Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 494 – 499

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Methodological Approach to the Design of Digital Discourse Corpora in Spanish. Proposal of the CÓDICE Project

Cristina Vela Delfa[a], Lucia Catamutto[b]*

[a]*Universidad de Valladolid, Campus María Zambrano. Plaza Alto Leones 1. 40005 Segovia*
[b]*Universidad Nacional del Sur- CONICET- 12 de octubre y San Juan. 8000 Bahía Blanca*

Abstract

Having analyzed the current situation of Spanish corpora—and the scarce representativeness of digital communication in them—and corpora from different types of interactions on digital platforms (e-mail, chats, SMSs), we noticed the need to create a repository of stable language samples aiming to solve this deficiency. Before implementing the CODICE, an open and collaborative repository of language samples from the digital discourse in Spanish, it becomes necessary to deal with the specific problems of compiling and transcribing this type of data. The present work addresses this approach aiming towards two goals: 1) establishing common standards, mainly concerning contextual and situational factors, in order to facilitate sociopragmatic analysis, and 2) developing ethical standards to ensure the anonymization of participants.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: corpus, digital discourse, Spanish, methodological aspects

1. Introduction

Certain properties of digital discourse need special consideration in the process of collecting and compiling this type of language samples. Among them, two stand out: multimodality and multisimultaneity. The nature of these data —their complexity and the difficulty to access them— compels researchers to recurrently dispute basic notions. Such is the case with the dichotomy between oral and written discourses and the question of where between them

* Corresponding author.
E-mail address: vela@fyl.uva.es

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.472

does digital discourse stand (Marcoccia & Gauducheau, 2007). The design of protocols for collecting digital data shows that, on the one hand, increasing multimodality detaches from the epistemology of written discourse and, on the other, poses profound challenges for conversational models (Llisterri, 1999, Briz, 2003). Ruling out oral and written-specific protocols, analysts have to reflect deeply on methodological aspects. The creation of a corpus is a challenge that many give up on and others face with a very poor methodological approach.

The field resents this situation and urgently calls for the proposal of methodologies specifically designed for digital data. In this sense, the CODICE project was elaborated in response to two needs: 1) creating a space for data collection and systematization, open to the general scientific community, and 2) proposing a collection protocol that responds to the complexity of such data. The second of these goals is an essential first step to address the first. Therefore, in these pages, we present a preliminary draft of the CODICE, through a methodological reflection on the process of collecting and compiling digital discourse data.

2. Dialogic digital discourse in Spanish corpora

We consider digital discourse that which is produced in contexts of technological mediation. This area covers many social needs and is therefore reflected in a range of different discourse genres (Swales, 1990). In particular, the present article focuses on the sub-genres belonging to the digital dialogic discourse, characterized by collaborative construction, changing voices through alternating turns and collective negotiation of meaning. This type of interactions poses additional difficulties to the process of data collection. Problems raised by the study of digital discourse must be added those arising from the participation of several interlocutors, which also falls within the sphere of private personal interactions (usually hard to access for researchers).

This situation translates into the absence of dialogic digital discourse samples in the general Spanish corpora. For example, they are not included in the CORPES (a general corpus recently compiled by the Real Academia Española). This deficiency can be justified from two complementary perspectives. It could be argued that the scientific community does not give due importance to the interactional phenomenon of digital communication. However, the increasing number of pragmalinguistic-oriented works focused on this discursive phenomenon refutes this idea. Therefore, as a plausible cause we could posit the difficulty of collecting and compiling digital discourse data, which would discourage their inclusion. Given this lack of samples in general corpora, researchers of digital Spanish discourse are forced to collect ad hoc samples for their particular works (Cantamutto & Vela, *forthcoming*).

The circumstances are considerably different in other languages, as spaces for massive collection of digital discourse data have been laid forward. For instance, in French a corpus of SMSs was compiled thanks to Sud4science project; for Chinese and English SMSs, NUSSMSCorpus is also available; regarding the study of e-mails in English, ApacheSpamAssassin was created. This effort on compiling data is also globally reflected in methodological considerations, and there are some extensive works on building digital communication corpora (Hundt, Nesselhaulf y Biewer, 2007; King, 2009; Tagg, 2009).

The situation of Spanish is particularly compromised, which is surprising, if we take into account the fact that, according to the Cervantes Institute, Spanish is the third most-used language on the net (Moreno Fernández, 2012). Without rapid intervention, enabling the systematization of digital discourse data in Spanish, we will miss the opportunity to document discourse genres in permanent evolution, which are essential to elaborate a XXI century history of the Spanish language. For example, SMSs once represented a genre of great influence but they fell into disuse, replaced by other discursive models, and it all happened without a compiling process of sufficiently representative samples. This situation highlights the urgent need to design stable collection protocols for digital discourse data, and that is the direction our proposal aims towards.

