Oral expression in Spanish by low-intermediate learners: a computer-aided error analysis

Leonardo Campillos Llanos

Computational Linguistics Laboratory

Autonomous University of Madrid (UAM)

Learner Corpus Research Conference (LCR2013) – Bergen, 28th September 2013

Index

- Background of Spanish learner corpus research
- Goals
- Methodology
  - ❑ Corpus design
  - ❑ Error typology
  - ❑ Error analysis
- The corpus search interface
- Results and discussion
- Conclusions

Background of Spanish learner corpus research

- Most learner corpora comprise written data: e.g. **International Corpus of Learner English**.
- Few research projects on spoken learner corpus:

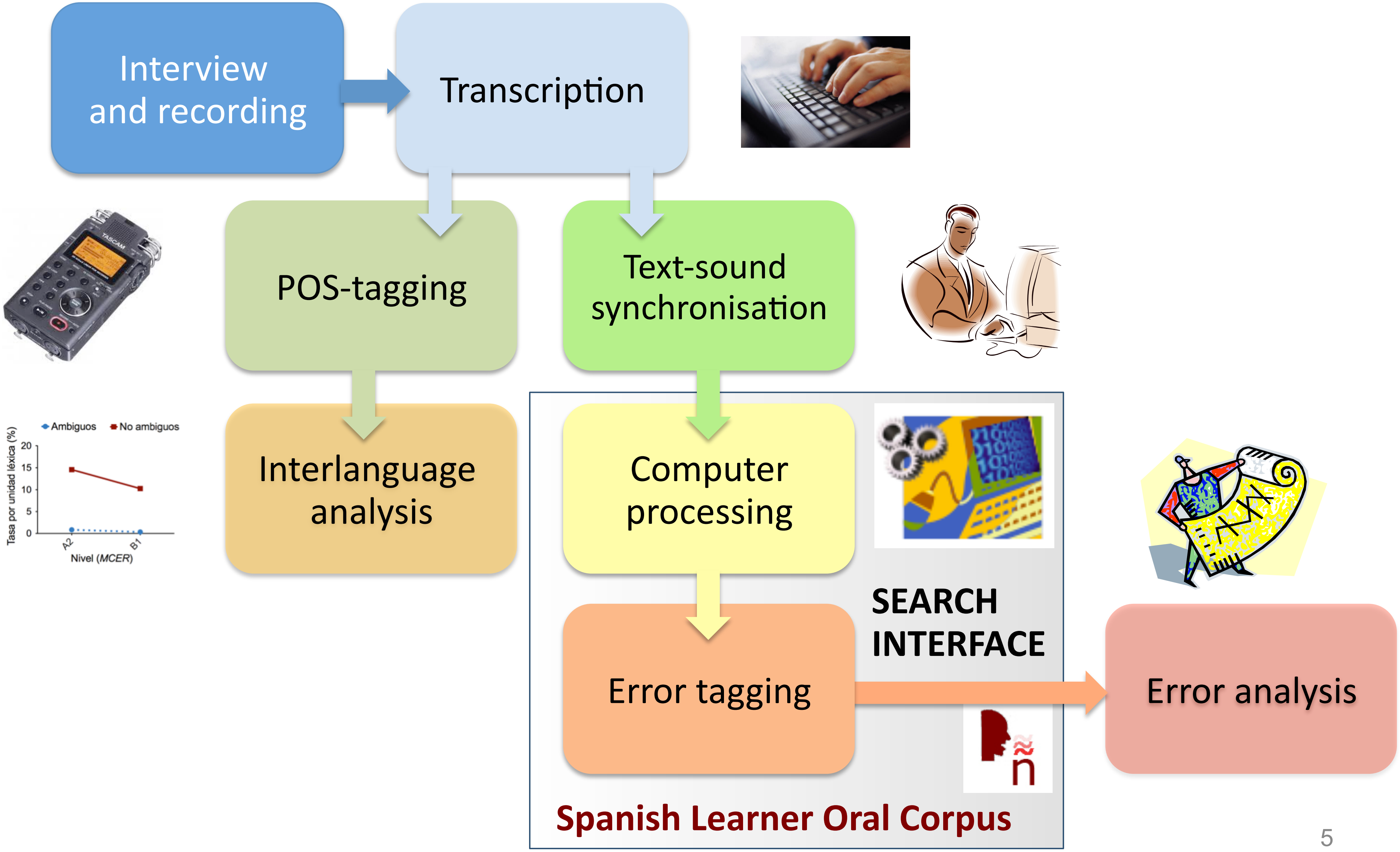
ENGLISH	FRENCH	SPANISH
LINDSEI (Louvain International Database of Spoken English; Gilquin et al., 2010)	FLLOC (French Learner Language Oral Corpus; Myles, 2005)	The Díaz Corpus (Díaz Rodríguez, 2007)
NICT JLE (NICT Japanese Learner Corpus; Izumi et al., 2004)		SPLLOC (Spanish Learner Language Oral Corpora; Mitchell et al., 2008)

Goals

- Fulfil the lack of **oral corpus** and **computerised resources** for Learner Corpus Research

- To understand the acquisition of the oral expression by different groups of learners of Spanish at **A2** and **B1 levels** (*CEFR*).

