Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 911–918
Marseille, 11–16 May 2020

Voices of the Great War: A Richly Annotated Corpus of Italian Texts on the First World War

Alessandro Lenci[†], Simonetta Montemagni[*], Federico Boschetti[*], Irene De Felice[†], Stefano dei Rossi[♠], Felice Dell'Orletta[*], Michele Di Giorgio[†], Martina Miliani[†], Lucia C. Passaro[†], Angelica Puddu[†], Giulia Venturi[*], Nicola Labanca[♣]

[†] University of Pisa, Computational Linguistics Laboratory (CoLing Lab)
[*] Istituto di Linguistica Computazionale "Antonio Zampolli", CNR, Pisa
[♠] WebSoup s.n.c.
[♣] University of Siena, Dipartimento di Scienze Storiche e dei Beni Culturali
{alessandro.lenci@unipi.it, simonetta.montemagni@ilc.cnr.it, federico.boschetti@ilc.cnr.it, irene_def@yahoo.it, stefano@websoup.it, felice.dellorletta@ilc.cnr.it, digiorgio2@gmail.com, martina.miliani@fileli.unipi.it, lucia.passaro@fileli.unipi.it, a.puddu4@studenti.unipi.it, giulia.venturi@ilc.cnr.it, nicola.labanca@unisi.it}

Abstract

Voci della Grande Guerra ("Voices of the Great War") is the first large corpus of Italian historical texts dating back to the period of First World War. This corpus differs from other existing resources in several respects. First, from the linguistic point of view it gives account of the wide range of varieties in which Italian was articulated in that period, namely from a diastratic (educated vs. uneducated writers), diaphasic (low/informal vs. high/formal registers) and diatopic (regional varieties, dialects) points of view. From the historical perspective, through a collection of texts belonging to different genres it represents different views on the war and the various styles of narrating war events and experiences. The final corpus is balanced along various dimensions, corresponding to the textual genre, the language variety used, the author type and the typology of conveyed contents. The corpus is annotated with lemmas, part-of-speech, terminology, and named entities. Significant corpus samples representative of the different "voices" have also been enriched with meta-linguistic and syntactic information. The layer of syntactic annotation forms the first nucleus of an Italian historical treebank complying with the Universal Dependencies standard. The paper illustrates the final resource, the methodology and tools used to build it, and the Web Interface for navigating it.

Keywords: Historical Corpora, Linguistic and Meta–linguistic Annotation, Information Extraction

1. Introduction and Motivation

Historical corpora are still vastly underrepresented within the landscape of language resources, for several reasons. They are *hard to collect*, because of the time-consuming effort of digitizing printed sources, and they are *hard to process and annotate* automatically, because the performances of existing Natural Language Processing (NLP) tools dramatically drop when applied to diachronic varieties of a language. Moreover, they are not economically attractive, due to their limited use to develop downstream applications with a large-scale economic impact. Still, historical corpora represent an invaluable asset in the era of Digital Humanities, given the growing interest in applying quantitative and computational methods to diachronic linguistics and historical text analysis (Tahmasebi et al., 2019). Italian makes not exception to this trend, as diachronic corpora are still few, among which it is worth pointing out the *Corpus OVI dell'Italiano antico* and the *Corpus TLIO* (by OVI–CNR), the *DiaCORIS* corpus (Onelli et al., 2006) and the *MIDIA* corpus (Gaeta et al., 2013). One major shortcoming of such resources is the extremely large timespan they cover in comparison to their limited size.

The project *Voci della Grande Guerra* ("Voices of the Great War") (in short, VGG)[1] aimed at filling this gap by creating the largest digital corpus to date of Italian texts at the time of World War I (WWI). The corpus includes a selection of texts representative of different textual genres and registers, including popular Italian. VGG 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 XXth century Italian. These characteristics make the *VGG* corpus unique in the panorama of existing Italian historical corpora. The only other corpus covering the same period and with multi–level "silver" annotation is *ALCIDE* (Moretti et al., 2016), but it only contains the writings of a single author, Alcide De Gasperi.

Why a corpus on WWI? The reason is that the Great War in Italy is extremely relevant both from the historical and the linguistic points of view. If WWI as a factual event is quite well-known, much less known are the different narrative and experiential perspectives on this war. The texts produced (with different purposes) in that period have had a crucial role in shaping the images of war before, during and after the conflict. Such texts express the attempt to persuade people to accept or refuse the war, or simply represent the way in which people tried to make sense of the tremendous experiences of their time. VGG thus aims at providing historians with a new digital tool to explore this wealth of extremely different (and often forgotten) voices.

Linguists have always ascribed a very important function to the Great War as a decisive time in the process leading to the linguistic unification of Italy (De Mauro, 1963), because imposing 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 the national language as the main communicative medium,

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

in contact with more educated officers possessing a higher level of Italian. The comparison between the language varieties that are included in VGG will facilitate a deeper understanding of these issues, and will provide new evidence on how language difficulties were overcome in those dramatic circumstances. Moreover, the corpus will allow scholars to study the influence on Italian by various linguistic models, the literary ones for the educated class and the dialectal ones for the less literate authors.

