COMPUTATIONAL TOOLS AND SPOKEN CORPORA DESIGN: AN ONGOING DIALOGUE*

LES EINES COMPUTACIONALS I EL DISSENY DE CORPUS ORALS: UN DIÀLEG VIGENT

Victoria Vázquez Rozas
Universidade de Santiago de Compostela
victoria.vazquez@usc.es

Mario Barcala
NLPgo Technologies S.L.
barcala@nlpgo.com

Abstract: The design of an oral corpus and the processes of registering, codifying and treating the materials in order to build a useful resource for linguistic analysis prompt numerous decisions regarding theory and methodology. This article is focused on those stages of corpus construction which are more clearly conditioned by the computational processing necessary to make it functional. In order to adequately match the initial expectations and the real possibilities of using the tool, each feature we intend to codify must be measured against the workload and the means required to do so. Therefore, it is essential to take into account the available possibilities of processing and exploitation as they have a crucial impact on decisions regarding the corpus' construction. Based on experience acquired in the construction of the ESLORA corpus, the present article looks into some of the problems arising in the process of designing an oral corpus, such as the delicacy

(*) This study was financed by the *Agencia Estatal de Investigación* (AEI) 'Spanish State Research Agency' and by the *Fondo Europeo de Desarrollo Regional* (FEDER) (European Regional Development Fund) through the ESLORA+ project (FFI2017-86379-P). The authors are members of the research group *Gramática del español* 'Spanish Grammar' from the University of Santiago de Compostela, which has been awarded a grant for the *Strengthening and Organisation of Research Groups with Potential for Growth* by the Regional Government's Education Department (ED431B 2017/39). The study has also benefited from the participation of the ESLORA project in the *Red temática en estudios de Análisis del Discurso* (FFI2017-90738-REDT).

with which oral phenomena are represented, the segmentation of the discourse, the coexistence of different simultaneous tagging systems and the particularities of annotation in a bilingual or multilingual context.

Key words: oral corpora, stand-off annotation, in-line annotation, segmentation, POS tagging

Resum: El disseny d'un corpus oral i els processos de registrar, codificar i tractar els materials per construir un recurs útil per a l'anàlisi lingüística comporta nombroses decisions pel que fa a la teoria i la metodologia. Aquest article s'ocupa d'aquelles etapes de la construcció d'un corpus que més clarament estan condicionades pel processament informàtic necessari que ha de fer el corpus funcional. Per tal de conjugar les expectatives inicials i les possibilitats reals quan usem l'eina, cada característica que pretenem codificar ha de ser mesurada quant a la càrrega de treball que comporta i els mitjans que són requerits per fer-ho possible. Per això, és essencial tenir en compte els recursos disponibles a l'hora de processar i explotar el corpus, ja que tenen un impacte fonamental en les decisions pel que fa a la construcció del corpus.
Basat en l'experiència adquirida en la construcció del corpus ESLORA, l'article analitza alguns dels problemes que sorgeixen en el procés de dissenyar un corpus oral, com ara el grau de detall en què és representat el fenomen oral, la segmentació del discurs, la convivència de diferents sistemes d'etiquetatge simultanis i les particularitats de l'anotació en un context bilingüe o multilingüe.

Paraules clau: corpus oral, anotació *stand-off*, anotació en línia, segmentació, etiquetatge morfològic.

1. INTRODUCTION

The literature on Corpus Linguistics highlights the relevance attributed to the design as a distinctive trait of corpora, as opposed to other compilations possessing a more random nature (known as *archives* or *text collections*)[1] (see, e.g., Atkins *et al.* 1992; Biber 1993; Sinclair 1995, 2005; Biber *et al.* 1999: 4; Rojo 2016; Egbert 2019). Corpus design requires an explicit formulation of the criteria guiding the selection, organization, and codification of materials; these criteria are simultaneously determined by the objectives of the corpus and its intended use (see, e.g., Tognini-Bonelli 2001; Hunston 2002; McEnery *et al.* 2006; Gries & Newman 2013; Rojo 2014; Weisser 2016; Torruella 2017).

1. For instance, in Cohen *et al.* (2005: 38) specific conditions are established to distinguish a corpus from a text collection: «By text collection we mean textual data sets that may include metadata about documents, but do not contain mark-up of the document contents».

Reflections surrounding the theoretical and methodological implications of the elaboration of corpora have allowed for an ever more precise definition of the adequate conditions for the creation of new resources. Research in this field has also brought to light diverse difficulties in the process of construction and subsequent use of corpora for linguistic analysis. Nevertheless, issues concerning design, preparation, and annotation of particular corpora are yet to be solved. By doing so, it will be possible to explicitly and systematically document the criteria applied and the solutions attained in the different phases of elaboration.

Building a corpus is not an easy task. It is common for corpus developers to come across unforeseen problems that question the criteria they themselves have established as initial benchmarks. Complications might become more burdensome when corpus developers find themselves confronted with difficulties at a later stage of the building process, when the coding and annotation system has already been set.

