Portuguese Multiword Expressions: data from a learner corpus

Sandra Antunes & Amália Mendes
{sandra.antunes, amalia.mendes}@clul.ul.pt

Center of Linguistics of the University of Lisbon
(UID/LIN/00214/2013)

1. Introduction

The proper usage of **Multiword Expressions (MWE),** i.e. sequences of words with a syntactic and semantic cohesion, plays an important role in FL/L2 studies. The learners frequently struggle to choose the right combination of words and produce errors related to the lexical-grammatical, semantic or stylistic aspects of MWE.

Our poster focuses on the use of MWE in a subset of COPLE2 – Corpus of Portuguese FL/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** subpart of COPLE2:

(i) 966 free handwritten essays collected in evaluation tests, in a total of **156.691 words**;

(ii) 424 students aged between 18-40 years (80% are aged between 18-30 years old; 68% are female);

(iii) different genres: opinion (36%), recount (19%), personal letter (13%), formal letter (10%), informative (10%), dialogue (6%), message (6%);

(iv) all levels of proficiency: beginner (7%), elementary (40%), intermediate (31%), advanced (19%) and proficient (3%);

(v) 14 different mother tongues;

L1	Informants	L1	Informants
Chinese	129	Italian	20
English	65	Dutch	11
Spanish	52	Tetum	9
German	39	Arabic	8
Russian	25	Polish	8
French	23	Korean	6
Japanese	23	Romanian	6

➢The transcriptions are encoded in **XML**, following TEI guidelines.

➢ Each file is composed by a header and the transcription, which includes: (i) all the changes made by the student (deletions, additions, transposition segments, etc); (ii) the correction and comments made by the teacher.

➢ The corpus will be lemmatized and annotated with information on PoS and error type, following a typological scheme.

➢ The interface platform TEITOK enables the **visualization of different versions of the text** (XML; transcription, faithful to the handwritten document; final version of the student; correction of the professor) and the **corpus search**, using the CQP query system.

3. Data Analysis

We restrict our analysis to learners of Portuguese with Spanish (Romance), English (Germanic) and Chinese (Sinitic) as L1.

L1	Inf.	Male	Female	Age	Tests	Texts	Words	Words/Text
Chinese	129	33%	67%	21	277	323	57.385	178
English	65	34%	66%	25	118	142	21.610	152
Spanish	52	42%	58%	28	102	139	21.200	153
TOTAL	**246**	**36%**	**64%**	**25**	**497**	**604**	**100.195**	**161**

➢ The MWE were extracted and annotated during the transcription process of the essays.

➢ We used a Contrastive Interlanguage Analysis approach to compare native and non-native data (L1-L2) and different non-native data (L2-L2).

➢ We organized the data according to different error types, having into account the MWE typology established by Sag et al. (2002).

❑ Substitution

- **For phonologically/morphologically similar words:**
  - ✓ **Collocations**

#comida populosa (Chinese) vs. *comida popular*
'populous food' 'popular food'

- **For (quasi-)synonyms or semantically related words:**
  - ✓ **Collocations**

#maneiras de transporte (Chinese) vs. *meios de transporte*
'ways of transport' 'means of transport'

#animais preciosos (Chinese) vs. *animais em vias de extinção*
'precious animals' 'endangered species'

- **For periphrasis**
  - ✓ **Collocations**

#as diferenças e as coisas iguais (Chinese) *vs.* *as diferenças e as semelhanças*
'the differences and the equal things' 'the differences and the similarities'

  - ✓ **Idiomatic expressions**

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

❑ Transposition of semantic relations:

✓ **Collocations**

#fechadura nórdica (English) vs. *abertura nórdica*
'Nordic closenness' 'Nordic openness'

❑ L1/L2 transfer at both lexical and syntactic levels:

✓ **Collocations**

#parada de metro (Spanish) vs. *estação de metro*
'subway *parada'* 'subway station'

#balança da natureza (Chinese via English) vs. *equilíbrio da natureza*
'nature scale' 'nature balance'

#especialistas biológicos (Chinese) vs. *especialistas em biologia*
'biological experts' 'experts in biology'

✓ **Lexically-syntactically fixed expressions**

#música viva (English) vs. *música ao vivo*
'live music' 'live music'

✓ **Routine formulae**

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

❑ Lexical mismatch:

✓ **Light verbs constructions** (students use them interchangeably)

#dar uma grande influência (Chinese) *vs.* *ter uma grande influência*
'to give a large influence' 'to have a large influence'

✓ **Lexically-syntactically fixed expressions**

#dia com dia (English) vs. *dia a dia*
'day with day' 'day after day'

4. Conclusion

Collocations are especially difficult for L2 learners because, even though they are semantically compositional, they pose degrees of lexical or pragmatic restrictions that are not easily acquired.

Substitution (for synonyms, periphrasis, etc.) is one of the most frequent errors.

L1/L2 transfer also 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 (either in their native language or adapted to Portuguese) at lexical, syntactic and pragmatic levels.

Contactos
Av. Professor Gama Pinto, 2, 1649-003 Lisboa, Portugal
Tel.: +351 21 790 47 00 | Fax: +351 21 796 56 22
http://www.clul.ul.pt

CLUL Centro de Linguística da Universidade de Lisboa
FCT Fundação para a Ciência e a Tecnologia MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
LISBOA UNIVERSIDADE DE LISBOA FLUL LETRAS LISBOA FUNDAÇÃO CALOUSTE GULBENKIAN ICLP Instituto de Cultura e Língua Portuguesa