Automatização no diagnóstico de nível de língua: anotação e versatilidade dos recursos

O diagnóstico e a análise automáticos da produção de aprendentes de língua estrangeira são atualmente um tópico de investigação muito relevante na medida em que permitem responder diretamente e de forma mais imediata a necessidades decorrentes das migrações de populações. As técnicas de automatização deste diagnóstico estão em pleno desenvolvimento desde há já alguns anos (Meurers, 2009) e inserem-se, *grosso modo*, em dois grandes grupos – análise de erro e análise de complexidade –, servindo-se de sistemas mais ou menos complexos do ponto de vista computacional e de processamento (Ripley, 2009; Amaral *et al*., 2006; Curto *et al*., 2014; Chen & Meurers, 2019).

No entanto, quer as técnicas baseadas em métodos de aprendizagem automática (supervisionada ou não) que extraem/identificam traços relevantes a partir de dados anotados, quer as técnicas de análise multidimensional de vetores de medida de complexidade linguística pressupõem a análise e a anotação manual de dados, seja para construir os *corpora* de treino e teste necessários aos sistemas de aprendizagem automática, seja para testar e indiretamente informar os sistemas de análise no que respeita aos vetores de complexidade. Apesar disso, os fenómenos e as tipologias de anotação necessárias são distintas:

i) anotação de erro: e.g., Ortografia: nasalidade, acentuação; Morfossintaxe: flexão verbal, concordância nominal (projeto Por Nível, CLUNL; COPLE2, Mendes *et al*., 2016).
ii) anotação de complexidade: e.g., dimensão média da oração; constituintes coordenados por oração; nomes complexos por oração; rácio de orações subordinadas (projeto SyB, EKUT).

Por outro lado, a análise de dados linguísticos de aprendizagem de língua estrangeira, e respetiva compilação e construção de *corpora* de aprendizagem, pela riqueza e complexidade dos fenómenos que abrangem e pela multiplicidade de objetivos que servem, são em si temas de estudo produtivos e, mais importante ainda, dependentes da(s) língua(s) em análise (por exemplo, Alexandre & Pinto, 2014; Alexandre & Gonçalves, 2015; Antunes & Mendes, 2015; Cabrera & Zubizarreta, 2005; Castelo *et al*., 2015; Mendes *et al*., 2016, Talhadas, 2016). É essencialmente a partir desta investigação que os sistemas de anotação são desenhados (Tono, 2003; Nicholls, 2003; Dagneaux *et al*., 2005).

Esta conjugação de fatores demonstra-nos, por um lado, a inevitabilidade da anotação humana dos dados, um processo moroso e dispendioso, e por outro, a importância de garantir a versatilidade dos recursos criados, de modo a maximizar a sua usabilidade e o investimento realizado. A análise dos sistemas de anotação de *corpora* de aprendizagem e das necessidades dos sistemas automáticos é essencial e implica perceber que formato terá uma anotação que permita viabilizar ambas as técnicas, ou seja, como desenhar o sistema de modo a que este permita uma anotação de erro e de estruturas associadas à complexidade e que permita também associar os dados de produção a níveis de proficiência, de modo a permitir o diagnóstico de nível.

A presente comunicação visa, assim, o contraste das necessidades dos sistemas de diagnóstico automático e a análise dos fenómenos refletidos nas atuais anotações para o Português, tendo como base o COPLE2 (Mendes *et al*., 2016) e os resultados da análise conduzida no âmbito do projeto POR Nível (Gramacho *et al*., 2018), propondo um sistema de anotação que contemple a anotação de erro (negativa) e a anotação de estruturas associadas à complexidade (positiva). Para além de sistematizar os fenómenos em causa em ambas as estratégias de diagnóstico, o trabalho pretende também potenciar a usabilidade dos recursos, valorizando-os e fomentando o tão necessário investimento no seu desenvolvimento.

Referências

Alexandre, N. & Gonçalves, A. (2015). Copular constructions in Portuguese as a second language (PL2) by Chinese learners: Do typological differences matter? In: Workshop on Copulas across Languages. June 18-19, University of Greenwich, London, England.