The nature of unexpected problems is subject to variation, as it inherently depends on the nature of the corpus. With regard to the different phases (design, codification, text annotation), the following instances may occur:

i. The corpus must integrate a text that does not match the initially defined textual structure.
ii. The markup and annotation system has neglected relevant information for a particular use of the corpus, deemed necessary afterwards.
iii. The markup and annotation system is excessively detailed and complex, significantly slowing down the execution of the project and complicating both the coherence and consistency of codification and text revision.
iv. Certain linguistic criteria attached to codification and annotation imply a computational cost beyond the project's reach.

When these difficulties are spotted at the initial phase of the construction of the corpus, solutions can be found more easily. However, obstacles at later stages are more challenging, as the possibilities to implement changes are reduced. For this reason, the planning and construction of a given corpus must always be coupled with regular evaluations, in order to consider different options in terms of the markup, annotation and processing available. These periodical assessments bring about an adequate balance between workload and potential results of the corpus at hand.

This article contributes to the field of Corpus Linguistics conceived as «the study of the properties of the corpora» (Gries 2011: 83) -and not to Corpus Linguistics understood as linguistic research based on corpus data. Taking the experience acqui-

red in the design and construction of the ESLORA corpus as our starting point, we look into some features of oral corpora. These elements are determined by the tools employed in its processing and exploitation.

This paper is structured as follows. In section 2, the ESLORA corpus is introduced, touching upon the objectives motivating its compilation (§2.1), as well as its structure, composition and elaboration process (§2.2 and §2.3). In section 3, we delve into questions related to corpus annotation standards and how they were applied to the ESLORA corpus. Section 3.1 comments on the cyclic character of the phases that make up the process of corpus building, section 3.2 reflects on the pros and cons of different annotation alternatives, section 3.3 explains the decisions taken to build the ESLORA corpus format and sections 3.4 and 3.5 concern the specific particularities involved in the POS tagging task and a multilingual context environment, respectively. Finally, section 4 presents the conclusions of this paper.

2. THE ESLORA CORPUS

2.1 objectives

The compilation of the ESLORA corpus (<http://eslora.usc.es>) was envisaged with a threefold aim. First, we intend(ed) to increase both the amount and variety of available materials in spoken Spanish by registering the use of Galician speakers. This objective has an additional outcome: gathering data from an under-documented linguistic community. Furthermore, the ESLORA corpus aims at examining methods for eliciting speech. The two techniques most often employed are thus analyzed: sociolinguistic interviewing and recording of spontaneous conversation. Finally, our goal is to contribute to the development of new tools for enriching, accessing and retrieving corpus data, such as a morphosyntactic (POS) tagger and a powerful search engine to give access to all the information provided by the materials (data, metadata and annotations).

ESLORA is a corpus of the Spanish language as spoken in Galicia. The majority of the recorded speech is produced by speakers born and living in this region. Most of our informants can also speak Galician and many of them alternate the use of Galician and Spanish in their daily life. The corpus therefore includes instances of code-switching, and also comprises a few samples of bilingual conversations.

According to the latest survey published by the IGE (*Instituto Galego de Estatística*) 'Galician Statistics Agency' based on data collected in 2013, the percentage of inhabi-

tants that speak Spanish in some or all contexts is 68.8%, and the proportion of those who always or more often speak Spanish reaches 48.51%. As for Galician, it is spoken by 73.75% of the population in some or all situations; more specifically, speakers who only speak Galician plus those who use Galician more often than Spanish, statistically amounts to 51.40%.[2] In terms of percentages, there is a slight advantage of Galician language users over Spanish language users (2.91%). However, this advantage is not noticeable when we look at the recent global evolution of language use.[3]

On the other hand, the use of Spanish in Galicia has been noticeable for years. The use of Spanish spread to the detriment of Galician, particularly in urban areas and certain sectors of society (administration, business, culture, and education). Spanish proliferated at a greater pace over the 20th century, and the most recent data indicate that it continues to do so in the 21st (see footnote 3).

Despite the increasing use of Spanish in Galicia, its study has not sparked the interest of traditional Spanish dialectologists. The scarce references found in 20th and 21st century literature about this variety were mainly aimed at identifying —and correcting— 'interferences' caused by Galician-Spanish bilingualism. This normative perspective considers the Spanish language in Galicia not worthy of individual study, unless the intention is to correct 'mistakes' and 'solecisms'. Another reason for the insufficient acknowledgement of non-standard varieties, such as the Spanish spoken in Galicia, is linked to the fact that they are solely found in speech —not in written form— and particularly in informal contexts, such as casual conversations and spontaneous interactions.

The ESLORA corpus is intended not only to document underrepresented uses in already available corpora, but also to evaluate the most widely used techniques for eliciting informal speech: sociolinguistic interviews and secret recordings of spontaneous conversation. We adhere to the hypothesis that the use of a given technique has a relevant effect on the characteristics of the data. Given the interest in conversational data for linguistic analyses and applications, we must identify differing aspects by looking into the samples registered with the two methods mentioned above, and determine if and to what extent both types of data are comparable. For that purpose, both methods were alternatively used to collect two samples (sub-corpora) of Spanish as it is spoken in Galicia.

