How Software Features and Linguistic Analyses Add Value to Orthographic Markup in Transcriptions of Multilingual Recordings for Digital Archives

Enrique Rodríguez
Department of Spanish & Portuguese
Indiana University
Bloomington, USA
enrodri@iu.edu

Robert E. Vann
Department of Spanish
Western Michigan University
Kalamazoo, USA
robert.vann@wmich.edu

Abstract— **This report discusses the importance of accounting for language contact and discourse circumstance in orthographic transcriptions of multilingual recordings of spoken language for deposit in digital language archives (DLAs). Our account provides a linguistically informed approach to the multilingual representation of spontaneous speech patterns, taking steps toward documenting ancestral and emergent codes. Our findings lead to portable lessons learned including (a) the conclusion that transcriptions can benefit from a bottom-up approach targeting particular linguistic features of sociocultural relevance to the community documented and (b) the implication (for researchers developing transcriptions for other DLAs) that the principled implementation of particular software features in tandem with systematic linguistic analysis can be helpful in finding and classifying such features, especially in multilingual recordings.**

Keywords—language documentation, digital language archives, spoken language corpora, orthographic transcription, multilingualism, linguistic analysis, transcodic markers, discourse patterns, emergent codes, Spanish, Catalan, DARDOSIPCAT

I. Introduction and Research Questions

This report focuses on accounting for language contact and discourse circumstance in markup of orthographic transcriptions of multilingual recordings of spoken language (henceforth MRSL) for subsequent deposit in digital language archives (henceforth DLA). We address two research questions. First, what linguistic phenomena should such orthographic transcriptions account for in terms of language contact and discourse circumstance? Second, what sort of linguistic analyses may aid in classifying such linguistic phenomena?

In exploring these research questions, we discuss the need to account for words that were unambiguously spoken in languages other than the base-language of the transcription as well as words that were spoken transcodicly [1] in some way. Moreover, we also discuss the need to account for basic discourse phenomena such as overlapped words and interrupted words and turns. Subsequently, we report on useful software features and linguistic analyses that can help in the accurate representation of such linguistic phenomena. Specifically, we discuss strategic use of standard functions in ExpressScribe and Microsoft Word and we explain the importance of phonetic-phonological, morphological, and discourse-pragmatic analyses in identifying and categorizing particular contact phenomena and discourse circumstances.

Discussions throughout this paper are based on transcriptions and digital recordings on deposit in the ***D***igital ***AR***chive to ***DO***cument ***S***panish ***I***n the ***P***aïsos ***CAT***alans, henceforth DARDOSIPCAT. DARDOSIPCAT is a DLA dedicated to collecting, preserving, annotating, cataloging, and disseminating language resources from The Països Catalans. Resources on deposit include longitudinal audio recordings of spoken language made in both Spanish and Catalan as well as orthographic transcriptions of these recordings. The audio recordings represent interviews about language and society in Barcelona. The fact that most research participants spoke in both Spanish and Catalan during their interviews presented multiple challenges for transcribing the recordings, including how to identify and represent potential discourse and contact phenomena perceived in the recordings. With the goal of producing transcriptions that represented the original recordings as faithfully as possible while maintaining easy readability, the DARDOSIPCAT research team innovatively exploited particular software features and carried out diverse linguistic analyses.

II. Literature Review

In our view, all DLAs are a form of language documentation, an interdisciplinary endeavor that aims to create lasting, multipurpose records of language [2]. Given that a central goal of all language documentations is the archiving of the linguistic practices of specific speech communities, systematic recordings of spoken language collected in appropriate sociocultural contexts are vital, as are transcriptions that apply linguistic knowledge to create practical representations, adding value to such primary data [3]. Best practice recommendations regarding the content of such samples [4, p. 571] advocate for comprehensive digital language resources that are “sufficiently broad in scope, rich in detail, and authentic in portrayal that future generations will be able to experience and study the language, even if no speakers remain”. Accordingly, for multilingual communities in which the dynamic interaction of languages may lead to all manner of translanguaging, best practices in language documentation include the archiving of transcriptions of spoken language samples of “ancestral and emergent codes” [5] whose very existence may depend on such usage.

Contact-induced language phenomena can manifest in different linguistic systems, hence the need for implementing

This research was partially funded by Western Michigan University.

linguistic analyses in the transcription of MRSL on deposit in DLAs. For example, while codeswitching as defined by Jakobson, Fant and Halle [6] can be either intentional or spontaneous, Poplack [7] has argued that this practice may be governed by morphosyntactic and phonotactic constraints from either language. Nevertheless, linguistic boundaries between two or more languages often blur, leading to situations in which speakers produce utterances that can be interpreted in multiple languages simultaneously. In this regard, the term *bivalency* refers to “the use by a bilingual of words or segments that could ‘belong’ equally, descriptively, and even prescriptively, to both codes” [8]. More generally, the term *transcodic marker* [1], a catchall for linguistic innovations that occur in language contact situations, may denote codeswitching, bivalency, borrowings, calques, semantic extensions, and/or spontaneous speech innovations.

