Information Design Journal 18(1), **69–73**

DOI: 10.1075/idj.18.1.08par

Giovanni Parodi

Research challenges for corpus cross-linguistics and multimodal texts

Introduction

In this article we argue that corpus linguistics is a powerful methodology that only recently has started to explore languages other than English, such as Spanish. At the same time, in developing automated tools to analyze Spanish and other languages researchers face some common challenges, even more so when the texts are multimodal in nature. Here we will explore key research problems in corpus linguistics for the Spanish language, identify emerging niches, and highlight issues in the automatic description of multimodal texts. We will, however, not move into the discussion about the status of corpus linguistics, the debate between corpus-based studies versus corpus-driven approaches (Tognini-Bonelli, 2001), the difference between light and strong corpus linguistics (Thompson & Hunston, 2006) or the distinctions between corpus linguistics research and discourse analysis (Biber, Connor, & Upton, 2007; Parodi, 2008). For a review of these distinctions, we refer to Parodi (2009).

In short, we will discuss two research challenges for cross-linguistic corpus analyses of multimodal texts. The first challenge concerns issues regarding non-English corpora, specifically Spanish. The second challenge concerns the overcoming of the monopoly of the verbal language by facing automatic analysis of multimodal texts.

Challenge 1: Corpus linguistic research on Spanish

Much of computational linguistic research is primarily concerned with the English language (see Jurafsky & Martin, 2001). Indeed, corpus linguistic studies using Spanish are not very common, not even in the Spanish scientific research community itself. Fortunately, there has been a growing interest in this area and the status quo is rapidly changing. Over the last decade, large and diversified corpora have been compiled and software has been developed to cover the needs of researchers working in Spanish (Briz & Grupo Val.Es.Co., 2002; De Kock, 2001; Moreno Fernandéz, 2006; Parodi, 2007; 2008; Pons & Ruiz, 2005; Venegas, 2008).

However, from a corpus linguistics approach, the limited attention for a language such as Spanish is surprising. Spanish has been rapidly growing as an international language, making the need for empirical studies of language use more urgent than ever. Admittedly, there are a significant number of studies concerning language description and variation in Spanish, but they tend to focus on examples taken from a small set of original corpora or are based only on made-up sentences. As for many languages, few studies on Spanish follow the principle of collecting and analyzing large and diversified corpora, covering not only register but also genre

and disciplinary variations. Almost no research describes contemporary Spanish in terms of language diversity and language unity, identifying major patterns of systematization and variation. For example, dictionaries have only recently given an account of dialectal variation and much work is still needed in this direction. What is more, no research team has undertaken the enterprise of producing a grammar of Spanish that identifies and describes similarities and differences of the kinds mentioned above. There is a strong tendency to appeal to a norm or standard Spanish and to overlook the variation across the many countries and populations that speak Spanish. It is true that from the Royal Academy of Spanish and the Association of Academies of the Spanish Language there has been a strong impulse to a compromise with a "unity in diversity." However, it is equally important to consider "diversity in unity." Fortunately, significant steps have been taken towards overcoming some of these problems with the production of grammars and dictionaries for Spanish (RAE, 2010; DUECH, 2010).

There are many opportunities for research on Spanish. For instance, researchers have free online access to the database of the Royal Academy of the Spanish Language (RAE), a query interface of concordances from two corpora, the Reference Corpus of Contemporary Spanish (CREA; 140 million forms) and the Diachronic Corpus of Spanish (CORDE; 180 million word forms) (http://www.rae.es/rae.html). More computational linguistic analytical tools are expected to be available online in the near future on this website.

Another example is the PRESEEA Project (*Proyecto para el Estudio Sociolingüístico del Español de España y de América*). This project aims at creating a corpus of spoken Spanish representing varieties of the world along geographical and social dimensions. The project is organized around research in parallel and coordination of researchers engaged in a common methodology for collecting a bank of materials that will enable its implementation consistent with educational and technological purposes. In this context, the project PRESEEA brings together a group of sociolinguistic research teams in different parts of the world (Moreno Fernández, 2006). It is worth noting that the material is compiled taking into account the sociolinguistic variety of Spanish-speaking communities.

Among several other groups, the Group Val.Es.Co. in Spain offers research opportunities for spoken register and colloquial conversational varieties (Briz & Grupo Val.Es.Co., 2002; Pons & Ruiz, 2005). Mention should also be made of the work of the research team from the University of Santiago de Compostela with a syntactic database of contemporary Spanish (www.bds.usc.es) and the Group of the Institute of Applied Linguistics at the Pompeu Fabra University (http://bwananet.iula.upf.edu). Another important contribution has been the computational resources developed by The Group for Data Structures and Computational Linguistics, Department of Information Technology and Systems, at University of Las Palmas de Gran Canaria, Spain. They have been working since 1986 on the analysis of data structures applied to the associative retrieval of information. Since 1990, the team has expanded its areas of interest to natural language processing and computational linguistics, developing tools for computational morphology, syntax, automated text analysis and lexicography (http://www.gedlc.ulpgc.es). These advances reveal that there are already a number of databases and resources for Spanish freely available on the Internet, created as institutional academic or personal initiatives. Some of these are reported in Instituto Cervantes (1996), De Kock (2001), and Parodi (2007).

