Theory and Applications
of Natural Language Processing

Aims and Scope

The field of Natural Language Processing (NLP) has expanded explosively over the past decade: growing bodies of available data, novel fields of applications, emerging areas and new connections to neighboring fields have all led to increasing output and to diversification of research.

"Theory and Applications of Natural Language Processing" is a series of volumes dedicated to selected topics in NLP and Language Technology. It focuses on the most recent advances in all areas of the computational modeling and processing of speech and text across languages and domains. Due to the rapid pace of development, the diversity of approaches and application scenarios are scattered in an ever-growing mass of conference proceedings, making entry into the field difficult for both students and potential users. Volumes in the series facilitate this first step and can be used as a teaching aid, advanced-level information resource or a point of reference.

The series encourages the submission of research monographs, contributed volumes and surveys, lecture notes and textbooks covering research frontiers on all relevant topics, offering a platform for the rapid publication of cutting-edge research as well as for comprehensive monographs that cover the full range of research on specific problem areas.

The topics include applications of NLP techniques to gain insights into the use and functioning of language, as well as the use of language technology in applications that enable communication, knowledge management and discovery such as natural language generation, information retrieval, question-answering, machine translation, localization and related fields.

The books are available in printed and electronic (e-book) form:

* Downloadable on your PC, e-reader or iPad
* Enhanced by Electronic Supplementary Material, such as algorithms, demonstrations, software, images and videos
* Available online within an extensive network of academic and corporate R&D libraries worldwide
* Never out of print thanks to innovative print-on-demand services
* Competitively priced print editions for eBook customers thanks to MyCopy service http://www.springer.com/librarians/e-content/mycopy

For other titles published in this series, go to
www.springer.com/series/8899

Caroline Sporleder • Antal van den Bosch
Kalliopi Zervanou
Editors

Language Technology for Cultural Heritage

Selected Papers from the LaTeCH Workshop Series

Springer

Editors
Caroline Sporleder
Computational Linguistics/MMCI
Saarland University
P.O. Box 15 11 50
66041 Saarbrücken
Germany
csporled@coli.uni-sb.de

Antal van den Bosch
Tilburg center for Cognition
and Communication
Tilburg School for Humanities
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands
Antal.vdnBosch@uvt.nl

Kalliopi Zervanou
Tilburg center for Cognition
and Communication
Tilburg School for Humanities
Tilburg University
P.O. Box 90153
5000 LE Tilburg
The Netherlands
k.zervanou@uvt.nl

ISSN 2192-032X e-ISSN 2192-0338
ISBN 978-3-642-20226-1 e-ISBN 978-3-642-20227-8
DOI 10.1007/978-3-642-20227-8
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011932565

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The task I set myself in this Foreword is to sketch out an historical context for the contributions to Language Technology for Cultural Heritage so as to illumine their wider intellectual significance. The problems they identify are fascinating in and of themselves, and the work on which they report brings most welcome benefits to the cultural heritage sector. But beyond the technical fascinations and the new affordances is a slower moving, much less immediately visible shift in relations between techno-scientific and humanistic ways of knowing. That's the huge subject at which I take a momentary glance here.

It would be bad historiography to say that the metamorphic device now less and less known as "the computer" is the cause of this shift. It is far better, I think, to regard the device as among the most prominent strands in a complex ravelling and unravelling of developments about which we can only speculate—or, as Hugh Kenner did in *The Counterfeiters* (2005/1968), write "an historical comedy" [17]. But without doubt the machine which brings us together here is a powerful and influential part of a great change.

To call our device "the computer" (singular noun with definite article) can be seriously misleading, though the convenience this term offers is at times irresistible. I succumb here repeatedly. It is wrong for two reasons. First, Michael S. Mahoney taught us, there's little that is singular about the machinery derived from Alan Turing's scheme, which specifies an indefinitely large number of actual machines; however finite they are, their number is limited only by the human imagination [21, 23]. Second, Alan Newell and Herbert Simon taught us, what counts is the symbol manipulation, not the calculation [37]. We must return to the etymological sense of "computer" – L. *cum* "with" + *putare* "put, place", paying attention to the implicit kinaesthesis – to make sense of computing more broadly, especially given the growing interest in gestural interfaces, e.g. as depicted in *Minority Report* (2002) and implemented on the iPad. What's happening is not just some game of logic in the head.

Do I digress? Hardly, and not to bring up anything very new. Here is Terry Winograd and Fernando Flores in the Preface to *Understanding Computers and Cognition* [53]:

> All new technologies develop within the background of a tacit understanding of human nature and human work. The use of technology in turn leads to fundamental changes in what we do, and ultimately what it is to be human. We encounter deep questions of design when we recognize that in designing tools we are designing ways of being. (xi)

I want to explore some of these deep questions of design ever so briefly now.

Clearing a Space for New

There seem to be moments when an emergent way of thinking or acting must be separated definitively from its origins so that it may be seen not as deviant but in its own terms, as something new, and so have a chance to survive. Here I can only indicate a temporal sequence suggestive of the historical meaning I want to draw out.

Early in the third century Tertullian of Carthage proclaimed the absolute divorce of Christianity from its pagan forbear, sternly asking, *Quid ergo Athenis et Hierosolymis? quid academiae et ecclesiae?*, "What then have Athens and Jerusalem in common? what have the Academy and the Church?" (*de praescr. haeret.* 7.9). In a similar but opposite act of separation, the Renaissance scholars who were nicknamed "humanists" (*humanistae*) spurned those they caricatured as the "schoolmen" (*scholastici*) so as to establish the *literae humaniores*, the concerns of man, in distinction to the *literae divinae*, the concerns of theology [5, 22–4, 35–8]. Then, Galileo Galilei founded quantitative, scientific epistemology by separating it from the qualitative (which we have come to champion in the humanities), declaring the book of nature to be written not in authoritative words but "in the language of mathematics… without which it is humanly impossible to understand a single word of it" (*The Assayer*, 1623). And again, more than three centuries later, in his Rede Lecture at Cambridge, physicist, public servant and novelist Sir Charles Percy Snow defended science (culturally weak in mid-century Great Britain) by declaring it a distinct, more vigorous and progressive culture than that of the privileged humanities [47]. His rhetorical act drew upon a tradition of distinguishing the humanities from the sciences going back at least to Wilhelm Windelband's contrast of law-seeking and particularizing disciplines [52]. The debate Sir Charles kindled with The Two Cultures in 1959 has died down and flared up more than once since he spoke, its remarkable persistence worth note. To my mind it has been most persuasively re-articulated by psychologist Jerome Bruner as a matter of divergent but similarly imaginative trajectories [4]. Bruner's account, along with several others, makes distinctions that clarify relations and so help us in a bridge-building whose time, it seems, has come.

