Word-level and higher level annotation of the Sardinian Medieval Corpus

Nicoletta Puddu & Achim Stein

Università di Cagliari & Universität Stuttgart

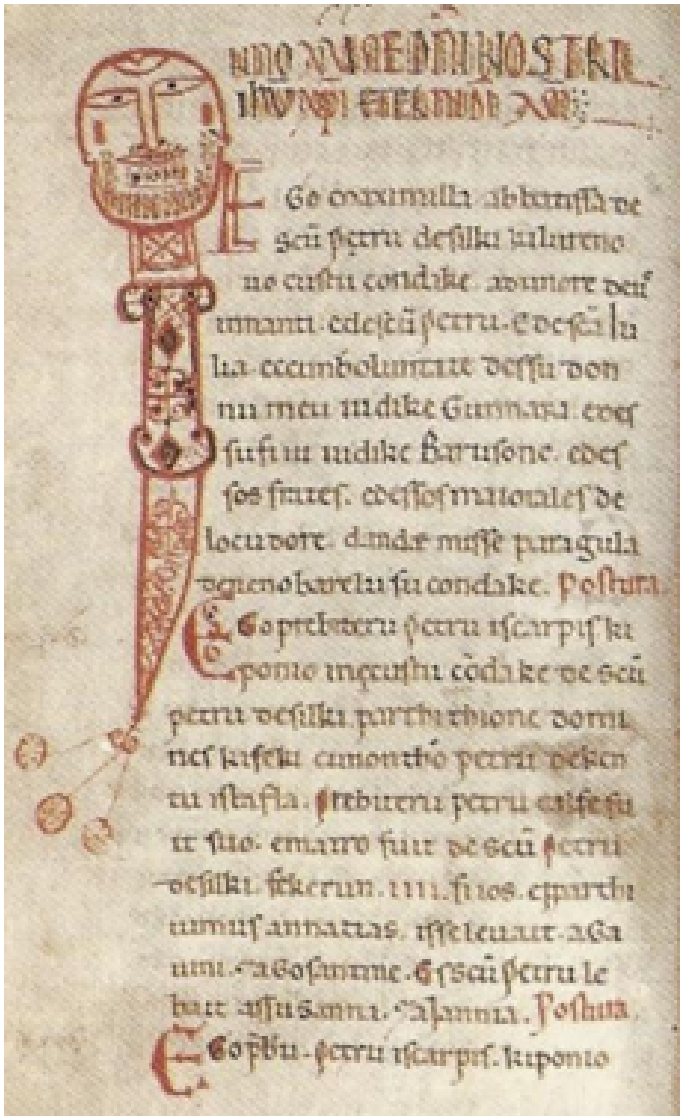
Workshop on Corpus-based Research in the Humanities (CRH)
Vienna, January 25-26, 2018

Medieval Sardinian

- Sardinian was the official language of the "Giudicati" (independent kingdoms of Sardinia) while Latin was used in international relationships.
- Italian spread gradually in commercial interactions.
- In 1392 Sardinia was conquered by the Aragonese kingdom losing its independence. Catalan gradually substituted Sardinian as the official language.

The *condaghes*

- The *condaghes* are documents recording acts of donation, transactions and income of churches and monasteries, often containing transcriptions of legal disputes called *kertos*
- The *Condaghe of Saint Nicholas of Trullas* registers acts from the first half of the 12th century to the second half of the 13th century



Condaghe di San Pietro di Silki (1065-1180)

The *kertos*

- An example of a *kertu* from the "condaghe di San Nicola di Trullas" is given in (1) to (4).

(1) osca mandaiti=mi donna Muscu et ego andai=bi=li ad Amendalas.
then sent=me lady Muscu and I went=there=to her to Amendalas
Then lady Muscu had me called and I went to her to Amendalas.

(2) et issa narraiti=mi ca: donnu, cussu kertu ki amus unpare, pr'onore de
and she told=me that lord this controversy that have together for the honour of
Sanctu Nichola, si vos placet, campaniemusi=lu kene iura vestra nen mea .
saint Nicolas if to you pleases settle=it without oaths yours nor mine .
And she told me "Milord, for the honour of Saint Nicholas, let's settle this controversy, if you like, without oaths, neither mine, nor yours."

