Italian in the Trenches: Linguistic Annotation and Analysis of Texts of the Great War

Irene De Felice•, Felice Dell'Orletta◇, Giulia Venturi◇, Alessandro Lenci•, Simonetta Montemagni◇
• University of Pisa, CoLing Lab
irene_def@yahoo.it, alessandro.lenci@unipi.it
◇Istituto di Linguistica Computazionale "A. Zampolli", ItaliaNLP Lab
{felice.dellorletta, giulia.venturi, simonetta.montemagni}@ilc.cnr.it

Abstract

English. The paper illustrates the design and development of a textual corpus representative of the historical variants of Italian during the Great War, which was enriched with linguistic (lemmatization and pos-tagging) and meta-linguistic annotation. The corpus, after a manual revision of the linguistic annotation, was used for specializing existing NLP tools to process historical texts with promising results.

Italiano. *L'articolo illustra la progettazione e la costruzione di un corpus rappresentativo delle varietà di italiano in uso durante la prima Guerra Mondiale, annotato con dati linguistici (lemmatizzazione, analisi morfo-sintattica) e meta-linguistici. Il corpus, a seguito della revisione manuale dell'annotazione linguistica, è stato utilizzato per l'adattamento degli strumenti NLP esistenti, con risultati promettenti.*

1 Introduction

World War I (WWI) represents a crucial period in the history of Italian. In fact, De Mauro (1963) claimed that Italian as a national language was born in the trenches of the Great War. Since masses of men from different regions of the peninsula were forced to live together for months in the trenches and behind the lines, and were forced to use Italian as the main communicative medium instead of regional dialects, WWI produced a decisive step forward in the process leading to the linguistic unification of Italy.

The project *Voci della Grande Guerra* (*VGG*)[1] provides scholars with a new text corpus to investigate the structure and different varieties of Italian at the time of the Great War. The corpus includes a selection of texts representative of different textual genres and registers, including popular Italian. All texts have been automatically annotated with state-of-the-art NLP tools. A large subset of the corpus has then been manually corrected and enriched with metadata to classify a broad range of phenomena relevant for the study of the linguistic features of early XX century Italian. These characteristics make the *VGG* corpus unique in the very limited panorama of existing Italian historical corpora, among which it is worth pointing out the corpus dell'*Opera del Vocabolario Italiano* (OVI), the *DiaCORIS* corpus (Onelli *et al.*, 2006), the *MIDIA* corpus (Gaeta *et al.*, 2013), and the *Letteratura italiana Zanichelli* (LIZ). Moreover, the developed *VGG* corpus was used in an interesting case-study for the application and adaptation of NLP tools to process historical texts. The aim of this paper is to present the results of the annotation and linguistic analysis of the *VGG* corpus.

[1] http://www.vocidellagrandeguerra.it/

2 The Corpus *Voci della Grande Guerra*

The *VGG* corpus consists of 91 texts (ca. 1M tokens) that were written in Italian in the period of the World War I or shortly afterwards (most of them date back to the years 1915-1919). The texts were selected by historians and linguists in order to represent the 'polyphony'of the different voices of people who were affected by World War I. The corpus is balanced with respect to genre, style, and authors' profession: it collects discourses, reports and diaries of politicians and military chiefs; letters written by men and women, soldiers and civilians; literary works of intellectuals, poets, and philosophers; writings of journalists and lawyers.

Most documents existed only in printed form and were scanned and digitized with OCR tools. Once digitized, the documents were codified in the TEI-XML standard format. A significant part of the corpus of about 650,000 tokens, for which the

output of the OCR was manually corrected line–by–line with a correction tool specially designed for this purpose, constitutes our textual gold standard (Boschetti *et al.*, 2018).[2]

As a second step, documents were exported to be processed with NLP tools (cf. Section 3). Automatic linguistic annotation has been manually checked and corrected for more than 500,000 tokens for sentence splitting, tokenization, and lemmatization. For one fifth of this revised part of the corpus (ca. 103,000 tokens), manual revision has also targeted PoS tagging and morphological analysis. The revised documents belong to different genres and styles (see Table 1).

3 Method

The annotation methodology we have employed for the construction of the *VGG* corpus was articulated in the following steps:

1. the whole *VGG* corpus was automatically annotated using *UDPipe*, a trainable pipeline for tokenization, pos-tagging, lemmatization and dependency parsing with a transition based parser based on a non-recurrent neural network, with just one hidden layer, with locally normalized scores (Straka and Straková, 2017). The pipeline was trained on the Italian Universal Dependency Treebank (IUDT), version 2.0 (Bosco *et al.*, 2013);

2. the linguistic annotation of the *VGG* subcorpus reported in Table 1 was manually revised and whenever needed corrected. As fully described in Section 4, it was also enriched with metalinguistic information aimed to highlight features characterizing the variety of Italian used in the historical period considered. Correction was performed with a UD-compliant annotation tool specifically designed for the project.