Computer and Human

The computer (allow me this backsliding) sits ambiguously, significantly in the middle, not unlike the nineteenth century technologies of communication that alternately modelled and provided models for human physiology — telegraph and nervous system, for example [40, 46]. Thus Alan Turing began his foundational article with the actions of a man doing his sums [49, 231]. In turn Warren McCulloch and Walter Pitts used Turing's scheme to model the brain [31]. Their model then informed John von Neumann's sketch of digital computing architecture [36][1], which we follow to this day. Its structure and functions have subsequently permeated the neurological and cognitive sciences. Biological ideas, neurophysiological, evolutionary and genetic, have subsequently influenced developments in computing [22]. Feedback and feed-forward, as the cyberneticists said.

For me at least the iconic image of this telling traffic between human and machine is a microphotograph of brain cells grown in tissue culture on a Motorola 68000 chip, taken by Toronto neurologist Judy Trogadis to illustrate her colleague John K. Stevens' feature article in Byte Magazine [48]; cf. [43, 92]. Since that photograph was taken, the juxtaposition Trogadis and Stevens engineered to illustrate their – and our – preoccupation with human-machine relations has been turned into a mass-produced tool for connecting neuronal and nano-electronic circuits[2]. Imagination and implementation in a virtuous circle, or rather, progressive spiral.

A Slow and Halting Progress

The imprinting of machine by human and human by machine would seem to favour computing as an obvious means for bridging the human and non-human sciences. Bridging would seem to be implicit in a tool that gives Galilean epistemology some purchase on human cultural artefacts and returns to students of these artefacts the benefits of the knowledge thus obtained. But realising the potential has not been without its delays, difficulties, mistakes and false directions. This is, in other words, a story of a long struggle that seems now to be paying off.

At the outset not so, however. For quite obvious reasons of background, education and disciplinary specialisation, few scholars in the humanities had the training to get engaged with computing in the early years or even to see the possibilities. The technical expertise required first to build and then to use those early machines excluded most if not all non-scientists from direct involvement. But computers were hardly unknown even to the least technically inclined during the incunabular period (the time from the end of World War II to the public release of the Web in 1991).

[1] For von Neumann's use of McCulloch and Pitts see [30, 31] The borrowing is made obvious by the neurophysiological vocabulary of the Draft Report [36].

[2] See the work of the National Research Council Canada (www.nrc-cnrc.gc.ca/eng/education/innovations/spotlight/brain.html) and the Nanobio Convergence Laboratory, Interuniversity Microelectronics Centre, Belgium (www.imec.be, reported in Science Daily, 26 November 2009).

In fact digital computing was a hugely popular subject, widely if not always accurately reported in the popular media throughout the period. A catalogue of items in the news or otherwise publicised would swamp this Foreword, but a few examples will give you an idea.

Notice of the new "electronic brain", as the computer was popularly known (not without reason, we have seen), appeared in *The Times* of London in November 1946. Two years later IBM put its Selective Sequence Automatic Calculator (SSEC) in the front window of its World Headquarters in Madison Avenue, New York, where it remained until 1952. Passers-by nicknamed it "Poppa"; the *New Yorker*'s "Talk of the Town" featured it in 1950. That same year *Time Magazine*, which paid close attention to computers from the beginning, ran a cartoon of the Harvard Mark III on its cover. Kits for children went on the market soon thereafter: by 1955, if not earlier, the GENIAC, a "simple electric brain machine"; by 1963 the Digi-Comp 1, "first real operating digital computer in plastic". The next year the Toronto *Globe and Mail* featured "Computers: The new age of miracles: hundreds of brains in a thimble" (16 November), imagining, in terms astonishingly similar to current dreams of a "semantic web", a world made ever so convenient to family life.

Hence by the mid 1960s, at least, humanists – except those without children, neighbours, magazines, newspapers, radio, television or spouses aware of the world – could hardly plead ignorance.

University computer centres were established for scientific research following rapid commercialisation of the computer in the 1950s. But digital computing had already been started in or near the humanities, in two projects first conceived ca. 1949: Fr Roberto Busa's *Index Thomisticus* [7] and Machine Translation, proposed in a memo by Warren Weaver [50] and then funded lavishly by the American government for its Cold War purposes. Within the following ten to fifteen years a relatively small cohort of humanists had taken to computing with great enthusiasm, as is attested for example by literary scholar Jess B. Bessinger, Jr., in his Foreword to Literary Data Processing Conference Proceedings, sponsored by IBM in September 1964 [2], and by the articles and reports on activities across the disciplines in Computers and the Humanities, founded in 1966 by another literary scholar, Joseph Raben.

Not all embraced the computer so readily, even when it emerged as a general purpose machine programmable in languages such as Fortran and Cobol. Stigma was attached to involvement with it. Indeed, through the 1980s association with computing could delay if not end a young scholar's academic career or stain the reputation of a senior academic. The severely negative reaction can be attributed to a number of causes apart from mere resistance to change: the hype from "early adopters" as well as salesmen; prominent use by the military during the Cold War, which spanned the entire incunabular period and profoundly affected daily life, especially in the United States [10, 51]; and perhaps most significant of all, the challenge to human identity of a device that from the beginning was thought one day to be capable of thinking. Even before digital computing moved Herman Goldstine and John von Neumann to define programming as "the technique of providing a dynamic background to the automatic evolution of meaning" [14, 2] electronic devices had begun to seem disturbingly, autonomously intelligent, for example the

Sperry gyropilot, known as "Metal Mike" [34, 72], and Norbert Wiener's Anti-Aircraft Predictor, which spooked engineer George Stibitz when he visited Wiener's laboratory in 1942 [12, 242–3].

In the humanities and among public intellectuals expressions of unease, even fear, echoed those in the popular literature. Only some of this had to do with Cold War paranoia and with the threats to jobs from automation and to personal autonomy from mechanisms of surveillance and control. The sense of being made insignificant, even redundant – "a threat less defined than [those others] but even more profound, arising out of the alleged capacity of these machines to develop into Homo Sapiens clones" [42, ix] – suggests that the cultural force Sigmund Freud had attributed to his own work a generation earlier could be applied to digital computing as well: a deeply disturbing "and most irritating insult... flung at the human mania of greatness" [11, 246f]; cf [29, 305]. Freud had put psychoanalysis in the lineage of two great predecessors, Copernicus, who displaced humans from an imagined centrality in the physical cosmos, and Darwin, whose "dangerous idea" unseated humankind from uniqueness in the living world [9, 26]. To that list of great insults we must now add Turing's. It is true that the machines propagated from his scheme have become furniture of ordinary life and are now taken for granted when not entirely undetected. But at the same time they have informed every aspect of how we think and talk about being human. The metaphors computing has provided us seem irresistible. Computers have made possible scientific research that continues eating away at the mania which Freud attacked with his revelations of the profound degree to which we do not know who and what we are.

