Part-of-Speech Tagging a Spanish Learner Oral Corpus

Criteria, Procedure, and a Sample Analysis

Leonardo Campillos Llanos

LIMSI & UAM

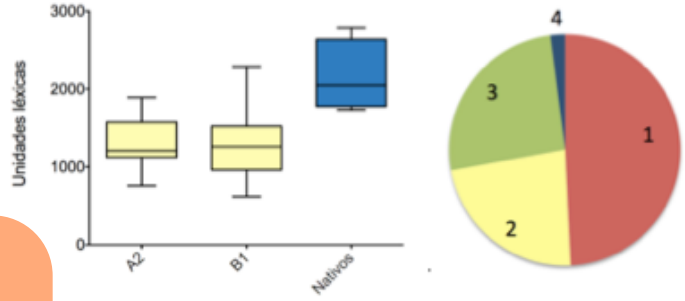
Spanish Learner Corpus Workshop

La Coruña, 14 July 2015

Outline

- Introduction
- Literature review
  - Background of Spanish learner corpus research
  - Part-of-Speech tagging learner corpora
  - Studies on the acquisition of Spanish articles
- Methodology
  - Corpus design
  - Part-of-Speech tagging
  - The corpus search interface
  - Analysis
- A sample analysis of the production of Spanish articles
- Results and discussion
- Conclusions

Introduction

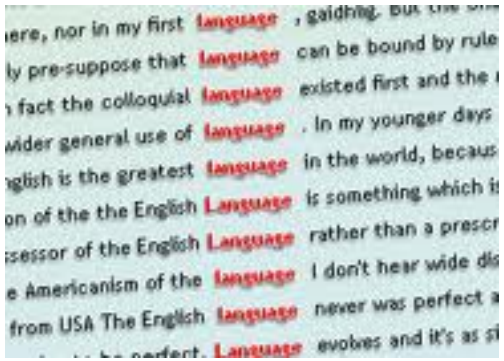


Second language acquisition research

Interlanguage analysis

Corpus linguistics

Computer tool



Background of Spanish learner corpus research

- Spanish learner text corpora available:
  - **CEDEL2** (Lozano and Mendikoetxea, 2013)
  - **CAES** (http://galvan.usc.es/caes)
  - **The Aprescrilov corpus** (Buyse et al. 2012)
- Few research projects on spoken learner corpus:

<table>
<tr><th>ENGLISH</th><th>FRENCH</th><th>SPANISH</th></tr>
<tr><td>LINDSEI
(Louvain International Database of Spoken English; Gilquin et al., 2010)</td><td rowspan="2">FLLOC
(French Learner Language Oral Corpus; Myles, 2005)</td><td>The Díaz Corpus
(Díaz Rodríguez, 2007)</td></tr>
<tr><td>NICT JLE
(NICT Japanese Learner Corpus; Izumi et al., 2004)</td><td>SPLLOC
(Spanish Learner Language Oral Corpora; Mitchell et al., 2008)</td></tr>
</table>

- Phonetic corpora: **FonoELE** (Blanco 2012, 2014)
- **Multimodal** corpora: www.laits.utexas.edu/spt/

Literature review

- **Annotation of Interlanguage/Learner corpora**:
  - Computer-aided Error Analysis (CEA, Dagneaux et al. 1998)
    -  Search and consult errors in context
    -  Error tagging is time-consuming
    -  Learners avoid structures (Schachter 1974)
  - Contrastive Interlanguage Analysis (CIA, Granger 1996)
    -  Compare L1-L2 (most common), or different L2s
    -  Comparative fallacy (Bley-Vroman 1983)
    - ➡ **Native speakers as a benchmark?**
    - ➡ Complementary approaches

*Also: Automatic tagging to describe errors (Diaz-Negrillo et al . 2010)

Part-of-Speech tagging learner corpora

- **Annotation of Interlanguage/Learner corpora**:
  - Initiatives to PoS tag learner data:
    - Mostly English texts: e.g. Tono (2002), Díaz-Negrillo et al. (2010)
    - Other languages: Italian (Rastelli 2006), Norwegian (Tenfjord et al. 2006), Czech (Rosen et al. 2014)
    - Oral data: FLLOC (Myles 2005), SPLLOC (Mitchell et al. 2008)
  - Parsed data (e.g. Rosén & De Smedt 2010; Dickinson & Raheb 2011)
    - ➡ syntactic properties

See an overview of NLP for learner data in Meurers (2015)

Part-of-Speech tagging learner corpora

- **Annotation of Interlanguage/Learner corpora**:
  - Performance of taggers on learner data **decreases**
  - How do native PoS categories fit with **non-native** data?
  - ➡ attempts to refine tagsets (Gaillat 2013; Gaillat et al. 2014)
  - CIA analyses have proliferated:
    - Modals
    - Vocabulary and collocations
    - Connectors
    - Sequences of grammatical categories (e.g. Tono 2000)
    - Clitics and word order...

See panorama in Granger 2004

Goals

- Have an insight into learners' production of L2 categories
  - ➡ a sample study on learners' production of articles

- Fulfil the lack of **computerised, annotated corpora** for Learner Corpus Research.

- ➡ an online interface to perform exploratory analysis of learner corpus data.