The project *Voci della Grande Guerra* was funded by a two-year grant from the Special Mission for the Celebrations of the 100^{th} Anniversary of WWI at the Presidenza del Consiglio dei Ministri of the Italian Government.[2] The project started in May 2016 and ended in November 2018. The project partners were: i) *University of Pisa*, Computational Linguistics Laboratory – CoLing Lab (project coordinator); ii) *Istituto di Linguistica Computazionale "Antonio Zampolli"*, CNR, Pisa, CoPhi Lab and ItaliaNLP Lab; iii) *University of Siena*, Dipartimento di Scienze Storiche e dei Beni Culturali; iv) *Accademia della Crusca*, Firenze.

This paper is organized as follows: in Section 2. we describe the overall composition of the collected corpus and we detail its internal composition. Section 3. reports results and challenges met during the process of corpus annotation at different levels. The query functionalities of the corpus offered by the dedicated web platform we developed are presented and exemplified in Section 4.

2. The "Voices of the Great War" Corpus

VGG texts, most of which never digitized before, belong to a wide range of registers, textual genres and linguistic varieties: the corpus composition was aimed at maximizing the representativeness of the corpus with respect to the various perspectives on the war and the various styles of narrating war events and experiences. VGG was designed with the aim of reconstructing the "polyphony" of the languages of Italy at war: the official voice of propaganda and the voice of soldiers, the voice of newspapers and the voice of letters, the voice of the elite of intellectuals and the popular voice, the voice of consensus and the voice of dissent, male voices and female voices. As a consequence, the final corpus is balanced along various dimensions, corresponding to the textual genre, the language variety used, the author type (e.g., sex, education, profession, political orientation etc.), or the typology of conveyed contents. Moreover, significant corpus samples representative of the different "voices" have also been enriched with multi–level linguistic (namely, lemmatization, morpho–syntactic, syntactic and semantic) annotation (both silver and gold) as well as with meta-linguistic information.

2.1. Corpus Composition

VGG consists of 99 texts for a total of more than 22,000 pages that were written in Italian during the period of the First World War or shortly afterwards. In particular, they go from 1914 up to 1923[3], in order to cover not only the years of the war, but also the cultural and social environment leading to the war and the aftermath of the Great War. Texts in the corpus have been selected with respect to the various war years, in order to create the prerequisites for investigating the immediate impact of the war and of its different phases (e.g., before and after the Caporetto battle, which represents the key turning point of the Great War in Italy) on language, communicative styles and contents.

The selection of texts has been jointly carried out by historians and linguists in order to represent the "polyphony" of the different voices of people who were affected by World War I, directly (i.e., who participated in it) or indirectly.

From the perspective of textual genre, VGG collects discourses, reports and diaries of politicians and military chiefs; letters written by men and women, soldiers and civilians; literary works by intellectuals, poets, and philosophers; writings of journalists and lawyers. The first three columns in Table 1 report the distribution of texts by textual genre in the final corpus.

The variety of VGG texts can also be looked at from another perspective, that of register which only partially overlaps with genres. Within the corpus, the following typology of linguistic registers is represented:

- the *official military language*, testified by the full collection of war bulletins, books of military strategy and analysis of war conduct, as well as propaganda texts and court martial records;
- the *language of the middle class*, represented by samples of officers' diaries and memoirs, most likely written in a high-level Italian inspired to major literary examples of the time, such as Gabriele D'Annunzio;
- the *popular language*, exemplified in letters, diaries and memoirs from soldiers;
- the *language of the political class*, testified in parliamentary proceedings and official political speeches;
- the *language of the intellectual elite*, represented by samples of pamphlets, literary journals, etc.;
- the *standard language of public opinion*, as testified in newspaper articles, magazines, news reports from the front, etc.

In the construction of the corpus particular attention has also been devoted to the tackled topics and the expressed opinions which are needed to have a comprehensive view of the historical period. Essays, letters, diaries and memory of War speaks about socialism, communism, Italian rebuilding, ideals, hopes, feminism, but also fears and feelings for the family. Some texts are official documents from the Italian House of Representatives (*Camera dei Deputati*) or the parliamentary Secret Committee, others are poems or political speeches.

The corpus is representative of different genres and registers: behind them there are many different types of authors, men vs. women, educated vs. illiterate, with different professional profiles, political and cultural backgrounds, and last but not least visions of war. For instance, authors of

[2] http://www.centenario1914-1918.it

[3] Actually, 18 volumes were published between 1929 and 2000, but they include texts (e.g., letters) originally produced during the war or in the years immediately after it.

