IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

aAccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

direttore responsabile
Michele Arnese

isbn 978-88-99982-26-3

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_2_2

IJCoL

Volume 2, Number 2
december 2016

Special Issue:
Digital Humanities and Computational Linguistics

Guest Editors:
John Nerbonne, Sara Tonelli

CONTENTS

Voci della Grande Guerra
An Annotated Corpus of Italian Texts on World War I

Alessandro Lenci*
Università di Pisa

Nicola Labanca**
Università di Siena

Claudio Marazzini†
Accademia della Crusca, Firenze

Simonetta Montemagni‡
ILC – CNR, Pisa

Voci della Grande Guerra *(Voices of the Great War) is a scientific and cultural initiative with the aim of preserving and promoting the memory of Italy in World War I through the creation of a corpus of digital texts selected by historians and linguists in order to be representative of the different ways to experience and describe the Italian war by its protagonists. With the help of advanced techniques of computational linguistics, semantic web and information visualization, the digitized historical materials will be explored with an online interface to enable easy but effective and innovative search modalities. The project will allow experts as well as non-experts to become acquainted with "linguistic polyphony" of Italy during World War I.*

1. Motivations and goals

World War I (WWI) represents a crucial landmark in the history of mankind. It has changed the destiny of whole generations and its geopolitical consequences still affect contemporary world. Unfortunately, the knowledge of the Great War is progressively fading away, especially among young generations. The first centenary of WWI raises the moral issue of how to preserve the historical memory of these events, making them accessible to a larger audience, not limited to scholars and experts. A better appreciation of the different and often contrasting motivations that brought millions of people to war, as well as of the political and cultural legacy of this conflict, can also provide important interpretive keys for the present historical phase.

The Great War is the first war of mass death, but it is also the first war of mass text production. An important part of such texts are also first-hand accounts by people endeavouring the experience of writing for the first time, to make sense of the dramatic and disruptive events they witnessed. Now, increasing amounts of such sources are available in digital form, but many historical documents still need to be digitized. This situation is even more difficult when one considers the historical sources concerning

* Computational Linguistics Laboratory (CoLing Lab), Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, via S. Maria 36, 56126, Pisa, Italy. E-mail: alessandro.lenci@unipi.it
** Dipartimento di Scienze Storiche e Beni Culturali, Università di Siena, Complesso dei Sevi, 53100, Siena, Italy. E-mail: nicola.labanca@unisi.it
† Accademia della Crusca, via di Castello 46, 50141, Florence, Italy. E-mail: claudio.marazzini@teletu.it
‡ Istituto di Linguistica Computazionale CNR, "Antonio Zampolli" - Via Moruzzi 1, 56124 Pisa, Italy. E-mail: simonetta.montemagni@ilc.cnr.it

the Italian war. First of all, most international actions have focused on the Western front, generally regarding the Italian one as a sort of "side show" in comparison to what happened in France, the Flanders or Russia. Secondly, the process of digitization of historical documents in Italy notoriously lags behind.

Voci della Grande Guerra (Voices of the Great War)[1] is a scientific and cultural initiative with the aim of preserving and promoting the memory of Italy in WWI through the creation of a corpus of digital texts selected by historians and linguists in order to be representative of the different ways to experience and describe the Italian war by its protagonists. With the help of advanced techniques of computational linguistics, semantic web and information visualization, the digitized historical materials will be explored with an online interface to enable easy but effective and innovative search modalities. These will allow experts as well as non-experts to become acquainted with and appreciate the "linguistic polyphony" of Italy during WWI: the official voices of the propaganda and the voices of soldiers, the voices of newspapers and the voices of the letters, the intellectual elite's and the people's voices, the voices of consensus to war and the voices of dissensus.

This enterprise is extremely relevant both from the historical and the linguistic point 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. *Voci della Grande Guerra* thus aims at proving 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, in contact with more educated officers possessing a higher level of Italian. The comparison between the language varieties that will be represented in the *Voci della Grande Guerra* corpus 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 of rhetorical and literary models on the official language.