Studies on L2 articles

- Most research on L2 English.

See an overview in Díez-Bedmar & Papp 2008; Díez-Bedmar & Pérez Paredes 2012

- Summary of results:
  - Learners whose L1 lacks articles use this category less accurately.

Studies on articles in L2 Spanish

- Error analyses

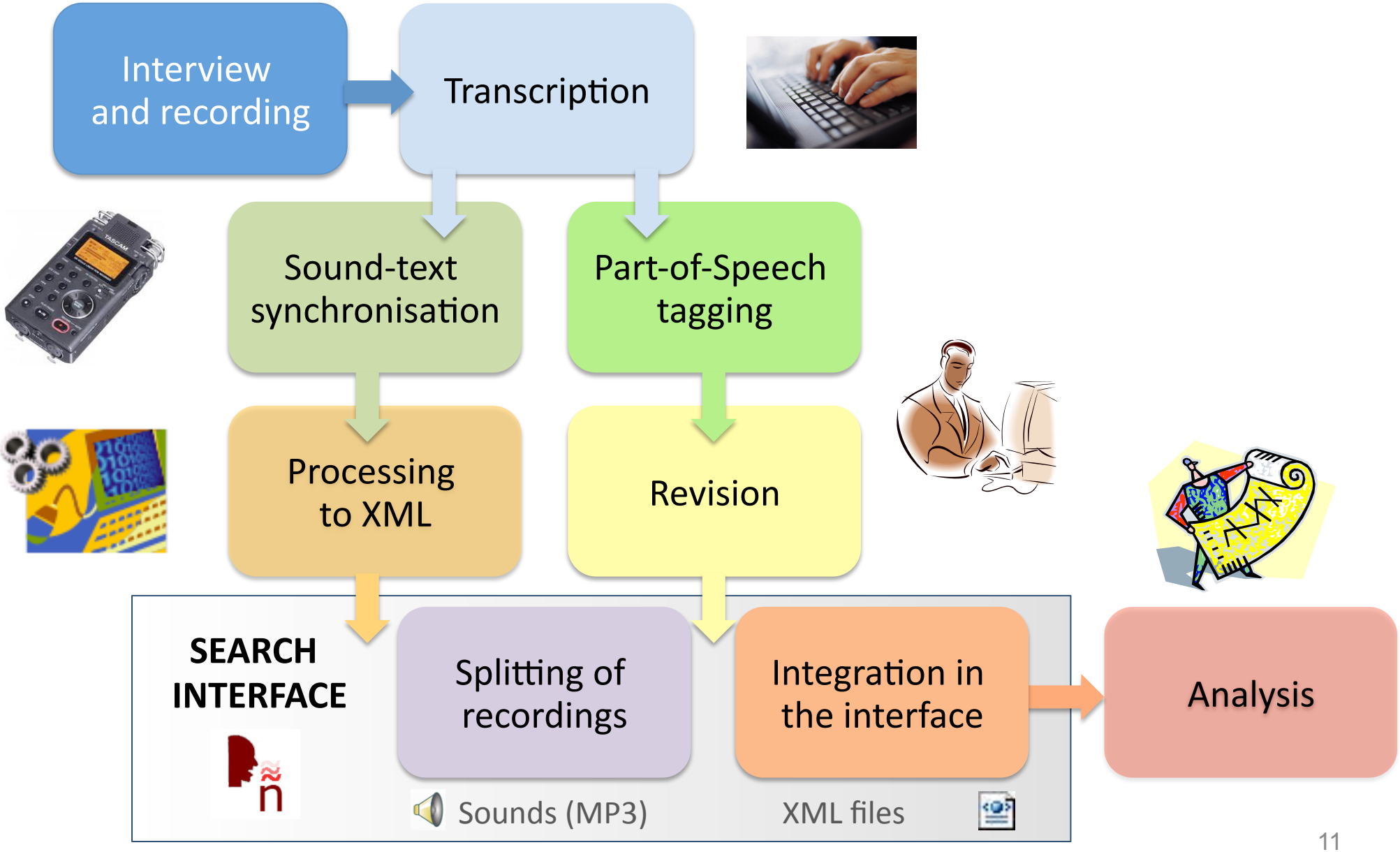
  e.g. Fernández 1990; Santos 1991; Vázquez 1991

- Focused studies on the acquisition of definite articles, or definite and indefinite articles

  e.g. Ramírez-Mayberry (1998), Said-Mohand (2007), Goitia (2007), Lin (2005), Tarrés (2007), Lu & Hsueh (2012) Valverde & Ohtani (2014)

- Summary of results:
  - Learners use more definite than indefinite articles
  - Both are underused
  - Interlanguage phenomena (e.g. Polish, Chinese and Japanese learners)

Methodology



Corpus design

- Participants: students of Spanish/FL (20-26 year old).
- **Low-intermediate level** (A2 and B1, *CEFR*).
- **N = 40**:
  - 9 groups of 4 students with the same L1:

Italian	English	Japanese
French	German	Chinese
Portuguese	Dutch	Polish

  - 1 heterogeneous group of 4 students with other L1s:

