Creating and designing a corpus of rural Spanish

Carlota de Benito Moreno
Universität Zürich
Romanisches Seminar
CH-8032 Zürich, Switzerland
carlota.debenitomoreno@uzh.ch

Javier Pueyo
College of the Holy Cross
Worcester
01610 Masssachusets, USA
javier.pueyo@gmail.com

Inés Fernández-Ordóñez
Universidad Autónoma de Madrid / Real Academia Española
Departamente de Filología Española
28049 Madrid, Spain
ines.fernandez-ordonnez@uam.es

Abstract

In this paper we address some of the difficulties that arise when compiling a corpus of rural varieties (namely, the COSER corpus of Rural Spanish). These difficulties affect mainly two different aspects of the corpus-building process, i.e., the transcription process, especially regarding the conventions used, and the lemmatization process. We describe the main problems that affected the COSER corpus during these two processes and the solutions that were adopted.

1 Introduction

In this paper we aim to describe the processes of transcription and lemmatization of the COSER corpus, which documents rural varieties of spoken Peninsular Spanish; the difficulties associated to these two processes, and how they were addressed. In section 2 a brief description of the corpus, its compilation process and its main purpose is provided. Section 3 focuses on the transcription process, especially on the transcription rules that were designed specifically for the representation of rural Spanish within this corpus. Section 4 presents the lemmatization process, which had to be adapted to the specific transcription conventions used. Finally, some conclusions are summarized in section 5.

2 The COSER corpus

COSER (an acronym that stands for *Corpus Oral y Sonoro del Español Rural —Audible Corpus of Spoken Rural Spanish* in English) was designed by Inés Fernández-Ordóñez with the goal of documenting rural varieties of Peninsular Spanish in a format that enabled the morphosyntactic study of these varieties. COSER consists of spoken interviews to old rural non-mobile speakers of different villages in Spain that have been recorded *in situ* (that is, not in a lab setting). In its current composition, the mean of the duration of the interviews is 75 minutes – interviews must be lengthy in order to document sufficient instances of different morphosyntactic structures (Fernández-Ordóñez 2009, 2010a & b).

The interviews have been being recorded from 1990 on (and they are still ongoing). So far 1124 villages of 44 different provinces have been interviewed, which amount to 1434 hours of audio and 2248 recorded speakers. Currently, 147 interviews from 141 villages (ca. 184 hours) have been transcribed and are available online (see http://corpusrural.es/) – these amount to 2,727,967 tokens and 1,853,141 words, which comprise 106,505 conversation turns.

So far, the transcription process has been carried out manually by a number of collaborators in the project. Manual transcriptions are highly costly in both economic and time terms, but they also have advantages, especially when dealing with substandard speech, where a human transcriber is more likely to understand and transcribe correctly difficult fragments.[1] As will be seen in section 3 (cf. especially subsection 3.1), the fact that transcriptions are done manually has had a strong impact in the transcription rules.

[1] Now that a significant proportion of the interviews have been manually transcribed, collaboration with private partners to automatically transcribe the rest of the corpus is being sought. So far, we have established contact with Verbio, a firm that specializes in natural language processing.

The available transcriptions are currently being lemmatized automatically using FreeLing, a process that will be explained in detail in section 4.

3 Transcription rules

One of the main decisions that has to be made when compiling a corpus of substandard speech is which phenomena should be included in the transcriptions and which can be left out (for the impossibility of including every possibly relevant detail of the original data in a corpus edition, cf. López Serena 2006). The main guidelines for such a decision must be the purposes of the corpus, but their secondary uses can also be taken into account.

As said above, the main purpose of COSER is to provide a database for research on dialectal morphosyntax of Peninsular Spanish, which advises against providing a phonetic transcription of the interviews. However, some salient phonetic substandard phenomena can interact with morphosyntactic phenomena, which in turn suggests that phonetic phenomena should be included in the transcription.

Hence, COSER adopts an intermediate solution, using "regular" spelling (as opposed to phonetic alphabets) to reflect some phonological (but not phonetic) substandard phenomena. The two main phonological changes included in the transcription are the omission and the addition of phonological segments. For instance, the dialectal pronunciation of *mucho* [ʻmut͡ʃo] 'much' as [ʻmunt͡ʃo] is transcribed *muncho*, adding the extra <n> that reflects the substandard extra [n], and the colloquial pronunciation of *comprado* [kom'praðo] 'bought' as [kom'prao] is transcribed *comprao*, supressing the <d> also in the spelling. The suppression of phonological segments due to the concatenation of sounds within the sentence is marked by a single quote (ʻ): the fast pronunciation of *que has* 'that you have' /ke as/ as /kas/ is hence transcribed as *qu'has*.

