Spontaneous Speech Corpora for language learners of Spanish, Chinese and Japanese

[*]Antonio Moreno-Sandoval, [*]Leonardo Campillos, [*#]Yang Dong, [*]Emi Takamori, [*+]José M. Guirao, [*]Paula Gozalo, [*]Chieko Kimura, [¢]Kengo Matsui, [*]Marta Garrote
[*]Computational Linguistics Lab – Autonomous University of Madrid
[#]Beijing International Studies University
[¢]Tokyo University of Foreign Studies
[+]University of Granada

antonio.msandoval@uam.es, leonardo.campillos@uam.es, sofiayy@hotmail.com, emi.takamori@uam.es, jmguirao@ugr.es, paula.gozalo@uam.es, chieko.kimura@uam.es matkenv@gmail.com, marta.garrote@uam.es

Abstract

This paper presents a method for designing, compiling and annotating corpora intended for language learners. In particular, we focus on spoken corpora for being used as complementary material in the classroom as well as in examinations. We describe the three corpora (Spanish, Chinese and Japanese) compiled by the Laboratorio de Lingüística Informática at the Autonomous University of Madrid (LLI-UAM).

.

Keywords: spoken corpora, learner corpora, Spanish, Chinese, Japanese

1. Language resources for language learners

The recent call for papers for the Special Issue on Resources and Tools for Language Learners, by the LRE Journal is an indication of the current interest in using corpora in language teaching. However for almost 20 years the TALC conferences have brought together practitioners and theorists with a common interest in the use of corpora for language teaching and learning (including data-driven learning materials, Johns 1991). In 2012 the 10th bi-annual Teaching and Language Corpora Conference will take place in Warsaw, Poland. This paper presents a method for designing, compiling and annotating corpora intended for language learners. In particular, we focus on spoken corpora to be used as complementary material in the classroom as well as in examinations.

The teaching of foreign languages requires knowledge of the language's real situation and the use of real materials. However, it is a common practice in the process of writing textbooks to choose language examples that fulfil the grammar rules, communicative functions, vocabulary, etc. of each lesson to facilitate learning. Considering both the advantages and disadvantages of using corpora in language learning (Sinclair 2004 and Gabrielatos 2005), we propose the use of spontaneous speech corpora to provide a source of real-life language examples and serve as a tool to access the reality of language in use. In addition, digitized data facilitate the analysis for the pedagogic adaptation of the corpus and to develop the necessary linguistic materials in teaching.

We first start with the adaptation of a general-purpose spontaneous speech corpus, the Spanish corpus of the C-ORAL-ROM project (Cresti & Moneglia 2005, Moreno et al. 2005). From this resource, 200 fragments of audio and transcription were selected, analysed and annotated for different learner levels, according to the Common European Framework of Reference for Languages. Then we follow with the design and compilation of two specialised corpora of Chinese and Japanese, based on the C-ORAL-ROM methodology but adapted to the features of language learners: selection of interesting topics, choice of young speakers, focus on the standard variety and good pronunciation. The Laboratorio de Lingüística Informática at the Autonomous University of Madrid (LLI-UAM) has compiled the three corpora, with support from the Beijing International Studies University (BISU) and the Tokyo University of Foreign Studies (TUFS).

2. The Spanish corpus

C-ORAL-ROM has two main drawbacks in being used directly in language teaching: some conversations are difficult to use in teaching/learning contexts and the computer tools included in the dvd published by John Benjamins are not user-friendly. Therefore, an adaptation was necessary to make it usable for teachers in the classroom or by students in self-learning contexts. The adaptation involved five stages:

1. Design of a specific syllabus for teaching Spanish as a foreign language. After documentation based on manuals of Spanish, a selection of structures was carried out at three levels: grammatical, communicative and lexical contents.
2. Selection of those examples most adequate for teaching Spanish from the Spanish subcorpus of C-ORAL-ROM, following a list of structures, with a concordance program developed by us.
3. Selection of fragments of files from C-ORAL-ROM and their classification according to the reference levels set by the *Common European Framework of Reference for Languages* (henceforth, *CEFR*). Several features were taken into account: the grammatical structures, communicative functions and lexical topics; the speed of speech; the register; the linguistic variety; and the diction clarity. In total, 200 texts between 1 to 3 minutes were selected.
4. Development of a tool to consult the selected texts and

its audio, following the design of a web-interface. In the adaptation, the metadata and the context for every text are explicitly shown, to help users to understand better.

