

# 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  
*dar un paseo/ faire une promenade* '[to] take a **walk**'  
*fumador empedernido / gros fumeur* 'heavy **smoker**'
- It is **collocations** which are difficult to master by the learners! Typical errors:  
*hacer un paseo/ donner une promenade* '[to] take a **walk**'  
*big smoker/ lourd fumeur* 'heavy **smoker**'
- Current learner error annotation schemata tend to group collocation errors into one single subclass of lexical errors

**BUT**

# The Problem

- A look at a learner corpus of Spanish (CEDEL2)

<http://www.uam.es/proyectosinv/woslac/cedel2.htm>

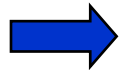
shows that collocation errors of rather different types can be identified

*salvar* dinero ‘to save money’ (instead of *ahorrar* dinero)

recibir un *llamo* ‘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!

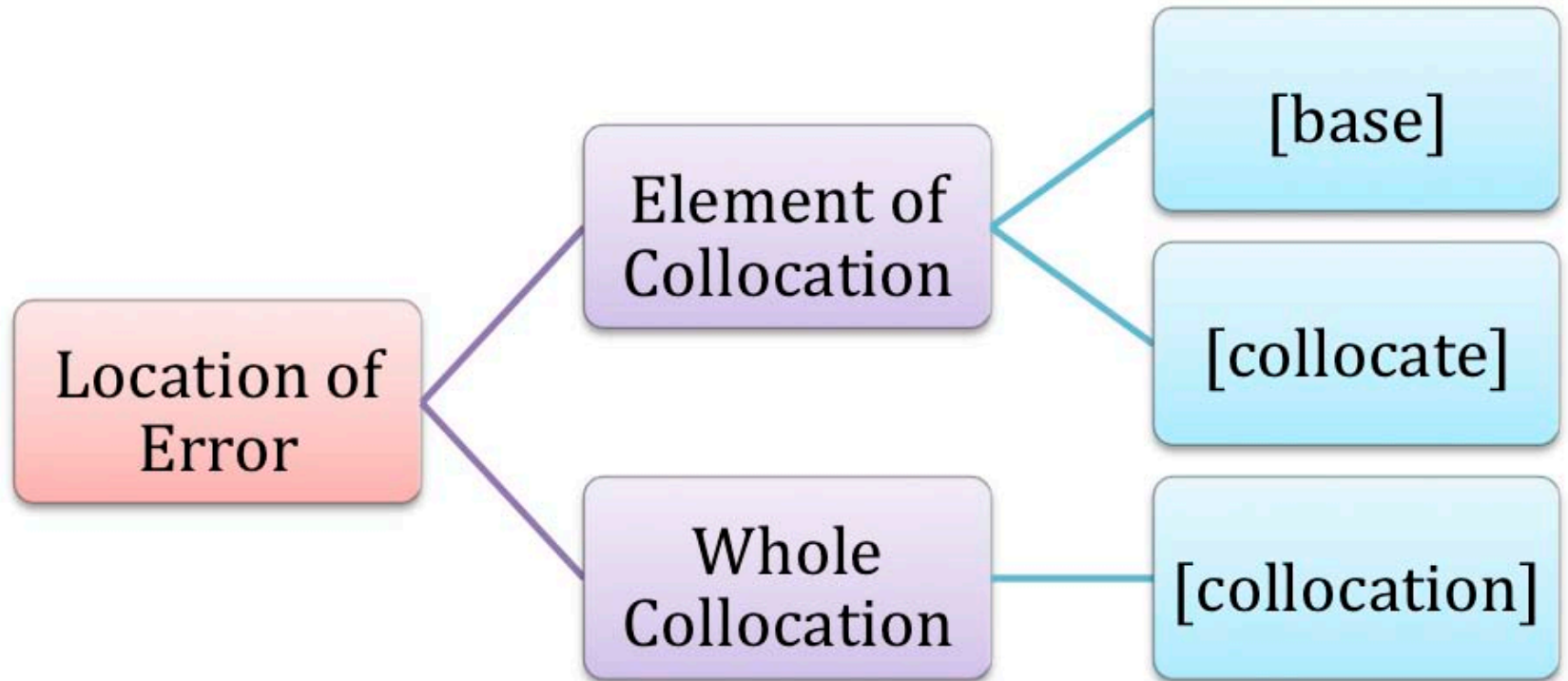
# Outline

- 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**

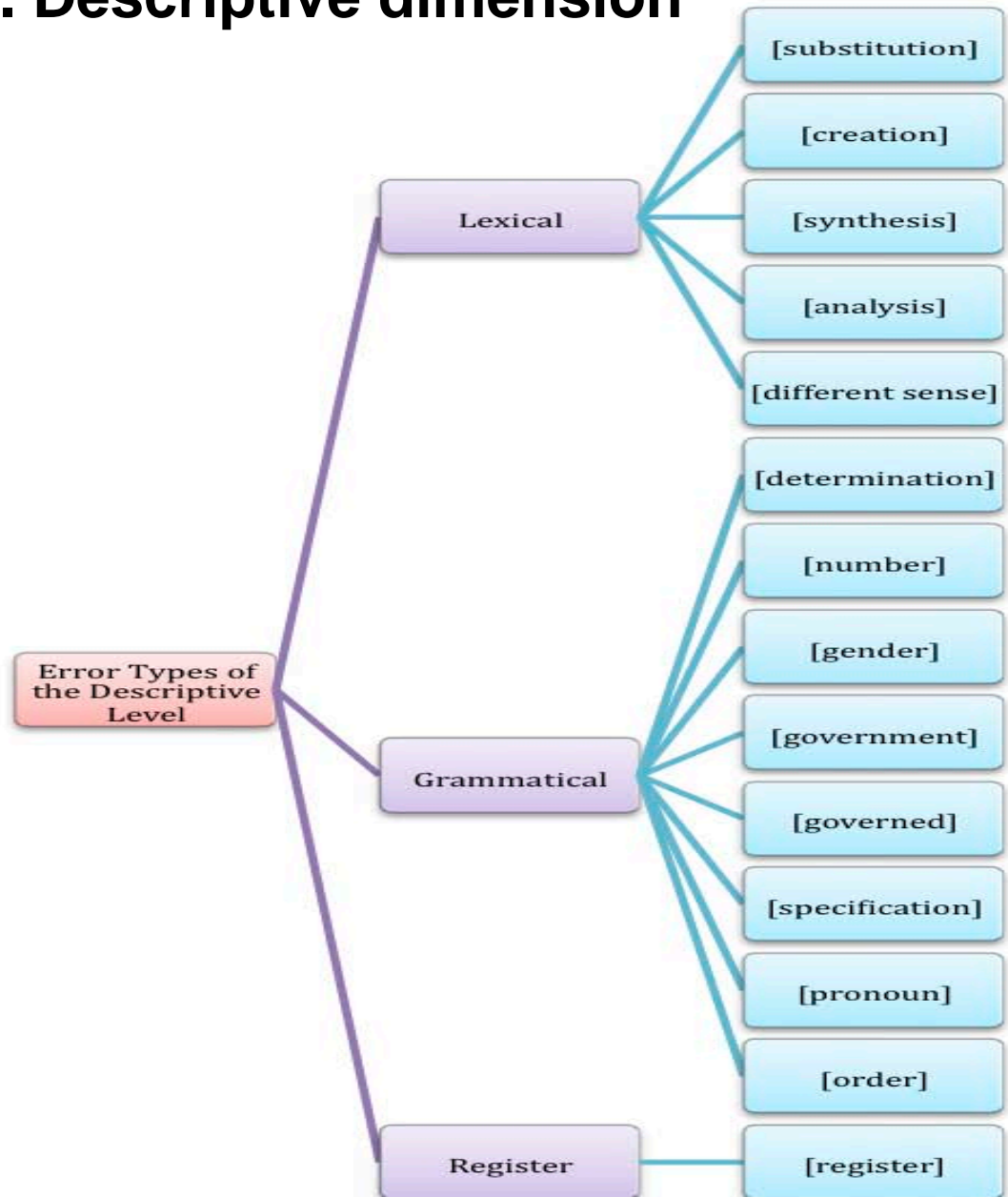
# **1. Three-dimensional Collocation Error Typology:**

