Theory and Applications of Natural Language Processing
Edited volumes

Caroline Sporleder
Antal van den Bosch
Kalliopi Zervanou *Editors*

# Language Technology for Cultural Heritage

## Selected Papers from the LaTeCH Workshop Series

Springer

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)

# Part I
# Pre-Processing

# Strategies for Reducing and Correcting OCR Errors

Martin Volk, Lenz Furrer and Rico Sennrich

**Abstract** In this paper we describe our efforts in reducing and correcting OCR errors in the context of building a large multilingual heritage corpus of Alpine texts which is based on digitizing the publications of various Alpine clubs. We have already digitized the yearbooks of the Swiss Alpine Club from its start in 1864 until 1995 with more than 75,000 pages resulting in 29 million running words. Since these books have come out continuously, they represent a unique basis for historical, cultural and linguistic research. We used commercial OCR systems for the conversion from the scanned images to searchable text. This poses several challenges. For example, the built-in lexicons of the OCR systems do not cover the 19th century German spelling, the Swiss German spelling variants and the plethora of toponyms that are characteristic of our text genre. We also realized that different OCR systems make different recognition errors. We therefore run two OCR systems over all our scanned pages and merge the output. Merging is especially tricky at spots where both systems result in partially correct word groups. We describe our strategies for reducing OCR errors by enlarging the systems' lexicons and by two post-correction methods namely, merging the output of two OCR systems and auto-correction based on additional lexical resources.

**Key words:** OCR error correction, OCR merging, historical corpus, multilingual corpus, Alpine texts

Martin Volk
Institute of Computational Linguistics, University of Zurich, e-mail: volk@cl.uzh.ch

Lenz Furrer
Institute of Computational Linguistics, University of Zurich, e-mail: lenz.furrer@access.uzh.ch

Rico Sennrich
Institute of Computational Linguistics, University of Zurich, e-mail: sennrich@cl.uzh.ch

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_1, © Springer-Verlag Berlin Heidelberg 2011

## 1 Introduction

In the project Text+Berg[1] we digitize the heritage of Alpine literature from various European countries. Currently our group digitizes all yearbooks of the Swiss Alpine Club (SAC) from 1864 until today. Each yearbook consists of 300 to 600 pages and contains reports on mountain expeditions, culture of mountain peoples, as well as the flora, fauna and geology of the mountains.

Digitization of this corpus requires a large-scale scanning effort followed by converting the images to text, a procedure known as optical character recognition (OCR). There are a few commercial OCR products (Abbyy FineReader, Nuance OmniPage) and one open-source product (previously named Tesseract, now called OCRopus[2]). Initial experiments indicated that the open-source tool's recognition quality is far lower than the commercial products, therefore we decided early-on to focus on Abbyy FineReader, the alleged market leader. Of course, the commercial OCR systems only deliver error-free text under ideal conditions like modern font, evenly printed on spotless white paper. For most of our input texts, these conditions are not given.

Some of the books of our corpus have yellowed pages, sometimes even grey spots from the paper, the printing or the handling. The books were typeset in Antiqua from the start but the letters are not always evenly printed. In addition the books contain special symbols (e.g. the clock time 15:45h is often written as 15 $^{3}/_{4}$ where the fraction is one symbol). Moreover, they comprise 19th century spelling variants, a complex layout with a mix of images and text, and text in multiple languages. Challenges for the OCR systems abound. We have noticed that this results in a recognition accuracy of several errors per page. Our aim was to reduce the error rate.

In principle there are three ways to improve the text accuracy resulting from OCR. We can improve the input to the OCR system, the OCR system itself, or the output of the OCR system. On the input side we can try to improve the image so that the contrasts are ideal and the image is clean. We have experimented with greyscale vs. black-and-white scanning and tried various contrast settings. The OCR results did not differ much, so we decided in favor of 300 dpi greyscale images, not least because Abbyy claims that their OCR system is optimized for these. We have not tried cleaning the images with despeckle programs. Experiments reported in [5] indicate that this does not improve the accuracy noticeably.

As a second option, we can try to improve the OCR system. Abbyy FineReader allows three ways of tuning the system. The user can train certain characters (or variants thereof) so that they are added to the recognition alphabet. This is cumbersome and time-consuming. The user can select additional characters from a list so that they are added to the recognition alphabet of the selected language. For example, if the German alphabet is chosen, the user may add Icelandic diacritic characters when processing a text about an expedition to Iceland that

[1] See http://www.textberg.ch.

[2] See http://code.google.com/p/ocropus/

contains geographical names with these special characters. Finally, the user can enlarge the systems' built-in lexicon to add domain-specific vocabulary. This can be expected to increase the recognition accuracy since the OCR system decides on word hypotheses based on language-specific word lists. We have experimented with lexicon enlargement and report on our results in Sect. 4.2.

Thirdly, we can improve the text accuracy by repairing the OCR system output. This amounts to automatic spelling and grammar correction and thus can be tackled in a large number of different ways. We present two methods in this paper. One method is based on merging the output of two different OCR systems, the other is based on producing spelling variants for "unknown" words and predicting which variant is most likely correct.

This paper first describes the Text+Berg project with its multitude of challenges for OCR. We then describe our experiments with a number of strategies for reducing and correcting OCR errors.

## 2 The Text+Berg Project

The Text+Berg project at the University of Zurich aims at building a large corpus of Alpine texts. As a first step we digitize the yearbooks of the Swiss Alpine Club (SAC). The SAC was founded in 1863 as a reaction to the founding of the British Alpine Club a year before. From the very start it produced a sizable yearbook documenting its mountaineering activities. Thus our corpus has a clear topical focus: conquering and understanding the mountains. The articles focus mostly on the Alps, but over the 145 years the books have probably covered every mountain region on the globe.

Some examples from the 1911 yearbook may illustrate the diversity. It has the typical reports on mountain expeditions: "*Klettereien in der Gruppe der Engelhörner*" (English: *Climbing in the Engelhörner group*) or "*Aus den Hochregionen des Kaukasus*" (English: *From the high regions of the Caucasus*). But the 1911 book also contains scientific articles on the development of caves ("*Über die Entstehung der Beaten- und Balmfluhhöhlen*") and on the periodic variations of the Swiss glaciers ("*Les variations périodiques des glaciers des Alpes suisses*").

The corpus is thus a valuable knowledge base to study the changes in all these areas. But the corpus is also a resource to catch the spirit of Switzerland in cultural terms: What does language use in Alpine texts show about the cultural identity of the country and its change over time? See [3] for our research in this area.

Let us briefly describe how we processed the books. Initially we have collected all books in two copies (as a result of a call for book donations by the Swiss Alpine Club). One copy was cut open so that the book can be scanned with automatic paper feed. The other copy remains as reference book.

Then all books were scanned and processed by the OCR systems. The main challenges for OCR which we encountered were the multilingual nature of the text,


# Wanderungen im nordamerikanischen Felsengebirge.

Von

Dr. *A. Emch* (Sektion Weißenstein).

Mit Ausnahme des kanadischen Teiles ist die Alpinistik in den Rocky Mountains noch sehr wenig entwickelt. In Kanada wurde kürzlich ein „Alpine Club“ gegründet, der sich, durch schweizerische und englische Verhältnisse angeregt, die alpine Erforschung der nordamerikanischen, speziell der kanadischen Gebirgsregionen, sowohl in touristischer als in wissenschaftlicher Beziehung zum Ziele gesetzt hat [1]). In den Vereinigten Staaten besteht keine solche, das ganze Land umfassende Vereinigung, und von einem bergsportlichen Interesse, einem Hang nach den seelischen Genüssen einer großartigen Gebirgswelt ist im allgemeinen nicht die Rede [2]). Dem Amerikaner genügt es meistens, wenn er das Bewußtsein hat, daß seine Berge durchschnittlich höher sind, als diejenigen der Alpen. Das nordamerikanische Felsengebirge ist also in touristischer und alpinistischer Beziehung ein noch durchaus jungfräuliches Gebiet. Es gibt dort noch keine speziell für den Bergsteiger gebaute Schutzhütten und an Touristenorten stationierte Führer und Träger. Aus diesem Grunde hat es für jeden eine unverfälschte Natur bewundernden Hochtouristen einen besondern Reiz, führerlos in die riesig

[1]) Jahrbuch S. A. C. 1908, pag. 417.

[2]) *Anm. der Redaktion.* Immerhin besteht seit 1876 ein „Appalachian Mountain Club“ und seit 1903 ein „American Alpine Club“. Ein 1876 gegründeter „Rocky Mountain Club“ scheint schon lange eingegangen zu sein.

**Fig. 1** Example page from the SAC yearbook 1909-10, with decorated initial letter and footnotes

diachronic changes in spelling and typesetting, and the wide range of proper nouns. In Sect. 3, we will give a detailed account of our efforts to improve OCR quality.

After text recognition we added a mark-up of the text structure. Specially developed programs annotated the text with XML tags for the beginning and end of each article, its title and author, subheaders and paragraphs, page breaks, footnotes and caption texts. For example, footnotes are recognized by their bottom position on the page, their smaller font size and their starting with any character followed by a closing parenthesis. Figure 1 shows the start page of an article from the 1910 yearbook with title, author and two footnotes.

Some of the text structure information can be checked against the table of contents and table of figures in the front matter of the yearbooks. We manually corrected these tables as the basis for a clean database of all articles in the corpus. Matching entries from the table of contents to the articles in the books is still not trivial. It requires that the article title, the author name(s) and the page number in the book are correctly recognized. Therefore, we use fuzzy matching to allow for OCR errors and small variations between table of content entries and the actual article header in the book.

## *2.1 Language Identification*

Proper language identification is important for most steps of automatic text analysis, e.g. part-of-speech tagging, lemmatization and named entity classification. The SAC yearbooks are multilingual, with most articles written in German and French, but also some in Italian, Romansch and Swiss German[3]. We use a character-n-gram-based language identification program[4] to determine the language for each sentence.

While language identification may help improve automatic text analysis, the dependency is circular. OCR, tokenization and sentence boundary recognition need to precede language identification so that we are able to feed individual sentences to the language identifier. But high quality tokenization relies heavily on language-specific abbreviation lists and conventions. We therefore perform an initial tokenization and sentence boundary recognition before language identification. Afterwards, we retokenize the text in order to correct possible tokenization errors.

OCR is performed without prior language identification. We configured the OCR systems to use the dictionaries for the following four languages: German, French, Italian and English.

[3] See Sect. 2.3 for information on the amount of text in each language.

[4] We use Michael Piotrowski's language identifier Lingua-Ident from http://search.cpan.org/dist/Lingua-Ident/ .

## 2.2 Further Annotation

Apart from structural mark-up and language identification, the corpus is automatically tagged with Part-of-Speech information. We also aim to provide a fine-grained annotation of named entities.

Named entity recognition is an important aspect of information extraction. But it has also been recognized as important for the access to heritage data. For example, Borin et al. [2] argue for named entity recognition in 19th century Swedish literature, distinguishing between 8 name types and 57 subtypes.

In the latest release of our corpus (all yearbooks from 1864 to 1995), we have annotated all mountain names that we identified unambiguously through exact matching. We have obtained a large gazetteer with 156,000 toponyms from the Swiss Federal Office of Topography. It contains geographical names in 61 categories. We have extracted the SwissTopo mountain names from the 4 highest mountain classes plus the names classified as ridges (*Grat*). This resulted in 6227 names from which we have manually excluded 50 noun homographs. For example *Ofen* (English: *oven*) is a Swiss mountain name, but in order to avoid false hits we eliminated it from the list. The resulting gazetteer triggered the identification of 95,400 mountain names in our corpus.

## 2.3 Aims and Current Status

In the final processing phase, the corpus will be stored in a database which can be searched via the internet. Because of our detailed annotations, the search options will be more powerful and lead to more precise search results than via the usual search engines. For example, it will be possible to find the answer to the query "List the names of all glaciers in Austria that were mentioned before 1900." We also annotate the captions of all photos and images so that they can be included in the search indexes.

[15] emphasize that advanced access methods are crucial for Cultural Heritage Data. They distinguish different user groups having different requirements (Historians, Practitioners, Laypersons, Computational Linguists). We will provide easy access to the texts and images through a variety of intuitive and appealing graphical user interfaces. We plan to have clickable geographic maps that lead to articles dealing with certain regions or places.

As of December 2010, we have scanned, OCR-converted and annotated 168 books from 1864 to 1995 (cf. [4]).

We have 90 books from 1864 to 1956. (In 1870, 1914 and 1924 no yearbooks were published.) From 1957 to 1995 we have parallel French and German versions of the yearbooks. Overall we have scanned around 75,000 pages. The corpus is made up of around 6000 articles in German, 2700 in French, 140 in Italian, 11 in Romansch, and 3 in Swiss-German. Our parallel corpus currently contains 950 articles amounting to 3.3 million tokens in French and 2.9 million tokens in German.

**Table 1** Token counts (rounded) in the Text+Berg corpus

| | German | French | Italian | Other | **Total** |
|---|---|---|---|---|---|
| tokens in entire corpus | 18,700,000 | 9,770,000 | 320,000 | 100,000 | 28,890,000 |
| tokens in parallel subcorpus | 2,910,000 | 3,300,000 | | | |

Table 1 gives an overview of the token frequencies per language. Work on scanning and converting the yearbooks from 1996 is ongoing and will be finished soon. More details on the project phases can be found in [14].

## 3 Scanning and OCR

Let us return to the OCR step. After scanning the pages in greyscale with 300 dpi we embarked on converting all pages to text. We started by using Abbyy-FineReader 7. We have initially evaluated Abbyy-FineReader version 7 to version 9, but found the older version more stable and of equal OCR quality.

Even though the performance of OCR applications is satisfactory for most purposes, we are faced with thousands of OCR errors in large text collections. Since we aim to digitize the data as cleanly as possible, we wish to minimize the number of errors. Additionally, OCR errors can be especially damaging for some applications. The numerous named entities, i.e. names of mountains, streams and Alpine cabins are especially prone to OCR errors, especially because many of them do not occur in the dictionaries used by OCR tools. At the same time, these named entities are highly relevant for our goal of building a searchable database. In our project, OCR is complicated by the fact that we are digitizing a multilingual and diachronic corpus with texts spanning from 1864–1995.

### *3.1 Enlarging the OCR Lexicon*

The OCR software comes with two lexicons for German, one for the spelling after 1901 and one for the new orthography following the spelling reform of the late 1990s. The system does not have a lexicon for the German spelling of the 19th century (e.g. old *Nachtheil*, *passiren* and *successive* instead of modern *Nachteil*, *passieren* and *sukzessive*). We have therefore added 19th century word lists to the system. We have manually corrected one book from 1890, and subsequently extracted all words from that book that displayed old-style character sequences (such as 'th', 'iren', and 'cc'). In this way we added 1500 word forms to the OCR lexicon.

The 20th century books follow the Swiss variant of German spelling. In particular, the Swiss spelling has abandoned the special character 'ß' in favor of 'ss'. For example, the word *ließ* (English: *let*) is spelled *liess* in Switzerland. The OCR

lexicons list only the spelling from Germany. We have therefore compiled special word lists with Swiss spelling variants taken from the GNU Aspell program and added around 5000 entries to the OCR lexicon.

Names that are not in the system's lexicon pose another problem to character recognition. Our books contain a multitude of geographical names many of which are unknown to the OCR system. We have therefore purchased a large list of geographical names from the Swiss Federal Office of Topography[5] and extracted the names of the major Swiss cities, mountains, valleys, rivers, lakes, hiking passes and mountain cabins. In total we added 14,800 toponyms to the OCR system. In Sect. 4.2 we present the results of this lexicon enlargement.

## *3.2 Post-correcting OCR Errors*

Following the OCR we have experimented with three post-correction methods. Here we introduce the methods, while Sect. 4.2 has the comparative evaluation results.

### 3.2.1 Pattern-based Corrections

When the annotation process has completed tokenisation, a polisher module is invoked. Here, some heuristics are applied to catch and correct common OCR errors, such as misrecognised *Htttte* for correct *Hütte* (English: *cabin*) or erroneous *Eichtung* for correct *Richtung* (English: *direction*). Using regular expressions, a closed set of substitution patterns is applied to the corpus.

In this way we post-correct errors caused by graphemic similarities which have been missed by the OCR engine. This automatic correction happens after tokenization with heuristics that check each word. For example, a word-initial 'R' is often misinterpreted as 'K', resulting in e.g. *Kedaktion* instead of *Redaktion* (English: *editorial office*). To minimize false positives, our rules fall in three categories: First, strict rule application: The tentative substitute must occur in the corpus and its frequency must be at least 2 times as large as the frequency of the presumably mistyped word. The above *K*→*R* example falls in this category. Second, normal rule application: The tentative substitute must occur in the corpus. Substituting 'ii' by either 'n', 'u', 'ü', 'li' or 'il' (of the five tentative substitutes the word with the highest frequency is selected; *iiberein* → *überein*, English: *in agreement*) falls in the normal category. Third, unconditional substitution. For example, substituting *Thai* with *Thal* (the 19th century spelling of *Tal*, English: *valley*) is an example of the unconditional rule category.

[5] http://www.swisstopo.ch

### 3.2.2 OCR Merging

In an attempt to automatically detect and correct the OCR errors, we exploit the fact that different OCR systems make different errors. Ideally, we can eliminate all OCR errors that are only made by one of two systems. We have created an algorithm that compares the output of two OCR systems (Abbyy FineReader 7 and OmniPage 17) and performs a disambiguation, returning the top-ranking alternative wherever the systems produce different results.

### 3.2.3 The Merging Algorithm

For our task, we can avoid potential complexity problems since we do not have to compute a global alignment between the two OCR systems. Three factors help us keep the search space small: Firstly, we can extract differences page-by-page. Secondly, we ignore any differences that cross paragraph boundaries, defaulting to our primary system FineReader if such a large discrepancy should occur. Thirdly, the output of the two systems is similar enough that differences typically only span one or two words.

For each page, the algorithm traverses the two OCR-generated texts linearly until a difference is encountered. This point is then used as starting point for a longest common subsequence search in a 40-character-window. We extract as difference everything up to the start of the longest subsequence, and continue the algorithm from its end.

For selecting the best alternative, we consider the differences on a word level. If there are several differences within a short distance, all combinations of them are considered possible alternatives. As a consequence, we not only consider the output of FineReader (*Recensione-»,*) and OmniPage (*Rccensionen*), but also the combinations *Rccensione-»,* and *Recensionen*. In this way, the correct word form *Recensionen* can be constructed from two wrong alternatives.

Our decision procedure is based on a unigram language model trained on the latest release of the Text+Berg corpus. The choice to bootstrap the decision procedure with noisy data generated by Abbyy FineReader bears the potential risk of skewing the selection in Abbyy FineReader's favor. However, the language model is large (25.7 mio words), which means that possible misreadings of a word are far outnumbered by the correct reading. For instance, *Bergbauer* (English: *mountain farmer*) is twice misrecognized as *ßergbauer* by Abbyy FineReader. Still, *Bergbauer* is more than 20 times as frequent as *ßergbauer* in the corpus (47 vs. 2 occurrences), which lets the language model make a felicitous judgment.

It is worth noting that OCR merging is performed before language identification, and that we do not use one model per language, but a language model trained on the whole corpus, irrespective of language.

Words containing non-alphabetical characters have been removed from the language model, with the exception of hyphenated words. Punctuation marks and

other special characters are thus penalized in our decision module, which we found to be an improvement.

A language model approach is problematic for cases in which the alternatives are tokenized differently. Generally, alternatives with fewer tokens obtain a higher probability. We try to counter this bias with a second score that prefers alternatives with a high ratio of known words. This means that *in Göschenen* is preferred over *inGöschenen*, even if we assume that both *Göschenen* (the name of a village) and *inGöschenen* are unknown words in our language model[6].

The alternatives are ranked first by the ratio of known words, second by their language model probability. If there are several candidates with identical scores, the alternative produced by Abbyy FineReader is selected.

### 3.2.4 First Evaluation of the OCR Merging

We have performed a manual evaluation of the merged algorithm based on all instances where the merged system produces a different output than Abbyy FineReader. The cases where Abbyy's system wins are not as interesting since we regard them as the baseline result. Out of the 1800 differences identified between the two systems[7] in the 1899 yearbook, the FineReader output is selected in 1350 cases (75%); in 410 (23%), the OmniPage reading is preferred; in 40 (2%), the final output is a combination of both systems. We manually evaluated all instances where the final selection differs from the output of Abbyy FineReader, which is our baseline and the default choice in the merging procedure.

**Table 2** Examples where OmniPage is preferred over FineReader by our merging procedure

| Abbyy FineReader | OmniPage | correct alternative in context | jugdment |
|---|---|---|---|
| Wunseh, | Wunsch, | entstand in unserem Herzen der **Wunsch**, | better |
| East | Rast | durch die **Rast** neu gestärkt | better |
| Übergangspunkt,. das | Übergangspunktr das | ist Hochkrumbach ein äußerst lohnender Übergangspunkt, das | equal |
| großen. Freude | großen, Freude | zu meiner **großen Freude** | equal |
| halten | hatten | Wir **halten** es nicht mehr aus | worse |
| là | la | c'est **là** le rôle principal qu'elle joue | worse |

Table 2 shows some examples and our judgment. We see clear improvements where non-words produced by Abbyy FineReader (e.g. *Wunseh*) are replaced with a known word produced by OmniPage (*Wunsch*, English *wish*). On the other hand, there are cases where a correctly recognized Abbyy word (e.g. *halten*, English: *hold*)

[6] Unknown words are assigned a constant probability $> 0$.

[7] Note that one difference, as defined by our merging algorithm, may span several words. Also, frequent differences that would be resolved in later processing steps (i.e. differences in tokenization or hyphenation) are ignored by the merging algorithm.

is overwritten by the OmniPage candidate (*hatten*, English: *had*) because the latter is more frequent in our corpus. As a third possibility, there are neutral changes where the Abbyy output is as wrong as the OmniPage output, as in the two examples judged as "equal", where the systems suggest different punctuation symbols where none is intended in the text.

In our manual evaluation, we found 277 cases where OCR quality was improved, 82 cases where OCR quality was decreased, and 89 cases where combining two systems neither improved nor hurt OCR quality.

We noticed that performance is worse for non-German text. Most notably, OmniPage tends to misrecognize the accented character *à*, which is common in French, as *A* or *a*, or to delete it. The misrecognition is a problem for words which exist in both variants, especially if the variant without accent is more common. This is the case for the French article *la* (English: *the*) and the adverb *là* (English: *there*), and leads to a miscorrection in the example shown in table 2. We are lucky that in our language model, the French preposition *à* (English: *to*) is slightly more probable than the French verb *a* (English: *has*); otherwise, we would encounter dozens of additional miscorrections.[8] Word deletions are relatively rare in the evaluation set, but pose a yet unsolved problem to our merging algorithm. In 8 cases, *à* is regrettably deleted by OmniPage. These alternatives always obtain a higher probability than the sequences with *à*[9], and are thus selected by our merging procedure, even though the deletion is incorrect in all 8 instances.

Considering that we are working with a strong baseline, we find it encouraging that using the output of OmniPage, which is considerably worse than that of Abbyy FineReader, allows us to further improve OCR performance.

### 3.2.5 Corrections based on External Resources

Our third approach for cleaning up the corpus in post-correction is based on external lexical resources. The aforementioned methods base their decision about "wrong" and "correct" word forms on frequencies of the words in the corpus itself. This entails the risk of categorizing frequent errors as "good" words, which is not an unlikely scenario since OCR systems misrecognize unknown words quite often; e. g. the Alpine toponym *Schneehorn* is rendered thrice as misspelt *Sehneehorn* or its genitive form *Sehneehorns*.

In order to reduce the danger of propagating OCR errors in post-correction, the following approach makes use of external resources for the correctness categorization. It is partly comparable to the pattern-based approach of 3.2.1 as it is all about substituting potentially misspelt words by close orthographical variants which are assumed to be correct. The main differences are in the decision routine, which is not based on corpus frequencies but on lexicon data, and in a search space not restricted to predefined regular expressions.

[8] Of course, one could devise rules to disallow particular corrections.

[9] Since every word has a probability $< 1$, each additional token decreases the total probability of an alternative.

We use two resources as categorizers to divide all word types of our corpus into *known words* and *unknown words*. For the general German vocabulary we use Gertwol, which takes care of the unlimited number of compounds in German. A subcorpus of the SwissTopo list mentioned in 2.2 is then used to recognize toponyms unknown to Gertwol. Subsequently, unknown words are re-classified as known words if they show to be ancient-spelling variants of known words, or if they can be analyzed as compounds with a toponym as the head and any known word as the tail. All words shorter than a predefined length threshold are excluded. The remaining unknown words are now considered potential OCR-errors and they are exposed to a correcting algorithm.

The procedure of searching correction candidates is done in three steps of ascending complexity. In the first and the second step, a small set of character substitutions found in a high number of OCR-errors is used to derive hypothetical spelling variants. For example, it is common to find mistaken *u* for correct *n* in the output of OCR systems, as in recognized *Küustlergesellschaft* for correct *Künstlergesellschaft* (English: *artist association*). For every unknown word, this substitution is applied to every occurrence of *u*, producing correction hypotheses. For example, from unknown *Tourenverzeichuissen* the variants *Tonrenverzeichuissen* and *Tourenverzeichnissen* are derived, the latter being the correct form meaning *tables of tours*. Of course, all of the substitution pairs (such as the inverse: $n \rightarrow u$) are applied to all words, as well as combinations of them up to a predefined recursion depth, rendering a large number of variants.

In the first step, the known words of the initial categorization serve as a lexicon. All hypothetical spelling variants derived from unknown words are looked up in this corpus-specific lexicon and, if present, are considered a spelling correction of the underlying unknown word. With this method, we can correct misrecognised words if their correct form is also present in the corpus (and has been judged correct by the categorizing mechanism) and if the differences of the known- / unknown-word pair can be described by means of the predefined substitution set. We found this to be safe (i. e. has a high precision), but only very few corrections are done overall (low recall).

In the second step, the hypothetical variants of unknown words that have not been corrected in the first step are analysed by Gertwol to find corrections beyond the corpus-derived lexicon. Word forms appearing only a few times throughout the corpus are likely to have no correct version in the corpus, if they happen to be misrecognised, so they cannot be caught in the first step. Therefore, the hypothetical variants are sent through the categorizing process as described above to find correct words among the bulk of non-words. The words found are taken as a correction.

In pre-evaluation during building, this method turned out to be error-prone in terms of categorizing non-words as correct words and hence introducing bad substitutions to the correction table. The reason for these erroneous jugdements lies in the “creativity” of Gertwol in analyzing words as compounds. This shall be illustrated by an example: The name of a famous gorge at the old route across the Gotthard pass, *Schöllenenschlucht*, had been correctly recognised by the OCR systems, but it was tagged ‘unknown’ by our categorizing method. As a

consequence, the word is passed to the correction tool which produces, among others, the hypothetical variant *Sehölleuenschlucht* by substituting one *c* by *e* and one *n* by *u*. Unfortunately, Gertwol claims to know this word, analyzing it as `Seh#\"ol#leu\en#schlucht`, which can only be interpreted as a fantastic word creation, to be approximately translated as *gorge of the seeing-oil lions*. In order to suppress the sometimes amusing, but undesired analyses of this kind, we apply a filter to the Gertwol output, stopping compound segmentations with tiny elements like *-seh-*.

The processing configuration of the post-correction system including the first and the second step of searching for corrections, is referred to as the 'basic' configuration in the evaluation. To keep the number of searching variants small, only a limited number of character substitutions is followed in the basic configuration. Of course, real-life OCR-errors are not limited to a closed set of character substitutions, but rather show a wide range of deviations from their original word. The aim of the third step is to account for a broad variety of character operations to find correction candidates for unknown words.

This last task is done by a module that partly implements Martin Reynaert's TICCL algorithm (see [10]). As the algorithm is already well described there, only the basic idea is given here: With the anagram hashing technique, Reynaert found a way to treat words as "bags of letters" to reduce searching complexity. The words are stored by a numerical hash key and character operations (such as deletion, insertion, substitution) can be modelled with arithmetic operations. By addition and subtraction one can easily get from one word to all other words having a majority of characters in common.

## 4 Evaluation

The aim of our evaluation was to measure the influence of several methods for improving OCR accuracy in the Text+Berg project. At different stages of the digitization process, various attempts were made to reduce the rate of OCR-errors. First we briefly specify the evaluation method. The results are then presented in Sect. 4.2.

### *4.1 Evaluation Setup*

The evaluation is based on the Text+Berg *Release 131* [4]. The test corpus was compiled from four volumes of different periods, namely the SAC yearbooks from 1890, 1899, 1912 and 1950. Based on the automatically recognized text, these four books had been manually corrected. The corrected books seem to be of high quality and serve as a gold standard for this evaluation. However, it cannot be ruled out that a certain amount of OCR-errors remained undetected in the manual correction process.