Korean	Finish
Turkish	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!**

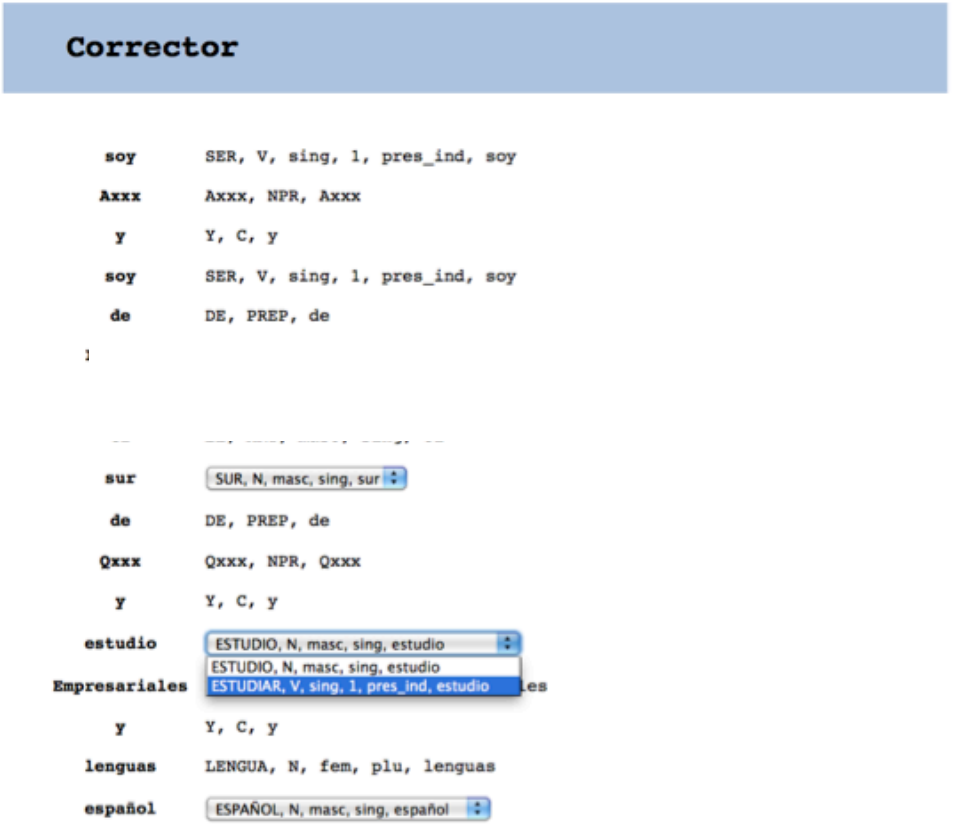
Spanish Learner Oral Corpus	Corpus Oral de Español como Lengua Extranjera (ELE)

PoS tagging

- Automatic tagging with GRAMPAL (Moreno Sandoval and Guirao, 2006).

  → 95.3% precision

- Semi-automatic revision (by the author)



PoS tagging

- Manual revision:
  - Correct **ambiguities**, **categories** and phrase **chunking**:

    e.g. *vino* (V ~ N) ('wine')

  - Fixed other aspects:
    - Add **unrecognised** categories/lemmas: e.g. *wasabi*
    - **Enrich** the annotation: e.g. code for foreign words
    - Fixing **multiword units**: e.g. *darse cuenta* ('to realize')

PoS tagging

- Annotation criteria:
  - Codes for **interlanguage phenomena**:
    - *ext:* borrowings: e.g. *fecha* (Port. 'cierra', 'to close')
    - *def:* wrong forms: e.g. **escribido* ('written')
  - **Morphological variants**: e.g. *superbien* (form) → *bien* (lemma)
  - **Ambiguous** tokens: *UNKN* lemma or category
  - **Multiword units**: nested entities

    e.g. *tiene que hervirlos* ('he/she has to boil them')

    (verbal periphrasis: auxiliary + conjunction + verb + clitic)
  - **Difficult** task, many **ambiguities → results** with **caution**

PoS tagging

- PoS tags:

Category		Examples
ADJ	Adjective	*blanco* ('white')
ADV	Adverb	*bien* ('well'), *mal* ('wrong')
ART	Article	*el* ('the'), *un* ('a')
AUX	Auxiliary verb	*haber* ('to have'), *ser* ('to be')
C	Conjunction	*pero* ('but'), *que* ('that')
DEM	Demonstrative adjective or pronoun	*este* ('this'), *aquel* ('that')
MD	Discourse marker	*entonces* ('so'), *quiero decir* ('I mean')
FORM	Formula	*lo siento* ('I am sorry'), *por favor* ('please')
INTJ	Interjection	*¡eh!* ('hey!')