(3) et ego narai=li ca: donna, in benedictione! .
and I told=her that: lady in blessing
And I told her: "Milady, bless you"

(4) et issa narraiti=mi: Prossa canpania date=mi su latus de Istefane Pira et
and she told=me for the settlement give=me the half of I. Pira and
levate vois sas filias: Furata intrega et anbos sos filios, et latus de Maria.
take you the daughters Furata complete and both the sons and half of Maria
And she told me: "For the settlement, give me the half of Istefane Pira and take his daughters: the whole Furata and her sons and the half of Mary"

(4, SNT, 164; cf. Merci 1992)

Coding and annotating the SMC

Our texts present all the difficulties common to historical corpora.

- A high degree of graphical variation
- A limited and partial documentation
- A high degree of linguistic heterogenity

Coding and annotating the SMC

Text encoding and manual annotation (Lass 2004)

1. Maximal information preservation
2. No irreversible editorial intervention
3. Maximal flexibility

Coding and annotating the SMC

(5) cu<expan rend="italic">n</expan> corte et
cu<expan rend="italic">n</expan>
o<expan rend="italic">mn</expan>ia
ca<expan rend="italic">n</expan>tu vi aveat

CSNT, 281, 3

oia

Latin: *omnia*

Sardinian: *onnia*

Coding and annotating the SMC

```
<person xml:id="a30" sex="1">
<persName>Petrus</persName>
<forename>Petrus</forename>
</surname>
<occupation>prior</occupation>
</person>
<person xml:id="a050" sex="1">
<persName>
<forename>Petru</forename>
<surname>de Thori</surname>
</occupation>
</persName>
```

Coding and annotating the *kertos*

(6) Petru de Thori sued me because «Why are you taking Sardinia away, who is mine?»

```
<p n="1">Certait mecu Petru de Thori ka.
<q who="#a050">Procetiu mi la levas a Sardinia, ca es mea?
</q></p>
```

Petru de Thori, person a050, as listed in the TEI header

Goal

- The Digitalization of the SMC is work in progress.
- No quantitative evaluation of the tools for now.

Our goal:

Show how digitalization and annotation can be linked in a productive process that leads to a faster and more coherent creation of a linguistically relevant resource, even with medieval texts.

- Annotation on two levels:
  - lexical: part-of-speech (POS) tagging
  - syntactic: dependency parsing

Issues

- Multilingualism
  - Sardinian alternates with Latin.
  - Dialectal variation of Sardinian.
- Decisions:
  - We don't want Latin words in our lemma list.
  - We need to process Latin words correctly for the syntactic analysis.

Clitic pronouns: a problem for tokenization

Clitic pronouns ("clitics"):

Pronouns that attach to lexical forms ("hosts"), mostly verbs

- proclitics: pre-lexical attachment
- enclitics: post-lexical attachment

(7) Comporaili a cComita de Bosobe et assos frates su saltu de serra
I.bought him.DAT to Comita de Bosobe and to.the brothers the land of serra
de Iugale
de Iugale
'I bought the land in Serra de Iugale from Comita de Bosobe and his brothers'. (SNDT 17, quoted from Wolfe 2015)

- Issues for automatic annotation:
  - Split clitics from the host (semi-automatically, using a script).
  - Tokenize clitics as separate entities on word and syntactic level.

Clitic pronouns: a problem for tokenization

(8)

```
<w ... xml:id="w_621" rend="H:manu-yes3">partindo</w>
<w ... xml:id="w_622" rend="aggl">llu</w>
<w ... xml:id="w_623">ladus</w>
<w ... xml:id="w_624">a</w>
<w ... xml:id="w_625">pare</w>
<w ... xml:id="w_626">cun</w>
<w ... xml:id="w_627">clesia</w>
<w ... xml:id="w_628">,</w>
<w ... xml:id="w_629">cum</w>
<w ... xml:id="w_630">serbos</w>
<w ... xml:id="w_631">et</w>
<w ... xml:id="w_632">cun</w>
<w ... xml:id="w_633" rend="H:manu-stop-cont">anchillas</w>
```

temporary log tag: manual split, 3 chars

mark-up for detached clitic (*partindo-llu*)

temporary log tag: add to stop list, continue (no splitting)

Word-level annotation

- Tests with several POS taggers
- *TreeTagger*: slightly less accurate, but controllable lexicon-based lemmatization.
- First step:
  - Training of *TreeTagger* (Schmid, 1994) on *Trullas*: 30.000 words, manually pre-annotated.
  - Accuracy score: 94.6%
- Second step:
  - Application of the trained model to the *Condaghe di Santa Maria di Bonarcado* (Virdis, 2002)
  - Unknown words were added to the tagger lexicon (3705 types, 10429 tokens).
- Iterative improvement of the POS tagging.