Apart from Busa and a few other early figures (such as computational linguist Margaret Masterman[3] and literary critic Louis Milic), the potential of computing to enable radically new work was clear to a number of collaborating artists, engineers and artist-engineers of the early period, especially in Great Britain. They were not frightened away. Their enthusiastic and insightful projects, pronouncements and writings leave us in no doubt of this [3]. But for reasons yet adequately to be identified and explored, the time was not right. Just as computing entered the humanities, for example, those most involved were marooned professionally by the shift of the disciplines that computing could most readily serve, away from a concern with scholarly data (in literary studies, "close reading") to that which Jonathan Culler has called "just plain 'theory"' [8][4]. Most of those who persisted with computing, Milic complained at the time, were only mechanising what they conceived to be drudgery rather than using scholarly data to probe the unknown or to puzzle out what the new machine might be, and what it might be capable

[3] See her uncollected contributions to the Times Literary Supplement and especially [25]. She was a vigorous proponent of the kinds of experimental, imaginative work advocated by Milic [33] and widely spurned by the academic establishment, e.g. [19].

[4] Anthony Kenny has made the fascinating suggestion that the turn from a focus on texts to a preoccupation with theory in mainstream humanities research just as computing entered the scene might have been a negative reaction to all that computing represented – precisely to its power for symbol-manipulation [18, 9–10]. I remain suspicious of such simple, cause-effect formulations, however.

of [6, 33]. On the one hand the computer was widely assumed or thought to be a servant or slave, on the other to be imminently capable of enslaving humankind[5]. In other words, computing got caught in a master/slave dialectic. In an anonymous TLS review (probably written by Sir Charles Geoffrey Vickers, lawyer and pioneering systems scientist), the reader was warned that to regard computing in this way would be to bury its intellectual potential. This potential, the author wrote, could help resolve "the major epistemological problem of our time": "[w]hether and, if so, how the playing of a role differs from the application of rules which could and should be made explicit and compatible" [1]

Literary critic Jerome McGann has argued persuasively that for those disciplines focused chiefly on interpreting cultural artefacts, the major emphasis of the digital humanities for the last many years does little itself to liberate this potential [32], though it clearly benefits conventional research by supplying resources in convenient form. Exactly how best to engage computing directly in turning the hermeneutic circle remains an open question and the most difficult of challenges for the digital humanities to consider. Those other disciplines primarily devoted to reporting on, cataloguing and providing access to cultural heritage, such as epigraphy, are at present much better served.

The same year as that stern warning against enslavement, W. G. Runciman wrote consolingly in the TLS series Thinking by Numbers about the disappointing results from computational studies in sociology, recommending greater patience than a single generation or lifetime could measure [45, 943]. His recommendation remains a good antidote against the all the talk of great paradigm shifts and the hype that goes with it[6]. Steady progress of hardware and software together with online resources have in the intervening years slowly rendered some of the very hard computational problems of our cultural artefacts somewhat easier. At a similar pace, alarmingly less in anyone's spotlight, computing has changed how we and our students study, use and think about those artefacts. The chief cause of this change, I would argue, is not great analytic breakthroughs, not directly anyhow, but the theoretically simple and unsophisticated fact of access to great quantities of material. The "distant reading" Franco Moretti has described and which Mark Olsen pointed to several years earlier is one such new "condition of knowledge" brought about by quantitative access [35, 39, 56-8]. Another, anecdotal evidence suggests, is the rude juxtaposition of disciplines by keyword searches for articles e.g. in JSTOR, which implicitly brings the possibility of interdisciplinary interconnections

[5] The drudgery of computation could be very real and was certainly thought intolerable by the likes of Leibniz and Babbage, hence their common solution: the mechanical servant [42, 20–44]; [13, 8ff]. The anti-aircraft problem of World War II likewise made computation by hand unsupportable, hence stimulated development of machines for the task. The error I am describing occurs, however, when we identify the calculational power of computing machines as their essential nature, and having equated that power with intelligence then anthropomorphize our relationship to the machine as that between slave and master.

[6] The unthinking importation of Thomas S. Kuhn's idea of a "paradigm shift", from *The Structure of Scientific Revolutions*, brings with it the assumption of a complete break from one way of thinking to another incommensurable with it. However well that works for physics, it seems a highly dubious notion for the humanities.

into view and so encourages their exploration. Given constraints of time this in turn pushes research to go wide rather than deep, with implications Richard Rorty has opened up [44]. This is a largely unexplored question, it would seem: how actually to do genuine interdisciplinary research on one's own, or to put the matter pedagogically, how to train our students responsibly to handle what is already being thrust at them.

Humanities and Sciences

New analytical tools for the humanities, though slower to develop and still at a highly primitive stage, are advancing apace, as several of the contributions here suggest. My concluding question is, given the long, cultural view, what status do these tools have within the humanities? What are they doing to research?

Elsewhere I have argued that these tools create a conjectural space within the humanities in which cultural artefacts can be operated on *as if* they were natural objects [28]. This argument proceeds from the fact that to make a cultural artefact computationally tractable it must be rendered as data. Since data are qualitatively indifferent as to source, scientific methods of analysis apply. That which is lost in the rendering, and so does not affect the analysis, can then be brought into consideration by comparing the results with the researcher's pre-existing ideas, changes made and the hermeneutic cycle repeated. Thus modelling in the humanities [27, 20–72]. Meanwhile – and here is my overriding point – such analytic practices in the digitised humanities have implicitly established what Geoffrey Lloyd has called a "beachhead of intelligibility" joining the humanities with the sciences [20]. Borrowing liberally from Ian Hacking's work on "styles of scientific reasoning", I have argued that in effect computing has given us *humanistae* a way of tapping into centuries of scientific work and wisdom in our employment of these styles [15].

Again: the important matter here is that beachhead of intelligibility, or what I have called the bridge-building that the computer has greatly strengthened if not made possible. Earlier I devoted space to the incunabular fears of the machine in the humanities. I did so not simply to help explain our rather halting progress toward this time of bridge-building but to shine a light on the bridge under construction. Even if we no longer write articles entitled "Fear and Trembling: The Humanist Approaches the Computer" [38] or feel compelled to reassure ourselves that in the face of computing the scholar can still find "work which only he can accomplish" [41], we still, indeed especially, need to be most acutely aware – not to that fear (which continues)[7] but to that to which fear is a less than helpful reaction: the defamiliarizing perception of changed epistemic conditions.