PINT	Interrogative or exclamative pronoun		*qué* ('what'), *quién* ('who')	
N	Noun		*coche* ('car')	
NPR	Proper name		*Asturias, Universidad Autónoma*	
P	Personal pronoun		*tú* ('you'), *conmigo* ('with me')	
POSS	Possessive adjective or pronoun		*mi* ('my'), *tuyo* ('yours')	
PREP	Preposition		*a* ('to'), *por* ('for')	
Q	Quantifier	**NUM**	Numeral adjective or pronoun	*primer* ('first'), *dos* ('two')
		IND	Indefinite adjective or pronoun	*algunos* ('some')
REL	Relative pronoun		*que* ('which' or 'that'), *quien* ('who')	
UNKN	Unknown		*xxx*	
V	Verb		*amar* ('to love')	

PoS tagging

- Sample of annotated file (XML):

```
<w cat="C" lem="y" for="y" id="959">y</w>
<w cat="C" lem="y" for="y" id="960">y</w>
<w cat="MD" lem="pues" for="pues" id="961">pues</w>
<w cat="ART" lem="el" gen="masc" num="plu" for="los" id="962">los</w>
<w cat="N" lem="verdura" gen="fem" num="plu" for="verduras" id="963">verduras</w>
<w cat="V" lem="hervir" num="sing" per="3" tie="pres_ind" asp="modal" for="tiene que hervir" id="964">
<w cat="AUX" lem="tener" num="sing" per="3" tie="pres_ind" for="tiene">tiene</w>
<w cat="C" lem="que" for="que">que</w>
<w cat="V" for="hervirlos" pro="1">
hervirlos
<w cat="V" lem="hervir" tie="inf" for="hervir" />
<w cat="P" lem="lo" gen="masc" num="plu" per="3" for="los" id="965"/>
</w>
</w>
```

Category, lemma, form and id

Nested entity

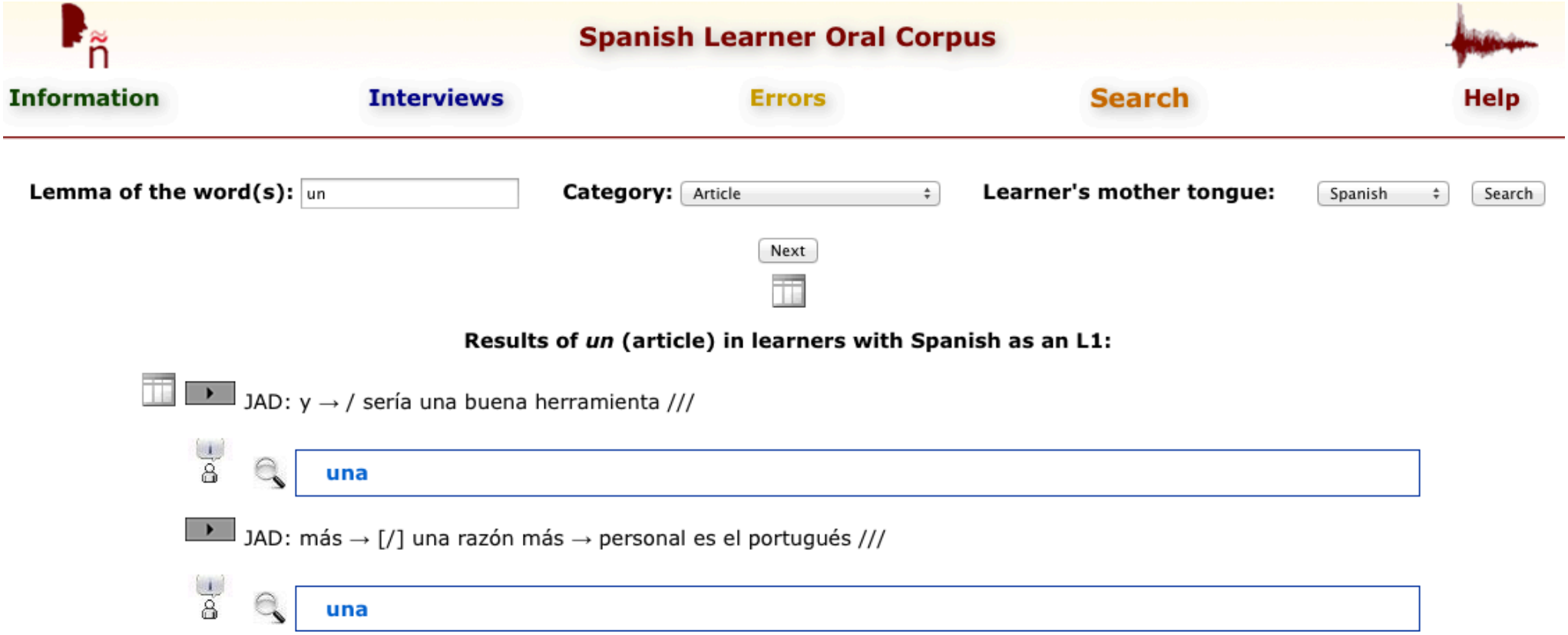
PoS tagging

- Unit of count: **lexical unit** (LU) → Fernández's criteria (1990):
  - **Single words**, except contractions: e.g. *del* (*de* + *el,* 'of the')
  - **Multiwords**:
    - idioms: e.g. *darse cuenta* ('to realize')
    - discourse markers: e.g. *es decir* ('I mean')
    - formulas: e.g. *de nada* ('you're welcomed')
    - proper nouns: e.g. *Puerto Rico*
    - foreign calques: e.g. **frutas de mar* ('marisco', Eng. 'seafood')
  - **Spoken phenomena:**
    - repetitions: 1 LU: e.g. *al [/] al principio* ('at [/] at first')
    - reformulations: 2 LUs: e.g. *es [/] está abierta* ('is [/] is open')

