Mapping Grammatical Structures onto Proficiency Levels

Rui Talhadas[1,2]

[1] Universidade do Algarve
Campus de Gambelas, P-8005-139 Faro, Portugal
[2] L^{2}F – Spoken Language Lab, INESC-ID Lisboa
Rua Alves Redol 9, P-1000-029 Lisboa, Portugal
rtalhadas@gmail.com

Abstract. In the development of scientifically validated *curricula* that would promote a consistent and appropriate learning process of progressive complexity, it is necessary to determine at what stage of this process are the students of Portuguese as a Foreign Language (PFL) linguistically prepared to learn and use the different language structures. This project intends to map the use of various grammatical and lexical structures, namely: (i) vocabulary; (ii) the use of verbal tenses and modes; (iii) the use of conjunctive adverbs, conjunctions and other discourse connectors; (iv) the clausal internal structures; and (v) the passive construction; in correlation with the learning levels defined in the Common European Framework of Reference for Languages (CEFR), and the evolution in the learning process of Portuguese as a Foreign Language.

Keywords: Portuguese Foreign Language, Grammatical Structures, Proficiency Levels, Learning *Corpora*, Natural Language Processing (NLP)

1 Introduction

The teaching of Portuguese as a Foreign Language (PFL) has had, in recent years, a notorious growth. The use of the Portuguese language evolved from the 8^{th} most spoken language in the Internet in 2008, to 5^{th} in 2013, maintaining its position in 2015[1]. Because of this, a number of countries are beginning to integrate Portuguese in their *curricula*[2,3]. One of the major players in this process is Instituto Camões[4] that coordinates the teaching of the Portuguese

[1] http://www.internetworldstats.com/stats7.htm (last checked: May 20^{th}, 2016; all other URL presented in this paper were checked in this date).
[2] http://www.publico.pt/sociedade/noticia/-galiza-aposta-no-ensino-do-portugues-para-entrar-no-mundo-lusofono-1631589
[3] http://www.observalinguaportuguesa.org/portugues-vai-ser-ensinado-no-ensino-publico-no-luxemburgo/
[4] http://www.instituto-camoes.pt/centros-de-lingua-portuguesa/root/lingua-e-ensino/centros-de-lingua

language abroad. According to the British Council, this trend will continue in the coming years [20], therefore a more intensive research in PFL is urgent. This project intends to map the use of various grammatical and lexical structures, onto the learning levels defined in the Common European Framework of Reference for Languages (CEFR) [4], thus contributing for a better understanding of the evolution in the learning process of Portuguese as a Foreign Language.

This paper is organised as follows: In Section 2, key concepts and related work are presented. Section 3 presents the methodology of this project. Finally, Section 4 presents the expected results.

2 Related Work

The Common European Framework of Reference for Languages [4] was built, among other purposes, to help the planning of language learning programs, mainly defined in a communicative perspective, defining a set of competences that students should acquire in order to attain a certain communicative proficiency level. However, the specific contents to be taught are not always made explicit, leaving that task to be defined for each particular language.

Although the CEFR mentions generically grammatical competences (*e.g.* in level B2, the student should have a "good grammatical control", p. 114), morphological competences (morphological processes the learner knows and uses, p.115) and lexical competences (*e.g.* the ability to use fixed expressions, p. 150), the framework does not establish any link between these competences and the different proficiency levels proposed, which is explained by Hawkins and Filipovic [10, p. 5]) by the purpose of maintaining a certain neutrality between the CEFR and, on one hand, the different languages of Europe, and, on the other hand, competing linguistics theories. As a result, for less studied languages such as Portuguese, the study of the learning processes of foreign languages grammatical structures for PFL, in view of the CEFR, is still incipient [11].

Though related research conducted for other languages is known (for example, [10,17]), their results can not be directly transposed into Portuguese, given the lexical, syntactic and semantic specificities of each language, as well as the possibility that some of those results may be the consequence of different learning contexts, in particular the influence of the mother tongue (L1) in the development of an interlanguage [18].

