A corpus study of Spanish as a Foreign Language learners' collocation production

Orsolya Vincze, Marcos García Salido,
Margarita Alonso Ramos
Universidade da Coruña

Workshop on Spanish Learner Corpus Research
A Coruña, July 14, 2015

Outline

1. **Introduction**
   - 1.1 What is a collocation?
   - 1.2 Why study collocations?
2. **The study**
   - 2.1 Research questions
   - 2.2. Methodology
3. **Results**
   - 3.1. Quantitative differences between the collocation use of SFL learners and native speakers of Spanish
   - 3.2. Collocation errors in SFL learners' writing

Introduction

1.1 Introduction: Collocations

What is a collocation? phraseological unit W_1W_2

W_1 = base selected according to its meaning

W_2 = collocate whose selection is determined by the base

pouring rain, dense fog, fierce wind

??? dense rain, fierce fog, pouring wind

Introduction: Why study collocations?

Aim:

Describe Spanish as a Foreign Language (SFL) learners' collocation use

Why are collocations important?

- **Native like production and fluency**
- **Often neglected in teaching**

Previous studies:

- **Number of studies on learners' collocation use in languages other than English is scarce**
- **Most studies are limited to a particular type of combination (verb+noun, adjective+noun)**
- **Analyses of collocation errors are also limited to particular collocation types**

The study

Research questions

Comparing learners' and native speakers collocation production:

- Do learners of Spanish produce a similar amount of collocations as native speakers?
- Do learners display a similar lexical diversity to native speakers when producing collocations?
- Is there any difference in the amount of collocations used and lexical diversity between collocations with different syntactic patterns (e.g. verb+noun or noun+adjective combinations)?
- Is there any difference in the amount of collocations used and lexical diversity in the case of collocations expressing different generic meanings (e.g. combinations expressing intensification)?

Research questions

Comparing learners' and native speakers collocation production:

- Which element of the collocation (the base or the collocate) is more commonly erroneous?
- What descriptive types of collocation errors can be identified and which of these is more common?
- To what extent does the native language of learners affect collocation production?

Methodology

Learner corpus

- CEDEL2 corpus
- 100 learner essays = 46420 words
- 102 native essays = 29935 words

Annotation

- Manual annotation of collocations
- Collocation error typology
- LF → syntactic and semantic properties

Results

Comparing learner and native collocation use

1. Do learners of Spanish produce a similar amount of collocations as native speakers?

	LEARNER SUBCORPUS	NATIVE SUBCORPUS
Corpus size (in number of words)	46420	29935
Number of collocation occurrences	1825	1138
Number of collocation lemmas	1127	935
Number of collocations/10000 words	39.31	38.02
Lexical diversity (Lemma/token ratio)	0.618	0.822
Proportion of most frequent 10% collocate lemmas	65.3%	49.7%

Summary of data regarding corpus size, number of collocations identified and lexical diversity of collocations

Comparing learner and native collocation use

2. Do learners display a similar lexical diversity to native speakers when producing collocations?

	LEARNER SUBCORPUS	NATIVE SUBCORPUS
Corpus size (in number of words)	46420	29935
Number of collocation occurrences	1825	1138
Number of collocation lemmas	1127	935
Number of collocations/10000 words	39.31	38.02
Lexical diversity (Lemma/token ratio)	0.618	0.822
Proportion of most frequent 10% collocate lemmas	65.3%	49.7%

Summary of data regarding corpus size, number of collocations identified and lexical diversity of collocations

Comparing learner and native collocation use

2. Do learners display a similar lexical diversity to native speakers when producing collocations?

	Learner subcorpus		Native subcorpus	
	Number of lemmas	**Lemma/token ratio**	**Number of lemmas**	**Lemma/token ratio**
Base	637	0.35	465	0.41
Collocate	433	0.24	567	0.50

Lemma/token ratio in the case of bases and collocates in the learner and native subcorpora

Comparing learner and native collocation use

3. Is there any difference in the amount of collocations used and lexical diversity between collocations with different syntactic patterns?

