A study on the production of collocations by European Portuguese learners

Ângela Costa
INESC-ID
CLUNL
Portugal
angela@l2f.inesc-id.pt

Luísa Coheur
INESC-ID
IST - Universidade de Lisboa
Portugal
luisa.coheur@inesc-id.pt

Teresa Lino
FCSH
CLUNL
Portugal
tlino@fcsh.unl.pt

Abstract

In this paper we present a study on the production of collocations by students of European Portuguese as a foreign language. We start by gathering several corpora written by students, and identify the correct and incorrect collocations. We annotate the latter considering several different aspects, such as the error location, description and explanation. Then, taking these elements into consideration, we compare the performance of students considering their levels of proficiency, their mother tongue and, also, other languages they know. Finally, we correct all the students productions and contribute with a corpus of everyday language collocations that can be helpful in Portuguese classes.

1 Introduction

Collocations are stable and mostly non-idiomatic combinations that fall under the category of multiword expressions. They are usually constituted by two or more words, in which one (the base) determines the other (the collocate) (Hausmann, 2004). For instance, in the collocation *strong coffee*, *coffee* is the base and *strong* is the collocate. Collocations can be seen as pre-fabricated blocks (Corpas Pastor, 1996), available as units on the minds of the speakers of a language, and used in oral and written production in the same way single words are. They are highly frequent in languages, and, thus, assume an important role in the teaching/learning process of a foreign language. However, if most non-native speakers of a given language are able to understand the meaning of a collocation, as these are relatively transparent structures, their production can be challenging, as the relation between their elements is, in most of the cases, arbitrary (Cruse, 2000). As an example, and considering the study of English as a foreign language, there is no way to know *a priori*, that a coffee with too much water is a *weak coffee* and not a **faint coffee* (Mackin, 1978).

In their study concerning the production of multiword expressions by European Portuguese learners, Antunes and Mendes (2015) concluded that collocations are the type of multiword expressions that had the largest number of inaccuracies, independently of the mother tongue. According to the authors, "collocations are particularly difficult for learners of Portuguese L2, because they pose degrees of restrictions that are not easily acquired". Considering that there is little information available in Portuguese dictionaries, compared with resources for English (Antunes and Mendes, 2015), lists of everyday language collocations can be a useful tool for these students. By the same token, documenting their errors when producing collocations, like done by Ramos et al. (2010) and Konecny et al. (2015), can help to identify specific difficulties students may have.

In this paper, we study the collocational performance of students of European Portuguese as a foreign language. We start by gathering a corpus with texts written by Spanish, French, English and German students learning European Portuguese (Section 3). Then (Section 4), we identify their production of collocations, and annotate the incorrect ones with information such as the location of the error, its description and a possible explanation. For the latter cases, we follow an adapted version of the taxonomy suggested in (Ramos et al., 2010). We analyse the attained data (Section 5) and identify the main difficulties. Although most of the results are in line with what can be found in the literature, some are, somehow, unexpected. Our last contribution is a corpus of 549 everyday language collocations, which

Proceedings of the 12th Workshop on Multiword Expressions, pages 91–95,
Berlin, Germany, August 7-12, 2016.

resulted from correcting the whole set of collocations provided by the students.

2 Related work

As a linguistic phenomenon, collocations have been the subject of numerous studies (Sinclair, 1991; Tutin, 2004; Hausmann, 2004); also, they have proven to be an extremely fruitful thematic of research in language technology (Smadja, 1993; Seretan, 2011; Wehrli, 2014).

Considering the Portuguese language, we detach the work of Leiria (2006), and Antunes and Mendes (2015). The former concerns lexical acquisition by students learning Portuguese as Foreign Language (L2). The author analysed a corpus of written material produced by French, German, Swedish and Chinese students, where she found "privileged co-occurrences" with a certain degree of fixedness, like *velhos amigos* "old friends" or *gastar dinheiro* "spend money", which matches our definition of collocation. However, each one of these elements was evaluated based mostly on the criteria of whether a native speaker would have used it or not (similarly to the work described in (Konecny et al., 2015)), which is different from the evaluation that we will conduct in this work.

Concerning the work of Antunes and Mendes (2015), it focuses on the multiword expressions found on a subset of a learner corpus of Portuguese[1]. The authors identify different types of multiword expressions (including collocations) produced by foreign students, and characterise the errors found according with a taxonomy they propose. In this work, we opted to follow (and extend) the taxonomy proposed by Ramos et al. (2010), as it was specifically tailored to collocations. In fact, having noticed that no theoretically-motivated collocation error tag set was available, and, in many corpora, collocation errors were simply tagged as "lexical errors", the aforementioned authors created a fine-grained three-dimensional typology of collocation errors. The first dimension captures if the error concerns the collocation as a whole or one of its elements (error location); the second dimension captures the language-oriented error analysis (error description); the third dimension exemplifies the interpretative error analysis (error explanation). Ramos and her team annotated the collocational errors on a learner corpus composed by texts produced by foreign students of Spanish that had English as their mother tongue. In this paper, we annotate erroneous productions of Portuguese collocations by using the lexical level of this taxonomy, to which we felt the need to add some categories.