Methodology



Interview and recording

Transcription

POS-tagging

Text-sound synchronisation

Interlanguage analysis

Computer processing

Error tagging

SEARCH INTERFACE

Error analysis

Spanish Learner Oral Corpus

Corpus design

- Cross-sectional corpus.
- Participants: Foreign students of Spanish (20-26 years old).
- Low-intermediate level: A2 (N=20) and B1 (N=20) (*CEFR*)
- N=40, clustered in:
- 9 groups of 4 students with the same L1:

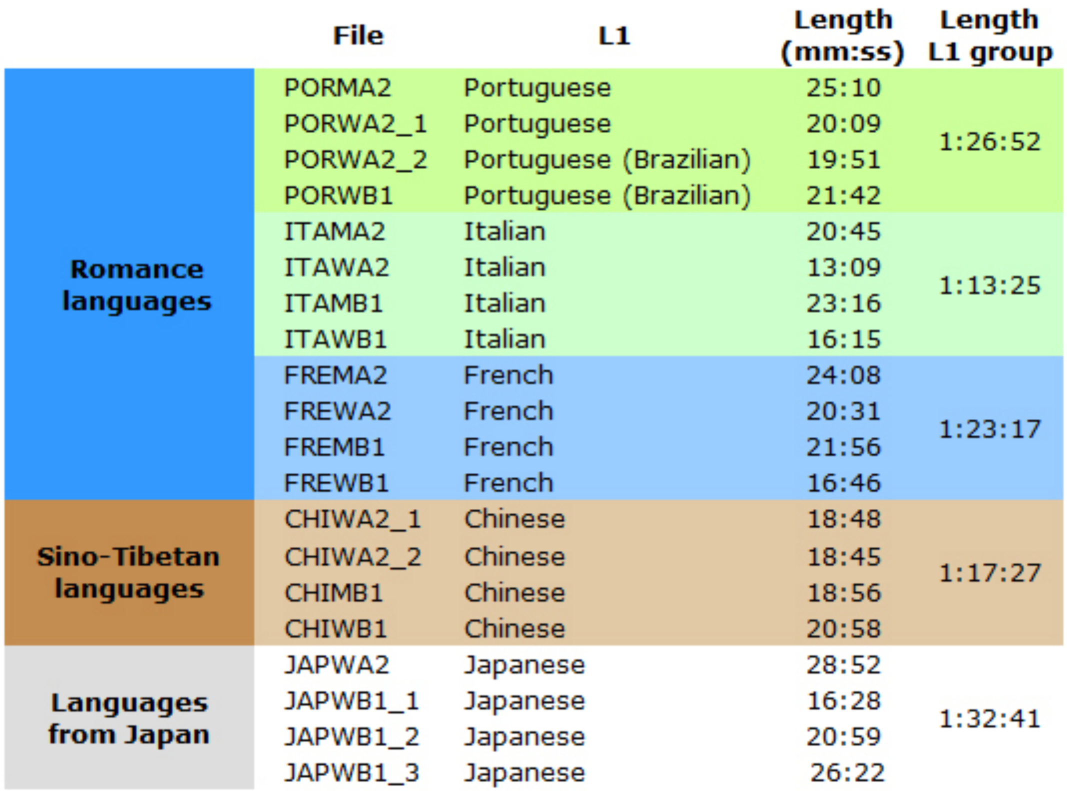
  Italian
  French
  Portuguese
  English
  German
  Dutch
  Japanese
  Chinese
  Polish

- 1 mixed group of 4 students with other L1s:

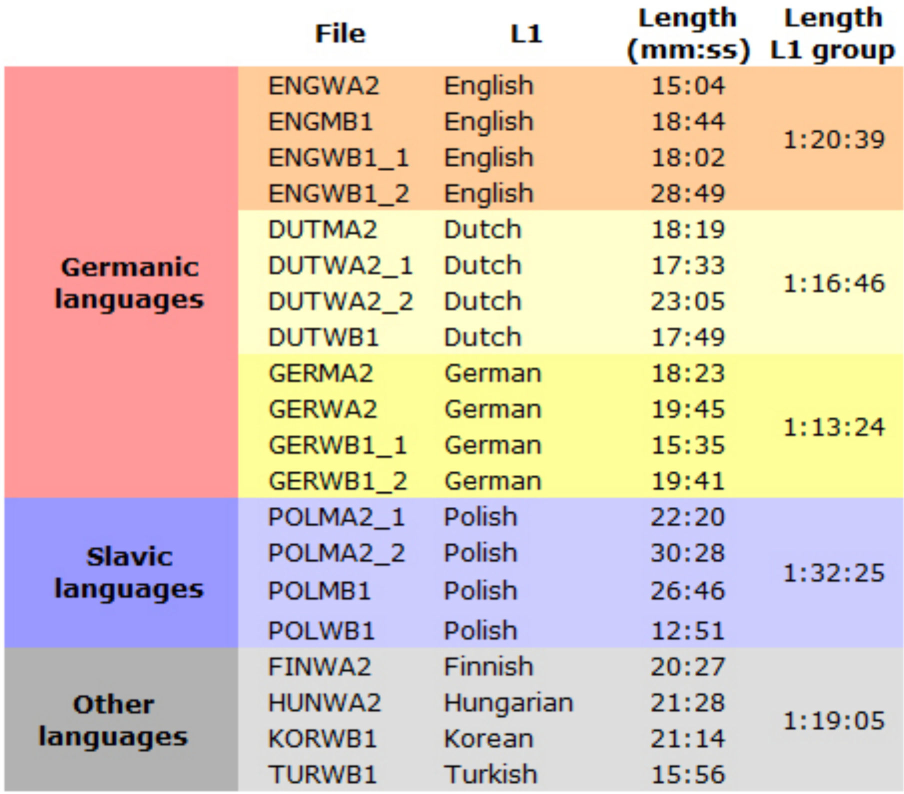
  Korean
  Turkish
  Finnish
  Hungarian