- (i) location**
- (ii) descriptive**
- (iii) explanatory**

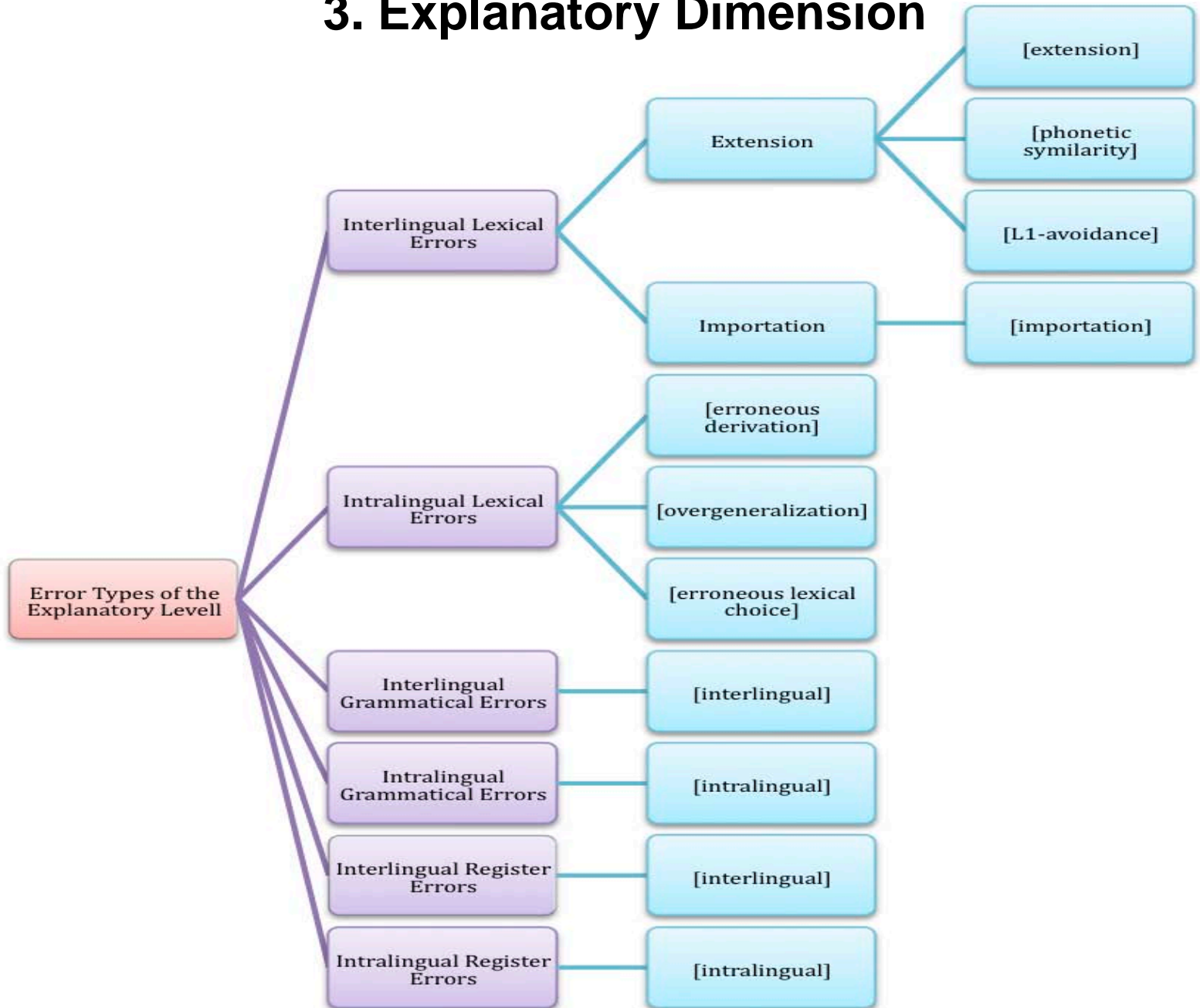
# 1. Location dimension



## 2. Descriptive dimension



# 3. Explanatory Dimension





# Illustration of *interlingual lexical errors* (affecting the *base* or the *collocate*)

		EXPLANATORY CRITERION			
LOCATION	DESCRIPTIVE CRITERION	Interlingual Errors			
		Importation	Extension		
				Phonetic similarity	L1 avoidance
BASE	Substitution	hicimos <u>wakeboarding</u>	<u>juego</u> de fútbol (from <i>game</i> )	Doy una <u>marca</u> (instead of <i>poner una nota</i> )	<u>Estado</u> económico (from <i>economic situation</i> )
	Creation	recibí un <u>llamo</u> (analogy with <i>call</i> )			
COLLOCATE	Substitution		<u>gastar</u> todo el año (from <i>spend</i> ) <u>tomé</u> examen (from <i>take</i> ) <u>Doy</u> una marca (from <i>give</i> )	lengua <u>maternal</u> (from <i>maternal language</i> ) <u>capturar</u> la atención (from <i>capture</i> )	<u>cambiar</u> a la verdadera religión (Eng. <i>convert to a religión</i> ) <u>acudir</u> el teléfono (Eng. <i>attend the telephone</i> )
	Creation				

# 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 <u>llamo</u> (analogy with <i>paseo, canto, salto..</i> )		Me da el <u>sentido</u> (instead of <i>la sensación</i> )
COLLOCATE	enseñanza <u>segundaria</u> (from <i>segundo</i> )	<u>hacer</u> citas (instead of <i>concertar</i> )  <u>malos</u> efectos (instead of <i>nocivos, dañinos</i> )	<u>toman</u> puestos (instead of <i>ocupar puestos</i> )  <u>escribir</u> el examen (instead of <i>hacer</i> )

# Illustration of *interlingual* and *intralingual* grammatical errors (affecting the *base* or the *collocate*)

		EXPLANATORY CRITERION	
