Natural Language Processing across time: an empirical investigation on Italian

Marco Pennacchiotti[1] and Fabio Massimo Zanzotto[2]

[1] Dept. of Computational Linguistics, Saarland University, Saarbrücken, Germany
pennacchiotti@coli.uni-sb.de,
[2] DISP, Universitá di Roma Tor Vergata, Roma, Italy
zanzotto@info.uniroma2.it

Abstract. In this paper, we study how existing natural language processing tools for Italian perform on ancient texts. The first goal is to understand to what extent such tools can be used "as they are" for the automatic analysis of old literary works. Indeed, while NLP tools for Italian achieve today good performance, it is not clear if they could be successfully used for the humanities, to support the critical study of historical works. Our analysis will show how tools' performance systematically vary across different time periods, and within literary movements. As a second goal, we want to verify whether or not simple customization methods can improve the tools performance over the old works.

1 Introduction

Natural Language Processing (NLP) tools for morphological and syntactic analysis guarantee today high standards in terms of performance and robustness, so that they can be successfully used in a wide range of applications. Yet, despite the large availability of electronic editions of old literary works in the context of the TEI initiative [1], and despite the potential benefits, few attempts have been made so far to adapt NLP tools to the field of humanities (e.g. [2, 3]), especially for Italian.

For example, researchers in philology mostly use simple keyword search in context (KWIC) and dictionary query engines working on human annotated material, where the NLP contribution is minimal. The application of deeper automatic linguistic techniques would be valuable to build more sophisticated and useful tools for philological and literary studies.

A natural question is then why not much effort has been spent so far in adapting NLP tools to ancient texts. This cannot be only explained by the skepticism of researchers in the humanities over NLP applications in general. In our view, the core issue is that the humanistic and NLP research areas have different objectives. NLP research aims to build language models that cover the *most frequent* phenomena of *contemporary* natural languages. On the contrary, the goal of humanistic studies is to discover and analyse *odd* phenomena of *historical* natural languages. This makes difficult to adapt existing NLP models and tools on the humanities, discouraging any effort in that direction.

Yet, a fruitful interaction between the two areas is possible. For example, researchers in the humanities could easily contribute in NLP whenever an odd phenomenon for an old grammar/lexicon is covered, by including its model in an existing NLP architecture. In turn, researchers in NLP could leverage these models, to customize the architecture to the particular language. In this framework, the first step to be done is to evaluate how far modern NLP machineries are from achieving good performance over ancient texts.

In this paper, we want to address the above issue, by studying the performance of basic NLP tools and resources (namely a standard dictionary, a morphological analyser and a part-of-speech tagger) over an historical language, and across different time periods and literary movements. Also, we want to verify whether or not simple customization techniques can help in achieving better tool performance over the ancient texts. We focus our attention on the Italian language, which for its lexical richness and its comparatively old story (the first forms of Italian date from the 10th century) represents an exemplar test set.

The paper is organised as follows. Section 2 describes work related to our investigation. In Section 3, we introduce the NLP resources we tested in our study, namely a syntactic parser and a dictionary of contemporary Italian. Section 4 describes the corpus of the ancient Italian texts we adopted as a test set. Section 5 reports and comments on the results of our experiments. Finally, in Section 6 we present possible future works to improve resource performance.

2 Related work

Studies on the portability of NLP tools and resources to historical languages is still fairly limited. Rocio et al. [3] applied a standard grammar for contemporary Portuguese to automatically parse Medieval Portuguese, along with a specific lexical analyser to extend the coverage of the lexicon on ancient words. The good results demonstrated that partial parsing of ancient Portuguese texts is feasible by relying on tools for contemporary languages. Britto et al. [4] used a PoS-tagger similar to the Brill tagger [5], trained on a corpus of 130,000 manually annotated words, to annotate 25 of the 52 texts (2 million words) contained in Tycho Brahe corpus for Historical Portuguese. They obtained a precision of 95.45%, four points below standard performance on contemporary languages. More recently, Moon and Baldridge [2] proposed a semi-automatic approach to induce a PoS-tagger for Middle English from material taken from Present-Day English, lavereging bilingual boostrapping techniques [6], achieving an accuracy on the low 80ies. All these works seem to indicate that, in order to successfully use NLP tools on ancient texts, major adaptations on the lexicon and at other levels are needed. We are here interested to verify this hypothesis, in particular for a Romance language as Italian.

