D(h)ante: A New Set of Tools for XIII Century Italian

Angelo Basile, Federico Sangati
Digital Humanities Joint Project - Fondazione Bruno Kessler -
via Sommarive 18, Trento(TN), Italy
angelo.basile@studenti.unitn.it, federico.sangati@gmail.com

Abstract

In this paper we describe 1) the process of converting a corpus of Dante Alighieri from a TEI XML format in to a pseudo-CoNLL format; 2) how a pos-tagger trained on modern Italian performs on Dante's Italian 3) the performances of two different pos-taggers trained on the given corpus. We are making our conversion scripts and models available to the community. The two other models trained on the corpus performs reasonably well. The tool used for the conversion process might turn useful for bridging the gap between traditional digital humanities and modern NLP applications since the TEI original format is not usually suitable for being processed with standard NLP tools. We believe our work will serve both communities: the DH community will be able to tag new documents and the NLP world will have an easier way in converting existing documents to a standardized machine-readable format.

Keywords: digital humanities, old italian, pos-tagging

1. Introduction

It is commonly known that Computational Linguistics (CL) originated in the 1950s with the efforts to use computers to automatically translate texts from foreign languages (Camburn, 2013). Not everyone is aware, however, that in the same period of time Roberto Busa[1] gave rise to the *humanities branch* of CL now known as Digital Humanities (DH). Since then the two branches have been developed rather independently. The DH community has focused on deriving standardized methods for working with digitized *literary* work (Burnard, 2014), while the NLP branch concentrated on computational models of language to be used in more general language tasks. Only in the last decade, the two fields started to converge (Pennacchiotti and Zanzotto, 2008).

In this work we attempt to contribute to the latter trend. More specifically, we build upon the linguistic annotation work of (Tavoni, 2010) to develop a Part of Speech Tagger (PoS) of XIII century Italian language. The objective of the work is twofold: (1) to provide the NLP community with a tool to perform automatic processing of ancient text and (2) to provide the literature community with more powerful tools for simplifying the annotation process and performing more advanced data analysis.

The rest of the paper is organized as follows: in section 2 we describe the corpus being used, in section 3 the transformation we had to perform to convert the corpus into a more consistent format suitable for performing NLP tasks. Finally, section 4 comprises the description of two PoS tagger models we have built and the experiments we have run to compare their performance.

2. Corpus Description

Documents The corpus we have used consists of the following documents, all belonging to Dante Alighieri: *Divina Commedia (Inferno, Purgatorio, Paradiso), Convivio, Detto d'amore, Rime, Vita Nuova, Fiore*. It is a subset of the DanteSearch corpus (Tavoni, 2005)[2].

It is not a complete corpus of the author: five documents are missing. Only one document is in prose form, one is in a mixed form and all the others are written in verse. The language is mostly Italian; some documents contain short snippets of Latin language: these tokens are marked as foreign words and all together add up to 312.

Annotation The annotation task has been performed manually by Italian native speakers (PhD/Master students in Italian literature).

The corpus is encoded in a TEI 2 XML format. Word tokens are encapsulated in a LM tag.

```
<LM lemma="il" catg="rdms">Nel</LM>
```

Listing 1: Example of LM tags.

An additional LM1 tag is used for those cases in which a single form can be mapped to two different lemmas.

```
<LM1>
<LM lemma="il" catg="rdms">nel</LM>
<LM lemma="in" catg="epaksl">nel</LM>
</LM1>
```

Listing 2: Example of LM1 tags.

Each word token is marked with a POS tag and lemmatized. Punctuation is not tokenized and thus not tagged: in some cases it is glued together with the preceding word, while in others it is left outside the XML word tag:

[1] He was an Italian Jesuit priest who in 1949 met with Thomas J. Watson, the founder of IBM, and was able to persuade him to sponsor the Index Thomisticus, a tool for performing text searches within the massive corpus of Aquinas's works (Busa, 1974 1980; David Bamman and Crane, 2008)

[2] See http://cibit.humnet.unipi.it/ for a description of the critical editions of the texts upon which the corpus documents are based on.

```
<LM [...]>alto</LM>,
<LM [...]>e</LM>
```

Listing 3: Example of punctuation outside tags.

```
<LM [...]>quivi</LM>
<LM [...]>regge;</LM>
```

Listing 4: Example of punctuation inside tags.