2. A summary of data of the latest IGE survey about the use of Spanish and Galician languages is available at <http://www.ige.eu/estatico/estat.jsp?ruta=html/gl/ecv/ECV_ResumoResultados_galego.html#02>.

3. The progressive loss of the Galician language comes to light in the comparison of the IGE statistical data from 2003, 2008, and 2013 (<https://www.ige.eu/web/mostrar_actividade_estatistica.jsp?idioma=gl&codigo=0206004>).

2.2 corpus make-up

The ESLORA corpus consists of semi-structured interviews and spontaneous conversations audio-recorded in Galicia between 2007 and 2015. The 1.2.2 version of November 2018 incorporates 56 documents including 647,758 orthographic words (776,260 grammatical elements).

The interview sub-corpus is part of the PRESEEA project, which aims at collecting comparable corpora for the sociolinguistic study of Spanish in Spain and in American countries. As the main objective of this macro-project is to gather representative samples of similar structure and size from each geographical area, it was necessary to put into practice a common methodology for eliciting speech. Employing semi-structured interviewing has proven to be a useful technique to gather a large amount of high-quality recordings to obtain balanced samples stratified by age, gender, and level of education.

The use of this structured method for collecting speech data leads to the production of comparable corpora across places and times. Nonetheless, the interview is essentially a formal situation that prevents the interviewees from spontaneously expressing themselves. As a result, the speech registered tends to display signs of self-control and homogeneity. This uniformity among speakers thus becomes less representative of the sociolinguistic variation.

The conversational sub-corpus gathers recordings of spontaneous interactions among friends and family members in informal contexts. In addition to being the first public corpus focused on naturally occurring data of Spanish spoken in Galicia, the conversation materials can be also compared with interview materials in order to determine the impact of using each of the two techniques of eliciting speech on the linguistic characteristics of the register.

Yet, each tool has its drawbacks: sociolinguistic interviewing cannot overcome the so-called Observer's Paradox (Labov 1972, 1984; cf. also Fernández Sanmartín 2018). Non-intrusive recording of conversational exchanges makes it difficult to achieve a stratified homogeneous sample according to sociolinguistic variables. Moreover, it supposes a higher degree of complexity ethically and technically speaking.

Regarding ethical questions, both interviews and conversations require the informed written consent of the participants, but only for the register of conversations signatures before and after the recording are needed.

Almost all the interviews and part of the conversations were recorded as WMA (Olympus) audio files and were then converted to MP3 or WAV formats in order to be manually aligned and transcribed with either Transcriber or ELAN tools. The

transcription was mainly orthographic, but it also represents features of spoken discourse such as repetitions, false starts, lengthened sounds, hesitations, pauses, etc.

2.3 METADATA

Selecting and structuring metadata are of utmost importance in the design of every corpus, as the information gathered about the situational context and the characteristics of the participants is crucial for further retrieval and analyses. Figure 1 shows the header of the XML view of a document of ESLORA.

```
- <cabecera>
    <nombre_fichero_audio>SCOM_H11_047</nombre_fichero_audio>
    <versión>58</versión>
    <fecha_versión>2010-11-15</fecha_versión>
    <tipo_texto>entrevista semidirigida</tipo_texto>
    <corpus>ESLORA</corpus>
    <subcorpus>entrevistas</subcorpus>
    <ciudad>Santiago de Compostela</ciudad>
    <zona>Galicia</zona>
    <país>España</país>
    <lugar_grabación>Domicilio del informante</lugar_grabación>
    <fecha_grabación>2010-10-22</fecha_grabación>
    <sistema_grabación>mp3</sistema_grabación>
  - <hablantes>
    - <hablante id="hab1">
        <nombre>SCOM_H11_047_hab1</nombre>
        <papel>informante</papel>
        <sexo>hombre</sexo>
        <nivel_educativo>bajo</nivel_educativo>
        <edad>21</edad>
        <grupo_edad>19-34</grupo_edad>
      </hablante>
    - <hablante id="hab2">
        <nombre>SCOM_E_08</nombre>
        <papel>entrevistador</papel>
        <sexo>mujer</sexo>
        <nivel_educativo>alto</nivel_educativo>
        <edad>27</edad>
        <grupo_edad>19-34</grupo_edad>
      </hablante>
    </hablantes>
</cabecera>
```

Figure 1. Header of a document of ESLORA

The markup section of the file represented in Figure 1 uses XML tags to encode data on

i. type of interaction: interview *vs*. naturally occurring conversation,
ii. setting: date, place,
iii. participants: age, gender, education, interactive role (interviewer / interviewee),
iv. and transcription process: format, audio file, date.

In addition to the data included in the current XML header, information was collected on speakers' professions and birthplaces, and the relationship between participants (relatives/family, friends, previously unacquainted), as well as details on transcribers, reviewers and review dates. In addition, as the corpus aims at documenting the variety of Spanish spoken in Galicia, we also collected sociolinguistic information on the speakers' use of Galician and Spanish. In order to do so, we included a sociolinguistic questionnaire to record fine-grained data and statements of the speakers on their linguistic uses and attitudes.[4]

3. METHODOLOGICAL ISSUES IN THE CONSTRUCTION OF THE CORPUS

3.1 phases of construction

In order to closely link the initial theoretical design of a corpus and its actual possibilities in terms of execution and use, Biber (1993: 256) defended that different steps of construction should be applied to fragments of documents, progressively readjusting the design on the basis of the needs that arise. Similarly, a cyclical point of view, in order to improve both tagger and corpus annotation, is advocated by Wallis (2007) (and cf. also Pustejovsky & Stubbs 2012 and Egbert 2019).