	VGG Corpus			VGG-Silver		
Text Genre	**Books**	**Pages**	**Tokens**	**Books**	**Pages**	**Tokens**
Diary	16	6318	1,687,353	16	1,026	253,095
Discourse	35	3,206	1,200,207	35	644	200,050
Essay	22	5,691	1,768,893	22	786	200,272
Letters	5	1,272	408,970	5	389	106,583
Memoir	15	4,181	890,051	15	1,443	326,016
Pamphlet	2	92	11,604	2	79	11,452
Poetry, Novel, Story	1	418	72,785	1	47	8,864
Report	3	1,488	409,392	3	290	85,016
Total	**99**	**22,666**	**6,449,255**	**99**	**4,704**	**1,191,348**

Table 1: VGG composition by textual genre

VGG texts are politicians like Benedetto Croce, Giovanni Gentile or Vittorio Emanuele Orlando, others are journalists or intellectual people like Teresa Labriola or Luigi Barzini, talking about rights and ideals, others describe the war from the trenches or hospitals, they are generals like Luigi Capello or Luigi Cadorna, voluntaries and soldiers like Giuseppe Prezzolini or Maria Luisa Perduca. Although most part of the authors are men, voices of women have also been included (besides those mentioned above, cfr. Carla Cadorna, Matilde Serao, Anna Soldati Manis).

2.2. Corpus Pre-processing and Organization

Most part of the texts in VGG existed only in printed form and were scanned and digitized with Tesseract (`http://bit.ly/380T79O`), an open source OCR engine, which is a valid competitor to commercial software, such as Abbyy FineReader. A small group of documents was available online as searchable pdfs, but we preferred to apply our OCR methods, for the sake of homogeneity. Once the corpus has been digitized as described above, part of it has been manually corrected line-by-line and word-by-word with a correction tool especially designed for this purpose (Boschetti et al., 2018). This tool is a collaborative proof-reading web application inspired to WikiSource for the management of multiple text reviews and minimal TEI P5 encoding (such as the division in paragraphs, the identification of footnotes, etc.). The result of this manual revision process is a sample corpus, which includes excerpts of variable size from all VGG texts. This corpus, henceforth referred to as *VGG-Silver*, constitutes our textual gold standard: in particular, VGG-Silver contains excerpts extracted from all 99 texts, for a total of 4,704 pages and more than 1M tokens (see the last three columns of Table 1 for its internal composition). We refer to the part of the VGG corpus without manual revision of the OCR output as *VGG-Bronze*.

VGG-Silver has also been enriched with different types of linguistic, meta–linguistic and semantic manually revised annotations, covering it partially (this is the case of the linguistic and meta–linguistic information) or fully (semantic annotation). Table 2 illustrates the internal composition of the linguistically and meta–linguistically annotated section of the corpus, which includes text samples representative of different textual genres and language registers. We refer to this part of the corpus as *VGG-Gold*. VGG-Gold is in turn subdivided into progressively smaller sections, whose internal composition is described in the different columns of the table. The first column describes the part of the corpus, containing 500,079 tokens, with manually revised annotation for what concerns sentence splitting, tokenization and lemmatization; in addition, this part of the corpus has been enriched with meta–linguistic information aimed at highlighting features characterizing the variety of Italian used in the historical period considered. For smaller sections of VGG-Gold, manual revision was carried out for morpho–syntactic annotation (second column) and syntactic annotation (third column).

The gold sections of VGG can be used to retrain automatic linguistic annotation pipelines in order to improve the performance of the automatic analysis tools on this historical variety of Italian. As illustrated in De Felice et al. (2018), retraining was done for morpho–syntactic annotation, with promising results: retrained models for morpho–syntactic annotation were used to tokenize, morpho–syntactically annotate and lemmatize the rest of VGG-Silver. The results of this fully automatic annotation can be used for querying purposes (see Section 4.).

3. Corpus Annotation

Corpus annotation was carried out on VGG-Silver, i.e. the section of the corpus with manually revised OCR results. In particular, multi-level annotation of the corpus has been performed in different ways, namely fully manually, fully automatically and semi-automatically (i.e. automatic annotation followed by manual revision).

It is a widely acknowledged fact that automatic linguistic analysis of historical texts is a complicated venture (Piotrowski, 2012), due to e.g. 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 and need to be specialized with respect to the peculiarities of the historical variety of language to be processed. The annotation methodology we have employed for the annotation of the VGG corpus was articulated in the following steps:

1. automatic annotation of representative sample texts using UDPipe (Straka et al., 2016) trained on the Italian Universal Dependency Treebank (IUDT), version 2.0 (Bosco et al., 2013);

