Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II

Federico Boschetti[*], Andrea Cimino[*], Felice Dell'Orletta[*], Gianluca E. Lebani[†], Lucia Passaro[†], Paolo Picchi[*], Giulia Venturi[*], Simonetta Montemagni[*], Alessandro Lenci[†]

[*]Istituto di Linguistica Computazionale "Antonio Zampolli", CNR- Pisa (Italy)
[†]CoLing Lab, Dipartimento di Filologia, Letteratura e Linguistica, University of Pisa (Italy)
E-mail: federico.boschetti@yahoo.com, andrea.cimino@ilc.cnr.it, felice.dellorletta@ilc.cnr.it, gianluca.lebani@for.unipi.it, lucia.passaro@for.unipi.it, paolo.picchi@ilc.cnr.it, giulia.venturi@ilc.cnr.it, simonetta.montemagni@ilc.cnr.it, alessandro.lenci@ling.unipi.it

Abstract

World War (WW) I and II represent crucial landmarks in the history on mankind: They have affected the destiny of whole generations and their consequences are still alive throughout Europe. In this paper we present an ongoing project to carry out a computational analysis of Italian war bulletins in WWI and WWII, by applying state-of-the-art tools for NLP and Information Extraction. The annotated texts and extracted information will be explored with a dedicated Web interface, allowing for multidimensional access and exploration of historical events through space and time.

Keywords: Digital History, War Bulletins, NLP, World War I, World WAR II, Information Extraction

1. Introduction

World War (WW) I and II represent crucial landmarks in the history of mankind: They have affected the destiny of whole generations and their consequences are still alive worldwide. Unfortunately, the knowledge of these events is progressively fading away, especially among young generations. The first centenary of the beginning of the 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. Methods and tools for Natural Language Processing (NLP) can play an important role to achieve this goal, by providing new way to access historical documents and events (cf. Ide and Woolner 2004, Cybulska and Vossen 2011).

In this paper we present an ongoing project to carry out a computational analysis of Italian war bulletins in WWI and WWII, by applying state-of-the-art tools for NLP and Information Extraction. The project relies on the collaboration between computational linguists and historians, in particular Prof. Nicola Labanca (University of Siena), one of the major experts on Italian military history during the WWs. This project has several elements of originality and challenge. To the best of our knowledge, this is the first computational analysis of this kind of historical texts. Moreover, WWI Italian war bulletins have never been digitalized before. The type of language (Italian of first half of the 20th century) and domain (military) require an intense effort of adaptation of existing NLP tools. Bulletins are annotated automatically with different types of information, such as simple and multi-word terms, named entities, events, their participants, time and georeferenced locations. The annotated texts and extracted information will be explored with a dedicated Web interface, allowing for multidimensional access and explorations of historical events through space and time.

The paper is organized as follows. In the next section, we provide an overview of the project, as well as a motivation of the text choice. In section 3, we describe some of the current works, mainly focusing on the digitalization of the WWI bulletins, the adaptation of existing NLP tools, the first experiments for term and event extraction carried out on WWII bulletins. In section 4 we present the next steps for the project implementation and some future plans.

2. Project overview

2.1 War bulletins and NLP

War bulletins (WBs) were issued by the Italian *Comando Supremo* "Supreme Headquarters" during WWI and WWII as the official daily report about the military operations of the Italian armed forces. WBs were published on major newspapers, and during WWII they were also radio broadcasted. WBs provide a dynamic picture of the unfolding of war events, from the official perspective of the Italian Government. They allow us to follow the complex series of events in the two WWs, respectively from the 24th May 1915 to the 11th November 1918, and then from the 10th June 1940 to the 8th September 1943 (date of the armistice between Italy and the Allies, and the dissolution of the Italian army). It is important to remark that WBs do not provide a faithful and objective picture of the war. Events can be missing or misrepresented for military or propaganda reasons. For instance there is a systematic overestimation of enemy losses and Italian achievements, and conversely the underestimation of Italian defeats and losses.