Analysis

http://cartago.lllf.uam.es/corele/search.html

Analysis



http://cartago.lllf.uam.es/corele/search.html

PoS tagging

- Learners' production (rate per 1000 LUs):

	PT	IT	FR	EN	DU	GE	PL	CH	JP	OT	TOT	SP
ADJ	41.1	35.9	36.2	28.3	37.3	37.2	47.7	46.6	47.0	44.5	40.2	34.7
ADV	142.3	135.5	170.4	187.4	152.9	196.3	166.7	186.6	180.7	186.5	170.5	108.2
ART	78.0	90.0	85.8	76.0	85.1	79.6	60.2	75.0	47.5	52.9	73.1	79.8
AUX	1.6	2.6	1.1	1.3	1.2	1.3	1.1	3.3	0.6	1.9	1.6	1.2
C	105.2	106.1	112.0	108.9	124.6	102.3	103.9	91.5	115.7	126.1	109.6	114.3
DE	13.2	11.4	5.6	4.9	8.8	9.4	12.8	6.1	9.4	7.8	8.9	11.3
DM	27.4	32.0	54.9	40.3	29.8	29.1	20.4	20.1	34.9	26.6	31.5	67.5
FOR	4.8	6.3	8.3	12.0	4.4	12.2	3.2	0.3	5.1	2.9	6.0	3.2
IND	36.6	34.7	23.4	32.1	38.6	22.7	40.9	32.1	26.9	33.6	32.2	33.9
INJ	9.7	2.4	9.5	20.9	10.4	17.7	6.6	15.3	18.0	9.4	12.0	3.7

PoS tagging

- Learners' production (rate per 1000 LUs):

	PT	IT	FR	EN	DU	GE	PL	CH	JP	OT	TOT	SP
INT	7.8	4.1	8.4	8.9	8.8	7.7	10.9	6.4	8.1	7.3	7.8	3.7
N	151.7	145.8	141.4	144.8	141.3	144.7	158.2	182.4	177.9	147.5	153.6	137.5
NM	9.7	12.3	15.1	8.7	13.8	13.1	9.8	15.3	14.5	10.7	12.3	10.5
POS	6.4	9.9	8.4	9.3	4.8	5.7	6.2	6.4	9.0	10.3	7.6	5.7
PRE	94.3	101.7	92.8	80.0	90.6	92.2	113.9	79.2	74.4	79.9	89.9	88.5
PRN	20.7	15.2	19.0	22.6	16.3	18.8	21.7	26.8	27.4	21.6	21.0	7.5
PRO	43.5	52.6	27.1	32.3	37.1	33.4	34.9	33.7	32.3	38.8	36.6	68.6
REL	12.8	20.9	15.3	7.0	15.2	5.0	8.3	4.2	1.5	5.9	9.6	24.2
UN	10.8	6.5	3.0	12.2	10.2	16.4	22.1	11.2	5.8	5.2	10.3	5.7
V	182.3	174.1	162.3	162.5	168.8	155.2	150.3	157.6	163.4	180.6	165.7	189.6

A sample analysis of articles in L2 Spanish - Motivation

- Dimensions of [±specific reference] and [±hearer knowledge]

→ use of article is related to the speakers' ability to identify the referent:

- **Definite article** (*el*): shared knowledge
- **Indefinite article** (*un*): new information

e.g. *He comprado, **un** coche, pero **el** vehículo no tiene airbag*

'I bought **a** car, but **the** vehicle does not have airbag'

→ **anaphoric relations**

- Other selection criteria:
  - General classes: e.g. *Los gatos cazan ratones* ('Cats chase mice')
  - Intensify statements: *e.g. ¡Tengo un hambre!* ('I am so hungry!')
  - Lexical restrictions: idioms, proper nouns…

See Leonetti (1999), Morimoto (2011), Brucart (2012)

A sample analysis of articles in L2 Spanish - Motivation

- Languages without articles:

  e.g. Polish, Japanese, Chinese, Finnish, Turkish, Korean…

  - Use demonstratives (definiteness)
  - Use numerals or quantifiers (indefiniteness)

   See Brucart (2012)

→ These learners face specific learning difficulties to use the article:

- Omission: e.g. *Pidió la cuenta ***a** camarero*
- Hypercorrection: e.g. *Está lleno de ***la** gente*

→ **Contrastive interlanguage analysis**

Results

- **Learners**:

  Mean (M) = 73.08 articles per 1,000 LUs (SD = 18.78)