[7] The fact that the American Association of Artificial Intelligence felt moved two years ago to convene a meeting to worry over "potential long-term societal influences of AI research and development" is perhaps evidence enough that familiarity has not superseded fear but only blanketed it [16]; see [24], which however intemperate makes the point).

These which follow are not just scientific papers. They are components of the bridge now visible to any who care to look.

London, January 2011 *Willard McCarty*

References

1. Anon: Keepers of rules versus players of roles. In: Rev. of Thomas L. Whisler, The Impact of Computers on Organizations, and James Martin and Adrian R. D. Norman, The Computerized Society., *Times Literary Supplement*, vol. 21, p. 585 (1971)
2. Bessinger, Jr., J.B.: Foreword. In: J.B. Bessinger, Jr., S.M. Parrish, H.F. Arader. (eds.) Literary Data Processing Conference Proceedings, pp. 1–2. IBM Corporation, Armonk NY (1964)
3. Brown, P., Gere, C., Lambert, N., Mason, C.: White Heat Cold Logic: British Computer Art 1960-1980. MIT Press, Cambridge MA (2008)
4. Bruner, J.: Possible castles. In: Actual Minds, Possible Worlds, pp. 44–64. Harvard University Press, Cambridge MA (1986)
5. Burke, P.: A Social History of Knowledge. Polity, London (2000)
6. Busa, R.: Guest editorial: Why can a computer do so little? Bulletin of the Association for Literary and Linguistic Computing **4**, 1–3 (1976)
7. Busa, R.: The annals of humanities computing: The index thomisticus. Computers and the Humanities **14**, 83–90 (1980)
8. Culler, J.: Literary theory: A very short introduction. In: Very Short Introductions. Oxford University Press, Oxford (1997)
9. Dennett, D.C.: Darwin's Dangerous Idea: Evolution and the Meanings of Life. Simon & Schuster, New York (1995)
10. Edwards, P.N.: The Closed World: Computers and the Politics of Discourse in Cold War America. MIT Press, Cambridge MA (1996)
11. Freud, S.: Eighteenth lecture. general theory of the neuroses: Traumatic fixation – the unconscious. In: A General Introduction to Psychoanalysis, pp. 236–47. Boni and Liveright, New York (1920)
12. Galison, P.: The ontology of the enemy: Norbert Wiener and cybernetics. Critical Inquiry **21**(1), 228–66 (1994)
13. Goldstine, H.H.: The Computer from Pascal to von Neumann. Princeton University Press, Princeton NJ (1972)
14. Goldstine, H.H., von Neumann, J.: Planning and coding of problems for an electronic digital computer. Report on the Mathematical and Logical aspects of an Electronic Computing Instrument, Part II, Vol. I-3, IAS ECP list of reports, 1946-57. 4, 8, 11, Institute for Advanced Study, Princeton NJ (1947)
15. Hacking, I.: 'Style' for historians and philosophers. Historical Ontology pp. 178–99 (2002)
16. Horvitz, E., Selman, B.: Interim report from the panel chairs, AAAI Presidential Panel on Long-Term AI Futures (2009). URL www.aaai.org/Organization/Panel/panel-note.pdf
17. Kenner, H.: The Counterfeiters: An Historical Comedy. Dalkey Archive, Scholarly Series. Dalkey Archive Press, Normal IL (2005/1968)
18. Kenny, A.: Computers and the humanities. The Ninth British Library Research Lecture. The British Library, London (1992)
19. Leavis, F.R.: 'Literarism' versus 'Scientism': The misconception and the menace, *Times Literary Supplement 3556 (23 April)*, vol. Repub. 1972 in Nor Shall My Sword: Discourses on Pluralism, Compassion and Social Hope, 137-60, pp. 441–5. Chatto & Windus, London (1970)

20. Lloyd, G.: History and human nature: Cross-cultural universals and cultural relativities. Interdisciplinary Science Reviews **35**(3-4), 201–14 (2010)
21. Mahoney, M.S.: The roots of software engineering. CWI Quarterly **3**(4), 325–34 (1990)
22. Mahoney, M.S.: Historical perspectives on models and modeling. In: XIIIth DHS-DLMPS Joint Conference on "Scientific Models: Their Historical and Philosophical Relevance. Zürich (2000). URL www.princeton.edu/~hos/Mahoney/articles/models/models.htm
23. Mahoney, M.S.: The histories of computing(s). Interdisciplinary Science Reviews **30**(2), 119–35 (2005)
24. Markoff, J.: Scientists worry machines may outsmart man. New York Times (2009). URL www.nytimes.com/2009/07/26/science/26robot.html
25. Masterman, M.: The intellect's new eye. Freeing the Mind. Times Literary Supplement **284** (1962)
26. Mayr, E.: This is Biology: The Science of the Living World. Belknap Press, Harvard University Press, Cambridge MA (1997)
27. McCarty, W.: Humanities Computing. Palgrave, Houndmills, Basingstoke (2005)
28. McCarty, W.: Being reborn: The humanities, computing and styles of scientific reasoning. In: W.R. Bowen, R.G. Siemens (eds.) New Technologies and Renaissance Studies, *Medieval and Renaissance Studies and Texts, Vol. 324*, vol. 1, pp. 1–22. Iter Inc. in collaboration with the Arizona Center For Medieval and Renaissance Studies, Tempe AZ (2008)
29. McCorduck, P.: Machines who think: A personal inquiry into the history and prospects of artificial intelligence. W. H. Freeman and Company, San Francisco CA (1972)
30. McCulloch, W.S.: What is a number, that a man may know it, and a man, that he may know a number? In: Embodiments of Mind, Intro. Seymour Papert, Pref. Jerome Y. Lettvin, The Ninth Alfrew Korzybski Memorial Lecture. MIT Press, Cambridge MA (1988/1961)
31. McCulloch, W.S., Pitts, W.H.: A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics **5**, 115–33 (1943)
32. McGann, J.: Marking texts of many dimensions. In: R.S. Susan Schreibman, J. Unsworth (eds.) A Companion to Digital Humanities, Blackwell Companions to Literature and Culture, pp. 198–217. Blackwell Publishing, Oxford (2004). URL www.digitalhumanities.org/companion/
33. Milic, L.T.: The next step. Computers and the Humanities **1**(1), 3–6 (1966)
34. Mindell, D.A.: Between Human and Machine: Feedback, Control, and Computing before Cybernetics. Johns Hopkins Studies in the History of Technology. Johns Hopkins University Press, Baltimore MD (2002)
35. Moretti, F.: Conjectures on world literature. New Left Review **1**, 54–68 (2000)
36. von Neumann, J.: First draft of a report on the EDVAC. Tech. rep., Moore School of Electrical Engineering, University of Pennsylvania, Philadelphia PA (1945)
37. Newell, A., Simon, H.A.: Computer Science as Empirical Inquiry: Symbols and Search - 1975 Turing Award Lecture. Communications of the ACM **19**(3), 113–26 (1976)
38. Nold, E.W.: Fear and trembling: The humanist approaches the computer. College Composition and Communication **26**(3), 269–73 (1975)
39. Olsen, M.: Signs, symbols and discourses: A new direction for computer-aided literature studies. Computers and the Humanities **27**, 309–14 (1993)
40. Otis, L.: Networking: Communicating with Bodies and Machines in the Nineteenth Century. University of Michigan Press, Ann Arbor MI (2001)
41. Pegues, F.J.: Editorial: Computer research in the humanities. The Journal of Higher Education **36**(2), 105–8 (1965)
42. Pratt, V.: Thinking Machines: The Evolution of Artificial Intelligence. Basil Blackwell, Oxford (1987)
43. Reddy, R.: The challenge of artificial intelligence **29**(10), 86–98 (1996)
44. Rorty, R.: Being that can be understood is language. London Review of Books pp. 23–5 (2000)
45. Runciman, W.G.: Thinking by numbers: 1. Times Literary Supplement pp. 943–4 (1971)
46. Sappol, M.: Dream Anatomy. National Institutes of Health, National Library of Medicine, Washington DC (2006). URL www.nlm.nih.gov/dreamanatomy/