More widely, as regards electronic resources for historical texts, great efforts have been spent in recent years to create manually annotated corpora with lemma, morphological and part-of-speech information. Major examples are the Penn-Helsinki Parsed Corpus of Middle English [7] containing 1.5 million words in the period 1100-1500; the Corpus of Early Modern English [8] of 1.8 million words extracted from text of different type in the period 1500-1710; and the York-Toronto-Helsinki Parsed Corpus of Old English Prose [9], consisting of 1.5 million words. Resources also exists for other languages, such as Early New High German and Latin.

As for Italian, the *Opera del Vocabolario Italiano*[3] is manually building TLIO [10], a dictionary for ancient Italian, from the 9th to the 14th century, currently containing 18,000 entries, extracted from 1,960 literary works. The dictionary builds up on a corpus of around 3.5 million lemmatized and morphologically annotated tokens. The related query software *Gattoweb* is today one of the most used automatic tools for Italian philology, allowing to carry out keyword searches and KWIC over the corpus. Finally, the Corpus Taurinense [11] consists of 21 Italian texts from the 13th century (260,000 tokens), lemmatized, morphologically annotated and PoS-tagged in the EAGLES/ISLE format.

Unfortunately, so far there have been no significant researches explicitly dedicated to the development or customization of NLP tools for historical Italian, apart from some exploratory attempts using statistical methods reported in [12], where machine learning techniques have been used to semantically annotate the Italian novel *"Gli indifferenti"*. NLP tools for contemporary Italian, including the Chaos parser [13] adopted in our study, are instead of great interest for the Italian NLP community, as demonstrated by Evalita[4], an evaluation campaign of Italian NLP tools.

3 Contemporary Italian dictionary and parser for NLP

In this section we describe the two resources that are evaluated on the historical Italian texts. In Section 3.1 we shortly introduce the dictionary, while in Section 3.2 we describe Chaos, a syntactic parser for Italian.

3.1 Dictionary

The recognition of syntactic and morphological classes of words is one of the most important tasks in sentence interpretation. Even apparently monolithic syntactic parsers (e.g., [14, 15]) perform part-of-speech tagging with specific models. For most part-of-speech taggers, one of the main problem is the treatment of unknown words. Morpho-syntactic lexicons are then one of the most important resources for the overall syntactic analysis. In romance languages morpho-syntactic lexicons are even more central. Unlike in English, in these languages each lemma may have a large number of forms. Simple stemming techniques cannot solve the problem, as forms of the same lemma can be very different. For example, the Italian form *"aiuterebbe"* (English: *"may help"*) and *"aiuta"* (*"helps"*) are two forms of the same lemma *"aiutare"* (*"to help"*).

[3] http://www.ovi.cnr.it/

[4] http://evalita.itc.it/

In our study we derive the dictionary from two morpho-syntactic lexicons included in the Chaos architecture: a manually-built generative morphology lexicon, and a corpus-induced lexicon.

Generative morphology lexicon

		#
roots	nouns	10,658
	verbs	5,104
	adjectives	5,288
forms	nouns	16,567
	verbs	84,610
	adjectives	11,644

Corpus-induced lexicon

	#
forms	12,132

Table 1. Italian morpho-syntatic lexicons

Generative morphology lexicon: This manually-built lexicon (see Table 1) includes ca. 22,000 lemmas: 10,658 nouns, 5,288 adjectives, 5,104 verbs, and other classes. Dictionary entries are organised as feature structures containing syntactic information and morphological information, specifying *gender*, *number*, *person*, *tense*, and *mood*. The generative lexicon produces 73,838 different forms, with an average ambiguity of 1.55. We included all the produced entries in our dictionary.

