Preface

Research in the Humanities is predominantly text-based. For centuries scholars have studied documents such as historical manuscripts, literary works, legal contracts, diaries of important personalities, old tax records etc. Manual analysis of such documents is still the dominant research paradigm in the Humanities. However, with the advent of the digital age this is increasingly complemented by approaches that utilise digital resources. More and more corpora are made available in digital form (theatrical plays, contemporary novels, critical literature, literary reviews etc.). This has a potentially profound impact on how research is conducted in the Humanities. Digitised sources can be searched more easily than traditional, paper-based sources, allowing scholars to analyse texts quicker and more systematically. Moreover, digital data can also be (semi-)automatically mined: important facts, trends and interdependencies can be detected, complex statistics can be calculated and the results can be visualised and presented to the scholars, who can then delve further into the data for verification and deeper analysis. Digitisation encourages empirical research, opening the road for completely new research paradigms that exploit `big data' for humanities research. This has also given rise to Digital Humanities (or E-Humanities) as a new research area.

Digitisation is only a first step, however. In their raw form, electronic corpora are of limited use to humanities researchers. The true potential of such resources is only unlocked if corpora are enriched with different layers of linguistic annotation (ranging from morphology to semantics). While corpus annotation can build on a long tradition in (corpus) linguistics and computational linguistics, corpus and computational linguistics on the one side and the Humanities on the other side have grown apart over the past decades. We believe that a tighter collaboration between people working in the Humanities and the research community involved in developing annotated corpora is now needed because, while annotating a corpus from scratch still remains a labor-intensive and time-consuming task, today this is simplified by intensively exploiting prior experience in the field. Indeed, such a collaboration is still quite far from being achieved, as a gap still holds between computational linguists (who sometimes do not involve humanists in

developing and exploiting annotated corpora for the Humanities) and humanists (who sometimes just ignore that such corpora do exist and that automatic methods and standards to build them are today available).
ACRH-2 aims to foster communication and collaboration between these two groups, in the same way that its predecessor ACRH-12 did. ACRH-12 was held at Heidelberg University on January 5, 2012, in conjunction with the 10th edition of the international workshop on "Treebanks and Linguistic Theories" (TLT-10). ACRH-2 is again co-located with TLT, this time at the University of Lisbon. We received thirteen submissions for ACRH-2. After a thorough reviewing process eight submissions were included in the workshop, addressing several important issues related to corpus annotation for the Humanities.

The papers in the proceedings concern several different topics. The task of resource creation is tackled by Koeva et al., who present an aligned parallel Bulgarian-English corpus for linguistic research, and Ferreira et al., who introduce a novel framework for annotating corpora with a particular focus on language documentation. Four papers are concerned with corpora of historical texts which pose particular challenges for language processing software. A major problem are spelling variations. Detecting and normalising these is addressed by two papers: Bollmann test several string distance methods for Early New High German, while Reynaert et al. compare two state-of-the-art error detection systems on old Portuguese. Historical texts also often lack consistent punctuation, which poses difficulties for automatic segmentation into linguistic units. The paper by Petran presents a method for segmenting texts that lack punctuation marks into sentences, clauses and chunks. In turn, Bouma and Hermans introduce an algorithm for syllabification in Middle Dutch text. Finally, two papers are concerned with deeper processing problems. Both focus on folktale corpora. Everhardus et al. present an approach for normalisation and consistency checking in semi-structured corpora, while Karsdorp et al. address the task of identifying actors and ranking them by importance. The workshop programme is completed by an invited lecture by Martin Wynne, who heads the Oxford Text Archive and has worked extensively in the areas of corpus linguistics and corpus infrastructure development.

We are grateful to everybody who made this event possible, including Erhard Hinrichs, the local and non-local organisers of TLT-11 (in particular Iris Hendrickx), the ACRH-2 programme committee, and the authors who submitted papers. We also acknowledge the endorsement of the AMICUS network.

The ACRH-2 Co-Chairs and Organisers
Francesco Mambrini (University of Cologne, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (Trier University, Germany)