We have focused our research on the computational analysis of WBs because they represent a particularly interesting source not only to reconstruct the military history of the two WWs, but also to study the propaganda strategies of the Italian Government, as well as the way the enemy was depicted by official sources. The collection of WBs for WWI was first published in 1923, and the one for WWII was published in 1970. The former has never been digitalized, while an html version of the latter is freely available on the Web.[1]

The reason for working simultaneously on WWI and WWII is twofold. First of all, nowadays historians commonly assume that, despite their several differences, WWI and WWII should not be regarded as two separated events, but rather as two episodes of a single 30-years European war. Secondly, the comparison between the bulletins of the two WWs is extremely interesting under many respects. From an historical point of view, we can observe the radical change in warfare between the two conflicts: for instance, WWI was mainly a static trench war, while WWII was a movement war fought on as different fronts as Greece, North Africa, Atlantic Ocean, etc. Some weapons, like gases, represent the hallmark of WWI, but were not used in WWII, which was instead dominated by tanks and aviation. These types of information easily emerge from the analysis of WBs. Moreover, during WWI Italy had a liberal, aristocratic government, while in WWII it was ruled by the dictatorial fascist regime. This difference has important consequences on the language, propaganda style, etc. to be found in WBs.

2.2 Work program

Our project of computational analysis of WBs include the following phases:

1. **text digitalization of WWI bulletins**. This phase is currently ongoing, and described in section 3.1.
2. **NLP annotation of WB**. The corpus of bulletins will be POS-tagged, lemmatized and dependency parsed with existing tools for Italian NLP. Waiting for the complete digitalization of WWI bulletins, we have started processing the WWII ones. This part of the project also involves an important work on NLP tools adaption to the target domain and genre, as illustrated in section 3.2.
3. **Statistical analysis and information extraction.** This is the core and most challenging part of the project. Its goal is to index texts with a large amount of linguistic and semantic information, to highlight significant text features as well as to identify the most prominent historical events. This part includes:
    a. *statistical profiling* – each bulletin will be assigned textual statistics like number of word tokens, type-token ratio, readability scores, etc. This information is also extremely useful to characterize the development of historical events. For instance, shorter bulletins correspond to periods of less intense military operations, successful operations are described in greater details, while defeats are usually reported in a more sketchy form, etc.;
    b. *term extraction* – simple and multi-word terms will be automatically extracted from texts, to provide users more advanced search keys. This activity is carried out by adapting existing term extraction tools (cf. section 3.3.);
    c. *named entity recognition* – proper names will be identified and classified into general (e.g., person, location, etc.) and domain-specific (e.g., ship, military unit, airplane, etc.) semantic categories (cf. section 3.4);
    d. *event extraction* – using standard Information Extraction techniques (cf. in particular Ide and Woolner (2004), Cybulska and Vossen (2011) for previous research on event extraction from historical texts) we will identify instances of major event types (e.g., *bombing*, *sinking*, *battles*, etc.) their participants, places and times. Event timestamps will be derived from the bulletin date, and will be used to reconstruct the event timeline.
4. **data linking** – various types of extracted data will be linked to external sources. First of all, location names will be georeferenced. This is particularly challenging given the spelling variations of many location names (e.g., Arabic ones), as well as the changes that some of them have undergone through time (e.g. various WWI locations were Italian at that time, and have now become Slovenian, etc.). Moreover, we plan to provide links between other extracted data external sources (e.g., Wikipedia pages describing weapon types or warships, etc.).
5. **browse and search interface** – a key aspect of the project is the development of an advanced and user-friendly interface to explore the texts. Our target users are not limited to scholars but also crucially include students. We intend to provide multidimensional access keys to WBs, which will be queried for simple words, multi-word terms, semantic classes, event types, locations, and persons, etc. The tools will also include user-friendly visualization modules such as "word clouds", event timeline, even projection on dynamical geographical maps, etc. The interface will allow experts as well as non-expert users to follow historical events in space and time, thereby gaining a new view of the parallel development of war actions across multiple fronts.

3. Ongoing work

3.1 Text digitalization of WWI bulletins

We are currently digitalizing the WBs of WWI published in *I Bollettini della Guerra 1915-1918*, preface by Benito Mussolini, Milano, Alpes, 1923 (pages VIII + 596). The exemplar in our possession is preserved in a

[1] http://www.alieuomini.it/pagine/dettaglio/bollettini_di_guerra

good state, due to the high quality of paper and printing. The book has been accurately unbound, in order to acquire images from loose pages. This technique drastically reduce scanning artifacts, avoiding the necessity to digitally straighten and deskew page images. The same conditions of brightness and contrast are assured not only to each page of the book, but also to each area of the page, increasing the accuracy of the Optical Character Recognition (OCR).