While these examples are only phonetic, the application of these rules allows for including phenomena whose precise nature (whether phonetic or morphological) is unclear or debated. This is the case of the dialectal pronunciation of the modal adverb *así* 'so' as *asín*, which can be due both to phonetic and to morphological reasons (cf. Rodríguez Molina 2015); the addition of a final -n to the combination of some verbal forms with the reflexive clitic (*sentarse* 'sit down' > *sentarsen*), which has been interpreted both as a phonetic process and as the addition of a plural morpheme (cf. Heap / Pato 2012) or the common reduction of the universal quantifier *todo* 'all' to *to*, which has important morphosyntactic consequences, such as the loss of gender agreement (cf. Fernández-Ordóñez 2015).

Similarly, changes in the stress position are another phonological process represented in COSER transcriptions. That is, substandard pronunciations such as the pronunciation of proparoxytone words as paroxytone, typical of Aragonese varieties, are transcribed by using an extra accent — that may or may not be in accordance with the standard accentuation rules. For instance, the pronunciation of *pájaro* [ʻpaxaro] as [pa'xaro] is transcribed as *pajáro* (despite the fact that standard spelling would dictate the spelling *pajaro* for such a form). The reason for this "extra" marking is to indicate that there was an actual change in the stress position and that the lack of the accent is not a typo (as most lemmatizer softwares would most likely assume). Once again, while this transcription rule mostly affects phonological phenomena, changes in the stress position may also have morphological consequences, as with verbal forms – the change from [kan'taramos] (*cantáramos*) to [kanta'ramos] (*cantarámos*) alters the morphological relationships within the verbal paradigm.

This systematic representation of phonological phenomena sets COSER apart from similar projects in Spanish, such as PRESEEA, and other languages, such as CORDIAL-SIN for Portuguese, FRED for English or the Nordic Dialect Corpus for Scandinavian languages, which rely only on standard orthography except for those forms that are relevant to morphosyntactic analysis (CORDIAL-SIN transcription conventions: 8), but "do not offer any consistent renderings of phonological features" (FRED user's guide: 10). While it requires substantially more work, the advantage of the approach adopted by COSER is that it does not make any assumptions on what phenomena are relevant for morphosyntactic analysis, hence allowing for the potential discovery of morphosyntacic phenomena that have not yet been described. As secondary effects, these transcription rules make COSER useful also for those who are interested in phonological variation in Peninsular Spanish and provide a more accurate image of the speech of the informants.

A second aspect that had to be dealt with for designing COSER transcription rules is not related to its purpose, but to its material. Transcribing spoken interviews requires some circumstances of the conversation to be taken into account, es-

pecially those that refer to turn-taking, interruptions and self-corrections.

Conversational turns are normally not distributed orderly within the participants of a conversation, but they typically imply the overlapping of at least two speakers during a few seconds. Reflecting properly such overlaps in the transcription is crucial to its alignment with the audio files. In corpora designed for the study the characteristics of spoken language, overlaps during turn-taking are typically transcribed with indented lines, as in the following example taken from conversation 146a of Val.Es.Co:

6	B: pues lo tenemos que celebrar→	[¿eeh?]
7	A:	[claaro]

Figure 1. Indented overlaps in Val.Es.Co.

This convention, however, makes the transcription hard to read – an undesirable effect for a corpus whose main purpose is the documentation of morphosyntactic variation. COSER transcription rules, then, try to avoid this problem by using written tags to indicate that a specific fragment was produced simultaneously to some other fragment and who produced it. The simultaneous fragment does not appear in a new line or paragraph, but is instead inserted within the speech of the first speaker, in the exact moment where overlapping occurs.[2] Example (1), for instance, depicts the overlap of the interviewer (E) with the informant (I1). Different colours are used to increase legibility:

(1) **I1:** Es que es la cueva los Moros… Había una farmacia donde estaba todos los… que dejaba el botiquín. Mira todo esto eran todo casas, en cada de esta hay con dos ventanitas era **[HS:E** Sí…, sí.**]** una… Ahora vive ahí un señor soltero, después aquí las tienen aquí… (COSER 2501, Ausejo, La Rioja)

This representation was chosen both due to readability reasons and because it easily allows for differentiating the speech of various participants in the lemmatization process. An acknowledged shortcoming, however, is the fact that it does not specify the end of the overlap, which is made up for by the fact that the audio files are provided together with the transcription.[3]

Two other fundamental spoken phenomena that must be marked when transcribing interviews for linguistic purposes are interruptions and self-corrections. Not only is the proper representation of these phenomena key for the accuracy of the transcription, but they also have linguistic significance – interruptions can be used as discourse-planning tools (López Serena 2007) and self-corrections can be indicators of sociolinguistic awareness. In COSER, the hyphen (-) indicates an interrupted word (see example (2)) and the vertical bar (|) indicates an interruption followed by a sequence that does not repeat the interrupted sequence, i.e., the first sequence has been altered or corrected (see example (3)). The use of these transcription conventions, then, makes the COSER corpus a useful tool for discourse researchers too.