In order to reduce the complexity of measuring and to increase the reliability of the results while working with a multilingual and diachronic corpus, only the German parts of the books were used. Tokens not containing at least one of the basic Latin alphabet's letters were rejected (i.e. numbers and punctuation were left out, as well as noisy tokens). The text was extracted from the XML files as found in the *Release 131* and compared using the ISRI OCR-Evaluation Frontiers Toolkit [11] for aligning and measuring word accuracy.

## *4.2 Evaluation Results*

### 4.2.1 Standard Processing Modules

To measure the influence of the dictionaries and the merging and polishing module, which are already standard parts of our corpus building process, the whole processing pipeline has been run four times with different configurations, each having one of the modules switched off. One pipeline run was done with standard settings including all modules. For every output, word accuracy was measured by determining agreement with the gold standard. Since the modules evaluated are expected to have an improving effect on the data, their switching off should lead to a lower score than the standard settings. This assumption is shown to hold in most cases.

**Table 3** Word accuracy of the standard modules

| | standard settings | | without ancient spelling dict | | without merging module | | without polisher module | |
|---|---|---|---|---|---|---|---|---|
| 1890 | 119008 | Words | | | | | | |
| | 94.38 | % Accuracy | 94.36 | % Accuracy | 94.17 | % Accuracy | 94.32 | % Accuracy |
| | 6689 | Errors | 6718 | Errors | 6944 | Errors | 6765 | Errors |
| | | | +29 | (0.43 %) | +255 | (3.67 %) | +76 | (1.12 %) |
| 1899 | 111967 | Words | | | | | | |
| | 99.49 | % Accuracy | 99.53 | % Accuracy | 99.35 | % Accuracy | 99.50 | % Accuracy |
| | 575 | Errors | 527 | Errors | 727 | Errors | 557 | Errors |
| | | | -48 | (9.11 %) | +152 | (20.91 %) | -18 | (3.23 %) |
| 1912 | 93750 | Words | | | | | | |
| | 99.23 | % Accuracy | | | 99.11 | % Accuracy | 99.17 | % Accuracy |
| | 720 | Errors | | | 835 | Errors | 776 | Errors |
| | | | | | +115 | (13.77 %) | +56 | (7.22 %) |
| 1950 | 135844 | Words | | | | | | |
| | 99.49 | % Accuracy | | | 99.42 | % Accuracy | 99.49 | % Accuracy |
| | 691 | Errors | | | 785 | Errors | 699 | Errors |
| | | | | | +94 | (11.97 %) | +8 | (1.14 %) |

The results are shown in table 3. As can be seen from the accuracy values, overall quality is high (> 99 %) as compared to the gold standard, with the oldest yearbook being an exception (approx. 94 %). The deviations are not exceedingly high when seen in relation to the book size, but they are remarkable with respect to the number of misrecognized words. The merging module achieves the best results. In all books of the test corpus, turning it off leads to a serious increase of misrecognized words; e.g. in yearbook 1912 a total of 727 words differ from what they should (according to the gold standard) if the merging is not done, but with the standard settings, where the merging is included, this number is reduced by a fifth to 575 differing words. Note that we found a higher number of word error corrections in the manual evaluation of the 1899 yearbook than in the automatic one. We attribute this discrepancy to errors in the gold standard, since we found several corrections being counted as new errors.

The results of the pattern-based correction module are less convincing, but it seems to be helping in the most cases, too. As for the yearbooks 1890, 1912 and 1950 a low to moderate improvement can be achieved by applying some regular-expression patterns during postprocessing; but as for 1899 the module had better not been used since its switching off leads to a higher agreement with the gold standard. The reasons for this unexpected result have not been analyzed yet, but see also 4.2.2 for concerns about the gold standard accuracy.

The results are bad for the semiautomatically built (and very incomplete) dictionary for the 19th-century spelling. The word list was based on a manually corrected list of old spelling variants found in the yearbook 1890. Nevertheless, adding this list to the OCR system has almost no effect on the recognition of that very yearbook (a plus of 29 words out of 6718 are correctly recognized), although the dictionary was overfitted for this data. One might conclude, that the OCR system does not trust the dictionaries provided by the user. This finding corresponds with the observations reported in [5]. The score of the other 19th-century yearbook, 1899, even shows worse results when adding the dictionary. Since this impairment cannot be explained by non-use of the dictionary by the OCR system (if the dictionary was simply ignored, the result would be the same as with the standard setting), the dictionary must be either misleading the system or we are dealing with gaps in the gold standard (cf. 4.2.2).

### 4.2.2 New Post-correction Module

Our post-correction tool using Gertwol and the SwissTopo list as categorizers (introduced in 3.2.5) has not yet been integrated into the processing pipeline. Therefore the new tool was evaluated against the existing standard settings to see if it leads to any improvement. Seven configurations have been evaluated, four of which will be presented here. The configurations differ in the minimal-length threshold and the combinations of the different correction steps as described in 3.2.5.

The post-correction tool was run with the corpus of the Text+Berg *Release 131*. Each configuration run led to a correction file, listing OCR-errors with a single

correction each. Every correction list was applied to a copy of the pipeline output for each of the four yearbooks of the test corpus, substituting every occurrence of an OCR-error by the proposed correction. The post-corrected books were then each compared with their corresponding gold standard book.

Table 4 sums up the results. The leftmost column shows the original standard-settings pipeline output with no further post-correcting and is identical with the first column of the evaluation table 3. The basic configuration includes steps 1 and 2, i. e. correction candidates were only searched by applying a small set of character substitutions. In the second configuration, a minimal version was tried, which performed only step 1 (corrections were only sought within the known words of the corpus). The last configuration was done using all of three correction steps, including the TICCL module, which has no limitations on character substitutions.

**Table 4** Word accuracy of the new post-correction module

| | standard settings | | basic | | corpus-derived lexicon only | | with TICCL | |
|---|---|---|---|---|---|---|---|---|
| 1890 | 119008 | Words | | | | | | |
| | 94.38 | % Accuracy | 94.37 | % Accuracy | 94.38 | % Accuracy | 94.22 | % Accuracy |
| | 6689 | Errors | 6698 | Errors | 6689 | Errors | 6884 | Errors |
| | | | +9 | (0.13 %) | +0 | | +195 | (2.83 %) |
| 1899 | 111967 | Words | | | | | | |
| | 99.49 | % Accuracy | 99.46 | % Accuracy | 99.49 | % Accuracy | 99.41 | % Accuracy |
| | 575 | Errors | 601 | Errors | 575 | Errors | 665 | Errors |
| | | | +26 | (4.32 %) | +0 | | +90 | (13.53 %) |
| 1912 | 93750 | Words | | | | | | |
| | 99.23 | % Accuracy | 99.21 | % Accuracy | 99.23 | % Accuracy | 99.14 | % Accuracy |
| | 720 | Errors | 736 | Errors | 722 | Errors | 802 | Errors |
| | | | +16 | (2.17 %) | +2 | (0.28 %) | +82 | (10.22 %) |
| 1950 | 135844 | Words | | | | | | |
| | 99.49 | % Accuracy | 99.48 | % Accuracy | 99.49 | % Accuracy | 99.41 | % Accuracy |
| | 691 | Errors | 703 | Errors | 691 | Errors | 803 | Errors |
| | | | +12 | (1.71 %) | +0 | | +112 | (13.94 %) |

Unlike the previous evaluation setup, where the modules were switched off to observe their effect, the post-correction tools here are added to existing settings. Improvements could therefore be seen in a decrease of the misrecognized words. Disturbingly, not a single instance of a reduced word error rate can be found in the evaluation table. In the reduced configuration doing only one correction step, hardly any substitutions can be seen at all. The other configurations introduce heaps of words judged as additional misrecognized ones.

Since observations and partial evaluations during building did not indicate such disastrous results, manual re-checking was done. All word replacements that have

actually been done while applying the correction lists to the test corpus had been saved to logfiles and could easily be reconstructed. It became clear that the tool measuring word accuracy ignored such substitutions as the deletion of noisy characters in word tokens, which makes us lose a good part of the good substitutions in the evaluation score. Then, the gold standard's reliability has to be questioned when measuring word accuracy in ranges above 99 %. For example, in the reduced configuration ('corpus-derived lexicon only'), the – truly – misrecognized word *Verkanfs-magazine* is replaced by correct *Verkaufsmagazine* (English: *vending magazine*). Unfortunately, the gold standard has the misspelt version here, so the good correction is falsely judged a bad replacement. Although the rate of OCR-errors still present in the gold-standard books is probably low, the few remaining errors have a great chance to be found by the post-correction module, which gives them considerable weight in the automatic evaluation.

Two random samples of each 50 replacements have been manually checked thereafter. In the TICCL-configuration, 18 replacements were found to be good, 29 lead to an impairment and 3 were judged neutral (the text is neither improved nor impaired by those substitutions). Among the undesired replacements, mostly correct words became another known word. A major problem is the haphazardness of correct words to be known or not known by the resources which we used. For example, the compound participle *bestgemachten* (English: *best made*) cannot be analyzed by Gertwol, but *selbstgemachten* (a compound participle as well, meaning *self made*) is known, and since the two words happen to be within a Levenshtein distance of 3, a replacement is triggered. The same holds for the two mountain names *Kienthaleralpen* and *Simmenthaleralpen* (with 19th-century spelling, but both correctly recognized), where the first is known, but not the latter.

While the full configuration of the post-correction tool – after a closer look to the data – still seems not to be performing too well as of now, the basic configuration turns out to be far better than at first sight. Out of the random sample of 50 replacements, 44 have shown to be perfectly well, 5 had better not happened and 1 could not be categorized. So the overall result is positive after all. Again, the two circumstances mentioned above lead to the poor evaluation scores: The – desirable – removing of non-alphanumeric characters is not reported in the accuracy values, and uncorrected OCR errors in the gold standard can be critical. Having the latter in mind, one might suppose that the evaluation results of the first evaluation (see 4.2.1) could be expected to be higher, if the gold-standard books were further improved.

## 5 Related Work

Holley [5] provides an excellent overview of the issues involved in improving the text quality when converting historic newspaper corpora with OCR. She concludes that the most promising way is collaborative correction and describes their approach of having users correct the Australian Newspaper collection in [6]. A project to

digitize a large collection of Romansch texts also aims at collaborative user input [9].

Other methods for automatic OCR-error correction include e.g. statistical approaches as described in [10] and [7], as well as lexical approaches as in [13]. As mentioned before, some of our experiments were inspired by Reynaert [10] who worked on cleaning a digitized collection of historical Dutch newspapers. He has developed an efficient way of mapping an unknown word to its most similar known words which can then be used as substitute in the text. In contrast Kolak et al. [7] present a finite-state character sequence transformation system. They are interested in improving OCR in new languages and show in their experiments that an OCR system for English can produce good results for French when combined with GIZA-style word alignment. Strohmaier [13] has collected domain-specific lexicons by crawling the web. He has shown that these lexicons help to improve the OCR accuracy (for the OCR systems of his time, Abbyy FineReader version 5 and OmniPage version 10). But he has also demonstrated that the combination of two OCR systems leads to improved accuracy which is well in line with our results.

As for the combination of multiple OCR systems, research has identified two main questions: how to efficiently align the output of multiple OCR systems (e.g. [8]), and how to select the optimal word among different candidates. The question of output alignment arises because multiple OCR systems will result in different tokenisations. The word selection step has been performed using voting algorithms [12], dictionaries [8], or human post-editing [1].

## 6 Conclusion

We are working on the digitization and annotation of Alpine texts. Currently we compile a corpus of German and French yearbooks from the Swiss Alpine Club that span 145 years. In the next step we will digitize the French yearbooks *L'Echo des Alpes* that were published in Switzerland from 1871 until 1924 to counterbalance the German language dominance in the SAC collection. We also have an agreement with the British Alpine Club to include their texts in our corpus.

In this paper we have presented various methods aimed at reducing OCR errors. We discussed the enlargement of the OCR system lexicon and various post-correction methods. The lexicon enlargement had surprisingly little impact on the results. It seems that the OCR system does not make good use of the additional lexical material and, annoyingly, the OCR company leaves the user in the dark as to when and how additional lexicons will improve the accuracy rate.

In addition, we have implemented three different post-correction methods. Our post-correction heuristics based on regular expression substitutions aim at obvious OCR errors and are thus a reliable correction method with low recall. The other two methods are more flexible, but only the merging of the output of two OCR systems leads to clearly improved text accuracy. Our final attempt, employing external

resources, has not resulted in clear improvements, although our manual inspections showed a number of interesting automatic corrections.

An obvious extension of our merging approach is the inclusion of further OCR systems. For this, Tesseract is an attractive candidate since it is open-source and can thus be tuned to handle those characters well where we observe special weaknesses in the commercial OCR systems.

Our merging procedure also triggered further ideas for combining other textual sources. Our parallel French and German books since the 1950s contain many identical texts. These books are only partially translated, and they partially contain the same article in both books. We have already found out that even the same OCR system (Abbyy FineReader) makes different errors in the recognition of the two versions of the (same) text (e.g. *in der Gipfelfaüinie* vs. *inj der Gipfelfallinie*). This gives us more variants of the same text which we can merge.

We are also wondering whether the same text scanned under different scanner settings, e.g. different contrasts or different resolution, will lead to different OCR results which could be merged towards improved results. For instance, a certain scanner setting (or a certain image post-correction) might suppress dirt spots on the page which may lead to improved OCR quality.

Finally we would also like to explore whether translated texts can help in OCR error correction. Automatic word alignment might indicate implausible translation correspondences which could be corrected via orthographically similar, but more frequent aligned words.

**Acknowledgements** The authors would like to thank Torsten Marek for his contributions to the OCR merging. Many thanks also to the many student helpers who have contributed to the Text+Berg project. We are especially grateful for the support by the Swiss Alpine Club and by Hanno Biber and his team from the Austrian Academy Corpus. Part of this research has been funded by the Swiss National Science Foundation in the project "Domain-specific Statistical Machine Translation".

## References

1. Abdulkader, A., Casey, M.R.: Low cost correction of OCR errors using learning in a multi-engine environment. In: Proceedings of the 10th International Conference on Document Analysis and Recognition (2009)
2. Borin, L., Kokkinakis, D., Olsson, L.J.: Naming the past: Named entity and animacy recognition in 19th century Swedish literature. In: Proceedings of The ACL Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007). Prague (2007)
3. Bubenhofer, N.: Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. No. 4 in Sprache und Wissen. de Gruyter, Berlin, New York (2009)
4. Bubenhofer, N., Volk, M., Althaus, A., Bangerter, M., Marek, T., Ruef, B.: Text+Berg-Korpus (Release 131). XML-Format (2010). Digitale Edition des Jahrbuch des SAC 1864-1923 und Die Alpen 1925-1995
5. Holley, R.: How good can it get? Analysing and improving the OCR accuracy in large scale historic newspaper digitisation programs. D-Lib Magazine **15**(3/4) (2009)

6. Holley, R.: Many hands make light work: Public collaborative OCR text correction in australian historic newspapers. Tech. rep., National Library of Australia (2009)
7. Kolak, O., Byrne, W., Resnik, P.: A generative probabilistic OCR model for NLP applications. In: NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pp. 55–62. Association for Computational Linguistics, Morristown, NJ, USA (2003)
8. Lund, W.B., Ringger, E.K.: Improving optical character recognition through efficient multiple system alignment. In: Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (JDLC09), pp. 231–240. Austin, TX (2009)
9. Neuefeind, C., Steeg, F.: Digitale rätoromanische Chrestomathie - Werkzeuge und Verfahren für die kollaborative Volltexterschließung digitaler Sammlungen. In: Poster bei der DGfS Jahrestagung. Göttingen (2011)
10. Reynaert, M.: Non-interactive OCR post-correction for giga-scale digitization projects. In: A. Gelbukh (ed.) Proceedings of the Computational Linguistics and Intelligent Text Processing 9th International Conference, CICLing 2008, Lecture Notes in Computer Science, pp. 617–630. Springer, Berlin (2008)
11. Rice, S.V.: Measuring the accuracy of page-reading systems. Ph.D. thesis, University of Nevada (1996)
12. Rice, S.V., Kanai, J., Nartker, T.A.: A report on the accuracy of OCR devices. Tech. rep., University of Nevada (1992). Technical Report
13. Strohmaier, C.M.: Methoden der lexikalischen nachkorrektur OCR-erfasster dokumente. Ph.D. thesis, Ludwig-Maximilians-Universität, München (2004)
14. Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., Ruef, B.: Challenges in building a multilingual alpine heritage corpus. In: Proceedings of LREC. Malta (2010)
15. Witte, R., Gitzinger, T., Kappler, T., Krestel, R.: A Semantic Wiki Approach to Cultural Heritage Data Management. In: Proceedings of LREC Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008). Marrakech, Morocco (2008)

# Alignment between Text Images and their Transcripts for Handwritten Documents

Alejandro H. Toselli, Verónica Romero and Enrique Vidal

**Abstract** An alignment method based on the Viterbi algorithm is proposed to find mappings between word images of a given handwritten document and their respective (ASCII) words in its transcription. The approach takes advantage of the underlying segmentation made by Viterbi decoding in handwritten text recognition based on Hidden Markov Models (HMMs). Two levels of alignments are considered: the traditional one at word level and the one at text-line level where pages are transcribed without line break synchronization. According to various metrics used to measure the quality of the alignments, satisfactory results are obtained. Furthermore, the presented alignment approach is tested on two different HMMs modelling schemes: one using 78 HMMs (one HMM per character class) and other using two HMMs (for blank space and no-blank characters respectively).

**Key words:** handwriting image and transcription alignments, forced recognition, digital libraries, handwritten text recognition, Viterbi algorithm

Dr. Alejandro H. Toselli
Instituto Tecnológico de Informática - Universidad Politécnica de Valencia
Camino de Vera s/n - 46022 Valencia - Spain,
e-mail: ahector@iti.upv.es

Dr. Verónica Romero
Instituto Tecnológico de Informática - Universidad Politécnica de Valencia
Camino de Vera s/n - 46022 Valencia - Spain,
e-mail: vromero@iti.upv.es

Dr. Enrique Vidal
Instituto Tecnológico de Informática - Universidad Politécnica de Valencia
Camino de Vera s/n - 46022 Valencia - Spain,
e-mail: evidal@iti.upv.es

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_2, © Springer-Verlag Berlin Heidelberg 2011

## 1 Introduction

Lately, many on-line digital libraries have been publishing large quantities of digitized handwritten documents, which allows both scholars and the general public to access this kind of cultural heritage resources. This is a new, comfortable way of consulting and querying this material. The *Biblioteca Valenciana Digital* (BiValDi)[1] is an example of such digital libraries, which provides an interesting collection of handwritten documents.

Many of these handwritten documents include both, the handwritten material and its proper transcription (in ASCII format for example). Generally speaking, most documents have transcriptions aligned only at the page level, but not at individual text lines, making it difficult the visualization and consulting of these documents for the paleography experts. In Fig. 1 an example of an original piece of manuscript (left) and its corresponding transcription without any kind of alignment (center) is shown.

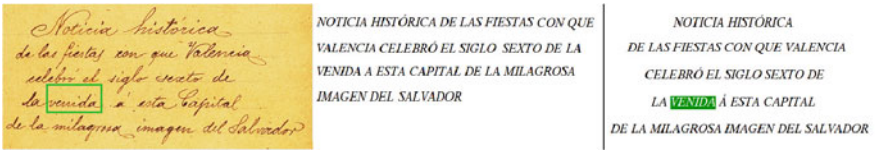


**Fig. 1** Left: original image. Center: a corresponding unaligned transcription. Right: the line and word-alignments have been computed. The lines in the manuscript have the same words that the lines at the transcription and an example of a word-alignment is shown with an outlined word (using a box) in the manuscript (left) and the corresponding highlighted word (in reverse video) in the transcript (right)

This fact has motivated the development of techniques to align these documents and their transcripts; i.e. to generate a mapping between each line or word image on a document page with its respective line or word on its electronic transcript. This kind of alignment can help readers to quickly locate image text while reading a transcript, with useful applications to editing, indexing, etc. In the opposite direction, the alignment can also be useful for people trying to read the image text directly, when arriving to complex or damaged parts of the document.

Two different levels of alignment can be defined: line level and word level as shown in Fig 1 (right). Line alignments attempt to obtain beginning and end positions of lines in transcribed pages that do not have synchronized line breaks. This information allows users to easily visualize the page image documents and their corresponding transcriptions. Moreover, using these alignments as segmentation ground truth, large amounts of training and test data for segmentation-free cursive handwriting recognition systems become available. On the other hand, word alignments allow users to easily find the place of a word in the manuscript when reading the corresponding transcript. On a graphical interface properly designed to show word alignments, for example, one can display both the handwritten page

[1] http://bv2.gva.es

image and the transcript and, whenever the mouse is held over a word in the transcript, the corresponding word in the handwritten image would be outlined using a box. In a similar way, whenever the mouse is held over a word in the handwritten image, the corresponding word in the transcript would be highlighted (see Fig. 2).

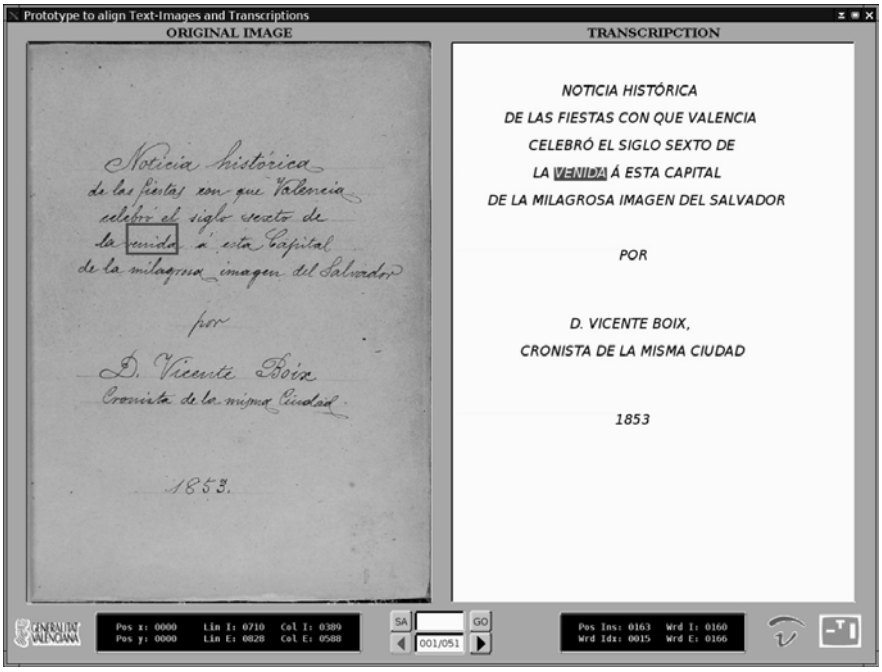


**Fig. 2** Screen-shot of the alignment prototype interface displaying an outlined word (using a box) in the manuscript (left) and the corresponding highlighted word in the transcript (right)

It is worth noting that word alignments can be displayed without need of line alignments. However, the visualization of the manuscript and its transcription with both alignments is friendlier for the readers than using only the word alignments.

Creating such alignments is challenging since the transcript is an ASCII text file while the manuscript page is an image. In the case of word alignments, some recent works address this problem by relying on a previous explicit image-processing based word pre-segmentation of the page image, before attempting the transcription alignments. For example, in [6], the set of previously segmented word images and their corresponding transcriptions are transformed into two different times series, which are aligned using *dynamic time warping* (DTW). In this same direction, [3], in addition to the word pre-segmentation, a (rough) recognition of the word images is attempted. The resulting word string is then aligned with the transcription using dynamic programming.

The alignment method presented here (henceforward called Viterbi alignment), relies on the Viterbi decoding approach to handwritten text recognition (HTR)

based on Hidden Markov Models (HMMs) [2, 9]. These techniques are based on methods originally introduced for speech recognition [4]. In such HTR systems, the alignment is actually a byproduct of the proper recognition process, i.e. an implicit segmentation of each text image line is obtained, where each segment successively corresponds to one recognized word. In our case, word recognition is not actually needed, as we do already have the correct transcription. Therefore, to obtain the segmentations for the *given* word sequences, the so-called "forced-recognition" approach is employed (see Sect. 2.2). This idea has been previously explored in [13].

As it has been explained previously, line alignments only make sense in document transcriptions where the beginning and end positions of the image lines are not registered. However, the word alignments can be computed both for line transcriptions or for page transcriptions. In this work, line and word alignment results are reported for a set of 53 pages from a XIX century handwritten document (see Sect. 5.2). To evaluate the quality of the obtained alignments, several metrics were used which give information basically at two different alignment levels: the accuracy of alignment mark placements and the the amount of erroneous assignments produced between word images and transcriptions (see Sect. 4).

The remainder of this paper is organized as follows. First, the alignment framework is introduced and formalized in Sect. 2. Then, an implemented prototype is described in Sect. 3. The alignment evaluation metrics are presented in Sect. 4. The experiments and results are commented in Sect. 5. Finally, some conclusions are drawn in Sect. 6.

## 2 HMM-based HTR and Viterbi Alignment

HMM-based handwritten text recognition is briefly outlined in this section, followed by a more detailed presentation of the Viterbi alignment approach.

### *2.1 HMM HTR Basics*

The traditional handwritten text recognition problem can be formulated as the problem of finding a most likely word sequence $\hat{\mathbf{w}} = \langle w_1, w_2, \ldots, w_n \rangle$, for a given handwritten sentence (or line) image represented by a feature vector sequence $\mathbf{x} = x_1^p = \langle x_1, x_2, \ldots, x_p \rangle$, that is:

$$\begin{aligned} \hat{\mathbf{w}} &= \arg\max_{\mathbf{w}} \Pr(\mathbf{w}|\mathbf{x}) \\ &= \arg\max_{\mathbf{w}} \Pr(\mathbf{x}|\mathbf{w}) \cdot \Pr(\mathbf{w}) \end{aligned} \tag{1}$$

where $\Pr(\mathbf{x}|\mathbf{w})$ is usually approximated by concatenated character Hidden Markov Models (HMMs) [2,4], whereas $\Pr(\mathbf{w})$ is approximated typically by an $n$-gram word language model [4].

Thus, each character class is modeled by a continuous density left-to-right HMM, characterized by a set of states and a Gaussian mixture per state. The Gaussian mixture serves as a probabilistic law to model the emission of feature vectors by each HMM state. Figure 3 shows an example of how a HMM models a feature vector sequence corresponding to character "a". The process to obtain feature vector sequences from text images as well as the training of HMMs are explained in Sect. 3.

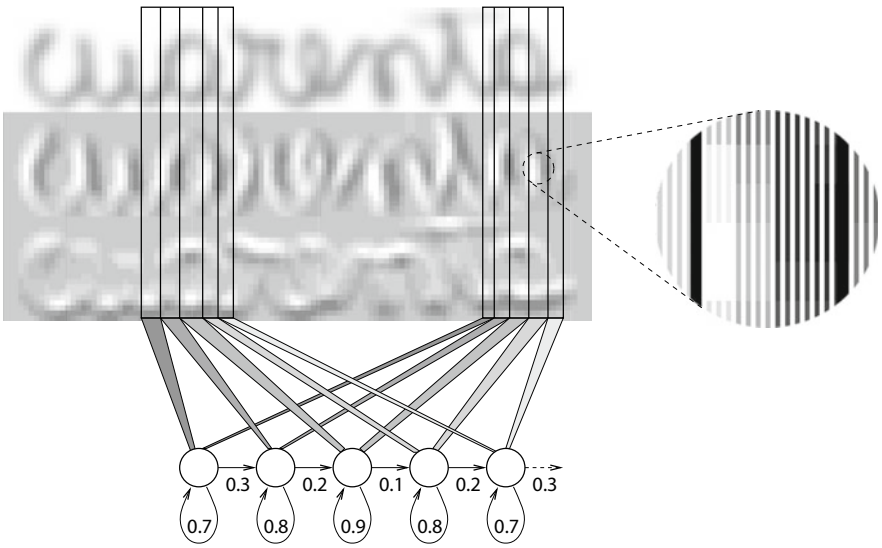


**Fig. 3** Example of 5-states HMM modeling (feature vectors sequences) of instances of the character "a" within the Spanish word "cuarenta" (forty). The states are shared among all instances of characters of the same class. The zones modelled by each state show graphically subsequences of feature vectors (see details in the magnifying-glass view) compounded by stacking the normalized grey level and its both derivatives features

HMMs as well as n-grams models can be represented by stochastic finite state networks (SFN), which are integrated into a single global SFN by replacing each word character of the n-gram model by the corresponding HMM. The search involved in the Equ. (1) to decode the input feature vectors sequence $\mathbf{x}$ into the more likely output word sequence $\hat{\mathbf{w}}$, is performed over this global SFN. This search problem is adequately solved by the Viterbi algorithm [4].

