Towards a Motivated Annotation Schema of Collocation Errors in Learner Corpora

M. Alonso Ramos, L. Wanner, O. Vincze, G. Casamayor, N. Vázquez, E. Mosqueira y S. Prieto

Universidade da Coruña
ICREA
Universitat Pompeu Fabra

XXVII LREC
Malta 2010
The Problem

- The relevance of collocations (in the sense of Hausmann, Mel’čuk et al.) in L2 learning is generally acknowledged

  \[ \text{dar un paseo/ faire une promenade} \ ‘[to] \text{take a walk}’ \]
  \[ \text{fumador empedernido / gros fumeur} \ ‘\text{heavy smoker}’ \]

- It is collocations which are difficult to master by the learners! Typical errors:

  \[ \text{hacer un paseo/ donner une promenade} \ ‘[to] \text{take a walk}’ \]
  \[ \text{big smoker/ lourd fumeur} \ ‘\text{heavy smoker}’ \]

- Current learner error annotation schemata tend to group collocation errors into one single subclass of lexical errors

  \[ \text{BUT} \]
• A look at a learner corpus of Spanish (CEDEL2) http://www.uam.es/proyectosinv/woslac/cedel2.htm shows that collocation errors of rather different types can be identified

  *salvar dinero* ‘to save money’ (instead of *ahorrar dinero*)
  *recibir un llamado* ‘to receive a call’ (instead of *recibir una llamada*)
  *asistir la universidad*, lit.‘to attend university’ (instead of *asistir a la universidad*)

  ...

• A more detailed collocation error classification is needed!
1. Towards a typology of collocation errors (based on a Spanish learner corpus)

2. Knowtator: Tool for annotating collocation errors in the corpus

3. The framework of our work: The research project COLOCATE
   3.1. Creation of collocation-oriented content in a web-based learning environment
   3.2. Automatic processing of collocations in a web-based learning environment

4. Preliminary findings

5. Conclusions and future work
1. Three-dimensional Collocation Error Typology:
   (i) location
   (ii) descriptive
   (iii) explanatory
1. Location dimension

- Location of Error
- Element of Collocation
  - [base]
  - [collocate]
- Whole Collocation
  - [collocation]
2. Descriptive dimension

- Substitution
- Creation
- Synthesis
- Analysis
- Different sense
- Determination
- Number
- Gender
- Government
- Governed
- Specification
- Pronoun
- Order
- Register
3. Explanatory Dimension

- Extension
  - Phonetic similarity
  - [I.1-avoidance]
- Importation
- [importation]
- Interalingual Lexical Errors
  - [erroneous derivation]
  - [overgeneralization]
  - [erroneous lexical choice]
- Intralingual Lexical Errors
  - Intralingual Grammatical Errors
  - [intralingual]
- Interlingual Register Errors
  - [interlingual]
- Intralingual Register Errors
  - [intralingual]
Illustration of interlingual lexical errors (affecting the base or the collocate)

| LOCATION | DESCRIPTIVE CRITERION | IMPORTATION | EXTENSION | PHONETIC SIMILARITY | L1 AVOIDANCE |
|----------|-----------------------|-------------|-----------|---------------------|---------------|
| BASE     | Substitution          | hicimos wakeboarding | juego de fútbol (from game) | Doy una marca (instead of poner una nota) | Estado económico (from economic situation) |
|          | Creation              | recibí un llamo (analogy with call) | | | |
| COLLOCATE| Substitution          | gastar todo el año (from spend) | lengua maternal (from maternal language) | cambiar a la verdadera religión (Eng. convert to a religion) | |
|          |                       | tomé examen (from take) | capturar la atención (from capture) | acudir el teléfono (Eng. attend the telephone) | |
|          | Creation              | Doy una marca (from give) | | | |
Illustration of interlingual lexical errors (affecting the base or the collocate)

| Explanatory Criterion | Intralingual Errors |
|-----------------------|---------------------|
| **LOCATION**          |                     |
| Erroneous derivation  |                     |
| Overgeneralization    |                     |
| Err. lexical choice   |                     |
| **BASE**              |                     |
| recibí un llamó (analogy with paseo, canto, salto..) | Me da el sentido (instead of la sensación) |
| **COLLOCATE**         |                     |
| enseñanza secundaria (from segundo) | hacer citas (instead of concertar) |
| | malos efectos (instead of nocivos, dañinos) |
| | toman puestos (instead of ocupar puestos) |
| | escribir el examen (instead of hacer) |
**Illustration of interlingual and intralingual grammatical errors**

(affecting the base or the collocate)

| Location | Descriptive criterion | Interlingual errors | Intralingual errors |
|----------|-----------------------|---------------------|---------------------|
| **Base** | determination         | • tienen el derecho de (from *to have the right to*)  
                              • terminé escuela (from *to finish school*) | • tomé examen (instead of *tomé un examen*) |
|          | number                |                     | • hicimos los esquís del agua (instead of *esquí de agua*)  
                              • tienen prejuicio (instead of *prejuicios*) |
|          | gender                |                     | • días festivas (instead of *días festivos*) |
|          | government            | • tengo interés in  
                              • tengo razones por | |
|          | specification         |                     | • probar comida (an article or an adjective should be used to specify the base, e.g. *probar comida exótica*) |
| **Collocate** | government         | • hablando al teléfono (from Italian, the student’s L2: *parlare al telefono*)  
                              • asisto un juego de fútbol (from *assist a game of...*)  
                              • entró la universidad (from *enter university*) | • montó el autobús (instead of *montar en..*) |
|          | pronoun               | • muero de ganas (instead of *por me muero de...*)  
                              • la película se trata.. (instead of *la película trata*) | |
2. The corpus annotation tool: Knowtator
The annotation schema in *knowtator*
Tagging collocations with **knowtator**
3. The research project: towards a Learning Environment COLOCATE
The Objectives of COLOCATE

A) Develop didactic means which support
   1) interactive learning with collocation error verification and NLP-based error correction
   2) data-driven active learning

B) Develop resources such as
   1) DiCE *Diccionario de colocaciones del español* (DiCE)
      http://www.dicesp.com
   2) personalized collocation dictionaries
   3) collocation-annotated learner corpus

C) Develop NLP-techniques
   1) For automatic recognition and classification of collocations
   2) For automatic error correction and learning material provision
4. Preliminary Findings
5. Conclusions and Future Work
5. Conclusions

1) Collocation errors in learner corpora are far from homogeneous and neither is their distribution!

2) A fine-grained collocation error typology is needed to capture the major error types

3) Targeted exercises and targeted supplementary teaching material (provided by automatic means) are needed to support active language learning

4) COLOCATE is about to address the important issues in L2 learning: (i) adequate didactic tools, (ii) collocation and collocation error resources; (iii) NLP techniques for tracing and classification of collocations and collocation errors
5. Future Work

✓ Continue with the annotation of the learner corpus with collocation errors
✓ Continue with the annotation of the learner corpus with collocations (Lexical Functions)
✓ Extend the DICE
✓ Provide resources for didactic material
✓ Continue to work on ML-based recognition/classification of collocations and collocation errors
✓ Etc., etc., etc.
Thank you very much for your attention!