Alexandre, N. & Pinto, J. (2014). Aspects of relative clauses in Portuguese as a foreign language by Chinese learners. In: 20th Conference of the European Association for Chinese Studies. July 22-26, Braga, Coimbra.

Amaral, L., Meurers, D. & Silva, G. (2006). Using Intelligent Computer-Assisted Language Learning (ICALL) Systems to Support Portuguese Instruction. In: The 5th International Conference of the American Portuguese Studies Association (APSA). University of Minnesota. Minneapolis, Minnessota, October 5 - 7, 2006.

Antunes, S., & Mendes, A. (2015). Portuguese Multiword Expressions: data from a learner corpus. In: LCR2015: 3rd Learner Corpus Research Conference. September 11-13, Radboud University, Nijmegen, The Netherlands.

Cabrera, M. & Zubizarreta, M. L. (2005). Overgeneralization of Causatives and Transfer in L2 Spanish and L2 English. In: D. Eddington (ed.). Selected Proc. of the 6th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages. Somerville, MA: Cascadilla Proceedings Project, 15--30.

Castelo, A., Santos, R. & Freitas, M. J. (2015). O uso de vogais ortográficas por aprendentes de Português como língua estrangeira: unidade na diversidade. In: Língua Portuguesa: Unidade na diversidade – Cultura, Literatura, História, Linguística, Tradução e Ensino. November 5-6, Lublin, Poland.

Gramacho, C., Madeira, A., Martins, C., Alexandre, N., Pinto, J . & Correia, S. (2018.). POR Nível: Construção e validação de um teste de colocação para o Português Língua Estrangeira – resultados de um estudo-piloto. Revista da Associação Portuguesa de Linguística (submetido).

Curto, P., Mamede, N., Baptista, J. (2014). Automatic readability classifier for European Portuguese. In: INFORUM 2014 – Simpósio de Informática, 309–324.

Dagneaux, E., Denness, S., Granger, S., Meunier, F., Neff, J. & Thewissen, J. (Eds.) (2005). Error Tagging Manual. Version 1.2. Centre for English Corpus Linguistics, Université Catholique de Louvain.

Mendes, A., Antunes, S. Janssen, M. & Gonçalves, A. (2016). The COPLE2 Corpus: A Learner Corpus for Portuguese. In: Proc. of the 10th Language Resources and Evaluation Conference – LREC'16, 23-28 May 2016, Portoroz, Eslovénia, 3207-3214.

Meurers, D. (ed.) (2009). Automatic Analysis of Learner Language. In: CALICO Journal 26(3). Equinox Publishing Ltd.

Nicholls, D. (2003). The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT. In: D. Archer, P. Rayson, A. Wilson & T. McEnery (eds.). Proc. of the Corpus Linguistics 2003 Conference. Lancaster University, 572-581.

POR Nível - Construção e validação de um teste de colocação em nível para o PLE, projeto de investigação do Centro de Linguística da Universidade Nova de Lisboa (CLUNL), em parceria com o

Centro de Linguística da Universidade de Lisboa. http://fabricadesites.fcsh.unl.pt/por_nivel/

Ripley, M. (2009). JISC case study: Automatic scoring of foreign language textual and spoken. http://community.dur.ac.uk/smart.centre1/jiscdirectory/media/JISC%20Case%20Study%20-%20Languages%20-%20v2.0.pdf

SyB – Complexity, projeto de investigação da Universidade de Tübingen (EKUT), http://sifnos.sfs.uni-tuebingen.de/SyB-0.1/.

Talhadas, R. (2016). Mapping Grammatical Structures onto Proficiency Levels. In Proceedings of 12th International Conference on Computational Processing of the Portuguese Language, http://propor2016.di.fc.ul.pt/wp-content/uploads/2016/07/RuiTalhadasPROPORSRW2016.pdf.

Tono, Y. (2003). Learner corpora: Design, development and applications. In: D. Archer, P. Rayson, A. Wilson e T. McEnery (Eds.), Proceedings of the Corpus Linguistics 2003 Conference. Lancaster University, 800-809.