3 Methodology

This study will follow a quantitative approach, in which the data frequency of the use in *corpora*, by students of PFL, of the following structures and transformations of Portuguese language will be analysed: (i) vocabulary; (ii) the use

of verbal tenses and modes; (iii) the use of conjunctive adverbs[5], conjunctions and other discourse connectors; (iv) the clausal internal structures; (v) the passive construction; and other that may be considered relevant. This structures and constructions are described in greater detail below:

Vocabulary: The moment in the learning process when a learner starts to use a word seems to be related to the frequency with which that word is used in language: the most frequent words are learnt before the rarest. Using language models, the vocabulary used by Portuguese native speakers will be compared to the one used by students of PFL, distributed by the various proficiency levels, in order to determine the correlation between the vocabulary and the student's level on the one hand, in comparison with the frequency of that vocabulary in *corpora* of Portuguese native speakers.

Verbal tenses and modes: The relative frequency of the use of verbal tenses is different. Frequency of use is known to correlate with language proficiency levels. Thus it is expected that a Present Simple, being more frequent, be learnt more quickly than the less frequently occurring Conditional. With the analysis of *corpora*, we expect to understand at which stage of learning the different tenses are used.

Discourse connectors: When beginning to learn PFL, learners use mostly simple sentences, because they are not entirely at ease in the use of FL. According to [10], the use of conjunctions, conjunctive adverbs and other discursive connectors can help to situate the learner at the adequate CEFR level.

Clausal internal structures: The clausal internal structures will also be analysed, based on the sequence of the elementary syntactic constituents (or chunks); *e.g.* noun phrase (`NP`), verb phrase (`VP`), prepositional phrase (`PP`), and their syntactic dependencies. This analysis will allow to assess if, and how, sentence complexity increases along learning levels.

Passive construction: The use of the passive constructions, with both auxiliary '*ser*' (to be) and '*estar*' (to be), and the pronominal passive (with the so-called "passive particle"), demonstrates knowledge of various lexical-grammatical structures. These processes are determined by complex lexical and syntactic phenomena, and have been extensively described in [2]. The frequency of use of these constructions and the lexicon involved will be studied, relating these aspects with the students' proficiency levels.

The strategy to be adopted in this project involves the intensive use of natural language processing (NLP) tools, for the automatic extraction of lexical and syntactic features from learning *corpora*, and the use of machine learning techniques [22] in order to determine how these features are projected on the different levels of the CEFR.

STRING (a hybrid **St**atistical and **R**ule-Based Natural Language Process**ing** Chain)[6] [12] is the NLP chain to be used in this project, and it performs all the

[5] We adopt the term used by Molinier and Levrier in [14]. The term *connective adverb* is also used by Raposo *et al.* in [16, pp. 1810-11] and by Azeredo in [1, pp. 302, 304, 308].

[6] http://www.propor2012.org/demos/DemoSTRING.pdf

basic tasks required for the processing purposes outlined here (namely, text segmentation and lexical analysis [21], morphosyntactic disambiguation [7], chunking, parsing and named-entity recognition [8], anaphora resolution, time expressions processing [9], word sense disambiguation and semantic role labeling [19]). Under the REAP.PT project [15][7], STRING already serves as the basis for developing various teaching/learning applications for PFL, including several lexical and grammar exercises. Currently, the REAP.PT system builds a student model [3] in which the level of proficiency is determined by using language models (unigrams) based on the lexicon, thus allowing, even if in an approximative way, to classify students according to their level of proficiency in PFL, and also to follow their progress. This project will make possible to integrate a higher level of complexity into this student model, representing the structures here studied in the L2/FL learning process, something that, as far as we know, has never been done for Portuguese.

