P-AWL: Academic Word List for Portuguese

Jorge Baptista, Neuza Costa, Joaquim Guerra, Marcos Zampieri, Maria Cabral, Nuno Mamede

Universidade do Algarve, Campus de Gambelas P 8005-139 Faro, Portugal
Spoken Language Laboratory, INESC ID Lisboa, R. Alves Redol 9, P 1000-029 Lisboa, Portugal
{jbaptis,ncosta,jguerra,mcabral@ualg.pt},
marcos_zampieri@uol.com.br, Nuno.Mamede@l2f.inesc-id.pt

Abstract. This paper presents and discusses the methodology for the construction of an Academic Word List for Portuguese: PAWL, inspired in its English equivalent. The aim of this linguistic resource is to provide a solid base for future studies and applications on Computer Assisted Language Learning, while maintaining comparability with other comparable resources.

Keywords: Keywords: Portuguese, academic word list, Computer Assisted Language Learning

1 Introduction

Computer Assisted Language Learning (CALL) is a growing area of research, in the intersection of Computational Engineering, Linguistics and Language Teaching [8]. One of the important applications developed on this field is the (automatic) assessment of students' language proficiency or of his/her lexical knowledge [16]. The later is usually based on carefully selected subset of the lexicon, which is deemed relevant for a specific purpose; for example, in view of enrolling the students in an adequate level at language courses [13].

The English Academic Word List [5], henceforward E-AWL, is the academic vocabulary most widely used today in language teaching, testing and materials development. E-AWL defines a subset of the English lexicon whose mastery is considered necessary for students at University level. E-AWL has been extensively used in many Natural Language Processing (NLP) studies [9,19], and more recently, in CALL research on reading practice [7,11,15]. To our knowledge, there is still no equivalent resource for Portuguese. Still, there is a significant awareness that university students are required to master a basic 'academese' vocabulary as an indispensable tool for their reading and writing practices [14]. The lack of such standard resource also makes difficult a consensual assessment of students' proficiency/lexical knowledge, especially, but not exclusively, for Portuguese as Second Language courses.

The purpose of this paper is to present and discuss the methodology followed in the construction of Portuguese-Academic Word List (P-AWL). P-AWL is a general purpose vocabulary, with current (but not colloquial) words, which has been designed for immediate application on a CALL web-based environment, currently devoted to improve students' reading practice and vocabulary acquisition [4,13].

2 Issues on selecting vocabulary

There is probably no consensus about how an academic word list should be drawn [12,17]. Corpus-based approaches may contribute to define a vocabulary intersecting different academic genres and subject matters [3,19]. For European Portuguese, there is still no freely available, digital repository of graduation/post-graduation monographs (theses). Therefore, establishing a (balanced) corpus for deriving the vocabulary intersection, across scientific domains and genres, may well be an impossible mission. Even if a particular list could be obtainable by some corpus-based methods, an inevitable selection would still have to be done, since frequency-based vocabulary is very unlikely to produce the balanced and homogeneous word lists adequately targeted for such specific purposes, such as language learning and vocabulary assessment. For example, consider French Dubois-Buise Scale [18]. Because of the (quantitative) methods used to produce this scale, certain relatively homogenous grammatical categories, such as numerals or the names of months, have been attributed not only different ranks in the scale, but they were deemed to belong to different teaching levels. This is direct contradiction with well established evidence (often reflected in teaching practices) that people structure their vocabulary in an associative manner, by grasping semantically or domain-related related words at the same [1, 6].

This paper assumes in full the manually crafting of an academic word list with the inevitable subjective nature of any vocabulary selection. It proposes, however, a set of provisions on procedures and methodology to minimize individual bias in that selection [10]. Our goal is that P-AWL may be used not only as a means of assessment of the students but also a tool to aid in the acquisition of the Portuguese language, hence permitting students to further develop their proficiency in the various levels of the language acquisition process and therefore become increasingly successful as Portuguese speakers.

3 Methods

Taking E-AWL as a starting point, and in order to select the adequate vocabulary for P-AWL, an initial list of 2,145 (2,136 different) entries was drawn adopting the following main criteria: 1. As far as possible, all meaning units of E-AWL were kept in the P-AWL; 2. As far as possible, the Portuguese entry or node-word is cognate of the English entry; 3. Where multiple senses or synonyms were involved, several entries were established for Portuguese; Brazilian Portuguese (BR) and European

Portuguese (LUS) orthographic variants were systematically contrasted (*gênero*/*género* ‘gender, type, kind’) but were not split into different entries; 4. The morphologic family of each entry was systematically drawn; 5. For each word (lemma), a part-of-speech (POS) and inflectional codes (FLEX) were given.

