A XML-tagged Spanish Learner Oral Corpus for Learner Language Research

Leonardo Campillos Llanos
Computational Linguistics Laboratory - Universidad Autónoma de Madrid
leonardo.campillos@uam.es

BACKGROUND AND GOALS

- Few research projects of spoken learner corpus:

ENGLISH	FRENCH	SPANISH
LINDSEI [6]	FLLOC [12]	The Díaz Corpus [3]
NICT JLE [9]		SPLLOC [11]

- A new resource to perform Error Analysis of Spanish learner's speech.

METHODOLOGY

Interview and recording → Transcription → Conversion to XML → Tagging of errors → Error analysis

CORPUS DESIGN

- Cross-sectional: 40 students of more than 9 mother tongues (L1).
- University students: almost all between 19-25 years old.
- 4 learners from the A2 and the B1 levels (*Common European Framework of Reference for Languages*, henceforth *CEFR*) for every L1.
- More than 13 hours collected.
- Control group: 2 men and 2 women, native speakers, 26-27 years old.

	File	Sex	L1	Level	Length (mm : ss)	Length L1 group
Romance languages	PORMA2	M	Portuguese	A2	25:10	1:26:52
	PORWA2_1	W	Portuguese	A2	20:09	
	PORWA2_2	W	Portuguese (Brazilian)	A2	19:51	
	PORWB1	W	Portuguese (Brazilian)	B1	21:42	
	ITAMA2	M	Italian	A2	20:45	1:13:25
	ITAWA2	W	Italian	A2	13:09	
	ITAMB1	M	Italian	B1	23:16	
	ITAWB1	W	Italian	B1	16:15	
	FREMA2	M	French	A2	24:08	1:23:17
	FREWA2	W	French	A2	20:31	
	FREMB1	M	French	B1	21:56	
	FREWB1	W	French	B1	16:46	
Germanic languages	ENGWA2	W	English	A2	15:04	1:20:39
	ENGMB1	M	English	B1	18:44	
	ENGWB1_1	W	English	B1	18:02	
	ENGWB1_2	W	English	B1	28:49	
	DUTMA2	M	Dutch	A2	18:19	1:16:46
	DUTWA2_1	W	Dutch	A2	17:33	
	DUTWA2_2	W	Dutch	A2	23:05	
	DUTWB1	W	Dutch	B1	17:49	
	GERMA2	M	German	A2	18:23	1:13:24
	GERWA2	W	German	A2	19:45	
	GERWB1_1	W	German	B1	15:35	
	GERWB1_2	W	German	B1	19:41	
Slavic languages	POLMA2_1	M	Polish	A2	22:20	1:32:25
	POLMA2_2	M	Polish	A2	30:28	
	POLMB1	M	Polish	B1	26:46	
	POLWB1	W	Polish	B1	12:51	
Sino-Tibetan languages	CHIWA2_1	W	Chinese	A2	18:48	1:17:27
	CHIWA2_2	W	Chinese	A2	18:45	
	CHIMB1	M	Chinese	B1	18:56	
	CHIWB1	W	Chinese	B1	20:58	
Japanese languages	JAPWA2	W	Japanese	A2	28:52	1:32:41
	JAPWB1_1	W	Japanese	B1	16:28	
	JAPWB1_2	W	Japanese	B1	20:59	
	JAPWB1_3	W	Japanese	B1	26:22	
Other languages	FINWA2	W	Finnish	A2	20:27	1:19:05
	HUNWA2	W	Hungarian	A2	21:28	
	KORWB1	W	Korean	B1	21:14	
	TURWB1	W	Turkish	B1	15:56	

DATA COLLECTION METHOD

- Semi structured, spontaneous interview.
- Retelling task (using pictures) and description of photographs.
- Opinions on different topics (e.g. differences in food habits).

TRANSCRIPTION MARKS

- An adaptation of the CHAT format (as used in C-Oral-Rom [2]) and the conventions from SPLLOC [11].
- Speech phenomena (overlappings, pauses, repetitions...) are coded:

```
bueno &mm / lo → [/] lo mezclas ///
```

- Utterances synchronized with the corresponding fragment of sound.