Corpus-induced lexicon: This lexicon has been built over a collection of articles of the Italian financial newspaper *Il Sole 24 Ore* and contains 12,132 words with an average ambiguity of 1.06. To find the interpretation of unknown words, we used a transformational part-of-speech tagger learner [5] producing 181 rules. Rules consist in a triggering condition and an emitted part-of-speech tag. For example, the rule *hassuf(ato)* → *VNP*, indicates that a word with the suffix *-ato* is likely to be a *VNP* – i.e. a verb in the past particle. Typical interpretations produced by this lexicon are impoverished compared to the generative morphology lexicon, as they include only the part-of-speech class information.

3.2 The Chaos Parser

Chaos [13] is a robust modular constituent-dependency parser for Italian, producing partial and possibly ambiguous syntactic analysis. In our study, we use the following module cascade: a *tokenizer*, matching words from character streams; a *yellow page look-up module* that matches named entities existing in catalogues; a *morphological analyser* that attaches (possibly ambiguous) syntactic categories and morphological interpretations to each word; a *named entity matcher* that recognizes complex named entities according to special purpose grammars; a rule-based *part-of-speech tagger*; a *PoS disambiguation module* that resolves potential conflicts among the results of the PoS tagger and the morphological analyser. Chaos also includes a *syntactic parser* based on modularisation and lexicalisation, whose study is not included in this paper.

We hereafter describe more in depth the morphological analyser M and the part-of-speech tagger POS, and their mutual interactions. Assume that $s = t_1 \ldots t_n$ is a tokenized sentence where t_i is a generic token. As a first step, Chaos activates the morphological analyser. The analyser is a function that works on tokens: given t, it produces the set of interpretations $M(t) = I$ that t has in the dictionary. These interpretations are considered unordered – i.e. the first interpretation is not necessarily the most plausible. Interpretations correspond to those described in the previous section. In a second step, the part-of-speech tagger is applied, by working on the whole sentence. Given a tokenized sentence $s = t_1 \ldots t_n$, it produces a sequence of PoS-tags $POS(s) = pos_1 ... pos_n$. At the end, for each token t_i there exists a unique interpretation pos_i. In a last step, the information from the PoS tagger and from the morphological analyser are harmonised. Given a token t_i in a sentence s, the preferred interpretation is the $l \in M(t_i)$ that is compliant with the PoS tag pos_i.

For example, consider the sentence *"the boat sinks"*. The morphological analyser produces the following interpretations for the token *sinks*: $M(sinks) = I_3 =$ {[lemma:sink,type:noun],[lemma:sink,type:verb]}. The PoS tagger, after analysing the overall sentence, assigns the PoS-tag $pos_3 = Verb$ to the token t_3. In the last step, the interpretations in I_3 are reduced to those compliant with pos_3, i.e. $I'_3 =$ {[lemma:sink,type:verb]}. However, the PoS tagger is not a word sense disambiguator. Homograph forms with the same PoS (e.g., the noun *bank* as *institution* or *river bank*) are not disambiguated at this stage.

4 A corpus of historical Italian texts

Our corpus of historical Italian texts is composed of 14 major Italian literary works, listed in Table 2.[5] We chose texts ranging across different time periods, literary movements, and genres, so that each of them could be somehow representative of a specific style. This allows to specifically evaluate the tools on movements, instead of generally on ancient Italian (though it must be clear that by studying a single piece it is possible to draw only indicative conclusions on a movement). The overall time range encompasses almost 700 year, starting with one of the first examples of written Italian (the *Rime* by the *Scuola Siciliana*), to a late work of the 19th century. It is here important to stress that by choosing such a different range of works, our goal is to give a very coarse-grained exploratory evaluation of the dictionary and the parser on different time periods, in order to investigate their applicability across time. However, we do not aim to draw final conclusions on the issue, which would require a much larger corpus of works. Also, we will not look in depth into philological explanations, as this should be left to an analysis requiring expertise in romance philology.