3. the manually revised sub-corpus was used to retrain the automatic linguistic annotation pipeline in order to improve the performance of the automatic analysis tools.

4 Manual revision and meta-linguistic annotation

The first phase of automatic linguistic analysis performed on the *VGG* corpus (see Section 3) did not prove to be sufficient to achieve an accurate annotation of the texts, for two main reasons. First of all, the *VGG* corpus represents a historical variety of language, therefore obsolete forms are frequently found at both the lexical and the morphological level. Moreover, the documents feature an impressive degree of linguistic variation, which reflects the level of education of the writers, the style and register of texts (which in turn depend on their targeted purposes and audience, and on the particular social settings in which they were written), and the regional diversification of the Italian language in the years of the WWI (which was still largely permeated with dialectal features). Current NLP tools, trained on texts representative of standard, contemporary Italian (cf. Section 5), are not able to handle such a huge linguistic variation (see the performance reported in Table 2). Therefore, we performed a manual revision of the automatic annotation on a gold subsection of the corpus and enriched it with additional data, in order to retrain and improve the language model.

4.1 Manual revision

Automatic annotation was manually checked and corrected for more than 500k tokens for sentence splitting, tokenization, lemmatization, and partly also for PoS tagging and morphological analysis (cf. Table 1). This operation allowed us to individuate the most relevant features of the *VGG* corpus that pose critical difficulties to automatic annotation, as briefly illustrated in what follows.

Major issues with tokenization:

1. *Pronominal clitics attached to verbs*. Although pronominal clitics regularly attach to verbs in Italian under particular conditions, some combinations (e.g., *abbiti*, *siasi*) are very rare in contemporary Italian and linguistic tools often fail in segmenting and analyzing them correctly. Such forms were manually identified and splitted (*abbi+ti*, *sia+si*).

2. *Hyposegmentation*. When two or more words appear erroneously unsegmented (as it frequently happens in texts written by uneducated people), they were manually split and analyzed separately (*sela=se+la*, *inmente=in+mente*), similarly to the tool that automatically splits articulated prepositions and verbs with clitics.

[2] We plan to extend the manual revision of the output of the OCR, which is still ongoing, to approximately 1M tokens.

Text genre	Tok. + Lemm.	Tok. + Lemm. + PoS
Diary (Gadda, Martini, Sonnino)	43,419	49,868
Discourse (D'Annunzio, Morgari, Salandra, Salvemini, Treves, Turati; dichiarazioni del Partito Socialista)	44,942	7,792
Essay (Croce, Gemelli, Gentile)	8,352	9,524
Letters (Fontana, Monteleone, Monti, Procacci, Raviele)	89,938	5,310
Memoir (Cadorna, Jahier, Monelli, Prezzolini, Soffici)	134,874	22,938
Report (Comitati Segreti della Camera dei Deputati)	75,549	7,573
Tot.	397,074	103,005

Table 1: For each genre, number of tokens manually revised (for tokenization and lemmatization only, or also for PoS and morphological features).

Major issues with lemmatization:

1. *Rare terms*. The VGG corpus is rich with terms that are rare or old-fashoned in standard contemporary Italian (e.g., *costí*, *ingramagliare*), and that for this reason are rarely analyzed correctly. For such forms, the correct annotation was manually entered.

2. *Variants of lemmas*. Automatic tools often fail in lemmatizing a word correctly, when it does not refer to a standard lemma of contemporary Italian, but to one of its possible variants (e.g., *comperare* for *comprare*, *spedale* for *ospedale*). In such cases, both the standard and the variant lemma are manually annotated (359 different variant lemmas were found so far, for a total of 1361 occurrences).

3. *Misspellings*. In informal texts, words are often lemmatized incorrectly because they are wrongly spelled. For instance, *o* and *anno* may be the misspelled inflected forms of the verb *avere* (*ho*, *hanno*), and not just the conjunction *o* and the noun *anno*. In these cases, the correct linguistic annotation was added.

Major issues with morphological analysis:

1. *Variants in inflectional morphology*. Words that present rare or old-fashioned morphological formations (e.g., 3pl. pres. subj. *sieno* for standard It. *siano*; 2sg. fut. ind. *anderai* for standard It. *andrai*) in most cases are wrongly analyzed by the automatic tool and were therefore manually corrected.

