Francesco Mambrini, Marco Passarotti, Caroline Sporleder

Preface

In almost every culture of the world, the sciences of language and the study of cultural-heritage documents are inextricably bound. In Western tradition, as in many others, the need to preserve the literary or historical legacy of the past gave the strongest input to the development of a formalized grammatical speculation.

Since the foundation of linguistics as a discipline, the interaction has proceeded in both directions. Linguistics has profited from the huge amount of material that was gathered by philologists and historians, along with the full apparatus of concepts and problems that originated from their work. Humanities, in their turn, have often seen in linguistics a model of a rigorous scientific approach to a social and historically complex phenomenon like human language.

It is not by chance, thus, that the work of scholars engaged in historical and literary studies was not alien to one of the most original development in contemporary linguistics, namely the creation and use of the first digital corpora. It is worth remembering that the *Index Thomisticus*, which is considered the starting point for both corpus-linguistics and digital humanities, was designed in order to allow a more rigorous approach to the philosophy of Thomas Aquinas.

From the time of the first pioneering projects, the concepts and methodologies of corpus linguistics (including the notion of "corpus" itself) have been widely debated; technologies for storing and processing digital information have also changed radically. Nowadays, computational and corpus-linguistics have grown into autonomous disciplines, with their own set of required expertises. As in many other scientific fields, autonomy means inevitably a certain degree of isolation.

The loss of contact between corpus-linguistics and humanities is particularly visible in one crucial aspect. Although quantitative or stylometric approaches to large collections of documents are increasingly frequent in literary or historical studies, the available resources are not quite at the same level as those used by linguists.

The "Workshop on Annotation of Corpora for Research in Humanities" (ACRH) was held in Heidelberg University on January 5^{th}. The event was co-located to the 10^{th} edition of the international workshop on "Treebanks and Linguistic Theories" (TLT-10), also held in Heidelberg on January 6^{th}-7^{th}. The ACRH workshop was conceived to address one special aspect where the aforementioned gap between the two disciplines is particularly visible: the creation and exploitation of annotated corpora for the needs of research in the different fields of humanities (philology, literary studies, history, philosophy etc.).

The availability of annotated corpora is indeed an area where humanities and corpus-linguistics are more distant. Many corpora that play a relevant role for research in humanities are today available in digital format (theatrical plays, contemporary novels, critical literature, literary reviews etc.). Yet, only a few of them are linguistically tagged, while most still lack any linguistic annotation. Standards for the encoding of a vast number of information on texts used in the humanities and social sciences, such as the TEI, are becoming increasingly popular and are adopted for the most recent efforts of digitization of collections. But although there is an agreement on the meta-language for the description of texts, what features an annotated corpus for research in the humanities must have and how annotation must be performed in order to conform to the strict requirements described by corpus-linguistics is a question that is still not sufficiently debated. Fostering this discussion was precisely the aim of our workshop.

As the work for the creation of annotated resources is in most cases still in a very early stage, it is only natural that most of the attention is focused on the annotation process. There are a number of peculiarities that distinguish corpora created for the special use of research in humanities from the standard linguistic corpora, such as the Brown Corpus or the Penn Treebank. These special features, which affect both the theoretical debate and the practical task of developing tools for the work of corpus creation, are well reflected by the papers that are collected here.

Two sets of problems arise from the peculiar nature of the documents used by scholars in the humanities, and these issues affect both the work of manual and (semi-)automatic annotation.

On the one hand, even the preliminary task of selecting and digitizing the relevant materials for the corpus must tackle peculiar problems, that can be generally ignored in the case of standard linguistic corpora. Often, as in the case of ancient or badly preserved documents, even the “raw” materials are controversial or need special philological care. Different kinds of expertise are required in order to build a corpus of such documents. This multidisciplinary program, though, can be too vast and too difficult to handle on a large scale. In such situations, the general competence of a native speaker with linguistic training (which is typically sufficient for the annotation of standard linguistic corpora) is inadequate; in most cases linguistic competence must be complemented by a strong training on the particular problems, such as philological, historical, epistemological issues, raised by the materials. New standards and models of manual annotation must be devised in order to integrate the different competences of corpus linguists and specialists for the texts. The challenge of adapting tools and concepts for manual annotation is manifest in several papers presented at ACRH (see especially the contributions by: Manca *et al.*; Bech and Eide; Kaplan *et al.*; Korkiakangas and Passarotti).