[5] All works are available in XML-TEI format at: *www.bibliotecaitaliana.it*

Author	Work	Year	Genre	Movement
Scuola Siciliana	Rime	1200	poetry	origins
Guido Cavalcanti	Rime	1275	poetry	Stilnovo
Giovanni Boccaccio	Decameron	1300	prose	
Dante Alighieri	Divina Commedia	1321	poetry	
Francesco Petrarca	Canzoniere	1348	poetry	
Lorenzo De'Medici	Canzoniere	1475	poetry	Renaissance
Ludovico Ariosto	Orlando Furioso	1532	poetry	Renaissance
Galileo Galilei	Dialogo sopra i due massimi sistemi	1632	prose	Baroque
G. Battista Basile	Le muse napolitane	1635	poetry	Baroque
Giuseppe Parini	Odi	1790	poetry	Illuminism
Vincenzo Monti	Poesie	1800	poetry	Neo-Classicism
Ugo Foscolo	Ultime lettere di Jacopo Ortis	1802	prose	Neo-Classicism
Vittorio Alfieri	Vita	1803	prose	Illuminism
Giovanni Verga	I Malavoglia	1881	prose	Verism

Table 2. Corpus of historical Italian texts adopted in the experiments.

5 Resource Evaluation

In this section, we present an empirical evaluation of the dictionary and the parser on the corpus of ancient Italian works presented in Section 4.

5.1 Experimental Setup

We implement two different evaluation tasks: one to check the dictionary coverage; one to evaluate the Chaos morphological analyser and part-of-speech tagger accuracies. For the first evaluation task, we extract from the XML files all tokens – i.e. lists of characters separated by space and punctuation. From the collected tokens we derive a list of unique words (tokens without repetitions). Finally, for each word we check if there is an entry in the dictionary. We evaluate the dictionary coverage as the number of unique words in a literary work which have at least an entry in the dictionary.

For the second evaluation task we build a gold standard dataset over which to compute Chaos accuracies. The gold standard consists of a random sample of 42 sentences (3 for each work), manually annotated by two human experts. The annotators were asked to select for each word in the sentence, the correct morphological and part-of-speech classes. In case of ambiguity, the class that fits the contexts was chosen. We computed inter-annotator agreement over 3 sentences randomly extracted from the corpus, in order to assess the reliability of the gold standard. We obtained a Kappa value agreement of 0.87 for morphology and 0.63 for part-of-speech, corresponding respectively to *almost perfect* and *substantial agreement*. The accuracy of the tools has been computed as the percentage of correct predictions over the gold standard.

In both tasks we compare the performance of the tools on the ancient texts with the performance obtained on a contemporary Italian text, namely an excerpt of the Italian newspaper *La Repubblica*. Such an evaluation will account for

Author	# Words	Dict. coverage		morpho accuracy	PoS accuracy
Scuola Siciliana	8,751	2,387	27,3%	0.48	0.54
Guido Cavalcanti	1,978	941	47,6%	0.66	0.73
Giovanni Boccaccio	18,785	6,736	35,8%	0.74	0.90
Dante Alighieri	12,610	5,136	40,7%	0.72	0.75
Francesco Petrarca	6,946	3,094	44,5%	0.69	0.71
Lorenzo De'Medici	3,805	2,068	54,3%	0.83	0.81
Ludovico Ariosto	20,120	6,889	34,2%	0.62	0.68
Galileo Galilei	13,027	6,674	51,2%	0.77	0.77
G. Battista Basile	5,411	1,077	19,9%	0.52	0.56
Giuseppe Parini	4,030	2,250	55,8%	0.73	0.79
Vincenzo Monti	5,050	2,625	52,0%	0.74	0.84
Ugo Foscolo	8,567	4,610	53,8%	0.69	0.76
Vittorio Alfieri	13,277	6,627	49,9%	0.72	0.77
Giovanni Verga	8,250	4,019	48,7%	0.68	0.68
La Repubblica	*16.520*	*10.328*	*62.5%*	*0.91*	*0.97*