47. Snow, C.P.: The Two Cultures. Intro. Stefan Collini. Cambridge University Press, Cambridge (1993/1959)
48. Stevens, J.K.: Reverse engineering the brain. Byte Magazine, special issue on Artificial Intelligence pp. 286–99
49. Turing, A.M.: On computable numbers, with an application to the entscheidungsproblem. In: Proceedings of the London Mathematical Society, *2*, vol. 42, pp. 230–65 (1936)
50. Weaver, W.: Translation. In: W.N. Locke, A.D. Booth (eds.) Machine translation of languages: fourteen essays, pp. 15–23. Technology Press, Massachusetts Institute of Technology, Cambridge MA (1949/1955)
51. Whitfield, S.J.: The Culture of the Cold War, 2nd edn. Johns Hopkins University Press, Baltimore MD (1996)
52. Windelband, W.: History and natural science. In: History and Theory: Classics in the Philosophy of History. Transl. with intro. Guy Oakes, vol. 19, pp. 165–85 (1980/1894)
53. Winograd, T., Flores, F.: Understanding Computers and Cognition: A New Foundation for Design. Addison-Wesley, Boston MA (1986)

Contents

Language Technology for Cultural Heritage, Social Sciences and Humanities: Chances and Challenges

Caroline Sporleder, Antal van den Bosch and Kalliopi Zervanou

1 From Quill and Paper to Digital Knowledge Access and Discovery

For the most part of their long history, the Social Sciences and Humanities (SSH) have essentially been pen and paper based disciplines. Researchers worked on paper-based data sources (manuscripts, books). Collections of such sources were catalogued in inventory books, or on file cards in libraries and archives. Museums and other cultural heritage (CH) institutions also tended to manage their collections by entering metadata about their exhibits in register books. Researchers who wanted to gain access to an artefact or document had to find the object's entry in a paper-based inventory, look up its location, and then retrieve the object from the museum's depot or archive.

The advent of computers and information technology (IT) changed the methods for data access. As a first step, many cataloging systems carrying object metadata were computerised, and digital databases replaced register books and file cards. Locating an object of interest not only became faster, but also easier, since digital registries allow for indexing along several dimensions and facets (e.g., title words, author, keywords, publication year).

While digital cataloguing systems are now more or less standard and integrated into the work routines of CH/SSH researchers, they did not fundamentally change

Caroline Sporleder
Computational Linguistics/MMCI - Saarland University, PO Box 15 11 50, 66041 Saarbrücken, Germany e-mail: csporled@coli.uni-saarland.de

Antal van den Bosch
Tilburg center for Cognition and Communication, School of Humanities, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands e-mail: Antal.vdnBosch@uvt.nl

Kalliopi Zervanou
Tilburg center for Cognition and Communication, School of Humanities, Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands e-mail: K.Zervanou@uvt.nl

these routines—in the Foreword to this book, Prof. Willard McCarty offers background to the deeper scientific-philosophical movements behind this development. The past decade, however, set a profound change in motion: beyond object metadata, the primary object data themselves started to become available in digital form, either due to large-scale digitisation efforts (for instance, in the case of historical manuscripts, through optical character recognition, OCR), or because the data were born digital. The availability of primary data in digital form raises the possibility of much more sophisticated data access. For example, data sources can be retrieved based not only on keyword search, but also, after linguistic enrichment and disambiguation, via semantic search, which enables searching for meaningful patterns and relations between facets of objects.

Moreover, there is a growing awareness that digitisation provides opportunities which go far beyond sophisticated data access. Computer systems may support data analysis, by providing data visualisation in a suitable form, or even by automatically analysing data and discovering new knowledge in the form of interesting trends or higher-level patterns of interdependency, to present to researchers for verification, exploration, and serendipitous discoveries.

Yet, in order to make the most of the opportunities offered by digitisation, it is necessary to develop robust computational methods to clean, enrich, search, and mine digitised data. Since much of the primary and most of the secondary data in SSH/CH are textual, language technology (LT) has an important role to play in this endeavour.

2 Mutual Benefits

The combination of cultural heritage, social sciences and the humanities on the one hand, and information technology and language technology on the other, stands to benefit all disciplines.

For the curators of cultural heritage institutions, information technology can lead to significant time savings and efficiency in the curation work. Computers can provide support in ensuring consistency, (largest possible) completeness, and reliability of the metadata. This can be achieved by the implementation of controlled vocabularies, automatic consistency checks, and (semi-)automatic data completion and error detection methods. Since curation is often undertaken by researchers employed by museums and archives, the excess time and effort originally required for the curation process can be invested in performing original research. Moreover, research itself benefits from digitisation and information technology, both in terms of improved data access, as well as in terms of novel computational research tools and methodologies.