The project *Voci della Grande Guerra* is 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 has started in May 2016 and ends in May 2018. The project partners are:

- University of Pisa, Department of Philology, Literature and Linguistics, Computational Linguistics Laboratory – CoLing Lab (project coordinator);
- Istituto di Linguistica Computazionale CNR "Antonio Zampolli", Pisa;

1 http://vocidellagrandeguerra.it
2 http://www.centenario1914-1918.it

- University of Siena, Dipartimento di Scienze Storiche e dei Beni Culturali, Centro Interuniversitario di Studi e Ricerche Storico-Militari;
- Accademia della Crusca, Firenze.

The corpus and the digital platform for its navigation will be presented in a public cultural event organized at the Accademia della Crusa in the spring of 2018, devoted to the study of the linguistic varieties of Italian in the Great War.

1.1 Elements of innovation and expected results

Although some digital archives of Italian texts on the Great War are already available (e.g. *la Grande Guerra 1914 - 1918*),[3] these are typically limited to just to one text genre, mainly diaries. Moreover, the mere digitization of historical sources is not able by itself to guarantee the full access to their content. A real scientific breakthrough is offered by the application of natural language processing techniques to digitized historical sources, to enable truly semantic searches in the texts.

Voci della Grande Guerra presents at least four major elements of innovation with respect to the state of the art of digital text archives on WWI:

1. the project will create an archive of digital texts, most of which never digitized before, and belonging to a wide range fo 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. The digitized texts will be paired with the scanned images of the original documents, to guarantee a philological check of the sources;
2. state-of-the-art tools for computational linguistics, natural language processing and text mining will be used to annotate the digitized texts with semantic metadata that will enrich their informative value, multiplying the possibilities and modalities of accessing such information;
3. an online navigation tool will allow 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, despite in different forms and degrees;
4. the automatic methodology for text analysis and annotation will allow *Voci della Grande Guerra* to be not just a static and closed corpus, but rather an open and extendable digital platform for historical text analysis and processing.

The main expected results of *Voci della Grande Guerra* are:

- a substantial improvement of our knowledge of the structure and varieties of Italian language at the time of the Great War: How did Italians speak a century ago? How different was their language from ours? How did the

3 http://espresso.repubblica.it/grandeguerra/index.php

war change the language of the newborn Italian State? The project will contribute to answering these questions through a collection of digitized primary textual sources on WWI, enriched with semantic metadata;

- a digital platform to create, navigate and explore a digital archive, in order to evoke the wealth and polyphony of the voices represented therein;
- a public cultural event to promote the memory and knowledge of the Great War through the linguistic analysis of the texts produced by its protagonists (soldiers, intellectuals, politicians, civilians, etc.).

1.2 Impact of the initiative

Voci della Grande Guerra is an initiative characterized by a lasting impact and geared towards a wide audience:

- historians and linguists will cooperate in an interdisciplinary research focused on analyzing the "languages" of the Italians in the war, made possible by an innovative platform to process and access textual materials, which is going to be developed in the project. It is also worth highlighting that the corpus created by *Voci della Grande Guerra* will also be used to build the Great Vocabulary of Post-unification Italian under preparation by the Accademia dell Crusca;
- thanks to the exploration of the testimonies offered by linguistic texts in the new corpus, high school students and teachers will be able to reconstruct the complex and multifarious picture of the relationships between Italians and WWI, understanding at the same time its fundamental role in the formation of a linguistic and national cultural identity;
- history "buffs" and operators in cultural institutions such as museums and libraries will have a new resource to obtain an unprecedented view of the war. The digital platform of *Voci della Grande Guerra* will be used in cultural programs dedicated to the Great War, to provide new linguistic evidence coming from the "voices" of its actors.