Table 3. Coverage of the modern dictionary; accuracy of the morphological analyser; and accuracy of PoS-tagger over different works.

the portability of the tools – i.e. if there is a performance gap between historical and contemporary Italian.

5.2 Results

Results are reported in Table 3. Hereafter, we present both a *quantitative analysis* and a coarse-grained *qualitative study* of the results.

Quantitative Analysis. All ancient works show performance significantly lower than *La Repubblica*. Specifically, the average dictionary coverage on ancient works is 0.44, about 19% less than *La Repubblica*. The highest ancient work coverage is 0.56, still 7% less than the newspaper. Similar results are obtained for the Chaos' morphological analyser and PoS tagger, for which the average accuracies on ancient works are respectively 22% and 24% below *La Repubblica*.

The coverage of the dictionary is in general low. Regarding *La Repubblica*, this is due to the fact that the dictionary does not include proper nouns (which are clearly very common in newspapers) and modern foreign words (which are more and more present in contemporary Italian). Ancient texts present a much lower number of proper nouns, and no foreign words: in this case, the low performance are then completely due to the ancient lexicon.

The accuracy of the parser for *La Repubblica* is very high, somehow contrasting the above evidence on the dictionary. This indicates that the morphological analyser and the PoS-tagger successfully interact to find the correct analysis of words which are not present in the dictionary, by relying on the Pos-tagging rules encoded in the parser. For example the word *"logo"* is not present in the dictionary, but is still correctly recognized as a common noun by using contextual and morpho-derivational rules.

Parser accuracy over the ancient works is by contrast low, confirming the trend of the dictionary coverage. This suggests that PoS-tagging rules valid for contemporary Italian, cannot be straightforwardly applied to ancient texts. For example, given the fragment *"...d'amare domandassen pietanza"* (English: *"...would ask mercy for loving"*) (from *Rime della Scuola Siciliana*), the parser wrongly assigns the tag *common noun* to *"domandassen"*, because the word is not present in the dictionary, and then the parser backs-up to a PoS-tagging rule which states that a word following a transitive verb ("amare", *love*) must be a noun. Unfortunately, while this stands in general for contemporary Italian, it does not apply to many ancient examples.

Overall results support our initial claim that the dictionary and the Chaos parser for contemporary Italian are insufficient for the analysis of ancient texts, as there exists a significant gap in dictionary coverage between contemporary and ancient texts. PoS taggers cannot easily recover this gap. Default classification rules for unknown words learnt for contemporary Italian generally fail when used for historical Italian. We believe that this claim can be safely extended in general to all dictionaries and parsers for contemporary Italian. Indeed, our claim is in line with similar works for other historical languages. For example, Ricio et al. [3] show that the lexicons of Medieval and Contemporary Portuguese are substantially different, heavily impacting on parsing performance, despite the fact that the two grammars are quite similar. Also, Moon and Baldridge [2] prove that a straightforward application of parsers for contemporary English cannot be effective on Middle English, without applying strong adaptation strategies.

Diachronic/Synchronic analysis. We were somehow surprised that there seem to be no correlation between genres and coverage, suggesting that poetry is not more complex than prose, at least from a lexicon perspective. Also, there is no correlation between performance and literary movements (one could expect that works of the same movement have stylistic similarities and by consequence similar performance).