One of the largest online Spanish databases and computational tools covering a variety of genres is the El Grial Project (www.elgrial.cl). A part-of-speech (POS) tagger, a syntactic parser and a lexical database can be freely used by researchers. In the website of

the project, electronic documents with more than 400 million words in texts, all lexicogrammatically tagged, are stored. Among the most recently collected corpora of this research project are the *Academic and Professional Corpora of Contemporary Written Spanish PUCV-2006* (Parodi 2008; 2009; 2010). These corpora comprise all the reading materials given to students from Psychology, Social Work, Industrial Chemistry, and Construction Engineering during each five-year program in university settings. The corpora exceed 80 million words, separated by disciplines and academic domains (social sciences and humanities and basic sciences and engineering), and are also classified into discourse genres. At the same time, the research team is now in the process of collecting the Corpus PUCV-2010, which will include the reading materials of doctoral students in Biotechnology, Chemistry, Physics, Linguistics, Literature, and History. In this corpus, efforts are being made to compare multimodal corpora (www.linguistica.cl).

The development of online computational tools for the study of Spanish has resulted in very similar problems and challenges to those for other languages. These include, for example, the problem of deciding which kind of grammatical principles or grammar should underlie the tagger and parser (e.g. generative, structural, or functional) and the corresponding problem of deciding on the level of description (e.g. morphological, syntactic, prosodic, pragmatic, textual, or discursive) and ensuring the availability of descriptive resources; the problem of reaching a high degree of automaticity with high precision, avoiding in this way the time-consuming and demanding work of manual revision; and the problem of having POS taggers and syntactic parsers that can be improved incrementally, which means widening the starting principles based on the corpora they process.

Challenge 2: Multimodal texts

In a cross-linguistic analysis of multimodal corpora an additional challenge emerges. Most of the available analytical computational tools are restricted to linguistic information (e.g. Graesser, McNamara & Louwerse, 2004; Louwerse & Jeuniaux, 2009). This means that figures, photographs, diagrams, formulas, just to mention some non-verbal elements, as well as their layouts, are not considered in corpus linguistic analysis, even though most genres in almost all scientific disciplines are involved with multimodal texts (Martin & Rose, 2008; Parodi, 2008; 2010). Multimodal texts have become an area of increasing interest (Kress & van Leeuwen, 1996; Martin & Rose, 2008), although many challenges are to be faced.

Multimodal annotated corpora require the development of sophisticated computational tools. Some of them should use machine-readable digital texts (tagged and annotated corpora). Thus, contemporary corpus linguistics might have to move towards a "multimodal corpus linguistics" in order to fully account for all the meaning-making resources involved in most texts, thus overcoming the monopoly of a radical focus on verbal or lexicogramatical feature analysis.

Some important advancements in research on multimodal texts have been made (Delin, Bateman, & Allen, 2002/3; Kong, 2006; O'Halloran, 2008). For example, in the Project "Genre and Multimodality: A computer model of genre in document layout" (GeM) (Delin, Bateman, & Allen, 2002/3), a multimodal view of genre was pursued with the objective of producing an annotation scheme for multilayered description of illustrated documents with complex layout. More precisely, in the GeM project, the researchers attempt to establish empirically the extent to which there is a systematic and regular relationship between some genres (e.g. instruction manuals, newspapers, illustrated books, web pages) and their

realizations in complex texts which include together verbal and visual formats such as diagrams, pictures and graphics. Also, in the Multimodal Analysis Lab at the University of Singapore (O'Halloran, 2008), a team of researchers from social sciences and computer sciences work together to develop prototype software for modeling, analyzing, storing and retrieving meaning from images, video texts and interactive digital sites constructed through the use of multiple semiotic resources (e.g. language, visual imagery, gesture, movement, music, sound, three-dimensional objects and so forth). These researchers are interdisciplinary and explore the complex dynamics of integral meaning-making practices (http://multimodal-analysis-lab.org/) (editors note: see the article by O'Halloran, Tan, Smith and Podlasov starting on p. 2 of this issue).

