Monogràfic
«Spoken Corpus Linguistics in Romance: thoughts, design and results»
coordinat per Miquel Esplà-Gomis i Andreu Sentí

PRESENTATION*

Corpus linguistics is a consolidated area of study that, according to McEnery & Hardie (2012: 1), consists of «dealing with some set of machine-readable texts which is deemed an appropriate basis on which to study a specific set of research questions». It is a heterogeneous discipline whose objective is to study any aspects of a language by means of a corpus. Very different corpora are used in this field, and it is the objectives of the study that condition the features of the corpora to be used: spoken/written, monomodal/multimodal, monolingual/multilingual, diachronic/synchronic, annotated or otherwise, L1/L2 speakers, etc.

Spoken corpora are the basic resources used in corpus linguistics to study spoken language. Spoken language is a critical source of information for linguistics, as this mode of communication makes use of behavioral patterns that are substantially different to those used in written language (Carter & McCarthy 1997; McEnery & Hardie 2012: 4.6; Adolphs & Carter 2013). In fact, this is the topic dealt with by the first paper in this special issue, written by Voghera. It is not possible to achieve complete knowledge of natural language without studying and understanding speech. However, most of the efforts in the field of corpus linguistics that have appeared in the last decade have focused on studying written language. This has led to a growing number of corpora for this mode of communication in most languages around the world. Unfortunately, this is not the case for spoken language corpora and, according to Adolphs & Carter (2013: 1), «there are still relatively few projects devoted to spoken corpus linguistics».

The main reason for the under-representation of spoken language in the field of corpus linguistics is the intrinsic difficulty and, in some cases, higher cost of building

(*) This work has been supported by the Valencian Government through the project «Elaboració d'un corpus oral dialectal del valencià col·loquial (CorDiVal)» (Ref. GV/2017/094, Generalitat Valenciana).

such corpora (Newman 2008). In contrast to written corpora, the aim of spoken corpora is to contain monological or conversational samples of speech that, in most cases, have to be recorded from spontaneous interactions. The recording phase is one of the first stages in the creation of a spoken corpus, and may require having special equipment that, particularly when aiming to attain high quality records, can be expensive. Transcription is also a challenging and time-consuming task, and involves a series of methodological decisions, such as the type of transcription (orthographic, phonetic or prosodic) or a segmentation criterion that conditions the actual transcription process and its resulting quality. In fact, a number of authors in the bibliography report the inherent subjectivity of this process and the difficulty involved in defining a consistent and robust methodology (Cole *et al*. 1994; Raymond *et al*. 2002; Glenn *et al*. 2010). This special issue has, therefore, been conceived as a forum in which different authors can share their experiences of and insights into the most relevant aspects of spoken corpora building and their usefulness for corpus linguistics. In order to delimit the field of discussion, all the works compiled in this special issue focus on corpus linguistics for Romance languages.

Despite the fact that the development of spoken corpora has not been as fast as in the case of written corpora, it has gained momentum during the last few years (Wichmann 2008; Adolphs & Carter 2013; Sauer & Lüdeling 2016; Love *et al.* 2017). This growing interest in spoken language has also promoted and, simultaneously, taken advantage of the new technologies and tools available for the creation and exploitation of spoken corpora. On the one hand, there are tools that support the creation of spoken corpus linguistics, such as the well-known transcription software ELAN (Wittenburg *et al.* 2006), in addition to other more-comprehensive suites of tools that make it possible to not only transcribe, but also convert between different formats, edit media files or perform automatic and manual annotation. Two good examples in this category are the EXMARaLDA (Schmidt 2009) and *corpus-tools.org* (Druskat 2016) tool sets. The adoption of these tools has contributed to improving linguists' experiences when creating corpora, and has also facilitated the adoption of standardized formats and methodologies (Baude *et al*. 2006) that make modern corpora easier to maintain, extend and exploit. On the other hand, the fast evolution of automatic speech recognition (Yu & Deng 2016) and other sub-fields in natural language processing that enable the automatic annotation of corpora (Swift *et al.* 2004) have paved the way toward the use of automatic and semi-automatic techniques to ease the task of linguists in some phases of corpora building.

Several spoken corpora are available for Romance languages, particularly from the socio-linguistically strong languages, such as French, Italian, Spanish and Por-

tuguese (see for example Clarin - European Research Infrastructure for Language Resources and Technology). Two good models are, for example, available in Spanish: the dialectal *Corpus Oral y Sonoro del Español Rural* (COSER) and the conversational informal corpus by Val.Es.Co. Nevertheless, in the case of minoritized languages, such as Occitan, Aragonese, Asturleonese, Arpitan, Sardinian…, building spoken corpora is a challenge for the future.

