Voci della Grande Guerra
Preserving the Digital Memory of World War I

Alessandro Lenci[1], Nicola Labanca[3], Claudio Marazzini[4], Simonetta Montemagni[2], Federico Boschetti[2], Irene De Felice[1], Stefano Dei Rossi[5], Felice Dell'Orletta[2], Michele Di Giorgio[1,2,3], Lucia Passaro[1], Giulia Venturi[2]

[1] Università di Pisa, Laboratorio di Linguistica Computazionale (CoLing Lab), Pisa, Italia
[2] Istituto di Linguistica Computazionale - CNR, Italia
[3] Università di Siena, Dipartimento di Scienze Storiche e Beni Culturali
[4] Università di Siena, Dipartimento di Scienze Storiche e Beni Culturali
[5] WebSoup

ABSTRACT
Voci della Grande Guerra 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 an annotated 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.

KEYWORDS
Computational linguistics, Digital history, World War I, annotated corpus

1. INTRODUCTION
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. Increasing amounts of such sources are available in digital form, but many historical documents still need to be digitized. *Voci della Grande Guerra* (Voices of the Great War: http://www.vocidellagrandeguerra.it/) is a scientific and cultural initiative with the aim of preserving and promoting the memory of Italy in World War I (WWI) through the creation of an annotated 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, 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. The project *Voci della Grande Guerra* is part of the official initiatives for the Celebrations of the 100th Anniversary of WWI: it started in May 2016 and will end in September 2018.

2. ASSEMBLING THE VOICES OF WWI
The project *Voci della Grande Guerra* 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. 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 [2], 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.
Voci della Grande Guerra aims at creating an archive of digital texts, most of which never digitized before. Given the practical impossibility of keeping track of all the varieties of the Italian of a century ago, the corpus consists of a selection of texts representative of the most relevant communicative situations to characterize the language of the time:

- the official military language: books of military strategy, analysis war conduct, senior officers' memoirs and diaries; propaganda texts and court martial records;
- the language of the middle class: samples of low rank officers' diaries and memoirs, most likely written in a high-level Italian inspired to major literary examples of the time;
- the popular language: examples of letters, diaries and memoirs from soldiers (and from the common people);
- the language of the political class: samples of parliamentary proceedings, official speeches and secret sessions of the parliament;
- 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.

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.
In the great majority of cases, when not available in digital form, the texts selected for *Voices of the Great War* are 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 are used with a "voting" system, already experienced in previous work on historical documents, to increase the accuracy of character recognition. The final output is checked and corrected manually, and later codified in the TEI-XML standard format. Collaborative proof-reading involves not only the members of the project, but also volunteers, through the WikiSource platform, in collaboration with the Biblioteca del Comune di Trento.

The digitized texts undergo the following computational processing [1]:

- 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 land", *gas asfissianti* "poisonous gas", etc.) significantly associated with different texts types, and named entity recognition (recognized named entities will include person names like *Luigi Cadorna*, location names like *Ortigara*, military units like *9 Reggimento Bersaglieri*, etc). The locations mentioned in the texts are also normalized with respect to spelling variations and associated with their geographic coordinates. Named entities will be linked to existing knowledge bases such as DBpedia;

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 be annotated automatically with a random checking for errors. The pilot project *Memorie di Guerra* (War Memories: http://www.memoriediguerra.it) [1] shows in action a part of the tools and of the analyses that will be applied to the new, extended corpus.

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 in 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 [3].

3. ACKNOWLEDGEMENTS

The project *Voci della Grande Guerra* is funded by a two-year grant from the Special Mission for the Celebrations of the 100th Anniversary of WWI at the Presidenza del Consiglio dei Ministri of the Italian Government.

4. REFERENCES

[1] 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 Heritage*" (LRT4HDA 2014), Reykjavik.
[2] De Mauro, Tullio. 1963. *Storia linguistica dell'Italia unita*. Laterza, Bari.
[3] Moretti, Franco. 2013. *Distant Reading*. Verso, London.