3 Corpora

We gathered a corpus with students productions of collocations in European Portuguese, by considering four corpora, namely: a) *Corpus de Produções Escritas de Aprendentes de PL2 from Centro de Estudos de Linguística Geral e Aplicada* (CELGA) (Pereira, 2014); b) *Recolha de Dados de Aprendizagem de Português Língua Estrangeira* collected by Centro de Linguística da Universidade de Lisboa (CLUL)[2]; c) two other corpora collected by the authors while teaching at Ciberescola da Língua Portuguesa[3], and at Faculdade de Ciências Sociais e Humanas (FCSH)[4].

CELGA and FCSH corpus were collected in the classroom, and the Ciberescola corpus in online classes. Data from CLUL was collected in Portuguese courses given in 18 universities from different countries (Austria, Bulgaria, South Korea, Spain, USA, etc.). Students that participated in CELGA and CLUL corpus were presented with the same stimuli, divided in three main topics: the individual, the society and the environment. Students from FCSH and Ciberescola had more diversified topics, such as description of their house, their last holidays, their city or their hobbies, among others. From these corpora we selected all texts from students that had Spanish, French, English and German as their native language, and organize them in three levels: Level 1 for A1 and A2 students, Level 2 for B1 and B2 students, and Level 3 for C1 and C2 students.

4 Annotation process

We manually annotated all the correct and incorrect productions of collocations in the collected corpus. We followed Tutin and Grossman (2002) definition of collocation: a "privileged lexical co-occurrence of two (or more) linguistic elements that together establish a syntactic relationship".

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

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

[3] http://www.ciberescola.com/

[4] http://www.fcsh.unl.pt/clcp/

Each incorrect collocation was associated with its correct production and the respective syntactic form, as well as with information concerning the student mother tongue and other foreign languages that the student may know. Then, we annotated the incorrect collocations considering: a) its location (base, collocate, or whole collocation); b) its description and c) its explanation, based on an adapted version of the lexical level of Ramos et al. (2010) taxonomy, as previously mentioned.

In what concerns the description of the error, two new error types were added: preposition and better choice. The first is used when the learner selects the wrong preposition, adds or elides it[5] (*apanhar do avião* for *apanhar o avião* ("take the plane")). Better choice is used when the collocation is not wrong, but there is a better choice (*cozinhar uma receita* for *fazer uma receita* ("make a recipe")). The remaining types are a subset of the ones described in (Ramos et al., 2010): a) Substitution captures the incorrect replacement of a collocate or a base by another existing word (*cabelos vermelhos* for *cabelos ruivos* ("red hair")); b) Creation is used when a student creates a word that does not exist, in this case, in the Portuguese lexicon, which is the case of the word *tiempo* in *passar o tiempo* for *passar tempo* ("spend time"); c) Synthesis is applied when a language unit is used instead of a collocation (*descripção* for *fazer uma descrição* ("to make a description")); d) Analysis covers the case in which the learner creates a new expression with the structure of a collocation instead of using a single word (*tomei o almoço* for *almoçar* ("to have lunch")); e) Different sense is used when the learner uses a correct collocation, but with a different meaning from the intended one (*ter uma escolha* for *fazer uma escolha* ("make a choice")).

Regarding the explanation of the error, we add an extra type to Ramos' taxonomy, in order to cover the situation in which the student mixes European and Brazilian Portuguese (*fazer regime* for *fazer dieta* ("to be on a diet")). The remaining types are the following ones: a) Importation deals with the case in which a collocation is created from an expression in another language known by the student (*fazia a merenda* for *lanchar* ("have a snack")), which shows an importation from Italian ("fare merenda"); b) Extension is used when the learner extends the meaning of an existing word in Portuguese (*faz chuva* for *chover* ("to rain")). A more specific case of this type, that we also use in this work is extension – spelling, which should be used when spelling is influenced by the pronunciation of the misspelled word, as in *loungar um carro* for *alugar um carro* ("rent a car"); c) Erroneous derivation addresses the case when the leaner produces an inexistent form in L2 as a result of a process of erroneous derivation, in many cases by analogy with another form in L2 (*modelos teoréticos* for *modelos teóricos* ("theoretical models')'); d) Overgeneralization handles the scenario in which the learner selects a vaguer or more generic word than required (*fazer sms* for *mandar um sms* ("send a message")); e) Erroneous choice is used when the student selects a wrong word without a clear reason and without intervention of the L1 or another L2 (*memória de pula* for *memória de peixe* ("short memory")).