As an example, imagine a social historian working on the public sector strikes in the UK during the winter of 1978–79 (known as the Winter of Discontent) and on the role of the then Chancellor of the Exchequer, Denis Healey, in the lead-up to this event. In the pre-digital age the historian would have had to read

the transcripts of all UK parliamentary debates held between 1975 and 1979 to find the relevant passages. Nowadays this information can be easily retrieved by searching for the keywords "Denis Healey" or "Chancellor of the Exchequer" in the digitised transcripts for the relevant time period. If the document metadata are (automatically) enriched by means of various types of linguistic analysis the search results would be even better. For instance, co-reference resolution could be used to disambiguate the deictic expression "Chancellor of the Exchequer" and link it to the correct person depending on the time period to which it refers. Topic tracking could automatically detect all passages that relate to a given event. Speaker attribution could link direct or indirect expressions of opinions to their originator [4]. Word sense disambiguation could help to distinguish between the 'work stoppage' sense of "strike" and the 'attack' sense.

Moreover, if our imaginary historian is also interested in how the *Winter of Discontent* was perceived in the British or international media at that time, digitisation together with automatic data linking allows for all these data sources to be browsed and visualised in an intuitive way, thus enabling scientists to explore unprecedented amounts of information from their own workspace. The fact that digitised data are available non-locally and instantly is clearly another key benefit in terms of data access, saving researchers a lot of time originally required for travelling to archives around the world, or waiting for the arrival of documents ordered on long-distance loan. Additionally, digitised data may be accessed concurrently, thus allowing more researchers from different locations to work simultaneously on the same data set.

Information technology may influence research in CH/SSH domains in ways which go beyond data access, though. It could have an impact on the respective research practices and methodologies. Computer programmes can be used to automatically visualise the data e.g., by highlighting which amount of press coverage the *Winter of Discontent* received in different countries, or to detect trends and interdependencies, e.g. between increased press coverage in a given country and similar strike events. Going one step further, IT researchers may develop computer tools which not only support the visualisation of interdependencies, but also but also detect and infer new knowledge, such as for example, similar periods of social unrest which had a different outcome [6], or the role of other non-intuitive factors, such as public opinion of less prominent figures than the Chancellor of the Exchequer in the respective or other similar events.

Digitisation does not only offer advantages to curators and researchers, it also provides benefits to cultural heritage institutions. Digital data can be published on websites, thus allowing museums, archives and libraries to reach out to new user groups who might eventually visit in person. Moreover, objects which are not currently on public display, either due to their fragility or—more likely—due to limited exhibition space, can be easily exhibited in a virtual museum, thus providing a solution to a common problem for the vast majority of objects in many collections and allowing for these, virtual exhibitions, to be more complete.

Such virtual exhibitions can also be personalised, based on the interests and preferences of a particular user or any special needs of a group of users. For example, a user might be predominantly interested in paintings from a given period

or in a given style; object descriptions for laypeople need to be considerably less technical than those for experts; and visually impaired users will benefit from a more vivid and detailed description of what the object looks like. Interests, preferences, and needs of different users or user groups can be pre-specified or learnt automatically, e.g., based on the user's browsing history or by analogy with other user groups. Natural language descriptions of artefacts can then be created on the fly, which not only allows tailor-made presentations for different users but also makes it possible to take into account the viewing history; thus similarities and differences between an object and a previously seen piece can be pointed out explicitly, bringing the virtual exhibition tour much closer to a personalised tour by a museum guide.

Personalisation need not be restricted to virtual exhibitions, though; it can also be applied when a user visits a museum in person. For instance, audio guides can automatically generate object descriptions, instead of using pre-recorded texts. Additionally, object description generation may be combined with eyetracking to detect where a visitor is looking, so as to highlight interesting features in that area.

Finally, digitisation also enables user participation, e.g., in the form of user-generated content. For instance, museums may ask their (real or virtual) visitors to contribute photos and descriptions of objects of historical significance which they own,[1] to describe how they or their parents and grandparents were affected by a particular historical event, or to post questions to the museum's curators via a website.

While the benefits of computerisation for CH and SSH are clear, information technology, and language technology in particular, also stand to benefit from working on CH/SSH data. Language technology has been predominantly focused on relatively small and well-curated data sets from a handful of domains, such as newswire and biomedicine. CH/SSH data are typically much more challenging. For example, the digitisation process can introduce errors, metadata often come in note-form and not in carefully written complete sentences, and the language in old manuscripts is archaic and non-standard. Furthermore, the provenance of the data, that gives researchers vital information on its origins in place and time, its author, and its purpose, is often a research topic by itself. The analysis of such data requires robust natural language processing tools. To this end, LT research is impelled towards more sophisticated fallback and adaptation strategies and towards intelligently combining all available resources.

The particularities of the CH/SSH domains also entail that the adaptation of existing tools and resources is far from trivial. Standard techniques, such as supervised machine learning using annotated training data, make too strong assumptions about these data, thus forcing researchers to develop hybrid or even completely new techniques to meet the challenges of the CH/SSH domains. The proposed research solutions, once available, are bound to have a beneficial impact on the natural language processing field and its applications in general, because techniques dealing successfully with such challenging domains are likely to be robust enough to also work for numerous others.

[1] For example, the British Museum recently encouraged this form of user participation in the context of a BBC radio programme. See http://www.bbc.co.uk/ahistoryoftheworld/.

3 Challenges

While the development of information technology for the CH/SSH domains brings about great opportunities and benefits, it is not without difficulties. The challenges associated with research on IT applications for these domains are not only technical; they are also related to the communication and understanding between researchers coming from different disciplines and established research traditions.

Some of the technical difficulties of developing language technology for CH/SSH data were already mentioned above. Furthermore, the digitisation itself is not a trivial process. Optical character recognition of hand-written manuscripts, for example, is still very much an unsolved problem which is further exacerbated by old manuscripts containing stains, faded letters, non-standard orthography, and generally by the use of old variants of current languages. In terms of metadata, standardisation presents another challenge: in order to ensure interoperability and access to linked data from different sources, metadata descriptions should be standard. However, numerous metadata schemes currently exist, both general purpose and domain specific. The IT department of a museum may currently choose, for example, between the Dublin Core Metadata standard, the MIDAS Heritage standard [3], or the CIDOC-CRM [2] among others. Moreover, many CH institutes may decide that none of the existing standards really meets their needs and develop their own, in-house standard. Unfortunately, mapping between different metadata standards automatically is far from straightforward and constitutes in itself the topic of a considerable amount of ongoing research.

Another major technical challenge concerns the preservation of digitised data. While non-digital data can have surprising longevity, despite being stored on fragile material such as paper, digital data typically have an extremely short life-span of only a few years if they are not carefully managed. Digital data are endangered by various factors. Storage media such as CD-ROMs can become unreadable due to material aging and decaying; outdated media cannot be read by modern hardware (e.g., floppy disks), and outdated formats may become inaccessible (e.g., old word processing formats no longer supported). Digital data management and preservation is extremely challenging and also rather expensive. Guaranteed reliable long-term digital data management, i.e., spanning several decades, is currently out of reach for most CH institutes. The list of spectacular failures of data preservation is long, including the BBC's Domesday Project,[2] a multimedia collection from the mid-1980s which had become virtually unreadable by 2000 due to an outdated storage format, and NASA's 1976 Viking Mars mission, whose data were stored on magnetic tape that was later found to have become partly unreadable due to material aging [1]. Digital durability, data management and preservation is the focus of much ongoing research.