The construction of an oral corpus entails a costly investment in terms of time as well as of human and technical resources. As a consequence, it would take more time than expected for the project to attain any palpable results. We therefore adopted Biber's (1993) idea to distribute the workload into phases, which turned out to be fruitful in ESLORA. By focusing on codification tasks in a partial segment of the

4. See <http://gramatica.usc.es/proxectos/presegal/att/Cuestionario.pdf>.

documents and reducing the degree of detail in annotations in the early phases, we are able to later use that part of the corpus more rapidly. Most importantly, this procedure allowed us to have all the processing phases completed for parts of the corpus, thus making them readily available. This allowed us to jointly analyse all the steps and leave relevant changes for further phases, the latter including more documents and a higher degree of detail or granularity in tagging.

Transcription and annotation tasks in particular slow down the work pace. Complications of this kind became acute when, during the first orthographic-transcription phase in the ESLORA case, we tried to account for a wide range of oral phenomena in detail. In consequence, the annotation guidelines turned out to be far too complex, complicating both homogeneity and coherence in document mark-up not only between different annotators but also for each annotator individually.

We began with a standard orthographic transcription and long pause indications, and the structuring of annotation into phases or levels opened up a window for the definition and sophistication of the representation criteria of oral phenomena that lack a conventional written representation, such as the lengthening of sounds, laughs and vocalisations. In this way, we could release an initial version within a reasonable timespan, despite not yet exploiting all the possibilities that the corpus would offer in its final form. Besides, having the corpus function at an early stage also allowed for the testing of computational processing phases, before addressing all the representation elements. This enabled us to rapidly acquire a global vision of the development of the construction, and consequently to correct errors and take decisions to complete the definitive guidelines for annotation.

3.2 annotation

There is a general consensus concerning the legitimacy of Extensible Markup Language (XML), being an adequate choice for text annotation and giving way to XML-based standards specifically designed for linguistic encoding and annotation such as TEI, XCES or LAF.

Employing a given standard in a consistent manner would be the ideal situation in order to correctly manipulate documents; nevertheless, this is sometimes difficult to achieve in reality. What often takes place is that each tool we use to build or modify the corpus texts uses a different format. For instance, Figure 2 shows how we started building our documents with the transcription tool Transcriber in ESLORA, and we switched to another one, ELAN, as time went by. Each of these widely used

tools for transcription storage has its own XML based format: .trs files for Transcriber and .eaf files for ELAN. This forces us to make changes accordingly so as to have all documents in one single format.

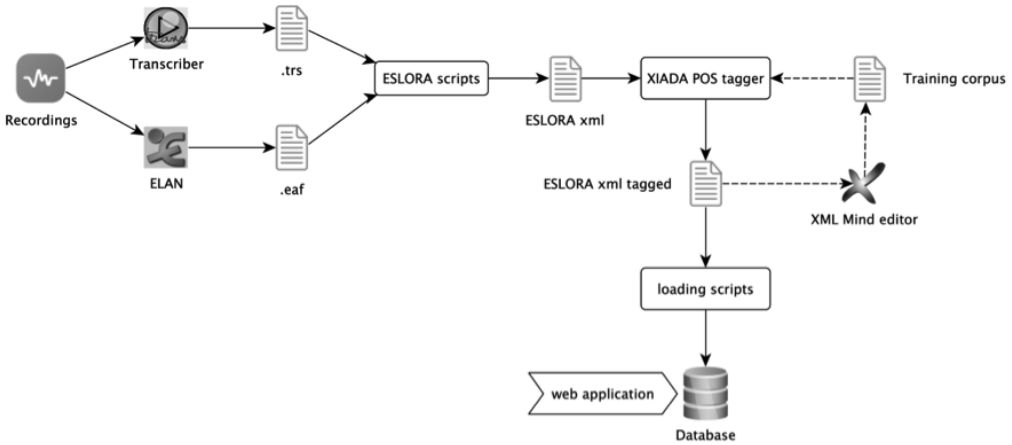

Figure 2. Workflow diagram of the construction of the ESLORA corpus

Our decision was to convert both formats into a common one, which we named ESLORA corpus format (ESLORA xml in Figure 2), in order to allow us to make further processing for Transcriber and ELAN files. We asked ourselves the following questions:

1. What XML encoding format should we be using? A new XML document format created by us or other available linguistic annotation standards?
2. What annotation scheme should we apply? An in-line markup or a stand-off one?