Mark-up languages such as SGML (Standard Generalized Markup Language) and XML (Extensible Markup Language) (Bryan, 1988; CES, 2000) are extremely valuable resources to automatically identify some of the semiotic features in multimodal text. These tools offer preliminary standards and frameworks for corpus annotation, but nowadays they do not guarantee a fully automatic process for analyzing visual artifacts with high precision and robust consistency and correctness. What is more, machine-readable digital multimodal automatic text identification lacks a robust theory of (multimodal) language in the framework of the so-called "visual turn."

Final remarks

The current paper discussed two research challenges, one related to cross-linguistic analyses, the other related to multimodal discourse. The first challenge, however, is not restricted to Spanish but applies to other languages of the world too. This cross-linguistic challenge is directly linked to the second challenge discussed here, that of the analysis of multimodal discourse. In a nutshell, in order for corpus linguistics to be ecologically valid, it should consider more than one language and more than one kind of discourse. It should consider cross-linguistic analysis of multimodal discourse. Addressing each of these challenges will make the future of research into document design more exciting than ever.

Acknowledgments

This research is funded by the FONDECYT Research Project 1090030.

References

Biber, D., Connor, U., & Upton, T. (2007). *Discourse on the move*. Amsterdam: John Benjamins.

Briz, A., & Grupo Val.Es.Co. (2002). *Corpus de conversaciones coloquiales*. Madrid: Arco.

Bryan, M. (1988). *SGML: An Author's guide to the standard generalized markup language*. New York: Addsson-Wesley.

CES (Corpus Encnding Standart. (2000). Corpus encoding standatd. Version 1.5. http//www.cs.vassar.edu/CES

De Kock, J. (ed.) (2001). *Lingüística con corpus: Catorce aplicaciones sobre el español*. Serie Gramática Española, 1. Apuntes Metodológicos, 7. Salamanca, España: Universidad de Salamanca

Delin, J., Bateman, J., & Allen, P. (2002/3). A model of genre in document layout. *Information Design Journal*, 11(1), 54-66.

DUECH (2010). *Diccionario de uso del español de Chile*. Santiago de Chile: Academia Chilena de la Lengua & Editorial MS.

Graesser, A., McNamara, D., & Louwerse, M. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, 36, 193–202.

Instituto Cervantes (1996). *Informe sobre recursos lingüísticos para el español. Corpus escritos y orales disponibles y en desarrollo en España.* (Vol. I y II). Alcalá de Henares: Instituto Cervantes.

Jurafsky, D., & Martin, J. (2001). *Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition*. New Jersey: Prentice Hall.

Kong, K. (2006). A taxonomy of the discourse relations between words and visuals. *Information Design Journal*, 14(3), 20–230.

Kress, G., & van Leeuwen, T. (1996). *A grammar of visual imagery*. London: Routledge.
Louwerse, M. M., & Jeuniaux, P. (2009). A computational psycholinguistic algorithm to measure cohesion in discourse. In J. Renkema (Ed.), *Discourse, of course* (pp. 213–226). Amsterdam: John Benjamins.
Martin, J., & Rose, D. (2008). *Genre relations: Mapping culture*. London: Equinox.
Moreno Fernández, F. (2006). Información básica sobre el Proyecto para el Estudio Sociolingüístico del Español de España y de América – PRESEEA (199–2010). *Revista Española de Lingüística,* XXVI, 12–126.
O'Halloran, K. (2008). Systemic functional-multimodal discourse analysis (SF-MDA): Constructing ideational meaning using language and visual imagery. *Visual Communication*, 7(4), 443–475.
Parodi, G. (ed.) (2007). *Working with Spanish corpora*. London: Continuum.
Parodi, G. (ed.) (2008). *Géneros académicos y géneros profesionales*. *Accesos discursivos para saber y hacer*. Valparaíso: EUV.
Parodi, G. (2009). *Lingüística de corpus. De la teoría a la empiria*. Frankfurt: Iberoamericana/Vervuert.
Parodi, G. (ed.) (2010). *Academic and professional discourse genres in Spanish*. Amsterdam: John Benjamins (in press).
Pons, S., & Ruiz, L. (2005). Corpus para el estudio de la conversación coloquial. El corpus Val.Es.Co. (Valenci. Español Coloquial), *Oralia*, 8, 243–263.
RAE (2010). *Nueva gramática de la lengua española*. Madrid: Espasa-Calpe.
Thompson, G., & Hunston, S. (2006). System and corpus: Two traditions with a common ground. In G. Thompson & S. Hunston (Eds.), *System and corpus: Exploring connections* (pp. 1–14). London: Equinox.
Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.
Venegas, R. (2008). Interfaz computacional de apoyo al análisis textual: "El Manchador de Textos". *Revista de Lingüística Teórica y Aplicada*, 46(2), 53–79.

Contact

Pontificia Universidad Católica de Valparaíso
Escuela Lingüística de Valparaíso
Av. Brasil 2830, 9th Floor, Valparaíso
Chile
gparodi@ucv.cl