EModSar: A Corpus of Early Modern Sardinian Texts

Nicoletta Puddu
University of Cagliari
Cagliari, Italy
nicoletta.puddu@unica.it

Luigi Talamo
Saarland University
Saarbrücken, Germany
luigi.talamo@uni-saarland.de

Abstract

English. The article introduces the Early Modern Sardinian Corpus (EModSar), a corpus featuring nine manuscripts from the Early Modern Period (16th-17th centuries) written in Sardinian with passages in Catalan and Latin. Manuscripts are encoded according to the TEI-P5 guidelines, annotated for bibliographic, philological and linguistic features and published on-line using TEITOK, a software aimed at combining digital philology and corpus linguistics.

Italiano. Presentiamo EModSar (Early Modern Sardinian), un corpus composto da nove manoscritti della prima età moderna (XVI-XVII secolo) scritti in sardo con inserti di catalano e latino cancelleresco. I manoscritti sono codificati secondo le linee-guida TEI-P5, annotati per caratteristiche bibliografiche, filologiche e linguistiche, e resi disponibili on-line tramite il software TEITOK, che combina le esigenze della filologia digitale con la flessibilità di ricerca degli strumenti della linguistica dei corpora.

1 Introduction

In this paper[1] we present the Early Modern Sardinian Corpus (EModSar: http://corpora.unica.it/TEITOK/emodsar)[2], a historical corpus developed within a more general project whose aim is to describe the linguistic repertoire of Sardinia in the Modern Era[3] (see section 2.1). Our main research question addresses the impact of language contact on Sardinian, and, in order to answer this question, we decided to build a pos tagged and lemmatized corpus covering texts from the 16th to the 17th century which also contains extralinguistic information about the chosen texts (see section 2.2). Moreover, given that our texts are written in Sardinian, but also contain sections written in Catalan and Latin, we wanted to both preserve multilingualism, and also ensure that our corpus tools focused on the linguistic analysis of Sardinian. As Pahta et al. 2018:10 point out, multilingual historical corpora are rarer than monolingual ones, and have not been used extensively in historical linguistics. However "embracing a multilingual approach to language history leads the researcher to look beyond the main language of a text and consider what a holistic overview of all the languages in it reveals about the 'grammar' of non-monolingual writing on the one hand or individual identity or social practice on the other" (Pahta et al. 2018:5). Consequently, we decided to adopt the TEI-P5 guidelines to code our documents in order to accomplish Lass 2004's three desiderata for a proper historical corpus (i.e. "maximal information preservation", "no irreversible editorial intervention", and "maximal flexibility"). On the one hand, the use of TEI-P5 for our corpus allowed editorial choices to be preserved in the text at the philological level, while, all the relevant information could be inserted in the header. The use of a TEI-P5 encoding is not a common standard in historical corpora. As Jenset and McGillivray 2017:125 note, "TEI is not very widely used for historical corpora, where there is a stronger emphasis on linguistic annotation rather than on paleographic and historical markup. However, in the case of historical texts, the information contained in these tags can

[1] For Italian academic purposes only, Nicoletta Puddu was responsible for Sections 1 and 2 and Luigi Talamo for Sections 3 and 4.

[2] The corpus is currently composed of nine manuscripts, for a total of 6495 tokens.

[3] EModSar has been developed under the project *System for developing and annotating a corpus of ancient Sardinian texts*, funded by the Regione Autonoma della Sardegna (*Capitale Umano ad alta qualificazione*, L.R. 7/07, year 2015).

be crucial to the interpretation of the text and should be considered by the language processing tools. [. . .].” A convenient solution is the use of softwares such as TEITOK (Janssen 2016), a tool which can handle both textual mark-up and linguistic annotation. Since our texts have been annotated at three different levels (at the document-level, at the section-level and at the token-level (see section 3.2), queries in EModSar can combine different levels in order to connect linguistic information with extralinguistic information.