Text genre	Tok. + Lemm. + Meta–ling.	Morpho–synt.	Dep. Syntax
Diary (Gadda, Martini, Sonnino)	93,287	49,868	3,050
Discourse (D'Annunzio, Morgari, Salandra, Salvemini, Treves, Turati; dichiarazioni del Partito Socialista)	52,734	7,792	2,627
Essay (Croce, Gemelli, Gentile)	17,876	9,524	3,487
Letters (Fontana, Monteleone, Monti, Procacci, Raviele)	95,248	5,310	–
Memoir (Cadorna, Jahier, Monelli, Prezzolini, Soffici)	157,812	22,938	2,445
Report (Comitati Segreti della Camera dei Deputati)	83,122	7,573	3,050
Total	**500,079**	**103,005**	**14,659**

Table 2: For each genre, number of tokens manually revised for tokenization and lemmatization (column 1), or also for morpho–syntactic annotation (column 2), or also for syntactic annotation (column 3)

2. the output of step 1. was manually revised and, whenever needed, corrected;

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 tool. The retrained model was used for the fully automatic annotation of the remaining parts of VGG-Silver.

The results of fully manual and semi-automatic (i.e. "gold") annotations are briefly illustrated below, together with the used methods and tools.

3.1. Lemmatization and Morpho-syntactic Annotation

Morpho–syntactic annotation was first carried out with UDPipe trained on IUDT v2.0. Once the manually revised gold corpus has been available, it was combined with the IUDT training data to retrain UDPipe: the new specialized model has been used to automatically annotate the rest of the Silver-VGG corpus (De Felice et al., 2018).

A number of characteristics specific to the linguistic variety of the documents posed critical difficulties to the automatic linguistic annotation at different levels. Sentence splitting, tokenization and lemmatization have been manually checked and corrected for 500,079 tokens representative of different genres (see column *Tok.+Lemm.+Meta–ling.*, Table 2). For a subset of this revised part of the corpus (namely 103,005 tokens, see column *Morpho–synt.*, Table 2), manual revision has also targeted morpho–syntactic annotation (i.e. part–of–speech information and morpho–syntactic features).

Major issues involved with tokenization concern segmentation of pronominal clitics attached to finite verbs, that are very rare in contemporary Italian (e.g. *abbi+ti*, *sia+si*), and hyposegmentation of words which were erroneously written (possibly by uneducated people) as a unique word (e.g. *inmente=in+mente*). Rare or old-fashioned terms (e.g. *costì*, *ingramagliare*), non-standard variants of lemmas (e.g. *comperare* for *comprare*, *spedale* for *ospedale*) and misspellings (e.g. *o* and *anno* for *ho* and *hanno*) represented the main source of errors in the automatic lemmatization and they were manually corrected. Old-fashioned morphological formations (e.g. 3pl. pres. subj. *sieno* for standard It. *siano*) in most cases are wrongly analyzed by the automatic tool. For a more detailed account of the issues tackled and the solutions adopted for these annotation levels, see De Felice et al. (2018).

We also studied the vocabulary composition of the different sub–corpora by extracting out–of–vocabulary words, i.e. not contained in the *Dizionario Macchina dell'Italiano* (Gruppo di Pisa, 1976), the wide coverage dictionary used here as a reference resource, containing 317,845 lemmata corresponding to 1,268,442 inflected forms. It turned out that the genres Diary, Memoir and Essay (typically written by educated people) include a higher proportion of out–of–vocabulary words: many of them represent neologisms such as *munizionamento*, *promettitore*, *austriacante*, *triplicista* / *triplicistico* or old–fashioned connectives such as *checché*, *dimodoché*, *dappoiché*. On the other hand, Discourses, Letters and Reports are characterized by a significantly higher percentage of words belonging to the Basic Italian Dictionary (De Mauro, 2000): following Giuliani et al. (2005) who demonstrate the stability and conservativeness of the Italian lexicon throughout the centuries, we took the Basic Italian Dictionary as a reference, starting from the assumption that the most important areas of experience continue to be mostly indicated by a stable nucleus of words in use for at least seven centuries.

Passa un morto della [13.a]$_{ORG}$. Bombardamento di un'ora a shrapnel. Conversazione col Capitan [Bon]$_{PER}$.
A dead man of the 13a is passing. One hour bombing with shrapnels. Conversation with Captain Bon.

Figure 1: A sample text annotated with NEs.

3.2. Named Entities Annotation

The Named Entity annotation of VGG-Silver was carried out with an adaption of the CoLingLab Named Entity Recognizer (NER) described in Passaro and Lenci (2014) and based on the Stanford NER (Finkel et al., 2005). Our NE label set was composed of three tags: PER for people's proper names (e.g., *Luigi Cadorna*), LOC for locations (e.g., *passo di Monte Croce*), and ORG for organizations, including military formations (e.g., 2° *Reggimento Bersaglieri*) (see Figure 1). To annotate VGG-Silver we concatenated two distinct NER models, both trained on the same dataset. In order to adapt the NER to the Italian variety represented in the VGG corpus, the training

data consisted of the ICAB corpus (Magnini et al., 2006), in which the Geo-Political Entities (GPE) were converted into the LOC tag, and the Italian war bulletins of World War I (Boschetti et al., 2014), in which the military formations originally tagged as MIL were converted into ORG. Moreover, we added three texts from VGG (a sample from Mussolini's diary, a parliamentary report and a speech by the politician Claudio Treves), whose NEs were manually annotated. Only the second model was also trained on a gazetteer including person names collected from the online Italian Treccani Encyclopedia,[4] WWI military organization names gathered from Wikipedia,[5] and location names provided by the *I Luoghi della Grande Guerra* project.[6]