To answer the first question, we kept in mind one of the main objectives of the ESLORA project, namely to develop a web application that can query the corpus using words and morphosyntactic information (from POS tagging) in combination with other phenomena included in the transcription files (word lengthening, fragmentation of words, laughs, quotes, etc.). We used Transcriber and ELAN for transcription and annotation, our idea being to process the documents via a Galician language POS tagger (XIADA). Once the POS tagger was adapted to the Spanish language, it would systematically add the POS information. Finally, all the information would be

uploaded to a database for further use by the web application, as shown in Figure 2. The desired characteristics for the ESLORA XML format were therefore:

i. to be systematically readable and understandable by the application itself, without any need for specific interpretation tools;
ii. to be suitable to be used by a tagger to add POS information;
iii. and to be edited in a user-friendly way.

It was the third item on this list that led us to reject any of the available standards, because we had some problems in trying to achieve a user-friendly view in the XMLMind editor for TEI encoded documents, i.e. to restrict the editor to only show TEI tags and attribute values relevant for the ESLORA project.

Regarding the second question, there are two main approaches to make linguistic annotations in an XML structured text. While in the in-line annotation scheme transcription text and annotations share the same XML document, the stand-off model keeps annotations in an independent file referencing the transcription text.

Current standards recommend using stand-off annotations, mainly due to the fact that multiple overlapping hierarchies can be applied easily (Thompson & McKelvie 1997; Stührenberg 2012).

However, there are many projects that prefer to use in-line annotations (see, e.g., Kavanagh 2019), a preference derived from several factors:

1. Some projects started before stand-off annotation was widely available.
2. There are not many tools that work easily with stand-off annotations.
3. In-line annotations can be very easy to read, use and edit, without any specific application.
4. In stand-off annotations it can be very expensive to change the source documents, yet source-text modifications happen frequently in the process of corpus building.

Choosing an available annotation scheme depends on each project's particularities and needs. As explained above, we started off our work using a transcription tool (Transcriber) and then changed to another tool (ELAN), but we continue to use in-line annotations for different kinds of phenomena (lengthenings, fragmented words, laughs, etc.). This allows us to easily modify the source text as often as we like. In addition, the XIADA POS tagger also works with XML documents and in-line annotations. As a result, we have since been using our own XML format with

in-line annotations. In the next section we discuss some of the choices made concerning the representation of different annotation problems.

3.3 ESLORA XML format

An ESLORA corpus format sample document is shown in (1).[5]

(1)
<document>
<fragment>this is the text transcription of a fragment</fragment>
<fragment>this is <lengthening>another</lengthening> one</fragment>
<fragment>and <fragmented_word>ano</fragmented_word> another more</fragment>
<fragment>and <other_language>el último</other_language></fragment>
</document>

An ESLORA document is an XML document, i.e. a text file with a defined structure composed of opening (<tag>) and closing (</tag>) tags placed before and after sections. It can be graphically represented as a «tree» in computing terms, wherein the tags of the document determine the levels of the tree. For example, the XML document shown in (1) can be represented as the tree in Figure 3.

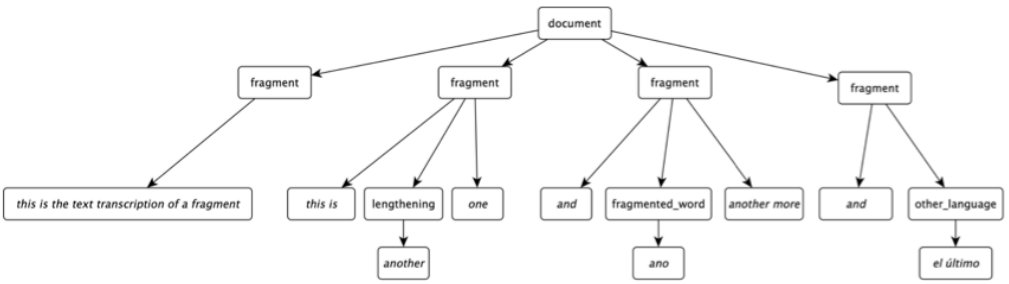

Figure 3. Tree representation of example (1)

If we look carefully at example (1), we will observe that most of the internal tags (the ones found in the lower levels in the tree shown in Figure 3) are closed before their parent tags, which is precisely the fact that gives way to representing these documents

5. We have simplified the real document structure for explanation purposes: e.g. the XML preamble has been removed, time sequence attributes have been omitted, and document headers are not shown.

as trees. On the contrary, the document displayed in (2) is not a valid XML document, because the opening tag <other_language> is not closed before the <fragment> tag.

(2)
<document>
<fragment>this is the text transcription of a fragment</fragment>
<fragment>this is <lengthening>another</lengthening> one</fragment>
<fragment>and <fragmented_word>ano</fragmented_word> another **<other_language>**otro más</fragment>
<fragment>y el último**</other_language>**</fragment>
</document>

Keeping this in mind, we must first decide how to represent a phenomenon that starts in one fragment and ends in another. There are two ways to solve this. The first is to represent the phenomenon in all the affected fragments, as shown in (3) below. Another option would be using different empty tags to set starting and ending points, as in example (4). In the second case, we make use of a special kind of tag (<tag/>) which means the same as <tag></tag>, that is, a tag with empty content.