(1) verb+noun: *ahorrar dinero* 'save money'
(2) noun+verb: *la temperatura se refresca* 'the temperature cools down'
(3) noun+modifier: *razón principal* 'main reason'
(4) noun+de+noun: *paquete de tabaco* 'pack of cigarettes'
(5) verb+adverb: *querer sinceramente* 'love sincerely'
(6) verb+adjective: *poner nervioso* 'make nervous'
(7) verb+adverb combinations: *creer firmemente* 'firmly believe'

Comparing learner and native collocation use

3. Is there any difference in the amount of collocations used and lexical diversity between collocations with different syntactic patterns?

- Verb+noun (*ahorrar dinero* 'save money') and noun+modifier (*razón principal* 'main reason) combinations were the most frequent in both corpora
- Verb+noun combinations are overused by learners
  - Overuse of combinations with *tener* 'have' (*tener derecho* 'have right', *tener problema* 'have a problem', *tener oportunidad* 'have an opportunity')
- Noun+modifier combinations are underused by learners

Comparing learner and native collocation use

4. Is there any difference in the amount of collocations used and lexical diversity in the case of collocations expressing different generic meanings (e.g. combinations expressing intensification)?

- The five most frequent LFs in both the learner and native corpora:
  - $Oper_1$: support verb+noun → *tener un problema* 'to have a problem'
  - Non-standard Adjective: noun+modifier with not easily generalizable meaning → *vida privada* 'private life', *padre biológico* 'biological father'
  - Real1: fulfillment verb + noun → *andar en bicicleta* 'to ride a bike', *ganar un premio* 'to win a prize'
  - Magn: intensifier → *estándares altos* 'high standards'
  - Bon: modifier expressing positive evaluation → *plato delicioso* 'delicious dish'

Comparing learner and native collocation use

4. Is there any difference in the amount of collocations used and lexical diversity in the case of collocations expressing different generic meanings (e.g. combinations expressing intensification)?

	Learner subcorpus		Native subcorpus	
Lexical Function	**Number of occurrences**	**% of all collocation occurrences**	**Number of occurrences**	**% of all collocation occurrences**
Oper1	501	27.45	219	19.24
Real1	219	12	105	9.23
Non-Standard A	128	7.01	124	10.9
Bon	111	6.08	50	4.39
Magn	92	5.04	112	9.84

Collocation errors

Number of correct vs. erroneous collocations

	Number of collocations	%
correct	1390	76.16%
erroneous	435	23.84%

Collocation errors

1. Which element of the collocation (the base or the collocate) is more commonly erroneous?

Element affected by the error	Number of error instances	% of all error instances
base	170	35.34%
collocate	248	51.56%
collocation	63	13.10%

Collocation errors

2. What descriptive types of collocation errors can be identified and which of these is more common?

→ What linguistic categories are affected by the error?

1) **Lexical collocation errors**, e.g.:

Incorrect collocate: **capturar la atención* instead of e.g. *captar la atención* 'catch sb's attention

Synthesis: **misinterpretaciones* instead of e.g. *malas interpretaciones* 'wrong interpretations'

2) **Grammatical collocation errors**, e.g.:

Governed preposition: **montar una bicicleta* instead of *montar en una bicicleta* 'ride a bike'

Number: **dimos bienvenidas* lit. 'we gave welcomes' instad of *dimos la bienvenida* 'we gave a welcome'

Collocation errors

2. What descriptive types of collocation errors can be identified and which of these is more common?

Descriptive error type	Error instances	
	Number	%
Lexical	277	57.47%
Grammatical	203	42.12%
Register	1	0.21%

	Lexical		Gramm.	
	Num.	%	Num.	%
Base	61	35.88%	109	64.12%
Collocate	164	66.13%	84	33.87%
Collocation	52	82.54%	10	15.87%

Collocation errors

3. To what extent does the native language of learners affect collocation production?

Error type	Number of error instances	% of all error instances
Interlingual	241	50.10%
Intralingual	240	49.90%

	Interlingual		Intralingual	
	Number	%	Number	%
Lexical	178	74.17%	99	41.25%
Gramm.	62	25.83%	141	58.75%

Thank you for your attention!

This research has been supported by Ministerio de Economía y Competitividad (FFI2011-30219-C02-01), and the FPU grant (AP2010-4334).