On the other hand, automated and semi-automated annotation is difficult due to another special feature of corpora of cultural-heritage documents. The high level of intra-linguistic variation, that is observed, especially in diachronic corpora, is clearly a crucial factor. Almost all the available NLP tools are developed for modern corpora and tested on standardized varieties of a languages. Yet the documents collected for their cultural significance often predate the establishment of modern writing conventions. In other cases, historical documents can reflect a large spectrum of regional dialects that are not yet unified in a national standard language; other times, literary documents are deliberately violating the rules of ordinary languages. How NLP tools (like parsers or pos-taggers) can be effectively adapted to this situation is a crucial practical question, which in its turn can provide new evidence to scholars working in historical linguistics or literary studies. This is perhaps the most debated topic of the workshop (see especially the papers by: Dipper; Hendrickx and Marquilhas; Sukhareva *et al.*; Rögnvaldsson *et al.*; Ekbal *et al.*; Piotrowski and Höfler; Skjærholt).

The motivation for creating and annotating the collections is also a factor that must be kept in mind. The need of preserving endangered cultural-heritage documents often precedes (and determines) the scientific motivation of creating resources for scholars. The representativeness of those corpora is another difficult question which is tied to that of the motivations. Very often, for corpora of historical languages or corpora designed for specific research purposes on narrow domains in the humanities, there is not much to be chosen; corpus designers have rather to deal with the scanty material that is available. This is another important difference with standard linguistic corpora, where the size of the collection can be determined in advance and adjusted to the needs of researchers. The small extension of the corpora used for research in the humanities is a big challenge both for the quantitative analysis of data and for the performances of the NLP tools, of which many of the papers presented here are well aware. There are many other questions that remain open to discussion. What and how to annotate is another crucial problem in a pluralistic and often very controversial area like the different fields of the humanities. The dilemma between the adherence to a strict theoretical frame and the needs of portability of theoretically independent resources is of course not unknown to linguists, but is even more challenging for scholars in the humanities.

A typical example of resources discussed in many papers of ACRH is a corpus created by digitizing collections of manuscripts (e.g. see the paper by Dipper), archives of hand-written papers (e.g. Hendrickx and Marquilhas), or ancient printed books. In such cases, the textual content that must be tagged is only a part of information that define these cultural artifacts. Linguistic analysis must therefore be associated with other types of annotation, that should encode aspects such as description of e.g. the medium, the script, the collocation of the text in the page, and so on. The interaction of these different levels of information and with the standard markup languages whose aim is the comprehensive description of textual artifacts (of which TEI is perhaps the most widespread) is a problem which requires further attention.

The quantity and quality of the submitted papers are an unmistakable proof of the great interest in the area of annotated digital corpora for humanities.

The call for papers for ACRH requested unpublished, completed work. We received 23 submissions. The submissions were authored by researchers from 15 different countries in America, Asia and Europe. Each submission was evaluated by three reviewers. The Programme Committee consisted of 18 members (including the 3 co-chairs) from 8 different countries. They all worked as reviewers. Based on their scores and comments on the content and quality of the papers, 14 papers were accepted for presentation and publication, which corresponds to an acceptance rate of 58.3%. 8 papers were presented orally, 5 as posters; finally, 1 paper, which was proposed and accepted to both workshops (ACRH and TLT), was considered by the two Programme Committees to be more appropriate for presentation at TLT10.

As we have mentioned, the accepted submissions cover a wide range of topics. The languages studied are: English, French, German, Greek, Icelandic, Italian, Japanese, Latin, Portuguese and Spanish. Some of the corpora that were discussed in the paper cover a significant variety of diachronic phases of these languages; the corpora presented at ACRH extend over a very broad time span (from Antiquity, to Early and Late Middle-ages, to Modern era and our days) and come from different domains (religious texts, private letters, epic poems and chronicles etc.).

The ACRH Workshop was introduced by a keynote lecture by Professor Gregory R. Crane, editor in chief of the Perseus Project and chair of the Classics Department of Tufts University (Medford, USA). Professor Crane has long distinguished himself in pursuing a high-level scholarly activity in the field of Greek and Latin literature; but, especially, the Perseus Project has been providing classicists of all levels and from all over the world with access to the primary sources for their work (both texts and archaeological artifacts) within a digital cyberinfrastructure that relies largely on the implementation of specific NLP technologies. Recently, the Perseus Project has undertaken the creation of the first treebank for Classical Latin and Ancient Greek (The Greek and Latin Dependency Treebank). While actively promoting the work of annotation to the community of scholars and students in Classics, Professor Crane is particularly engaged in showing how the inclusion of a treebank into the larger context of a cultural heritage library could benefit both the digital library and the annotated corpus itself.

The event was made possible by the work and generosity of the local organizers and hosts at the University of Heidelberg. Our gratitude and acknowledgements go to them. We would also like to thank to board of the Journal for Language Technology and Computational Linguistics for accepting the publications of the ACRH proceedings.

The ACRH Co-Chairs and Organizers
Francesco Mambrini, Università Cattolica del Sacro Cuore, Milan, Italy
Marco Passarotti, Università Cattolica del Sacro Cuore, Milan, Italy
Caroline Sporleder, Saarland University, Saarbrücken, Germany