3. Properties of digital discourse language samples

As previously mentioned, much of the idiosyncrasies of digital discourse data can be explained from two angles: multimodality and multisimultaneity.

Multimodality refers to the property of integrating multisensory data present in certain discourses, exceeding what is traditionally considered as verbal language (Jewitt & Kess, 2001). With the evolution of interfaces, multimodality is increasingly present in the digital discourse. In a first step, Herring (1999, 2001) conceived and characterized digital discourse as text-only, i.e., as a strictly textual phenomenon, while in more recent work, she preferred to approach it from a multimodal perspective (Bourlai & Herring, 2014, Herring, 2015). Currently, she distinguishes platforms where the exchange is mainly textual from Interactive Multimodal Platforms (IMPs) which comprise multimodal content within a conversational thread and orient the exchange towards that combination. This aspect, essential to digital communication, poses a profound challenge both for the collection and the transcription of samples.

Multisimultaneity refers to the maintenance of co-occurring interactions, i.e., the fact that speakers are involved simultaneously in several parallel exchanges. This is an intrinsic condition of the digital medium, enabled by the confluence of several factors: 1) textual persistence, 2) the deferred nature of the exchanges, and 3) the lack of physical co-presence. In addition, it annuls discursive linearity, thus hindering data compilation.

These properties result in difficulties that necessarily raise several questions concerning the processes of data collection and compilation. The possibility of compiling parallel interactions, identifying disarrays in the (linear) order of exchanges produced through different devices, recovering visual representations of the devices used by each speaker, and retrieving the feedback incorporated into the interactions should be addressed. Moreover, questions arise concerning the possibility of reflecting multimodality in Computer-mediated communication (CMC) corpus, the creation of guidelines for collecting the paralinguistic component in CMC, among others. These are some of the specific problems faced by digital discourse analysts in the design of digital discourse corpora. Some possible solutions are proposed in the CODICE project.

4. Problems and solutions in the design of a digital discourse corpus: the proposal of the CODICE project

The particularities of digital discourse here described are correlated, as we have seen, with a number of difficulties raised by the tasks of collecting and compiling data. In order to systematize their complexities, we have organized them around two spheres: 1) the limitations in the collection stage, and 2) the ethics and representativeness issues posed by the data.

4.1. Problems in the stage of sample collection

In general terms, in any process of data collection, it is necessary to address the constraints posed by the type of support mediating the exchange, since they condition the message and define certain aspects of the context as well as the interaction rituals (Gobato, 2014). Each communicative context entails certain enunciation conditions: the type of support affects the time management of the exchange and creates diffuse spaces between synchrony and asynchrony, more or less marked towards either end depending on the specific medium. This should be considered in the methodology used for establishing the boundaries of the macro and micro segmentation units. Thus, with instant messaging conversations, it is not always easy to distinguish the boundaries between units, since the criterion of temporal continuity does not operate as in a co-presence context. Therefore, as a factor within the contextual issues, the support mediating the communication has specific characteristics affecting the actants' interventions and therefore the language samples collected by the researcher. In each interaction, the characteristics of the platform impose conditions on the message, both from a paratextual perspective and from the type and amount of data that can be sent.

Unlike face-to-face communication, where it is relatively easy to record the participation of both interactants at the moment of production, digital communication does not occur in a common physical or temporal space. Exchanges occur in parallel, simultaneously or sometimes overlapping, and not necessarily through the same application. As a consequence, the representation of the linguistic signs and the interactional structure is often not equivalent among the different interfaces (interface of reception vs. interface of production), thus breaking linearity criteria affecting, for example, the procedures of discursive cohesion and the order of interventions. However, users are usually able to replenish each intervention as a reply to a previous pair. These nuances are not obvious to the eye of the researcher, who generally accesses communications through the messages produced and received by one of the interactants.

What happens in an interaction where the exchanges occur in a transition over time without preset continuity intervals? Faced with this challenge, researchers must reformulate old units (Goffman, 1959) to adapt to new conditions. Vela-Delfa (2006) proposed the term interactional history to overcome the obstacles posed by determining the maximum unity, which affects analysis and collection. Alcántara-Pla (20014:231) uses the term conversation for the units comprised within sequences with explicit opening and closing, which, as he acknowledges, are very scarce, and proposes the term relation to refer to “the bond that people create through a particular computer application so that they can communicate (through conversations in the case of instant messages) whenever they want”.