This (non)-tagging schema does not contain end-of-sentence marks. Multi-word expressions and proper nouns are not tokenized:

```
<LM lemma="Guglielmo Borsiere" catg="np">
Guiglielmo Borsiere
</LM>

<LM lemma="quando che sia" catg="bl">
quando che sia
</LM>
```

Listing 5: Example of multiple tokens inside one element.

The tagset cardinality amounts to 2244: the count is high because the schema includes fine-grained morphological information.

Statistics Table 1 highlights some basic statistics about the corpus. We used the LM tags to extract these statistics. The lexical richness is computed with a simple formula:

$$\frac{n(types)}{n(tokens)} \quad (1)$$

The transformed corpus, as shown in the next section, allows for refined statistics and more details.

3. Corpus Transformation

While perfect for some tasks (i.e. fine-grained search, manual lookup), the tagset and the format used in the original corpus are not fit for other NLP tasks. The partial tokenization and the rich tagset are not appropriate for performing advanced statistics (e.g., stylometry) and also are not appropriate for direct use with standard NLP tools.
To address these problems we performed the following transformations:

1. full tokenization
2. 1-word–2-lemma nodes merging
3. punctuation tagging
4. sentence segmentation
5. tagset conversion to ELRA set

Except for the last one, all the operations have been performed via XSL transformation. The XSL code used to convert this specific corpus can be used with small changes to handle different documents encoded in the TEI format. The code is open source and can be found at `https://github.com/anbasile/DH`.
Since punctuation was not tagged, in the original corpus there were no sentence boundaries. After properly tagging the punctuation, we decided to split sentences on periods only: the corpus is extremely rich in dialogues, so using exclamation marks and questions marks as well would have produced incorrect results: indirect speech is often used to introduce direct speech and these often terminates with exclamation or question marks; splitting the sentence in a naive way on these marks too would leave the rest (indirect speech) in a meaningless form.
The tagset conversion task has been completed in a *Python* environment: we defined a dictionary that maps each new tag to the original via regular expression. The following is an excerpt of the dictionary, showing how singular and plural determined articles have been handled:

```
post_conversion = {
    '^r..p$':'RP',
    '^r..s$':'RS',
```

This method can be seen as a middle way between a full manual conversion (which can be extremely time consuming) and statistical/automatic tagset conversion method. For this corpus a small amount of noise is introduced since the tag categorization conflicts for some elements, but our method allows for certain flexibility in the translation. We used the Morph-it! lexicon (Zanchetta and Baroni, 2005) to resolve the conflicts. A working description of the ELRA tagset can be found at `http://hlt-services2.fbk.eu/textpro/?p=89`.
The output of the transformation is a pseudo-CoNLL format. See Table 2 for a sample output taken from the *incipit* of *Purgatorio*.
The tagset conversion process will need additional work in the future: the ELRA set was choosen only because it is the one used by TextPro[3] and since we wanted to evaluate its performance on this corpus we were forced to this choice. It remains to be investigated if the following could be a better solution: converting both the original corpus tagset and the ELRA tagset to a middle layer, namely the Universal Pos Tagset.