## 2.2 Viterbi Alignment

As a byproduct of the Viterbi solution to (1), the feature vector subsequences of $\mathbf{x}$ aligned with each of the recognized words $w_1, w_2, \ldots, w_n$ can be obtained. This subsequence alignment is implicit or "hidden" in the Equ. (1), which can thus be rewritten as:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum_{\mathbf{b}} \Pr(\mathbf{x}, \mathbf{b} \mid \mathbf{w}) \cdot \Pr(\mathbf{w}) \tag{2}$$

where $\mathbf{b}$ is an *alignment*; that is, an ordered sequence of $n+1$ marks $\langle b_0, b_1, \ldots, b_n \rangle$, used to demarcate the subsequences belonging to each recognized word. The marks $b_0$ and $b_n$ always point out to the first and last components of $\mathbf{x}$ (see Fig. 4).

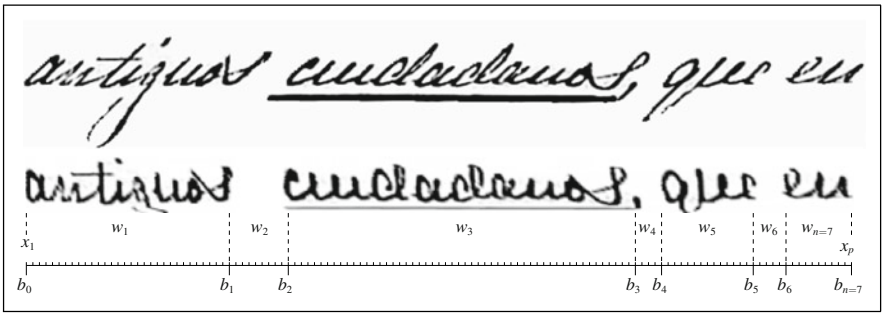


**Fig. 4** Example of segmented text line image along with its resulting deslanted and size-normalized image. Moreover, the alignment marks ($b_0 \ldots b_8$) which delimit each of the words (including word-spaces) over the text image feature vectors sequence $\mathbf{x}$.

Now, approximating the sum in (2) by the dominant term, $\max_{\mathbf{b}} \Pr(\mathbf{x}, \mathbf{b} \mid \mathbf{w})$:

$$(\hat{\mathbf{b}}, \hat{\mathbf{w}}) \approx \arg\max_{\mathbf{b},\mathbf{w}} \Pr(\mathbf{w}) \cdot \Pr(\mathbf{x}, \mathbf{b} \mid \mathbf{w}) \tag{3}$$

where $\hat{\mathbf{b}}$ is the optimal alignment. In our case, we are not really interested in text recognition proper, because the transcription is known beforehand. Let $\tilde{\mathbf{w}}$ be the given transcription. Now, $\Pr(\mathbf{w})$ in Equ. (3) is zero for all $\mathbf{w}$ except $\tilde{\mathbf{w}}$, for which $\Pr(\tilde{\mathbf{w}}) = 1$. Therefore,

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} \Pr(\mathbf{x}, \mathbf{b} \mid \tilde{\mathbf{w}}) = \Pr(x_{b_0}^{b_1}, x_{b_1}^{b_2}, \ldots, x_{b_{n-1}}^{b_n} \mid \tilde{\mathbf{w}}) \tag{4}$$

which can be expanded to,

$$\hat{\mathbf{b}} = \arg\max_{\mathbf{b}} \Pr(x_{b_0}^{b_1} \mid \tilde{\mathbf{w}}) \Pr(x_{b_1}^{b_2} \mid x_{b_0}^{b_1}, \tilde{\mathbf{w}}) \ldots \Pr(x_{b_{n-1}}^{b_n} \mid x_{b_0}^{b_{n-1}}, \tilde{\mathbf{w}}) \tag{5}$$

Assuming that each subsequence $x_{b_{i-1}}^{b_i}$ is independent from that it its predecessors and also depends only of the ith word of $\tilde{w}_i$, Equ. (5) can be rewritten as,

$$\hat{\mathbf{b}} \approx \arg\max_{\mathbf{b}} \Pr(x_{b_0}^{b_1}|\tilde{w}_1) \ldots \Pr(x_{b_{n-1}}^{b_n}|\tilde{w}_n) \tag{6}$$

This problem is optimally solved by using Viterbi search, and this is known as "forced recognition".

### *2.3 Word and Line Alignments*

The word alignments are obtained directly from Equ. (6). If the transcription is given at the line level, the input feature vector sequence $\mathbf{x}$ represents a handwritten text line image and $\tilde{\mathbf{w}}$ its corresponding line transcription. If the transcription is given only at page level, the input is a very long feature vector sequence that represents the whole page. This sequence is obtained by concatenating the feature vector sequences of all the successive text line images of the document page. Accordingly, $\tilde{\mathbf{w}}$ in this case is considered a single, correspondingly long word sequence, without line breaks.

Obviously, when the feature vector sequences of the different text line images are concatenated into the whole-page sequence, we know the positions in $\mathbf{x}$ that identify the joints between lines. Using this information the line alignments are easily computed. Let $\mathbf{l} = \langle l_1, \ldots, l_M \rangle$ be this sequence of positions, where $M$ is the number of lines in the page, and let $x_{b_{i-1}}^{b_i}$ be the feature vector subsequence belonging to the word $w_i$. This word is considered to belong to the line $j$ if $b_{i-1}$ and $b_i$ are between $l_{j-1}$ and $l_j$. Sometimes, it may happen that word boundaries ($b_{i-1}$ and $b_i$) are in two different lines; $b_{i-1}$ is between $l_{j-1}$ and $l_j$, but $b_i$ is between $l_j$ and $l_{j+1}$. In this case if $l_j - b_{i-1} \geqslant b_i - l_j$ the word $w_i$ is considered to belong to the line $j$; otherwise it is considered to belong to the line $j+1$.

This way, once the word alignment has been computed for the whole page, the line alignment is obtained by visiting sequentially each word in the transcript and deciding which line it belongs to.

## 3 Overview of the Alignment Prototype

The implementation of the alignment prototype involved four different parts: document image preprocessing, line image feature extraction, HMMs training and alignment map generation.

Document image preprocessing encompasses the following steps: first, skew correction is carried out on each document page image; then background removal and noise reduction is performed by applying a bi-dimensional median filter [5] on the whole page image. Next, a text line extraction process based on local minimums of the horizontal projection profile of page image, divides the page into separate line images [7]. In addition connected components has been used to solve the situations

where local minimum values are greater than zero, making it impossible to obtain a clear text line separation. Finally, slant correction and non-linear size normalization are applied [8, 9] on each extracted line image. An example of extracted text line image is shown in the top panel of Fig. 4, along with the resulting deslanted and size-normalized image. Note how non-linear normalization leads to reduced sizes of ascenders and descenders, as well as to a thiner underline of the word "ciudadanos".

As our alignment prototype is based on Hidden Markov Models (HMMs), each preprocessed line image is represented as a sequence of feature vectors. To do this, the feature extraction module applies a grid to divide each line image into $N \times M$ squared cells. In this work, $N$=40 is chosen empirically (using the corpus described further on) and $M$ must satisfy the condition $M/N = \textit{original image aspect ratio}$. From each cell, three features are calculated: normalized gray level, horizontal gray level derivative and vertical gray level derivative. The way these three features are determined is described in [9]. Columns of cells or *frames* are processed from left to right and a feature vector is constructed for each *frame* by stacking the three features computed in its constituent cells. Hence, at the end of this process, a sequence of $M$ 120-dimensional feature vectors (40 normalized gray-level components, 40 horizontal and 40 vertical derivatives components) is obtained. An example of feature vectors sequence, representing an image of the Spanish word *"cuarenta"* (forty) is shown in Fig. 3.

As it was explained in Sect. 2.1, characters are modeled by continuous density left-to-right HMMs with 6 states and 64 Gaussian mixture components per state. This topology (number of HMM states and Gaussian densities per state) was determined by tuning empirically the system on the corpus described in Sect. 5.1. Once a HMM "*topology*" has been adopted, the model parameters can be easily trained from images of continuously handwritten text (*without any kind of segmentation*) accompanied by the transcription of these images into the corresponding sequence of characters. This training process is carried out using a well known instance of the EM algorithm called *forward-backward or Baum-Welch re-estimation* [4].

The last phase in the alignment process is the generation of the mapping proper by means of Viterbi "forced recognition", as discussed in Sect. 2.2.

## 4 Alignment Evaluation Metrics

Several kinds of measures have been adopted to evaluate the quality of alignments at line and at word level.

On the one hand, line alignments are evaluated by means of the line error rate (LER), the average number of words assigned to erroneous lines (AEW) and the maximum number of erroneous words by line (MWE). The LER measures the number of system proposed lines that have different word count than their corresponding reference line-synchronized transcriptions, divided by the total number of lines. The AEW is the number of words that have been assigned to an incorrect line, divided by the total number of words. The MWE, finally, measures the maximum

difference between the word count of the proposed line and that of the correct line-synchronized transcription. Figure 5 shows an example of a line alignment for the image shown in the Fig. 1. The LER, in this example, would be 40%, the AEW 4% and the MWE to 1. A perfect line alignment would have LER=AEW=MWE=0.

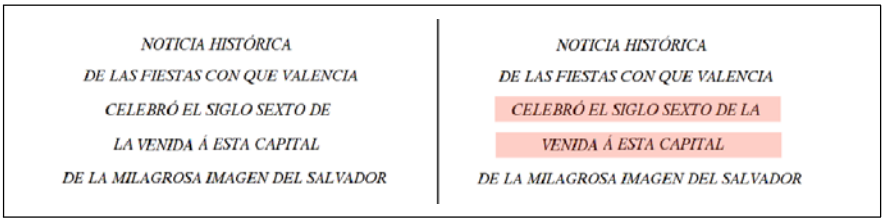


**Fig. 5** Example of LER, WAE and MWE computation for the image show in the Fig. 1. On the left we can see the correct lines and in the right the system proposed line alignments. In this case 2 lines have been erroneously aligned (the highlighted lines). The word "LA" has been assigned to the line 3 instead of the line 4. The resulting LER is 40%, the AEW is 4% and the MWE is 1.

On the other hand, quality of word alignments are measured by the alignment error rate (AER) and the average value and standard deviation (henceforward called MEAN-STD) of the absolute differences between the system-proposed word alignment marks and their corresponding (correct) references. The MEAN-STD gives us an idea of the geometrical accuracy of the word alignments obtained, whereas the AER measures the amount of totally erroneous assignments produced between word images and transcriptions.

Given a reference mark sequence $\mathbf{r} = \langle r_0, r_1, \ldots, r_n \rangle$, along with an associated word token sequence $\mathbf{w} = \langle w_1, w_2, \ldots, w_n \rangle$, and an automatic segmentation mark sequence $\mathbf{b} = \langle b_0, b_1, \ldots, b_n \rangle$ (with $r_0 = b_0$, $r_n = b_n$), we define the MEAN-STD and AER metrics as follows:

**MEAN-STD:** The average value and standard deviation of absolute differences between reference and proposed alignment marks, are given by:

$$\mu = \frac{\sum_{i=1}^{n-1} d_i}{n-1} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n-1} (d_i - \mu)^2}{n-1}} \qquad \text{where} \quad d_i = |r_i - b_i| \tag{7}$$

**AER:** Defined as:

$$\text{AER}(\%) = \frac{100}{N} \sum_{j:w_j \neq s} e_j \qquad \text{with} \quad e_j = \begin{cases} 0 & b_{j-1} < m_j < b_j \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

where $s$ stands for the blank-space token, $N < n$ is the number of real words (i.e., tokens which are not $s$) and $m_j = (r_{j-1} + r_j)/2$.

A good word alignment would have a $\mu$ value close to 0 and small $\sigma$. Thus, MEAN-STD gives us an idea of how accurate are the automatically computed word alignment marks. On the other hand, AER assesses word alignments at a

higher level; that is, it measures mismatches between word-images and ASCII transcriptions (tokens), excluding word-space tokens. This is illustrated in Fig. 6, where the AER would be 25%.

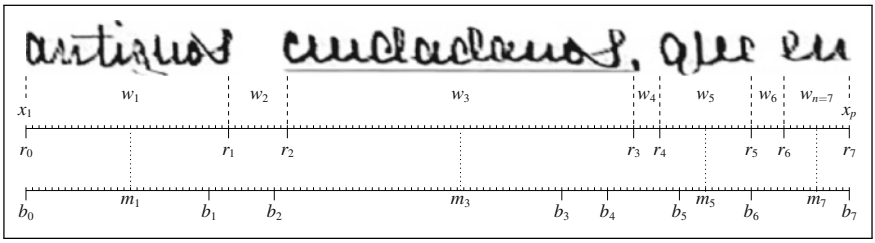


**Fig. 6** Example of AER computation. In this case $N=4$ (only no word-space are considered: $w_1, w_3, w_5, w_7$) and $w_5$ is erroneously aligned with the subsequence $x_{b_5}^{b_6}$ ($m_5 \notin (b_4, b_5)$). The resulting AER is 25%.

## 5 Experiments

In order to test the effectiveness of the presented alignment approach, different experiments were carried out. The corpus used, as well as the experimental setup carried out and the obtained results, are explained in the following subsections.

### *5.1 Corpus Description*

The corpus was compiled from the legacy handwriting document identified as *Cristo-Salvador*, which was kindly provided by the *Biblioteca Valenciana Digital* (BiValDi). It is composed of 53 text page images, scanned at 300dpi and written by only one writer. Some of these page images are shown in the Fig. 7.

As has been explained in Sect. 3, the page images have been preprocessed and divided into lines, resulting in a data-set of 1,172 text line images. In this phase, around 4% of the automatically extracted line-separation marks were manually corrected. All the page transcriptions are available with synchronized line breaks containing 10,911 running words with a vocabulary of 3,408 different words.

To test the quality of the computed alignments, 12 pages were randomly chosen from the whole corpus to be used as references. For these pages the true locations of alignment marks were set manually. Table 1 summarized the basic statistics of this corpus and its reference pages.

**Fig. 7** Example of page images of the corpus "Cristo-Salvador" (CS), which show backgrounds of big variations and uneven illumination, spots due to the humidity, marks resulting from the ink that goes through the paper (called bleed-through), etc.

**Table 1** Basic statistics of the CS corpus (book)

| Number of: | References | Total | Lexicon |
|---|---|---|---|
| pages | 12 | 53 | – |
| text lines | 312 | 1,172 | – |
| words | 2,955 | 10,911 | 3,408 |
| characters | 16,893 | 62,159 | 78 |

## *5.2 Experiments and Results*

Since only forced-recognition is needed to obtain line-level word alignments, training and test are carried out on identical data; namely, the whole set of document lines. In principle, a similar process could be applied for the page-level alignments; that is, training the HMMs on the long, whole-page feature vector sequences and the corresponding page transcriptions. However, this process has proved not to be straightforward. On the one hand, training accuracy tends to degrade with the length of the training sequences. On the other hand, training time becomes prohibitive, since the Baum-Welch computing time is proportional to the length of the feature vector sequence times the number of words in the transcription. Therefore, to obtain page-level alignments, HMM models are trained on line data, as in the case of the line-level alignments, but using only a part of the document that has been marked with synchronized line-breaks.

In our experiments, the 41 non-reference pages (860 lines) have been used for training and the test has been carried out with the 12 reference pages. Table 2 shows the line alignment results. The results obtained clearly suggest that, for large documents with hundreds of pages, it would be quite profitable to devote some work

to line-synchronize a (relatively small) part of the document by hand and let the rest be automatically aligned by the system.

**Table 2** Line alignment evaluation results: *Line Error Rate* (LER), *Average Number of Words Assigned to Erroneous Lines* (AEW) and *Maximum number of Erroneous Words by line* (MWE).

| LER % | MEW | AEW % |
|---|---|---|
| 6.4 | 1 | 0.3 |

Word alignments can be computed, in the CS corpus, using either line-synchronized transcriptions or whole page transcriptions. In the first case, HMMs were trained with all the 1,172 lines (53 pages) of the corpus and the test has been carried out with the 312 reference lines (12 pages). In the second case, the same HMM models trained for the page-level experiments have been used. Additionally, for the first case (word-alignments at line level), two different HMM modeling schemes were employed: one modeling each of the 78 character classes using a different HMM per class, or other modeling separately the blank "character" class and all the 77 no-blank character classes using two HMMs respectively. Whichever the case, the HMM topology was identical for all HMMs in both schemes: left-to-right with 6 states and 64 Gaussian mixture components per state. Table 3 reports the results of the quality of the obtained word alignments at line level (for the both above-mentioned HMM modeling schemes) and at page level.

**Table 3** Word alignment evaluation results at line and page levels: *Alignment Error Rate* (AER) and *Average Value and Standard Deviation* of the absolute differences between the system-proposed word alignment marks and their corresponding (correct) references (MEAN-STD).

| | Line Level | | Page Level |
|---|---|---|---|
| | 78-HMMs | 2-HMMs | |
| AER (%) | 7.20 | 25.98 | 7.88 |
| $\mu$ (mm) | 1.14 | 2.95 | 1.15 |
| $\sigma$ (mm) | 3.90 | 6.56 | 3.43 |

Several comments about these results are in order. Firth, the system-proposed line alignments are quite acceptable. The results show that from every 100 automatically aligned lines only 6 need to be corrected by the user. Moreover, the error of each incorrect line is due to one word at most. It is worth noting that although the number of erroneous lines is 6%, the number of incorrect system-proposed line breaks is 3%, because each incorrect line break involves two erroneous lines. Overall, the number of words that have been assigned to an erroneous line is just 0.3% of the total number of words.

Furthermore, from the word alignment results using the 78 HMMs scheme, we can see that computing the alignments at line level the best AER is obtained (7.20%). Moreover, the relative low values of $\mu$ and $\sigma$ show that the obtained alignment

marks are quite accurate; that is, they are very close to their respective references. This is illustrated on the left histogram of Fig. 8. The two typical alignment errors

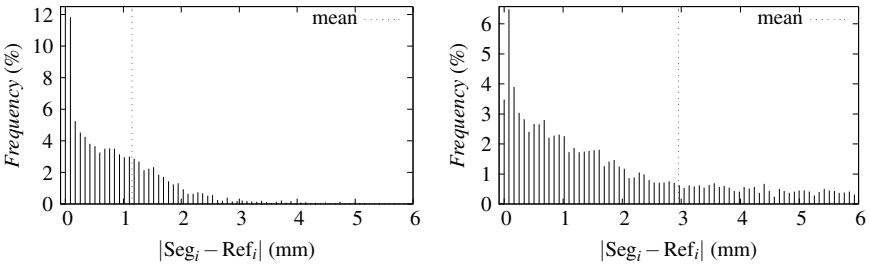


**Fig. 8** $|r_i - b_i|$ distribution histograms for 78-HMMs (left) and 2-HMMs (right) modelling schemes.

are known as over-segmentation and under-segmentation respectively. The over-segmentation error is when one word image is separated into two or more fragments. The under-segmentation error occurs when two or more images are grouped together and returned as one word. Figure 9 shows some of them.

**Fig. 9** Word alignment for 6 lines of a particularly noisy part of the corpus. The four last words on the second line as well as the last line illustrate some of over-segmentation and under-segmentation error types.

If the word alignments are computed at the page level the results are quite acceptable as well, showing that from every 100 automatically aligned words only 8 are misaligned.

## 6 Remarks, Conclusions and Future Work

Given a manuscript and its transcription, we propose an alignment method to map every line or word image on the manuscript with its respective line or word on the electronic (ASCII or PDF) transcript. This method takes advantage of the implicit

alignments made by Viterbi decoding used in forced text image recognition with HMMs.

Experiments have been carried out with the CS corpus, which is a legacy handwritten document written in 1853. Despite the difficulty that entails the task, the results achieved in this work are encouraging.

In future works, we plan to carry out the training of the different HMMs models using whole pages, trying to solve the previously explained problems that this training entails. In the experiments carried out in this work, the different HMM models have been trained using text line images, but this way to perform the training is not viable for documents that do not have, at least, a part of their transcriptions synchronized at this line level. Other interesting issue is to follow the new interactive framework presented in [12]. This framework integrates the human activity into the recognition process taking advantage of the user's feedback. This idea has been previously applied to computer assisted translation (CAT) [1], Computer assisted speech transcription (CATS) [12] and computer assisted transcription of handwritten text images (CATTI) [10, 11] with good result.

**Acknowledgements** Work supported by the EC (FEDER), the Spanish MEC under the MIPRCV "Consolider Ingenio 2010" research programme (CSD2007-00018) and the Spanisg Government (MICINN and "Plan E") under the MITTRAL (TIN2009-14633-C03-01) research project.

## References

1. Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Ney, A.L.H., Tomás, J., Vidal, E.: Statistical approaches to computer-assisted translation. Computational Linguistics p. In press (2008)
2. Bazzi, I., Schwartz, R., Makhoul, J.: An Omnifont Open-Vocabulary OCR System for English and Arabic. IEEE Trans. on PAMI **21**(6), 495–504 (1999)
3. Huang, C., Srihari, S.N.: Mapping Transcripts to Handwritten Text. In: S. Ltd. (ed.) Tenth International Workshop on Frontiers in Handwriting Recognition, pp. 15–20. La Baule, France (2006)
4. Jelinek, F.: Statistical Methods for Speech Recognition. MIT Press (1998)
5. Kavallieratou, E., Stamatatos, E.: Improving the quality of degraded document images. In: DIAL '06: Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL'06), pp. 340–349. IEEE Computer Society, Washington DC, USA (2006). DOI http://dx.doi.org/10.1109/DIAL.2006.23
6. Kornfield, E.M., Manmatha, R., Allan, J.: Text Alignment with Handwritten Documents. In: First International Workshop on Document Image Analysis for Libraries (DIAL), pp. 195–209. Palo Alto, CA, USA (2004)
7. Marti, U.V., Bunke, H.: Using a Statistical Language Model to improve the preformance of an HMM-Based Cursive Handwriting Recognition System. Int. Journal of Pattern Recognition and Artificial In telligence **15**(1), 65–90 (2001)
8. Romero, V., Pastor, M., Toselli, A.H., Vidal, E.: Criteria for handwritten off-line text size normalization. In: Procc. of The Sixth IASTED Int. Conf. on Visualization, Imaging, and Image Processing (VIIP 06). Palma de Mallorca, Spain (2006)
9. Toselli, A.H., Juan, A., Keysers, D., González, J., Salvador, I., H. Ney, Vidal, E., Casacuberta, F.: Integrated Handwriting Recognition and Interpretation using Finite-State Models. Int. Journal of Pattern Recognition and Artificial Intelligence **18**(4), 519–539 (2004)

10. Toselli, A.H., Romero, V., Pastor, M., Vidal, E.: Multimodal interactive transcription of text images. Pattern Recognition **43**(5), 1814–1825 (2009)
11. Toselli, A.H., Romero, V., Rodríguez, L., Vidal, E.: Computer Assisted Transcription of Handwritten Text. In: 9th Int. Conf. on Document Analysis and Recognition (ICDAR 2007), pp. 944–948. IEEE Computer Society, Curitiba, Paraná (Brazil) (2007)
12. Vidal, E., Rodríguez, L., Casacuberta, F., García-Varea, I.: Interactive pattern recognition. In: Proc. of the 4th Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, *LNCS*, vol. 4892, pp. 60–71. Springer, Brno, Czech Republic (2007)
13. Zimmermann, M., Bunke, H.: Automatic Segmentation of the IAM Off-Line Database for Handwritten English Text. In: Proc. of the 16 th Int. Conf. on Pattern Recognition (ICPR'02) Volume 4, p. 40035. IEEE Computer Society, Washington, DC, USA (2002)

# Part II
# Adapting NLP Tools to Older Language Varieties

# A Diachronic Computational Lexical Resource for 800 Years of Swedish

Lars Borin and Markus Forsberg

**Abstract** The goal of the work presented in this chapter is to create a set of computational lexical resources, interlinked on the lexical sense level using the persistent sense identifiers designed for the Present-Day Swedish lexical resource SALDO. In this way, all the diverse linguistic information available in our individual lexical resources – modern and historical – becomes available for all resources where the interlinking has been completed. Using this mechanism, we have been able to devise a semantic search application for 19th century fiction which combines a morphological component for this language variety – Late Modern Swedish – and a lexical-semantic resource for Present-Day Swedish through the lexical sense identifiers. At present, we are extending and cross-linking the modern lexical resources, as well as steadily integrating the 19th century resource into this emerging diachronic lexical resource. We are also working on the intricate and challenging problem of incorporating the most deviant historical language variety – Old Swedish – in this resource, which we clearly have to do before we can truthfully refer to it as a diachronic computational lexical resource for 800 years of Swedish.

**Key words:** lexical resources, language tools, language resources, cultural heritage, language technology

Lars Borin
Språkbanken, Department of Swedish Language, University of Gothenburg,
Box 200, SE-405 30 Gothenburg, Sweden
e-mail: lars.borin@svenska.gu.se

Markus Forsberg
Språkbanken, Department of Swedish Language, University of Gothenburg,
Box 200, SE-405 30 Gothenburg, Sweden
e-mail: markus.forsberg@gu.se

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_3, © Springer-Verlag Berlin Heidelberg 2011

## 1 Introduction

Our cultural heritage takes many forms, one of the most important being *text*. Indeed, the most common definition of history has it starting with the invention of writing – the long period before the emergence of writing being referred to as "prehistory" – and the subsequent recording of events that this invention enabled. Furthermore, and for obvious reasons, any form of any language older than a bit over a century will be accessible *only* in writing, i.e., in the form of text.

An important access venue to our past is consequently the texts accumulated over the centuries in language communities with a long written tradition, making up a rich cultural heritage reflecting different periods in the history of the language community. In these texts, we find information on many aspects of the origins and historical development of our culture. However, these texts will also inevitably reflect the fact that language changes over time. This means that if there are texts that go sufficiently far back in time, these texts – even though they are nominally in our own language – will be increasingly difficult to understand to a modern reader.

Languages change on all levels. Vocabulary turnover tends to be rapid, so that even texts written only a generation or two ago will make a slightly odd impression. Given enough time, the grammar and sound system of a language will change beyond recognition by others than specialists. Indeed, the oldest extant texts in many European languages must in fact be translated in order to be accessible to a modern reader.

This of course has consequences for those attempting to apply modern language tools to the massive amounts of digitized older text that now have started to appear in digital libraries and other digital cultural heritage collections, not least on the internet. The motivation for undertaking this kind of work is obviously to make our cultural heritage accessible to the public as well as to provide state-of-the-art tools to researchers who wish to utilize this rich text material as primary research data. These tools should ideally provide at least the same kind of functionality as the tools that we develop for working with modern texts, i.e., linguistically aware information retrieval, information extraction and text mining, machine translation, etc.

However, since older forms of the language – as we noted above – are increasingly different from the modern language, language tools need to be adapted or built from scratch in order to deal successfully with these historical language stages. Which approach is chosen for any particular case may of course vary depending on the availability of language resources and tools for the modern language, linguistic expertise in the research group, and whether the knowledge residing in the language tools is primarily in the form of manually formulated rules or statistical, acquired through machine learning.

For our work on computational lexical resources for processing texts written in older forms of Swedish, we have chosen a methodology where we adapt resources and tools originally developed for the modern language. This is partly for pragmatic reasons: We have at our disposition a rich array of lexical resources and tools for the modern language, which are under active development in several interconnected projects. It is however also partly methodologically motivated. One way of looking

at historical forms of Swedish – or indeed any language – is as different, but closely related languages. Arguably, language tools are to a considerable extent reusable across closely related languages, especially the kind where the linguistic knowledge resides in explicitly manually formulated rules [19]. We are also in the fortunate position of having in our research unit several digitized large dictionaries of historical forms of Swedish, and are actively in search of more such resources. Thus, our strategy is to create a *diachronic computational lexical resource for Swedish* by interlinking our individual lexical resources and extending the linguistic description format used for the modern language to earlier language stages.

The historical periods of Swedish relevant to the work described in this chapter are the following:[1]

**Early Old Swedish (1225–1374 CE):** This is a historical form of Swedish very similar to Old Norse or Old English in its general language structure (see Sect. 4).

**Late Old Swedish (1375–1526 CE):** In this period the complex inflectional morphology characteristic of the previous stage undergoes a radical simplification in the direction of the modern system. In the phonology the vowel system is thoroughly restructured.

**Early Modern Swedish (1527–1732 CE):** During this period, a written standard language slowly emerged, in no small part due to the development of book printing.

**Late Modern Swedish (1733–1906 CE):** This is the forerunner of the contemporary language. A standard orthography and written language grammar are firmly consolidated during this period.

**Contemporary Swedish (1906 CE – present):** The starting point of what Swedish linguists refer to as "Contemporary Swedish" is conveniently placed in the year 1906, the date of the latest Swedish spelling reform, which established the orthography used today.

One more systematic and pervasive change in the written language is dated after the start of the Contemporary Swedish period, namely the abandonment of subject-verb number agreement, which had run its course only in the 1950s. Of course, vocabulary has also been changing all through the century-long Contemporary Swedish period. Because of this, we find it convenient to introduce the label **Present-Day Swedish** for the language of the last half-century, which is roughly the period covered by our lexical and other resources for the modern language.