Problems of an entirely different nature arise from the fact that developing information technology for CH/SSH data is a highly interdisciplinary endeavor. IT and CH/SSH researchers do not only work in very different areas, they also have

[2] http://www.domesday.org.uk/

very different styles of doing research. Humanities researchers tend to work on their own and pursue long-term research and publication goals. IT research tends to be more dominated by collaborative work, relatively fast publication cycles, and strong requirements for measurable, reproducible results. For these reasons, and for the common reason that the vocabularies of the different areas need considerable translation work in order to be mutually intelligible, communication across such different disciplines can often be challenging, but the result can be enlightening. For instance, to understand that the concepts of 'completeness' and 'correctness' (e.g., of data in historical research) are analogues of the technical concepts of 'recall' and 'precision' can lead to a more fundamental understanding of the relation between human and automated information retrieval.

Intense communication remains crucial, because IT researchers are typically not aware of the needs of CH/SSH researchers, while the latter are not aware of the full extent of the opportunities offered by technology. In this respect, the IT problems which a CH/SSH researcher thinks unsolvable (and thus possibly does not even mention to the IT specialist) may be trivial to solve, while other problems, seemingly trivial to a CH/SSH expert, might pose great problems for the IT researcher (e.g., data being available, but not in digital form).

Furthermore, many CH/SSH researchers do not really trust automatic methods as much as they trust their own research methodology, partly due to a lack of understanding of the way a specific technology works. In order to establish trust, computer tools are required to make their analysis more transparent. For instance, a program which automatically detects errors in CH databases could provide information about why a given piece of information is considered erroneous [5]. In other cases, the lack of trust relates to the accuracy of automatic results: since the integrity of data and metadata is of crucial importance in CH/SSH domains, IT researchers are required to provide means of process documentation and data provenance, so that manual annotations are distinguished from automatic ones, and the 'original' data can always be recovered.

The lack of understanding between CH/SSH and IT researchers is particularly problematic since making the most of digitisation does not simply involve the adaptation of existing technology to new domains, but rather the development of new methodologies, new research questions, and new applications. This requires close collaboration across disciplines, over a long period of time. Joint research projects are ideal to foster such collaboration. Fortunately there are now a number of initiatives that encourage cooperation, such as CLARIN[3] and DARIAH,[4] and as an example of a national programme, CATCH[5] in the Netherlands.

A related problem is that the development of technology to enrich, access, and mine CH/SSH data involves a number of different communities even within the disciplines (natural language processing, artificial intelligence, semantic web research, museum informatics, archival science, digital humanities etc.). These

[3] http://www.clarin.eu/

[4] http://www.dariah.eu

[5] http://www.nwo.nl/catch

communities tend to be more or less disjoint, and attend different workshops and conferences. This means that there is often little opportunity to meet and exchange ideas. Joint events could help to overcome this problem. Even though some joint events start to be regularly organised, they are just gathering momentum to get accepted as a communal outlet of research for different communities.

In addition to collaborative research projects and workshops, it would also be beneficial to increase cross-disciplinary awareness already in the educational system. On the one hand, this can be done by broadening the curricula of different subjects, e.g., offering more IT courses to humanities students and more exposure to CH/SSH data to IT students. On the other hand, recent years have seen the creation of new interdisciplinary study areas. For example, several universities offer degrees or at least specialisations in "Digital Humanities".

Language technology researchers may have a special role to play in bridging the gap between science and humanities. Natural language processing and computational linguistics are traditionally highly interdisciplinary research areas that have found homes both in computer science and linguistics departments. Language technology researchers are familiar with integrating people from different backgrounds (linguistics, logic, computer science, cognitive science, etc.) and have a foot in both worlds: the sciences and the humanities. Arguably, they are therefore uniquely placed to drive forward the IT revolution in cultural heritage, the humanities and social sciences.

4 This Volume

There is a growing interest in the language processing community to meet the challenges posed by the CH/SSH domains. To provide an outlet for various types of language technology research carried out in these areas, we initiated the "Language Technology for Cultural Heritage, Social Sciences, and Humanities" (LaTeCH) workshop series in 2007. In the past four years, LaTeCH has been held in conjunction with various language technology and artificial intelligence conferences (ACL-07, LREC-08, EACL-09, ECAI-10), and has attracted high-quality papers on a wide variety of topics from all over the natural language processing community. The current book provides an overview of some of the highlights of the past four editions of LaTeCH. The book covers a large spectrum of applications and illustrates how language technology can be employed in key task areas in the CH/SSH domain, ranging from preprocessing, over tool adaptation, to personalisation, automatic data analysis, and data management and retrieval. The papers also relate to different application domains, some being more concerned with museums and other cultural heritage institutes, while others relate more to work in the humanities and social sciences. This volume consists of six parts:

Preprocessing

The first part is concerned with the digitisation process itself. It contains two chapters highlighting two important aspects of the digitisation process. Digitisation of written data is typically done by optical character recognition. This is, however, an error-prone process, especially for old manuscripts which can contain old-fashioned fonts, stained or faded text passages, and archaic orthography. *Volk, Furrer and Sennrich* are concerned with the digitisation of a large multilingual corpus of Alpine texts, dating back to the 19th century. To reduce the number of OCR errors, they propose a technique which merges the output of two OCR systems and exploits lexical resources for further correction. *Toselli, Romero and Vidal* also deal with digitisation, but their paper is concerned with hand-written texts. Hand-written historical documents are typically difficult to digitise by optical character recognition and are consequently often transcribed manually. Usually, the transcripts are aligned with digital images of the original documents on a page-level. For researchers, however, it is often desirable to have a more fine-grained alignment, e.g., on the word level. Toselli et al. propose a technique based on Hidden Markov Models to automatically align original texts and their transcripts on the word-level.