3.3. Terminology Annotation

VGG-Silver was annotated and indexed with simple and multiword terms automatically extracted with EXTra (Passaro and Lenci, 2016). The EXTra term extractor takes into account the linguistic structure of multiword terms by implementing a candidate selection step that uses manually-defined structured PoS-patterns. Moreover, in order to tackle the complexity of term phrases, EXTra adopts a new association measure that promotes terms composed by one or more sub-terms. The intuition is that the degree of termhood of a candidate pattern is a function of the statistical distribution of its parts, and of the presence of highly weighted sub-terms. The last step of EXTra applies a filtering function to separate real terms from wrong candidates. EXTra includes various parameters that allow users to optimize the extracted terms with respect to the target corpus and domain. In particular, users can specify the set of structured patterns that guide the extraction process, a list of stopwords, the association measure to be used by the weighting algorithm (in VGG, local mutual information, LMI), as well as the thresholds for the association measure and the n-gram frequency. In the present case, EXTra was fed with a list of 24 fully-specified PoS-patterns covering terms with a noun head and an adjectival or prepositional phrase modifier. For instance, the pattern `['s','a']` retrieves the term *comando supremo*, and the pattern `['s','e','s']` the term *onore della patria*.

The terms identified by EXTra were manually inspected to filter out possible errors. A list of extracted terms is reported in Table 3. A final list of simple and multiword terms were annotated on the VGG-Silver corpus. Moreover, each document in the corpus was indexed with the terms associated with their raw frequency and their tf-idf value. Tf-idf is a popular weighting scheme used in information retrieval, directly proportional to how many times the term appears in the document, but inversely proportional to its global frequency in the collection. The tf-idf weight scheme is used in the Web interface to identify the most characteristic terms of a document (cf. Section 4.).

[4] http://www.treccani.it/biografico/

[5] https://it.wikipedia.org/wiki/Organizzazione_del_Regio_Esercito_durante_la_Prima_Guerra_Mondiale

[6] http://luoghigrandeguerra.iia.cnr.it/

Term	LMI
presidente del consiglio	1231.92
generale Cadorna	977.51
comando supremo	969.46
corpo di armata	761.91
onorevole collega	727.61
viva approvazione	724.72
ordine del giorno	656.78
zona di guerra	610.72
estrema sinistra	606.54
sottosegretario di stato	466.43

Table 3: Sample of multiword terms identified by EXTra in VGG-Silver.

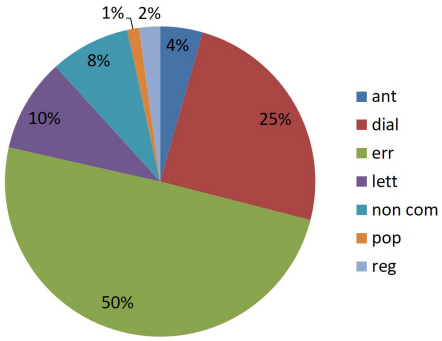

Figure 2: Distribution of meta–linguistic labels in VGG.

3.4. Meta–linguistic Annotation

Meta-linguistic annotation was aimed at manually identifying and classifying 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); reference lexical resources used for this task are listed in the References section. Annotation was performed with respect to the following classes: **lit**, for literary forms (e.g. *pelago*, *nocumento*); **dial**, for dialectal forms (e.g. *batajun*, *preive*); **uncomm**, for rare forms (e.g. *impinguire*, *sconcordia*); **ant**, for obsolete forms (e.g. *imperocche*, *tardanza*); **reg**, for diatopically marked forms, typical of a regional variety of Italian (e.g. *cocuzza*, *mencio*); **pop**, for popular or vulgar forms (e.g. *pisciare*, *minchione*). Finally, single misspelled or hyposegmented forms (e.g., *Cavur*, *cuatro*, *inmente*) were also marked with a specific label: **err**(or). Meta–linguistic annotation has been carried out manually, for a total of 10,594 annotations, associated with a lemma (e.g. *tardanza*, *pelago*), or a variant lemma (e.g., *imaginazione*) or individual inflected forms (e.g. *dieno*), sometimes combined together (e.g. *periglioso*, marked as "ant/lit").