5. Creation of exercises for each text in order to practice and evaluate listening comprehension skills.

The result is *Español Oral en Contexto* (Campillos et al 2010), complementary teaching material of Spanish, aimed at teachers or students of intermediate, advanced and proficiency levels (B1-B2 or C1-C2 according to the *CEFR*). Every text is provided with the transcription, the metadata and the audio (see Figure 1), as well as a series of exercises on the recording. As the adaptation process has already been explained in Campillos et al. (2007), we will briefly address these issues in the following.

2.1. Design of a specific syllabus for teaching Spanish as a foreign language

To design the syllabus, a descriptive analysis and communicative teaching approach was adopted, with contributions from Pragmatics. For the selection of the grammar contents (see appendix, Table 4), normative textbooks were not sufficient, being it necessary to look up grammars for foreign students (Cerrolaza: 2005). The choice and organisation of communicative contents (appendix, Table 5) was based on communicative grammars (Matte Bon, 2004) and indexes of functions (van Ek, 1975; van Ek et al, 1977; Gelabert et al, 1996).

2.2. Selection and extraction of samples from the Spanish C-ORAL-ROM

Two methodologies were followed to collect the samples:

- Extraction and search of a list of structures for each category by means of a concordance program.
- Reading and listening the whole documents.

Samples were selected following with these criteria:

- Linguistic criteria:
  - Grammar, communicative and lexical contents.
  - Linguistic variety: samples from a standard peninsular Spanish were preferred, but samples from American or Southern Spanish were also selected.
- Extralinguistic criteria:
  - Number of samples: up to 50 samples for each category, though it was surpassed for some of them
  - Acoustic quality: some structures were not collected if they were not heard properly, due to a lack of quality of the recording or speech overlappings.

The selection was manually performed and could not be automated by means of any program.