[5]This type of mistake could have been considered a subtype of Substitution, but in that case additions and elisions would not have been taken into account.

5 Data analysis

Studies like the one presented by Nesselhauf (2005) state that: a) a higher proficiency level in a language is usually characterised by a higher rate in the use of collocations; b) this quantitative gain does not mean a qualitative improvement. Our results, shown in Table 1, do not corroborate the first statement as students from higher levels did not produce collocations in a higher rate. However, the second statement is in line with our results, as only for English students collocational knowledge seems to improve with higher levels of proficiency (that is, considering the total number of produced collocations, the percentage of incorrect collocations decreases with the level).

In our study, 16.53% of the errors concern the base, 74.25% the collocate, and 9.21% the whole collocation (this tendency is observed in all levels and all mother tongues), which is in accordance with Ramos et al. (2010).

Among the deviant collocations, the syntactic form most used by the students was V + N. In fact, that is the most studied sequence in learner corpus research, as students have difficulties selecting the correct verb not only inside a collocation, but also in free sequences of V + N. In Nesselhauf (2005) study with German students of English, one third of the V + N combinations analysed were not acceptable, mainly due to a wrong choice of the verb, which is also in accordance with what we have ob-

L1	l	Txt	Wds	Corr	Incorr
es	1	148	18002	495/83%	98/17%
	2	92	19615	350/84%	66/16%
	3	7	1354	30/83%	6/17%
fr	1	24	2992	76/87%	11/13%
	2	29	8117	135/93%	10/7%
	3	3	896	12/86%	2/14%
en	1	29	4371	49/69%	22/31%
	2	57	14774	236/82%	52/18%
	3	10	2079	26/90%	3/10%
de	1	64	8174	167/83%	34/17%
	2	73	20304	353/84%	65/16%
	3	1	523	10/100%	0/0%

Table 1: Texts, words, correct (Corr) and incorrect (Incorr) collocations and the corresponding percentage, by L1 and level (l).

L1	l	1	2	3
es	1	26/27%	25/26%	15/15%
	2	25/38%	14/21%	2/3%
	3	1/17%	1/17%	0/0%
fr	1	4/36%	3/27%	0/0%
	2	3/30%	3/30%	0/0%
	3	2/1%	0/0%	0/0%
en	1	8/36%	11/50%	0/0%
	2	16/31%	10/19%	2/4%
	3	3/100%	0/0%	0/0%
de	1	19/56%	9/26%	2/6%
	2	25/38%	9/14%	1/2%
	3	0/0%	0/0%	0/0%

Table 2: Substitutions (1), creations (2), analysis (3) by L1 and level (l).

L1	l	4	5	6	7
es	1	1/1%	11/11%	10/10%	10/10%
	2	3/5%	10/15%	3/5%	9/14%
	3	1/17%	1/17%	0/0%	2/33%
fr	1	0/0%	2/18%	2/18%	0/0%
	2	0/0%	3/30%	0/0%	1/10%
	3	0/0%	0/0%	0/0%	0/0%
en	1	0/0%	1/5%	2/9%	0/0%
	2	2/4%	6/12%	7/13%	9/17%
	3	0/0%	0/0%	0/0%	0/0%
de	1	2/6%	0/0%	2/6%	0/0%
	2	3/5%	11/17%	10/15%	6/9%
	3	0/0%	0/0%	0/0%	0/0%

Table 3: Synthesis (4), different sense (5), preposition (7) and better choice (8) by L1 and level (l).

served. Collocations that include adjectives and adverbs seem to be less frequent. A possible explanation is that learners master nouns and verbs before they get to master adjectives and adverbs whose presence augments at higher proficiency levels (Palapanidi and Llach, 2014).

In what concerns description and explanation of the errors, on Table 2 and 3, substitution was the most common error in all the three levels and for all mother tongues (*música forte* for *música alta* ("loud music") or *cabello largo* for *cabelo comprido* ("long hair")). Creation is the second most common error type also for the three levels and four languages. In the following example, *coger um táxi* for *apanhar um táxi* ("take a taxi"), the word *coger* was created, as it does not exist in Portuguese.