Figure 2 illustrates the distribution of meta–linguistic labels in the corpus: half of them correspond to errors, 25% are dialectal variants and 18% are literary and obsolete forms. Interestingly, meta–linguistic labels mainly concentrate in letters (65%), memoirs (22%) and diaries (6%), i.e. those genres which constitute the so–called "scritture popolari" (Renzi, 2017), namely writings by laypersons. Meta–linguistic annotation also offers an interesting insight into

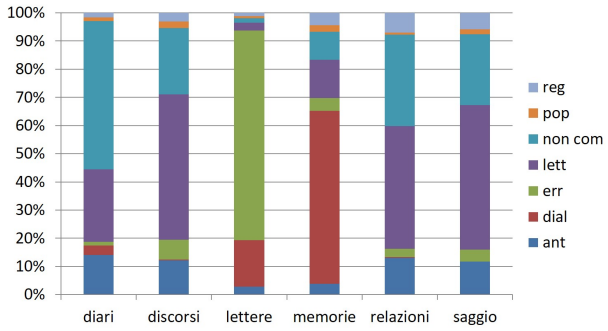

Figure 3: Normalized distribution of meta–linguistic labels across genres.

Finalmente, una *disciplina* esteriore anche molto severa, ma | che non sia accompagnata dalla disciplina interiore, | quando sorgano straordinarie contingenze, | quando occorra reagire contro eccezionali forze dissolventi come quelle, ad esempio, | che ci condussero a Caporetto, non **basta** | a mantener salda la compagine degli eserciti ed | a conservare al complesso organismo tutta la sua potenza.

Finally, a very severe external discipline, but that is not accompanied by inner discipline, when extraordinary contingencies arise, when it is necessary to react against exceptional dissolving forces such as, for example, those that led us to Caporetto, is not enough to keep the team firmly and to maintain all its power to the complex organism.

Figure 4: A sample text from Cadorna vol. I, pp. 27-39, 1921 (Memoir genre).

the different dimensions of linguistic variation of the Italian language in the period of the First World War, from a diachronic, diatopic, diaphasic and diastratic points of view. For instance, it is worth noting that most part of errors are found in letters, that dialectal forms are typically associated with memoirs and then with letters, and literary forms with discourses, essays and reports.

3.5. Syntactic Annotation

A small part of the corpus, representative of the different genres corresponding to 14,659 tokens (see column *Dep. Syntax* in Table 2), was enriched with a syntactic annotation level which was performed according to the Universal Dependency scheme, a *de facto* standard for dependency syntax: this will constitute the first small treebank for historical Italian among the set of UD treebanks[7]. Currently, the annotation of a sample of letters, which poses specific challenges due to the massive presence ill-formed constructions, is still being completed and is not part of the syntactically annotated section of the corpus.

The boxed text in Figure 4, followed by a semi-literal translation into English, is meant to give the reader an idea of the type of sentences occurring in the VGG corpus. The sentence, extracted from Luigi Cadorna's memoirs[8], is long

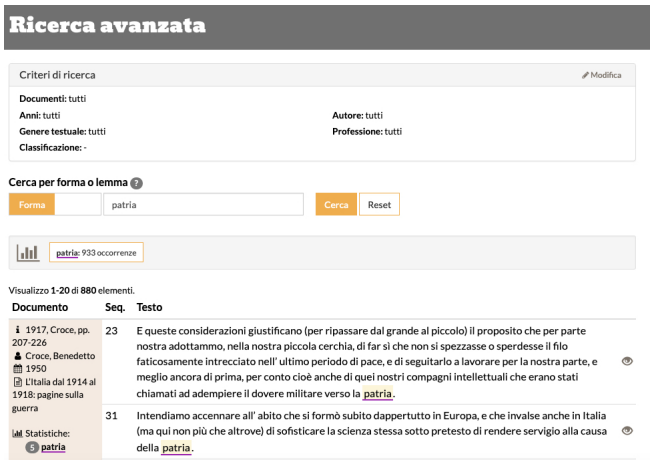

Figure 5: Free text search of *patria* "homeland".

61 tokens, contains 6 subordinate clauses (2 relatives, 2 infinitives and 2 temporal adverbials) whose boundaries are marked with |, and contains quite complex dependencies, e.g. the distance between the nominal subject (*disciplina*) and the main verb (**basta**) is 39 tokens. The major issues we had to tackle for the dependency annotation of this subset of the corpus are concerned with:

- the length of sentences, which on average are significantly longer than contemporary Italian (25.08 vs 21.04 tokens per sentence) as resulting from the Italian Universal Dependencies Treebank (Bosco et al., 2013), with memoirs and essays showing a much higher average value, namely 35.41 and 31.65 tokens per sentence;
- widespread use of subordinate constructions, often recursively embedded, and which in letters and diaries represent more than 40% of the clauses; as in the previous case, this is a peculiarity of the Italian language of the early XXth century, which differs from contemporary Italian where the recourse to subordination is more limited, especially for what concerns embedded structures;
- long distance dependency relations: the average distance between the head and the dependent is longer with respect to contemporary Italian (i.e. 3.36 vs 2.67), with genres like memoirs and essays showing much longer dependency links.
- other problematic syntactic annotation areas are represented by an anomalous usage of punctuation with respect to the contemporary one, and atypical syntactic constructions such as absolute clauses without any grammatical connection to the rest of the sentence.