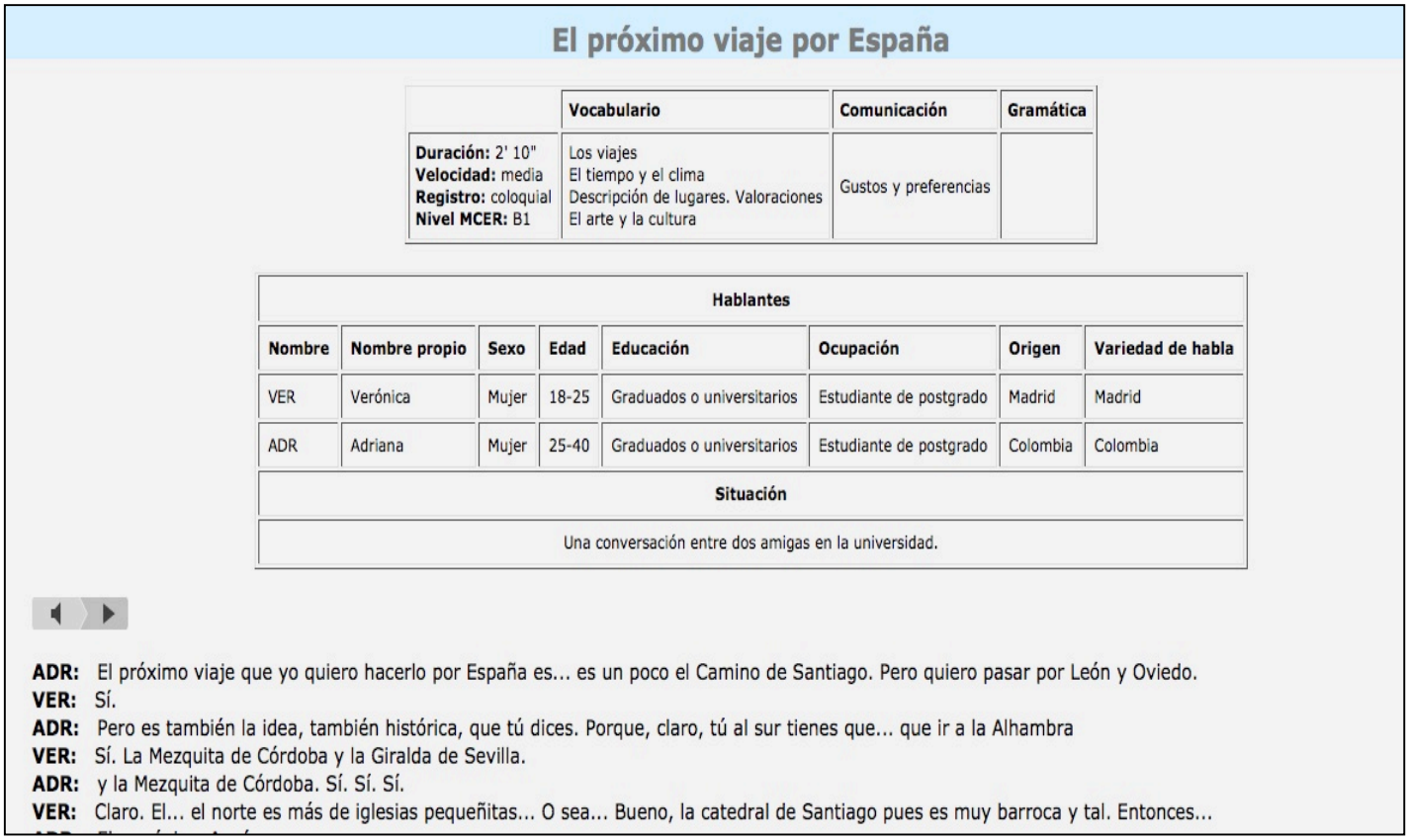

Figure 1: Screenshot of a Spanish text (including metadata)

2.3. Selection of the files from C-ORAL-ROM and classification in levels

Apart from the samples, we selected 92 texts (along with their aligned sound files) out of the 183 in the original Spanish subcorpus (which represents approximately 50%). These files were manually fragmented to make it easier its use, and we finally obtained 200 texts, each one having a total unity of sense. These fragments were classified according to the reference levels established in the *CEFR*, after considering the characteristics of each document:

- Grammar, communicative functions and lexical topics.
- Speed of speech and diction clarity
- Register
- Linguistic variety

As the graph shows, the B2 level predominates in our selection, followed by the C1 level. (see Figure 2 below)

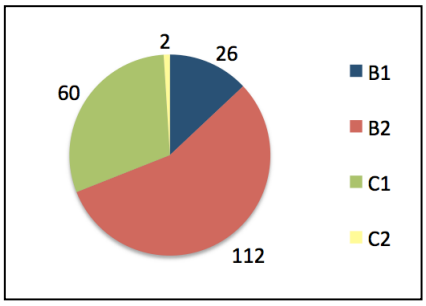

Figure 2: Number of fragments selected for every difficulty level (*CEFR*)

3. The Chinese corpus

C-ORAL-CHINA is a corpus oriented to Chinese learning for intermediate and advanced students. The corpus consists of 71 texts, 140,703 Chinese characters, and over 10 hours of recordings. The 47 participants, 33 female and 14 male, were asked to speak Putonghua (standard) Chinese, without dialectal forms. All of them gave their written consent for transcribing and publishing their speech. The recordings were collected on the campuses of the Beijing International Studies University, the Autonomous University of Madrid and the University of Alcala of Henares. All the participants are university workers or students, native speakers of Putonghua (Modern Standard Chinese, also known as Mandarin). The acoustic quality was an essential requirement, because language learners need quality input to avoid unnecessary complication in listening comprehension.

In order to focus on interesting topics for the students, the speakers were prompted to talk on selected topics including entertainment, art, culture, sport, travel, health, shopping, hobbies, cooking, pets, etc. Transcription in Chinese characters is provided, and a subset of 5 texts has also been transcribed into pinyin for intermediate students. The pinyin subset has been used for studying phoneme frequency. Table 1 shows the distribution of the text register in the Chinese corpus.

Table 2 shows the age distribution of participants. As we have volunteers mostly from university campuses and the media recordings belong to a Chinese university, the number of participants in age A (between 18-25 years old) is significant, at 74.5%. One advantage of this is that, as the speakers are mostly young, the vocabulary used is very contemporary, reflecting innovations in technology and society. We also assume that most of the potential users of this corpus are university students.