2 Language and texts

2.1 Sardinian in the Modern Era

The linguistic repertoire of Sardinia in the Modern Era is largely understudied, but it is extremely interesting since it sees the presence of many different languages within the same period. From 1324 onwards, the kingdom of Aragon gradually took possession of the Island, and as a consequence Catalan became the official language. After the unification of the Kingdom of Aragon with the kingdom of Castile, Castilian began to spread, but Catalan actually remained in use for juridical and administrative purposes, while Castilian became the language of Universities and of the Church (Virdis 2017). Thus, between 1324 and 1720, when the Island was conceded to the House of Savoy and started its process of Italianization, Sardinia was under Iberian domination. However, Sardinian continued to be used in juridical documents of both a public and, especially, private nature particularly in the countryside. Both Catalan and Castilian deeply influenced Sardinian during the Iberian domination.

The Sardinian language of this period is documented through two typologies of documents: literary sources and juridical sources. The Sardinian literati in the Modern Era usually wrote in the dominant languages (mainly Castilian). However, some of them (like Antonio Lo Frasso) inserted some Sardinian sections in their works or even wrote entire compositions in Sardinian (like Girolamo Araolla). What is clear, however, is that all the literati living in Sardinian were highly plurilingual (Marci 2006).

As for juridical documents, Sardinian was used during trial courts, not only for testimonies, but also for other stages of the trial. In private documents Sardinian appears in notary deeds mainly containing sales, donations, debit notes, last wills and testaments (Cadeddu 2013). While we have critical editions of literary texts from the Modern Era, juridical documents are mainly kept in a number of archives in Sardinia. Only a small part of these documents have been published, mainly in historical studies: there are very few critical editions and no systematic linguistic studies.

2.2 The choice of texts

In our project, we wanted to study Sardinian of the Modern era, in the perspective of historical sociolinguistics in Romaine 1992’s terms. In order to do so, we decided to create the Early Modern Sardinian Corpus by encoding and annotating juridical documents of the Modern Era, annotated by POS and lemma, and accompanied by contextual information. To date, we have encoded nine documents written in Sardinian dating from the 16th to the 17th century retrieved from the *Archivio storico del Comune di Cagliari* and the *Archivio di Stato di Cagliari*. Most of the retrieved documents come from villages in the Northern Sardinian area and from the towns of Sassari and Bosa. However, we know for certain that documents written in Sardinian datable to those centuries also exist in southern Sardinia. We do not expect to find any documents in Sardinian for the city of Cagliari where Catalan was widespread in all the written domains.

Our documents have presented many problematic aspects typical of historical corpora which we will exemplify by discussing document Osp250 which contains the last will of Canonigu Montixi, the priest of the diocese of Arborea, who, in 1569, leaves a “fellowship” to one of his relatives so that he can study grammar, philosophy and theology.

First of all, our documents are characterized by a high level of orthographic variation, both between different documents and within the same document. For instance, in Osp250 the preposition ’in’ can have different orthographic realizations (*in*, *jn*, *en*). Moreover, we have many cases of univerbation, such as *insu* ’in the’, *inpodere* ’in power’, *etinsu* ‘and in the’.

Secondly, our documents are multilingual and we can have code-mixing both at the intersentential level and at the intrasentential level (on different levels of code-switching in historical texts see Kopaczyk 2018). Different codes often correlate with different sections of the document. If we adopt the traditional subdivision in the *formulae* which make up the document, we can see that the *datatio* and the *dispositio* (the core of the document) in Osp250 are written in Sardinian, while the *roboratio testes* and the *completio* are in Catalan. However, we also have intrasentential code-mixing. First of all, as could be expected in juridical documents, we have Latin expressions, such as *ut supra*, *qui supra fidem facio*. But, even more interestingly, we have Catalan and Sardinian code mixing. The *datatio* in Osp250 is in Sardinian, but we find the form *en* for the preposition 'in' , and the name of the month 'June' in the Catalan form *junny*. By contrast, in the *completio*, written in Catalan, the name of the month 'July' is in the Sardinian form *treulas*. Given the close affinity between the different languages present in the document, it is worth noting that, it is not always simple to identify the instances of code-switching, nor to distinguish code-mixing from borrowing.

Finally, our documents are 'stratified', since they have come to us via several passages. Osp250 contains the last will of Canonigu Montixi, but the codicil was redacted by another scribe-priest, Antiogo Molarja. Moreover, the document we have was actually copied by the scribe Sebastià Polla in 1648 at the request of another citizen from Villanovafranca. The document finally arrived in the Archives of the Hospital of Sant'Antonio, since Canonigu Montixi had decided that, were the chain of heirs to die out, his house would have gone to the hospital.