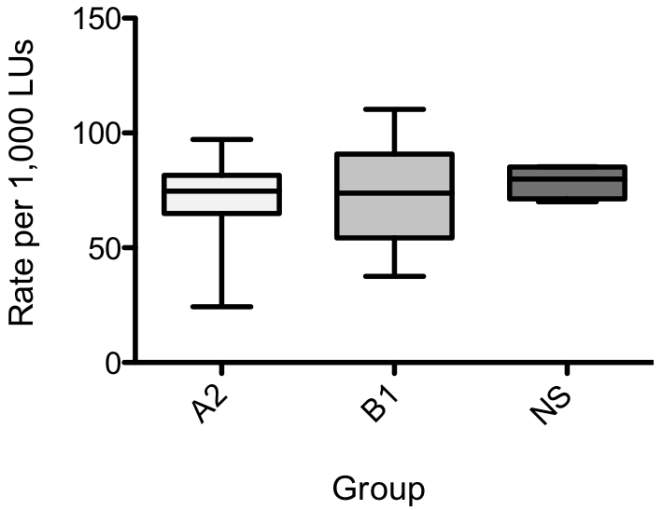
- **Native speakers** (NS):

  Mean (M) = 78.80 articles per 1,000 LUs (SD = 7.56)

Similar

- Differences between groups **not statistically significant**:

  Kruskal-Wallis $\chi 2$ (2, 44) = 0.44, p = 0.80



Results

- **Lower production** of indefinite article in both groups
  - → in line with previous analyses
- **Foreign articles (FA)** very frequent at A2
  - → especially, Portuguese learners (*a/o,* 'el'/'la')

	el	M	SD	*un*	M	SD	FA	M	SD	Total	M	SD
A2	1,291	49.16	12.02	566	21.40	6.67	64	1.79	5.74	1,921	72.34	16.39
B1	1,378	51.84	15.01	605	21.92	9.62	2	0.07	0.21	1,985	73.82	21.31
Total	2,669	50.50	13.49	1,171	21.66	8.17	66	0.93	4.10	3,906	73.08	18.78
NS	487	55.32	8.31	203	23.49	1.70	0	0	0	690	78.80	7.56

Articles produced by each group (NS vs. NNS)

Results

- **Higher SD** in **learners** usage (SD = 18.78)

  → some groups of learners produced fewer articles

  Not tested statistically (not enough data)

Group	*el*	M	SD	*un*	M	SD	FA	M	SD	Total	M	SD
PT	369	92.25	31.15	139	34.75	8.26	64	16.00	19.51	572	77.53	19.15
IT	361	90.25	49.63	194	48.50	27.01	0	0.00	0.00	555	86.74	17.92
FR	393	98.25	36.13	157	39.25	10.81	0	0.00	0.00	550	84.59	11.79
DU	280	70.00	13.44	162	40.50	8.39	1	0.25	0.50	443	85.21	3.95
EN	247	61.75	15.00	113	28.25	15.52	0	0.00	0.00	360	80.09	18.53
GE	247	61.75	16.94	117	29.25	7.50	0	0.00	0.00	364	78.57	11.22
PL	211	52.75	16.05	108	27.00	13.14	0	0.00	0.00	319	59.18	6.26
CH	207	51.75	11.27	62	15.50	4.65	0	0.00	0.00	269	76.38	17.76
JP	176	44.00	13.29	46	11.50	4.04	0	0.00	0.00	222	48.94	8.52
OT	178	44.50	17.08	73	18.25	9.95	1	0.25	0.50	252	53.59	23.16
Total	2,669	50.50	13.49	1,171	21.66	8.17	66	0.93	4.10	3,906	73.08	18.78

Results

- Two groups **clustering learners**:
  - [+article] L1: Germanic and Romance languages
  - [-article] L1: Polish, Chinese, Japanese, Finnish, Korean, Turkish

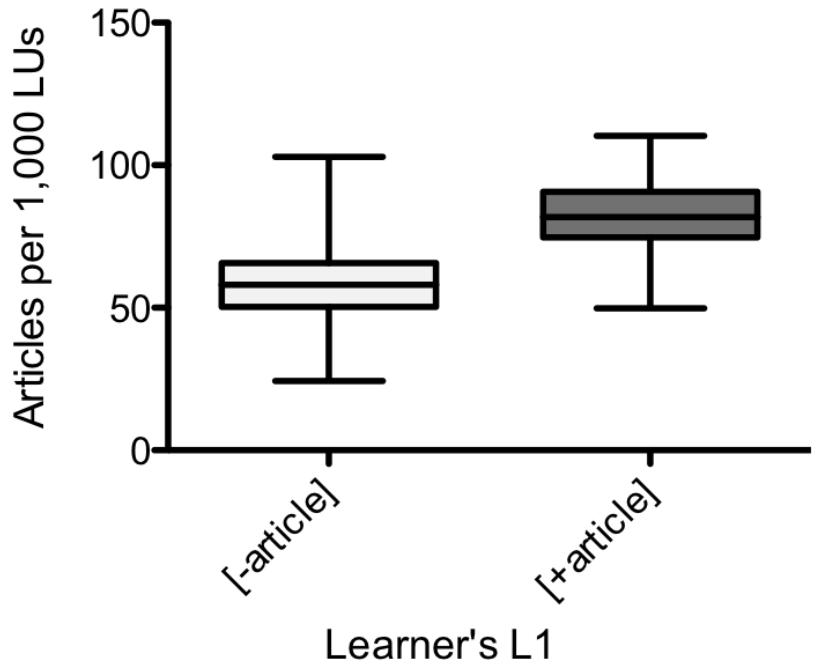
Group	*el*	M	SD	*un*	M	SD	FA	M	SD	Total	M	SD
[-Art]	714	42.90	13.68	257	15.15	7.00	1	0.06	0.21	972	58.10	17.17
[+Art]	1,955	55.06	11.35	914	25.56	6.13	65	1.45	5.16	2,934	82.08	13.28
Total	2,669	50.50	13.49	1,171	21.66	8.17	66	0.93	4.10	3,906	73.08	18.78

Results

- Results **statistically significant**:

  Mann-Whitney U = 44, p < 0.0001, ***

  → Negative influence of the lack of articles in learners' L1



Discussion

- Evidence that **learners whose L1 lacks articles underused** this category.
- **Pedagogical implications** for Chinese, Polish, Japanese, Finnish or Korean learners.