Adapting NLP Tools to Older Language Varieties

Dealing with older language varieties is one of the main challenges of the field. While tools for standard linguistic tasks such as part-of-speech tagging, syntactic parsing, or word sense disambiguation exist for many modern day language varieties and a small set of domains, the performance of these tools tends to drop significantly when they are applied to older language varieties or new domains. The two papers in the second part of this volume both deal with the problem of how existing tools can be adapted for older language varieties, but they offer two different solutions. *Borin and Forsberg* discuss a rule-based approach for adapting a morphological component for Present-Day Swedish to Late Modern Swedish and Old Swedish. The work was carried out in the context of creating a diachronic lexical resource that enables semantic search in historical texts by using Present-Day Swedish as a pivot to which the lexical entries of older language varieties are semi-automatically linked. *Rögnvaldsson and Helgadóttir* are also concerned with morphosyntactic analysis, but they use a machine learning approach. They first train a statistical tagger on a Modern Icelandic corpus and apply it to Old Icelandic texts. A sample of the tagged Old Icelandic corpus is then manually corrected and the tagger is retrained on mixture of Old Icelandic and Modern Icelandic data.

Linguistic Resources for CH/SSH

The availability of linguistic resources for CH/SSH is a recurrent issue for the application of language technologies in those domains, as linguistic resources are

essential both for the development and the adaptation of language technology methods in such domains and language varieties. In this part, two approaches to the development of such resources are discussed. *Bamman and Crane* start with a presentation of the Ancient Greek and Latin Dependency Treebanks, novel historical corpora resources comprising of works of classic Greek and Roman authors, and discuss issues related to the development and applications of such resources both in linguistics, as well as in classical philology research. Subsequently, *Giouli, Simov and Osenova* are concerned with the development of a bilingual Greek/Bulgarian corpus comprising of literary works, folktales and legends, as well as texts presenting the customs, rituals, everyday-life habits of the people living in the cross-border area of Thrace. The authors discuss issues related to tool adaptation for corpus linguistic annotation and those related to metadata creation intended to facilitate comparative cultural, linguistic and literary studies, for the neighbouring areas of Greece and Bulgaria.

Personalisation

One of the advantages in using language technologies for the CH/SSH domains lies in providing support for personalised access to CH/SSH information. *Konstantopoulos, Karkaletsis, Vogiatzis and Bilidas* present the ELEON/NATURALOWL system which exploits language technologies and linguistic adaptation resources for providing multi-lingual and personalised conceptual representations of cultural heritage objects. In the ELEON/NATURALOWL approach, personalised profiles allow the specification of, for example, whether technical vocabulary should be used for expert audience, or whether shorter and simpler sentences should be generated for children and gear the system towards achieving different interaction goals, such as targeting specific objects and facts.

Structural and Narrative Analysis

Part five of this volume includes three papers that are concerned with the automatic analysis of texts in terms of structure and narrative content. The developed tools support linguistic, literary, historical, sociological and ethnological research by segmenting texts in meaningful ways, detecting topics or narrative schemas, and finding common content elements in different texts. *Hendrickx, Généreux and Marquilhas* are concerned with detecting boundaries between discourse segments in Portuguese letters dating from the 16th to 19th century. The segments are subsequently also automatically labelled according to their discourse function, e.g., opening, introduction, main part, and closing. The authors model the task as a supervised machine learning task, using additional resources to overcome data sparseness and problems relating to inconsistent spelling. *Declerck, Scheidel and Lendvai* describe a markup language schema for annotating fairy tales with a semantic, narrative analysis of motifs according to the theory of Vladimir Propp. They also investigate in how far such an analysis can be integrated with other

semantic annotation schemes, e.g., frame semantics. *Reiter, Hellwig, Frank, Gossmann, Larios, Rodrigues and Zeller* are also concerned with the semantic analysis of texts. They work on descriptions of Indian rituals, and aim to automatically analyse such texts and discover common elements that could provide useful starting points for ritual researchers. In order to detect regularities and differences in rituals, the authors perform a frame semantic analysis of the texts, adapting various NLP tools to this domain.

Data Management, Visualisation and Retrieval

The final part of the present volume deals broadly with the management of digitised data and with strategies for retrieval and visualisation of information. The paper by *Berzak, Richter, Ehrler and Shore* describes a system for searching, browsing, and visualising a large collection of speeches by Fidel Castro. At the core of their system lies a method for computing the semantic relatedness between different documents. The collection can then be displayed as a graph structure, highlighting similarities between documents as well as their relevance for a given user query.

The work discussed by *Witte, Kappler, Krestel and Lockemann* attempts to make heritage documents more flexibly accessible by transforming them into a semantic knowledge base. They apply semantic analysis technologies to the historic Encyclopedia of Architecture, so as to automatically populate an ontology which allows building historians to navigate and query the encyclopedia, while architects can directly integrate it into contemporary construction tools. The content is also made accessible in a user-friendly Wiki interface, combining original text with NLP-derived metadata and annotation capabilities for collaborative use.

Acknowledgements The LaTeCH workshops—and by extension this book—would not have been possible without the dedication and hard work of a large number of people, including:

- past co-organisers of the workshop series, particularly: Lars Borin, Claire Grover, Piroska Lendvai, Martin Reynaert, and Kiril Ribarov;
- the authors who submitted papers to the workshop;
- the reviewers;
- the invited speakers: Martin Doerr, Douglas Oard, Martin Reynaert, and Tamás Váradi;
- and the organisers of the conferences which hosted LaTeCH.

We would like to take this opportunity to thank everybody for their enthusiasm and commitment over the past four years, and look forward to the continuation of the LaTeCH workshop series.

We would like to express our gratitude for the financial support the LaTeCH workshop series has received. The European Union's MultiMatch project partly sponsored the first edition of LaTeCH, while our work on the present book was supported by the CATCH programme of the Netherlands Organisation of Scientific Research (NWO), and the Cluster of Excellence "Multimodal Computing and Interaction" within the Excellence Initiative of the German Federal Government.

Last but not least, we would like to thank Willard McCarty for his support and for sharing his thoughts in the foreword of this book.

References

1. Besser, H.: Digital longevity. In: M. Sitts (ed.) Handbook for Digital Projects: A Management Tool for Preservation and Access, pp. 155–166. Northeast Document Conservation Center (2000)
2. Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M. (eds.): Definition of the CIDOC Conceptual Reference Model. ICOM/CIDOC CRM Special Interest Group (2009)
3. Lee, E. (ed.): MIDAS: A Manual and Data Standard for Monument Inventories. English Heritage, Swindon (1998)
4. Ruppenhofer, J., Sporleder, C., Shirokov, F.: Speaker attribution in cabinet protocols. In: The seventh international conference on Language Resources and Evaluation (LREC), pp. 2510–2515 (2010)
5. Van den Bosch, A., Van Erp, M., Sporleder, C.: Making a clean sweep of cultural heritage. IEEE Intelligent Systems **34**(2), 54–63 (2009)
6. Van den Hoven, M., Van den Bosch, A., Zervanou, K.: Beyond reported history: Strikes that never happened. In: S. Darányi, P. Lendvai (eds.) Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts, pp. 20–28. Vienna, Austria (2010)