4. Experiments

As a first experiment, we have assessed how a state-of-the-art PoS tagger trained on modern Italian would perform on the XIII century text of the DanteSearch corpus. For this we have employed the PoS tagger of the TextPro NLP Suite (Pianta and Zanoli, 2009; Pianta et al., 2008).
Next, we have trained two PoS taggers on the converted corpus: TreeTagger and the Stanford tagger using standard settings.[4]

TreeTagger We have employed TreeTagger version 3.2 (Schmid, 1995; Schmid, 1997).[5]

[3] TextPro is the one tagger trained on contemporary Italian.
[4] Different settings may yield slightly different results.
[5] We have used a standard context length of 2 and the following list of closed-classed tags: XPS, XPW, XPB, XPO, RS, RP, C, CCHE, CCHI.

doc	tokens	types	lex. richness
convivio.xml	73457	6826	0.09
dettodamore.xml	2503	766	0.31
fiore.xml	23698	4420	0.19
inferno.xml	34280	6704	0.19
paradiso.xml	33717	6339	0.19
purgatorio.xml	34146	6591	0.19
rime.xml	12102	2733	0.22
vitanuova.xml	18988	3004	0.16
	232891	20562	0.09

Table 1: Corpus statistics

id	token	lemma	pos	mwe	eos	sentenceid
0	Per	per	epsf	0	0	1
1	correr	correrre	vta2fp	0	0	1
2	miglior	migliore	a2fp	0	0	1

Table 2: Sample output

Stanford POS tagger We have employed the MaxEnt Stanford Tagger part of the CoreNLP version 3.5.2 (Toutanova and Manning, 2000; Toutanova et al., 2003).[6]
For the training and test set selection we shuffled the order of all the sentences from the corpus and then we divided this shuffled set in three parts: 80% for the training set, 10% for the development set and 10% for the test set.

4.1. Results

Table 3 shows the overall results of the PoS taggers. The two models which were trained on the XIII century corpus are clearly outperforming the baseline model trained on modern Italian. Among the two best performing models, the Stanford tagger is the one with the highest accuracy. The result of TextPro is not unexpected: it is known that Dante's Italian contains from 65% to 70% of the words used in currently spoken Italian.
For a more detailed comparison between the TreeTagger and the Stanford tagger, we have produced two *confusability matrices* in figure 1 (see caption for more explanation).

Model	Accuracy
TextPro*	0.72
TreeTagger	0.90
Stanford	0.92

Table 3: Overall accuracy of the three PoS taggers. *TextPro has been trained on modern Italian.

5. Conclusions

In this paper we described the process of transforming the annotated corpus of DanteSearch provided by (Tavoni, 2005) into a format which is more suitable for developing NLP applications.

[6]We have used the 'bidirectional5words' architecture and the same list of closed-class tags as in the Stanford tagger.

We have assessed how an out-of-the-box PoS tagger (TextPro) trained on modern Italian performs on this corpus, and showed how state-of-the-art PoS taggers (TreeTaggers and Stanford) properly trained on this resource largely outperform the previous model.
We are going to release the tool behind the transformation procedure (TEI2CONLL) which could assist other researchers who wish to convert TEI encoded resources into a format which is more suitable for NLP analysis.
Additionally, we are going to release the tagger models to the DH community at large, which can be used to simplify the process of annotating additional material belonging to Dante and other XIII century Italian poets.
We believe that these tools can be a good starting point for the construction of a usable pipeline for handling old documents: digital humanities is rapidly growing field and there will be the need for these tools.

5.1. Future work

Some points will require additional work in the future. First, we need to try using a middle layer for the tagging conversion process. Second, we are going to analyze the errors of the taggers and see where the one trained on modern Italian fails. Third, sentence segmentation needs a more robust approach.

6. Acknowledgement

We would like to thank Mirko Tavoni and his research group for having provided the original corpus.

7. Bibliographical References

Burnard, L. (2014). *What is the Text Encoding Initiative?: How to Add Intelligent Markup to Digital Resources*. Encyclopédie numérique. OpenEdition Press.