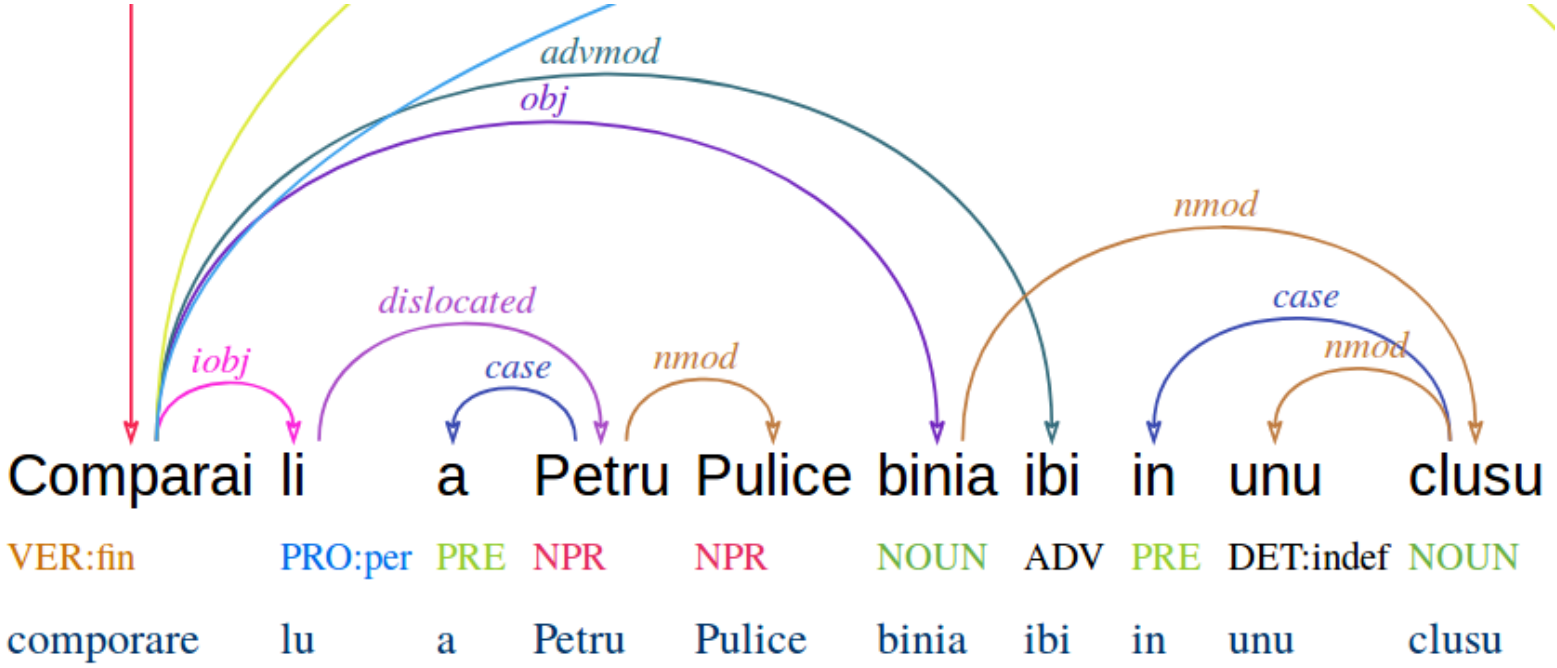
Syntactic parsing

- *Condaghes*: few and rather fixed word order patterns
- Manual dependency annotation
  - using the grammatical principles of the Old French *SRCMF* corpus[1] (Prévost & Stein, 2013)
  - adapted to (most) principles of Universal Dependencies[2]
- Parser training using the joint transition-based parser (JTP) of *mate tools* (Bohnet et al., 2013)
  - on 100 sentences of *Trullas*
  - iterative improvement by correcting and re-training
- Application of the trained model to *Libellus Iudicum Turritanorum* (Sanna, 1957)
  - late copy of a chronicle dating back presumably to the 13th c.
  - larger variety of word orders than *Trullas*

Syntactic parsing

- No gold standard ⇒ no accuracy scores
- The parser copes well with both the verb-initial structures of the *Condaghes* (*Trullas*) and the more variable verb order of *Libellus*.
- It is not necessary to train separate parser models for different text types (cf. Stein 2014).

Syntactic parsing of dislocations



'I bought from P. Pulice a vinyard there in a valley' (graph drawn by *Arborator* (Gerdes, 2013))

Fig. 1: Dislocated elements and resumptive clitic pronouns

- The verb governs the clitic *li* as an indirect object (*iobj*).
- The clitic in turn governs its referent, the prepositional phrase *a Petru Pulice* (*dislocated*).
- Dependencies clitic–host and clitic–referent are predicted well.