4.2 Metalinguistic annotation

During the manual revision of the annotation (conducted on more than 500k tokens), an additional level of metalinguistic annotation was added. Words that can be considered as 'marked' with respect to standard contemporary Italian, and that are explicitly signaled as such in dictionaries (e.g., as literary or archaic forms), were manually identified and classified according to how they are labeled in the lexical resources consulted (*Dizionario De Mauro*, *Dizionario Hoepli*, *Dizionario Sabatini-Coletti*, and *Vocabolario Treccani*). We adopted the following labels:
dial: for forms classified as dialectal (e.g. *batajun*, *preive*; tot. 1,536 annotations).[3]
lit: for forms classified as literary or poetic (e.g. *pelago*, *nocumento*; tot. 1,046 annotations).
uncomm: for forms classified as rare and unfrequent (e.g. *impinguire*, *sconcordia*; tot. 891 annotations).
ant: for forms classified as obsolete or archaic (e.g. *imperocché*, *tardanza*; tot. 474 annotations).
reg: for forms classified as regional, i.e. typical of a regional variety of Italian (e.g. *cocuzza*, *mencio*; tot. 232 annotations).
pop: for forms classified as popular or vulgar (e.g. *pisciare*, *minchione*; tot. 134 annotations).

These labels (tot. 4,313 annotations) can be associated: (i) to a lemma (e.g. *tardanza*, *pelago*); (ii) to a variant lemma, in which case we add to the label the feature **var** (e.g., *imaginazione*, 'lit. var.' of the standard lemma *immaginazione*); (iii) to a single inflected form marked at the morphological level, in which case we add to the label the feature **morph** (e.g., *dieno*, 'morph. ant.' form of the 3pl. pres. subj. of the verb *dare*). Moreover, the same form may also receive two labels (e.g., *periglioso*, marked as 'ant./lit.').

Finally, misspelled or wrongly segmented forms (e.g., *Cavur* for *Cavour*, *cuatro* for *quat-*

[3]Not all dialectal forms are listed in Italian dictionaries. Nevertheless, they can be confidently identified in texts, since dialectal elements mostly appear in sequences, for instance in proverbs, songs, or poems. Moreover, authors often enclose dialectal forms in double quotation marks, or write them in italics.

tro, *inmente* for *in mente*) were also marked with a specific label: **err** (tot. 5,251 annotations).

It is evident that the metalinguistic annotation of marked forms is particularly relevant from a (socio-)linguistic point of view, since it offers an insight into the different dimensions of linguistic variation of the Italian language of the years of the WWI, from a diachronic, diatopic, diaphasic and diastratic points of view.

5 Automatic Linguistic Annotation

Automatic linguistic analysis of historical texts is a complicated venture. As reported in Piotrowski (2012), the main challenge is high variation on all levels both across and within texts, for instance due to the absence of standardized spelling, the occurrence of historical variants of words as well as peculiar syntactic structures. For these reasons, contemporary tools for linguistic analysis are generally not suitable for processing historical texts. This is the problem we faced in the project: as reported in Section 4 the texts of the VGG corpus differ in many respects from modern Italian.

Table 2 reports the performance recorded for the different levels of automatic linguistic annotation of the *VGG* corpus, using general and specialized language models. We tested the whole annotation pipeline on two test sets representative of two very different textual genres, i.e. discourses and letters, in order to assess the impact of different language varieties on the performance of the analysis tools.

We first trained UDPipe on IUDT v2.0: a significantly high drop of accuracy can be observed with respect to the state-of-the-art performance on modern Italian (Straka and Straková, 2017). In particular, for the letters collected by Monteleone very low performance is reported at all levels of analysis. This is mainly due to the features of this language variety: the letters were often written by uneducated people, they are characterized by a colloquial style, reminiscent of spoken language that is quite different from the typology of texts used for training. The split of sentences is the least accurate level of analysis: a non canonical use of punctuation both in Salandra's discourses and in the corpus of letters can be the main cause. On the other hand, token segmentation resulted to be less negatively affected in both cases.

Once the sub-corpus of $\sim$100k manually revised tokens was available, which included documents representative of the different textual genres considered, it was combined with the IUDT training data to retrain UDPipe. As expected, a general improvement was achieved at all analysis levels. For the two textual genres chosen for testing, the highest improvement turned out to be concerned with lemmatization. As discussed in Section 4, the *VGG* corpus contains several rare lexical items, old lemma variants, misspellings due to uneducated or informal use of language. The manual correction of the lemma helped to improve lemmatization and, similarly, PoS tagging.

6 Conclusions and current developments

Voices of the Great War is the first large corpus of documents in Italian dating back to the period of WWI. This corpus differs from other existing resources because it gives account of the wide range of varieties in which Italian was articulated in the years of WWI, namely from a diastratic (educated vs. uneducated writers), diaphasic (low/informal vs. high/formal registers) and diatopic (regional varieties, dialects) points of view. The linguistic variety subsumed in the corpus posits a number of challenges for current NLP tools, which are trained on texts representative of standard contemporary Italian. In this paper, we showed how we faced such challenges, by developing a more efficient model for the analysis of Italian texts of the period of WWI.