The case of Catalan is in a middle position between both groups since there are, on the one hand, several spoken corpora, as summarized in the second paper in this special issue by Payrató and Nogué, but a spoken reference corpus does not yet exist. On the other, meanwhile, there are also several specific varieties of the language that have not been covered. For example, a conversational corpus with spontaneous interactions from speakers from all Catalan dialects is still lacking. The CorDiVal[1] project has, therefore, begun to build a spoken and informal corpus, that is, overall, conversational, into which a dialectal vision is incorporated: the *Parlars Corpus*. In the current stage, the corpus comprises only interactions from the Catalan Valencian dialect and from those elderly native speakers who have had little mobility throughout their lives.

Building a colloquial spoken corpus from a dialectal perspective is especially relevant in the case of minoritized languages that are losing speakers, and for which the interference of dominating languages is high, especially for two groups of speakers: younger generations and speakers living in urban areas. This is also crucial in cases such as Italo-Romance languages, for which some dialectal varieties have remained exclusively in spoken language for centuries, as pointed out by Voghera in the second paper.

This special issue on Spoken Corpus Linguistics (SCL) was partially motivated by the CorDiVal project and the *Parlars Corpus*. One of the outcomes of this publication is the interchange of ideas and experiences with other projects and researchers that have worked on the creation of spoken corpora. The contributions of other research teams as regards tasks such as the planning, design or annotation of spoken corpora will undoubtedly be a reference that will hopefully help and guide other projects, as is the case of the *Parlars Corpus*.

The first paper in the special issue is a general reflection on the specific contributions made by Spoken Corpus Linguistics (SCL) to linguistic knowledge. The author, Miriam Voghera (University of Salerno), has been the coordinator of the VoLIP corpus and other spoken corpora of Italian, which have been made available through the *Parlare Italiano* portal. Voghera claims that SCL is a useful tool with

1. More information about the project is available at: <https://www.uv.es/corvalc>.

which to study the general grammar in a unitary framework that fully integrates spoken language. This author first focuses on the approach to the sociolinguistics of spoken Italian: spoken corpora make it possible to study the presence of dialect forms in several contexts: different levels of spontaneity and formality. The second part, which is the longest, introduces a discussion regarding the contribution of SCL to the explanation of grammar. Gestuality and prosody are highlighted as playing a central role in communication, together with the interactive (and interpersonal) dimension and turn-taking. SCL has made relevant contributions to the characterization of spoken language, from both a qualitative and a quantitative point of view; as Voghera summarizes, constituents of a spoken language have the minimum level of specification for all the components: phonetic, syntax and semantic. Hypo-specification is, therefore, a fundamental feature in spoken grammar. Conversely, at a textual level, language is characterized by an increase in redundancy and the number of discursive markers. The paper provides relevant data regarding different types of phrases and part of speech that can be found in speech or written texts, both in dialogic interactions and monologues, or for the presence or absence of verbs in the sentences or fragments. Spoken corpora from different languages have equivalent results, which are genuinely revealing. The paper closes with some reflections on the concept of grammaticality from a global perspective of the study of the language, both spoken and written.

The second paper in this special issue is devoted to the spoken corpora in Catalan that have been created at the University of Barcelona, and its authors are Lluís Payrató and Neus Nogué. In the context of the Catalan language, the *Corpus del Català Contemporani de la Universitat de Barcelona (CCCUB)* is the most relevant and immediate precedent that serves as a reference for the *Parlars Corpus*. This paper introduces the CCCUB, a collection of corpora for the study of variations in Catalan, paying particular attention to those corpora whose purpose is the study of functional variation. The authors describe the objectives, structure and main features of three corpora: the *Corpus Oral Col·loquial (COC)*, which focuses on colloquial conversations between native speakers from the central-dialectal area of Catalan; the *Corpus Oral de Registres (COR)*, which consists of recordings of spoken language from different registers; and the *Corpus Audiovisual Plurilingüe (CAP)*, a multimodal corpus that allows the analysis of the connections between verbal and nonverbal communication. The paper additionally describes several techniques for elicitation, and particularly for semi-directed interviews, in several textual typologies. Speakers of Catalan and Spanish as language L1 and L2, and with English as a foreign language have been interviewed. Readers of this work will find a comprehensive characterization of these corpora, along with some insights into the creation of spoken corpora: transcription

and data management criteria, sociolinguistic variables, textual typologies, discursive genders, etc. In conclusion, this is a collection of valuable contributions that may provide relevant information regarding spoken language.

The third paper in the special issue is by Lucia de Almeida Ferrari and Giulia Bossaglia and describes the C-ORAL-BRASIL project, which compiled a set of Brazilian-Portuguese spoken corpora included in the C-ORAL corpus family. The paper starts by discussing some features and methodological requirements to take into account for the compilation of spoken corpora. After a review of the main spoken corpora in Brazilian Portuguese, the paper moves on to provide a comprehensive description of the C-ORAL-BRASIL corpora from several perspectives. The article introduces the project that produced the corpus, in addition to a comparison with other corpora in the C-ORAL family for a better contextualization. A detailed characterization of the corpora is provided, including insights into the internal organization of the team that developed and maintains them, the amount of data available, and the typology of the data compiled. According to the authors, there are four main corpora: formal natural context, informal natural context, telephonic and media. They provide detailed information on the amount of data available for each of these corpora and describe their subsections and contents. With regard to the creation of these resources, the authors share especially interesting insights regarding the methodological practices followed, mostly describing strategies for quality assurance that cover most phases of corpus building, from recording to annotation. In the last section, a set of minicorpora are presented: an innovative tool for studies focused on information structure and its interfaces. These microcorpora have been informationally tagged and allow crosslinguistic comparison.