The rest of this chapter is organized as follows: In Sect. 2, we describe the Present-Day Swedish lexical resource that has been chosen as the 'pivot' – both conceptually and in practice – for our modern and historical lexical resources and accompanying language processing tools. For completeness' sake, we will briefly mention the other modern resources that we are merging in the Swedish FrameNet++ project – since they will be important for the historical resources in the longer perspective – but for our present purposes, the historical resources are in focus: In Sects. 3 and 4 we describe our work with lexical resources and morphology tools for

[1] The listed periods are preceded by **Runic Swedish** (or Old East Norse), the language of the runestones of Sweden and Denmark. This language stage is extant mainly in the form of a small number of short epigraphs written in the runic alphabet.

nineteenth-century and medieval Swedish, respectively. In Sect. 5 we sum up our work so far, say something about our experiences from the particular approach that we have chosen, and present some of our plans for the future.

## 2 Lexical Resources for Present-Day Swedish

### *2.1 SALDO*

Our central lexical resource for Present-Day Swedish is SALDO [3–5, 7], or SAL version 2, a free modern Swedish semantic and morphological lexicon intended for language technology applications. The lexicon is available under a Creative Commons Attribute-Share Alike license and LGPL 3.0.

SALDO started its life as *Svenskt associationslexikon* [21] – 'The Swedish associative thesaurus', a so far relatively unknown Swedish thesaurus with an unusual semantic organization, reminiscent of, but different from that of WordNet [5]. SAL has been published in paper form in two reports, from the Center for Computational Linguistics [22], and the Department of Linguistics [21], both at Uppsala University. Additionally, the headwords and their basic semantic characterizations have been available electronically, in the form of text files, from the very beginning.

The history of SAL has been documented by Lönngren [20] and Borin [3]. Initially, text corpora were used as sources of the vocabulary which went into SAL, e.g., a Swedish textbook for foreigners and a corpus of popular-scientific articles. A small encyclopedia and some other sources provided the large number (over 3000) of proper names found in SAL. Eventually, a list of the headwords from *Svensk ordbok* [28] was acquired from the NLP and Lexicology Unit at the University of Gothenburg, and the second paper edition of SAL [21] contained 71,750 entries. At the time of writing, SALDO contains over 108,000 entries, the increased number being because a number of new words have been added, but also because a number of entries belong to more than one part of speech or more than one inflectional pattern.

The central semantic relation of SALDO is *association*, a 'non-classical' lexical-semantic relation [23]. SALDO describes *all* words semantically, not only the open word classes. By way of illustration, Fig. 1 shows the semantic 'neighbors' (rendered in blue/non-bold) in SALDO of the word *telefon* 'telephone (noun)'. It is associated i.a. with words like *samtala* 'hold a conversation', *telefonledes* 'by phone', *pulsval* 'pulse dialling', *ringa* 'call (verb)', *mobiltelefon* 'mobile phone', the proper name *Bell*, and many others, as shown in Fig. 1.[2]

We soon realized that in order to be useful in language technology applications, SAL would have to be provided at least with part-of-speech and inflectional morphological information – both entirely absent from SAL in its original form – and SALDO was created. The morphological component of SALDO has been defined

[2] From http://spraakbanken.gu.se/ws/saldo-ws/lid/html/telefon..1.

| | |
|---|---|
| **lex:** | **telefon** |
| **l:** | telefon+nn |
| **fm:** | samtala |
| **fp:** | PRIM |
| **mf**(19): | **PRIM:** fingerskiva hörtelefon kobra[2] pulsval ringa telefonautomat telefonera telefonledes telefonlur telefonör tonval<br>**bild:** bildtelefon<br>**knapp**[3]**:** knapptelefon<br>**lokal**[2]**:** lokaltelefon<br>**lyssna:** hörlur<br>**mobil:** mobiltelefon<br>**port:** porttelefon<br>**trådlös:** radiotelefon<br>**vägg:** väggtelefon |
| **pf**(18): | **abonnent:** telefonabonnent<br>**anrop:** telefonanrop<br>**apparat:** telefonapparat<br>**avgift:** teleavgift<br>**central:** telefonstation<br>**elledning:** telefonledning<br>**fingerskiva:** petmoj<br>**förbindelse:** telefförbindelse<br>**katalog:** telefonkatalog<br>**kontakt**[2]**:** jack[2]<br>**samtal:** telefonsamtal<br>**signal:** telefonsignal<br>**sladd:** telefonsladd<br>**svara:** telefonsvarare telefonvakt<br>**teknisk:** teleteknisk<br>**ton:** kopplingston<br>**uppfinnare:** Bell |

**Fig. 1** Semantic neighbors (rendered in blue/non-bold) of *telefon* 'telephone (n)' in SALDO

using Functional Morphology (FM) [10, 11], a tool that provides a development environment for computational morphologies (see Sect. 4.1).

As one of its distribution channels, SALDO is published as web services, updated daily. Web services provide clean interfaces and instant updates, but are restricted to small amounts of data because of network latency. Presently available web services include incremental fullform lookup, semantic lookup, compound analysis, and an inflection engine service.[3]

Crucially for the purposes of the present chapter, the information model of SALDO has been carefully designed with reusability and resource cross-linking in mind. This is accomplished by the use of persistent identifiers for the four kinds of lexical entities defined in SALDO: word senses, lemgrams (our term for a combination of lemma and inflectional class), inflectional classes/paradigms, and semantic relationships among senses. Consequently SALDO is used as the pivot resource both in our work on an integrated lexical resource for Present-Day Swedish described in the next section and in our work on a diachronic lexical resource for Swedish which is the topic of the rest of this chapter. Using a common pivot resource in this way, we hope to be able to reuse in many ways the large amounts of hard-earned lexical information now residing in various diverse resources.

[3] See http://spraakbanken.gu.se/eng/saldo/ws.

## 2.2 Swedish FrameNet++

The Swedish FrameNet++ project (SweFN++) is a recent computational lexical project undertaken by our research unit.[4] SweFN++ has two goals:

1. to merge a number of existing, freely available lexical resources, by harmonizing their data formats and information models. The merging is ultimately done by linking them on the sense or lemgram level using SALDO identifiers. In other words – as mentioned in the previous section – SALDO is used as the 'pivot' or 'hub' resource in all cases. The most important of these existing resources are:
   - the Swedish PAROLE lexicon resulting from the EC project PAROLE (1996–1998), containing 29,000 syntactic units representing syntactic valence information;
   - the Swedish SIMPLE lexicon resulting from the EC project SIMPLE (1998–2000), containing 8,500 semantic units being characterised with respect to semantic type, domain and selectional restrictions;
   - Swesaurus [6], a kind of wordnet with graded synonymy relations among senses;
   - the Gothenburg Lexical Database (GLBD), a lexical database for modern Swedish covering 61,000 entries with an extensive description of their inflection, morphology and semantics. SDB (Semantic Database) is a version of GLDB where many of the verb senses have been provided with semantic valence information using a set of about 40 general semantic roles [16] and linked to example sentences in a reference corpus of Present-Day Swedish.
2. to compile a Swedish FrameNet, building on the Berkeley FrameNet [1], again using SALDO sense and lemgram identifiers as the 'atomic' lexical units of the resource. At the time of writing the Swedish FrameNet contains 347 frames and 14,600 lexical units.

Importantly for the work presented in this chapter, the interlinking of the Present-Day Swedish lexical resources through SALDO, together with a similar linking of lexical resources for older stages of Swedish through SALDO, potentially promises to make the rich grammatical and semantic information present in the modern resources available also for the processing of older texts. To which extent this promise will actually be borne out is an exciting empirical question which we have only started to explore.

[4] See http://spraakbanken.gu.se/eng/swefn.

## 3 A Lexical Resource for 19th Century Swedish

*A.F. Dalin: Ordbok öfver svenska språket (1850-1855)* [8] is a dictionary that has been digitized at our research unit, and is the starting point for our work on 19th century Swedish.

The language of Dalin – Late Modern Swedish according to the traditional linguistic periodization of Swedish given above – is close to the modern language of SALDO, where the differences are in minor spelling variations, such as the use of the letter <f> instead of <v>, and some morphological differences, such as inflection of verbs in person and number, completely absent in Present-Day Swedish. Moreover, there is a difference in the vocabulary, since the vocabulary of a dictionary reflects the society it was produced in, e.g., many of the words in Dalin have to do either with agriculture or with religious matters.

We have created a morphology for Dalin by adapting the morphological component developed for SALDO. With a comparatively small effort we have been able to provide around 80% of the entries with an inflectional pattern based on the inflectional information provided in Dalin. However, the inflectional information in Dalin is underspecified, which means that there are erroneous word forms that we need to weed out. Also, the remaining 20% of the entries will require a considerably larger effort.

The linking of Dalin to SALDO has been done by analyzing modernized headwords of the entries in Dalin using SALDO. Connecting on the entry level is a first, over-generating, approximation, since the senses of Dalin is given within an entry. For the linking we were able to reuse an existing resource, as a manual spelling translation for the entries in Dalin had been made in an earlier project in our department and preserved in a database, which by a stroke of good luck still existed in one of our servers.

Since the vocabulary of Dalin reflects another time and another societal structure, many of its content words are not in SALDO. However, in a large number of cases they are compound words where the constituents of the compound are in SALDO. An example could be the headword *bäfverhund* with a modern spelling *bäverhund* 'beaver dog', meaning a dog used for hunting beavers, a word that would normally not find its way into a modern lexical resource – since the practice it refers to is no longer pursued – and adding the word to the modern resource would be unsatisfactory, since to a modern reader *bäverhund* would at most be a completely transparent compound, meaning 'a dog in some way connected to beavers', but without the conventionalized or lexicalized meaning the word had earlier.

In cases like this we instead choose another approach for linking the resources. Even though *bäverhund* is not in SALDO, both *bäver* 'beaver' and *hund* 'dog' are, so we use SALDO to do a compound analysis of *bäverhund* to *bäver+hund*, and link with respect to the head of the compound, i.e., *hund*.

The resulting entry for *bäfverhund*, which is analogous to the other linked entries in Dalin, is summarized in the table below. Every dictionary entry has been given a persistent identifier, here *bäfverhund..e.1*. We have the headword, its modern spelling, the inflectional information in Dalin, *m. 2.*, and the paradigm identifier

it has been associated to, *nn_2m_ulf*. The paradigm identifier together with the headword defines the inflection table, which will be explained in more detail in Sect. 4.1. Finally, we have the connection to the sense identifier in SALDO, *hund..1*.

| *headword* | *modern* | *pos* | *paradigm* | *gram. desc.* | *saldo* | *id* |
|---|---|---|---|---|---|---|
| bäfverhund | bäverhund | nn | nn_2m_ulf | m. 2. | hund..1 | bäfverhund..e.1 |

Dalin contains many verb entries consisting of prefix plus verb written as one word, which in Present-Day Swedish generally correspond to phrasal verbs, i.e., verb plus separate particle/adverb. E.g., *påspäda* 'add to', where *på* 'on, onto' is the particle, would in modern Swedish be a phrasal verb, *späda på*, or in its more common short form, *spä på*. We deal with these cases by allowing adverbs as the first constituent of a compound, given that the head of the compound is a verb. For example:

| *headword* | *modern* | *pos* | *paradigm* | *gram. desc.* | *saldo* | *id* |
|---|---|---|---|---|---|---|
| påspäda | påspäda | vb | vb_2a_ärfva | v. a. 2. | spä_på..1 | påspäda..e.1 |

There is still much work to be done on the 19th century resource, but it is now in such a shape that we have been able to harvest the fruits of its creation. We have performed some encouraging experiments with semantic search in a large archive of digitized 19th century Swedish fiction using semantic relations in the modern resources, which are available in the 19th century resource through the links to SALDO.

## 4 A Lexical Resource for Old Swedish

As mentioned above, the Old Swedish period (1225–1526) – conventionally subdivided into Early (1225–1374) and Late (1375–1526) Old Swedish – covers a time span of 300 years. The language of the extant Old Swedish texts exhibits considerable variation, for at least the following reasons:

1. The orthography was not standardized in the way that we expect of modern literary languages;
2. the language itself was not standardized, in the sense, e.g., that a deliberate choice would had been made about which of several competing forms should be used in writing; and
3. the Middle Ages was a time of rapid language change in Swedish, perhaps more so than any subsequent period of similar length.

The first shows itself in a great variation in spelling; the 'same' word can be spelled in a number of different ways, even on the same page of a document. Not only is there variation in the orthography itself, but also geographical variation, because no unified standard variety had been established at the time when the texts were produced.

The second factor makes itself felt in the number of variant forms in our inflectional paradigms (Sect. 4.1).

As for the third factor, during the second half of the Old Swedish period the language underwent a development from the Old Norse (or Modern Icelandic, or Old English) mainly synthetic language type to the present, considerably more analytical state. In addition (or perhaps compounding this process), the sound system of Swedish was thoroughly reorganized.

For instance, in the nouns, the case system changed profoundly during this period, from the old four-case system (nominative, accusative, dative, genitive, in two numbers) to the modern system with a basic form and a genitive clitic which is the same in all declensions (as opposed to the old system where there were a number of different genitive markers), and where most functions that the older case forms expressed by themselves have been taken over by a combination of free grammatical morphemes and a much more rigid constituent order.

In the available collections of digitized Old Swedish texts, these changes are in full swing, which manifests itself as variation in inflectional endings and in the use of case and other inflectional categories and in the distribution of the corresponding forms.

It is not always easy to tease out the contributions of these different factors to the linguistic motley evinced by the texts. Without doubt, the diachronic component is important – the texts are after all from a period three centuries in length – but it is also probable that the lack of standardization simply allows normal synchronic language variation to 'shine through' in the texts, as it were, rather than being eliminated as is normally the case with modern, normalized written standard languages.

**fisker** (Söderwall) ( fysker Lg 3: 301; -ar BSH 5: 5067 (1512). fiisker: fiisk RK 3: 4179. -ar), m. [Isl. fiskr] L. **1)** *fisk*. han tok w fiske tolpänigh Bu 100. taka fiska KL 12. thz förста han katadhe vt sin krok tha fik lhan en storan fisk ib. ib 13. Bo 240. Lg 546, 3: 9, 10, 301, 302. i slike watne äru tholka fiska GO 978. ätin the fiska oc hwitan maat Bir 4: 15. färska ällir salte fiska ib 5: 32. tw pund skarpa fisca SD NS 1: 656 (1407). - koll. han (qvarndammen) skal vara open þree vikur vm varenä þaa fisken gaar vpp ok swa lenge vm hösten. þaa vatneth er mykith ok fiksen gar vpp FH 3: 4 (1352). ib 4: 15 (1451), 16. SD 5: 699 (1347, gammal afskr.). äta fisk oc hwitan maat Bir 4: 15. VKR 17, 62. fäghin är han som fyrme ok findher han fikh (för fiskh) a diska GO 105. tw stykke fisk Bir 5: 31. tw stykke färskan fisk ib 32. eet stykke stekan fisk Bo 234. ii pund fiisk RK 3: 4179. **2)** ?iiij (4) lösa järn bultar, item xi (11) lösa fyskar, item 1 fangabult BSH 5: 506 (1512). - Jfr arbeidis-, bnären-, flat-, horn-, hval-, skal-, skarp-, skat-, sma-, spit-, stok-fisker. — **fiska bater** (-baater: -baat Su 363), m. *fiskarebåt*. Su 363. — **fiska ben**, n. *fiskben*. eet fiska ben sath fast j hans halse Bil 900. KL 370. ST 102. — **fiska dike**, n. *fiskdamm*. ST 299. — **fiska drät**, f.L. — **fiska fäl**, n. *fiskfjäll*. aff rutnom fiska fiällom Bir 3: 203. — **fiska fänge** ( fiske-), n. *fiskafänge*. aff the fiske fängeno Lg 3: 11. aff hwario fiske fänge ib. — **fiska hovudh** ( hwffwd LB 7: 265), n. *fiskhufvud*. LB 7: 265. PM XLVIII. — **fiska kyn** ( -kön), n. *fiskslag*. alla handa fiska kön Al 6495. — **fiska lim**, n. *fisklim*. tak fiska lim giorth aff maghommen PM XLVII. — **fiska liver** ( -leffwer: leffrenas PM XXXVIII), f. *fisklefver*. PM XXXVIII. — **fiska läghe**, n. *fiskläge*. RK 1: (Yngre red. af LRK) s. 263. Jfr fiskeläghe. — **fiska skal**, f. *musselskal, snäckskal*. trykte han ällir wredh vth aff vlla fätthen ena fiska skal äller eeth kar fwlt mz daagh (concham rore implevit) MB 2: 88. — **fiska slagh**, n. *fiskslag*. mang the fiska slagh, som aldrig fingos ther förra Lg 3: 11. — **fiska sudh** (fiiska sodh LB 7: 159), n. *fiskspad*. aff fersko fiska sudhi LB 3: 182. ib 7: 159. — **fiska thiuver**, m. L.

**Fig. 2** The entry *fisker* 'fish' in Söderwall's dictionary of Old Swedish

Our point of departure for creating a lexical resource for Old Swedish are digitized versions of the full text of the three main reference dictionaries of Old Swedish:

- Söderwall's dictionary of Old Swedish [29] (23,000 entries);
- the supplement to Söderwall's dictionary [30] (21,000 entries); and
- Schlyter's dictionary of the language of the Old Swedish laws [27] (10,000 entries)

The overlap between the three dictionaries is great, so that we are actually dealing with less than 25,000 different headwords. On the other hand, compounds – whether written as one word or separately – are not listed as independent headwords, but as secondary entries under the entry of one of the compound members. Thus, a full morphological description reflecting the vocabulary of the three dictionaries will contain many more entries, possibly an order of magnitude more.

As an example of the kind of information that is available in the dictionaries, we will briefly discuss the entry for the word *fisker* 'fish', as it appears in Söderwall's dictionary [29], as shown in Fig. 2. From this entry we learn that *fisker* is a masculine noun (indicated by "m.", in the second line of the entry), and that it has been attested in a number of variant spellings (*fysker*, *fiisker*, *fiisk*). We also find references to occurrences of the word in the classical texts, and finally there is a listing of the compounds in which it occurs, e.g. *fiska slagh* 'type of fish'.

**Table 1** The inflection table of *fisker*

***nn_m_fisker fisker*** ⇒

| **Lemma** | fisker | | | **Traditional normalized form** |
|---|---|---|---|---|
| **POS** | nn | | | |
| **Gender** | m | | | |
| **Number** | **Def** | **Case** | **Word form** | |
| sg | indef | nom | *fisker* | *fiskęr* |
| sg | indef | gen | *fisks* | *fisks* |
| sg | indef | dat | *fiski, fiske, fisk* | *fiski, fisk* |
| sg | indef | ack | *fisk* | *fisk* |
| pl | indef | nom | *fiska(r), fiskæ(r)* | *fiska(r)* |
| pl | indef | gen | *fiska, fiskæ* | *fiska* |
| pl | indef | dat | *fiskum, fiskom* | *fiskum* |
| pl | indef | ack | *fiska, fiskæ* | *fiska* |
| sg | def | nom | *fiskrin* | *fiskrin* |
| sg | def | gen | *fisksins* | *fisksins* |
| sg | def | dat | *fiskinum, fisk(e)num* | *fiskinum* |
| sg | def | ack | *fiskin* | *fiskin* |
| pl | def | nom | *fiskani(r), fiskæni(r)* | *fiskani(r)* |
| pl | def | gen | *fiskanna, fiskænna* | *fiskanna* |
| pl | def | dat | *fiskumin, fiskomin* | *fiskumin* |
| pl | def | ack | *fiskana, fiskæna* | *fiskana* |

### *4.1 Developing a Computational Morphology for Old Swedish*

For the morphological component of all our lexical resources we are using Functional Morphology (FM) [10,11] as our development framework. We chose this tool for a number of reasons: it provides a high-level description language (namely the modern functional programming language Haskell [12,18]), supporting modern programming constructs such as a strong type system and higher order functions; it supports tasks such as testing, (compound) analysis and synthesis; and, perhaps most importantly, it supports compilation to many standard formats, such as XML, JSON, CSV, LMF, SQL, LexC and XFST [2], GF [26], and full-form lexicons, and also provides facilities for the user to add new formats with little work.

The morphological model used in FM is *word and paradigm*, a term coined by the American structural linguist Charles Hockett [13]. A paradigm is a collection of words inflected in the same manner and is typically illustrated with an inflection table.

An FM lexicon consists of words annotated with paradigm identifiers from which the inflection engine of FM computes the full inflection tables.

Consider, for example, the citation form *fisker*, which is assigned the paradigm identifier *nn_m_fisker*. The paradigm identifier carries no meaning, it could just as well be any uniquely identifiable symbol, e.g. a number, but a mnemonic encoding is preferred. The encoding is read as: "This is a masculine noun inflected in the same way as the word *fisker*" (which is trivially true in this case). If the paradigm name and the citation form are supplied to the inflection engine, it will generate the information in table 1. To keep the presentation compact, we have contracted some word forms, i.e., the parenthesized letters are optional.

We also show (in the last column of table 1) how this paradigm is presented in traditional grammatical treatises of Old Swedish, e.g. those by Wessén [32] and Pettersson [25]. For a discussion of the differences between our paradigms and those found in traditional grammatical descriptions, see Sect. 4.1.1.

The starting point of the paradigmatic specification, besides the dictionaries themselves, are the standard grammars of Old Swedish mentioned above, i.e., those by Noreen [24], Wessén [31–33], and Pettersson [25]. The number of paradigms in the current description by part of speech are as follows:

| *Part of speech* | *# of paradigms* |
|---|---|
| Noun | 38 |
| Adjective | 6 |
| Numeral | 7 |
| Pronoun | 15 |
| Adverb | 3 |
| Verb | 6 |
| Total | 75 |

We will now give some technical details of the implementation by explaining how some of the verb paradigms in our morphology were defined. The main

objective is not to give a complete description, but rather to provide a taste of what is involved. The interested reader is referred to one of the FM papers.

An implementation of a new paradigm in FM involves: a type system; an inflection function for the paradigm; an interface function that connects the inflection function to the generic lexicon; and a paradigm name. Note that if the new paradigm is in a previously defined part of speech, then no new type system is required.

A paradigm in FM is represented as a function, where the input is one or more word forms (typically the citation form or principal parts) and a set of morphosyntactic encodings, and the output of the function is a set of inflected word forms computed from the input word forms. It is a set instead of a single word form to enable treatment of variants and missing cases, very important in the case of Old Swedish.

More concretely, if we represent the paradigm of regular nouns in English as a function, and only consider a morphosyntactic encoding for number, we would then define a function that expects a regular noun in nominative singular. If this function is given the word `"elephant"`, then the result would be another function. This function would, if an encoding for singular is given to the function, return `{"elephant"}`, and if an encoding for plural is given, return `{"elephants"}`. The resulting function may be translated into an inflection table given that the morphosyntactic encoding is ensured to be enumerable and finite (how this is ensured in FM will not be discussed here).

Turning now to the verb paradigms of Old Swedish, a `Verb` is a function from a morphosyntactic encoding, `VerbForm`, to a set of word forms with the abstract name `Str`.

```
type Verb = VerbForm -> Str
```

The type `VerbForm` defines the inflectional parameters of Old Swedish verbs. We only include those parameter combinations that actually exist, which will ensure, by type checking, that no spurious parameter combinations are created. A morphosyntactic encoding in FM is an algebraic data type, consisting of a list of constructors, where a constructor may have zero or more arguments. The vertical line should be interpreted as disjunction. The arguments here are also algebraic data types (only the definition of `Vox` is given here). A member of this type is, for example, `Inf Active`, where `Active` is a constructor of the type `Vox`.

```
data VerbForm =
 PresSg Modus Vox          |
 PresPl Person Modus Vox   |
 PretInd Number Person Vox |
 PretConjSg Vox            |
 PretConjPl Person Vox     |
 Inf Vox                   |
 ImperSg                   |
 ImperPl Person12

data Vox =
 Active |
 Passive
```

The `VerbForm` expands into 41 different parameter combinations. These parameter combinations may be given any string realization, i.e., we are not stuck with these rather artificially looking tags, we can choose any tag set. For example, instead of `PretConjSg Passive`, we have *pret konj sg pass*.

The next step is to define some inflection functions. We start with the paradigm of the first conjugation, exemplified by the word *ælska* 'to love'. The inflection function `aelska_rule` performs case analysis on the `VerbForm` type. There is one input word form, which will be associated with the variable `aelska`. The function `strs` translates a list of strings to the abstract type `Str`. The function is built up with the support of a set of helper functions, such as `passive` that computes the active and passive forms, `tk` that removes the nth last characters of a string, and `imperative_pl` that computes the plural imperative forms (inflected for person).

```
aelska_rule :: String -> Verb
aelska_rule aelska p =
 case p of
 PresSg Ind Act ->
  strs [aelska++"r",aelska]
 PresSg Ind Pass -> strs [aelska ++"s"]
 Inf v -> passive v [aelska]
 ImperSg -> strs [aelska]
 ImperPl per ->
  imperative_pl per aelsk
 PresPl per m v   ->
  indicative_pl (per,m,v) aelsk
 PretInd Pl per v ->
  pret_ind_pl (per,v) aelsk
 PretConjPl per v ->
  pret_conj_pl (per,v) (aelsk++"a")
 PresSg Conj v ->
  passive v [aelsk++"i",aelsk++"e"]
 PretInd Sg _ v -> passive v [aelska++"þi"]
 PretConjSg v ->
  passive v [aelska++"þi", aelska++"þe"]
 where aelsk = tk 1 aelska
```

The inflection function `aelska_rule` computes 65 word forms from one input word form, e.g. *kalla* 'to call'.

Given that we now have defined an inflection function for a verb paradigm, we can continue by defining the other paradigms in relation to this paradigm, i.e., we first give the parameter combinations that differ from `aelska_rule` and finalize the definition with a reference to `aelska_rule`. This is demonstrated in the inflection function `foera_rule`, the paradigm of the third conjugation.

```
foera_rule :: String -> Verb
foera_rule foera p =
 case p of
 PresSg Ind Act ->
  strs [foer++"ir", foer++"i"]
 PresSg Ind Pass -> strs [foer++"s"]
```

```
 Inf v -> passive v [foera]
 ImperSg -> strs [foer]
 PretInd Pl per v ->
  pret_ind_pl  (per,v) foer
 PretConjPl per v ->
  pret_conj_pl (per,v) foer
 PretInd Sg _ v -> passive v [foer++"þi"]
 PretConjSg v ->
  passive v [foer++"þi", foer++"þe"]
 _  -> aelska_rule foera p
 where foer = tk 1 foera
```

The last inflection function we present, representing the fourth conjugation paradigm, is `liva_rule`, defined in terms of `foera_rule`. Note that we use two different forms when referring to `foera_rule`: we use `lif++"a"`, i.e., the input form where "v" has been replaced with "f", for the past tense forms, and the input form for all other cases.

```
liva_rule :: String -> Verb
liva_rule liva p =
 case p of
 PresSg Ind Act  ->
  strs [liv++"er", liv++"ir", liv++"i"]
 PresSg Ind Pass -> strs [lif++"s"]
 ImperSg         -> strs [lif]
 p | is_pret p   -> foera_rule (lif++"a") p
 _               -> foera_rule liva p
 where liv = tk 1 liva
       lif = v_to_f liv
```

When the inflection functions are defined, we continue with the interface functions. An interface function translates one or more input words, via an inflection function, into an entry in the generic dictionary. This is done with the function `entry` that transforms an inflection function into an inflection table. If the current part of speech has any inherent parameters such as gender, those would be added here. The inherent parameters are not inflectional, they describe properties of a word, which is the reason why they appear at the entry level.

```
vb_aelska :: String -> Entry
vb_aelska = entry . aelska_rule

vb_foera :: String -> Entry
vb_foera = entry . foera_rule

vb_liva :: String -> Entry
vb_liva = entry . liva_rule
```

The interface functions need to be named to connect them with an external lexicon. This is done with the function `paradigm`. The names are typically the same as those of the interface functions. Every paradigm is also given a list of example word forms, which provides paradigm documentation and enables automatic generation of an example inflection table, which is done by applying the

current interface function to its example word forms. The list of paradigm names, denoted here with `commands`, is later plugged into the generic part of FM.

```
commands = [
  paradigm "vb_aelska" ["ælska"] vb_aelska,
  paradigm "vb_foera"  ["føra"]  vb_foera,
  paradigm "vb_liva"   ["liva"]  vb_liva
           ]
```

We can now start developing our lexicon. The lexicon consists of a list of words annotated with their respective paradigm, e.g. the word *røra* 'to touch' and *føra* 'to move', which are inflected according to the paradigm `vb_foera`. A lemgram identifier for each entry is also introduced here, *føra..vb.1* and *røra..vb.1*, identifying the individual inflection tables. The identifier is constructed in a mnemonic manner from the headword, the part of speech, and a disambiguating digit.

```
vb_foera "føra" {id("føra..vb.1")};
vb_foera "røra" {id("røra..vb.1")};
```

The lines above are put into an external file that is supplied to the compiled runtime system of FM.