3 Corpus building and annotations

3.1 Corpus building

Due to the mixed nature of our corpus, we needed a software that was able to combine philological aspects i.e., faithful rendering of the manuscripts, bibliographic and historical information with the standard tools used in corpus linguistics i.e., a powerful and flexible query engine. Our choice fell on TEITOK[4] (Janssen 2016), a software developed by Marteen Janssen at the CELTA-ILTEC institute (University of Coimbra, Portugal); in a nutshell, TEITOK is organized in two main components: (i) a web-based application that renders XML files annotated according to the TEI-P5 guidelines and (ii) a suite of executable binaries that convert XML files into the Open Corpus WorkBench (CWB: Evert and Hardie 2011) file format. The first component of Teitok fits our philological needs, as we were able to reproduce our manuscripts with the original page and line breaks, ligatures and graphic variants of linguistic forms (words), while the second component allows us to search our corpus using the Corpus Query Processor (CQP), either from the standard command line facility or using the web application.

Although Teitok is also a powerful XML editor, we employed external XML editors such as oXygen in order to deal with the TEI encoding and annotation processes. Once annotated according to the TEI-P5 guidelines[5], TEI-XML files are uploaded to the web application where they are automatically split into tokens by the Teitok tokenizer. As for the linguistic annotations, Teitok contains some in-development pos-tagging and lemmatization facilities, which have been proven to perform well on historical varieties of languages (Janssen et al. 2017); however, the parts of speech tagging and lemmatization processes, as well as the difficult process of the annotation of graphic variants are all performed manually: at the moment the creation of annotation tools for Sardinian is work in progress (Puddu and Stein 2018) and no annotated corpus is available even for contemporary Sardinian.

Summing up, our corpus building process can be summarized as follows:

1. creation of the XML files: encoding of manuscripts;
2. XML files become TEI-XML files: text annotation according to the TEI-P5 guidelines (TEI header and text elements);
3. automatic tokenization of the TEI-XML files, which are stored in the web application (Teitok);

[4] http://www.teitok.org

[5] The EModSar corpus complies with the latest version of the TEI-P5 guidelines, 3.6.0 released on 16/07/2019. Whenever relevant, we have indicated the URL for the online documentation in the footnotes.

4. manual pos-tagging, lemmatization and annotation of graphic variants.

3.2 Annotations

The annotations featured in EModSar can be conveniently divided into three types: (i) document-level annotation, (ii) section-level annotation and (iii) token-level annotation.

The first type of annotation corresponds to the TEI element known as ‘header’ and contains bibliographic and, to a lesser extent, linguistic and sociolinguistic information; out of the five principal components described by the TEI-P5 guideline[6], we have compiled the ‘file description’, the ‘text profile’ and the ‘revision history’ components. The ‘file description’ component[7] contains bibliographic information such as the repository, collection and archival reference of the manuscript, a brief history of the manuscript tradition and the name(s) of the author and copyist. In the ‘text profile’ component[8], we have gathered information about the place and redaction of the manuscript, the language(s) employed and a summary of the content. As we have pointed out in the previous section, this kind of information is of paramount importance for historical corpora. Finally, the ‘revision history’ component[9], as the name suggests, works as a change log displaying the date when the TEI-XML file was last changed; the component is most useful during the process of corpus building, which is usually characterized by many versions of the same TEI-XML file, often shared between several collaborators.