Yet, as expected, there is a fairly high correlation between the age of the work and tools' performance: ancient works tend to have lower coverage than more recent ones. An exception to this trend is *Le Muse Napolitane*, which overall shows the lowest coverage, 19.9%. This is due to the fact that it is written in a dialect, whose lexicon contains many words which are not standard Italian.

The most ancient work is the *Rime* by the *Scuola Siciliana*, a collection of poetries of different Sicilian authors from the 13th century. The language of the Scuola is characterized by a richness in both quality and quantity. Indeed, these poets used to mix and assimilate different regional dialects, Latin, Langue d'Oc and Langue d'Oil. The result is a lexicon rich of different influences and forms, which is very distant from contemporary Italian. The performance of the tools are in facts very low: the dictionary covers only 27% of words, while the parser has accuracies close to 50%.

The following work, the *Rime* by *Guido Cavalcanti*, signals an important change in the Italian language, which in that time was evolving strongly towards the contemporary form. In the late 13th century the literary movement called *Dolce Stil Nuovo* put the basis for modern Italian, by importing substantial changes in the language phonology and morphology. This is why the performance of the tools on the Scuola Siciliana are so low if compared to all later works. For example, in the Scuola Siciliana we still find expressions derived from Latin such as *“flamma”*, *“plaser”* and *“dovero”’*, which in Cavalcanti are already changed in the contemporary Italian variants *“fiamma”*, *“piacere”*, and *“dovro”’*.

As for the 14th century, *Dante Aligheri*, *Francesco Petrarca* and *Giovanni Boccaccio* represent the final achievement of a strong and stable Italian linguistic system. The performance obtained by our tools are here difficult to explain with a coarse-grained analysis. Yet, what is interesting to notice is that the dictionary coverage is lower for Boccaccio, probably because of the extensive use of dialectal expressions to describe everyday life, that are today disused. The higher coverage on Petrarca can be justified by the fact that he tended to use a short and stable lexicon, without any concession to dialectal expressions and neologisms. This also explains why coverage on Dante is in between the other two authors. Dante, especially in the *Divina Commedia*, tended to introduce many neologisms, which today are in part lost and in part accepted.

The variable performance on the works from the Renaissance period (*Orlando Furioso* by *Ludovico Ariosto* and *Canzoniere* by *Lorenzo de'Medici*), reveal that there is no high consistency in the tools' performance among works of the same movement. Both the dictionary and the parser show much better results for the latter work than for the former. This supports the observation that it is not possible to draw conclusions on the applicability of tools for automatic analysis even on works of the same literary movement. A closer look at the two works reveals that the lexicons used by the two authors are very different. From the one side, the *Orlando Furioso* (edition 1532) was mainly written to address the taste of the overall Italian audience, and its vocabulary is then very tied to the Italian language of the 16th century, which was highly influenced by the old Latin and Greek languages. Indeed, the poem contains common content words such *“haver”* (contemporary Italian *“avere”*, English *“to have”*) and *“huom”* (contemporary Italian *“uomo”*, English *“man”*) which today are disused, and which are a direct derivation of Latin (respectively *“habeo”* and *“homo”*). On the contrary, the vocabulary used by *Lorenzo De'Medici* appears closer to contemporary Italian. A possible explanation is that one of the goal of *Lorenzo* was to disseminate the use of the “Fiorentino” dialect, which is the base of contemporary Italian, and that was in contrast with the tendency of that period.