4. Exploring VGG

The VGG project also developed a software platform to assist researchers during the phases of corpus building, and

[7] https://universaldependencies.org/

[8] Luigi Cadorna, *La guerra alla fronte italiana fino all'arresto sulla linea della Piave e del Grappa: 24 maggio 1915-9 novembre 1917*, vol. 1, Milano, Treves, 1921.

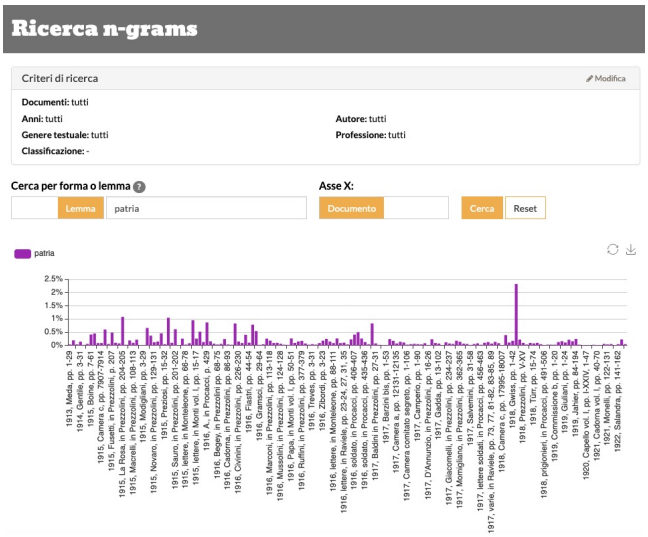

Figure 6: N-gram search of *patria* "homeland".

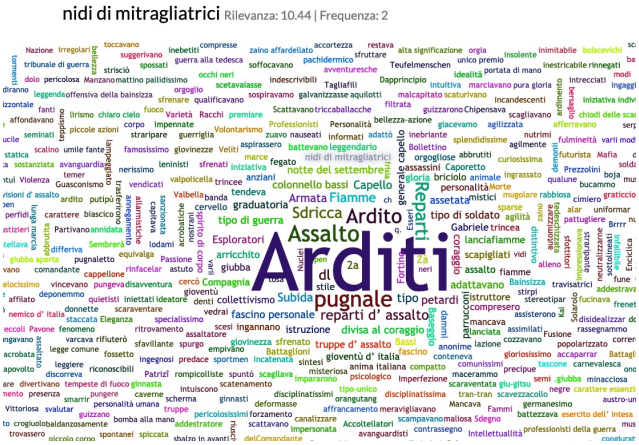

Figure 7: Term word cloud from Carli M. (1919), *Noi arditi*.

to provide various search functionalities. The tool consists of a back-end module to support the correction of digitized and automatically annotated texts, and a front-end module for the exploration of the corpus with advanced forms of information visualization and query, to perform both "close" and "distant" readings of the texts (Moretti, 2013). Currently, the web interface implements the following functionalities:

- **sub-corpus definition** – it is possible to define subsets of VGG according to several criteria, such as text genre, author, and year;
- **free text search** – selected texts can be searched by (unigram or n-gram) word and lemma. The interface returns all their occurrences in context (Figure 5);
- **NE search** – VGG can be explored by browsing the list of occurring person, location, and organization names;
- **n-gram search** – frequency analysis of simple and complex terms plotted by document or by year, similar to *Google Ngram Viewer* (Figure 6);[9]
- **term search** – simple and multiword terms of each text are shown as a word cloud representing their frequency or tf-idf values (Figure 7).

5. Conclusion

We presented the first large corpus of Italian historical texts dating back to the period of First World War, "Voices of the Great War". The main elements of novelty can be summarized as follows:

1. the design and construction of an archive of digital texts, most of which never digitized before, and belonging to a wide range of registers, textual genres and linguistic varieties, with the aim of maximizing the representativeness of the corpus with respect to the various perspectives on the war, and the various styles of narrating war events and experiences;
2. state-of-the-art tools for Optical Character Recognition, Natural Language Processing and text mining have been used and, whenever needed, customized to annotate the digitized texts with linguistic and semantic metadata that enrich their informative value, multiplying the possibilities and modalities of accessing the information contained in them;
3. an online navigation tool allows personalized search paths by a wide audience going from scholars in contemporary history and linguistics, to teachers, students, up to common people interested in knowing more about an event that has been so crucial for the Italian cultural identity, and affected practically every family, albeit in different forms and degrees;
4. a digital platform to create, navigate and explore a digital archive, in order to evoke the wealth and polyphony of the voices represented therein. In this way, "Voices of the Great War" is not just a static and closed corpus, but rather an open and extendable digital platform for historical text analysis and processing.