The presence of the grammatical features and the structural transformations here studied will be examined in the existing PFL learning *corpora* available, namely the COPLE2 - Corpus of Portuguese as a Foreign Language/ Second Language [13], the Corpora of PLE (RePLE)[8] and the PEAPL2 - Corpus of Writing Production of Learners of PL2[9], and, eventually, other *corpora* that may become available in the meanwhile.

The compilers of RePLE and PEAPL2 used different names for the CEFR levels. Given these differences, the nomenclature adopted in this work is that of the Council of Europe (A1, A2, B1, B2, C1, C2). Furthermore, COPLE2 and PEAPL2 are divided in 5 proficiency levels, while RePLE is divided in only 3 levels. This difference will be taken into account. Table 1 presents the data relative to the dimensions (number of texts and words) of the three *corpora* (after preprocessing).

Table 1. *Corpora* data

CEFR	**COPLE2**		**RePLE**		**PEAPL2**		**Total**	
levels	**Texts**	**Words**	**Texts**	**Words**	**Texts**	**Words**	**Texts**	**Words**
A1	70	6,236	236	32,717	111	13,963	417	52,916
A2	382	50,302			117	20,914	499	71,216
B1	305	52,731	163	31,332	251	65,719	719	149,782
B2	183	39,539			91	11,337	274	50,876
C1	26	7,290	72	12,727	59	6,552	157	26,569
Total	**966**	**156,188**	**471**	**76,776**	**629**	**118,485**	**2,066**	**351,359**

COPLE2 consists of approximately 1,000 texts written by about 500 learners and speakers of 14 different native languages. RePLE is composed of 476 texts,

[7] https://www.l2f.inesc-id.pt/wiki/index.php/READ.PT_%28Computer_Aided_Language_Learning_-_Reading_Practice%29

[8] http://www.clul.ul.pt/pt/recursos/314-corpora-of-ple

[9] http://www.uc.pt/fluc/rcpl2/

produced by 397 students, speakers of 28 different L1, with a total of approximately 76,800 words. PEAPL2 comprises 629 texts written by PFL students from 50 different nationalities and 39 different L1. Besides these L1, 13 pairs of bilingualism are also represented in this *corpus*.

Though the *corpus* to be used is the union of three *corpora*, it is still relatively small (approx. 351,500 words) when compared to other *corpora* used in similar studies for other languages such as the Cambridge Learner Corpus[10] (45 million words).

To proceed with the analysis of lexical information, structures and transformations, the texts of the *corpora* will be processed by STRING [12] and the CLAVIS feature extraction tool [5,6] will be used, adapting it to cover the structures here referred. These data will then be used to build language models based on different machine learning algorithms. These models are applied as automatic text classifiers, using the set of tools provided by the Weka platform [22]. The classes of these models correspond to the CEFR levels. Among other learning algorithms, we specifically intend to use *decision trees* that allow a clearer view of the factors (in this case, the extracted linguistic characteristics) that produce the achieved results.

4 Expected results

It is expected that, at the end of this project, there will be a mapping of the above mentioned European Portuguese structures and transformations onto the different CEFR proficiency levels.

After this mapping, it will be possible to examine whether the positioning of these grammatical structures and processes within the curricular structure of current PFL syllabuses is adequate to the sequence of the learning of students and whether it reflects the gradual progression in their linguistic proficiency as proposed by the CEFR.

Throughout the project, whenever it is possible and appropriate, the study of other linguistic phenomena, other than those mentioned above, will be considered.

The findings of this study will also pave the way for the development of different computer-aided educational applications, in the PFL domain, for example, to create gaming exercises that allow the learner to practice some processes of the Portuguese language. These games can also help to track the student's proficiency evolution in correlation to the CEFR levels.