A similar observation stands for the Baroque period, where we find two works (the *Dialogo* by *Galileo* and the *Muse* by *Giovan Battista Basile*) which highly differ from a lexical perspective. Indeed, the latter is a collection of dialectal poetries, which are distant from standard Italian. The former is a prose work whose main intent was to disseminate a scientific theory to the largest audience possible. It then sticks to the spoken language of the 17th century, that

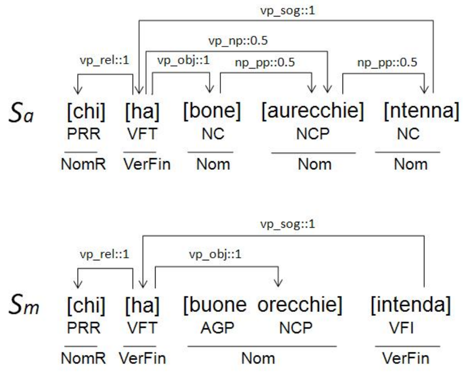

Fig. 1. Incorrect analysis of historical Italian (S_a) and correct analysis of contemporary Italian (S_m).

had consistently changed from the Italian of the 16th century, toward a form which is much more similar to contemporary Italian. This also explains why the performance over the *Dialogo* are much higher than over the *Orlando Furioso*.

More difficult is to follow the development of the Italian language in later periods, as different influences and movements start to merge together. A coherent analysis of the works from the Illuminism on (18th-19th century) would then require a deeper philological investigation, which is out of the scope of this work. We here only notice that in average the performance tend to increase with time, with the only exception of the latest work (*"I Malavoglia"*), whose performance are lower, due to the presence of many dialectal dialogs.

Reported results show that the accuracy of the tools is too low on historical Italian. This would strongly affect the overall syntactic analysis produced by the parser, because an incorrect PoS tagging has negative effects on the subsequent phase of the parsing process. For example, Fig. 1 reports the syntactic analysis for the ancient sentence S_a = *"chi ha bone auricchie 'ntenna"*, and the corresponding contemporary sentence S_m = *"chi ha buone orecchie intenda"* (English: *"if you have good ears try to listen"*). S_a and S_m have a similar grammatical structure. Yet, the fact that words are different leads to a completely different analysis. The analysis for sentence S_a is incorrect, while the analysis for S_m is *more* correct. The main problem in the analysis of S_a is the morpho-syntactic lexicon. Words such as *"bone"* (English: *"good"*), *"auricchie"* (English: *"ears"*), and *"ntenna"* (English: *"*try to listen"*) are not contained in the contemporary Italian lexicon. Two of them receive an incorrect part-of-speech tag: common noun (NC) instead of adjective (AGP) and common noun (NC) instead of verb (VFI). This is due to the fact that the PoS-tagger gives the noun tag (NC) as first hypothesis for unknown words. These type of errors completely mislead the syntactic analysis, as the figure shows.

6 Enhancing resources for ancient Italian

The previous section showed that the dictionary coverage and the accuracy of the morphological analyser and of the PoS-tagger are too low on historical Italian, thus strongly affecting the overall analysis produced by the parser. Hereafter, we propose some possible solutions to this problem, to improve the portability of the tools to historical Italian.

Manually build a lexicon for each period. This would be the most effective, but more costly solution, as the annotation should be carried out independently on every time period or literary movement having a definite lexicon. Previous similar experiences suggest that the effort is not feasible in a short time: for example, up to now and after more than 10 years of work, the TLIO dictionary contains lemmas only for the letters *A B C D E*, with a projected time to market of almost 50 years.

Leverage manually annotated corpora. A morpho-syntactically annotated corpus of historical Italian texts could be used to train reliable corpus-induced lexicons and PoS-taggers, as done in [4]. This solution is much more feasible than the previous one. Indeed, previous studies for contemporary and historical languages indicate that small sized corpora are sufficient to learn reliable NLP models. In this setting, active learning techniques could be highly valuable, as they allow to achieve a good compromise between accuracy and annotation effort.