In addition, we verify that Level 1 students mostly use importation from L1 or another L2 (Table 4). In Level 2, importation and extension have similar proportions, and represent 40% of the errors. Level 3 errors have their origin mostly in extensions. This may show that lower level students tend to rely more on other languages, while higher level students use more sophisticated mechanisms, like extending the meaning of a known word. An example is the extension of the delexical verb *fazer* in *fazer uma photo* for *tirar uma foto* ("take a picture"). In line with Leiria (2006), who observed that, regarding combinations of words, the majority of the students use their mother tongue when they are lacking the correct expression, we also conclude that students use their mother tongue as their first support, being the Spanish students the ones that do it the most (46.47%), and English students the ones that do it the least (25.97%). Spanish and French students also use Italian and English, and German students rely in Spanish. Other than German, no other students use German as support language. From this we can conclude that the closest the students native language is to Portuguese, more the language will be used as support, and students clearly are aware of this distance.

6 Conclusions and future work

In this paper we presented a study on the production of collocations by foreign students of European Portuguese. This corpus was annotated, analysed and then corrected, resulting in a corpus of

L1	l	fr	es	it	en	de
es	1	0	52	1	1	0
	2	1	27	6	1	0
	3	0	0	0	2	0
fr	1	5	1	1	0	0
	2	1	0	1	0	0
	3	0	0	0	0	0
en	1	2	11	0	4	0
	2	0	7	0	16	0
	3	0	0	0	0	0
de	1	0	4	0	2	2
	2	0	3	0	2	14
	3	0	0	0	0	0

Table 4: Collocations imported by L1 and level (l).

colocations. As future work, we want to enlarge our corpus, especially with Level 3 students, but also with texts produced by students with other native languages, like Italian. We also intend to study the production of collocations by native speakers of Portuguese. Finally, we want to ask a second annotator to use the same error categories so that we are able to calculate an inter-annotator agreement.

Acknowledgments

This work was partially supported by national funds through FCT - Fundação para a Ciência e a Tecnologia, under project UID/CEC/50021/2013 and under project LAW-TRAIN with reference H2020-EU.3.7. – 653587. Ângela Costa is supported by a PhD fellowship from FCT (SFRH/BD/85737/2012).

References

Sandra Antunes and Amália Mendes. 2015. Portuguese multiword expressions: Data from a learner corpus. In *Third Learner Corpus Research Conference*, Radboud University Nijmegen, September.

Gloria Corpas Pastor. 1996. *Manual de fraseología española*. Biblioteca Románica Hispánica, Madrid.

David Alan Cruse. 2000. *Meaning in Language. An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Franz Josef Hausmann. 2004. Was sind eigentlich kollokationen? In Kathrin (Hrsg.) Steyer, editor, *Wortverbindungen ? mehr oder weniger fest.*, pages 309–334. Institut fr Deutsche Sprache, Berlin/New York.

Christine Konecny, Erica Autelli, and Andrea Abel. 2015. Identification, classification and analysis of phrasemes in an l2 learner corpus of italian. In Gloria Corpas Pastor, Miriam Buendía Castro, and Rut Gutiérrez Florido, editors, *Europhras2015: Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, Malaga, Spain, july. EUROPHRAS.

Isabel Leiria. 2006. *Léxico, Aquisição e Ensino do Português Europeu língua não materna*. Fundação Calouste Gulbenkian/ Fundação para a Ciência e a Tecnologia, Lisboa.

Ronald Mackin. 1978. On collocations. words shall be known by the company they keep. In Peter (ed.) Strevens, editor, *In Honour of A. S. Hornby.*, pages 149–165. Oxford University Press, Oxford.

Nadja Nesselhauf. 2005. *Collocations in a Learner Corpus.* Amsterdam and Philadelphia: Benjamins.

Kiriakí Palapanidi and María Pilar Agustín Llach. 2014. Can lexical errors inform about word class acquisition in fl?: evidence from greek learners of spanish as a foreign language. *Revista de Lingüística y Lenguas Aplicadas*, 9(1):67–78.

Isabel Pereira. 2014. Ensino de português língua estrangeira e investigação em pl2 na fluc. In Graça Rio-Torto (ed), editor, *90 Anos de Ensino de Língua e Cultura Portuguesas para Estrangeiros na Faculdade de Letras da Universidade de Coimbra*, pages 39–47. Imprensa da Universidade de Coimbra, Coimbra.

Margarita Alonso Ramos, Leo Wanner, Orsolya Vincze, Gerard Casamayor del Bosque, Nancy Vzquez Veiga, Estela Mosqueira Surez, and Sabela Prieto Gonzlez. 2010. Towards a motivated annotation schema of collocation errors in learner corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Agnès Tutin and Francis Grossmann. 2002. Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif. *Revue française de linguistique appliquée*, 7(1):7–25.