Portuguese Multiword Expressions: data from a learner corpus[1]

Sandra Antunes
Centro de Linguística da Universidade de Lisboa
sandra.antunes@clul.ul.pt

Amália Mendes
Centro de Linguística da Universidade de Lisboa
amalia.mendes@clul.ul.pt

1 Introduction

The proper usage of Multiword Expressions (MWE), i.e., sequences of words with a syntactic and semantic cohesion (Mel'cuk, 1984; Sinclair, 1991, Cowie, 1998; Sag et al., 2002) is crucial in L2 studies. Indeed, L2 learners frequently struggle to choose the right combination of words and eventually produce errors related to the lexical-grammatical, semantic or stylistic aspects of MWE (Nesselhauf, 2004; Gilquin, 2007; Granger and Paquot, 2010; Paquot, 2013).

Our paper focuses on the use of MWE in a subset of COPLE2, a new learner corpus of Portuguese L2, and addresses the following issues: (i) how significant is the difficulty for the learners to produce MWE; (ii) what are the major errors students make when dealing with constrained expressions.

2 Corpus constitution

Our analysis is based on data from the written register of COPLE2[2], which is composed by: (i) 966 free handwritten essays from different genres (the most frequent being opinion), collected in evaluation tests; (ii) 424 students (18-40 years); (iii) 14 different mother tongues; (iv) all levels of proficiency (the most frequent being elementary). The corpus will be lemmatized and annotated with information on PoS and error type (Nicholls, 2003; Dagneaux et al., 2005).

We restrict our analysis to learners of Portuguese with Spanish, English and Chinese as L1 (Table 1).

L1	Inf.	Age	Texts	Words	Proficiency
Chinese	129	22	323	57.377	Int. (34%)
English	65	24	142	21.610	Elem. (41%)
Spanish	52	29	139	21.200	Elem. (57%)
TOTAL	**246**	**25**	**604**	**100.195**	------

Table1: COPLE2 subcorpus

3 Data analysis

Since all the essays were handwritten, and had to be digitalized and transcribed, the MWE were extracted and annotated during the transcription process. We organized the data according to the typology established by Sag et al. (2002), slightly adapted to our data, and, using a Contrastive Interlanguage Analysis approach (Granger, 1996), we identified different error types:

(i) Collocations (expressions semantically compositional but lexically and/or pragmatically constrained).

- Substitution for (quasi-)synonyms or words belonging to the same semantic field: #*maneiras de transporte* 'ways of transport' vs. *meios de transporte* 'means of transport' (Chinese).
- Substitution for phonologically or morphologically similar words: #*comida populosa* 'populous food' vs. *comida popular* 'popular food' (Chinese).
- Substitution for periphrasis or semantically related words: #*as diferenças e as coisas iguais* 'the differences and the equal things' vs. *as diferenças e as semelhanças* 'the differences and the similarities' (Chinese); #*animais preciosos* 'precious animals' vs. *animais em vias de extinção* 'endangered species' (Chinese).
- L1/L2 transfer at both lexical and syntactic levels: #*parada de metro* 'subway *parada'* vs. *estação de metro* 'subway station' (Spanish); #*especialistas biológicos* 'biological experts' vs. *especialistas em biologia* 'experts in biology' (Chinese). The last example shows that Portuguese non-predicative adjectives pose restrictions regarding the nouns they modify, requiring prepositional phrases.
- Mismatch of the copulative verbs *ser* and *estar* 'to be': #*estamos cruéis* vs. *somos cruéis* 'we are cruel' (English).
- Transposition of semantic relations: #*fechadura nórdica* 'Nordic closenness' in contrast with *abertura nórdica* 'Nordic openness' (English).

(ii) Light verbs constructions (as these verbs carry no significant meaning, the students frequently use them interchangeably): #*dar uma grande influência* 'to give a large influence' vs. *ter uma grande influência* 'to have a large influence' (Chinese).

[1] The corpus compilation is funded by Fundação para a Ciência e Tecnologia (UID/LIN/00214/2013), Fundação Calouste Gulbenkian (Proc. nr. 134655) and ADFLUL.

[2] http://www.clul.ul.pt/en/research-teams/547

(iii) Lexically-syntactically fixed expressions.

- Lexical mismatch: #*dia com dia* 'day with day' vs. *dia a dia* 'day after day' (English).
- L1 transfer: #*música viva* 'live music' vs. *música ao vivo* (English).

(iv) Routine formulae.

- L1 transfer (#*sem outras coisas para reclamar* 'there being no other things to complaint' vs. *sem outro assunto de momento* 'there being no other matter to discuss' (Chinese).

(v) Idiomatic expressions.

- Substitution for semantically related words: #*faca sempre tem dois lados* 'knife always has two sides' vs. *faca de dois gumes* 'double-edged sword' (Chinese).

4 Conclusion

Our data show that collocations are especially difficult for learners of Portuguese L2 because, even though they are semantically compositional, they pose degrees of restrictions that are not easily acquired, generating obvious errors. The few cases of idiomatic expressions in our corpus are also problematic. A possible explanation for their low frequency is that learners have elementary proficiency and are not yet familiarized with them. To target this subtype, other methods, such as translations or elicitation tests, would be required.

L1/L2 transfer plays an important role in the students' productions and is particularly noticeable in expressions with equivalent forms in their L1. We identified cases of transfer of lexical units (either in their native language or adapted to Portuguese), syntactic constructions and register.

We believe that a clear description of the categories of MWE and the identification of factors that correlate with the learners' difficulties may be the key to their lexical accuracy. It is our aim to provide such a typology for Portuguese.

References

Cowie, A. P. 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press.

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

Gilquin, G. 2007. "To err is not all. What corpus and elicitation can reveal about the use of collocations by learners". *Zeitschrift für Anglistik und Amerikanistik*, 55.3. Pp. 273-291.

Granger, S. 1996. "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora". In K. Aijmer, B. Altenberg and M. Johansson (eds.) *Languages in Contrast. Text-based cross-linguistic studies*. Lund Studies in English 88. Lund: Lund University Press. Pp. 37-51.

Granger, S. and Paquot, M. 2010. "Customising a general EAP dictionary to meet learner needs". In *eLexicography in the 21st century: New challenges, new applications*. Proceedings of ELEX2009. Cahiers du CENTAL N°7. Louvain-la-Neuve, Presses universitaires de Louvain.

Mel'cuk, I. 1984. *Dictionnaire explicatif et combinatoire du français contemporain*. Les Presses de L'Université de Montréal. Montréal. Canada.

Nesselhauf, N. 2004. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins Publishing Company.

Nicholls, D. 2003. "The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT". In Archer, D., Rayson, P., Wilson, A. and McEnery T. (eds.) *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster University (UK). University Centre for Computer Corpus Research on Language. 28-31 March. Pp. 572-581.

Paquot, M. 2013. "Lexical bundles and L1 transfer effects". *Language Learning and technologt 14(2)*. Pp. 30-49.

Sag, I., Baldwin T., Bond F., Copestake A. and Flickinger D. 2002. "Multiword Expressions: A Pain in the Neck for NLP", in A. Gelbukh (ed.) *Proceedings of CICLing-2002*, Mexico City, Mexico. Vol. 2276, pp. 1-15.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.