Adapt existing models. A third viable solution consists in adapting current models for contemporary Italian, without going through a new learning phase and costly annotations. Rocio et al. [3] show for example that a simple **lexical analyzer** can turn a lexicon of contemporary Portuguese into a reliable lexicon for Medieval Portuguese, by using simple inflection rules. A similar approach could be used for historical Italian, lavereging adaptation rules for capturing morphological variations, such as : *-are* → *-ar*, to map *"amare"* and *"amar"*. Another adaptation strategy could rely on simple heuristic **string matching** functions. In facts, many contemporary words are small variations of ancient words – e.g. *"orecchio"* is adapted from *"auricchio"*. One of the best way of capturing these type of variations is using the Levensthein edit distance. We experimented such an approach on a dataset of 200 forms randomly extracted from the ancient Italian corpus (ca. 20 forms from each text). We obtained a coverage of 0.478 and an accuracy of 0.345.[6] Results indicate that string matching contributes to the task to some extent (it finds a good mapping for almost half ancient words), but at the cost of introducing potential noise (ancient words are often mapped to wrong entries in the dictionary).

As a future work, we will explore in particular the second and the third solutions. Also, we will measure the parser and dictionary performance over larger

[6] Given an ancient word w, we say that the Levensthein function *covers* the word if it finds the correct corresponding word(s) in the contemporary dictionary. Coverage is then defined as the percentage of ancient words which are covered, over the total number of words in the dataset. Accuracy is defined as the percentage of correct corresponding words over the dataset.

corpora, as the TLIO and the Corpus Taurinense, and investigate the performance of the whole parser chain, including a full syntactic analysis. Finally, we will activate collaborations with philologists, with the further goal of formalizng grammatical and lexical models for ancient Italian, and for studying a possible implementation of NLP-based tools for philological studies.

References

1. TEIconsortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium (2005)
2. Moon, T., Baldridge, J.: Part-of-speech tagging for middle English through alignment and projection of parallel diachronic texts. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). (2007) 390–399
3. Rocio, V., Alves, M.A., Lopes, J.G.P., Xavier, M.F., Vicente, G.: Automated creation of a partially syntactially annotated corpus of medieval portuguese using contemporary portuguese resources. In: Proceedings of the ATALA workshop on Treebanks, Paris, France (1999)
4. Britto, H., Finger, M., Galves, C. In: Computational and linguistic aspects of the construction of the Tycho Brahe Parsed Corpus of Historical Portuguese. Gunter Narr Verlag, Tubingen, Germany (2002)
5. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part of speech tagging. Computational Linguistics **21**(4) (1995)
6. Yarowsky, D., Ngai, G.: Inducing multilingual pos taggers and np bracketers via robust projection across aligned corpora. In: Proceedings of NAACL 01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies, Morristown, NJ 1–8
7. Kroch, A., Taylor, A.: Penn-helsinki parsed corpus of middle english. (2000)
8. Kroch, A., Santorini, B., Delfs, L.: Penn-helsinki parsed corpus of early modern english. (2004)
9. Taylor, A., Warner, A., Pintzuk, S., Beths, F.: The york-toronto-helsinki parsed corpus of old english prose. (2003)
10. Pollidori, V., Larson, P. In: Il Tesoro della Lingua Italiana delle Origini(TLIO): il progetto lessicografico e i suoi risultati attuali. Franco Cesati Editore, Dordrecht, Germany (2005)
11. Barbera, Manuel Barbera, C.M., Marello, C. In: Corpus Taurinense: italiano antico annotato in modo nuovo. Bulzoni Editore,Roma, Dordrecht, Germany (2003)
12. Basili, R., Di Stefano, A., Gigliucci, R., Moschitti, A., Pennacchiotti, M.: Automatic analysis and annotation of literary texts. In: Wokshop on Cultural Heritage, 9th AIIA Conference, Milan, Italy (2005)
13. Basili, R., Zanzotto, F.M.: Parsing engineering and empirical robustness. Natural Language Engineering **8/2-3** (2002)
14. Collins, M.: Head-driven statistical models for natural language parsing. Computational Linguistics **29**(4) (December 2003)
15. Charniak, C.: A maximum-entropy-inspired parser. In: NAACL, Seattle, Washington (2000)