This special issue closes with a paper written by Victoria Vázquez Rozas and Mario Barcala. In contrast to the previous papers in this special issue, that focus on describing linguistic aspects of different spoken corpora, this paper is rather aimed at sharing the experience of the authors in the creation of the ESLORA corpus, a spoken corpus of Spanish in Galicia, paying particular attention to the computational processes that supported this task. In this way, the ESLORA corpus is described but also used as a driving thread by the authors to share considerations and advice related to methodologies, formats and technologies with the reader.

The authors introduce their work by highlighting the relevance of accurately designing the methodologies to be followed during the creation of a spoken corpus. They emphasize the importance of doing so in the early stages of this process in order to avoid becoming constrained by limitations and wrong assumptions in later stages, when solutions are more expensive. The introduction is followed by a description of

the ESLORA corpus, including a motivation of the project of building this corpus, and an accurate description of the data and metadata included. After describing the corpus, the authors move on to a section in which they share with the reader their experience in the process of building spoken corpora, highlighting the challenges that they have confronted in the ESLORA project, and proposing strategies with which to tackle them. The authors cover several aspects of the creation of a corpus, such as planning, transcription, annotation, or format definition. The entire discussion is an extremely valuable resource for anyone planning to build a spoken corpus.

Miquel Esplà-Gomis
Universitat d'Alacant
mespla@dlsi.ua.es
ORCID 0000-0002-2682-066X

Andreu Sentí
Universitat de València
andreu.senti@uv.es
ORCID 0000-0002-4470-0469

BIBLIOGRAPHIC REFERENCES

Adolphs, S. & R. Carter (2013) *Spoken Corpus Linguistics: From monomodal to multimodal*, New York / Abingdon, Routledge.

Baude, O., C. Blanche-Benveniste, M. F. Calas, P. Cappeau, P. Cordereix, L. Goury, … & L. Mondada (2010) *Spoken Corpora Good Practice Guide 2006*. [Online: <https://www.researchgate.net/publication/280700484_Spoken_Corpora_Good_Practice_Guide_2006>.]

Carter, R. & M. McCarthy (1997) *Exploring Spoken English*, Cambridge, Cambridge University Press.

Cole, R., B. T. Oshika, M. Noel, T. Lander & M. Fanty (1994) «Labeler agreement in phonetic labeling of continuous speech», in *Third International Conference on Spoken Language Processing*.

Druskat, Stephan, Volker Gast, Thomas Krause & Zipser Florian (2016): «Corpus-tools.org: An Interoperable Generic Software Tool Set for Multi-layer Linguistic Corpora», in *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (*LREC*).

Glenn, M. L., S. M. Strassel, H. Lee, K. Maeda, R. Zakhary & X. Li (2010) «Transcription Methods for Consistency, Volume and Efficiency», in *LREC*.

McEnery, T. & A. Hardie (2012) *Corpus linguistics*, Cambridge, Cambridge University Press.

Love, R., C. Dembry, A. Hardie, V. Brezina & T. McEnery (2017) «The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations», *International Journal of Corpus Linguistics*, 22:3, p. 319-344.

Newman, J. (2008) «Spoken corpora: rationale and application», *Taiwan journal of linguistics*, 6.2.

Raymond, W. D., M. Pitt, K. Johnson, E. Hume, M. Makashay, R. Dautricourt & C. Hilts (2002) «An analysis of transcription consistency in spontaneous speech from the Buckeye corpus», in *Seventh International Conference on Spoken Language Processing*.

Sauer, S. & A. Lüdeling (2016) «Flexible multi-layer spoken dialogue corpora», *International Journal of Corpus Linguistics*, 21.3, p. 419-438.

Schmidt, T. (2009) «Creating and Working with Spoken Language Corpora in EXMARaLDA», in *LULCL II: Lesser Used Languages & Computer Linguistics*, II, p. 151-164.

Swift, M. D., M. O. Dzikovska, J. R. Tetreault & J. F. Allen (2004) «Semi-automatic Syntactic and Semantic Corpus Annotation with a Deep Parser», in *LREC*.

Wichmann, A. (2008) «Speech corpora and spoken corpora», in A. Lüdeling & M. Kytö (eds.) *Corpus linguistics. An International Handbook*, Berlin / New York, Walter de Gruyter.

Wittenburg, P., H. Brugman, A. Russel, A. Klassmann & H. Sloetjes (2006) «ELAN: a Professional Framework for Multimodality Research», in *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (*LREC* 2006).

Yu, D. & L. Deng (2016) *Automatic Speech Recognition*, Springer London Limited.