Busa, R. (1974–1980). *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis*

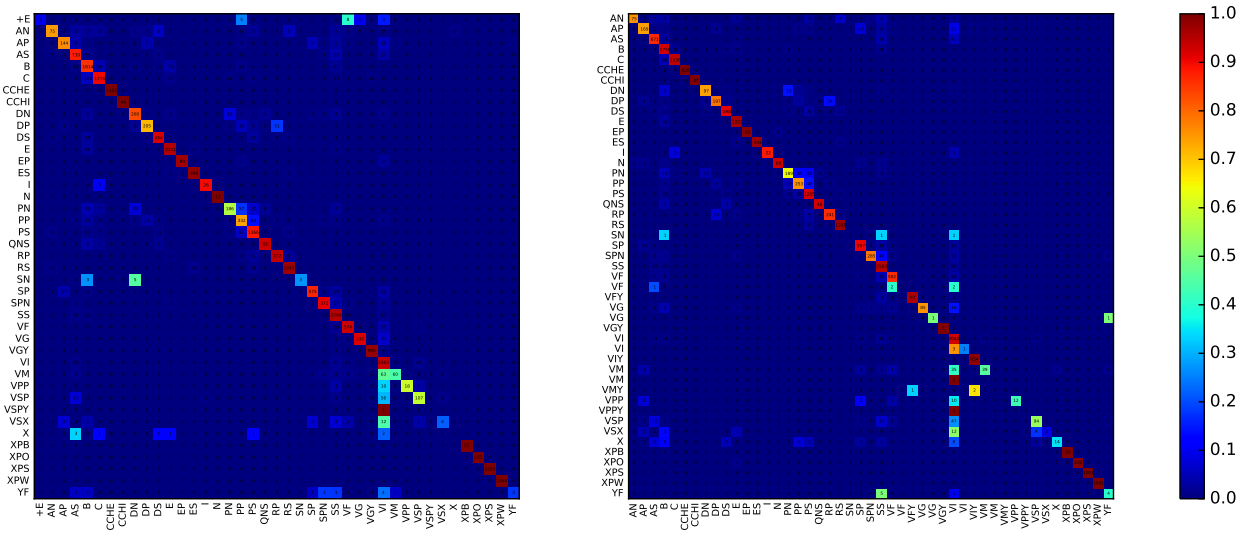

Figure 1: PoS confusability matrices for the TreeTagger (left) and the Stanford PoS-tagger (right). For every pair $\langle X, Y \rangle$ of PoS (X for gold PoS in rows, Y for test PoS in columns) it shows the number of words whose PoS is X in the gold file and Y in the test file. Colors are related to the precision percentage.

referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SJ. Stuttgart - Bad Cannstatt: Frommann - Holzboog.

Camburn, R. (2013). A short history of computational linguistics. California State University, Fresno, CA, December.

David Bamman, Marco Passarotti, R. B. and Crane, G. (2008). The annotation guidelines of the latin dependency treebank and index thomisticus treebank: the treatment of some specific syntactic constructions in latin. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Pennacchiotti, M. and Zanzotto, F. (2008). Natural language processing across time: An empirical investigation on italian. In Bengt Nordström et al., editors, *Advances in Natural Language Processing*, volume 5221 of *Lecture Notes in Computer Science*, pages 371–382. Springer Berlin Heidelberg.

Pianta, E. and Zanoli, R. (2009). A multistage pos-tagger at the evalita 2009 pos-tagging task. In *Proceedings of the EVALITA 2009 Workshop on Evaluation of NLP Tools for Italian*, Reggio Emilia, Italy.

Pianta, E., Girardi, C., and Zanoli, R. (2008). The textpro tool suite. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer.

Schmid, H. (1997). Probabilistic Part-of-Speech Tagging Using Decision Trees. In Daniel B. Jones et al., editors, *New Methods in Language Processing*. Taylor & Francis, Inc., Bristol, PA, USA.

Tavoni, M., (2005). *Dante in lettura*, chapter Un nuovo strumento informatico per lo studio di Dante (con una proposta interpretativa per Inf. IV 69), pages 217–229. Longo editore, Ravenna.

Tavoni, M. (2010). Dantesearch.

Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

Zanchetta, E. and Baroni, M. (2005). Morph-it! a free corpus-based morphological resource for the italian language. *Corpus Linguistics 2005*, 1(1).