Subcorpora	**Length**	**Contents**	**Recording location**
Formal in public context	2 hours	Lectures on Chinese or Chinese language, etc.	China and Spain
Media (television/radio)	4 hours	Interviews, news reports, etc.	China
Informal	4 hours	Dialogues in public and private locations	China and Spain

Table 1: Register distribution in the Chinese corpus

Participants (47)											
Sex				Age							
M		F		A: 18-25		B: 26-40		C: 41-60		D: >60	
14	29.8%	33	70.2%	35	74.5 %	8	17 %	1	2.1 %	3	6.4 %

Table 2: Participant data in the Chinese corpus

Participants (59)											
Sex				Age							
M		F		A: 18-25		B: 26-40		C: 41-60		D: >60	
27	45,76%	32	54,23%	33	55,93%	8	13,55%	8	13,55%	10	16,94%

Table 3: Participant data in the Japanese corpus

4. The Japanese corpus

C-ORAL-JAPON has more than 12 hours of recordings, divided into three types of texts: monologues, dialogues and conversations. There is a total of 39 text files (transcription) aligned with their corresponding audio files. In total, 59 Japanese native speakers of different ages participated. The recordings were made mainly in Tokyo, with participants from Tokyo University of Foreign Studies (both students and professors), but some recordings were collected in Shizouka and in Madrid (lectures and conversations between colleagues).

Table 3 shows data regarding the speakers' age and sex. It reflects the absolute numbers (number of participants) and the relative figures.

Language in use varies as a social phenomenon. Sociolinguistic factors that cause language variation are the communicative situation and the participants. One characteristic of Japanese is the honorary form of address. Our intention is that this feature, so significant in the Japanese language, is reflected in our corpus. For this purpose, we obtained samples of the same participant in different communicative situations, where his/her position and attitude with respect to the hearer changes regarding respect and courtesy. All these factors are included in the metadata that precede each transcription. Figure 3 shows a fragment of a Japanese text.

4. Conclusion

This paper has presented the compilation of three corpora for language learners.
The goal is to provide a linguistic immersion environment for those learners that do not have direct access to spontaneous speech in Spanish, Chinese or Japanese.

The methodology is based on previous experience in compiling spoken corpora such as C-ORAL-ROM (Moreno et al 2005) and CHIEDE (Garrote & Moreno 2010), and adapted to Chinese and Japanese languages. The corpora are divided into three registers: informal, formal and the media.

The transcription is synchronised with the original sound, in a segmented way. The transcription also provides the metadata of the recording (sociolinguistic features of the participants, contextual information, duration, etc.). All the recordings have written consent for publication.

The content of a corpus has to conform to the purpose for which the corpus is built. Here, it is important that the topics are suitable for our teaching purposes. Spontaneous speech is framed by a set of spatial, cultural, social circumstances, etc. We intended that the topics be more related to everyday life, language and culture, etc. The content must be interesting and easy to understand, and words most frequently used in everyday life should be represented.

A web-based concordance tool has been used to search for examples in the corpus, and providing the text along with the corresponding audio. With this tool, relevant grammatical examples have been extracted and classified by learning level. In the last phase, we are developing teaching materials from the corpus, consisting the texts, the audio files and exercises on them. The Spanish materials have been published as a book plus CD-rom. At the moment, we are preparing the materials for Chinese in collaboration with the Chinese Department at BISU.

The Japanese and Chinese corpora can be accessed at the LLI-UAM website (http://www.lllf.uam.es). Both corpora are available for research and academic purposes.