(2) y lo echas a una especie de banco, entonces **le cla-, le clavan** el cuchillo y sacan la sangre. (COSER 4128-2, Perales de Alfambra, Teruel)

(3) Hace dos años… me parece, no sé si son dos o tres, teníamos **una ce- | una cosecha** que era la, la, la mayor. (COSER 4128-2, Perales de Alfambra, Teruel)

3.1 The disambiguation convention

A special transcription rule is the so-called disambiguation convention, which was especially designed for easing the difficulties that COSER's substandard orthography could cause in the lemmatization process. The phonological processes included in the transcritpion (i.e. omissions of phonological segments and changes in the stress position) contribute to the proliferation of ambiguous forms in the final text and hence pose a potential problem to the lemmatization process. This proliferation of ambiguous forms is especially troubling insofar it affects substandard forms, i.e., the potentially most interesting forms of the corpus.

For instance, the loss of intervocalic /d/ and final /r/, extremely common in Southern varieties, produce the identical pronunciation of the infinitive and the participial adjective feminine of verbs in the 1st conjugation: *cantar* /kan'tar/

[2] Overlaps of more than two speakers are represented in the same way, with two consecutive "overlap tags" inserted within the speech of the primary speaker.

[3] Audio-text alignment is not yet provided in COSER, although it is planned for future stages.

'to sing' becomes *cantá* /kan'ta/, as does *cantada* /kan'tada/ 'sung.FEM'. Similarly, substandard pronunciation of the locative adverb *adonde* /a'donde/ 'where' renders the spelling *ande* /'ande/, a form identical to the 1st person singular of the present subjunctive of *andar* 'to walk'.

Since stress has distinctive value in Spanish, changes in the stress position can also result in ambiguities for the lemmatization tool. For example, a paroxytone pronunciation of *cántara* /'kantara/ 'jug', expected in Aragonese varieties, would become *cantára* /'kantara/, phonetically identical to the 1st and 3rd person singular of the subjunctive imperfect of *cantar*. Although these two forms are not spelled identically (the spelling of the verbal form would be *cantara*, with no accent, see above), *cantára* does not represent an unequivocal dialectal pronunciation of *cántara*, since a third possibility exists: *cantará* /kanta'ra/ is the 3rd person singular of the indicative future of *cantar* and a hypothetical change of stress could be also represented as *cantára*.

Lemmatization tools normally have disambiguation resources, but since these ambiguous forms are only ambiguous because of the special COSER transcription rules, a disambiguation convention was designed in order to help the lemmatization tool with these examples. That is to say, transcribers manually indicate whether a dialectal form is ambiguous and which is the standard reading of the form. This disambiguation convention is quite intuitive and uses the equality sign to identify the standard form, placing both between parentheses. The substandard form is placed at the left of the equality sign, while the standard form is placed at the right. That is to say, /kan'ta/ can be transcribed *(cantá=cantar)* or *(cantá=cantada)*, the adverb /'ande/ is transcribed *(ande=adonde)*, and the substandard pronunciation of the noun *cántara* is transcribed *(cantára=cántara)*, as opposed to a hypothetical *(cantára=cantará)*. The second word in the parenthesis is used by the lemmatization software to assign a tag to the first word, which in turn is the one maintained in the transcription (as available to the public).

4 The linguistic annotation process

The transcription system outlined above normalizes in some degree the language recorded in the interviews. However, as said above, it still preserves many of the phonetic, morphological, lexical, and even syntactic features of oral and rural Spanish, which prevents COSER from being fully lemmatized and PoS annotated with Natural Language Processing (NLP) tools developed for standard written Spanish. Typically, tools for the analysis and annotation of modern languages are trained on and applied to orthographically standardized varieties of such languages. Therefore, lemmatization and PoS annotation of the rural and conversational Spanish encoded in our particular transcription system is still a challenging process for any standard NLP tool.