Following Vann [9], digital recordings and orthographic transcriptions on deposit in DLAs represent our best hope of finding such phenomena in contact dialects, as well as our best way to reference them, and multifaceted linguistic analyses provide the best way to identify them and the discourse-pragmatic strategies they may represent. Correspondingly, such transcriptions also need to address relevant discourse phenomena that deal with turn sequence and organization in social interaction, such as overlapped and simultaneous talk, interruptions, and realignment across turns and sequences [10, 11], whose discovery and identification is greatly facilitated by linguistic analyses that take into account context-dependent conversational actions.

III. Methodology

A. *Software*

Audio recordings were played in ExpressScribe while corresponding transcriptions were written in Microsoft Word. Functions of these two software applications were key to our accounting for both language contact and basic discourse circumstance in orthographic transcriptions of MRSL on deposit in DARDOSIPCAT. Practical implementations are discussed below.

ExpressScribe is freeware that features constant-pitch, variable-speed playback on the fly, as well as user-configurable options for rewinding and advancing playback. These features, particularly the ability to play audio smoothly even at speeds as slow as 25%, were critical to the discovery of the linguistic phenomena under investigation, as the research assistant (henceforth RA) was able to listen meticulously and repeatedly to segments of each recording. Moreover, in a small window within the ExpressScribe interface itself, the RA was able to annotate potential contact and discourse phenomena observed in each recording. These notes were later exported as text files and stored for future research and transcription-related discussions between the Principal Investigator (henceforth PI) and the RA.

As the RA listened to the audio playback in ExpressScribe, the RA typed into Word the utterances that the RA perceived as the RA perceived them, spelling all utterances the way native speakers of Castilian Spanish would typically write down the spoken language heard on the recordings. Word dictionaries set to Spanish were then used to spellcheck the transcriptions. Utterances that the spellchecker flagged as spelled incorrectly in Spanish were considered as potential transcodic markers or discourse phenomena such as false starts or interrupted words. These utterances were then spellchecked in Word dictionaries of other languages to ascertain whether they were in fact words in a language other than Spanish.

B. *Linguistic Analyses*

Once we had uncovered potential language contact and discourse phenomena thanks to strategically implementing the functions of these two software applications, we used different levels of linguistic analysis to categorize these linguistic phenomena accordingly. Corresponding transcriptional markup followed. Once transcripts were finalized, they were converted to PDF format for deposit in DARDOSIPCAT, where master copies are stored in PDF/A format and access copies are served in basic PDF format.

1) Phonetics and Phonology: In transcribing DARDOSIPCAT interviews, phonetic and phonological features were used to distinguish between Catalan and Spanish words in potential situations of bivalency. Bivalent words and expressions such as *Esquerra Republicana* ‘Republican Left’ (the name of a political party in Catalonia) and *Polònia* ‘Poland’, despite ostensibly being Catalan words, were deliberately not always regarded as such by the speakers. Though the decision to determine whether an utterance was being spoken in Spanish or Catalan was a principled one based on phonological criteria, determining the language in which a particular word or expression was being spoken was not always straightforward even when the RA and the PI strongly agreed on the phonology used, because many people in Catalonia speak Spanish with a phonology that may reflect varying degrees of influence from Catalan.

2) Morphology: Morphological criteria were used to determine the language to which a given word or expression belonged in situations of speech innovations due to language contact between Spanish and Catalan. Two examples that illustrate how such criteria were used in transcription are *bastoneres* and *foguerones*. In both cases, we have Catalan-based words, *bastoners* ‘emcees’ and *foguerons* ‘bonfires’, that have been borrowed into Spanish with concomitant morphological change (adoption of the Spanish plural agreement suffix -*es*) that make them appear as Spanish to the transcriptionists. These two cases reflect the sort of contact-induced linguistic innovations that abound among bilingual speakers of Spanish and Catalan in the Països Catalans. Without transcodic morphological analyses, such borrowings might have been deemed spontaneous speech errors or, worse, gone undetected entirely.

3) Discourse and Pragmatics: In the transcription of DARDOSIPCAT’s audio recordings, pragmatic and prosodic analyses played a role in determining how discourse was organized and co-constructed across turns by participants in the conversations transcribed. Extract (1) illustrates an interrupted question in Spanish (an English translation follows):

R:*¿Puede haber gente castellanohablante* (1)
en las manifestaciones <de>-
X: *<Claro.> Sí, sí.*
R: *De independencia?*
X: *La hay, de hecho.*

R: Is it possible to find Spanish-speaking people
at the demonstrations <for>-
X: <Of course.>. Yeah, yeah.
R: For independence?
X: In fact there are.

Extract (1) begins with the PI asking a question to the interviewee, who answers before the question is concluded, thus interrupting the turn in which the question originated. In this and similar examples, pragmatic analysis was used to determine how to categorize such discourse patterns and mark them accordingly in the transcriptions in a principled way that respected the illocutionary force of the utterances spoken despite subsequent discursive interruptions and force abandonments.