### 4.1.1 Developing the Old Swedish Morphology Component

For the specification of the inflectional morphological component of the Old Swedish lexical resource, we have collaborated with a linguist who is an expert on orthographic and morphological variation in Old Swedish. In the first phase of the project, she defined the inflectional paradigms on the basis of the reference dictionaries and the actual variation empirically observed in texts.

The FM description was developed in parallel with this work. The linguist selected a set of sample words from the dictionaries and annotated those with the appropriate paradigms. The full inflection tables could then be generated immediately and the result evaluated by the linguist.

At the time of writing, about 3,000 lexical entries have been provided with inflectional information in the form of a paradigm identifier. The number of paradigms defined so far are 75 (see above).

In SALDO, described in Sect. 2.1, the number of inflectional classes (paradigms) turns out to be on an order of magnitude more, i.e., around 1,000 rather than around 100.[5] Note that this holds equally for the written standard language and colloquial spoken Swedish.[6] This is something that calls for an explanation, since under the (generally accepted, at least in some form) assumption of uniformitarianism [15],

[5] The distribution of inflectional patterns in the modern language is Zipfian in shape: Nearly half the paradigms are singletons, almost a fifth of them have only two members, etc.

[6] Although for slightly different reasons in the two cases: In the written standard language, it is generally the low-frequency words that have unique paradigms, e.g. learned words and loanwords. In the spoken language, high-frequency everyday words show variation in their inflectional behavior. There is some overlap, too, e.g. the strong verbs.

we would not expect to find less diversity in Old Swedish than in the modern language.

First, we may note that our morphological description is not yet complete. For example, while it covers all four weak verb conjugations, as yet it accounts for only two out of the nine or so classes of strong and irregular verbs. However, even standard grammars of Old Swedish like that by Wessén [32] list somewhere in the vicinity of 100 paradigms, and no more. We believe that the main factor here is our lack of information. For many lexical entries it is even difficult to assign an inflectional class, because the crucial forms are not attested in the extant texts, and of course, there are no native speakers on whose linguistic intuitions we could draw in order to settle the matter.

In the standard reference grammars of Old Swedish, inflectional paradigms are consistently idealized in the direction of a (re)constructed Old Swedish, arrived at on the basis of historical-comparative Indo-European and Germanic studies. In this connection, the actual variation seen in texts has been interpreted as a sign of language change, of 'exceptional' usages, etc.[7] [17]. In our paradigms, we have endeavored to capture the actual variation encountered in the texts and in the dictionary examples (but see Sect. 4.2).

## *4.2 The Computational Treatment of Variation in Old Swedish*

As mentioned above, the source of variation in Old Swedish texts are of three kinds: no standardized spelling; no standardized forms; and language change (diachronic drift). For our work on the computational morphological description of Old Swedish, we have found it natural and useful to make an additional distinction, namely that between *stem variation* and *ending variation*, since it has seemed to us from the outset that we need to treat stems and endings differently in this regard.

This gives us altogether six possible combinations of factors, as shown in the following table:

[7] and possibly even of carelessness or sloth on the part of the scribes; cf. the following quote, which well captures an attitude toward linguistic variation traditionally prevalent among linguists:

> Variation in Navajo pronunciation had long disturbed Haile (to Sapir, 30 March 1931: SWL): "Sometimes I do wish that the informants would be more careful in pronunciation and follow some system which would conform to theory. ... Apparently no excuse, excepting that informants are too lazy to use it correctly." Sapir responded (6 April 1931: SWL) that—at least in collecting texts—it was "not absolutely necessary to have the same words spelled in exactly the same way every time."
> [9, 257]

| | spelling variation | lack of lg standardization | language change |
|---|---|---|---|
| stem variation | $S_1$ | $S_2/L_1$ | $S_3/L_2$ |
| ending variation | $M_1$ | $M_2$ | $M_3$ |

(Legend: *S*=spelling rules; *L*=lexical component; *M*=morphological component)

The table reflects the fact that we have decided already to handle all inflectional ending variation – regardless of its origin – in the morphological component, i.e., our paradigms contain all attested ending variants, still a finite and in fact rather small set, which partly motivates their uniform treatment.

Representing a dead language with a finite corpus of texts, the Old Swedish stems could in theory be treated in the same way. The corpus is big enough, however (on the order of millions of words), that we will need to treat it as unlimited in practice, and hence the stems as a set that cannot be enumerated.

In order not to bite off more than we can chew, we have tentatively decided to treat all stem variation as a spelling problem (with one exception; see below). It will then be natural to look to some kind of solution involving edit distance or other string similarity measures.

However, the spelling is not completely anarchistic, far from it: For example, the /i:/ sound will be written <i>, <y>, <j>, <ii>, <ij>, and possibly in some other ways, but not, e.g., <a> or <m>, etc. Thus, a rule-based method may be more appropriate, or possibly a hybrid solution should be sought.

In the table above, the use of subscripts ($S_1$, $M_3$, etc.) hints at the possibility of distinguishing formally among different types of information even within a component. The present morphological description does not make a distinction between ending variation due to spelling variants of the 'same' ending (from a historical-normative point of view – e.g., indef sg dat *fiski/fiske* – and ending variation whereby 'different' endings occupy the same paradigmatic slot, e.g., indef sg dat *fiski/fisk*. However, there is no technical reason that we could not make this and other distinctions on the level of paradigms or even on the level of individual lexical entries. In fact, our work on the Old Swedish morphological description has clearly indicated the need for this kind of facility.[8]

There is one kind of stem variation which does not fit neatly into the picture painted so far, namely that brought about by inflectional morphological processes, in our case those of Ablaut and Umlaut. At the moment, the strong verb class paradigms do not account for variation in the realization of the Ablaut grades of the stem vowels – which of course we find in the texts – and we are still undecided as to how to treat them, by a separate normalization step or in the FM description. In the latter case we would then probably need to duplicate some information already present in the spelling rules component.

[8] It is not difficult to think of situations where this would be useful in modern language descriptions as well; for instance, it would be useful to be able to record the frequency of occurrence of homographs according to which lexical entry they represent.

### *4.3 Linking the Old Swedish Lexical Resource to SALDO*

The diachronic lexical resource under development in our research unit relies for its maximal usability on all resources being linked through SALDO. In the case of Dalin's 19th century dictionary such a linkage is straightforward, from the point of view of both language form and language content: The two language varieties involved are relatively close both in form and in semantics, which means that the interlinking can be largely automatic with a limited amount of manual post-editing.

Not so in the case of Old Swedish. This is simply a variety too far removed from Present-Day Swedish for this to be possible. There is possibly some headway to be gained by analyzing the entries of Old Swedish dictionaries, which are written in 19th century Swedish, using the 19th century resource, and, by transitivity, use the SALDO links appearing in that resource. This still remains to be tested, however.

The table below provides the information defined for the entry *aptanbakka* 'bat (animal)'. We have the headword, the SALDO identifier, the lemgram identifier, and the paradigm identifier. There is no linguistic connection between *aptanbakka* and *fladdermus* (the Present-Day Swedish word for 'bat') – the words are not related, historically or otherwise – and this is in no way exceptional. In other words, an attempt to use a rule-based approach to connect the headwords to their modern equivalents would probably not get us very far.

| *headword* | *saldo* | *lemgram* | *paradigm* |
|---|---|---|---|
| aptanbakka | fladdermus..1 | aptanbakka..nn.1 | nn_f_gata |

The linking poses many methodological questions, e.g., *bakvaþi* is an Old Swedish word meaning 'fatal accident resulting from a sword being struck backwards without the striker looking in that direction beforehand'. Naturally, there is no modern variant of this word, so the question is where to link. We could, e.g., link to 'sword strike', or to 'accident', or to both. Intuitively, linking to many may seem like a good idea, but we run the risk of poor precision.

Our attitude is that of the experimentalist, we just select one approach, try it in the wild, and refine according to the accumulated experience. We have, however, not reached the stage in the work on Old Swedish where we can do any extended experiments.

## 5 Summary and Conclusions

In the introduction we stated that access to our cultural heritage to a large extent equals access to old texts. With increasing digitization of historical texts, the need for language tools for accessing such text collections has become obvious. A central component of any language processing application is a rich lexical resource providing morphological, syntactic, semantic, pragmatic and in the best case also

encyclopedic knowledge about preferably upwards of a hundred thousand lexical units (single words and multi-word units).

We have made good headway toward our goal of creating such lexical resources for several historical varieties of Swedish: Present-Day Swedish, Late Modern Swedish, and Old Swedish. Our intention is to make these resources into one diachronic computational lexical resource. This will be accomplished by interlinking all resources using the persistent lexical sense identifiers of SALDO. We are partly there with the 19th century Late Modern Swedish lexical resource, and we have already been able to utilize the linked semantic information in a practical text processing application. The Old Swedish case is considerably more complex, however, and further research is needed.

The resources are permantly available in their most current state, together with the other Swedish FrameNet++ resources,[9] where it is also possible to search all the resources using the diachronic morphologies. The search interface collects all SALDO identifiers associated to the input word form, in any of the morphologies, and renders the lexical information connected to these SALDO identifiers. The resources are furthermore available via an open web service, which the search interface is built upon.

We are constantly on the lookout for additional lexical resources, and negotiations are currently ongoing about adding a digital version of an early 18th century dictionary – *Swensk Ordabok* by Jesper Swedberg [14] – to our lexical resources. At present we have no lexical resource for Early Modern Swedish, and Swedberg's dictionary would bridge this gap in our emerging diachronic computational lexical resource for 800 years of Swedish.

**Acknowledgements** The Old Swedish morphological description on which our computational morphology is based was made by Rakel Johnson, Department of Swedish Language, University of Gothenburg.

The work presented here was financed in part by Swedish Research Council grant 2007-7430 (2008–2010) awarded to Lars Borin for the infrastructure project *Safeguarding the future of Språkbanken*, and by Swedish Research Council grant 2005-4211 (2006–2008) awarded to Aarne Ranta, Chalmers University of Technology, for the research project *Library-Based Grammar Engineering*, and in part by the Faculty of Arts, University of Gothenburg, through its support to Språkbanken (the Swedish Language Bank).

We would also like to acknowledge CLT – the Centre for Language Technology, Göteborg http://www.clt.gu.se – for providing a creative atmosphere in which multidisciplinary collaborations such as this come naturally.

## References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proceedings of the 17th international conference on Computational linguistics, pp. 86–90. Morristown, NJ, USA (1998)

[9] See http://spraakbanken.gu.se/eng/sblex (updated daily)

2. Beesley, K.R., Karttunen, L.: Finite State Morphology. CSLI Publications, Stanford University, United States, (2003)
3. Borin, L.: Mannen är faderns mormor: *Svenskt associationslexikon* reinkarnerat. LexicoNordica **12**, 39–54 (2005)
4. Borin, L., Forsberg, M.: Saldo 1.0 (svenskt associationslexikon version 2). (2008). Språkbanken, Göteborgs universitet
5. Borin, L., Forsberg, M.: All in the family: A comparison of SALDO and WordNet. In: Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources—between Lexical Semantics, Lexicography, Terminology and Formal Ontologies. Odense (2009)
6. Borin, L., Forsberg, M.: From the people's synonym dictionary to fuzzy synsets - first steps. Proceedings of the LREC 2010 workshop Semantic relations. Theory and Applications (2010)
7. Borin, L., Forsberg, M., Lönngren, L.: The hunting of the BLARK – SALDO, a freely available lexical database for Swedish language technology. In: J. Nivre, M. Dahllöf, B. Megyesi (eds.) Resourceful language technology. Festschrift in honor of Anna Sågvall Hein, no. 7 in Acta Universitatis Upsaliensis: Studia Linguistica Upsaliensia, pp. 21–32. Uppsala University, Department of Linguistics and Philology, Uppsala (2008)
8. Dalin, A.F.: Ordbok öfver svenska språket. Vol. I–II. Stockholm (1853–1855)
9. Darnell, R.: Edward Sapir: Linguist, Anthropologist, Humanist. University of California Press, Berkeley / Los Angeles / London (1990)
10. Forsberg, M.: Three tools for language processing: BNF converter, Functional Morphology, and Extract. Ph.D. thesis, Göteborg University and Chalmers University of Technology (2007)
11. Forsberg, M., Ranta, A.: Functional morphology. In: ICFP'04. Proceedings of the ninth ACM SIGPLAN international conference of functional programming. ACM, Snowbird, Utah (2004)
12. Haskell: Haskell homepage. www.haskell.org (2010)
13. Hockett, C.: Two models of grammatical description. Word **10**, 210–234 (1954)
14. Holm, L. (ed.): Jesper Swedberg: Swensk Ordabok. Skara stiftshistoriska sällskaps skriftserie. Stiftelsen för utgivande av Skaramissalet, Skara (2009)
15. Janda, R.D., Joseph, B.D.: On language, change, and language change – or on history, linguistics, and historical linguistics. In: B.D. Joseph, R.D. Janda (eds.) Handbook of Historical Linguistics, pp. 3–180. Blackwell, Oxford (2003)
16. Järborg, J.: Roller i Semantisk databas. Tech. Rep. GU-ISS-01-3, Department of Swedish Language, University of Gothenburg (2001)
17. Johnson, R.: Skrivaren och språket. Ph.D. thesis, Department of Swedish Language, Göteborg University (2003)
18. Jones, S.P.: Haskell 98 Language and Libraries: The Revised Report. Cambridge University Press (2003)
19. Lene Antonsen, T.T., Wiechetek, L.: Reusing grammatical resources for new languages. In: Proceedings of LREC'10. ELRA, Valletta (2010)
20. Lönngren, L.: Svenskt associationslexikon: Rapport från ett projekt inom datorstödd lexikografi. Centrum för datorlingvistik. Uppsala universitet (1989). UCDL-R-89-1
21. Lönngren, L.: Svenskt associationslexikon. Del I-IV. Institutionen för lingvistik. Uppsala universitet (1992)
22. Lönngren, L.: A Swedish associative thesaurus. In: Euralex '98 proceedings, Vol. 2, pp. 467–474 (1998)
23. Morris, J., Hirst, G.: Non-classical lexical semantic relations. In: D. Moldovan, R. Girju (eds.) HLT-NAACL 2004: Workshop on Computational Lexical Semantics, pp. 46–51. ACL, Boston (2004)
24. Noreen, A.: Altschwedische Grammatik. Halle (1904)
25. Pettersson, G.: Svenska språket under sjuhundra år. Studentlitteratur, Lund, Sweden (2005)
26. Ranta, A.: Grammatical Framework: A Type-theoretical Grammar Formalism. The Journal of Functional Programming **14(2)**, 145–189 (2004)
27. Schlyter, C.: Ordbok till Samlingen af Sweriges Gamla Lagar. (Saml. af Sweriges Gamla Lagar 13). Lund, Sweden (1887)

28. SO: Svensk ordbok. Esselte Studium, Stockholm (1986)
29. Söderwall, K.F.: Ordbok Öfver svenska medeltids-språket. Vol I–III. Lund, Sweden (1884)
30. Söderwall, K.F.: Ordbok Öfver svenska medeltids-språket. Supplement. Vol IV–V. Lund, Sweden (1953)
31. Wessén, E.: Svensk språkhistoria: Grundlinjer till en historisk syntax. Stockholm, Sweden (1965)
32. Wessén, E.: Svensk språkhistoria: Ljudlära och ordböjningslära. Stockholm, Sweden (1969)
33. Wessén, E.: Svensk språkhistoria: Ordböjningslära. Stockholm, Sweden (1971)

# Morphosyntactic Tagging of Old Icelandic Texts and Its Use in Studying Syntactic Variation and Change

Eiríkur Rögnvaldsson and Sigrún Helgadóttir

**Abstract** We describe experiments with morphosyntactic tagging of Old Icelandic (Old Norse) narrative texts using different tagging models for the TnT tagger [3] and a tagset of almost 700 tags, originally developed for Modern Icelandic. It is shown that by using a model that has been trained on both Old and Modern Icelandic texts, we can get 92.7% tagging accuracy which is considerably better than the 90.4% that have been reported for Modern Icelandic. Although our tagging is morphological in nature, the tags carry a substantial amount of syntactic information and the tagging is detailed enough for the syntactic function of words to be more or less deduced from their morphology and the adjacent words. We show that the morphosyntactic tags can be very useful in locating certain syntactic constructions and features in a large corpus of Old Icelandic narrative texts. We demonstrate this by searching for—and finding—previously undiscovered examples of a number of syntactic constructions in the corpus. We conclude that in a highly inflectional language, a morphologically tagged corpus can be an important tool in studying syntactic variation and change, in the absence of a fully parsed corpus which of course gives more possibilities.

**Key words:** morphosyntactic tagging, bootstrap training, Old Icelandic, syntactic variation

## 1 Introduction

Until recently, no part-of-speech tagger was available for Icelandic. In a previous project [11, 12], we trained the TnT tagger written by Brants (cf. [3]) on a corpus that was created in the making of the Icelandic Frequency Dictionary (*Íslensk*

Eiríkur Rögnvaldsson
University of Iceland, Árnagarði við Suðurgötu, IS-101 Reykjavík, e-mail: eirikur@hi.is

Sigrún Helgadóttir
Árni Magnússon Institute, Neshaga 16, IS-107 Reykjavík, e-mail: sigruhel@hi.is

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_4, 

*orðtíðnibók*, henceforth IFD; [27]). The IFD corpus is considered to be a carefully balanced corpus consisting of 590,297 tokens with 59,358 types—both figures including punctuation.

The corpus contains 100 fragments of texts, approximately 5,000 tokens each. All the texts were published for the first time in 1980–1989. Five categories of texts were considered, i.e. Icelandic fiction, translated fiction, biographies and memoirs, non-fiction (evenly divided between science and humanities) and books for children and youngsters (original Icelandic and translations).

The texts were pre-tagged using a specially designed computer program and the tagging was then carefully checked and corrected manually. Thus, this corpus is ideal as training material for data-driven statistical taggers, such as the TnT tagger.

In the present project, we applied the TnT tagger trained on the Modern Icelandic corpus to Old Icelandic (Old Norse) texts.[1] This paper describes the results of this experiment, and also describes our experiments with using the tagged Old Icelandic corpus to search for syntactic constructions. We conclude that in an inflectional language like Old Icelandic, a morphosyntactically tagged corpus like this can be an important tool in studying syntactic variation and change.

## 2 Tagging Modern Icelandic

In this section, we describe the tagset used in our research, and give a brief overview of our experience with the training of the TnT tagger on Modern Icelandic texts.

### *2.1 The Tagset*

The tagset developed for the IFD corpus is very large, compared to tagsets designed for English at least, such as the Penn Treebank tagset [23]. The size of the tagset of course reflects the inflectional character of Icelandic, since it is for the most part based on the traditional Icelandic analysis of the parts of speech and grammatical categories, with some exceptions where that classification has been rationalized.

In the tag strings, each character corresponds to a single morphosyntactic category. The first character always marks the part of speech. Thus, the sentence *Hún hefur mætt gamla manninum* 'She has met the old man' will be tagged as shown in Table 1. The first column shows the word being tagged, the second column shows the tag, and the third column shows the meaning of each individual character in the tag string.

[1] It is customary to use the term 'Old Norse' for the language spoken in Norway, Iceland and the Faroe Islands up to the middle of the 14th century. The overwhelming majority of existing texts written in this language is either of Icelandic origin or only preserved in Icelandic manuscripts. For the purposes of this paper, 'Old Icelandic' is thus synonymous with 'Old Norse'.

**Table 1** A tagging example

| | | |
|---|---|---|
| Hún | fpven | pronoun (f)–personal (p)–feminine (v)–singular (e)–nominative (n) |
| hefur | sfg3eþ | verb (s)–indicative (f)–active (g)–3rd person (3)–singular (e)–past (þ) |
| mætt | ssg | verb (s)–supine (s)–active (g) |
| gamla | lkeþvf | adjective (l)–masculine (k)–singular (e)–dative (þ)–definite (v)–positive (f) |
| manninum | nkeþg | noun (n)–masculine (k)–singular (e)–dative (þ)–suffixed article (g) |

Of the word forms in the IFD corpus, 15.9% are ambiguous as to the tagset within the IFD. This figure is quite high, at least compared to English, which reflects the fact that the inflectional morphology of Icelandic is considerably more complex than English. Icelandic nouns can have up to 16 grammatical forms or tags, verbs up to 106 different tags, and adjectives up to 120 tags. Altogether, 639 different tags occur in the IFD corpus, but the total sum of possible tags is around 700.

Some of the ambiguity is due to the fact that inflectional endings in Icelandic have many roles, the same ending often appearing in many places (e.g. *-a* in *penna* for all oblique cases in the singular (acc., dat., gen.), and accusative and genitive in the plural of the masculine noun *penni* 'pen', producing 5 different tags for one form of the same word). The most ambiguous of word forms in the IFD, *minni*, has 24 tags in the corpus, and has not exhausted its possibilities [2].[2]

## *2.2 Training the Tagger*

The computer files for the IFD corpus each contain one text excerpt. Each file was divided into ten approximately equal parts. From these, ten different disjoint pairs of files were created. In each pair there is a training set containing about 90% of the tokens from the corpus and a test set containing about 10% of the tokens from the corpus. Each set should therefore contain a representative sample from all genres in the corpus. The test sets are independent of each other whereas the training sets overlap and share about 80% of the examples. All words in the texts except proper nouns start with a lower case letter.

Results for ten-fold cross-validation testing for the TnT tagger on the IFD corpus are shown in Table 2 (cf. [11, 12]). It is worth noticing that these results show lower performance rates when the tagger is applied to the Icelandic corpus than is achieved for example for Swedish as reported in [25]. In that study, TnT was applied to and tested on the SUC corpus with 139 tags compared to the Icelandic tagset of almost 700 tags. Performance rates are also considerably lower than have been reported for taggers trained on the Penn treebank.[3]

[2] *minni* can be a noun meaning 'memory', present tense of the verb *minna* 'remind', comparative of the (irregular) adjective *lítill* 'small'. In all of these words we find extensive syncretism, resulting in many different tag strings for this word form in each part of speech.

[3] Since 2005, Hrafn Loftsson has been developing a rule-based tagger for Icelandic, IceTagger. The highest accuracy that has been reached using this tagger is 91.6% [22]. By correcting some errors

**Table 2** Mean tagging accuracy for all words, known words and unknown words for TnT

| Type | Accuracy % |
|---|---|
| All words | 90.4 |
| Known words | 91.7 |
| Unknown words | 71.6 |

Table 2 shows results for known words, unknown words and all words. Mean percentage of unknown words in the ten test sets was 6.84. This is similar to what was seen in the experiment on Swedish text [25] and indicates that the major difficulty in annotating Icelandic words stems from the difficulty in finding the correct tag for unknown words. Words belonging to the open word classes (nouns, adjectives and verbs) account for about 96% of unknown words in the test sets whereas words in these word classes account for just over 51% of all words in the test sets.

## 3 Tagging Old Icelandic Texts

Having trained the TnT tagger on Modern Icelandic texts, we wanted to find out whether the tagger could be of help in tagging Old Icelandic narrative texts, with the purpose of facilitating the use of these texts in research on syntactic variation and change. To create a manually annotated training corpus for Old Icelandic from scratch would have been a very time-consuming task. Thus, the possibility of using the bootstrapping method that we describe in this section was a key factor in realizing this project.

Bootstrapping is of course a common approach in training taggers and parsers. To our knowledge, however, this approach has not been used in historical linguistics to develop tagging models for a different stage of language than the tagger was originally trained on. Our method somewhat resembles the experiments of Hwa et al. [15], who used parallel texts to build training corpora by projecting syntactic relations from English to languages for which no parsed corpora were available. The training corpora created using this method were then in turn used to develop stochastic parsers for the languages in question. The whole process took only a small fragment of the time it would have taken to create a manually corrected corpus to train the parsers.

The common factor in our project and the work reported in [15] is the use of another language, or (in our case) another stage of the same language, as a starting point in the bootstrapping process. Our experiments with bootstrapping the tagging of Old Icelandic texts are described in this section.

in the IFD corpus and combining IceTagger with a number of data-driven taggers that had been trained on the corpus, the accuracy reached 92.5%, and by reducing the tagset somewhat, without losing any important semantic distinctions, the accuracy reached 93.63%.

### 3.1 Old vs. Modern Icelandic

At a first glance, it may seem unlikely that a tagger trained on 20th century language could be applied to 600–700 years old texts. However, Icelandic is often claimed to have undergone relatively small changes from the oldest written sources up to the present. The sound system, especially the vowel system, has changed dramatically, but these changes have not led to radical reduction or simplification of the system and hence they have not affected the inflectional system, which has not changed in any relevant respects. Thus, the tagset developed for Modern Icelandic can be applied to Old Icelandic without any modifications.

The vocabulary has also been rather stable. Of course, a great number of new words (loanwords, derived words and compounds) have entered the language, but the majority of the Old Icelandic vocabulary is still in use in Modern Icelandic, even though many words are confined to more formal styles and may have an archaic flavor.

On the other hand, many features of the syntax have changed (cf. [8, 31]). These changes involve for instance word order, especially within the verb phrase, the use of phonologically "empty" NPs in subject (and object) position, the introduction of the expletive *það* 'it, there', the development of new modal constructions such as *vera að* 'be in the process of' and *vera búinn að* 'have done/ finished', etc.

In spite of these changes, we found it worthwhile to try to adapt the tagging model that we had trained for Modern Icelandic to our Old Icelandic electronic corpus. Our motive was not to get a 100% correct tagging of the Old Icelandic texts, but rather to facilitate the use of the texts in syntactic research, cf. Sect. 4 below.

### 3.2 The Old Icelandic Corpus

Our Old Icelandic corpus consists of a number of narrative prose texts (sagas), which are assumed to have been written in the 13th and 14th centuries—a few of them probably later. Among these are many of the most famous Old Icelandic sagas. The division of the corpus is shown in Table 3:

**Table 3** The Old Icelandic Corpus

| Text | Tokens |
|---|---|
| Family Sagas (around 40 sagas) (Íslendingasögur) | 1,074,731 |
| Sturlunga Saga ("Contemporary Sagas") | 283,002 |
| Heimskringla (Sagas of the Kings of Norway) | 250,920 |
| The Book of Settlement (Landnámabók) | 42,745 |
| Total | 1,651,398 |

The texts we use are (with the exception of The Book of Settlement) taken from editions which were published between 1985 and 1991 [9, 16, 17]. In these

editions, the text has been normalized to Modern Icelandic spelling. This involves, for instance, reducing the number of vowel symbols ('æ' is used for both 'ae ligature' (æ) and 'oe ligature' (œ), 'ö' is used for both 'o with a slash' (ø) and 'o with a hook'), inserting *u* between a consonant and a word-final *r* (*maðr* 'man' > *maður*), shortening word-final *ss* and *rr* (*íss* 'ice' > *ís*, *herr* 'army' > *her*), changing word-final *t* and *k* in unstressed syllables to *ð* and *g*, respectively (*þat* 'it' > *það*, *ok* 'and' > *og*), etc. Furthermore, a few inflectional endings are changed to the Modern Icelandic form.

It must be emphasized, however, that these changes do not in any way simplify the inflectional system or lead to the loss of morphological distinctions in the texts. Thus, the texts are just as good as sources of syntactic evidence as texts that are published in the normalized Old Icelandic spelling.

On the other hand, we must point out that the original versions of these texts do not exist; the texts are mostly preserved in vellum manuscripts from the 13th through the 15th centuries, but some of them only exist in paper manuscripts from the 16th and 17th centuries. This makes it extremely difficult to assess the validity of these texts as linguistic evidence, since it is often impossible to know whether a certain feature of the preserved text stems from the original or from the scribe of the preserved copy, or perhaps from the scribe of an intermediate link between the original and the preserved manuscript.

It is well known that scribes often did not retain the spelling of the original when they made copies; instead, they used the spelling that they were used to. In many cases, two or more manuscripts of the same text are preserved, and usually they differ to a greater or lesser extent. Furthermore, it is known that not all of the editions that our electronic texts are based on are sufficiently accurate (cf., for instance, [5]).

Even though this may to some extent undermine the validity of the texts as sources of syntactic evidence, it does not directly concern the main subject of this paper, which is to show that we can use a tagging model developed for Modern Icelandic to assist us in making the Old Icelandic corpus a usable tool in studies of syntactic variation and change. There is no reason to believe that possible inaccuracies and errors in the texts—cases where they fail to mirror correctly the syntax of the manuscripts—have any effects on the tagging accuracy. That is, the use of more accurate editions would not lead to less accurate tagging.