- Control group of native speakers (N=4): 2 men and 2 women.

Corpus design



	File	L1	Length (mm:ss)	Length L1 group
Romance languages	PORMA2	Portuguese	25:10	1:26:52
	PORWA2_1	Portuguese	20:09	
	PORWA2_2	Portuguese (Brazilian)	19:51	
	PORWB1	Portuguese (Brazilian)	21:42	
	ITAMA2	Italian	20:45	1:13:25
	ITAWA2	Italian	13:09	
	ITAMB1	Italian	23:16	
	ITAWB1	Italian	16:15	
	FREMA2	French	24:08	1:23:17
	FREWA2	French	20:31	
	FREMB1	French	21:56	
	FREWB1	French	16:46	
Sino-Tibetan languages	CHIWA2_1	Chinese	18:48	1:17:27
	CHIWA2_2	Chinese	18:45	
	CHIMB1	Chinese	18:56	
	CHIWB1	Chinese	20:58	
Languages from Japan	JAPWA2	Japanese	28:52	1:32:41
	JAPWB1_1	Japanese	16:28	
	JAPWB1_2	Japanese	20:59	
	JAPWB1_3	Japanese	26:22	

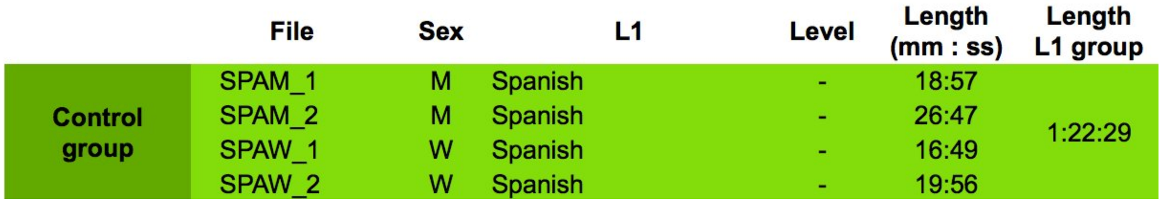


	File	L1	Length (mm:ss)	Length L1 group
Germanic languages	ENGWA2	English	15:04	1:20:39
	ENGMB1	English	18:44	
	ENGWB1_1	English	18:02	
	ENGWB1_2	English	28:49	
	DUTMA2	Dutch	18:19	1:16:46
	DUTWA2_1	Dutch	17:33	
	DUTWA2_2	Dutch	23:05	
	DUTWB1	Dutch	17:49	
	GERMA2	German	18:23	1:13:24
	GERWA2	German	19:45	
	GERWB1_1	German	15:35	
	GERWB1_2	German	19:41	
Slavic languages	POLMA2_1	Polish	22:20	1:32:25
	POLMA2_2	Polish	30:28	
	POLMB1	Polish	26:46	
	POLWB1	Polish	12:51	
Other languages	FINWA2	Finnish	20:27	1:19:05
	HUNWA2	Hungarian	21:28	
	KORWB1	Korean	21:14	
	TURWB1	Turkish	15:56	

NAME OF THE FILE

Key of the 3 letter code: L1 + M: man / W: woman + level *CEFR* (A2 or B1) + file number (optional)

e.g. PORWA2_1: woman, Portuguese as L1, A2 level, file 1.