For approximately 20,000 tokens of the manually revised part of the corpus, we are building a syntactic annotation level performed according to the Universal Dependency scheme, which will constitute the first small treebank for historical Italian.

At the end of the project, the texts not covered by copyright will be freely dowloadable together with their annotations. The other texts will instead be browsable online with a dedicated interface.

References

Federico Boschetti, Andrea Cimino, Felice Dell'Orletta, Gianluca E. Lebani, Lucia Passaro, Paolo Picchi, Giulia Venturi, Simonetta Montemagni, and Alessandro Lenci. 2014. Computational analysis of historical documents: An application to italian war bulletins in WWI and WWII. In *Proceedings of the LREC 2014 Workshop on Language resources and technologies for processing and linking historical documents and archives - Deploying Linked Open Data in Cultural*

Test	Sentence Spitting	Tokenization	Lemmatization	PoS Tagging
Training: IUDT				
IUDT	**97.1%**	**99.8%**	**97.03%**	**97.02%**
Training: IUDT				
Discourses (Salandra-1922)	82.20% (-14.90)	99.53% (-0.27)	85.80% (-11.23)	83.36% (-13.66)
Letters (Monteleone)	65.58% (-31.52)	99.15% (-0.65)	83.05% (-13.98)	79.45% (-17.57)
Training: IUDT+VGG				
Discourses (Salandra-1922)	**92.46% (+10.26)**	99.74% (+0.21)	**95.20% (+9.40)**	90.82% (+7.46)
Letters (Monteleone)	69.46% (+3.88)	99.69% (+0.54)	**90.80% (+7.75)**	84.93% (+4.98)

Table 2: Comparison of F-scores in different annotation tasks using IUDT (*IUDT Training*) and combining out- and in-domain training data (*IUDT+VGG Training*) on different test sets. In parenthesis the relative improvement or drop of accuracy with respect to Straka and Straková (2017).

Heritage (LRT4HDA 2014, Reykjavik, Iceland), 70–75.

Cristina Bosco, Simonetta Montemagni and Maria Simi. 2013. Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank". *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria.

Federico Boschetti, Michele Di Giorgio and Nicola Labanca. 2018. Bisogna farli parlare: la formazione di un corpus di Voci della Grande Guerra e il comma 22. In Mirko Volpi (ed.), *Atti del primo convegno "Voci della Grande Guerra"*, Firenze, Accademia della Crusca, pp. 14-31.

Felice Dell'Orletta and Giulia Venturi. 2016. ULISSE: una strategia di adattamento al dominio per l'annotazione sintattica automatica. In Edoardo Maria Ponti and Marco Budassi (eds.), *Compter parler soigner: tra linguistica e intelligenza artificiale. Atti del convegno 15-17 dicembre 2014*, 55-79. Pavia University Press, Pavia.

Tullio De Mauro. 1963. *Storia linguistica dell'Italia unita*. Laterza, Bari.

Dizionario De Mauro = *Il Nuovo De Mauro*. Available online at: `https://dizionario.internazionale.it`.

Dizionario Hoepli = Aldo Gabrielli. *Grande Dizionario Italiano Hoepli*. Available online at: `http://dizionari.hoepli.it`.

Dizionario Sabatini-Coletti = Francesco Sabatini and Vittorio Coletti. *Il Sabatini Coletti Dizionario della lingua italiana*. Available online at: `http://dizionari.corriere.it/`.

L. Gaeta, C. Iacobini, D. Ricca, M. Angster, A. De Rosa, G. Schirato. 2013. MIDIA: a balanced diachronic corpus of Italian. 21st International Conference on Historical Linguistics, Oslo.

Alessandro Lenci, Nicola Labanca, Claudio Marazzini, Simonetta Montemagni. 2016. Voci della Grande Guerra: An Annotated Corpus of Italian Texts on World War I. *Italian Journal of Computational Linguistics*, 2(2):101–108.

Corinna Onelli, Domenico Proietti, Corrado Seidenari, Fabio Tamburini. 2006. The DiaCORIS project: a diachronic corpus of written Italian. Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006, pp. 1212–1215.

Lucia Passaro and Alessandro Lenci. 2014. "Il Piave mormorava...": Recognizing locations and other named entities in Italian texts on the great war. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini (eds.), *Proceedings of the First Italian Conference on Computational Linguistics*, 286–290. Pisa University Press, Pisa.

Michael Piotrowski. 2012. Natural Language Processing for Historical Texts. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers.

Milan Straka and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada, pp. 88–99.

Vocabolario Treccani = *Il Vocabolario Treccani*. Available online at: `http://www.treccani.it/vocabolario/`.