OCR has been performed using the open source application Tesseract, in bundle with the Italian training set.[2] The accuracy and the F-score have been calculated on a random sample of 10 pages out of 604. In order to calculate accuracy and F-score, the texts of the sample have been manually corrected and aligned to the OCR output applying the Needleman-Wunsch algorithm, in order to identify the number of exact matches, the number of substitutions (e.g., "m" instead of "n"), the number of omissions (i.e., characters neglected by OCR) and, finally, the number of insertions (i.e., artifacts added by the OCR engine, such as an "i" at the end of the line).

Accuracy is defined as the ratio between the matches and the sum of the matches (m) with all the other phenomena, substitutions (s), insertions (i) and deletions (d). Precision (P) is defined as $m/(m+s+i)$, Recall (R) as $m/(m+s+d)$, and F-score as $2PR/(P+R)$. The accuracy on the test sample is 97.87% and the F-score is 98.68%. OCR performances will be improved by using three different OCR engines: the aforementioned Tesseract accompanied by OCRopus[3] and Gamera.[4] The results will be aligned and a voting system will be applied, according to the methods described in Lund-Ringger (2009) and in Boschetti et al. (2009). Finally, manual corrections will be performed.

3.2 Text processing and NLP tools adaptation

Parallely to the digitalization of WWI bulletins, we are carrying out experiments for the automatic linguistic annotation of WWII bulletins with NLP tools. The bulletins were automatically downloaded and cleaned of HTML tags and boilerplates. The resulting corpus was automatically POS tagged with the Part-Of-Speech tagger described in Dell'Orletta (2009) and dependency-parsed with the DeSR parser (Attardi et al., 2009) using Support Vector Machines as learning algorithm. They represent state-of-the-art tools for Italian NLP. In particular, the POS tagger achieves a performance of 96.34% and DeSR, trained on the ISST–TANL treebank (consisting of articles from newspapers and periodicals), achieves a performance of 83.38% and 87.71% in terms of Labeled Attachment Scores (LAS) and Unlabeled Attachment Scores (UAS) respectively when tested on texts of the same type.

However, since Gildea (2001) it is widely acknowledged that statistical NLP tools have a drop of accuracy when tested against corpora differing from the typology of texts on which they were trained. This is also the case with WBs: they contain lexical and syntactic structures characterising the Italian of the past century and they contain domain-specific lexicon. Sentences are typically shorter than in newspapers, but on the other hand they are often quite elliptic and full of omissions due to the telegraphic style. They also contain lots of old-fashioned syntactic constructions that may hamper linguistic annotation. The percentage of lexical items contained in the WWII corpus and in the training of DeSR parser (74%) is much lower than in the corpus of contemporary newspaper articles used as test set (about 90%). We expect this trend to be even in WWI bulletins, since Italian of early 20th century was very different from contemporary one. In fact, standard Italian was still very much under formation, due to the recent political unification of the country just 50 years before the Great War. Assuming that new lexicon introduces new syntactic constructions, we can assume that the parser tested on Bulletins can have a quite high drop of accuracy with respect to the accuracy achieved on the reference test set.

In order to overcome this problem, in the last few years several methods and techniques have been developed to adapt current NLP systems to new kinds of texts. They can be broadly divided in two main typologies: Self-training (McClosky et al., 2006) and Active Learning (Thompson et al., 1999). To adapt NLP tools to WBs, we are using the self-training approach to domain adaptation described in (Dell'Orletta et al., 2013), based on ULISSE (Dell'Orletta et al., 2011). ULISSE is an unsupervised linguistically-driven algorithm to select reliable parses from the output of dependency annotated texts. Each dependency tree is assigned a score quantifying its reliability based on a wide range of linguistic features. After collecting statistics about selected features from a corpus of automatically parsed sentences, for each newly parsed sentence ULISSE computes a reliability score using the previously extracted feature statistics. From top ranked parses according to their reliability score, different pools of parses are selected for training. The new training set extends the original one with the new selected parses including lexical and syntactic characteristics specific to the target domain, in this case the bulletins. We expect that the NLP tools trained on this new training set can improve their performance when tested on the target domain.