### *3.3 Training the Tagger on the Old Icelandic Corpus*

We started by running TnT on the whole Old Icelandic corpus using the tagging model developed for Modern Icelandic [11, 12]. We then measured the accuracy by taking four randomly chosen samples of 1,000 words each from different texts in the corpus—one from the Family Sagas, one from Heimskringla, and two from Sturlunga Saga—and checking them manually. Counting the correct tags in these samples gave 88.0% correct tags, compared to 90.4% for Modern Icelandic.

Even though these results were worse than those we got for Modern Icelandic, we considered them surprisingly good. The syntax of Old Icelandic differs from Modern Icelandic syntax in many ways, as mentioned above, and one would especially expect the differences in word order to greatly affect the performance of a trigram based tagger like TnT. However, sentences in the Old Icelandic corpus are often rather short, which may make them easier to analyze than the longer sentences of Modern Icelandic.

We then selected seven whole texts (sagas) and two fragments from the Sturlunga collection for manual correction—around 95,000 words in all. This amounts to one third of the Sturlunga collection. The manual correction was a time-consuming task, but the time and effort spent on checking and correcting the output of TnT was only a small fragment of the time and effort it would have taken to tag the raw text.

We trained TnT on the corrected text (95,000 words), tagged the whole corpus again with the resulting model, and measured the accuracy on the same four samples of 1,000 words each as in the first experiment. Now the results were much better—91.7% correct tags, which is better than the 90.4% accuracy that we got for Modern Icelandic. It may seem surprising how much the accuracy improved when we used this model, especially when we consider that the training corpus was much smaller than the training corpus for Modern Icelandic (95,000 words compared to more than 500,000). On a closer look, however, this is understandable.

First, many of the errors occurring in the first experiment could be predicted and were easy to correct. For instance, the word *er* was always classified as a verb in the third (or first) person singular present indicative ('is, am'), as it usually is in Modern Icelandic. In Old Icelandic, however, this word is very often a temporal conjunction ('when') or a relative particle ('that, which'). When the tagger was trained on a corrected Old Icelandic text, it could quickly and easily learn the correct tagging of these words, due to their frequency.

Second, it is well known that tagging accuracy is usually very much lower for unknown words than for known words, and the number of unknown words was much lower in the second experiment. In the first experiment, using the model for Modern Icelandic, the unknown word rate was 14.6%, reflecting the fact that a number of Old Icelandic words are rare or do not occur in Modern Icelandic. In the second experiment, using the model for Old Icelandic, the unknown word rate dropped to 9.6%, even though the training corpus was much smaller as pointed out above. This reflects the relatively small vocabulary of the Old Icelandic texts, which in turn reflects the narrow universe that the texts describe (cf. also [28]).

Finally, we trained TnT on a union of the corrected Old Icelandic texts and the Modern Icelandic texts. Thus, the training set for the final experiment consists of around 500,000 words from Modern Icelandic texts plus 95,000 words from Old Icelandic texts. When we tagged the Old Icelandic corpus using this model, we got 92.7% accuracy for the same four samples as in the first two experiments. The results of the three experiments are shown in Table 4.

It is possible to improve the results by tagging the texts using all three models and combining the results of different models in various ways. All three models agree

**Table 4** Tagging accuracy for Old Icelandic texts using three different tagging models

| Tagging model | Accuracy % |
|---|---|
| Modern Icelandic model | 88.0 |
| Old Icelandic model | 91.7 |
| MI + OI model | 92.7 |

on the tags for 84.6% of the words. In 80.9% of the cases, they agree on the correct tag, but for 3.7% of the words, all three models agree on a wrong tag.

For 15.4% of the words, the models disagree. In most cases, two of them assign the same tag and the third model assigns a different tag. In a few cases, each model assigns a separate tag. Thus, if we assume that the tag is correct when all three models agree, we only need to look at 15.4% of the whole corpus. This means that the highest possible accuracy to be obtained using this method is 96.3%, since all models agree on a wrong tag in the remaining cases as pointed out above.

We could also choose to disregard the model that is trained only on Modern Icelandic texts, since it gives much lower accuracy than the other two models. The remaining models agree on the tagging of 93.5% of the words—incorrectly for 4.3% of the words. If we only look at the 6.5% where the models disagree, we are down to around 107,000 words that we have to correct manually. This is a manageable task. We think that performance may exceed 95% after manual revision of the training set, assuming that about half of the disagreements can be correctly resolved. This is an acceptable result in our view, and should be sufficient for most uses of the corpus.

In this connection, it must be pointed out that a majority of the tagging errors only involve one morphosyntactic feature. Thus, nouns are often tagged as accusative instead of dative, or vice versa, whereas gender and number are correctly tagged; verbs are often tagged as 3rd person instead of 1st person, whereas mood, voice, number, and tense are correctly tagged; etc. This means that by using fuzzy search, we should in many cases be able to find what we are looking for, even if the words are not quite correctly tagged.

## 4 Tagged Texts in Syntactic Research

Over the past two decades, interest in historical syntax has grown substantially among linguists. Accompanied by the growing amount of electronically available texts, this has led to the desire for—and possibility of—creating syntactically parsed corpora of historical texts, which could be used to facilitate search for examples of certain syntactic features and constructions. A few such corpora have been built, the most notable being the Penn Parsed Corpora of Historical English, developed by Anthony Kroch and his associates [18, 19]. These corpora have already proven their usefulness in a number of studies of older stages of English (cf., e.g., [20, 21]).

We wanted to know whether our tagged Old Icelandic corpus could be used in syntactic research in a similar manner as syntactically parsed corpora. We had been

using the raw unannotated texts for this purpose (cf., for instance, [29, 30]) but the search for certain syntactic constructions and features had proven to be cumbersome and give insufficient results. Although our tagging is morphological in nature, the tags carry a substantial amount of syntactic information and the tagging is detailed enough for the syntactic function of words to be more or less deduced from their morphology and the adjacent words.

Thus, for instance, a noun in the nominative case can reasonably safely be assumed to be a subject, unless it is preceded by the copula *vera* 'to be' which is in turn preceded by another noun in the nominative, in which case the second noun is a predicative complement. A noun in the accusative or dative case can in most instances be assumed to be a (direct or indirect) object, unless it is immediately preceded by a preposition (cf. also [32]). As is well known, Modern Icelandic also has accusative and dative subjects, and even some nominative objects [35], but these can easily be identified from their accompanying verbs.

As a search interface for the corpus, we use the Glossa system from the Text Laboratory at the University of Oslo [26], which in turn builds on the IMS Corpus Workbench [4]. The system has been adapted to the IFD tagset by Guðmundur Örn Leifsson and Sigrún Helgadóttir at the Árni Magnússon Institute. The search options in Glossa are very flexible. It is possible to search for word forms, lemmas, and morphosyntactic categories, and these search options can be combined in various ways to create complex queries.

To test the usefulness of the tagging of Old Icelandic texts in syntactic research, we have made a small study of two controversial and disputed features of Old Icelandic syntax; Object Shift and Passive. These studies are described in the remainder of this Sect. The tagged Old Icelandic texts are available online at http://mim.hi.is where they can be searched using the Glossa system.

## *4.1 Object Shift*

As originally described by Holmberg [13], Object Shift is the process of moving a (direct or indirect) object to the left across a negation. In Modern Icelandic, this process applies both to pronouns and full NPs (or DPs), as shown in (1), whereas in the "Mainland" Scandinavian languages (Danish, Norwegian, and Swedish), it only applies to pronouns, as the Danish sentences in (2) show (examples from [35]). The "shifted" object is italicized whereas the negation is in boldface and the "place of origin" of the shifted object is shown by a dash:

(1) a. Nemandinn las *bókina* **ekki** —
the-student read book not —
'The student didn't read the book'

b. Nemandinn las *hana* **ekki** —
the-student read she not —
'The student didn't read it'

(2) a. *Studenten læste *bogen* **ikke** —
the-student read book not —
'The student didn't read the book'
b. Studenten læste *den* **ikke** —
the-student read she not —
'The student didn't read it'

It has been suggested that this difference between Icelandic and the Mainland Scandinavian languages is somehow related to the fact that Icelandic has a much richer case morphology than the Mainland Scandinavian languages (cf. [14]). If this were so, one would expect to find both types of Object Shift in Old Icelandic, since the case system of Icelandic is in all relevant respects the same as in Old Icelandic. The Mainland Scandinavian languages would then be assumed to have lost Object Shift of full DPs due to the loss of case inflections.

However, it has been claimed that Object Shift of full DPs does not occur in Old Icelandic. Mason [24] claims to have found two examples of shifted full DP objects in his study of nine Old Icelandic sagas. Sundquist [34], on the other hand, claims "that these two examples do not provide evidence for a full DP Object Shift like in modern Icelandic". Haugan [10] did not find any examples of full DP Object Shift in his study of Old Icelandic, and neither did Sundquist [34] in a study of Middle Norwegian. Thus, Sundquist concludes that "full DP Object Shift is not an option in earlier stages of Mainland Scandinavian".

It is therefore of considerable theoretical interest to search for examples of full DP Object Shift in Old Icelandic texts. However, this is a tedious and time-consuming task. Even though this is a perfectly grammatical construction in Modern Icelandic, it appears to be very rare in texts. Thus, one can read dozens or even hundreds of pages without finding a single example. When the constructions that we are looking for are that rare, it is easy to overlook the few examples that actually occur in the texts that we read. Given the rarity of full DP Object Shift in Modern Icelandic, one may wonder whether those who have studied Object Shift in Old Icelandic have looked at a large enough corpus.

We have searched for examples of full DP Object Shift in our morphosyntactically tagged Old Icelandic corpus. We search for a verb in the indicative or the subjunctive, followed by a noun, an adjective, or a demonstrative pronoun in an oblique case, followed by a negation (one of the words *eigi*, *ei*, *ekki* 'not', *aldrei*, *aldregi* 'never'). We allow for up to two words between the noun/adjective/demonstrative pronoun and the negation. Thus, in addition to simple sentences with a noun immediately following the verb and preceding the negation, we will find sentences where both a demonstrative pronoun and an adjective precedes the noun, and sentences where a prepositional phrase consisting of a preposition and a noun follows the head noun.

Of course, we will neither get 100% precision nor 100% recall by using this pattern. It will miss some potential examples of Object Shift; for instance, sentences with an adverb modifying a prenominal adjective when a demonstrative pronoun is also present, or sentences with an adjective modifying an object of a preposition, which

follows the head noun. Furthermore, this search pattern will return a number of sentences that are not instances of Object Shift.
When we run this search pattern on the Old Icelandic corpus, it returns 245 examples. The majority of these examples are not real examples of Object Shift but it doesn't take long to clean the search results and throw these sentences away. When we have finished this cleaning, it appears that we really are left with some genuine examples of full DP Object Shift:

(3) a. Nú leita þeir um skóginn og finna *Gísla* **eigi** —
now search they about the-forest and find Gisli not —
'Now they search through the forest and don't find Gisli'
b. er hann dræpi *Þórð* **eigi** — og förunauta hans
when he killed Thord not — and companions his
'if he didn't kill Thord and his companions'
c. og fundu *Þórð* **eigi** — sem von var að
and found Thord not — as expectance was at
'and not surprisingly, they didn't find Thord'

Using this method, we found at least 9 indisputable examples of full DP Object Shift. This may not be the exact number of such sentences in our corpus. First, in addition to these examples, there are some borderline cases, which may or may not be interpreted as instances of Object Shift. Second, our searching method does not guarantee 100% recall, as explained above. However, this doesn't really matter for our purposes. We have shown conclusively that full DP Object Shift existed in Old Icelandic, contrary to what has previously been claimed in the literature; and we have demonstrated the efficiency of our searching method.

## *4.2 Passive*

Another controversial feature of Old Icelandic syntax is the nature of the passive. It has sometimes been claimed [6, 7] that all passive sentences in Old Icelandic are lexical but not derived by NP-movement (or chain-formation). This claim has been disputed [1], and it has been claimed that the existence of agentive prepositional phrases (by-phrases) would be an argument against this analysis, since such phrases presuppose a derivational analysis of passive sentences [29].
Be that as it may, it is quite clear that agentive prepositional phrases in passives are rather rare in Modern Icelandic, and hence, one would not expect to find many of them in Old Icelandic. Faarlund [8], for instance, quotes two such examples but concludes: "This is very rarely found, however."
It is very difficult to search for such examples in an unannotated electronic text. However, once we have a morphosyntactically tagged text, it is relatively easy to search for agentive prepositional phrases. We can search for a past participle, followed by *af*, followed by a nominal (noun, pronoun, adjective) in the dative. Since the distinction between past participle forms and adjectives in the neuter singular is

not always clear, and the tagger makes a number of errors in this classification, we also search for adjectives in addition to past participles.
This search returns some 130 sentences. Most of them are not instances of agentive phrases, since the preposition *af* can also have other functions. Nevertheless, we have found at least 15 sentences with agentive prepositional phrases, only a few of which have previously been quoted in the literature on this subject. Three of these sentences are shown below—the agentive phrases in boldface:

(4) a. að Þorvarður ... hafi skírður verið **af Friðreki biskupi**
that Thorvard ... has baptized been by Fridrek bishop
'that Thorvard has been baptized by bishop Fridrek'
b. Og er þetta mál var rannsakað **af lögmönnum**
and when this case was investigated by lawyers
'and when lawyers investigated this case'
c. Óttar gerði sem honum var boðið **af Sighvati**
Ottar did as him was ordered by Sighvat
'Ottar did what Sighvat ordered him'

Thus, our searching method has enabled us to strengthen the evidence for the existence of derivational passive in Old Icelandic.

## 5 Conclusion

In this paper, we have demonstrated that it is possible to use a tagging model trained on Modern Icelandic texts to facilitate tagging of Old Icelandic narrative texts. By using this method, we are able to tag a large corpus of Old Icelandic with acceptable accuracy in a relatively short time—only a fragment of the time it would have taken to build a tagging model for Old Icelandic from scratch. Furthermore, we have shown that a corpus that has been tagged using a rich tagset based on morphosyntactic features can fruitfully be used in the search for a number of syntactic constructions, and hence is a valuable tool in studying syntactic variation and change.
Of course, a morphologically tagged corpus like the one we have built doesn't amount to a fully parsed corpus. Therefore, we are now in the process of building a million word historical treebank of Icelandic (Icelandic Parsed Historical Corpus, [33, 36]). This treebank will enable us to search for various complex constructions that cannot possibly be searched for using the methods described in this paper. However, given the tremendous effort it takes to build a large parsed corpus, we think our method is an alternative that must be taken seriously.

**Acknowledgements** This project was partly supported by grants from the University of Iceland Research Fund to the projects "The syntactic use of Old Icelandic POS tagged texts" and "Icelandic Diachronic Treebank" and by a grant from the Icelandic Research Fund to the project "Viable

Language Technology Beyond English". Thanks to the Text Laboratory at the University of Oslo for giving us access to the Glossa system and especially to Anders Nøklestad for valuable assistance. Thanks are also due to three anonymous reviewers who made many valuable comments on a previous version of this paper.

## References

1. Benediktsson, H.: The Old Norse passive: Some observations. In: E. Hovdhaugen (ed.) The Nordic Languages and Modern Linguistics 4, pp. 108–119. Universitetsforlaget, Oslo (1980)
2. Bjarnadóttir, K.: The Icelandic mu-tbl experiment:preparing the corpus (2002). Paper presented at NLP1 final session, January 9. GSLT, Växjö
3. Brants, T.: TnT—a statistical part-of-speech tagger. In: Proceedings of the 6th Applied NLP Conference, ANLP-2000, pp. 224–231. Seattle (2000)
4. Christ, O.: A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX'94. 3rd Conference on Computational Lexicography and Text Research, pp. 23–34. Budapest (1994)
5. Degnbol, H.: Hvad en ordbog behøver—og andre ønsker [what a dictionary needs—and others wish for]. In: The Sixth International Saga Conference. Workshop Papers I, pp. 235–254. Det Arnamagnæanske Institut, University of Copenhagen, Copenhagen (1995)
6. Dyvik, H.: Har gammelnorsk passiv? [does Old Norse have the passive?]. In: E. Hovdhaugen (ed.) The Nordic Languages and Modern Linguistics 4, pp. 82–107. Universitetsforlaget, Oslo (1980)
7. Faarlund, J.T.: Syntactic Change. Toward a Theory of Historical Syntax. Mouton, Berlin (1990)
8. Faarlund, J.T.: The Syntax of Old Norse. Oxford University Press, Oxford (2004)
9. Halldórsson, B., Torfason, J., Tómasson, S., Thorsson, Ö. (eds.): Íslendinga sögur [The Icelandic Family Sagas]. Svart á hvítu (1985–86)
10. Haugan, J.: Old Norse word order and information structure. Ph.D. thesis, NTNU, Trondheim (2001)
11. Helgadóttir, S.: Testing data-driven learning algorithms for pos tagging of Icelandic. In: H. Holmboe (ed.) Nordisk Sprogteknologi. Årbog 2004, pp. 257–265. Museum Tusculanums Forlag, Copenhagen (2005)
12. Helgadóttir, S.: Mörkun íslensks texta [tagging Icelandic text]. Orð og tunga **9**, 75–107 (2007)
13. Holmberg, A.: Word order and syntactic features in the Scandinavian languages and English. Ph.D. thesis, University of Stockholm, Stockholm (1986)
14. Holmberg, A., Platzack, C.: The Role of Inflection in the Syntax of Scandinavian Languages. Oxford University Press, Oxford (1995)
15. Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., Kolak, O.: Bootstrapping parsers via syntactic projection across parallel texts. Natural Language Engineering **11**(3), 311–325 (2005)
16. Kristjánsdóttir, B., Halldórsson, B., Sigurðsson, G., Grímsdóttir, G.Á., Ingólfsdóttir, G., Torfason, J., Tómasson, S., Thorsson, Ö. (eds.): Sturlunga saga [The Sturlunga Collection]. Svart á hvítu, Reykjavík (1988)
17. Kristjánsdóttir, B., Halldórsson, B., Torfason, J., Thorsson, Ö. (eds.): Heimskringla [The Sagas of the Kings of Norway]. Mál og menning, Reykjavík (1991)
18. Kroch, A., Santorini, B., Delfs, L.: Penn–Helsinki parsed corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1/ (2004)
19. Kroch, A., Taylor, A.: Penn-Helsinki parsed corpus of Middle English. http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/ (2000). Second edition
20. Kroch, A., Taylor, A.: Verb-object order in Early Middle English. In: S. Pintzuk, G. Tsoulas, A. Warner (eds.) Diachronic Syntax: Models and Mechanisms, pp. 132–163. Oxford University Press, Oxford (2001)

21. Kroch, A., Taylor, A., Ringe, D.: The Middle English verb-second constraint: a case study in language contact and language cange. In: S.C. Herring, P. van Reenen, L. Schøsler (eds.) Textual Parameters in Older Language, pp. 353–391. John Benjamins, Philadelphia (2000)
22. Loftsson, H., Kramarczyk, I., Helgadóttir, S., Rögnvaldsson, E.: Improving the POS tagging accuracy of Icelandic text. In: K. Jokinen, E. Bick (eds.) Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA-2009), pp. 103–110. Odense (2009)
23. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn treebank. Computational Linguistics **19**(2), 313–330 (1993)
24. Mason, L.: Object shift in Old Norse. Master's thesis, University of York, York (1999)
25. Megyesi, B.: Data-driven syntactic analysis—methods and applications for Swedish. Ph.D. thesis, Department of Speech, Music and Hearing. KTH, Stockholm (2002)
26. Nygaard, L., Priestley, J., Nøklestad, A., Johannessen, J.B.: Glossa: A multilingual, multimodal, configurable user interface. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), pp. 617–621. European Language Resources Association (ELRA), Paris (2008)
27. Pind, J., Magnússon, F., Briem, S. (eds.): Íslensk orðtíðnibók [Icelandic Frequency Dictionary, IFD]. Orðabók Háskólans, Reykjavík (1991)
28. Rögnvaldsson, E.: Orðstöðulykill Íslendinga sagna [the concordance to the Icelandic family sagas]. Skáldskaparmál **1**, 54–61 (1990)
29. Rögnvaldsson, E.: Old Icelandic: A non-configurational language? NOWELE **26**, 3–29 (1995)
30. Rögnvaldsson, E.: Word order variation in the VP in Old Icelandic. Working Papers in Scandinavian Syntax **58**, 55–86 (1996)
31. Rögnvaldsson, E.: Setningafræðilegar breytingar í íslensku. [syntactic changes in Icelandic.]. In: H. Thráinsson (ed.) Setningar. Handbók um setningafræði [Sentences: A Handbook on Syntax], Íslensk tunga III, pp. 602–635. Almenna bókafélagið, Reykjavík (2005)
32. Rögnvaldsson, E.: The corpus of spoken Icelandic and its morphosyntactic annotation. In: P.J. Henrichsen, P.R. Skadhauge (eds.) Treebanking for Discourse and Speech, Proceedings of the NODALIDA 2005 Special Session on Treebanks for Spoken Language and Discourse, Copenhagen Studies in Language 32, pp. 133–145. Samfundslitteratur, Copenhagen (2006)
33. Rögnvaldsson, E., Ingason, A.K., Sigurðsson, E.F.: Coping with variation in the Icelandic diachronic treebank. Oslo Studies in Language (2011). Forthcoming.
34. Sundquist, J.D.: Object shift and Holmberg's generalization. In: D. Lightfoot (ed.) Syntactic Effects of Morphological Change, pp. 326–347. Oxford University Press, Oxford (2002)
35. Thráinsson, H.: The Syntax of Icelandic. Cambridge University Press, Cambridge (2007)
36. Wallenberg, J., Ingason, A.K., Sigurðsson, E.F., Rögnvaldsson, E.: Icelandic parsed historical corpus (IcePaHC). http://www.linguist.is/icelandic_treebank (2010). Version 0.2

# Part III
# Linguistic Resources for CH/SSH

# The Ancient Greek and Latin Dependency Treebanks

David Bamman and Gregory Crane

**Abstract** This paper describes the development, composition, and several uses of the Ancient Greek and Latin Dependency Treebanks, large collections of Classical texts in which the syntactic, morphological and lexical information for each word is made explicit. To date, over 200 individuals from around the world have collaborated to annotate over 350,000 words, including the entirety of Homer's *Iliad* and *Odyssey*, Sophocles' *Ajax*, all of the extant works of Hesiod and Aeschylus, and selections from Caesar, Cicero, Jerome, Ovid, Petronius, Propertius, Sallust and Vergil. While perhaps the most straightforward value of such an annotated corpus for Classical philology is the morphosyntactic searching it makes possible, it also enables a large number of downstream tasks as well, such as inducing the syntactic behavior of lexemes and automatically identifying similar passages between texts.

**Key words:** treebanks, dependency grammar, digital libraries, Ancient Greek, Latin

## 1 Introduction

The definitive Classical reference grammars of the 19th and 20th centuries, such as Herbert Smyth's *Greek Grammar* [44] and Raphael Kühner's *Ausführliche Grammatik der lateinischen Sprache* [29], are monuments of scholarship that distill lifetimes of reading and linguistic observation into succinct aphorisms such as the following:

> "Apodotic δέ is very common in Homer and Herodotus, not rare in Attic poetry, but infrequent in Attic prose." (Smyth 2837).

David Bamman
Perseus Project, Tufts University e-mail: david.bamman@tufts.edu

Gregory Crane
Perseus Project, Tufts University e-mail: gregory.crane@tufts.edu

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_5, © Springer-Verlag Berlin Heidelberg 2011

On occasion these works offer a window into the traditional philological practice that lies behind them, as in Kühner's comparison of the Latin *accusativus cum infinitivo* construction with subordinate clauses containing an overt complementizer (e.g., *quod*):

> "So hat nach meiner Zählung bei *doleo* 57 Stellen mit *Acc. c. Inf.* gegen 4 *quod*, bei *miror* 110 gegen 8, bei *glorior* 19 gegen 2, bei *queror* 71 gegen 15, bei *gaudeo* 84 gegen 9 usw." (1914:77)[1]

In its most basic form, classical philology of this sort is by definition a data-driven science: it relies on a fixed dataset (the extant corpus of Ancient Greek and Latin) and builds larger arguments by the simple act of counting. Kühner here publishes his tally of ACI vs. *quod*-clauses in order to advance the argument that the ACI is more frequent in indirect discourse than subordinate clauses are, and one can assume that either such an explicit tally or an implicit one (collected over a lifetime of reading) is what drives Smyth's observations on relative frequency as well.

Where classical philology has so far diverged from data-driven science, however, is in its reliance on the authority of the editor rather than on the data itself. As much as the judgment of Kühner and Smyth may far exceed our own, the cornerstone of the scientific method is the reproducibility of experiments, and as P. Cuzzolin [20] notes about this very passage of Kühner:

> "...it is difficult to say what he meant by the word "Stelle" and impossible to say which texts his counting is based upon."

Ideally, what we want to see is the evidence that drives such linguistic observations – not simply knowing that the ACI is used in some unknown sample of 57 sentences containing *doleo*, but exactly which sentences those are, which textual editions they come from, and how that small sample relates to the corpus at large (if only to measure its significance). While such a work may not have been possible in the print culture of the past, we are at a transformative moment now where we can begin leveraging the scientific method in the service of classical philology.

## 2 Treebanks

Our work in developing treebanks for Ancient Greek and Latin are our own efforts to help move classical philology into this scientific space. A treebank is a large collection of sentences in which the syntactic relation for every word is made explicit – where a human has encoded an interpretation of the sentence in the form of a linguistic annotation. While much of the research and labor in treebanks over the past twenty years has focused on modern languages such as English [33], Arabic [22, 32] and Czech [21], recent scholarship has seen the rise of a number

[1] "And so, by my count, with *doleo* there are 57 sentences with the accusative + infinitive against 4 with *quod*, with *miror* 110 against 8, with *glorior* 19 against 2, with *queror* 71 against 15, with *gaudeo* 84 against 9 etc."

of treebanks for historical languages as well, including Middle English [28], Early Modern English [27], Old English [46], Medieval Portuguese [39], Ugaritic [49], Latin [2, 37] and several Indo-European translations of the New Testament [23].

Treebanks have been annotated under a variety of grammatical frameworks, with tthe most dominant being the phrase structure grammar employed by the Penn Treebank of English [33] and the dependency grammar in use by the Prague Dependency Treebank of Czech [21]. The defining feature of dependency grammar (Mel'cuk [35], Sgall [41], Tesniere [47]) that distinguishes it from constituent-based formalisms is the absence of non-terminal nodes common in $\overline{\text{X}}$ theory (Chomsky [15]) – while individual words under a phrase structure grammar form part of abstract structures such as NP (noun phrase) and VP (verb phrase), in a dependency grammar they are directly linked to each other via asymmetrical dependency relationships, with each word being the child (or "dependent") of exactly one other word. Dependency grammars deal especially well with languages involving relatively free word order (which in a transformational grammar would otherwise involve a high degree of scrambling), and this flexibility led us to adopt it as the annotation style for our treebanks as well.

Figure 1 represents one such example of a dependency annotation from an elegy of Propertius.

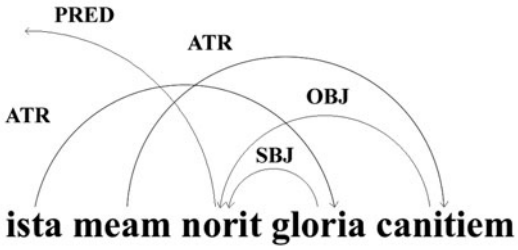


**Fig. 1** Dependency graph of the treebank annotation for *ista meam norit gloria canitiem* ("that glory will know my old age"), Propertius I.8.46. Arcs are directed from children to their parents

Classical texts have been a focused object of study for over two thousand years, with schoolchildren and tenured professors alike scrutinizing their every word; a treebank is simply an effort to capture such analysis in a quantified format that can provide a measurable dataset for reproducing linguistic experiments while also driving a new generation of computational analysis.

## 3 Building the Ancient Greek and Latin Dependency Treebanks

The Ancient Greek and Latin Dependency Treebanks are the work of over 200 researchers from around the world. The backgrounds of the annotators range from advanced undergraduate students to recent PhDs and professors, with the majority

being students in graduate programs in Classics. Annotators undergo an initial training period in which they learn the application of dependency grammar to Greek and Latin texts (as encoded in the guidelines for syntactic annotation[2]) and are actively engaged in new learning afterwards by means of an online forum in which they can ask questions of each other and of project editors; this allows them to be kept current on the most up-to-date codifications to the annotation guidelines while also helping bring new annotators up to speed. In the "standard" model of production, every sentence is annotated by two independent annotators and the differences are then reconciled by a third in order to filter out the biases (and errors) of any single individual.[3] This reconciliation (or "secondary" annotation as it is encoded in the XML release) is undertaken by a more experienced annotator/editor, typically a PhD with specialization in the particular subject area (such as Homer).