	File	Sex	L1	Level	Length (mm : ss)	Length L1 group
Control group	SPAM_1	M	Spanish	-	18:57	1:22:29
	SPAM_2	M	Spanish	-	26:47	
	SPAW_1	W	Spanish	-	16:49	
	SPAW_2	W	Spanish	-	19:56	

Total:

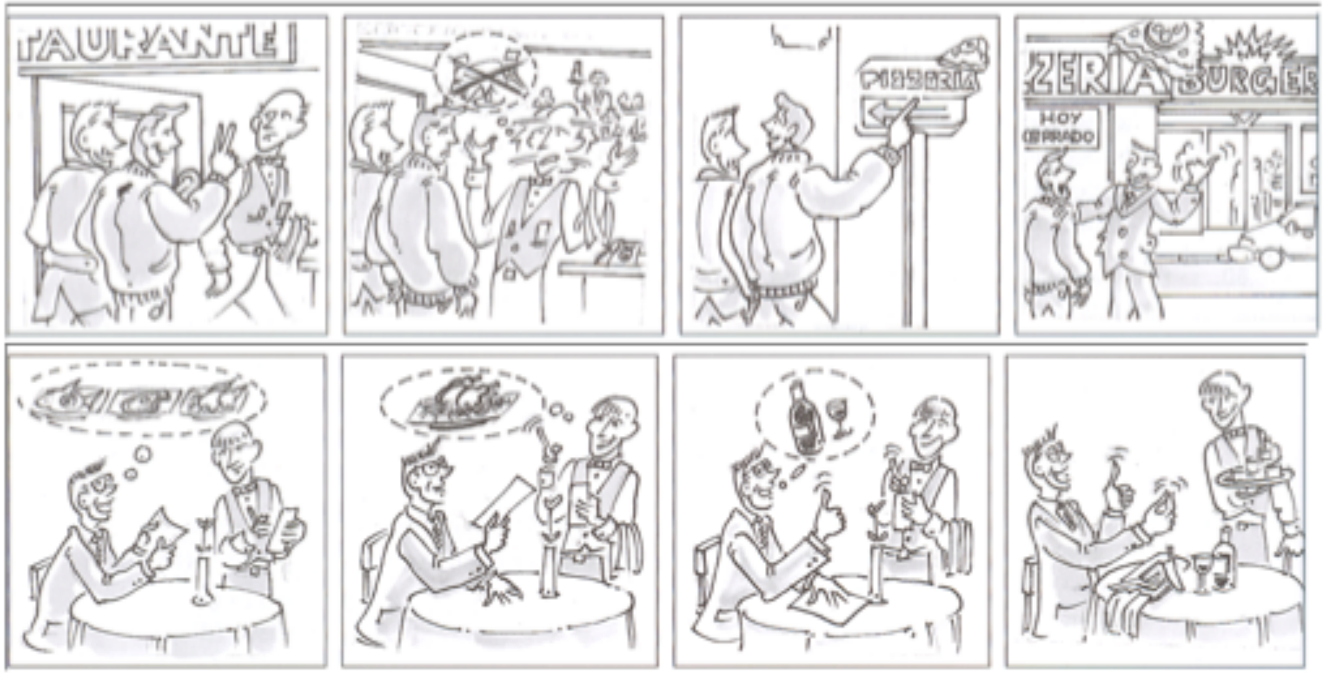
13 hs 36’

Data collection method

- One-to-one **semi-controlled** spoken interviews.
- **15-20** minutes long each recording.
- Tasks: (similar to foreign language examinations)
  - □ **Description of two photographs** about food.

Data collection method

- Tasks (cont.):
  - **Story retelling task** from pictures.



  - A question about two **speech acts.**
  - **Spontaneous dialogue**: opinion about topics related to food.

The corpus search interface

- http://cartago.lllf.uam.es/corele/index.html

Welcome!

¡Bienvenido!

Spanish Learner Oral Corpus
Corpus Oral de Español como Lengua Extranjera (ELE)

Error typology

- Classification according to several **criteria**:
  - **Linguistic level**: e.g. Grammar: *la casa ****blanco*** → *blanca*

    ('the white house')
  - **Target modification**: e.g. Unnecesary: ******un*** *mi amigo* ('a my friend')
  - **Category**: e.g. verb: ****tiengo*** → *tengo* ('I have')
  - **Type**: e.g*. ser/estar*: ****soy*** *satisfecho* → *estoy* ('I am satisfied')
  - **Etiology**: e.g. interlinguistic: e.g. *to realise* ('darse cuenta')

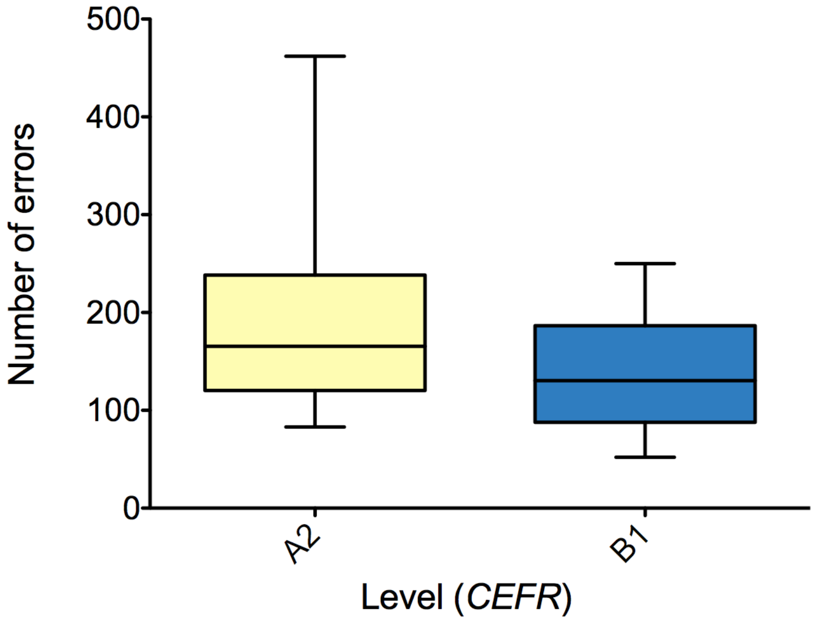
    ≠ *realizar* ('to make')