3.3 Term extraction

Single–word terms, e.g. *velivolo* (aircraft), and multi–word terms (complex terms), e.g. *velivolo da ricognizione marittima* (*maritime reconnaissance aircraft*) are the first types of information we intend to extract from WBs. They will be used for text indexing and querying. We are currently applying to WWII bulletins two methods for automatic term extraction from Italian texts, T2K[2] (Dell'Orletta et al., 2014) that follows

[2] https://code.google.com/p/tesseract-ocr/downloads/list

[3] https://code.google.com/p/ocropus

[4] http://gamera.informatik.hsnr.de

the methodology described in (Bonin et al. 2010) and EXTra (Passaro et al. 2014), both combining NLP techniques, linguistic and statistical filters.

Term extraction with T2K^2 is articulated in three main steps. In the first step, the POS-tagged and lemmatized text is searched for on the basis of linguistic filters aimed at identifying a) nouns, expressing candidate single terms and b) POS patterns covering the main morphosyntactic patterns expressing candidate complex terms: e.g., noun + adjective (e.g., *velivoli britannici*, British aircraft), noun + preposition + noun (e.g. *velivolo d'assalto, Aircraft of Assault*), etc. In the second step, the candidate terms are ranked according to their C-NC Value, a statistical filter described in (Frantzi et al. 1999) and (Vintar 2004). C-value is a method for term extraction which aims to improve the extraction of nested terms. The method produces a list of candidate terms that are ordered by their termhood. Then, the NC-value incorporates context information to the C-value method, improving term extraction. In the last step, a contrastive method is applied against the list of ranked terms using a contrastive function *CSmw* newly introduced in (Bonin et al. 2010). This function is applied to the top list of the terms resulting from the statistical filtering step. This procedure is oriented to a) prune common words from the list of domain-relevant terms and b) rank the extracted terms with respect their domain relevance. This method is based on the comparison of the distribution of terms across corpora of different domains. We also used T2K^2 to extract relevant domain-specific verbs representing instances of some of the major events. Table 1 reports a sample of the top-ranked domain verbs extracted with T2K^2 using contemporary newspaper collections as reference corpora in the contrastive method.

mitragliare	“to machine-gun”
spezzonare	“to bomb with incendiary devices”
bombardare	“to bomb”
abbattere	“to shoot down”
silurare	“to torpedo”
incendiare	“to set on fire”
affondare	“to sink”
attaccare	“to attack”

Table 1 – Sample of domain verbs extracted with T2K^2

Term extraction with EXTra is carried out in a very similar way, except for two major differences. First of all, instead of using flat POS-sequences, EXTra identifies candidate multi-word terms with structured patterns that take into account the internal syntactic structure of term phrases. For instance, the term *bomba di grosso calibro* “heavy bomb” is identified as an instance of the pattern *[noun, preposition [adjective, noun]]*, while the term *apprestamenti difensivi del nemico* “enemy defensive works” is identified as an instance of the pattern *[noun, adjective, preposition [noun]]*. Pattern structure is then used to guide the process of statistical term weighting. Terms are weighted using a new measure that recursively applies standard association measures (e.g., Pointwise Mutual Information, Local Mutual Information, Log-Likelihood Ratio, etc.) to the internal structure of complex terms. The intuition is that the degree of termhood of a candidate pattern depends not only the statistical association between its parts, but also on whether these parts are also terms. The EXTra term weighting algorithm works as follows:

i) *base step* - we measure the association strength σ of each candidate 2-word term $<w_1,w_2>$, and we then select the set of terms $T=\{t_1,\ldots,t_n\}$ whose score σ is above an empirically fixed threshold;

ii) *recursive step* - we measure the association strength σ of any *n*-word candidate term $<c_1,c_2>$, where either c_1, or c_2 or both belong to T:

$$\sigma(<c_1,c_2>) * S(c_1) * S(c_2)$$

where $S(c_i)=1$, if c_i is a word, while $S(c_i) = (\log_2 \sigma(c_i))/k$ if $c_i \in T$. The parameter *k* controls the length of complex terms: The smaller the *k*, the higher weight is assigned to longer terms. The candidate terms whose score σ is above an empirically fixed threshold are then added to T. The recursive step ii) is repeated for any extracted pattern, so that multi-word terms of any length are assigned a weight. Table 2 reports a sample of the top ranked complex terms extracted with EXTra from the WWII bulletins, using Local Mutual Information (as the association score σ (Evert 2008).