LOCATION	DESCRIPTIVE CRITERION	Interlingual errors	Intralingual errors
BASE	determination	<ul style="list-style-type: none"> <li>• tienen <u>el</u> derecho de (from <i>to have the right to</i>)</li> <li>• terminé escuela (from <i>to finish school</i>)</li> </ul>	<ul style="list-style-type: none"> <li>• tomé examen (instead of <i>tomé un examen</i>)</li> </ul>
	number		<ul style="list-style-type: none"> <li>• hicimos los <u>esquí</u>s del agua (instead of <i>esquí de agua</i>)</li> <li>• tienen <u>prejuicio</u> (instead of <i>prejuicios</i>)</li> </ul>
	gender		<ul style="list-style-type: none"> <li>• <u>días</u> festivas (instead of <i>días festivos</i>)</li> </ul>
	government	<ul style="list-style-type: none"> <li>• tengo interés <u>in</u></li> <li>• tengo razones <u>por</u></li> </ul>	
	specification		<ul style="list-style-type: none"> <li>• probar comida (an article or an adjective should be used to specify the base, e.g. <i>probar comida exótica</i>)</li> </ul>
COLLOCATE	government	<ul style="list-style-type: none"> <li>• hablando <u>al</u> teléfono (from Italian, the student's L2: <i>parlare al telefono</i>)</li> <li>• asisto un juego de fútbol (from <i>assist a game of...</i>)</li> <li>• entró la universidad (from <i>enter university</i>)</li> </ul>	<ul style="list-style-type: none"> <li>• montó el autobús (instead of <i>montar en..</i>)</li> </ul>
	pronoun		<ul style="list-style-type: none"> <li>• muero de ganas (instead of <i>por me muero de...</i>)</li> <li>• la película se trata.. (instead of <i>la película trata</i>)</li> </ul>