Although the latter representation is simpler and easier to manipulate for the annotator, its computational treatment is more complex, as it breaks away from the usual XML representation.

(3)
<document>
<fragment>this is the text transcription of a fragment</fragment>
<fragment>this is <lengthening>another</lengthening> one</fragment>
<fragment>and <fragmented_word>ano</fragmented_word> another **<other_language>**otro más**</other_language>**</fragment>
<fragment>**<other_language>**y el último**</other_language>**</fragment>
</document>

(4)
<document>
<fragment>this is the text transcription of a fragment</fragment>
<fragment>this is <lengthening>another</lengthening> one</fragment>
<fragment>and <fragmented_word>ano</fragmented_word> another **<other_language_begin/>**otro más</fragment>
<fragment>y el último**<other_language_end/>**</fragment>
</document>

The complexity of this problem increases in the representation of overlapped phenomena. For example, a fragment in another language as well as a quote (reported discourse), might last a considerable amount of time. The representation of these phenomena varies in complexity, as it depends on their duration and the moment

when they occur. For instance, the fragment in a different language may start before the beginning and end after the ending of the quote or the fragment in a different language may start after the beginning of the quote and end after the ending of the quote.

In the sample document shown in (5) we want to annotate a fragment in a different language starting just before word2 and ending right after word8, and a quote starts immediately after word1 and ends exactly after word7. By using the repetition task solution, we end up with the document shown in (6).

(5)
<document>
<fragment>word1 word2 word3</fragment>
<fragment>word4 word5 word6 word7</fragment>
<fragment>word8 word9</fragment>
<fragment>word10</fragment>
</document>

(6)
<document>
<fragment>word1 <quote>word2</quote> <quote><other_language>word3</other_language></quote></fragment>
<fragment><quote><other_language>word4 word5 word6 word7</other_language></quote></fragment>
<fragment><other_language>word8</other_language> word9</fragment>
<fragment>word10</fragment>
</document>

However, this approach can sometimes be cumbersome, which is why we prefer to apply the start/end tags approach to represent this kind of phenomena, as shown in (7).

(7)
<document>
<fragment>word1 <quote_begin/>word2 <other_language_begin/>word3</fragment>
<fragment>word4 word5 word6 word7<quote_end/></fragment>
<fragment>word8 <other_language_end/>word9</fragment>
<fragment>word10</fragment>
</document>

What we did in ESLORA to adjust human computing and linguistic efforts was to use tag repetition for phenomena of a limited scope (lengthening and fragmented words, for example) and the start/end tags for those with a wider scope (other language fragments and cites, for example). This procedure avoids the continuous repetition of the same task, which can easily lead to mistakes.

A note on using the transcription tool for annotation

We have used the ESLORA corpus format to illustrate the solutions adopted to counteract problems encountered in the annotation process. But as mentioned above, we made the annotations on the transcription tool. So, once we knew what the ESLORA corpus format was like, we defined some guidelines for the annotators about how these must be applied inside Transcriber and ELAN tools.

Inserting XML tags may result in tedious and repetitive work if the transcription tool lacks some kind of assistance for this task. XML tags usually have long names, which must be typed on the keyboard every single time, a time-consuming and error-prone task. In this respect, Transcriber allowed us to create some macros in order to produce all the XML-based tag information. However, when using ELAN we were forced to combine XML tags with symbolic representation. Short-named start/end XML tags were used to represent multi-word phenomena while symbolic representation was applied to one-word elements.

3.4 part of speech (pos) tagging

3.4.1 Segmentation

It is common for a corpus building project to include a POS tagging stage, as this kind of processing enriches the corpus information to a considerable extent. However, POS taggers require full-sentence fragments to work correctly, generating problems to be tackled in oral corpora such as ESLORA. If we arbitrarily break the transcriptions into segments, we will obtain a higher POS tagger error rate, therefore we must mark fragments when the syntactic context changes.

Despite the existence of certain methods to solve this issue (Pietrandrea *et al.* 2014; Wang *et al.* 2014), we simply cut fragments on the basis of long pauses and silences yet keeping in mind the possibility of linguistic inaccuracy. Other risks are relatd to the consistency between annotators or the slow pace of the project. For these reasons, we do not discard the possibility of reconsidering this procedure in the future.

3.4.2 Tagset

In POS tagging it is of crucial to decide on the so-called tagset, the set of morphosyntactic tags used in the corpus. We believe the EAGLES guidelines to be

an appropriate frame for our tagset. It is necessary, however, to restrict the number of tags (450 in ESLORA) because if the tagset is too broad, the morphosyntactic information is highly detailed, but at the same time the error rate increases. It is therefore convenient to find the right balance between the degree of detail of the information retrieval application and the POS tagging success rate.

3.4.3 Intra-word annotation

We may sometimes need to make an annotation that affects only part of a word, for example, to specify in which part of the word a lengthening takes place. Unfortunately, this intra-word annotation turns out to be not very succesful mainly due to the limitations imposed by the POS tagger. This is due to the fact that the tagger must be instructed to ignore internal tags, which, as yet, is not feasible in XIADA or in any other tagger. For the time being, we annotate this type of phenomena to the entirety of the word until this limitation can be overcome.