A team of 4 annotators included linguists and language teachers, from different scientific backgrounds, including a native-speaker Brazilian linguist applied the selection criteria to the initial list, through a two-round selection procedure. After this, a 85,6% inter-annotator agreement was achieved. The final list for P-AWL was obtained by selecting all words with less than 2 ‘discard’ marks (i.e. only consensual or with only minimal disagreement vocabulary) and by carefully checking the consistency of that list (in order to avoid word/headword repetitions or incorrect pairings).

Table 1. Breakdown of P-AWL by PoS

PoS	*Count*
Noun	754
Adjective	451
Verb	409
Adverb	203
Conjunction	4
Preposition	2
Total	1812

In its current form, P-AWL contains 1,823 entries. Each entry of P-AWL is classified for its part-of-speech (PoS) category. Table 1 shows the breakdown of P-AWL by part-of-speech category. Words from the same morphological family (derivates) are grouped together under a headword. Currently, there are 814 headwords and each morphological family consists of an average of 2.23 words. Different word senses were distinguished using [2] semantic tags for nouns. The task of semantic annotation of P-AWL entries is still ongoing.

4 Conclusions and future work

This paper presented a new linguistic resource — Portuguese Academic Word List (P-AWL) — specifically built to be used in language tutoring system, but that can also be of interest for the Portuguese NLP community. It was directly inspired in English AWL [5], but its content and detail make it more comprehensive in scope and aim. Future work will include the completion of the semantic annotation of P-AWL nominal entries and calibrating the list by using frequency information from different sources.

5 References

1. Bauer, L.; Nation, P.: Word Families. International Journal of Lexicography 6(4):253-279 (1993)
2. Bick, E.: Noun Sense Tagging: Semantic Prototype Annotation of a Portuguese Treebank, In: Hajic, J. ; Nivre, J. (eds.), Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (December 1-2, 2006, Prague, Czech Republic), pp.127-138. (2006)
3. Chen, Q; Guang-Chun, G.: A Corpus-Based Lexical Study on Frequency and Distribution of Coxhead's AWL Word Families in Medical Research Articles (RAs) (2007) English for Specific Purposes, 26(4): 502-514 (2007)
4. Collins-Thompson, K.; Callan, J.: Information retrieval for language tutoring: An overview of the REAP project. Proceedings of the Twenty Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Sheffield, UK. (2004)
5. Coxhead, A: A New Academic Word List. TESOL, Quarterly, 34(2): 213-238 (2000)
6. Fry, E.: The Vocabulary Teacher's Book of Lists. Jossey Bass (2004)
7. Heilman, M.; Zhao, L.; Pino, J.;Eskenazi, M.: Retrieval of Reading Materials for Vocabulary and Reading Practice. 3rd Workshop on Innovative Use of NLP for Building Educational Applications. Association for Computational Linguistics (2008)
8. Hubbard, P.: Call and future of language teacher education. CALICO Journal (2008)
9. Kulkarni, A; Heilman, M.; Eskenazi, M.; Callan, J.: Word Sense Disambiguation for Vocabulary Learning. Ninth International Conference on Intelligent Tutoring Systems (2008)
10. Laporte, E.: Lexicons and Grammars for Language Processing: Industrial or Handcrafted Products? In: Rezende, L.; Dias da Silva, B.; Barbosa, J. (eds.) Léxico e gramática: dos sentidos à construção da significação. Cultura Académica, São Paulo (Brazil), pp. 51-83 (2009)
11. Liou, H., Chang, J., Kuo, C.; Chen, H.; Chang, C: Web-based Academic English Course Design and Materials Development. International English Teachers' Association Symposium (November 12, 2005, Chieh-Tan Activity Center, Taipei, Taiwan) (2005)
12. Nation, P.: *Learning Vocabulary in another Language.* Cambridge: Cambridge University Press (2001)
13. Luis Marujo, José Lopes, Nuno J. Mamede, Isabel Trancoso, Juan Pino, Maxine Eskenazi, Jorge Baptista, Céu Viana, Porting REAP to European Portuguese, In ISCA International Workshop on Speech and Language Technology in Education (SLaTE 2009), ISCA, Wroxall Abbey Estate, Warwickshire, England, September (2009)
14. Paquot, M.: Towards A Productively-Oriented Academic Word List. PALC2005 *Practical Applications in Language and Computers* (7-9 April 2005, Łódź, Poland) (2005)
15. Pino, J. and Eskenazi, M.: An Application of Latent Semantic Analysis to Word Sense Discrimination for Words with Related and Unrelated Meanings. Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications. NAACL/HLT (2009)
16. Read, J.: Assessing Vocabulary. Cambridge University Press, New York (2000)
17. Schmitt, N.: Vocabulary In Language Teaching. Cambridge University Press, New York (2000)
18. Ters, F., Mayer, G., Reichenbach, D.: L'Echelle Dubois-Buyse. Editions M.D.I (1995)
19. Vongpumivitcha, V.; Huang, J.; Chang, Y.: Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. English for Specific Purposes 25(1) (2008)

Acknowledgments. Research was funded by project REAP.PT ref.CMU-PT/HumMach/0053/2008.