These issues also affect the categories used in the analysis. The displacement of linguistic-oriented perspectives towards semiotic-oriented (Kress, 2033:35-36) approaches in the understanding of the communicative practices in digital contexts has met frictions over the use of the theoretical categories proposed for the study of face-to-face conversation. In this regard, the prototypical units of analysis of Conversation Analysis (Briz, 2003) should be rethought on account of the segmentation issues raised by digital communication. Accordingly, some studies have critically addressed the persisting problems in using face-to-face interaction as paradigmatic. For example, in the case of instant messaging or chat, Vela and Jiménez (2011) identify the constant overlapping of interventions as a structuring property of this type of interaction. Among other issues, this leads to “a structure of chaining adjacency pairs unresponsive to an external linear model” (íbid, 134). Something similar happens in the analysis of the conversational units of Whatsapp (Álcantara Plá, 2014). In short, if the question is how to delimit the dialogic units, there is a need to comprehensively understand the context of enunciation (Cantamutto, 2013); not only because the user expands its communication channels but because interactions are not necessarily linear.

In order to address some of these issues, during the process of data collection participants should be requested to elicit shared pragmatic information, which is the base of digital communication. When possible, the interaction should be transcribed from all the devices involved, in order to confront how each of the speakers receives it. In particular: the morphology of the linguistic signs and the establishment of turns and adjacency pairs. Ideally, collaborators should provide contextual information that helps to understand the exchange.

4.2. Sample design: ethical issues and representativeness

Finally, further conditioning factors derive from data representativeness and their ethical conditions. One problem is the possibility to meet all the sociolinguistic variables: mainly those comprised by lower socio-educational levels (Cantamutto, 2014). As often happens in the methodological discussion on digital communication language sampling, researchers face permanent restrictions causing endless debate. The difficulty in accessing data leads most researchers to work with social networks of family and friends (Tagg, 2009) or through introspection techniques, if the use of data from interactions held by the investigator in the study platform (Noblía, 2009). Thus, not having other interactants’ profiles (those not belonging to high or university education levels) can lead to erroneous observations, since the richest items of analysis lie in intragroup variation.

In relation to sampling ethics, it is essential to respect privacy and anonymity. A crosscutting question emerging from digital communication data derives from their public or private nature. This aspect is critical in most of the interaction platforms as they are, precisely, a medium for sensitive exchanges in the private sphere. This question seems to have its counterpart in face-to-face communication, since most of the discourses studied belong to institutional, educational or trade-related contexts. In contrast, as digital communication provides channels and amply satisfies the most intimate communication needs of the actants, investigations have tended toward the domains of language use in the private sphere. Paradoxically, a multiplicity of data seems to be available to the researcher as the nature of the platforms entails that all messages sent through them are public and have some degree of eternal permanence in the sites (in this regard, see further discussions around the "right to be forgotten" demanded by users). Both issues should be carefully observed.

In this sense, de-Matteis *(forthcoming*) proposes to address two different types of data: those available and those elicited. The first, despite being seemingly available to anyone, should not be minimized in terms of their potential repercussions on the speaker, even more so in Sociocultural Studies and Critical Discourse Analysis. The latter, those which the investigator requests, or submits, should be sensitive to the possibility of identifying who produced them. It is even more challenging to identify the textual persistence of digital interactions, with its counterpart in work based on large databases, through which it is possible to mask particular productions and thus avoid identifying the sender.

In this context, in order to discard ostensibly unstructured or systematic data and prevent constant justification due to the potential methodological deficiencies posed by samples (and consequently, conclusions), the proposal of developing a repository that gathers all these perfectible language samples will rectify the various deficiencies thanks to the confluence of fragmentary data. At the same time, the inclusion of data that researchers can collect on their own interactions will enrich the debate from a variationist and social pragmatics perspective. Again, the accumulation of data from different interactants (expert participants who can provide more metadata than voluntary collaborators) contributes to dissipating the ethnocentric issues that may arise from an introspective investigation.

5. Discussion and conclusions

What alternatives does the CODICE project propose to solve the obstacles mentioned? After analyzing the singular nature of digital discourse, we offer a sampling protocol with the following guidelines:

- The transcription template should include a space to indicate the device and the platform through which the interactions occurred, specifying the situation of each of the interactants.
- An iconic description of the application should be included in order to have paratextual data. This information may be submitted through the attachments of screenshots. If there is an intention to public this information (to include it in the repository), it is necessary to apply the same anonymity criteria as in written text, by blurring certain areas of the pictures, for example.
- The process comprises two levels. First, plaintext: transcribing or transferring the linguistic signs to a processor which enables subsequent analysis and HTML tagging. Second, multimedia information: consigning screenshots, audios, videos and pictures that may accompany textuality. Textual data will be mandatory, but multimodal data will be included as long as it is possible to have access to them or if they are submitted.
- Samples have to blur the details denoting the identification of informants at both levels.
- The researcher should aim towards sources of information that preserve feedback, when the developed application enables it.
- Access to these data is possible only for those who enter through an institutional account from a research centre, university or recognized entity.