3.5 bilingual and multilingual contexts

In a bilingual or multilingual context, speakers often switch from one language to another, usually more or less arbitrarily, leading to a certain level of complexity when transcribing and annotating. We first must determine the criteria concerning when and how annotation corresponds to different linguistic phenomena. In the case of a bilingual or multilingual context, we must establish which language changes will be annotated and how this task will be accomplished.

In the case of the ESLORA corpus, we are faced with a bilingual context wherein speakers switch from Spanish to Galician and vice versa in different circumstances. As our main objective is to study the use of Spanish in Galicia, we enclose Galician fragments in <lengua_inicio nombre="gl"/><lengua_fin/> tags, configuring the POS tagger in such a way that it ignores these fragments and to exclude them from the web application searches. Moreover, we have decided not to annotate Galician-isolated words in order to facilitate the fluency of a given syntactic structure and in order for the tagger to achieve a higher success rate.

Managing linguistic information in a context with more than two languages implies a high degree of complexity and, therefore, an in depth analysis of computational requirements and tools. However, if we can focus on one language ignoring the others, usually many things can be simplified.

4. CONCLUSIONS

This paper has focused on solving some issues in spoken corpus construction related to computing capabilities and limitations. A careful assessment of the ESLORA building process reveals the computational implications derived from the oral nature of the corpus and from the available tools that were employed in its coding and annotation. Working with oral material meant having to transcribe and align the audio to text, which in ESLORA was initially carried out using Transcriber and later ELAN. The characteristics and possibilities of these tools established different tagging formats, but despite these differences it was still possible to jointly process all the data in order to integrate them into the same query application. Furthermore, the representation of oral phenomena meant it was necessary to make the text tagging format compatible with the annotation of other specifically oral phenomena, frequently overlapping each other.

It is widely known that the XML standard is a good choice for corpus annotation; but without going beyond this standard we have found that there are several different approaches to solve the same problems. In the ESLORA project, instead of using some of the commonly used standards, we have built our own XML document structure in order to be able to integrate the transcriptions carried out with two different tools and codified in specific XML formats.

In the same way, we have used in-line annotation, rather than stand-off ones in order to achieve a good balance between the manual workload for annotators and the developing effort of our computing team.

The experience of building the ESLORA corpus confirms the importance of a well thought-out structure for the documents of a corpus. Choosing either an XML or any other standard is not as important as to set out an organized document structure, because it enables us to transform our documents to whatever format or standard we need at any time to achieve new aims.

Defining and testing the stages that the corpus documents must go through at the beginning of a corpus building project is very helpful in detecting and solving problems easily. But it is also useful to constantly review and improve all the processes involved in order to discover and overcome the difficulties which occasionally can arise.

Last but not least, keeping note of all the decisions made and criteria chosen as well as documenting all processes, stages and manual tasks is very useful for the coherence of the project and the minimization of mistakes.

Victoria Vázquez Rozas
Universidade de Santiago de Compostela
victoria.vazquez@usc.es
ORCID 0000-0001-8155-669X

Mario Barcala
NLPgo Technologies S.L.
barcala@nlpgo.com
ORCID 0000-0002-6736-2773

ELECTRONIC TOOLS AND STANDARDS CITED

EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora, EAGLES Document EAG-TCWG-MAC/R, 1996.

ELAN: ELAN [Computer software] (V. 5.7 and 5.7-FX), June 14, 2019). Nijmegen: Max Planck Institute for Psycholinguistics. [Online: <https://tla.mpi.nl/tools/tla-tools/elan/>.]

ESLORA: Corpus para el estudio del español oral (V. 1.2.2, November, 2018). ISSN: 2444-1430. [Online: <http://eslora.usc.es>.]

LAF: ISO 24612:2012 Language resource management - Linguistic annotation framework (LAF). [Online: <https://www.iso.org/standard/37326.html>.]

TEI: Text Encoding Initiative. <https://tei-c.org>

Transcriber: A tool for segmenting, labeling and transcribing speech. [Online: <http://transag.sourceforge.net/>.]

XCES: Corpus Encoding Standard for XML. [Online: <http://www.xces.org>.]

XIADA: Etiquetador/Lematizador do Galego Actual. [Online: <http://corpus.cirp.gal/xiada>.]

XML: Extensible Markup Language (XML) V. 1.0. [Online: <https://www.w3.org/TR/xml>.]

XMLmind editor: XML Editor. [Online: <https://xmlmind.com>.]

BIBLIOGRAPHIC REFERENCES

Atkins, S., J. Clear & N. Ostler (1992) «Corpus design criteria», *Literary and Linguistic Computing*, 7 (1), p. 1-16. DOI: 10.1093/llc/7.1.1.

Biber, D. (1993) «Representativeness in corpus design», *Literary and Linguistic Computing*, 8/4, p. 243-257. DOI: 10.1093/llc/8.4.243.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999) *Longman Grammar of Spoken and Written English*, London/New York, Longman.