  → Also for learners whose L1 lacks article: Russian, Czech

- **Variation** in article usage according to **task** or **deictic context**
- **Lack of data per L1 to generalise the results**
- Not taking into account the **obligatory/non-obligatory uses** (**obligatory occasion analysis approach**; see Ellis & Barkhuizen 2005: 73ff)

Conclusions

- This corpus allows users to explore interlinguistic phenomena.
- Available online:

  http://cartago.lllf.uam.es/corele/search.html

- Future needs:
  - **Increase corpus data**.
  - Have **more annotators** tag the data and **calculate inter-annotator agreement**.

Thank you for your attention!

Contact

Leonardo Campillos Llanos:

leonardo.campillos@uam.es / leonardo.campillos@limsi.fr

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

This project was founded by the Madrid Region Government and the European Social Fund (PhD grant)

Comunidad de Madrid
CONSEJERÍA DE EDUCACIÓN

EUROPEAN SOCIAL FUND

References

Blanco Canales, A. “Corpus oral para el estudio de la adquisición y aprendizaje del componente fónico del español como lengua extranjera”. *Revista de Lingüística Teórica y Aplicada* 50 (2): 3-37.

Blanco Canales, A. 2014. “Adquisición y aprendizaje del componente fónico del español como lengua extranjera”. In *Fonética experimental, Educación Superior e Investigación*, Yolanda Congosto Martín, Mª Luisa Montero Curiel & Antonio Salvador Plans (eds), 179-198. Madrid: Arco-Libros.

Brucart, J. M. 2012. “La adquisición del artículo: flujo informativo y cohesión discursiva”. Presentation held at the *XI Encuentro de Profesores de ELE*, Barcelona, 21 December 2012. <www.encuentro-practico.com/pdf12/adquisicion_articulo.pdf> (11 July 2015)

Buyse, K., Fernández Pereda, L., González Ruiz, P. 2012. “The learner corpus Aprescrilov and its utility in the didactics of SFL in multilingual Belgium”*. Meertaligheid in België, Nederland en Europa.* Luik, Belgium, 14 December 2012

Council of Europe. 2001. *Common European Framework of Reference for Languages.* Cambridge: Cambridge University Press.

Dagneaux, E., Denness, S., Granger, S. 1998. “Computer-aided error analysis”. *System* 26(2): 163-174.

References

Díaz Rodríguez, L. 2007. *Interlengua española: estudio de casos.* Barcelona: Printulibro Intergrup .

Díaz-Negrillo, A., Meurers, D., Valera, S. & Wunsch, H. 2010. "Towards interlanguage POS annotation for effective learner corpora in SLA and FLT". *Language Forum* 36(1-2): 139–154. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair, Mª. Moreno Jaén & C. Pérez Basanta (eds).

Dickinson, M. & Raheb, M. 2011. "Dependency Annotation of Coordination for Learner Language". In *Proc. of the International Conference on Dependency Linguistics (Depling 2011)*, 5-7 September 2011, Barcelona (Spain), 135–144.

Díez-Bedmar, Mª. B. & Pérez Paredes, P. 2012. "A Cross sectional Analysis of the Use of the English Article System in Spanish Learner Writing". In *Developmental and Crosslinguistic Perspectives in Learner Corpus Research* [T.U.F.S. Studies in Linguistics 4], Y. Tono, Y. Kawaguchi & M. Minegishi (eds), 139–158. Amsterdam: John Benjamins.

Díez-Bedmar, Mª. B. & Papp, S. 2008. The use of the English article system by Chinese and Spanish learners. In *Linking up contrastive and learner corpus research*, G. Gilquin, S. Papp & Mª. B. Díez-Bedmar (eds), 147–175. Amsterdam/New York: Rodopi.

References

Ellis, R. & Barkhuizen, G. 2005. *Analysing learner language.* Oxford: Oxford University Press.

Fernández, S. 1990. *Análisis de errores e interlengua en el aprendizaje del español como lengua extranjera*. PhD dissertation. Universidad Complutense. Publicada como *Interlengua y análisis de errores en el aprendizaje del español como lengua extranjera.* 1997. Madrid: Edelsa.

Granger, S. 1996. "From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora". In Aijmer K., Altenberg B., Johansson M. (eds.) *Languages in Contrast. Text-based cross-linguistic studies*, pp. 37-51

Gaillat, T. 2013. "This and that in native and learner English: From typology of use to tagset characterisation"*. Twenty years of learner research: looking back, moving ahead*. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier (eds.) *Corpora and Language in Use,* vol. 1, Louvain-la-Neuve: Presses universitaires de Louvain, 2013. 167-177.