Error analysis

- Error analysis of:
  - ❑ Grammar.
  - ❑ Lexis.
  - ❑ Pronunciation.
  - ❑ Pragmatics-Discourse.
- **Word counts** for each morphological category were obtained to **normalise** error frequencies.

Results

- **6,838** errors in 52,688 lexical units → **12.98** errors per **100** lexical units
- A mean of **170.63 errors per interview** (SD = 90.36).
- Progress from A2 to B1 shows a **diminution of errors**:



Results

- These data only **partially** reflect the acquisition process:

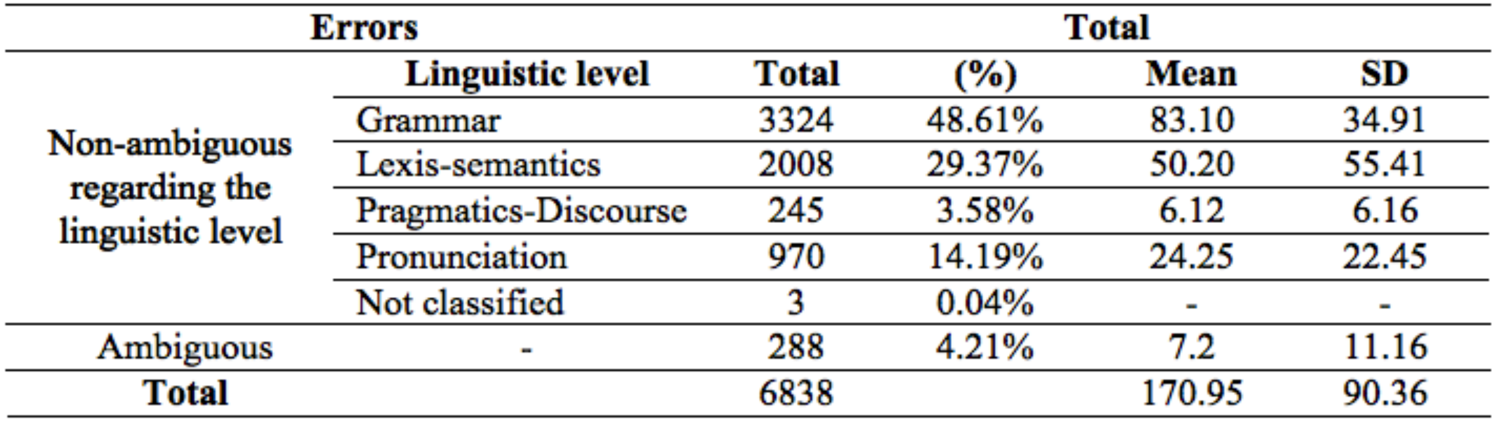
  - They can be related to the **avoidance** of difficult structures.

  - Learners at intermediate levels would be **expected** to make **more errors** than students at lower levels.

→ Students are trying or practising new structures.

Results

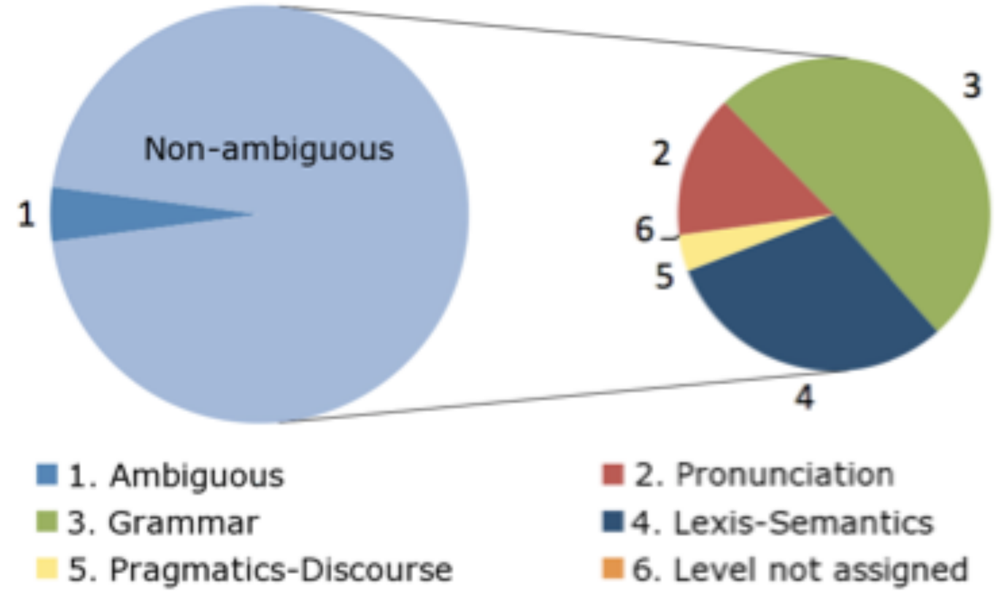
- Most errors affect:
  - **Grammar** (48.61%)
  - **Lexis** (29.37%)
- Fewer errors in:
  - **Pronunciation** (14.19%)
  - **Pragmatics-Discourse** (3.58%)