As Fig. 2 illustrates, all annotations are publicly released with the usernames of the primary and secondary annotators (which are then also associated with real names and institutional affiliations). By publicly acknowledging authorship, we are making our first steps toward an ownership model for annotation and hope to provide a means for students, both graduate and undergraduate alike, to engage in the act of scholarly research and in the production of scientific data that can be useful to the wider Classics community.

```
<sentence id="341" document_id="Perseus:text:1999.02.0066" subdoc="book=1:poem=8b" span="ista0:canitiem0">
  <primary>alexlessie</primary>
  <primary>sneil01</primary>
  <secondary>millermo</secondary>
  <word id="1" form="ista" lemma="iste1" postag="p-s---fn-" head="4" relation="ATR"/>
  <word id="2" form="meam" lemma="meus1" postag="a-s---fa-" head="5" relation="ATR"/>
  <word id="3" form="norit" lemma="nosco1" postag="v3srsa---" head="0" relation="PRED"/>
  <word id="4" form="gloria" lemma="gloria1" postag="n-s---fn-" head="3" relation="SBJ"/>
  <word id="5" form="canitiem" lemma="canities1" postag="n-s---fa-" head="3" relation="OBJ"/>
</sentence>
```

**Fig. 2** XML fragment from the Latin Dependency Treebank (Propertius I.8.46)

While the goal of the standard method of production is to filter out individual bias, we also want to provide the ability for scholars to publish a record of their own unique interpretation of the text – an interpretation that stands as theirs alone. In this spirit, we have also developed a "scholarly" method of annotation [7] and have published part of our treebank under this model. By publicly releasing data with citable attributions of ownership in this way, we hope to provide a fixed core around which other interpretations (by other scholars) can then be layered. Literary works very often license multiple valid syntactic annotations and, for ancient texts especially, scholarly disagreement can be found not only on the level of the correct syntactic parse, but also on the form of the text itself (since we do not have the original text in the author's own hand, but rather a series of copies by medieval

[2] The annotation guidelines are available in English (for Greek [5] and Latin [8]) and in Spanish (for Greek [6] and Latin [9]).

[3] The interannotator agreement rate for the Ancient Greek Dependency Treebank measures 87.4% for attachment agreement, 85.3% for label agreement, and 80.6% for labeled attachment [7].

scribes). Providing a quantified record of how these multiple interpretations differ can only help drive future research.

## 4 Ancient Greek Dependency Treebank

The current version of the Ancient Greek Dependency Treebank (v. 1.2) includes the entirety of Homer's *Iliad* and *Odyssey*, Sophocles' *Ajax*, and all of the works of Hesiod and Aeschylus for a total of 309,096 words, as distributed in Table 1.

**Table 1** Composition of the Ancient Greek Dependency Treebank (version 1.2)

| Method | Author | Work | Sentences | Words |
|---|---|---|---|---|
| Standard | Hesiod | Shield of Heracles | 255 | 3,834 |
| | | Theogony | 438 | 8,106 |
| | | Works and Days | 491 | 6,941 |
| | Homer | Iliad | 8,415 | 128,102 |
| | | Odyssey | 6,760 | 104,467 |
| Scholarly | Aeschylus | Agamemnon | 814 | 9,806 |
| | | Eumenides | 526 | 6,380 |
| | | Libation Bearers | 572 | 6,563 |
| | | Persians | 478 | 6,223 |
| | | Prometheus Bound | 589 | 7,045 |
| | | Seven Against Thebes | 518 | 6,206 |
| | | Suppliants | 529 | 5,949 |
| | Sophocles | Ajax | 785 | 9,474 |
| | | **Total:** | **21,170** | **309,096** |

In addition to the index of its syntactic head and the type of relation to it, each word is also annotated with the lemma from which it is inflected and its morphological code (a composite of nine different morphological features: part of speech, person, number, tense, mood, voice, gender, case and degree). All of the files have been freely released under a Creative Commons license.[4]

For the works of Homer and Hesiod, we have followed the standard production method of soliciting annotations from two different annotators and then reconciling the differences between them. Sophocles and Aeschylus, whose textual traditions are much more fragmentary, have presented an ideal case for annotation as scholarly treebanks.

[4] All treebank data can be found at: http://nlp.perseus.tufts.edu/syntax/treebank/.

## 5 Latin Dependency Treebank

Currently in version 1.5, the Latin Dependency Treebank is comprised of 53,143 words from eight texts, as shown in Table 2. As with the Ancient Greek Dependency Treebank, each word is also annotated with the lemma from which it is inflected and its morphological code. All of the texts in this release have been annotated under the standard model of production, with an editor reconciling the differences between two independent annotations.

**Table 2** Composition of the Latin Dependency Treebank (version 1.5)

| Method | Author | Work | Sentences | Words |
|---|---|---|---|---|
| Standard | Caesar | B.G. (Book 2 selections) | 71 | 1,488 |
| | Cicero | In Catilinam 1.1-2.11 | 327 | 6,229 |
| | Jerome | Vulgate: Apocalypse | 405 | 8,382 |
| | Ovid | Metamorphoses: Book I | 316 | 4,789 |
| | Petronius | Satyricon 26-78 (Cena Trimalchionis) | 1,114 | 12,474 |
| | Propertius | Elegies: Book I | 361 | 4,857 |
| | Sallust | Catilina | 701 | 12,311 |
| | Vergil | Aeneid (Book 6 selections) | 178 | 2,613 |
| | | **Total** | **3,473** | **53,143** |

## 6 The Influence of a Digital Library

The composition of historical treebanks is fundamentally different from that of modern ones. On the one hand, the efficient annotation of Ancient Greek and Latin is hindered by the fact that no native speakers exist and the texts that we have available are typically highly stylized in nature. On the other hand, however, while modern treebanks are generally comprised of newspaper articles,[5] the texts that make up historical treebanks have generally been the focus of scholarly attention for centuries, if not millennia. The Penn-Helsinki Parsed Corpus of Middle English [28], for example, includes Chaucer's 14th-century *Parson's Tale*, while the York Poetry Corpus [38] includes the entire text of *Beowulf*. The scholarship that has attended these texts since their writing has produced a wealth of contextual materials, including commentaries, translations, and linguistic resources.

In building a workflow for creating treebanks for Ancient Greek and Latin, we attempt to provide as much of this kind of contextualizing information for each

[5] To name just three, the Penn Treebank [33] is comprised of texts from the *Wall Street Journal*; the German TIGER Treebank [10] is built from texts taken from the *Frankfurter Rundschau*; and the Prague Dependency Treebank [21] includes articles from several daily newspapers (*Lidové noviny* and *Mladá fronta Dnes*), a business magazine (*Českomoravský Profit*) and a scientific journal (*Vesmír*).

sentence as possible, and embedding our annotation environment within the Perseus Digital Library has been crucial in this respect. Established in 1987 in order to construct a large, heterogeneous collection of textual and visual materials on the archaic and classical Greek world, Perseus today serves as a laboratory for digital library technologies and is also widely used by students, academics and others to access information on the Greco-Roman world [17–19].

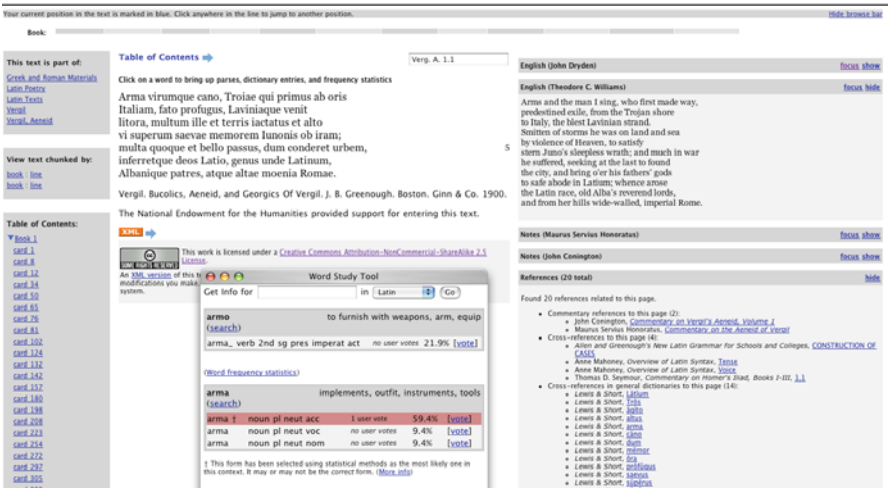


**Fig. 3** A screenshot of Vergil's *Aeneid* from the Perseus digital library

Figure 3 shows a screenshot from this digital library. In this view, the reader is looking at the first seven lines of Vergil's *Aeneid*. The source text is provided in the middle, with contextualizing information filling the right column. This information includes:

- Translations. Here two English translations are provided, one by the 17th-century English poet John Dryden and a more modern one by Theodore Williams.
- Commentaries. Two commentaries are also provided, one in Latin by the Roman grammarian Servius, and one in English by the 19th-century scholar John Conington.
- Citations in reference works. Classical reference works such as grammars and lexica often cite particular passages in literary works as examples of use. Here, all of the citations to any word or phrase in these seven lines are presented at the right.

Additionally, every word in the source text is linked to its morphological analysis, which lists every lemma and morphological feature associated with that particular word form. Here the reader has clicked on *arma* in the source text. This tool reveals that the word can be derived from two lemmas (the verb *armo* and the noun

*arma*), and gives a full morphological analysis for each. A recommender system automatically selects the most probable analysis for a word given its surrounding context, and users can also vote for the form they think is correct.[6]

A cultural heritage digital library has provided a fertile ground for our historical treebanks in two fundamental ways: by providing a structure on which to build new services and by providing reading support to expedite the process of annotation.

## *6.1 Structure*

By anchoring the treebank in a cultural heritage digital library, we are able to take advantage of a structured reading environment with canonical standards for the presentation of text and a large body of digitized resources, which include XML source texts, morphological analyzers, machine-readable dictionaries, and an online user interface.

### 6.1.1 Texts

The Perseus Digital Library contains 3.4 million words of Latin source texts along with 4.9 million words of Greek. The texts are all public-domain materials that have been scanned, OCR'd and formatted into TEI-compliant XML. The value of this prior labor is twofold: most immediately, the existence of clean, digital editions of these texts has saved us a considerable amount of time and resources in processing them for annotation, as we would otherwise have to create them before annotating them syntactically; but their encoding as repurposeable XML documents in a larger library also allows us to refer to them under standardized citations. The passage of Vergil displayed in Fig. 3 is not simply a string of unstructured text; it is a subdocument (*Book=1:card=1*) that is itself part of a larger document object (*Perseus:text:1999.02.0055*), with sisters (*Book=1:card=8*) and children of its own (e.g., *line=4*). This XML structure allows us to situate any given treebank sentence within its larger context.

### 6.1.2 Morphological Analysis

As highly inflected languages, Ancient Greek and Latin have an intricate morphological system, in which a full morphological analysis is the product of nine features: part of speech, person, number, tense, mood, voice, gender, case and degree. Our digital library has included a morphological analyzer from its beginning. This resource maps an inflected form of a word (such as *arma* above) to all of the possible

[6] These user contributions have the potential to significantly improve the morphological tagging of these texts: any single user vote assigns the correct morphological analysis to a word 89% of the time, while the recommender system does so with an accuracy of 76% [19].

analyses for all of the dictionary entries associated with it. In addition to providing a common morphological standard, this mapping greatly helps to constrain the problem of morphological tagging (selecting the correct form from all possible forms), since a statistical tagger only needs to consider the morphological analyses licensed by the inflection rather than all possible combinations.

### 6.1.3 User Interface

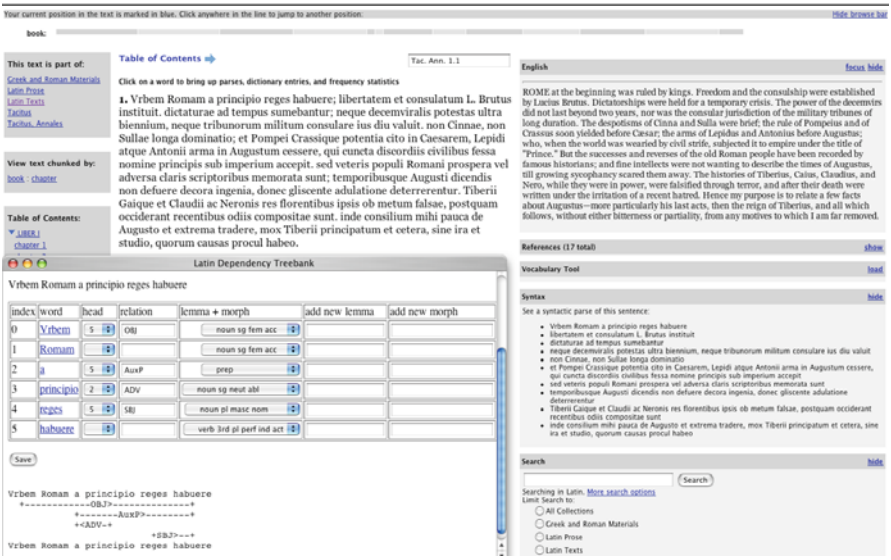


**Fig. 4** A screenshot of Tacitus' *Annales* from the Perseus digital library

The user interface of our library is designed to be modular, since different texts have different contextual resources associated with them (while some have translations, others may have commentaries). This modularity allows us to easily introduce new features, since the underlying architecture of the page doesn't change – a new feature can simply be added.

Figure 4 presents a screenshot of the digital library with an annotation tool built into the interface. In the widget on the right, the source text in view (the first chunk of Tacitus' *Annales*) has been automatically segmented into sentences; an annotator can click on any sentence to assign it a syntactic annotation. Here the user has clicked on the first sentence (*Vrbem Romam a principio reges habuere*); this action brings up an annotation screen in which a partial automatic parse is provided, along with the most likely morphological analysis for each word. The annotator can then correct this automatic output and move on to the next segmented sentence, with all of the contextual resources still in view.

Our collaboration with the Alpheios Project has also allowed us to integrate a graphical treebank editor into our annotation process to make the construction of trees more intuitive and to provide annotators with greater flexibility as to their preferred input method. Figure 5 shows a tree in the process of being constructed, with a single word (*Romam*) being dragged onto its syntactic head.

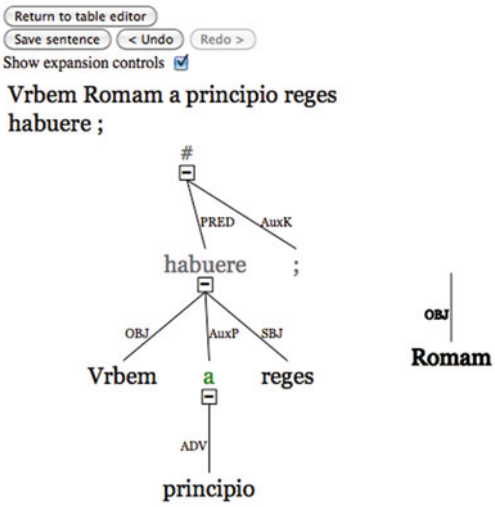


**Fig. 5** A screenshot of the first sentence of Tacitus' *Annales* being constructed using the Alpheios graphical editor

## *6.2 Reading Support*

Modern treebanks also differ from historical ones in the fluency of their annotators. The efficient annotation of historical languages is hindered by the fact that no native speakers exist, and this is especially true of Ancient Greek and Latin, both languages with a high degree of flexibility in word order. While the Penn Treebank can report a productivity rate of between 750 and 1,000 words per hour for their annotators after four months of training [45] and the Penn Chinese treebank can report a rate of 240-480 words per hour [14], our annotation speeds are significantly slower, ranging from 97 words per hour to 211, with an average of 124. Our best approach for these languages is to develop strategies that can speed up the annotation process, and here the resources found in a digital library are crucial. There are three varieties of contextual resources in our digital library that aid in the understanding of a text: translations, commentaries, and dictionaries. These resources shed light on a text, from the level of sentences to that of individual words.

### 6.2.1 Translations

Translations provide reading support on a large scale: while loose translations may not be able to inform readers about the meaning and syntactic role of any single word, they do provide a broad description of the action taking place, and this can often help to establish the semantic structure of the sentence – who did what to whom, and how. In a language with a free word order (and with poetry especially), this kind of high-level structure can be important for establishing a quick initial understanding of the sentence before narrowing down to individual syntactic roles.

### 6.2.2 Commentaries

Classical commentaries provide information about the specific use of individual words, often noting morphological information (such as case) for ambiguous words or giving explanatory information for unusual structures. This information often comes at crucial decision points in the annotation process, and represents judgments by authorities in the field with expertise in that particular text.

[4] **Vi superum** expresses the general agency, like 'fato profugus,' though Juno was his only personal enemy. Gossrau's fancy that 'vi superum' = βίᾳ θεῶν, 'in spite of heaven,' has no authority. For 'memorem iram' comp. **Livy 9. 29.** "Traditur censorem etiam Appium memori Deum ira post aliquot annos luminibus captum." So **Aesch. Ag. 155.** "μνάμων μῆνις". 'Ob iram,' below, v. 251, 'to sate the wrath.'

[5] **Passus,** constructed like 'iactatus.' 'Quoque' and 'et' of course form a pleonasm, though the former appears to be connected with 'multa,' and the latter with 'bello.' 'Dum conderet' like "dum fugeret," **G. 4. 457,** where see note. Here we might render 'in the struggle to build his city.' So Hom. Od. 1. 4. foll., πολλὰ πάθεν . . ἀρνύμενος κ.τ.λ. The clause belongs to 'multa bello passus,' rather than to 'iactatus.'

**Fig. 6** An excerpt from Conington's commentary on Vergil's *Aeneid* [16], here referring to Book 1, lines 4 and 5

### 6.2.3 Machine-Readable Dictionaries

In addition to providing lists of stems for morphological analyzers, machine-readable dictionaries also provide valuable reading support for the process of lemma selection. Every available morphological analysis for a word in the Perseus Digital Library is paired with the word stem (a lemma) from which it is derived, but analyses are often ambiguous between different lemmas. The extremely common form *est*, for example, is a third person singular present indicative active verb, but can be inflected from two different lemmas: the verb *sum* (to be) and the verb *edo* (to eat). In this case, we can use the text already tagged to suggest a more probable form (*sum* appears much more frequently and is therefore the likelier candidate),

but in less dominant cases, we can use the dictionary: since the word stems involved in morphological analysis have been derived from the dictionary lemmas, we can map each analysis to a dictionary definition, so that, for instance, if an annotator is unfamiliar with the distinction between the lemmas *occido1* (to strike down) and *occido2* (to fall), their respective definitions can clarify it.

Machine-readable dictionaries, however, are also a valuable annotation resource in that they often provide exemplary syntactic information as part of their definitions. Consider, for example, the following line from Book 6, line 2 of Vergil's *Aeneid*: *et tandem Euboicis Cumarum adlabitur oris* ("and at last it glides to the Euboean shores of Cumae"). The noun *oris* (shores) here is technically ambiguous, and can be derived from a single lemma (*ora*) as a noun in either the dative or ablative case. The dictionary definition of *allabor* (to glide), however, disambiguates this for us, since it notes that the verb is often constructed with either the dative or the accusative case.

**al-lābor** (**adl-** ), lapsus, 3, v. dep.,

**I.** *to glide to* or *toward* something, *to come to, to fly, fall, flow, slide*, and the like; constr. with dat. or acc. (**poet.**—oftenest in Verg.— "or in more elevated prose): viro adlapsa sagitta est," **Verg. A. 12, 319**: "fama adlabitur auris," *id. ib. 9, 474*: Curetum adlabimur oris, *we land upon*, etc., *id. ib. 3, 131*; cf. *id. ib. 3, 569*: "mare crescenti adlabitur aestu," *rolls up with increasing wave*, *id. ib. 10, 292*: "adlapsus genibus," *falling down at his knees*, *Sen. Hippol. 666*.—In prose: umor adlapsus extrinsecus, * *Cic. Div. 2, 27, 58*: "angues duo ex occulto adlapsi," **Liv. 25, 16**.

**Fig. 7** Definition of *allabor* (the dictionary entry for *adlabitur*) from Lewis and Short [30]

Every word in our digital library is linked to a list of its possible morphological analyses, and each of those analyses is linked to its respective dictionary entry. The place of a treebank in a digital library allows for this tight level of integration.

## 7 The Impact of Historical Treebanks

The traffic in the Perseus Digital Library currently exceeds 10 million page views by 400,000 distinct users per month. These users are not computational linguists or computer scientists who would typically make use of a treebank; they are a mix of Classical scholars and students. These different audiences have equally different uses for a large corpus of syntactically annotated sentences: for one group it can provide additional reading support, and for the other a scholarly resource to be queried. The Ancient Greek and Latin Dependency Treebanks currently yield

a powerful range of search options, including lemmatized and morphosyntactic searching, and have already been valuable for downstream research involving lexicography and identifying textual reuse.

## 7.1 Lemmatized Searching

The ability to conduct a lemma-based textual search has long been a desideratum in Classics,[7] where any given Latin word form, for example, has 3.1 possible analyses on average.[8] Locating all inflections of *edo* (to eat) in the texts of Caesar, for example, would involve two things:

1. Searching for all possible inflections of the root word. This amounts to 202 different word forms attested in our texts (including compounds with enclitics).
2. Eliminating all results that are homonyms derived from a different lemma. Since several inflections of *edo* are homonyms with inflections of the far more common *sum* (to be), many of the found results will be false positives and have to be discarded.

This is a laborious process and, as such, is rarely undertaken by Classical scholars: the lack of such a resource has constrained the set of questions we can ask about a text. Since a treebank encodes each word's lemma in addition to its morphological and syntactic analysis, this information is now easily accessible.

## 7.2 Morphosyntactic Searching

A treebank's major contribution to scholarship is that it encodes an interpretation of the syntax of a sentence, along with a morphological analysis of each word. These two together can be combined into elaborate searches, allowing scholars to find all instances of any particular morphosyntactic construction, such as the different types of subordinate clauses headed by the conjunction *cum* (when *cum* is the head of a subordinate clause whose verb is indicative, it is often recognized as a temporal clause, qualifying the time of the main clause's action; when that verb is subjunctive, however, the clause retains a different meaning, as either circumstantial, causal, or adversative). This type of searching allows us to gather statistical data on usage

[7] Both the Perseus Project and the Thesaurus Linguae Graecae (http://www.tlg.uci.edu) allow users to search for all inflected forms of a lemma in their texts, but neither filters results that are homonyms derived from different lemmas.

[8] Based on the average number of lemma + morphology combinations for all unique word tokens in our 3.4 million word corpus. The word form *amor*, for example, has 3 analyses: as a first-person singular present indicative passive verb derived from the lemma *amo* (to love) and as either a nominative or vocative masculine singular noun derived from *amor* (love).

while also locating individual examples for further qualitative analysis.[9] Figure 8 displays one tool for such analysis (Annis [48]) with a sample query from the Ancient Greek Dependency Treebank.

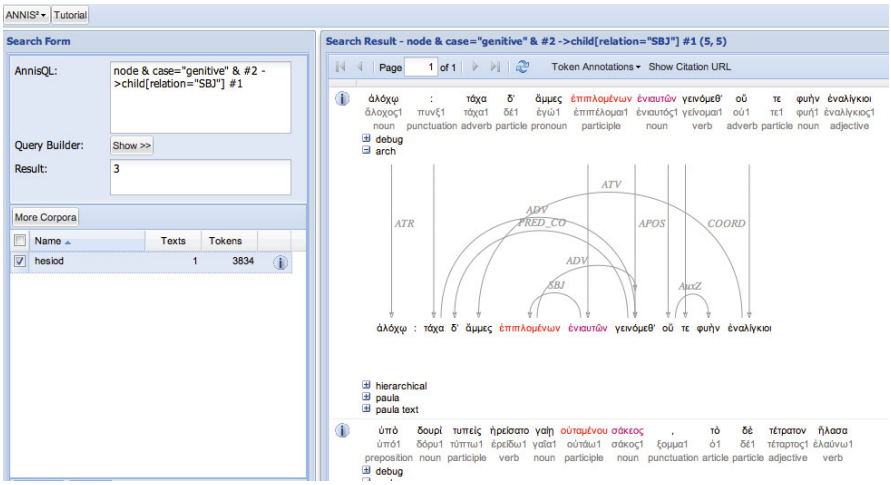


**Fig. 8** Morphosyntactic search for genitive absolutes in Hesiod using the Annis search tool [48]

## *7.3 Lexicography*

In addition to driving linguistic research on syntax itself, treebanks have been instrumental for several downstream computational tasks as well. One such task has been automatically inducing lexical information from large corpora in the service of automatically building bilingual dictionaries. Lexical information broadly defines what individual words "mean" and how they interact with others. Lexicographers have been exploiting large, unstructured corpora for this kind of knowledge in the service of dictionary creation since the COBUILD project [43] of the 1980s, often in the form of extracting frequency counts and collocations – a word's frequency information is especially important to second language learners, and collocations (a word's "company") are instrumental in delimiting its meaning. This corpus-based approach to lexicon building has since been augmented in two dimensions: on the one hand, dictionaries and lexicographic resources are being built on larger and larger textual collections: the German *elexiko* project [26], for instance, is built on a modern German corpus of 1.3 billion words, and we can expect much larger

[9] For the importance of a treebank in expediting morphosyntactic research in Latin rhetoric and historical linguistics, see Bamman and Crane [1].

projects in the future as the web is exploited as a corpus.[10] At the same time, researchers are also subjecting their corpora to more complex automatic processes in order to extract more knowledge from them. While word frequency and collocation analysis is fundamentally a task of simple counting, projects such as Kilgarriff's Sketch Engine [25] also enable lexicographers to induce information about a word's grammatical behavior as well.

Treebanks have helped drive this work by providing a dataset from which to induce syntactic behavior for individual lexemes [3]. While it is large collections of parallel texts (Latin/English and Greek/English) that provide the basic material for mining the dominant English senses of Greek and Latin words [13], the role of a treebank here is to provide the training material for an automatic parser (such as McDonald et al's MSTParser [34]), which can then provide a syntactic parse for all of the source texts in our comparatively much larger collection. With this syntactic information, we can far better calculate a word's relationships to the other words in a sentence, and more properly delimit what "company" we want to consider when inferring its meaning.

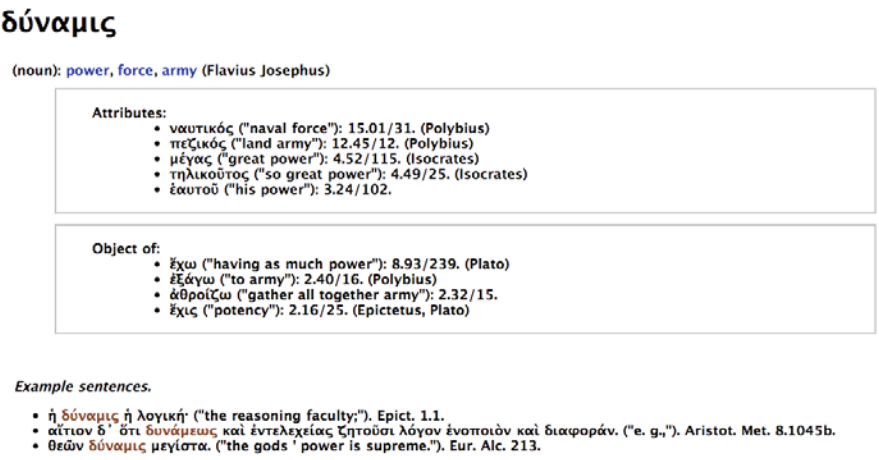


**Fig. 9** Automatically derived lexical information for the Greek word δύναμις

Figure 9 presents one example of such an automatically created lexical entry for the Greek noun δύναμις. While a traditional Greek lexicon such as the LSJ [31] can present much more detailed information about this word, we can here provide a quantitative measure of how frequently each sense appears in our corpus, and for which authors any given sense is dominant. δύναμις in general means "force" or "power" (the two most dominant senses found here), but it also retains a specialized meaning of "military power" as a consequence. Syntactic information lets us specify

[10] In 2006, for example, Google released the first version of its Web 1T 5-gram corpus [11], a collection of n-grams (n=1-5) and their frequencies calculated from 1 trillion words of text on the web.

not just what words it's commonly found with, but exactly *how* those words interact – for example, that when πεζικός ("on foot") modifies it as an attribute, it attains a new meaning of "army." Structural knowledge lets us distinguish between what surrounding words are merely descriptive attributes of a noun in question, and which words require that noun as part of their essential argument structure. While simple collocates induced from unstructured data provide information on what words accompany any individual lexeme, a treebank can specify the exact nature of their interaction on a much more detailed level.

## *7.4 Discovering Textual Similarity*

Most studies on text reuse focus on identifying either documents that are duplicates or near-duplicates of each other (e.g., web pages) or sentences in one document that have been sampled from another (e.g., in plagiarism detection). These studies generally employ variations of word-level similarity, including relative frequency measures (spotting similarities in the distribution of word patterns between two documents) [24], IR similarity methods based on the TF-IDF scores of individual words [36] and fingerprinting using n-grams [12, 40, 42]. While n-grams are good at approximating syntax in languages with a relatively fixed word order (such as English and German), they are much less effective in languages where the word order is more free, such as Greek and Latin.