On the basis of these items, we aim to provide digital discourse analysts with effective alternatives for the design of languages samples. To the extent that these corpora can be standardized and integrated into a common repository,

they will contribute to compile a parcel within discourse, the digital parcel, which is fundamental for establishing the last decades' language history. It shall thus benefit both synchronic and diachronic studies.

References

Alcántara Plá, M. (2014). Las unidades discursivas en los mensajes instantáneos de wasap. *Estudios de Lingüística del Español*, 35, 223-242

Boulari, E., & Herring, S. C. 2014. Multimodal communication on Tumblr: "I have so many feels!". En *Proceedings of WebSci'14*, June 23–26, Bloomington: IN (also in http://ella.slis.indiana.edu/~herring/tumblr.pdf)

Briz, A. (2003). Un sistema de unidades para el estudio del lenguaje coloquial, *Oralia*, 6, 7-61.

Cantamutto, L. (2013). La recursividad de las interacciones contemporáneas. Límites teórico-metodológicos del estudio de los SMS como conversación, *Revista de Ciencias Sociales de la Universidad Nacional de Quilmes "Al abordaje de la comunicación contemporánea. Sociedad y cultura en los mundos de la mediación digital"*, 23: 83-104.

Cantamutto, L. (2014). El discurso de los mensajes de texto en el habla adolescente del español bonaerense. *Lenguaje, discurso e interacción en los espacios virtuales*, 65-82. Parini, Alejandro y Giammatteo, M. (comp.). Mendoza: FFyL-UNCuyo-SA

Cantamutto, L. y Vela Delfa, C. (forthcoming), Repositorio colaborativo de comunicaciones digitales: aproximación a un corpus para el español, *I Jornadas Nacionales de Humanidades Digitales. Buenos Aires*: AAHD FyL-UBA

De-Matteis, L. (forthcoming). Ejes para una discusión del uso ético de datos interaccionales escritos y orales obtenidos en línea, *I Jornadas Nacionales de Humanidades Digitales. Buenos Aires*: AAHD-FyL-UBA

Gobato, F. (2014). *La escritura secundaria. Oralidad, grafía y digitalización en la interacción contemporánea*. Bernal: Universidad Nacional de Quilmes.

Goffman, E. (1959). *Presentation of Self in Everyday Life*. New York: Anchor.

Herring, S. C. (1999). Interactional Coherence in CMC, *Journal of Computer-Mediated Communication*, 4, 4. http://www.ascusc.org/jcmc/vol4/issue4/herring.html.

Herring, S. C. (forthcoming), 2015. "New frontiers in interactive multimodal communication". In A. Georgopoulou & T. Spilloti (Eds.), *The Routledge handbook of language and digital communication*. London: Routledge

Hundt, N. Nesselhauf & Biewer, C. (eds.) (2007). *Corpus Linguistics and the Web*, Amsterdam: Rodopi.

King, B. (2009). Building and Analysing Corpora of Computer-Mediated Communication. In *Contemporary Corpus Linguistics*, P. Baker (ed.). London: Continuum International Publishing Group, 301–20.

Kress, G. (2003). *Literacy in the new media age.* New York: Routledge.

Marcoccia, M., Gauducheau, N. (2007). L'analyse du rôle des smileys en production et en réception : un retour sur la question de l'oralité des écrits numériques, *Glottopol*, 10, 38-55. http://www.univ-rouen.fr/dyalang/glottopol/numero_10.html

Llisterri, J.(1999) Transcripción, etiquetado y codificación de corpus orales, in Panorama de la investigación en lingüística informática. RESLA, *Revista Española de Lingüística Aplicada*, Volumen monográfico. 53-82.

Moreno Fernández, F. (coord.) (2012), *El español, una lengua viva. Informe 2012*. España: Instituto Cervantes.

Noblía, V. (2009). Modalidad, evaluación e identidad en el chat. *Discurso & Sociedad*, 3(4), 738-768.

Swales, J. M. (1990). *Genre Analysis*, Cambridge: Cambridge University Press

Tagg, C. (2009). *A Corpus Linguistics Study of SMS Text Messaging*. UK: University of Birmingham.

Vela Delfa, C. (2006), *El correo electrónico el nacimiento de un nuevo género*. Madrid: Universidad Complutense de Madrid.

Vela Delfa, C., & Jiménez Gómez, J. (2011). El sistema de alternancia de turnos en los intercambios sincrónicos mediatizados por ordenador" *Pragmalingüística*, 0(19), 121-138.