References

1. Azeredo, J.: Gramática Houaiss da Língua Portuguesa. Publifolha (2009)

[10] http://www.cambridge.org/elt/corpus

2. Baptista, J.: ViPEr: A Lexicon-Grammar of European Portuguese Verbs. In: Radimsky, J. (ed.) Proceedings of the 31st Intl. Conf. Lexis and Grammar. pp. 10–16. U. Salerno (Italy)/U. South Bohemia, Nové Hrady (Czech Republic) (2012)
3. Correia, R.: Automatic Question Generation for REAP.PT Tutoring System. Master's thesis, IST/ UT Lisboa (2011)
4. Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press (2001)
5. Curto, P., Mamede, N., Baptista, J.: Automatic readability classifier for European Portuguese. In: INFORUM 2014 – Simpósio de Informática. pp. 309–324 (2014)
6. Curto, P., Mamede, N., Baptista, J.: Assisting European Portuguese Teaching: Linguistic features extraction and automatic readability classifier. In: Computer Supported Education, LNCS/CCIS, vol. 583, pp. 81–96. Springer (2015)
7. Diniz, C.: RuDriCo2 - Um Conversor Baseado em Regras de Transformação Declarativas. Master's thesis, IST/ UT Lisboa (2010)
8. Hagège, C., Baptista, J., Mamede, N.: Reconhecimento de entidades mencionadas com o XIP: Uma colaboração entre o INESC-L2F e a Xerox. In: Mota, C., Santos, D. (eds.) Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM, chap. 15, pp. 261–274. Linguateca (2009)
9. Hagège, Caroline; Baptista, J., Mamede, N.J.: Caracterização e Processamento de Expressões Temporais em Português. Linguamática 2(1), 63–76 (2010)
10. Hawkins, J., Filipović, L.: Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework. Cambridge University Press (2012)
11. Leiria, I.: Léxico, Aquisição e Ensino do Português Europeu Língua não Materna. FCG/FCT, Lisboa (2006)
12. Mamede, N., Baptista, J., Diniz, C., Cabarrão, V.: STRING - A Hybrid Statistical and Rule-Based Natural Language Processing Chain for Portuguese. In: Abad, A. (ed.) PROPOR 2012. PROPOR 2012 Demos, http://www.propor2012.org/demos/DemoSTRING.pdf (April 2012)
13. Mendes, A. et al: Corpus de Português Língua Estrangeira/Língua Segunda – COPLE2. In: XXX Encontro Nacional da APL. http://www.clul.ul.pt/files/anagrama/COPLE_APL2014.pdf (2014)
14. Molinier, C., Levrier, F.: Grammaire des adverbes: description des formes en 'ment'. Droz, Genève (2000)
15. Pellegrini, T., Ling, W., Silva, A., Correia, R., Trancoso, I., Baptista, J., Mamede, N.: Overview of Computer-assisted Language Learning for European Portuguese at L2F. In: Proceedings of the 4th Intl. Conf. on Computer Supported Education. pp. 538–543. Porto (2012)
16. Raposo, E., Nascimento, M., Mota, M., Segura, L., Mendes, A.: Gramática do Português. Fundação Calouste Gulbenkian, Lisboa, Portugal (2013)
17. Rau, V., Chang, L., Chien, Y.: From corpus to classroom: Investigating dative alternation of 'give'. In: Tseng, M. (ed.) Investigating Language at the Interface. pp. 27–76. National Sun Yat-sen University (2012)
18. Selinker, L.: Rediscovering Interlanguage. Applied linguistics and language study, Longman (1992)
19. Talhadas, R.: Automatic Semantic Role Labeling for European Portuguese. Master's thesis, U. Algarve, Faro (2014)
20. Tinsley, T., Board, K.: Languages for the future. Tech. rep., British Council (2014)
21. Vicente, A.: LexMan: um Segmentador e Analisador Morfológico com Transdutores. Master's thesis, IST/ U Lisboa (2013)
22. Witten, I., Frank, E., Hall, M.: Data Mining. Morgan Kaufmann (2011), 3^{rd} ed.