Errors		**Total**			
	Linguistic level	**Total**	**(%)**	**Mean**	**SD**
Non-ambiguous regarding the linguistic level	Grammar	3324	48.61%	83.10	34.91
	Lexis-semantics	2008	29.37%	50.20	55.41
	Pragmatics-Discourse	245	3.58%	6.12	6.16
	Pronunciation	970	14.19%	24.25	22.45
	Not classified	3	0.04%	-	-
Ambiguous	-	288	4.21%	7.2	11.16
Total		6838		170.95	90.36

Results

- Around **4.21%** are **ambiguous** errors.
- **49.21%** of errors would be due to **interference**.



Results

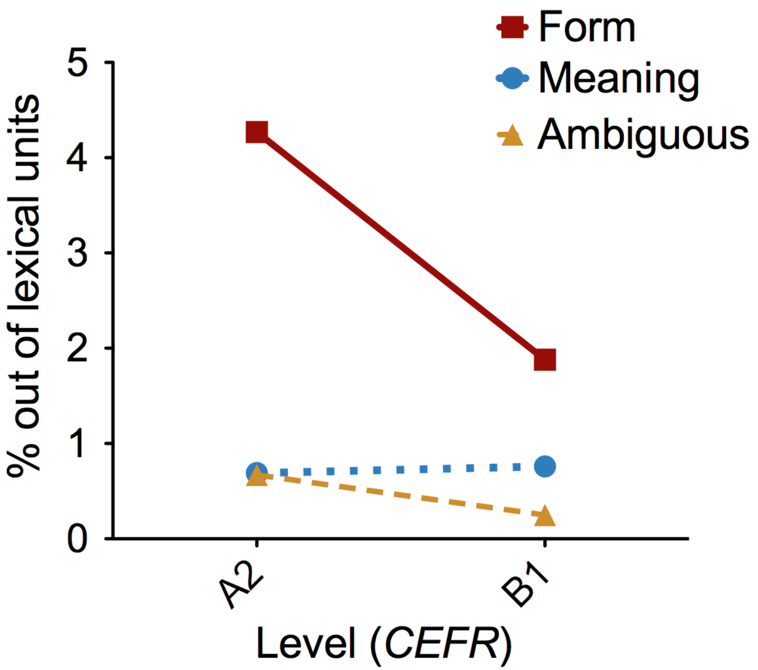
- **Lexical errors** at A2-B1 levels:
  - **Formal errors** are **more frequent** (80.73% of lexical errors)
    - e.g. borrowings, misformations, malapropisms, gender, calques…
  - **Semantic errors** are **less frequent** (19.12% of lexical errors)
    - e.g. semantic relation errors, false friends, collocations, register…

Note that in the following I will show only figures for non-ambiguous errors.

Results

- The **rate of formal errors decreases at B1**
- The **rate of semantic errors persists** and **slightly increases at B1**

→ **Semantics is more difficult to acquire**.



Results

- At **A2**, the most frequent **lexical errors** are:
  - **Borrowings**: M = 21.87 (SD = 46.99)

    e.g. *tocaba *guitare* ('I played guitar') → *guitarra*

    → There is a **large standard deviation** due to the fact that **borrowings are very frequent** among **Portuguese, German, and Dutch learners**.

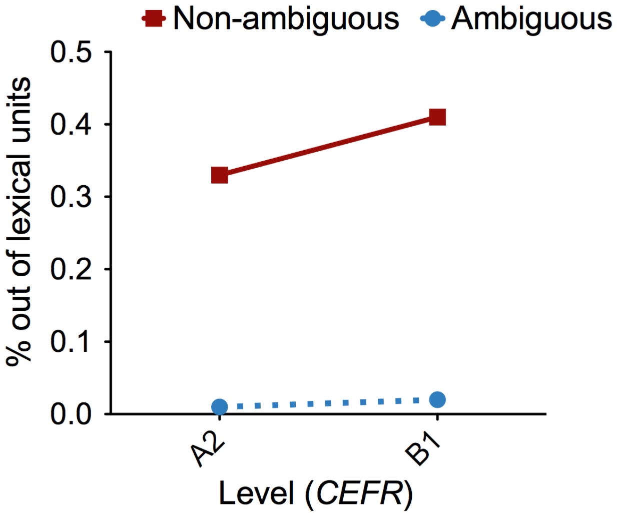
  - **Misformations**:

    e.g. **melijones* → *mejillones* ('mussels')