2. Description of the project and realization phases

The project *Voci della Grande Guerra* has an estimated duration of 24 months and consists of the following stages:

Phase I - *Corpus design and text collection*
Under the supervision of Nicola Labanca (University of Siena) and Claudio Marazzini (Accademia della Crusca), the project will define the composition of the corpus. Given the practical impossibility of keeping track of all the varieties of the Italian of a century ago, a selection of the most relevant communicative situations will be made to characterize the language of the time:

- **the official military language**: the full collection of war bulletins, books of military strategy and analysis war conduct; propaganda texts and court martial records (Forcella and Monticone 1968);

- **the language of the middle class**: 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**: examples of letters, diaries and memoirs from soldiers;
- **the language of the political class**: samples of parliamentary proceedings, official speeches;
- **the language of the intellectual elite**: samples of pamphlets, literary journals, etc.;
- **the standard language of public opinion**: samples of newspaper articles, magazines, news reports from the front, etc.

Most of these sources are easily accessible, partially already digitized, in part to be scanned yet. For instance, the whole corpus of WWI bulletins has been already linguistically annotated in the project *Memorie di Guerra* (Boschetti et al. 2014).[4] Diaries and memoirs will be obtained through an agreement with *Biblioteca Nazionale centrale* in Firenze (in addition to the libraries of the partner institutions). For soldiers' letters and diaries, the project will resort to the collections already published by the *Museo storico italiano della guerra* in Rovereto e by the *Museo storico* in Trento, as well as collections of unpublished series of the Archives of popular writing. A comprehensive collection of acts of the Italian Parliament is available in Pisa, Florence and Siena. As for newspapers, we will sample already digitized newspapers (e.g., *La Stampa*, *Il Corriere della Sera*) and others that will be specifically digitized (e.g., *Avanti!*). Issues concerning the text copyrights will be carefully evaluated and addressed. We expect that most of the texts to be included in the corpus are free from copyright. If necessary, specific agreements will be made with the copyright owners.

The final corpus will be balanced along various dimensions: textual genres, author type, time, education, etc. The corpus will include texts from the 1913 up to the early '20s, 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. We will also balance the texts with respect to the various war years, in order to investigate empirically the immediate impact of the war and of its different phases (e.g. before and after Caporetto) on language and communicative styles.

Phase II - *Text digitization*
When not available in digital form, the texts selected for *Voices of the Great War* will be digitized with high resolution scanners and then analyzed with optical character recognition software (OCR). OCR performance is closely dependent on the quality of the text to be scanned. For this reason, the most advanced techniques of multiple OCR output alignment will be used with a"voting" system, already experienced in previous works on historical documents, to increase the accuracy of character recognition. The final output will be checked and corrected manually, and later codified in the TEI-XML standard format.

Phase III - *Automatic linguistic annotation of texts and information extraction*

4 http://www.memoriediguerra.it/wwm/

The digitized texts will undergo the following computational processing:

- **automatic analysis of the linguistic structure of the texts**: lemmatization, morphological and syntactic analysis, etc.;
- **semantic information extraction from texts** - Extraction of simple terms (e.g., *irredentismo* "irredentism") and complex terms (e.g., *terre irredente* "unredeemed lands", *gas asfissianti* "poisonous gas", etc.) significantly associated with different texts types; named entity recognition (recognized named entities will include person names like *Luigi Cadorna*, location names like *Ortigara*, military units like *9 Reggimento Bersaglieri*, etc. Other potentially relevant semantic categories will be identified with the help of historians); recognition of dates and events, etc. The locations mentioned in the texts will also be normalized with respect to spelling variations and associated with their geographic coordinates;
- **text indexing with the extracted information**, in order to enrich the texts with semantic metadata for advanced content analysis and search.