The created resource, which will be made available in the ILC-CNR for CLARIN-IT repository, is now ready for substantially improving our knowledge about the structure and varieties of Italian language at the time of the First World War and how they differ from contemporary Italian, about how the war changed the language of the newborn Italian State, about the different ways to experience and describe the Italian war by its protagonists.

6. Acknowledgments

The project *Voci della grande Guerra* was supported with a grant by the Italian "Presidenza del Consiglio dei Ministri". We want to thank Claudio Marazzini and the Accademia della Crusca partners of the project, Paolo Plini, for sharing the location names collected in the project *I luoghi della Grande Guerra*, Paolo Picchi, for his support in the text digitization, Eusebia Parrotto, Mauro Hausbergher

[9] https://books.google.com/ngrams

and Francesco Serra, for contributing to the OCR correction, and Anna Caruso, Giulia Chiriatti, Clara D'Apoli, Pietro Dell'Oglio, Maria Iacono, Andrea Pedrotti, Chiara Pugliese, Selena Rorberi and Elisabetta Triolo for their help in the corpus annotation.

7. Bibliographical References

Boschetti, F., Cimino, A., Dell'Orletta, F., Lebani, G., Passaro, L., Picchi, P., Venturi, G., Montemagni, S., and Lenci, A. (2014). Computational analysis of historical documents: An application to italian war bulletins in world war i and ii. In *Workshop on Language resources and technologies for processing and linking historical documents and archives (LRT4HDA 2014)*, pages 70–75. ELRA.

Boschetti, F., Di Giorgio, M., and Labanca, N. (2018). *Bisogna farli parlare*: la formazione di un corpus di "voci della grande guerra" e il "comma 22". In M. Volpi, editor, *Atti del primo convegno "Voci della Grande Guerra", Firenze, Accademia della Crusca*, pages 14–31.

Bosco, C., Montemagni, S., and Simi, M. (2013). Converting italian treebanks: Towards an italian stanford dependency treebank. In *Proceedings of the ACL Linguistic Annotation Workshop & Interoperability with Discourse*, Sofia, Bulgaria, August.

De Felice, I., Dell'Orletta, F., Venturi, G., Lenci, A., and S.Montemagni. (2018). Italian in the trenches: Linguistic annotation and analysis of texts of the great war. In *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it)*, Turin, Italy, December.

De Mauro, T. (1963). *Storia linguistica dell'Italia unita*. Laterza, Bari.

De Mauro, T. (2000). *Grande dizionario italiano dell'uso (GRADIT)*. UTET, Torino.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Gaeta, L., Iacobini, C., Ricca, D., Angster, M., and Schirato, A. D. R. G. (2013). Midia: a balanced diachronic corpus of italian. In *21st International Conference on Historical Linguistics, Oslo*.

Giuliani, A., Iacobini, C., and Thornton, A. M. (2005). La nozione di vocabolario di base alla luce della stratificazione diacronica del lessico dell'italiano. In Tullio De Mauro et al., editors, *Parole e numeri. Analisi quantitative dei fatti di lingua*, pages 193–213. Aracne, London.

Gruppo di Pisa. (1976). Il dizionario di macchina dell'italiano. In D. Gambarara, et al., editors, *Linguaggi e Formalizzazioni*.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Lenzi, V. B., and Sprugnoli, R. (2006). I-cab: the italian content annotation bank. In *LREC*, pages 963–968. Citeseer.

Moretti, G., Sprugnoli, R., Menini, S., and Tonelli, S. (2016). Alcide. *Know.-Based Syst.*, 111(C):100–112, November.

Moretti, F. (2013). *Distant Reading*. Verso, London.

Onelli, C., Proietti, D., Seidenari, C., and Tamburini, F. (2006). The diacoris project: a diachronic corpus of written italian. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*, pages 1212–1215.

Passaro, L. C. and Lenci, A. (2014). Il piave mormorava. . . : Recognizing locations and other named entities in italian texts on the great war. In *Proceedings of 1st Italian Conference on Computational Linguistics (CLiC-it)*, pages 286–290, Pisa, Italy, December.

Passaro, L. C. and Lenci, A. (2016). Extracting terms with extra. In *Computerised and Corpus-based Approaches to Phraseology: Monolingual and Multilingual Perspectives*, pages 188–196. Editions Tradulex.

Piotrowski, M. (2012). Natural language processing for historical texts. In Morgan & Claypool Publishers, editor, *Synthesis Lectures on Human Language Technologies*.

Renzi, L. (2017). Philologica militaria. in margine alle "lettere dei prigionieri di guerra" di spitzer nella nuova edizione del 2016. *Linguistica e Filologia*, 37:7–52.

Straka, M., Hajic, J., and Strakova, J. (2016). UD-Pipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*.

Nina Tahmasebi, et al., editors. (2019). *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, Florence, Italy, August. Association for Computational Linguistics.