Cohen K. B., L. M. Fox, P. V. Ogren & L. Hunter (2005) «Corpus design for biomedical natural language processing», *Proceedings of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, Detroit, June 2005, p. 38-45. DOI: 10.3115/1641484.1641490

Fernández Sanmartín, A. (2018) «La entrevista libre como método para evitar la paradoja del observador. Un estudio de corpus», *CHIMERA. Romance Corpora and Linguistic Studies*, 5 (2), p. 141-196. DOI: <http://dx.doi.org/10.15366/chimera2018.5.2.001>.

Gries, S. Th. (2011) «Methodological and interdisciplinary stance in Corpus Linguistics», in V. Viana, S. Zyngier & G. Barnbrook (eds.), *Perspectives on Corpus Linguistics*, Amsterdam-Philadelphia, John Benjamins, p. 81-98. DOI: 10.1075/scl.48.06gri

Gries, S. Th & J. Newman (2013) «Creating and using corpora», in R. J. Podesva & D. Sharma (eds.), *Research Methods in Linguistics*, Cambridge, Cambridge University Press. DOI: 10.1017/CBO9781139013734.015.

Hunston, S. (2002) *Corpora in applied linguistics*, Cambridge, Cambridge University Press.

Jesse, E. (2019) «Corpus Design and Representativeness», in T. Berber Sardinha & M. Veirano Pinto (eds.), *Multi-Dimensional Analysis: Research Methods and Current Issues*, London, Bloomsbury, p. 27-42. DOI: 10.5040/9781350023857.0010

Kavanagh, K. (2019) «XML mark-up: an annotation tool for discourse analysis». [Online: <https://walesdtp.ac.uk/methodsblog/2019/05/21/xml-mark-up-an-annotation-tool-for-discourse-analysis/#more-116>, accessed: 2019-07-30.]

Labov, W. (1972) «Some principles of linguistic methodology», *Language in Society*, 1 (1), p. 97-120. DOI: 10.1017/S0047404500006576.

— (1984) «Field Methods of the Project on Linguistic Change and Variation», in J. Baugh & J. Sherzer (eds.), *Language in Use: Readings in Sociolinguistics*, Englewood Cliffs, NJ, Prentice Hall, p. 28-66.

McEnery, T., R. Xiao & Y. Tono, eds. (2006) *Corpus-Based Language Studies: An advanced resource book*, London / New York, Routledge.

Pietrandrea, P., S. Kahane, A. Lacheret-Dujour & F. Sabio (2014) «The notion of sentence and other discourse units in spoken corpus annotation», in H. Mello & T. Raso (eds.), *Spoken corpora and Linguistic Studies*, Amsterdam, John Benjamins, p. 331-364. DOI: 10.1075/scl.61.12pie

Pustejovsky, J. & A. Stubbs (2012) *Natural Language Annotation for Machine Learning. A Guide to Corpus-Building for Applications*, Sebastopol, California, O'Reilly Media.

Rojo, G. (2014) «Hispanic Corpus Linguistics», in M. Lacorte (ed.), *The Routledge Handbook of Hispanic Applied Linguistics*, New York, Routledge, p. 371-387.

— (2016) «Los corpus textuales del español», in J. Gutiérrez-Rexach (ed.), *Enciclopedia lingüística hispánica*, Oxon, Routledge, p. 285-296.

Sinclair, J. (1995) «Corpus typology - a framework for classification», in G. Melchers & B. Warren (eds.), *Studies in Anglistics*, Stockholm, Almqvist and Wiksell International, p. 17-34.

— (2005) «Corpus and Text - Basic Principles», in M. Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford: Oxbow Books, p. 1-16. [Online: <http://ota.ox.ac.uk/documents/creating/dlc>, accessed: 2019-07-26.]

Stührenberg, M. (2012) «The TEI and Current Standards for Structuring Linguistic Data: An Overview», *Journal of the text encoding initiative*, 3. DOI: 10.4000/jtei.523. [Online: <https://journals.openedition.org/jtei/523, accessed: 2019-07-30>.].

Thompson, H. S. & D. McKelvie (1997) «Hyperlink semantics for standoff markup of read-only documents», *Proceedings of SGML* Europe 1997: The next decade - Pushing the Envelope, Barcelona, p. 227-229. [Online: <http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>, accessed: 2019-07-27.]

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*, Amsterdam, John Benjamins.

Torruella Casañas, J. (2017) *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación lingüística*, Frankfurt, Peter Lang.

Wallis, S. (2007) «Annotation, retrieval and experimentation. Or: you only get out what you put in», *Studies in Variation, Contacts and Change in English* (*VARIENG*) 1: *Annotating Variation and Change*. [Online: <http://www.helsinki.fi/varieng/series/volumes/01/wallis>, accessed: 2019-08-05.]

Wang, I., S. Kahane & I. Tellier (2014) «Macrosyntactic Segmenters of a French spoken corpus», *Ninth Language Resources and Evaluation Conference* (LREC'14), May 2014, Reykjavík, European Languages Resources Association (ELRA), p. 3891-3896. [Online: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/889_Paper.pdf>, accessed: 2019-08-02.]

Weisser, M. (2016) *Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis*, Malden, MA: Wiley-Blackwell.