75**	**0.74**
CoLex_LR	0.62	0.61
CoMor_ NB	0.40	0.40
CoDesc_LR	0.60	0.59
COMP_LR	**0.65**	**0.64**
POS+Co_RF	0.74	0.74
POS+Dep+Co_LR	0.74	0.74
ALL_RF	**0.75**	**0.74**

Table 3: Classification including native texts.

Results per Class

Features	**A-F1**	**B-F1**	**C-F1**	**N-F1**
Baseline_LR	0.64	0.53	0	0.58
BOW_RF	**0.78**	**0.62**	**0.67**	**0.83**
POS_NB	0	0.25	0.52	0.42
Dep_LR	0.45	0.29	0.48	0.06
LING_LR	**0.75**	**0.63**	**0.70**	**0.88**
CoLex_LR	0.75	0.52	0.49	0.70
CoMor_NB	0.47	0.27	0.40	0.46
CoDesc_LR	0.73	0.42	0.48	0.74
COMP_LR	**0.73**	**0.56**	**0.50**	**0.78**
POS+Co_RF	0.75	0.62	0.67	0.88
POS+Dep+Co_LR	0.73	0.63	0.66	**0.90**
ALL_RF	**0.76**	**0.65**	**0.67**	**0.88**

Table 4: Classification including native texts, per level.

Experiment 3: Cross-lingual Spanish>Portuguese

Features	**Accuracy**	**F1-Score**
Baseline_LR	**0.57**	**0.54**
BOW_CART	0.47	0.40
POS_RF	**0.57**	**0.51**
Dep_LB	0.47	0.36
LING_RF	0.50	0.40
CoLex_NB	0.43	0.42
CoMor_SVM	0.39	0.22
CoDesc_NB	0.49	0.50
COMP_NB	0.44	0.44
POS+Co_RF	**0.57**	**0.52**
POS+Dep+Co_LR	0.55	0.48
ALL_RF	0.54	0.46

Table 5: General cross-lingual results for Spanish to Portuguese.

Results per Class

Features	**A-F1**	**B-F1**	**C-F1**
Baseline_LR	**0.67**	**0.55**	**0.54**
BOW_CART	0.61	0.33	0
POS_RF	**0.68**	**0.52**	**0**
Dep_LB	0.63	0.19	0
LING_RF	0.65	0.26	0
CoLex_NB	0.40	0.49	0.30
CoMor_SVM	0.48	0.48	0.25
CoDesc_NB	**0.60**	**0.49**	**0.25**
COMP_NB	0.48	0.48	0.25
POS+Co_RF	**0.66**	**0.55**	**0**
POS+Dep+Co_LR	0.65	0.30	0
ALL_RF	0.66	0.40	0

Table 6: Results per class for cross-lingual Spanish to Portuguese.

Experiment 3: Cross-lingual Portuguese>Spanish

Features	**Accuracy**	**F1-Score**
Baseline_NB	**0.56**	**0.54**
BOW_LB	0.50	0.49
POS_CART	0.47	0.46
Dep_LB	0.46	0.44
LING_RF	0.39	0.29
CoLex_NB	**0.60**	**0.58**
CoMor_NB	0.39	0.30
CoDesc_NB	0.57	0.55
COMP_NB	0.60	0.57
POS+Co_KNN	0.48	0.45
POS+Dep+Co_KNN	0.48	0.25
ALL_KNN	0.49	0.45

Table 7: General cross-lingual results for Portuguese to Spanish.

Results per Class

Features	A-F1	B-F1	C-F1
Baseline_NB	**0.69**	**0.37**	**0.62**
BOW_LB	0.57	0.40	0.52
POS_CART	**0.59**	**0.45**	**0.37**
Dep_LB	0.51	0.31	0.54
LING_RF	0.61	0.36	0
CoLex_NB	**0.74**	**0.38**	**0.67**
CoMor_NB	0.52	**0.45**	0
CoDesc_NB	0.71	0.38	0.63
COMP_NB	0.74	0.36	0.67
POS+Co_KNN	0.65	0.48	0.28
POS+Dep+Co_KNN	0.65	0.30	0
ALL_KNN	0.66	0.40	0

Table 8: Results per class for cross-lingual Portuguese to Spanish.

Outline

1. Introduction

2. Datasets

3. Automatic Proficiency Classification
   - Theoretical Framework: CEFR
   - Methodological Framework

4. Experiments
   - Features and Algorithms
   - Results

5. Conclusions and Future Work

Conclusions

- We got similar results to the state-of-the art for L2 Spanish (with only three classes)
- Lower results for the cross-lingual approach
- We investigated the relationship between different types of linguistic features and the three main levels of proficiency of the CEFR framework
- We concluded that the linguistic features that work better for the L2 Spanish model are not the same for the cross-lingual models
- POS representation performs better for monolingual L2 Spanish and cross-lingual Spanish to Portuguese
- Complexity features related to lexical and descriptive aspects perform better for cross-lingual Portuguese to Spanish
- Morphological-complexity features show a low performance in all the scenarios
- Comparing L2 and L1 Spanish texts, linguistic features work as better predictors than complexity features

Future Work

- Investigate in depth the causes for the low results in our cross-lingual experiments (homogeneity of CEDEL2 versus the diversity of NLI-PT?)

- Explore new features like metrics of syntactic, lexico-syntactic or discourse complexity

- Use of neural models in the classification task

Obrigada! Gracias! Thanks!

Iria del Río
igayo@letras.ulisboa.pt