XML representation of the annotation levels

- Projection of the tagger/parser output into the original TEI corpus file
- No TEI encoding standard for syntactic dependencies
  - grammatical information in the TEI guidelines only refers to morpho-syntactic features of lexical items
- *De facto* standard for parsing: CoNLL shared task format (http://www.conll.org/)
  - CoNLL columns *ID*, *HEAD*, *DEPREL*
    $\rightarrow$ XML attributes *conllid, head, deprel*)

XML representation of the annotation levels

(9)
```
<w...conllid="1" head="0" deprel="root" xml:id="w_40">Comparai</w>
<w...conllid="2" head="1" deprel="iobj" xml:id="w_41" rend="aggl">li</w>
<w...conllid="3" head="4" deprel="case" xml:id="w_42">a</w>
<w...conllid="4" head="2" deprel="dislocated" xml:id="w_43">Petru</w>
<w...conllid="5" head="4" deprel="nmod" xml:id="w_44">Pulice</w>
<w...conllid="6" head="1" deprel="obj" xml:id="w_45">binia</w>
<w...conllid="7" head="1" deprel="advmod" xml:id="w_46">ibi</w>
<w...conllid="8" head="10" deprel="case" xml:id="w_47">in</w>
<w...conllid="9" head="10" deprel="nmod" xml:id="w_48">unu</w>
<w...conllid="10" head="6" deprel="nmod" xml:id="w_49">clusu</w>
```

- *head*, *word*: attachment of nodes
  - *Comparai* (*word=1*) is marked as the root node of the graph: head=0
  - The clitic *li* (*word=2*) dependens on the root: head=1
  - This relation is labelled as 'indirect object': iobj
- Advantage: easy to switch between the more philological TEI markup and the CoNLL format used for dependency parsing

Conclusion

- We discussed the challenges of annotating a Medieval language with strong graphical and syntactic variation.
- The interplay between manual annotation (POS and lemmata) and NLP tools for word-level and syntactic annotation helps to build a medieval database in an efficient way.
- We created the first models for automatic part-of-speech annotation and dependency parsing of Medieval Sardinian.
- All added information is in TEI-XML compatible format.

Reference for the SMC:

Puddu, Nicoletta. 2015. "Costituzione del Sardinian Medieval Corpus: prime proposte per la codifica e l'annotazione."
In Piera Molinelli & Ignazio Putzu (eds.), *Modelli epistemologici, metodologie della ricerca e qualità del dato. Dalla linguistica storica alla sociolinguistica storica*, Milano: Edizioni FrancoAngeli, p. 282–299

Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter & Jan Hajic. 2013. Joint Morphological and Syntactic Analysis for Richly Inflected Languages. In *TACL 1*, 415–428.

Gerdes, Kim. 2013. Collaborative dependency annotation. *Dependency Linguistics* 88. URL http://www.aclweb.org/anthology/W13-3711.

Merci, Paolo (ed.). 1992. *Il condaghe di San Nicola di Trullas*. Sassari: Delfino.

Prévost, Sophie & Achim Stein (eds.). 2013. *Syntactic Reference Corpus of Medieval French (SRCMF)*. Lyon/Stuttgart: ENS de Lyon; Lattice, Paris; Universität Stuttgart. URL http://srcmf.org.

Puddu, Nicoletta. 2015. Costituzione del Sardinian Medieval Corpus: prime proposte per la codifica e l'annotazione. In Piera Molinelli & Ignazio Putzu (eds.), *Modelli epistemologici, metodologie della ricerca e qualità del dato. Dalla linguistica storica alla sociolinguistica storica*, 282–299. Milano: Edizioni FrancoAngeli.

Sanna, A. 1957. *Libellus Iudicum Turritanorum*. Sassari: S'Ischiglia.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging using Decision Trees. In Daniel Jones (ed.), *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP'94), Manchester September 1994*, 44–49. Manchester: UMIST.

Stein, Achim. 2014. Parsing Heterogeneous Corpora with a Rich Dependency Grammar. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 26.-31.5.2014*, Reykjavik, Iceland: European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/239_Paper.pdf.

Virdis, Maurizio. 2002. *Il Condaghe di Santa Maria di Bonarcado*. Sassari: Centro Studi Filologici Sardi.

Wolfe, Sam. 2015. The Old Sardinian Condaghes: A Syntactic Study. *Transactions of the Philological Society* 113(2). 137–205.