Results

- **Lexical errors decline** at **B1**, but some **persist** or **hardly decrease**:
  - **Semantic relation**:

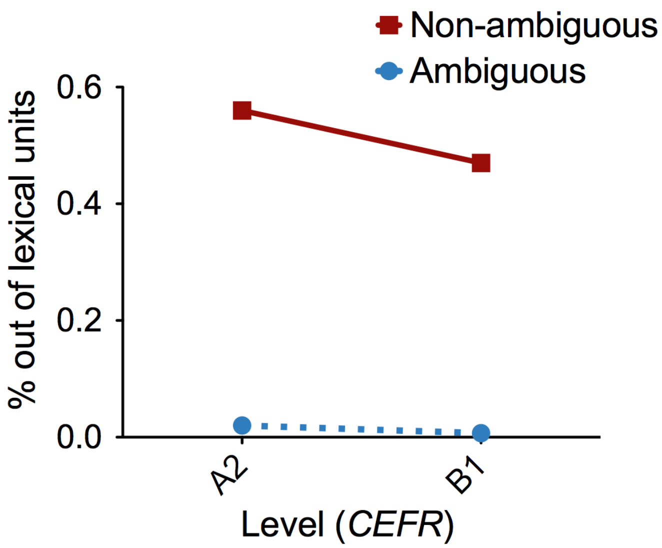
    e.g. confusion *ir* ('to go') ~ *venir* ('to come')



Results

- **Lexical errors decline** at **B1**, but some **persist** or **hardly decrease**:
  - **Gender**:

    e.g. *el bolso* ('handbag') ~ *la bolsa* ('bag')



Results

- **Grammar**: the most frequent and generalised **errors** affect:
  - **Articles**:

    e.g. *y* ***Ø*** *camarero está contento* → ***el*** *camarero*

    (‘and [the] waiter is happy’)

  - **Agreement**:

    e.g. *la comida *****famoso*** → ***famosa***

    (‘the famous food’)

Results

- **Grammar**: the most frequent and generalised **errors** (cont.):
  - ❑ **Prepositions**:

    e.g. *estoy aquí *__a__ Madrid* → ***en** Madrid*

    ('I am here in Madrid')

  - ❑ **Pronouns:**

    e.g. *a mí **Ø** encanta la pizza* → ***me** encanta* ('I love pizza')

Results

- **Grammar**: the most frequent and generalised **errors** (cont.):
  - **Sentence structure**:

    e.g. Blends: *estudio ****algo como se llama*** *Estudios de cultura*

    → *estudio* ***algo que*** *se llama…* or *estudio* ***algo como*** *…*

    (‘I study something called Culture studies’ or ‘I study something like…’)

  - **Past tense**:

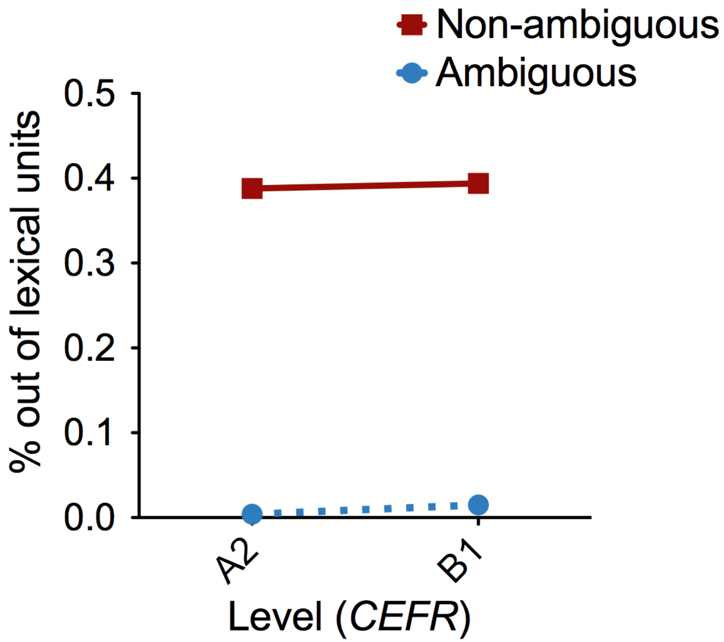
    e.g. *hace 30 años las mujeres no ****trabajaron*** → ***trabajaban***

    (‘women did not use to work 30 years ago’)