IV. Findings and Significance

The software and linguistic analyses carried out in aid of the transcriptions uncovered extensive contact and discourse phenomena that may be of interest to future users of our DLA. In terms of contact, our linguistic analyses revealed transcodic markers including codeswitching, codemixing, bivalency, borrowings, calques, semantic extensions, and spontaneous speech innovations. In terms of discourse, linguistic analyses revealed numerous performance errors, overlapped words, and interrupted words and turns.

To determine which of these phenomena to include in DARDOSIPCAT transcriptions, we considered the best practice recommendations described in Section II. In light of these recommendations, our findings regarding Question 1 are that orthographic transcriptions of MRSL on deposit in DLAs should account for ALL perceivable instances of language contact and spontaneous discourse patterns to the extent that they can do so in a user-friendly way. Simple orthographic conventions, easy-to-read formatting, and minimal markup should prevail so all users can easily understand the transcriptions without linguistic training. This finding is significant as it highlights the importance of transcriptions with sufficient detail to represent linguistic phenomena salient to the community under documentation. In DARDOSIPCAT, accounting for contact and discourse phenomena is key to faithful documentations of significant linguistic patterns in the community's ancestral and emergent codes [5]. These patterns may hold evidence of potential changes in progress. Additionally, this finding provides a straightforward way for researchers to locate and identify such phenomena within the transcripts themselves.

With regard to Question 2, we found that linguistic analysis in the areas of phonetics, phonology, morphology, and pragmatics was useful in identifying and categorizing relevant linguistic phenomena in the transcription of DARDOSIPCAT interviews. Given their compartmentalized nature, we believe such analyses could be useful in creating transcriptions for other DLAs as well, separately or in combination, depending on the phenomena of interest to the transcriptions' audience design. Incorporating linguistic analyses into the transcription process is important for accountability in research resources [2] insofar as accurate rendering of linguistic transcripts strengthens the empirical foundations of those branches of linguistics and related disciplines whose work depends on quality documentary resources.

V. Conclusions

The present report set out to discover what linguistic phenomena orthographic transcriptions of MRSL on deposit in DLAs should account for in terms of language contact and discourse circumstance and to describe software features and linguistic analyses that support the accurate representation of these linguistic phenomena in such transcriptions. While the issues addressed here relate to linguistic phenomena that are particularly pertinent to DARDOSIPCAT, our research questions and methods have implications for other DLAs, especially those that also document MRSLs. Our findings suggest that such DLAs can benefit from a bottom-up approach that targets specific linguistic features relevant to the speech community at hand. Accordingly, the principled implementation of particular software features in tandem with systematic linguistic analysis can be most helpful in this regard.

References

[1] G. Lüdi, "Les marques transcodiques: regards noveaux sur le bilinguisme," in Devenir Bilingüe - Parler Bilingüe. Actes du 2e Colloque sur le Bilinguisme, Université de Neuchâtel, Niemeyer, Tübingen: Max Niemeyer Verlag, 1984, pp. 1–19.

[2] N. P. Himmelmann, "Language documentation: What is it and what is it good for?," in Essentials of Language Documentation, J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Berlin: Mouton de Gruyter, 2006, pp. 1–30.

[3] P. Austin, "Language documentation 20 years on," in Endangered Languages and Languages in Danger: Issues of Documentation, Policy, and Language Rights, L. Filipović and M. Pütz, Eds. Amsterdam: Benjamins, 2015, pp. 147–170.

[4] S. Bird and G. Simons, "Seven dimensions of portability for language documentation and description," Language, vol. 79, no. 3, pp. 557–582, 2003, doi: 10.1353/lan.2003.0149.

[5] A. C. Woodbury, "Language documentation," in The Cambridge Handbook of Endangered Languages, P. K. Austin and J. Sallabank, Eds. Cambridge, UK: Cambridge University Press, 2011, pp. 159–186.

[6] R. Jakobson, G. Fant, and M. Halle, Preliminaries to Speech Analysis: The Distinctive Features and their Correlates. Cambridge, MA: MIT Press, 1952.

[7] S. Poplack, "Sometimes I'll start a sentence in Spanish y termino en español": Toward a typology of code-switching," Linguistics, vol. 18, no. 7/8, pp. 581–618, 1980, 10.1515/ling-2013-0039.

[8] K. Woolard. "Simultaneity and bivalency as strategies in bilingualism," Journal of Linguistic Anthropology, vol. 8, no. 1, pp. 3–29, 1998.

[9] R. E. Vann, "On the importance of spontaneous speech innovations in language contact situations," in Convergence and Divergence in

Language Contact Situations, K. Braunmüller and J. House, Eds. Amsterdam: Benjamins, 2009, pp. 153-182.

[10] H. Sacks, E. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking,” Language, vol. 50, no. 4, pp. 696–735, 1980.

[11] E. Schegloff, Sequence Organization in Interaction: A Primer in Conversation Analysis. Cambridge, MA: Cambridge University Press, 2007.