## **2. The corpus annotation tool: Knowtator**

# The annotation schema in *knowtator*

The screenshot displays the Knowtator application window, titled "copia consenso Protégé 3.3.1". The interface is divided into several panes:

- CLASS BROWSER:** Shows a class hierarchy for the project "copia consenso". The hierarchy includes:
  - :THING
  - :SYSTEM-CLASS
  - knowtator support class
  - Anotaciones Colocate
    - Colocación
      - Colocación Correcta
      - Colocación Incorrecta
    - Cuasiloc.
      - Cuasilocución
    - Error** (selected)

- CLASS EDITOR:** Configures the "Error" class (instance of :STANDARD-CLASS).
- Name: Error
- Role: Concrete
- Template Slots:
  - 1-localización
  - 2-descriptivo
  - 3-explicativo** (selected)
- 3-explicativo (instance of :STANDARD-SLOT):** A detailed configuration window for the selected slot.
- Name: 3-explicativo
- Value Type: Symbol
- Allowed Values: lex\_importación, lex\_extensión, lex semejanza fonética, lex\_evitación\_L1, lex\_derivación, lex\_falsa\_elección, lex\_sobregeneralización, interlingual, intralingual
- Cardinality:  required, at least 1
- Cardinality:  multiple, at most [ ]
- Domain: Error
- Superclasses:** Anotaciones Colocate

# Tagging collocations with *knowtator*

The screenshot displays the Knowtator application window. The main text area contains a paragraph of Spanish text with various words and phrases highlighted in different colors (green, red, purple, blue) to indicate different types of annotations. A blue circle highlights the phrase "comanda respeto de" in the text. On the right side, a panel titled "DELETED comanda respeto de" shows the annotated class "Colocación Incorrecta" and a list of 14 values for the slots of the annotated class. A "Symbol selection" dialog box is open in the foreground, showing a list of symbols with "ext semejanza\_fonetica" highlighted. The dialog box has a red circle around the highlighted symbol and a red arrow pointing to it from the right panel. The bottom of the window shows the text "comanda respeto de".

Colocate\_demo1 - Protégé 3.3.1 (file:\C:\Documents%20and%20Settings\Usuario\Escritorio\proyecto\_demo\Colocate\_demo1.pprj, Protégé Files (.pont and .pins)

File Edit Project Window Knowtator Help

Text source collection: coleccion nueva Filter: Ver colocaciones

Knowtator Classes

annotation schema

- Colocación Correcta (7)
- Colocación Incorrecta (14)
- Cuasilocución
- Base
- Colocativo