ERROR TYPOLOGY

- Error typology based on studies for English [5, 7, 13] or Spanish [4, 14]
- Criteria for error classification:
  - Part of Speech (e.g. article) or Syntactic category (e.g. verb phrase).
  - Target modification:
    - ☐ Blend ☐ Misselection ☐ Unnecessary
    - ☐ Misformation ☐ Missing ☐ Wrong order
  - Linguistic level:
    - ☐ Grammar ☐ Pragmatics-discourse
    - ☐ Lexis-semantics ☐ Pronunciation
  - Type (e.g. conjugation, *ser/estar*, indicative/subjunctive, etc.).
  - Etiology: ☐ Interlinguistic ☐ Intralinguistic ☐ Unknown

XML ERROR TAGS

- Error tags were set up taking in consideration previous tagsets [10, 7].
- Non-ambiguous and ambiguous errors have been differentiated.

```
<unit id="12">
<start time="57.539" />
<end time="59.407" />
<speech speaker="LIU">
<line><ER id="1">estudiendo</ER> / <AE id="2">español {%pho: [ispa&apos;&#626;ol]}</AE> ///</line>
    <error_description id="1">
        <linguistic_level tag="grammar" />
        <category tag="V" />
        <target_modification tag="misformation" />
        <type tag="conjugation" />
        <etiology tag="intralinguistic" />
        <correction tag="estudiando" />
    </error_description>
    <ambiguous_error_description id="2">
        <interpretation n="1">
            <linguistic_level tag="pronunciation" />
            <category tag="SEG" />
            <target_modification tag="misselection" />
            <type tag="-" />
            <etiology tag="unknown" />
            <correction tag="e [e]: [espa&apos;&#626;ol]" />
        </interpretation>
        <interpretation n="2">
            <linguistic_level tag="lexis-semantics" />
            <category tag="N" />
            <target_modification tag="misformation" />
            <type tag="-" />
            <etiology tag="intralinguistic" />
            <correction tag="español" />
        </interpretation>
    </ambiguous_error_description>
</speech>
</unit>
```

WHAT ERRORS TO MARK IN SPEECH?

- Reformulated structures were considered *mistakes* [1]
  ⇒ not included in the error count.
- Not marking constructions which are common in spontaneous speech.

LEARNER'S METADATA

- The following information about the learner has been coded:
  - Personal information:
    - ☐ Age ☐ Role in the recording
    - ☐ Level of education ☐ Geographical origin
    - ☐ Profession / occupation
  - Linguistic background and details:
    - ☐ Level of Spanish (*CEFR*) ☐ Time studying Spanish
    - ☐ Mother tongue ☐ Time in Spanish-speaking country
    - ☐ Languages spoken

CONCLUSIONS AND FUTURE WORK

- Small corpus ⇒ difficult to generalize our results.
- The aims of this work are to design, set up and evaluate a methodology for spoken learner language research.
- Development of a web-based interface to train teachers of Spanish.

Aknowledgements

This project is funded by the Comunidad Autónoma de Madrid and the European Social Fund (ESF)

REFERENCES

[1] Corder, P. (1971) Idiosyncratic Dialects and Error Analysis. *International Review of Applied Linguistics*, 9(2): 147-60.
[2] Cresti, E. & Moneglia, M. (2005) *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam/Philadelphia: J. Benjamins.
[3] Díaz Rodríguez, L. (2007) *Interlengua española*. Barcelona: Printulibro.
[4] Fernández López, S. (1997) *Interlengua y Análisis de Errores en el aprendizaje del español como lengua extranjera*. Madrid: Edelsa.
[5] James, C. (1998) *Errors in Language Learning and Use*. London/N.Y.: Longman.
[6] Gilquin, G., De Cock, S. & Granger, S. (2010) *Louvain International Database of Spoken English Interlanguage*. Presses universitaires de Louvain, Louvain-la-Neuve
[7] Granger, S. (2003) Error-tagged Learner Corpora and CALL: a promising synergy, *CALICO Journal*, 20 (3), pp. 465-480 .
[8] Granger, S., Kraifa, O., Pontona, C., Antoniadisa, G. & V. Zampa (2007) Integrating learner corpora and natural language processing. *ReCALL Journal*, 19, pp. 252-268.
[9] Izumi, E. et al. 2004. "SST speech corpus of Japanese learners' English and automatic detection of learners' errors". *ICAME Journal* 28, pp. 31-48.
[10] Lüdeling, A., Walter, M., Kroymann, E. & Adolphs, P. (2005) Multi-level error annotation in learner corpora. *Proceedings of Corpus Linguistics Conference 2005*.
[11] Mitchell, R., Dominguez, L., Arche, M. J., Myles, F. & Marsden, E. (2008) SPLLOC: A new database for Spanish second language acquisition research. *EuroSLA Yearbook*, 8, 287-304.
[12] Myles, F. (2005) Interlanguage corpora and Second Language Research. *Second Language Research*, vol. 21(4), pp. 373-391
[13] Nicholls, D. (2003) The Cambridge Learner Corpus – error coding and analysis for Lexicography and ELT. In Archer *et al.* (eds.) *Proceedings of the Corpus Linguistics Conference 2003*, pp. 572-581.
[14] Vázquez, G. (1999) *¿Errores? ¡Sin falta!* Madrid: Edelsa.