Automatic text annotation and information extraction will be performed with existing natural language processing tools for Italian. This activity will greatly benefit from the experience gathered in the project *Memorie di Guerra* (Boschetti et al. 2014). In particular, we expect a significant accuracy drop of the annotation tools trained on contemporary standard Italian, since the texts to be processed will contain lots of old-fashioned lexical items and syntactic constructions that may hamper linguistic annotation. In order to overcome these problems and increase accuracy, we will resort to self-training and active learning methods to adapt the tools to the various types of texts and language varieties to be analyzed. Passaro and Lenci (2015) adapted an existing named entity recognizer for Italian to annotate WWI bulletins, extending the standard repertoire of semantic classes to military units, ships and airplanes.

The linguistic annotations of about 1 million tokens will also be checked manually, thereby representing a sort of "gold" subset of the whole corpus. The remnant of the corpus will instead annotated automatically with a random checking for errors. The texts belonging to the "gold" and to the "silver" subsets will be marked with metadata, so that users will be fully aware of the "noise" to be expected in their text searches. The "gold" corpus will also be used in the process of domain adaptation of the annotation tools.

Phase IV - *Online tools for corpus annotation and exploration*
The project will also develop a software platform to assist researchers during the phases of corpus building, and to provide various search functionalities. The tool will consist 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):

- free text searches (e.g., words, lemmas, complex terms, etc.);
- searches for semantic categories (e.g. names of persons, locations, military units, etc.).

- frequency analysis of simple and complex terms in the texts (e.g. similar to *Google Ngram Viewer*);[5]
- possibility to browse texts starting from geographical maps marked with the places referred to therein;
- creation of "event timelines" to search texts belonging to different periods of the war.

The interface usability will be evaluated by a team of historians coordinated by Nicola Labanca. Usability evaluation will also involve: i) high schools, to encourage multidisciplinary training programs supported by new digital technologies; ii.) at least one museum and library, to test the *Voci della Grande Guerra* platform within educational tours and lifelong learning programs.

Phase V - *Organization of the cultural and scientific event to present the corpus and platform Voci della Grande Guerra*
The project *Voci della Grande Guerra*, its corpus and tools will be presented at a public event organized at the Accademia della Crusca, in Florence, involving historians of the Great War and linguists. In addition to present the project results, the event will be an opportunity for an interdisciplinary analysis on the role of the Great War in the history of the Italian language, and on the importance of linguistic evidence from textual memories for the understanding of the First World War.

3. Conclusions

Voci della Grande Guerra is an innovative project in digital humanities, applying advanced computational linguistic and natural language processing techniques to create the first large-scale annotated corpus of Italian texts about WWI. Therefore, the innovation of the project lies both in the content of the corpus, and in the scientific methodology used to build and explore it.

The types of texts targeted by *Voci della Grande Guerra* raise a great number of challenges to natural language processing and text analytics methods:

- highly noisy data, because of OCR errors;
- sub-standard or ill-formed linguistic expressions due to poor alphabetization (e.g., spelling errors, ungrammatical constructions, etc.);
- diachronic variation (e.g., spelling, lexical and syntactic differences);
- text genre variation (e.g., newspapers vs. diaries vs. letters).

Voci della Grande Guerra will deal with these challenges by applying and developing methods for text annotation and information extraction from noisy texts and adaptive to language variation (e.g., diachronic, diastratic, etc.).

Even if time and resources inevitably set the limits of the corpus size, the methodology and tools developed by the project will allow scholars to further expand and enrich the text archive. The project springs from the fruitful collaboration of historians

5 https://books.google.com/ngrams

and linguists, as a virtuous example of the potentialities of computational linguistics for digital humanities.

References

Boschetti, Federico, 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 Heritag" (LRT4HDA 2014)*, pages 70–75, Reykjavik.

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

Forcella, Enzo and Alberto Monticone. 1968. *Plotone di esecuzione. I processi della Prima Guerra Mondiale*. Laterza, Bari.

Moretti, Franco. 2013. *Distant Reading*. Verso, London.

Passaro, Lucia and Alessandro Lenci. 2015. "Il piave mormorava…": Recognizing locations and other named entities in italian texts on the great war. In *Proceedings of the First Italian Conference on Computational Linguistics*, pages 286–290, Pisa.