term	LMI
fronte greco	927.30
tenente di vascello	699.04
lieve danno	659.14
aereo nemico	623.10
capitano di corvetta	593.89
artiglieria contraerea	548.13
bomba di grosso calibro	500.12
velivolo nemico	496.32
bollettino odierno	456.01
caccia germanico	441.91
obiettivo militare	423.78
campo di aviazione	422.07
vasto incendio	416.86
caccia tedesco	413.63
piroscafo di medio tonnellaggio	366.60

Table 2 – Sample of terms extracted with EXTra

Extracted domain-specific entities are then organized into fragments of taxonomical chains, grouping entities which share the semantic head (e.g., *fronte cirenaico, fronte egiziano, fronte greco, fronte tunisino, fronti dello scacchiere, fronti* terrestri share the semantic head *fronte* “front”) or the modifiers defining their scope (eg. a*ereo britannico*, *apparecchio britannico*, *aviazione britannica*, *caccia britannico*, *incursione aerea britannica, velivolo britannico* share the modifier *britannico* “British”).

3.4 Named Entity Recognition

WBs report military events and therefore are full of proper names of places, persons and organizations (e.g. military formations). Therefore, NER plays a crucial role to obtain a semantic access to the content of these texts. The named entities are identified and classified using ItaliaNLP NER (described in Dell'Orletta et al., 2014). This module is a classifier based on Support Vector Machines using LIBSVM (Chang and Lin, 2001) that assigns a named entity tag to a token or a sequence of tokens. ItaliaNLP NER relies on 5 kinds of features:

- *orthographic features*: e.g., capitalized letters, presence of non–alphabetical characters, etc.;
- *linguistic features*: the lemma, POS, prefix and suffix of the analyzed token;
- *dictionary look-up features*: check if the analyzed token is part of an entity name occurring in People, Organization and Geo-political gazetteers;
- *contextual features*: these features refer to orthographic, linguistic and dictionary look–up features of the context words of the analyzed tokens;
- *non local features*: in the case of identical tokens , they take into account previous label assignments to predict the label for the current token (Ratinov and Roth, 2009).

ItaliaNLP NER is trained on I-CAB (Italian Content Annotation Treebank) (Magnini et al., 2006), the dataset used in the NER Task at EVALITA 2009 (Speranza, 2009) including four standard named entity tags, i.e. Person, Organization, Location and Geopolitical classes. The NE tagger accuracy is in line with the state of the art when compared with the systems that participated to the EVALITA shared task.(F-Measure: ~80%).

For the analysis of the Italian, we are currently adapting the NER under various respects. First of all, we plan to extend the range of semantic classes covered by the NER, for instance identify names of airplanes (e.g., *Gloster Gladiator*), ships (e.g., *Valiant*), and military organizations (e.g., *Divisione Ariete* “Ariete Division”) which frequently occur in this type of texts. Moreover, we intend to adapt the NER to the domain of WBs. To this purpose we plan to exploit the rich analytical index accompanying WWII bulletins. This index contains the names of places, persons and military formations mentioned in the WBs, with information about its semantic class. Using this index, we are automatically identifying named entities in the bulletins, thereby producing a fully annotated version of the WWII corpus with NE classes. This corpus will be used to train the NER before its application to the WWI bulletins (which instead lack this type of semantic indexing).

4. Conclusions and future plans

The short-term agenda of our project includes: i.) completing the digitalization and manual correction of the WWI bulletins; ii.) developing the module for event extraction and location georeference of WWII texts; iii.) applying the NLP and Information Extraction modules to WWI texts; iii.) designing and implementing the Web search interface. The output of text processing will consist of the two bulletin corpora annotated with XML and RDF metadata indexing texts with various types of linguistic and semantic information.