In order to linguistically annotate our corpus, we decided to extend an existing tool, FreeLing, which is an open-source NLP system, developed at the Universitat Politècnica de Catalunya (Padró, 2011). FreeLing is both a state-of-the art NLP library and a set of linguistic resources with multilingual capabilities that is used for the linguistic processing of standardized modern languages as English, Spanish, Catalan or Russian, among others. Being open-source, it is possible to freely modify its computational code and create new lexical resources, linguistic rules, and statistical information for the analysis of languages originally excluded. More interestingly, it is also relatively easy to extend and adapt the code and the linguistic resources provided by FreeLing in order to analyze non-standard varieties of a language already included, such as standard Spanish in our case.

In order to adapt and extend FreeLing to analyze oral and rural Spanish and to deal with our particular transcription system, we had to modify some key modules that were primarily designed to process standard Spanish written sources:

4.1 Tokenizer

As explained above, our transcriptions include a sheer number of conversational, and linguistic marks, which FreeLing is not able to understand out-of-the-box. In order to preserve the conversational structure and information included within the transcriptions, we pre-processed the transcriptions and converted those marks to XML tags and attributes. For example, indications of simultaneous speech such as [HS:E Sí…, sí.] (see example (1) above) were converted to <HS speaker="E"> Sí…, sí. </HS>. We then modified FreeLing's tokenizer module to include rules that preserve XML tags without splitting them. Additionally, we extended FreeLing's tagging mod-

ule, in order for the tagger to assign customized labels to each of the XML tags in the corpus.[4]

Our transcription conventions also required to modify FreeLing's tokenizer rules to deal with the punctuation marks used to transcribe interruptions and self-corrections (-, …, and |, see section 3), and also to allow the program to recognize lexical blends and contractions containing single quotation marks (*qu'has* for *que has*, *pa'l* for *para el*, etc., see section 3), an orthographical practice unknown to modern Spanish. The tokenized transcriptions contained 8,684 interrupted words marked by a hyphen, and 14,768 self-corrections marked by a vertical bar.

4.2 Lexical dictionary

The first task we needed to address in order to use FreeLing's standard Spanish analyzer was to extend its some 600,000 words/lemma/PoS dictionary with new entries reflecting the vocabulary of the semantic fields related to the rural life. We developed tools to identify all the terms in our corpus that were not included in FreeLing's Spanish lexical resources, and manually confirmed or modified the lemma and PoS tag initially proposed by the program. More than 3,000 words/lemmas/PoS were added to the massive Spanish dictionary shipped with FreeLing.

As explained above, we had marked potentially ambiguous non-standard realizations of common Spanish words by means of equal signs, mapping the non-standard occurrences of a given word to its corresponding normalized form. For example, the adverb *muy* 'very' is frequently shortened to *mu* in oral speech, which is reflected in our transcription system as *mu(0=y)*. All these cases – which amount to 12,750 items – were extracted from the transcriptions and were automatically duplicated as new entries in FreeLing's Spanish dictionary, so that the non-standard form was associated with the lemma and PoS tag of its standard counterpart: *mu(0=y): mu muy RG (< muy muy RG)*. We were also able to automatically duplicate entries and analyses of words with alternating stress patterns since they receive special marking in our transcriptions (see section 3) and, thus, were easily recognized and mapped to standard entries in FreeLing's dictionary.

[4] A total of 296,218 XML marks were obtained in the preprocessing of the 147 available interviews – 24, 309 of which correspond to overlapping fragments.

4.3 Affixation rules

Some of the dialectal varieties of Spanish recorded in the COSER corpus use derivative suffixes and verbal morphological endings that somehow differ of those of standard Spanish. We have extended FreeLing's affixation rules, so that those suffixes and verbal endings are properly recognized and the adequate lemmas and PoS tags are correctly assigned by the tagger. For example, the diminutive suffixes *-ico/a* (in Aragonese Spanish and western dialects) or *-ín/-ina* (in Asturian Spanish and eastern dialects) are much more frequent than the standard ending *-ito/a*. We introduced rules to detect these non-standard derivative suffixes, extract the root from the form, and re-analyze it (for example, *jugos-inos, grande-cico*, etc. for standard *jugos-itos, grande-citos*, etc. are now correctly analyzed as diminutive forms of the lemmas *jugoso* 'juicy' or *grande* 'big').

Furthermore, we also had to extend FreeLing's rules of clitic pronoun annotation since in some varieties of Spanish, both the form of the pronouns (*mos, sos, tos, vos* for standard *nos* 'us' and *os* 'you.OBJ'), and their position differ from standard Spanish. For example, postponed-clitic constructions like *trájo-me-lo* (lit. 'he.brought it to.me') instead of the standard Spanish syntax *me lo trajo* (lit. 'to.me it he.brought') are frequent in the Asturian variety of Spanish.