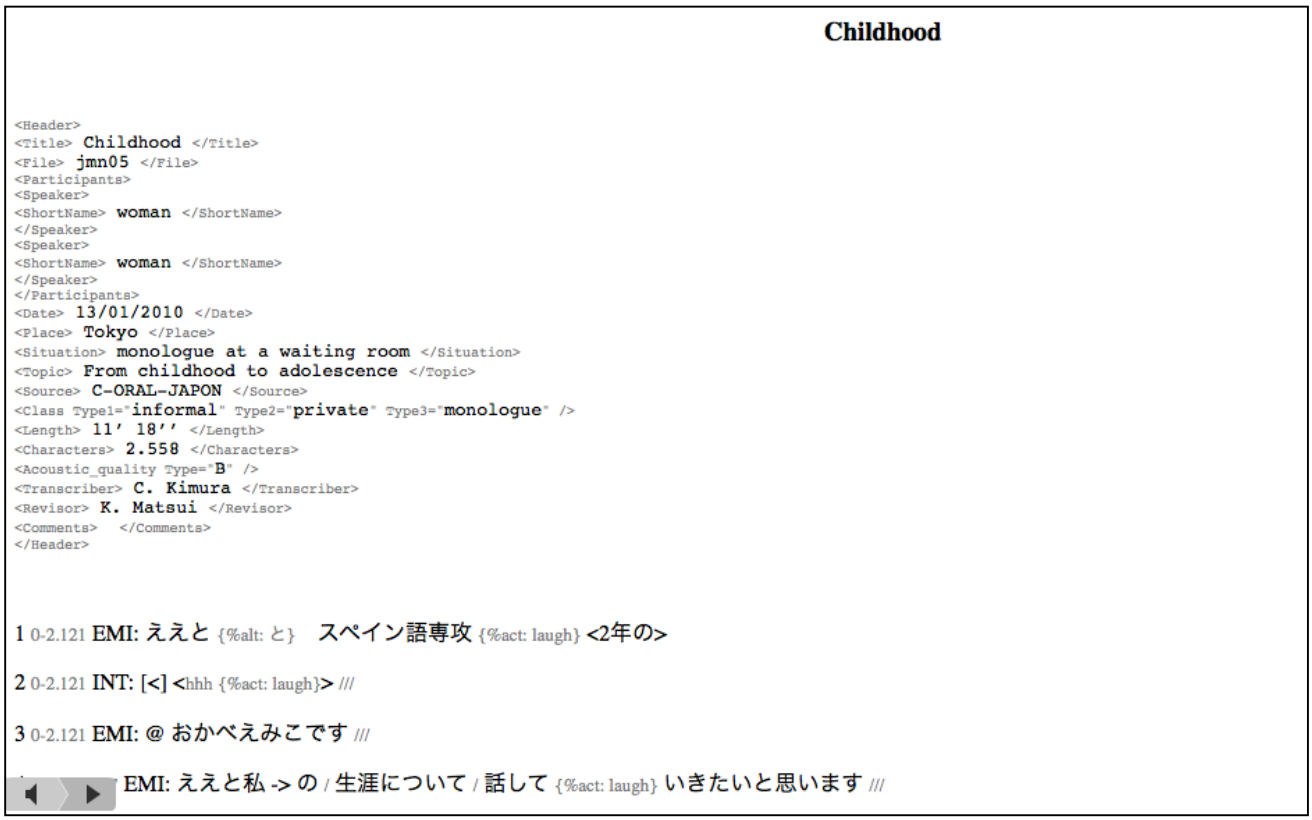

Figure 3: Screenshot of a Japanese fragment

5. Acknowledgements

This research has been funded by a grant from the Spanish Government (R&D National Plan, program TIN2010-20644-C03-03) and by a grant from UAM-Santander International Cooperation Program.

6. References

Campillos Llanos, L., Gozalo Gómez, P., Guirao Miras, J. Mª & Moreno Sandoval, A. (2007) Exploiting a spoken corpus in language teaching/learning: an advance web-based tool. *Proceedings of 4th Corpus Linguistics Conference*, University of Birmingham, 27-30 July 2007. http://ucrel.lancs.ac.uk/publications/CL2007/paper/99_Paper.pdf

Campillos Llanos, L., Gozalo Gómez, P., Guirao Miras, J. Mª & Moreno Sandoval, A. (2010). *Español oral en contexto. Vol. 1. Textos de español oral. Material de ELE basado en corpus. Comprensión auditiva*. Madrid: Universidad Autónoma de Madrid.

Cerrolaza Gili, O. (2005) Diccionario práctico de gramática. Madrid: Edelsa.

Cresti, E. & Moneglia, M. (eds) *C-ORAL-ROM. Integrated Reference Corpora for Spoken Romance Languages*. Studies in Corpus Linguistics, 15 (book + DVD). Amsterdam: John Benjamins, 2005.