text source: B.txt filter: Ver colocaciones

El año pasado durante las vacaciones del verano, trabajaba para los directores de unos grupos musicales de música punk. El trabajo era en Los Angeles, donde **antes de irme**. Empecé mi experiencia con la compañía, se llama Villam Artist Management, como voluntaria, porque **tenía ganas de** aprender eso tipo de trabajo. La compañía compartía un edificio en Hollywood con otra compañía que **produce discos** porque uno de nuestros grupos que **tiene el más éxito ha producido dos discos** **debajo de la marca** de la otra compañía (Side One Dummy Records). La compañía para que trabajara es muy pequeña. Ellos consisten de dos empleados- uno es él con el más poder, el **gran jefe** bromeábamos, y la otra es su única empleada que él no le ha despedido. El **problema es que él pide** demasiado de sus empleadas. El es un **gran hombre de negocios** y **comanda respeto de** todos que le conocen, pero sus **actitudes** son demasiado **altas**. Él quiere que sus empleadas **pasen un tiempo** de trabajar tan **fuerte** como él, y también, no acepta "no" para una respuesta a lo que pide. Cuando llegué a la compañía, había una secretaria sustituta porque él la ha despedido a su secretaria de dos años porque ella pedía una jornada más corta y él pensaba que esa pedida significaba que ella era perezosa. El próximo secretario era el primo de la otra empleada. Eso nuevo trabajador descasaba pronto de la cantidad de trabajo muy rápido y de los **estándares altos** que **era** Gary (el jefe). **Se le recomendó de** **dejarlo**, yo trabajaba como secretario. El me pagaba para esa trabaja, aunque era muy difícil **obtener el dinero** que él me ha prometido. En mi nueva posición, yo **me planea** de viajar para los grupos, **acompañar al teléfono** e **acompañar** para conferencias con otras compañías para Gary. También seguí con mis responsabilidades de antes- que todavía tenía que buscar nuevas maneras para hacer más eficiente la compañía. El trabajo era muy difícil—de hecho, todo de ese verano era muy difícil. Llegué a la oficina por la mañana a las 8 y salí a las 5, para que pudiera ir a mi otro trabajo como camarera en un restaurante. Por lo general, terminé **la medianoche** y como soy joven, y nada es más importante en este punto de mi vida que mis amigos y mi vida social, salí con ellos hasta muy temprano... yo sé que la falta de sueño es culpa mía, pero sin embargo, era difícil. Lo bueno del trabajo para Villam Artists es que Gary no me despidió tampoco quería ir yo cuando le dejé. Era un **gran chico** cuando Gary me invitó a almorzar conmigo y me contó que siempre habrá sitio para mi en su compañía si quería regresar para él después de escuela.

DELETED **comanda respeto de**

span edit: << >> <<< >>>

annotated class

- Colocación Incorrecta

slots of annotated class (14 values)

- 12-función Léxica
  - Caus
  - Func
  - 1
- 2-Error1: Localización Error
  - colocativo
- 21-Error1: Tipo Error Descriptivo
  - Lex\_miembro\_colocacion
- 22-Error1: Tipo Error Explicativo

Symbol selection

- importacion
- importacion
- extension
- ext semejanza\_fonetica
- ext evasivo\_1.1
- derivacion\_erronea
- comodin
- falsa\_eleccion
- gram\_transferencia