This project has also a great possibilities for future, long-term developments. First of all, we plan to enrich the linking of WBs to other types of external data. As we said above, WBs provide a very biased view of military events, because of propaganda or military reasons. It is therefore interesting to perform a kind of cross-text event co-reference to link the events extracted from the WBs to other historical sources reporting information about the same historical fact. Secondly, we intend to extend our project to cover WBs issued by other countries involved in WWI and WWII. This will again allow us to gain information about the way the same events are reported by different fighting countries (e.g., the battle of El Alamein described by the Axis Powers or the Allies). Computational linguistic methods have surely a great potential for applications on historical text. We believe that a project like ours can contribute to prove the possibilities offered by NLP to find new ways to study our past and to learn history.

Acknowledgements

We gratefully thank Prof. Nicola Labanca for his constant advice and support on the military history of World War I and II.

References

Attardi, G., Dell’Orletta, F., Simi, M., Turian, J. (2009). Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In *Proceedings of EVALITA 2009*, Reggio Emilia, Italy.

Bonin, F., Dell’Orletta, F., Montemagni, M., Venturi, G. (2010). A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta, pp. 19-21.

Boschetti, F., Romanello, M., Babeu, A., Bamman, D. and Crane, G. (2009). Improving OCR accuracy for classical critical editions. In *Proceedings of the 13th European conference on Research and advanced technology for digital libraries (ECDL'09)*, Berlin, Heidelberg: Springer-Verlag, pp. 156-167.

Chang C-C., Lin, C-J. (2001). LIBSVM: a library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Cybulska, A., and Vossen, P. (2011). Historical Event Extraction from Text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, Portland, OR, USA, pp. 39-43.

Dell’Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. In *Proceedings of EVALITA 2009*, Reggio Emilia, Italy.

Dell’Orletta, F., Venturi, G., Cimino, A., Montemagni, S.

(2014). T2K^2: a System for Automatically Extracting and Organizing Knowledge from Texts. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik (Iceland) (forthcoming).

Dell'Orletta, F., Venturi, G., Montemagni, S. (2011). ULISSE: an unsupervised algorithm for detecting reliable dependency parses. In *Proceedings of CoNLL 2011*, Portland, Oregon, pp. 115-124.

Dell'Orletta, F., Venturi, G., Montemagni, S. (2013). Unsupervised Linguistically-Driven Reliable Dependency Parses Detection and Self-Training for Adaptation to the Biomedical Domain. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013)*, Sofia, Bulgaria, pp. 45-53.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

Frantzi, K., Ananiadou, S. (1999). The C–value / NC Value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*, 6(3), pp. 145-179.

Gildea, D. (2001). Corpus Variation and Parser Performance. In *Proceedings of EMNLP 2001*, Pittsburgh, PA, pp. 167-202.

Ide, N., and Woolner, D. (2004). Exploiting Semantic Web Technologies for Intelligent Access to Historical Documents. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisboa, Portugal, pp. 2177-2180.

Lund, W. B., and Ringger, E. K. (2009). Improving optical character recognition through efficient multiple system alignment. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries (JCDL '09)*. ACM, New York, NY, USA, 231-240.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Bartalesi Lenzi, V., Sprugnoli, R. (2006). I-CAB: the Italian Content Annotation Bank. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

McClosky, D., Charniak, E., Johnson, M. (2006). Reranking and self-training for parser adaptation. In *Proceedings of ACL 2006*, Sydney, Australia, pp. 337-344.

Passaro, L., Lebani, G. and Lenci A. (2014), Extracting terms with EXTra. submitted.

Ratinov, L., Roth, D. (2009). Design challenges and misconceptions in named entity recognition, In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, pp. 147-155.

Thompson, C. A., Califf, M. E., Mooney, R. J. (1999). Active Learning for Natural Language Parsing and Information Extraction. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML'99)*, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc., pp. 406-414.

Vintar, Š. (2004). Comparative Evaluation of C-value in the Treatment of Nested Terms. In *Proceedings of Memura 2004 – Methodologies and Evaluation of Multi-word Units in Real-World Applications*, (LREC 2004 Workshop), pp. 54-57.