Annotations at the section-level are performed within the TEI element known as ‘text’, which in turn is divided into different sections, marked up by the <div> tag. Note that this text arrangement does not reproduce any formal elements of the original manuscript, but was carried out by the archivist during the encoding process. As mentioned earlier, we decided to mark this structure since it appears to be related to code switching. The <div> tag contains two attributes: the section attribute, describing one of the *formulae* in which a notary document is customarily arranged and the language attribute, giving the language used in the section. For instance, the following text snippet represents the section-level annotation of Osp250, whose *formulae* were mentioned in Sect. 2.2:

```
...
<div n="1" type="datatio" lang="srd" id="div-1"> ... </div>
<div n="2" type="dispositio" lang="srd" id="div-2">...</div>
<div n="3" type="notitia testium" lang="cat" id="div-3">...</div>
<div n="4" type="subscriptiones" lang="cat" id="div-4">...</div>
<div n="5" type="completio" lang="cat" id="div-5">...</div>
<div n="6" type="completio" lang="cat" id="div-6">...</div>
<div n="7" type="dispositio" lang="srd" id="div-7">...</div>
<div n="8" type="completio" lang="cat" id="div-8">...</div>
<div n="9" type="dispositio" lang="cat" id="div-9">...</div>
<div n="10" subtype="dorsale" lang="ita" id="div-10">...</div>
<div n="11" subtype="dorsale" lang="cat" id="div-11">...</div>
...
```

The third type of annotation takes place at the token level and, just like the previous section-level annotation, is implemented through the attributes of the <tok> tag; the tag is not described in the TEI-P5 guidelines and is added by Teitok during the automatic process of tokenization. Each token is annotated for graphic variants and for linguistic information, for a total of five different attributes; as for the graphic variants, we have distinguished between (i) ‘written form’, corresponding to the graphic variant as found in the manuscript, (ii) ‘extended form’, which is a written form with expanded abbreviations and (iii) ‘normalized form’, showing a tentative normalization of the graphic variant. For example, the annotation of the three different orthographic realizations of the preposition ‘in’, which we have discussed in Section 2.2 is given as follows:

[6] https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD1 Last accessed on 23/11/2019.
[7] https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD2 Last accessed on 23/11/2019.
[8] https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD4 Last accessed on 23/11/2019.
[9] https://tei-c.org/release/doc/tei-p5-doc/en/html/HD.html#HD6 Last accessed on 23/11/2019.

```
<tok id="w-10" form="en" fform="en" nform="in" pos="PRE" lemma="in">en</tok>
<tok id="w-100" form="jn" fform="jn" nform="in" pos="PRE" lemma="in">jn</tok>
<tok id="w-513" form="in" fform="in" nform="in" pos="PRE" lemma="in">in</tok>
```

As for the linguistic information, we provide annotations for (iv) parts of speech and (v) lemma; the parts-of-speech tagset is an adaptation of the tagset used in the Medieval Sardinian Corpus, which contains texts written in an earlier stage of Sardinian (Puddu 2015, Puddu and Stein 2018), and features 25 tags, some of which are specified for morpho-syntactic properties such as verbal mode and nominal definiteness.

Finally, let us just briefly mention how we handled linguistic expressions - mostly, noun and prepositional phrases - written without spaces between words in the manuscripts. In order to faithfully reproduce the manuscripts, these linguistic expressions are encoded without spaces in the written form of EModSar, with a correspondence between a linguistic expression and a single token; at the same time and for the purpose of linguistic queries, the linguistic expression is split into tokens in the normalized form of our corpus by means of another non-standard TEI tag, <dtok>, which is introduced by Teitok (Janssen 2016:4038) and nested into the <tok> tag. Take for instance the prepositional phrase *in podere*, which was originally written as a single word in one of the manuscripts:

```
<tok id="w-280" form="inpodere" fform="in podere" nform="in podere">
inpodere
<dtok id="d-280-1" form="in" fform="in" nform="in" pos="PRE" lemma="in"/>
<dtok id="d-280-2" form="podere" fform="podere" nform="podere" pos="NOUN"
lemma="podere"/></tok>
```

4 Further developments

In building the Early modern Sardinian Corpus we have already achieved several objectives, summarized as follows:

- we established an annotation schema for Early Modern Sardinian notary deeds which allows all the relevant external information to be preserved;
- we have inserted our documents into Teitok which, not onlymakes it easy to use for different kinds of users, but also permits linguistic searches to be performed with standard corpus tools;
- since the documents will be freely downloadable, they can be re-used for other searches (for instance, personalized queries through XPath, or through other platforms like TXM).