Gabrielatos, C. (2005) Corpus and Language Teaching: Just a fling or wedding bells?, *Teaching Language as a Second or Foreign Language* (*TESL-EJ*), vol. 8, n. 4, A-1

Garrote, M. (2010). *Los corpus de habla infantil. Metodología y análisis*. Madrid: Servicio de publicaciones de la Universidad Autónoma de Madrid.

Gelabert, Mª. J.; Herrera, M.; Martinell, E.; Martinell, F. (1996) *Repertorio de funciones comunicativas del español: Niveles umbral, intermedio y avanzado.* Madrid: SGEL.

Johns, T. (1991). Should you be persuaded-Two samples of data-driven learning materials, in T. Johns and P. King (eds.) *Classroom concordancing,* ELR Journal, 4, 1-16

Matte Bon, F. (2004) *Gramática comunicativa del español*. 2 vols. Madrid: Difusión.

Moreno, A., De la Madrid, G., Alcántara, M., González, A., Guirao, J.M. & De la Torre, R. (2005). The Spanish corpus. In E. Cresti & M. Moneglia (eds.), pp. 135-162).

Sinclair, J. M. (2004). *How to use Corpora in Language Teaching*. Philadelphia: John Benjamins.

van Ek, J. A. (1975) The Threshold Level. Strasbourg: Council of Europe. (Version of 1980, Oxford: Pergamon Press).

van Ek, J. A.; Alexander, L. G.; Fitzpatrick, M. A. (1977) Waystage English. Oxford: Pergamon Press.

Appendix – List of Grammar structures and Communicative functions for the Spanish corpus

I – Word level 1. Articles 2. Nouns 3. Determiners and pronouns 3.1. Demonstratives 3.2. Possessives 3.3. Indefinites 3.4. Numerals 3.5. Interrogatives and exclamatives 3.6. Personal pronouns 3.7. Relative pronouns y relative adverbs 4. Adjectives 5. Verbs and tenses. 5.1. Indicative tenses 5.1.1. Present, future, preterite, imperfect, past perfect 5.1.2. Conditional 5.2. Subjunctive tenses 5.3. Imperative 5.4. Infinitive, gerund and past participle	5.5. Verbal periphrasis 5.6. Passive 5.7. *Ser / estar* 6. Markers of space and time 7. Adverbs and adverbial idioms 8. Prepositions and prepositional idioms II – Enunciation and clause level 1. Coordinate clauses 2. Subordinate clauses 2.1. Noun clauses 2.2. Adjective clauses 2.3. Adverb clauses 3. Discourse markers 4. Uses of the *se* pronoun III - Orthography Orthography, accent and acronyms

Table 4: Grammar structures selected for teaching/learning with the Spanish corpus

I –Notions 1. Being and existing 2. Quantity 3. Time 4. Location and spatial relations 5. Relations among events or processes: 5.1. Conditional relations 5.2. Concessive relations 5.3. Causal relations 5.4. Consequence relations 5.5. Final relations 6. Expressing manner, means and instrument 7. Expressing comparison 8. Expressing ownership 9. Expressing intensity and exclaiming 10. Impersonality II - Communicative functions 1. SOCIAL CUSTOMS 1. Greetings and goodbye. Introducing people 2. Invitations. Dates and appointments 3. Giving thanks 4. Apologizing 5. Social functions (e.g. congratulations) and courtesy 2. PHYSICAL AND EMOTIONAL STATES 2.1. Physical states 2.2. Feelings and emotions	3. ATTITUDES AND KNOWLEDGE 3.1. Possibility/impossibility and ability/inability 3.2. Certainty and probability 3.3. Expressing knowledge, memory and oblivion 3.4. Opinion 3.5. Agreement and disagreement 3.6. Obligation and necessity 3.7. Likes and preferences 3.8. Wishes 4. INFLUENCE 4.1. Advice, warnings and recommendations 4.2. Suggestions and proposals 4.3. Requests 4.4. Complaints and reclamations 4.5. Arguments, threats and insults 4.6. Encourage to act 4.7. Promises, commitments and oaths 4.8. Instructions, orders, interdictions. Allow and request permission 5. COMMUNICA TION 5.1. Oral communication skills 5.2. Organising information 5.3. Controlling the language 5.4. Reported speech

Table 5: Communicative notions and functions selected for teaching with the Spanish corpus