Gaillat, T., Sébillot, P. & Ballier, N. 2014. "Automated Classification of Unexpected Uses of This and That in a Learner Corpus of English". In L. Vandelanotte, K. Davidse, C. Gentens, & D. Kimps (ed.) *Recent Advances in Corpus Linguistics: Developing and Exploiting Corpora*. Amsterdam/New York: Rodopi. 309–24

References

Izumi, E., K. Uchimoto, and H. Isahara. 2004. “SST speech corpus of Japanese learners’ English and automatic detection of learners’ errors”. *ICAME Journal* 28: 31–48.

Leonetti, M. 1999. El artículo. In *Gramática Descriptiva de la Lengua Española*, I. Bosque & V. Demonte (eds), vol. 1, 787–890. Madrid: Espasa Calpe

Lin, T-J. 2005. *La adquisición y el uso del artículo por alumnos chinos*. PhD dissertation, Universidad de Alcalá.

Lu, H.-C. & Hsueh, L. L. 2012. “Estudio del uso del artículo a partir de un corpus paralelo de aprendices”, *CPATEI. Revista de Lingüística y Lenguas Aplicadas* 7: 193–202. DOI: 10.4995/rlyla.2012.1135

Lozano, C. and A. Mendikoetxea. 2013. “Learner corpora and second language acquisition: the design and collection of CEDEL2”. In A. Díaz-Negrillo, N. Ballier and P. Thompson (eds.) A*utomatic Treatment and Analysis of Learner Corpus Data*, pp. 65–100. Amsterdam: John Benjamins.

Meurers, D. 2015. “Learner Corpora and Natural Language Processing”. In S. Granger et al. (eds.) *The Cambridge Handbook of Learner Corpus Research*. Cambrisge: CUP.

References

Mitchell, R., Domínguez, L., Arche, M. J., Myles, F. & Marsden, E. 2008. "SPLLOC: A new database for Spanish second language acquisition research". In L. Roberts, F. Myles, & A. David (eds.) *EuroSLA Yearbook* 8, 287–304. Amsterdam: John Benjamins.

Morimoto, Y. 2011. *El artículo en español*. Madrid: Castalia.

Myles, F. 2005. "Interlanguage corpora and second language research". *Second Language Research* 21 (4): 373–391.

Ramírez-Mayberry, M. 1998. "The acquisition of the Spanish definite articles by English-speaking learners of Spanish". *Texas Papers on Foreign Language Education* 3(5): 1–57.

Rastelli, S. 2006. "ISA 0.9 – Written Italian of Americans: syntactic and semantic tagging of verbs in a learner corpus". *Studi Italiani di Linguistica Teorica e Applicata (SILTA)* 1: 73–100.

Rosen, A., Hana, J., Štindlová, B., Škodová, S. & Feldman, A. 2014. "Evaluating and automating the annotation of a learner corpus". *Language Resources and Evaluation* 48: 65–92.

Rosén, V. & De Smedt, K. 2010. "Syntactic annotation of learner corpora". In *Systematisk, variert, men ikke tilfeldig*, H. Johansen, A. Golden, J. E. Hagen & A.-K. Helland (eds), 120–132. Oslo: Novus forlag.

References

Said-Mohand, A. 2007. “La adquisición del artículo definido: evidencia oral y escrita”. *RedELE* 10: 1–15.

Santos, I. 1991. *La enseñanza de segundas lenguas. Análisis de errores en la expresión escrita de estudiantes de español cuya lengua nativa es el serbo-croata*. PhD dissertation. Madrid, Universidad Complutense.

Schachter, J. 1974. “An error in error analysis”. *Language Learning* 24: 205-214.

Tarrés, I. 2002. *El uso del artículo por estudiantes polacos de ELE*. Master’s dissertation, Universidad de Barcelona. <http://www.mecd.gob.es/redele/Biblioteca-Virtual/2005/memoriaMaster/2-Semestre/TARRES-C.html>

Tenfjord, K., Meurer, P. & Hofland, K. 2006. The ASK corpus: A language learner corpus of Norwegian as a second language. In *Proc. of the 5th International Language Resources and Evaluation Conference*, 22-28 May, Genova (Italy), 1821–1824.

Tono, Y. 2000. A corpus-based analysis of interlanguage development: analysing POS tag sequences of EFL learner corpora. In *Practical Applications in Language Corpora*, B. Lewandowska-Tomaszczyk & P. J. Melia (eds), 323–343.

References

Tono, Y. 2002. *The role of learner corpora in SLA research and foreign language teaching: the multiple comparison approach*. PhD dissertation, University of Lancaster.

Valverde, Mª. P. & Ohtani, A. 2014. Annotating article errors in Spanish learner texts: design and evaluation of an annotation scheme. In *Proc. of the 28th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 12-14 December 2014, Phuket (Thailand), W. Aroonmanakun, T. Supnithi & P. Boonkwan (eds), 234–243

Vázquez, G. 1991. *Análisis de errores y aprendizaje de español/lengua extranjera* [Studia Romanica et Linguistica 25]. Frankfurt: Peter Lang.