Adapting an existing tool as FreeLing and its standard Spanish linguistic resources, both to our transcription system and to the oral and rural sources of the COSER has allowed us to fully lemmatize and annotate more than 180 hours of transcribed interviews. Furthermore, having been able to integrate this modified version of the tool into our own programs and workflow will allow our research team to keep updating FreeLing's linguistic resources for the COSER as the process of transcribing more interviews continues.

5 Conclusion

The substandard varieties documented in COSER pose a number of challenges to the adequate transcription and processing of the materials of the corpus. In this paper we have described how we have dealt with such challenges, both at the transcription (where we have resorted to a number of ad hoc conventions) and the lemmatization (where we have adapted previously available tools to such conventions) levels.

References

Araceli López Serena. 2006. La edición como construcción del objeto de estudio. El ejemplo de los corpus orales. In L. Pons Rodríguez (ed.), *Edición y crítica textual*, Madrid / Frankfurt, Iberoamericana / Vervuert, 301-334.

Araceli López Serena. 2007. *Oralidad y Escrituralidad en la Recreación Literaria del Español Coloquial*. Gredos, Madrid.

CORDIAL-SIN = Ana Maria Martins (coord.). 2000-2010. *CORDIAL-SIN: Corpus Dialectal para o Estudo da Sintaxe / Syntax-oriented Corpus of Portuguese Dialects*. Lisboa, Centro de Linguística da Universidade de Lisboa. http://www.clul.ul.pt/en/resources/411-cordial-corpus Transcription conventions available at: http://www.clul.ul.pt/english/sectores/variacao/cordialsin/manual_normas.pdf

COSER = Inés Fernández-Ordóñez. 2005-. *Corpus Oral y Sonoro del Español Rural*. http://corpusrural.es/

David Heap and Enrique Pato. 2012. Plurales anómalos en los dialectos y en la historia del español. In E. Montero Cartelle and C. Manzano Rovira (eds.), *Actas del VIII Congreso Internacional de Historia de la Lengua Española*. AHLE/Meubook, Santiago de Compostela, vol. 1, 829-840.

Inés Fernández-Ordóñez. 2009. Dialect grammar of Spanish from the perspective of the Audible Corpus of Spoken Rural Spanish (or Corpus Oral y Sonoro del Español Rural, COSER). *Dialectologia*, 3, 23-51.

Inés Fernández-Ordóñez, Inés. 2010a. La Grammaire dialectale de l'espagnol à travers le Corpus oral et sonore de l'espagnol rural (COSER, *Corpus Oral y Sonoro del Español Rural*). *Corpus: "La syntaxe de corpus / Corpus syntax*, 9, 81-114.

Inés Fernández-Ordóñez, Inés. 2010b. New methods for the study of grammatical variation and the Audible Corpus of Spoken Rural Spanish. In Gotzon Aurrekoetxea & José Luis Ormaetxea (eds.), *Tools for Linguistic Variation*, Bilbao, Universidad del País Vasco, 119-30.

Inés Fernández-Ordóñez. 2015. *Mucha trabajo*: sincretismo femenino en los cuantificadores evaluativos de Cantabria. In S. García *et al*., *Studium Grammaticae. Homenaje al profesor José Antonio Martínez*, EdiUNo, Oviedo, 337-349.

FRED = Bernd Kortmann et al. 2000-2005. *Freiburg English Dialect Corpus*. http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/ User's guide available at: https://www.freidok.uni-freiburg.de/fedora/objects/freidok:2489/datastreams/FILE1/content

The Nordic Dialect Corpus = Janne Bondi Johannessen, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. *The Nordic Dialect Corpus*. http://www.tekstlab.uio.no/scandiasyn/

Javier Rodríguez Molina. 2015. El adverbio *así* en español medieval: variantes morfofonéticas. In J. M. García (dir.), *Actas del IX Congreso Internacional de Historia de la Lengua Española*. Arco/Libros, Madrid, 1049-1064.

Lluís Padró. 2011. Analizadores Multilingües en FreeLing. *Linguamatica*, 3(2):13-20.

PRESEEA = Francisco Moreno Fernández (coord.). 2014–. *Proyecto para el Estudio Sociolingüístico del Español del España y de América*. Alcalá de Henares: Universidad de Alcalá. http://preseea.linguas.net/Inicio.aspx

Val.Es.Co = Pons Bordería, Salvador et al., *Corpus anotado de español coloquial*, available at http://www.uv.es/corpusvalesco/index.html.