annotation set membership

comanda respeto de

### **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

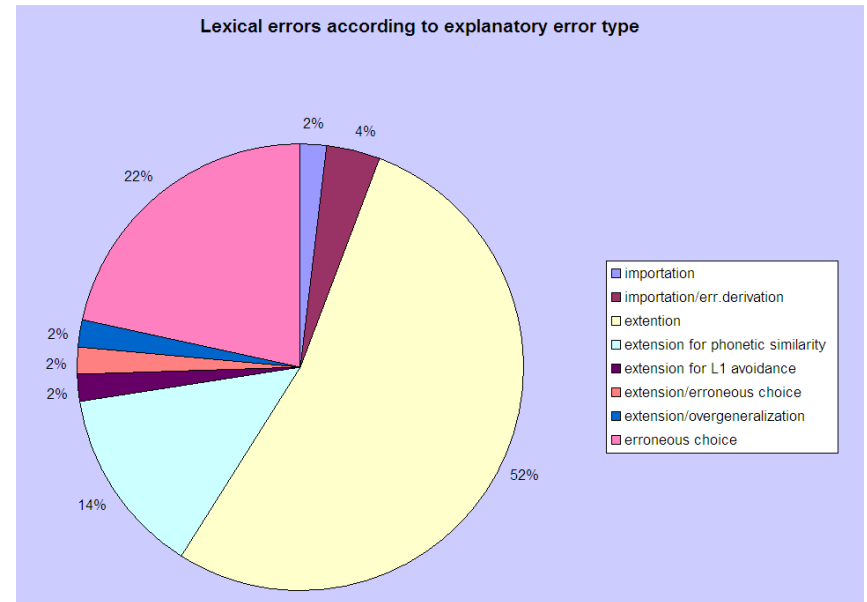
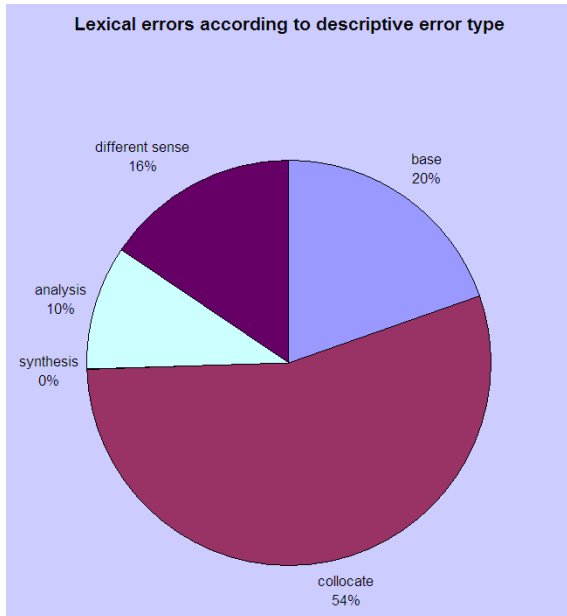
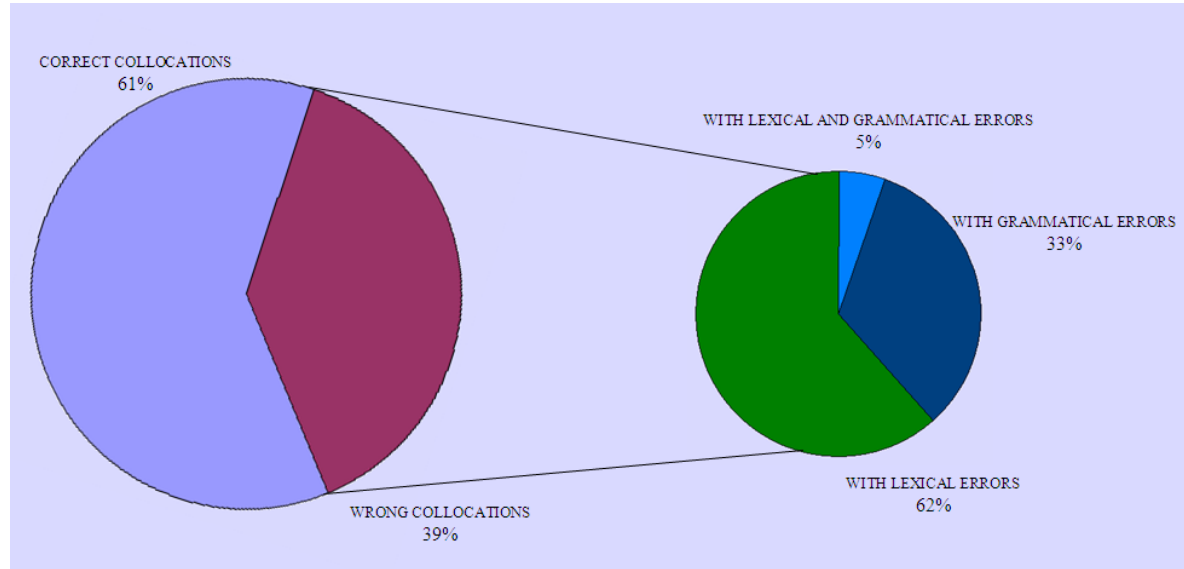
## C) Develop NLP-techniques

- 1) For automatic recognition and classification of collocations
- 2) For automatic error correction and learning material provision



## **4. Preliminary Findings**

# 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!**