The first studies on the languages used in the documents show the importance of being able to combine linguistic information and extralinguistic information and of considering texts in a multilingual perspective. For instance, we were able to confirm our idea that, some sections in our documents,such as the *completio* and the *subscriptiones*, are generally in Catalan while in others, like the *datatio*, Sardinian alternates with Latin. The use of Catalan and Latin thus seems to be correlated to more "formal" discourse moves and is used to add authority to the document. Moreover, since we also collected extralinguistic information, we were able to correlate linguistic phenomena with different levels of linguistic variation. For example, some of our documents show variants that mantain the original Latin consonant cluster *-pl-/-bl-* (as *complimentu* and *obligare*) while others have the innovative form in *-pr-/-br-* (like *comprimentu* and *obrigare*). Our corpus allowed us to see that the forms in *pr/br* tend to appear in documents which also show some other "lower" phenomena like the methathesis of *-r-* (as in *frimadu* for *firmadu*) and it can consequently be hypothesized that both correlate with diastratic variation.

Future work will focus on two points:

- at a more general level we need to develop the structural coding of more complex documents such as court trials, which arrived in the form of a summary report containing different documents such as letters, trial witness statements, and attestations relative to the delivery of convocations;

- some issues on normalization and lemmatization are still to be discussed, especially if we want to place our corpus in a diachronic and ambitious perspective as one of the steps for the construction of a diachronic corpus of Sardinian.

It goes without saying that, only by increasing the size of our corpus, can we confirm the already noticed tendencies and give a more detailed picture of the multilingual practices in Modern Sardinia.

Acknowledgements

The authors wish to thank Maarten Janssen for his wonderful support on Teitok.

References

Maria Eugenia Cadeddu. 2013. Reperti di plurilinguismo nell'Italia spagnola (sec.XVI-XVII). In T. Krefeld, W. Oesterreicher, and V. Schwägerl-Melchior, editors, *Scritture di una società plurilingue: note sugli atti parlamentari sardi di epoca moderna*, Berlin-Boston: DeGruyter, pages 13–26.

Stefan Evert and Andrew Hardie. 2011. Twenty-first century corpus workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference, Birmingham, UK*.

Maarten Janssen. 2016. Teitok: Text-faithful annotated corpora. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Maarten Janssen, Josep Ausensi, and Josep M. Fontana. 2017. Improving pos tagging in old spanish using teitok. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Linköping University Electronic Press, Linköpings universitet, 133, pages 2–6.

Gard B. Jenset and Barbara McGillivray. 2017. *Quantitative Historical Linguistics*. Oxford: Oxford University Press.

Joanna Kopaczyk. 2018. Administrative multilingualism on the page in early modern Poland. In Päivi Pahta, Janne Skaffari, and Laura Wright, editors, *Multilingual practices in language history*, De Gruyter, Berlin-Boston.

Roger Lass. 2004. Ut custodiant litteras. Editions, Corpora and Witnesshood. In M. Dossena and R. Lass, editors, *Methods and data in English historical dialectology*, Bern: Peter Lang, pages 21–48.

Giuseppe Marci. 2006. *In presenza di tutte le lingue del mondo. Letteratura sarda*. Cagliari: CUEC.

Päivi Pahta, Janne Skaffari, and Laura Wright. 2018. From historical code-switching to multilingual practices in the past. In Päivi Pahta, Janne Skaffari, and Laura Wright, editors, *Multilingual practices in language history)*, De Gruyter, Berlin-Boston.

Nicoletta Puddu. 2015. Costituzione del Sardinian Medieval Corpus: prime proposte per la codifica e l'annotazione. In Piera Molinelli and Ignazio Putzu, editors, *Modelli epistemologici, metodologie della ricerca e qualità del dato. Dalla linguistica storica alla sociolinguistica storica*, Franco Angeli, pages 282–299.

Nicoletta Puddu and Achim Stein. 2018. Word-level and higher level annotation of the sardinian medieval corpus. In Andrew U. Frank, Christine Ivanovic, Francesco Mambrini, Marco Passarotti, and Caroline Sporleder, editors, *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities)*. Gerastree Proceedings, Vienna.

Suzanne Romaine. 1992. *Socio-historical Linguistics. Its Status and Methodology*. Cambridge: Cambridge University Press.

Maurizio Virdis. 2017. Superstrato spagnolo. In M. Dossena and R. Lass, editors, *Manuale di linguistica sarda*, Berlin: DeGruyter, pages 168–183.