Results

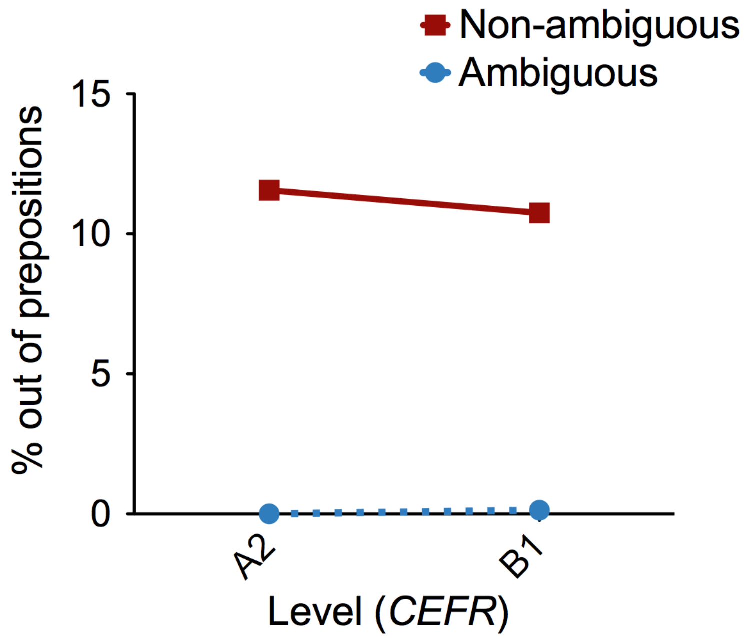
- Certain **grammar errors persist** or **hardly decrease** at **B1**:
  - **Pronouns**: e.g. *Él no sabe qué* ***se** quiere* → *Él no sabe qué quiere*

    (‘He does not know what he wants’)



Results

- Certain **grammar errors persist** or **hardly decrease** at **B1** (cont.):
  - ❑ **Prepositions**: e.g. *He venido ***en** Madrid* → *He venido **a** Madrid* (‘I have come to Madrid’)

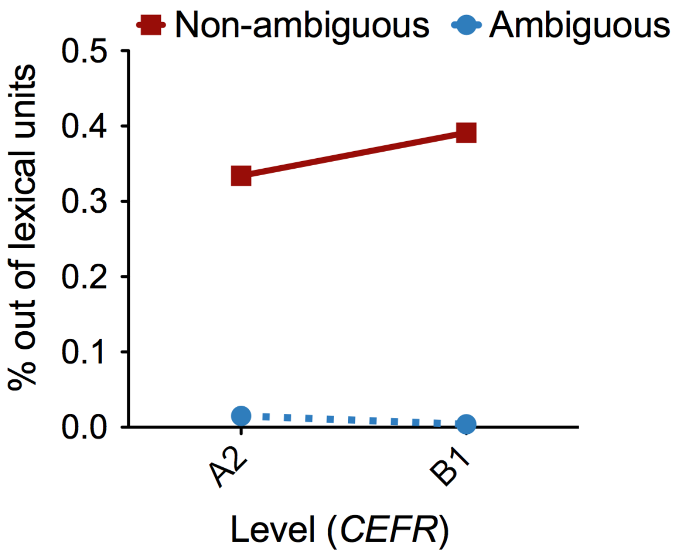


Results

- Certain **grammar errors persist** or **hardly decrease** at **B1** (cont.):
  - ❑ **Subordination**: e.g. *Espero ****que entiendo*** *qué pasa*

    → *Espero* ***entender*** *qué pasa*

    (‘I hope to understand what happens’)



Results

- The **characteristics of spoken discourse** may explain the **high number** of the following **grammar** errors:

  - ❑ **Sentence structure** errors, especially:
    - ➢ Omission: e.g*. su restaurante* ***Ø*** *muy bien* → ***está*** *muy bien* (‘his restaurant **is** very nice’)
    - ➢ Word order: e.g. *****no realmente*** *sé* → ***realmente no*** *sé* (‘I really do not know’)
  - ❑ **Agreement**: e.g. *****unos*** *amigas* → ***unas*** *amigas* (‘some friends’)
  - ❑ **Overuse of present tense**.

Results

- **Pronunciation errors**: **interference phenomena** tend to **strongly persist at B1**

  → The **L1** maybe has the **greatest** influence.

  ❑ However, learners from **every language background** commit certain errors: e.g. the articulation of ***/r/***: *perro* /'pero/ ~ *pero* /'pero/
  (‘dog’) (‘but’)

- **Pragmatics-Discourse errors** show a **wide individual variability**

  → each **learner’s rhetoric skills** in the L1 may explain these results

Discussion

- **Limitations of the study**:
  - Only oral data have been used → it is difficult to diagnose:
    - the **type** or the **linguistic level** of certain deviations
    - whether they are due to **competence** or **performance**
  - **Low number** of participants per L1 group, and **only** at **A2-B1 levels**:
    - results **cannot be generalised**
    - **conclusions** as to the possibility of **acquiring** an almost **bilingual** proficiency **cannot be inferred**

Discussion

- Some **results** are **similar** to error analyses of **written learner corpora of Spanish** (Fernández 1997) and **English** (Díez Bedmar 2011b):
  - The **most frequent** errors affected **grammar**, especially:
  - articles
  - verbs
  - pronouns
  - The **second most frequent** types of errors were **lexical errors**.
  - **BUT statistical significance** does not imply **pedagogical significance** (Díez Bedmar 2011a)

Thank you for your attention!

Contact

Leonardo Campillos Llanos:

leonardo.campillos@uam.es leonardo.campillos@gmail.com

Corpus interface: http://cartago.lllf.uam.es/corele/index.html