Additionally, when attempting to spot some of the more obliques classes of reuse – such as literary allusion – sometimes the strongest similarity can be found at a syntactic level. Consider, for example, the opening lines of the three great epics of Greco-Roman literature, Vergil's *Aeneid* and Homer's *Iliad* and *Odyssey*.

- arma virumque cano ("I sing of arms and the man") [Aen. 1.1]
- ἄνδρα μοι ἔννεπε, μοῦσα ("Tell me of the man, o Muse") [Od. 1.1]
- μῆνιν ἄειδε θεὰ ("Sing, goddess, of the rage") [Il. 1.1]

While there is a semantic similarity in all three examples (all three focus on the act of speaking and in two of the three it is a particular *man* that is spoken about), all three of them are most strongly similar by the explicit form of their structure. Figure 10 illustrates what these three phrases look like when annotated under a dependency grammar. In all cases, the initial phrase (arma/ἄνδρα/μῆνιν) is the direct object of the sentence predicate (cano/ἔννεπε/ἄειδε), wherever that happens to appear in the sentence.[11]

Our work in allusion detection [4] has focused on how to exploit the knowledge encoded in treebanks to automatically discover instances of textual reuse where the derived sentence bears some syntactic similarity to its source. Again, using our

[11] Note that we can also add later epics to this class as well, such as Milton's *Paradise Lost*: "Of man's disobedience, and the fruit of that forbidden tree ... sing, heavenly muse" (1.1-6), where the first syntactic phrase in the sentence is the object of the verb of telling.

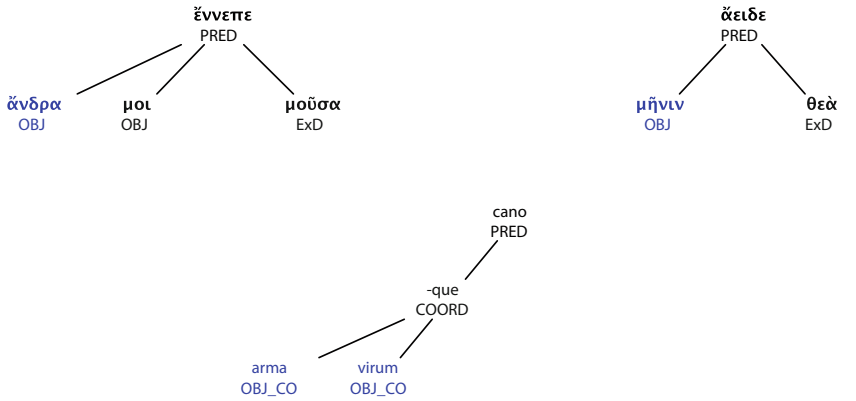


**Fig. 10** Syntactic trees of the opening of the *Odyssey*, *Iliad*, and *Aeneid*.

Latin treebank as training data for an automatic parser, we assigned a syntactic structure to all of the sentences in our larger textual collection. From this automatic structure we extracted 12 syntactic features for every word in the sentence – a combination of word-level representation (such as token, lemma or simply the part of speech), the length of the syntactic tree (including either just the parent or the parent and grandparent) and the presence or absence of an edge label (either simply specifying that a structural relation exists between a child and its parent, or also labeling that relationship as, e.g., an attributive one [ATR]). These features were then combined with other standard characteristics (such as word and lemma weights and n-grams) and used to calculate the similarity between two sentences, based on the cosine similarity between the two vectors that they constitute. Since each variable is weighted by TF-IDF, and syntactic features are relatively rare (with corresponding high IDF scores), syntactic features were generally found to be the most informative in establishing similarity. In its ability to generate this structural data, a treebank has enabled us to discover instances of text reuse even when the lexical similarity between two sentences is small and otherwise undetectable.

## 8 Conclusion

Treebanks already fill a niche in the computational linguistics community by providing valuable datasets for automatic processes such as parsing and grammar induction. Their utility, however, does not end there. The information that treebanks encode is of value to a wide range of potential users, including researchers not only in linguistics but in Classics as well, and we must encourage the use of these resources by making them available to such a diverse community. The treebanks so far are the work of hundreds of individuals who commit their interpretations of Greek and Latin sentences to a format that can be preserved for generations. While

this effort has resulted in the annotation of over 350,000 words of Classical texts, this is still only a small sample of the extant works in the Classical tradition; in the future, we plan to continue encouraging contributions to this ongoing work in order to strengthen, sentence by sentence, the foundation on which data-driven philology can stand.

**Acknowledgements** Grants from the Alpheios Project ("Building a Greek Treebank"), the National Endowment for the Humanities (PR-50013-08, "The Dynamic Lexicon: Cyberinfrastructure and the Automated Analysis of Historical Languages"), the Andrew W. Mellon Foundation ("The CyberEdition Project: Workflow for Textual Data in Cyberinfrastructure"), the Digital Library Initiative Phrase 2 (IIS-9817484) and the National Science Foundation (BCS-0616521) provided support for this work. This research used resources of the National Energy Research Scientific Computing Center, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. 

## References

1. Bamman, D., Crane, G.: The design and use of a Latin dependency treebank. In: Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006), pp. 67–78. ÚFAL MFF UK, Prague (2006)
2. Bamman, D., Crane, G.: The Latin Dependency Treebank in a cultural heritage digital library. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pp. 33–40. Association for Computational Linguistics, Prague (2007). URL http://www.aclweb.org/anthology/W/W07/W07-0905
3. Bamman, D., Crane, G.: Building a dynamic lexicon from a digital library. In: JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, pp. 11–20. ACM, New York, NY, USA (2008). DOI http://doi.acm.org/10.1145/1378889.1378892
4. Bamman, D., Crane, G.: The logic and discovery of textual allusion. In: Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008). Marrakesh (2008)
5. Bamman, D., Crane, G.: Guidelines for the syntactic annotation of Ancient Greek treebanks, version 1.1. Tech. rep., Tufts Digital Library, Medford (2009)
6. Bamman, D., Crane, G.: Pautas para la notación sintáctica del treebank de dependencia para el griego antiguo (1.1), traducción y adaptacón al español de Alejandro Abritta. Tech. rep., Tufts Digital Library, Medford (2010)
7. Bamman, D., Mambrini, F., Crane, G.: An ownership model of annotation: The Ancient Greek Dependency Treebank. In: The Eighth International Workshop on Treebanks and Linguistic Theories (2009)
8. Bamman, D., Passarotti, M., Crane, G., Raynaud, S.: Guidelines for the syntactic annotation of Latin treebanks, version 1.3. Tech. rep., Tufts Digital Library, Medford (2007)
9. Bamman, D., Passarotti, M., Crane, G., Raynaud, S.: Pautas para la notación sintáctica del treebank de dependencia para el latin (1.3), traducción y adaptacón al español de Alejandro Abritta. Tech. rep., Tufts Digital Library, Medford (2010)
10. Brants, S., Dipper, S., Hansen, S., Lezius, W., Smith, G.: The TIGER treebank. In: Proceedings of the Workshop on Treebanks and Linguistic Theories. Sozopol (2002)
11. Brants, T., Franz, A.: Web 1T 5-gram Version 1. Linguistic Data Consortium, Philadelphia (2006)
12. Brin, S., Davis, J., García-Molina, H.: Copy detection mechanisms for digital documents. SIGMOD Rec. **24**(2), 398–409 (1995). DOI http://doi.acm.org/10.1145/568271.223855

13. Brown, P.F., Pietra, V.J.D., Pietra, S.A.D., Mercer, R.L.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. **19**(2), 263–311 (1993)
14. Chiou, F.D., Chiang, D., Palmer, M.: Facilitating treebank annotation using a statistical parser. In: Proceedings of the First International Conference on Human Language Technology Research HLT '01, pp. 1–4 (2001)
15. Chomsky, N.: Remarks on nominalization. In: R. Jacobs, P. Rosenbaum (eds.) Reading in English Transformational Grammar. Ginn, Waltham (1970)
16. Conington, J. (ed.): P. Vergili Maronis Opera. The Works of Virgil, with Commentary. Whittaker and Co, London (1876)
17. Crane, G.: From the old to the new: Integrating hypertext into traditional scholarship. In: Hypertext '87: Proceedings of the 1st ACM conference on Hypertext, pp. 51–56. ACM Press (1987)
18. Crane, G.: New technologies for reading: The lexicon and the digital library. Classical World pp. 471–501 (1998)
19. Crane, G., Bamman, D., Cerrato, L., Jones, A., Mimno, D.M., Packel, A., Sculley, D., Weaver, G.: Beyond digital incunabula: Modeling the next generation of digital libraries. In: J. Gonzalo, C. Thanos, M.F. Verdejo, R.C. Carrasco (eds.) ECDL, *Lecture Notes in Computer Science*, vol. 4172, pp. 353–366. Springer (2006)
20. Cuzzolin, P.: On sentential complementation after *verba affectuum*. In: J. Herman (ed.) Linguistic Studies on Latin, pp. 167–178. Benjamins, Amsterdam-Philadelphia (1991)
21. Hajič, J.: Building a syntactically annotated corpus: The Prague Dependency Treebank. In: E. Hajičová (ed.) Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová, pp. 12–19. Prague Karolinum, Charles University Press (1998)
22. Hajič, J., Smrž, O., Zemánek, P., Šnaidauf, J., Beška, E.: Prague Arabic dependency treebank: Development in data and tools. In: Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools (2004)
23. Haug, D., Jøhndal, M.: Creating a Parallel Treebank of the Old Indo-European Bible Translations. In: Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008) (2008)
24. Hoad, T.C., Zobel, J.: Methods for identifying versioned and plagiarized documents. J. Am. Soc. Inf. Sci. Technol. **54**(3), 203–215 (2003). DOI http://dx.doi.org/10.1002/asi.10170
25. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The sketch engine. In: Proceedings of the Eleventh EURALEX International Congress, pp. 105–116 (2004). URL http://www.fit.vutbr.cz/research/view_pub.php?id=7703
26. Klosa, A., Schnörch, U., Storjohann, P.: ELEXIKO – a lexical and lexicological, corpus-based hypertext information system at the Institut für deutsche Sprache, Mannheim. In: Proceedings of the 12th Euralex International Congress (2006)
27. Kroch, A., Santorini, B., Delfs, L.: Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/ppceme-release-1 (2004)
28. Kroch, A., Taylor, A.: Penn-Helsinki Parsed Corpus of Middle English, second edition. http://www.ling.upenn.edu/hist-corpora/ppcme2-release-2/ (2000)
29. Kühner, R., Stegmann, C.: Ausführliche Grammatik der lateinischen Sprache II. Satzlehre. I. Teile Zweite Auflage. Hahnsche Buchhandlung, Hannover (1914)
30. Lewis, C.T., Short, C. (eds.): A Latin Dictionary. Clarendon Press, Oxford (1879)
31. Liddell, H.G., Scott, R., Jones, H.S., McKenzie, R. (eds.): A Greek-English Lexicon, 9th edition. Oxford University Press, Oxford (1996)
32. Maamouri, M., Bies, A., Buckwalter, T., Mekki, W.: The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In: Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools (2004)
33. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics **19**(2), 313–330 (1994)
34. McDonald, R., Pereira, F., Ribarov, K., Hajič, J.: Non-projective dependency parsing using spanning tree algorithms. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 523–530 (2005)

35. Mel'čuk, I.: Dependency Syntax: Theory and Practice. University of New York Press, Albany (1988)
36. Metzler, D., Bernstein, Y., Croft, W.B., Moffat, A., Zobel, J.: Similarity measures for tracking information flow. In: CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 517–524. ACM, New York, NY, USA (2005). DOI http://doi.acm.org/10.1145/1099554.1099695
37. Passarotti, M.: Verso il Lessico Tomistico Biculturale. La treebank dell'Index Thomisticus. In: P. Raffaella, F. Diego (eds.) Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, Settembre 2006, pp. 187–205. Roma, Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio (2007)
38. Pintzuk, S., Leendert, P.: York-Helsinki Parsed Corpus of Old English Poetry (2001)
39. Rocio, V., Alves, M.A., Lopes, J.G., Xavier, M.F., Vicente, G.: Automated creation of a Medieval Portuguese partial treebank. In: A. Abeillé (ed.) Treebanks: Building and Using Parsed Corpora, pp. 211–227. Kluwer Academic Publishers (2003)
40. Seo, J., Croft, W.B.: Local text reuse detection. In: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 571–578. ACM, New York, NY, USA (2008). DOI http://doi.acm.org/10.1145/1390334.1390432
41. Sgall, P., Hajičová, E., Panevová, J.: The Meaning of the Sentence in its Semantic and Pragmatic Aspects. Dordrecht: Reidel Publishing Company and Prague: Academia (1986)
42. Shivakumar, N., Garcia-Molina, H.: SCAM: A copy detection mechanism for digital documents. In: In Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries (1995)
43. Sinclair, J.M. (ed.): Looking Up: an account of the COBUILD project in lexical computing. Collins (1987)
44. Smyth, H.W.: Greek Grammar. Harvard University Press (1920)
45. Taylor, A., Marcus, M., Santorini, B.: The Penn Treebank: An overview. In: A. Abeillé (ed.) Treebanks: Building and Using Parsed Corpora, pp. 5–22. Kluwer Academic Publishers (2003)
46. Taylor, A., Warner, A., Pintzuk, S., Beths, F.: York-Toronto-Helsinki Parsed Corpus of Old English Prose (2003)
47. Tesnière, L.: Éleménts de syntaxe structurale. Klincksieck, Paris (1959)
48. Zeldes, A., Ritz, J., Lüdeling, A., Chiarcos, C.: Annis: A search tool for multi-layer annotated corpora. In: Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009. (2009)
49. Zemánek, P.: A treebank of Ugaritic: Annotating fragmentary attested languages. In: Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT2007), pp. 213–218. Bergen (2007)

# A Parallel Greek-Bulgarian Corpus: A Digital Resource of the Shared Cultural Heritage

Voula Giouli, Kiril Simov and Petya Osenova

**Abstract** There has been a long tradition in the digitization and manual documentation of cultural heritage data, yet the need for indexing and retrieval that goes beyond mere bibliographic information has only recently been recognized. This chapter reports on completed work aimed at highlighting textual cultural resources that, as of yet, remain under-exploited by creating the necessary infrastructure with the support and customization of Language Technologies (LT). The ultimate goal was to promote the study of cultural heritage of the neighboring areas of Greece and Bulgaria and to raise awareness about their common cultural identity, the focus being on literature, folklore and language. To this end, a bilingual collection of literary and folklore texts in Greek and Bulgarian was developed along with a number of accompanying resources. The authors present the methodology adopted for the automatic annotation of the textual data at various levels of linguistic analysis elaborating on the Greek and Bulgarian text processing tools that are integrated in the cross-lingual search and retrieval mechanisms, and discuss issues and problems encountered in the course of the project life-cycle.

**Key words:** digital collection, bilingual parallel texts, language technology, Greek, Bulgarian

Voula Giouli
Institute for Language and Speech Processing Epidavrou 6 & Artemidos, 15125, Greece, e-mail: voula@ilsp.gr

Kiril Simov
Institute of Parallel Processing, Bulgarian Academy of Sciences, Acad. G. Bonchev 25A, 1113 Sofia, Bulgaria e-mail: kivs@bultreebank.org

Petya Osenova
Institute for Parallel Processing, Bulgarian Academy of Sciences, Acad. G. Bonchev 25A, 1113 Sofia, Bulgaria e-mail: petya@bultreebank.org

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_6, © Springer-Verlag Berlin Heidelberg 2011

## 1 Introduction

In this paper we describe a suite of Language Resources (LRs) that were developed in an effort to highlight cultural assets that, as of yet, remain unexploited to their greatest extent, and at creating the necessary infrastructure for their effective access and visualization with the support of Language Technologies. The ultimate goal of this initiative was to promote the study of cultural heritage of the neighboring Greek and Bulgarian areas and raise awareness about the common cultural identity of the two people, the focus being on literature, folklore and language. The LRs produced in this context, therefore, comprise: (a) a bilingual parallel Greek (EL) and Bulgarian (BG) corpus of texts adhering to the afore-mentioned domains documented and appropriately processed at various levels of linguistic analysis, (b) meta-texts that document the primary textual data and complement the final textual collection, (c) bilingual glossaries (EL-BG) that were semi-automatically extracted from the textual data, and (d) a platform that integrated all the above LRs, facilitates access to them and provides a user-friendly visualization of the cultural heritage data. The LRs collected/developed (being primary texts, meta-texts, and glossaries) are publicly available over the internet to all interested parties, ranging from the research community to laypersons, school students and people interested in finding out more about the particular areas.

## 2 Background

There has been a long tradition in the digitization and manual documentation of cultural heritage data, yet the need for indexing and retrieval that goes beyond mere bibliographic information has only recently been recognized. Moreover, these initiatives are now targeted to wide-spread and complex audiences from researchers to lay-persons with disparate or even competing needs and interests [7]. Users' computer literacy and language barriers often hinder the exploitation of the full potential of the digital medium. One of the main streams in digital libraries projects is the metadata organization and extraction [12]. Various directions have been pursued for accessing and exploring the cultural heritage items also beyond information retrieval techniques. It is worth mentioning the wiki approach to the 19th century encyclopedia on architecture described in [20].

In another direction, a number of initiatives acknowledge that language technology can offer new ways of accessing and visualizing cultural content. This is also true for textual collections or archives, where automatic content annotation and indexing not only facilitates the retrieval processes, but can also give rise to new types of scholarship. Therefore, annotating documents with useful metadata such as person or location names, indexes of events, etc. is currently considered as an important – yet not trivial task. In fact, adding metadata manually is difficult, laborious and costly a task. And, while fully-automatic solutions towards indexing cultural content have been proved neither feasible nor practical, they are

currently used as a valuable aid towards bootstrapping such undertakings ([2–4, 13] etc.). Therefore, the need for language resources tailored to such undertakings is constantly growing.

Additionally, researchers have become aware of the fact that the digitalization of the cultural phenomena requires interdisciplinary attempts to familiarize the humanities and social studies as much as possible with information technologies. In connection with this, several large initiatives have been started in a multicultural context, such as European Cultural Heritage Online (ECHO), Common Language Resources and Technology Infrastructure (CLARIN), Digital Research Infrastructure for the Arts and Humanities (DARIAH), among others.

The work presented here might well be viewed as localization or a showcase of the above mentioned projects and initiatives. The focus is placed on the cultural inheritance of two neighboring countries – Greece and Bulgaria and data are constrained to literary texts, folktales and texts pertaining to the topic of folklore; finally, data selection has been performed on the basis of depicting the similarities between the two cultures in the bordering area.

## 3 The Bilingual Greek–Bulgarian Literary and Folklore Corpus: Selection and Description

### *3.1 Corpus Specifications*

One of the first requirements to be met while collecting the data was the creation of a cultural resource (textual collection) that would be appropriate for studying cultural similarities and/or differences between the neighboring areas of Greece and Bulgaria. To this end, three types of cultural content were deemed appropriate for inclusion: literary works, folktales and legends, as well as texts presenting the customs, rituals, everyday-life habits of people in the eligible areas (folklore texts). Indeed, the literary production in the eligible areas normally depicts in an unsolicited manner cultural aspects that would be of interest to the target users. More specifically, novels, poetry, etc. written by authors who were born and/or lived in the Greek and Bulgarian cross-border area of Thrace, and also literary works about Thrace or situated in Thrace were deemed ideal candidates. Moreover, studies, essays, and other folklore texts about Thrace, witnessing not only cultural but also linguistic elements of the two cultures were also considered for inclusion. Finally, folktales and legends, the former being the intermediate between literature and folklore, were also expected to provide a setting for the comparative study of cultural elements either shared by the two people or unique to each one.

To render the primary collection of texts meaningful to a wide audience, another requirement to be met was that the texts, especially those pertaining to the type literature, would be further documented with meta-texts, i.e., literary criticism, biographies of the selected authors, etc.

Finally, translations of primary data (literary/folklore texts, and folktales) from Greek to Bulgarian and vice versa were also deemed appropriate for inclusion. The rationale was that even users with little or no knowledge of one of the languages at hand would be able to read or simply browse the collection.

To cater for the above-mentioned requirements, an extensive research was conducted in order to record and roadmap the literary production spanning from the 19th century till the present days along with written records on folk culture and folktales from the eligible areas. Available translations of literary works were also researched. Both printed and digitized sources were exploited, i.e., (on-line and printed) anthologies of Bulgarian, Greek or Balkan literature, digital archives, web resources and library material. The outcome of this research was a wealth of literary works including titles by the most prominent authors in Greece (e.g. M. Loudemis, K. Varnalis, G. Xanthoulis, etc) and Bulgaria (Dalchev, A. Donchev, V. Mutafchieva, K. Topalov, etc.). The selection of the authors, who would finally participate in the cultural collection, was based on a range of criteria, as for example, the candidate author's impact to Greek or Bulgarian literature respectively, his overall contribution in culture or other major sectors such as journalism and education, etc.

Textual material representative of the two cultures and also suitable for their comparative study was selected from the resulting pool of candidate texts, according to a set of predefined criteria. In this way, texts which demonstrate the two people's cultural similarities and affinity were chosen, along with each author's most important and representative works. To ensure corpus "representativeness" to some extent, we tried to include the full range of the literary texts (poetry, fiction, short stories) and in proportion to the literary production with respect to the parameters of place, time and author. In this way, we think we have avoided biases and the corpus models all language varieties spoken in the areas and at different periods.

Finally, text selection was also determined to a great extent by such factors as the availability of a translation from the source to the target language as well as legal issues relevant to Intellectual Property Rights.

## *3.2 Collection Description*

Along with the aforementioned lines, the collection consists of bilingual parallel EL–BG literary and folklore texts. The collection of primary data currently comprises 135 literary works, 70 BG (Bulgarian) and 65 EL (Greek). Moreover, 30 BG folk texts and 30 EL folk texts along with 25 BG folktales and 31 EL folktales were added in order to build a corpus as balanced as possible and representative of each country's culture. In terms of tokens, the corpus amounts to 700,000 in total (circa 350,000 tokens per language): the literature part is about 550,000 tokens, whereas the folklore and legend sub-corpus is about 150,000 tokens. The collection covers EL and BG literary production dating from the 19th century until the present day, and also texts (both literary or folklore) that are written in the dialect(s) used in the eligible areas. This, in effect, reflects the language varieties represented in the textual

collection that range from contemporary to non-contemporary, and from normal to dialectical or even mixed language.

As already mentioned, the collection of primary data was also coupled with accompanying material (content metadata) for each literary work (literary criticism) and for each author (biographical information, list of works, etc.). Texts about the common cultural elements were also included.

Following text selection, digitization and extended manual validation were performed where appropriate. Normalization of the primary data was kept to a minimum so as to cater, for example, for the conversion from the Greek polytonic to the monotonic encoding system.

### *3.3 Metadata Descriptions*

Metadata descriptions and linguistic annotations were consequently added in order to serve a two-fold purpose: (a) indexing and retrieval, and (b) further facilitating the comparative study of textual data via the implemented platform. Further interoperability with other initiatives/tools was also taken into account while setting the relevant specifications. To this end, metadata descriptions and linguistic annotations compliant with internationally accepted standards were added to the raw material. The metadata scheme deployed in this project is compliant with internationally accredited standards with certain modifications that cater for the peculiarities of the data.

More specifically, the metadata scheme implemented in this project builds on the Text Encoding Initiative[1] which has served as the basis for almost all encoding initiatives (such as, for example, CLARIN). Metadata elements have been deployed which encode information necessary for text indexing with respect to text title, author, publisher, publication date, etc. (bibliographical information). Additionally, to ensure documentation completeness, and facilitate the inter-relation among primary data and the accompanying material (biographies, criticism, etc) the documentation scheme has been extended so as to include these elements (AuthorBio, etc). Information regarding text type (poetry, novel, folktale, etc.), language variety (contemporary / non-contemporary / idiomatic), and topic has also been retained, the latter being applicable to folklore texts and folk tales.

More specifically, topic definition for folktales has been based on the widely accepted Aarne-Thompson classification system [1]: (a) Animal Tales, (b) Tales of Magic, (c) Religious Tales, (d) Realistic Tales (Novelle), (e) Tales of the Stupid Ogre/Giant/Devil, (f) Anecdotes and Jokes, (g) Formula Tales. Folklore texts have been classified on the basis of the Library of Congress Classification scheme[2] simplified in order to cater for the data at hand (Table 1). Due to the small amount

[1] See http://www.tei-c.org, Text Encoding Initiative (TEI Guidelines for Electronic Text Encoding and Interchange).

[2] http://www.loc.gov/catdir/cpso/lcco/

of the selected data and the limited scope of the whole project we opted for a coarse classification scheme for both folktales and folklore texts.

**Table 1** Folklore texts classification scheme

| Material Life | Social Life | Spiritual Life |
|---|---|---|
| housing and architecture | Birth–marriage–death | folk faith and rituals |
| clothing | family life | superstitions |
| food and nutriment | community structure and administration | supernatural creatures |
| bucolic and rural life | feasts and festivals | magic |
| agriculture | customary law | folk medicine |
| | | folk language |
| | | (baptismal and family names, toponyms, nicknames) |

The encoding of the parallel texts has been based on the XCES, the XML version of the Corpus Encoding Standard (XCES),[3] and CES,[4] which has been proposed by EAGLES[5] and is compliant to the specifications of the TEI. From the total number of elements proposed by these guidelines, the annotation of the parallel corpus at hand has been restricted to the recognition of structural units at the sentence level, since this is necessary for the alignment processes. Alignments at the sentence level have been stored in files conformant to the internationally accredited TMX standard (Translation Memory eXchange),[6] which is an XML-compliant, vendor-neutral open standard for storing and exchanging translation memories created by Computer Aided Translation (CAT) and localization tools.

The external structural annotation (including text classification) of the corpus also adheres to the IMDI metadata scheme [11]. IMDI metadata elements for catalogue descriptions [10] were also taken into account to render the corpus compatible with existing standards (ELRA, and LDC). This type of metadata descriptions was added manually to the texts.

The aforementioned metadata descriptions are kept separately from the data in an XML header that is to be deployed by the web interface for search and retrieval purposes.

## 4 Text Annotation and Processing

As it has already pointed out, to further enhance the capabilities/functionalities of the final application, thus rendering the collection a useful resource to prospective users and researchers, further annotations at various levels of linguistic analysis

[3] http://www.cs.vassar.edu/XCES/

[4] http://www.cs.vassar.edu/CES/CES1-0.html

[5] http://www.ilc.cnr.it/EAGLES96/home.html

[6] http://www.lisa.org/tmx/

were integrated across two pillars: (a) semi-automatic indexing and retrieval; and (b) further facilitating the comparative study of textual data by means of bilingual glossaries which were constructed semi-automatically and via the visualization of aligned parallel texts.

Text processing at the monolingual level comprises the following procedures: (a) handling and tokenization, (b) Part-of-Speech (POS) tagging and lemmatization, (c) surface syntactic analysis, (d) indexing with terms/keywords and phrases/Named Entities (NEs) pertaining to the types Location (LOC) and Person (PER). Finally, alignment of parallel texts (primary source documents and their translations) has also been performed at both sentence and phrase level. As expected, poems posited the major difficulties due the fuzziness in identifying sentence boundaries, and alignments at the phrase level were favored instead.

Annotations at these levels were added semi-automatically, by deploying existing generic Natural Language Processing (NLP) tools that were developed for the languages at hand, whereas extensive and intensive validations were performed via several ways. Indeed, although the tools deployed have reported to achieve high accuracy rates in the domains/genres they were intended for, the specific nature of the data led to a significant reduction. To this end, half of the annotations were checked manually. After the identification of the errors in this part of the corpus, we have performed a manual check in the second part of the corpus only for these cases which were recognized as errors during the validation of the first part. For some of the cases relevant constraints in the systems were written, which automatically find places where some rules were not met. Tools customization was also performed by adding new rules applicable for the language varieties to be handled, and also by extending/modifying the resources used (word and name lists, etc.). In what follows the Greek and Bulgarian Text Processing Components will be described.

## *4.1 The Greek Pipeline*

Text processing of the EL data was applied via an existing pipeline of shallow processing tools for the Greek language that were developed at the Institute for Language and Speech Processing. These include:

Handling and tokenization: following common practice, the Greek tokenizer makes use of a set of regular expressions, coupled with precompiled lists of abbreviations, and a set of simple heuristics [15] for the recognition of word and sentence boundaries, abbreviations, digits, and simple dates.

POS-tagging and lemmatization: a tagger that is based on Brill's TBL architecture [6], modified to address peculiarities of the Greek language [16] was used in order to assign morphosyntactic information to tokenized words. Furthermore, the tagger uses a PAROLE-compliant tagset of 584 different part-of-speech tags. Following POS tagging, lemmas are retrieved from a Greek morphological lexicon.

Surface syntactic analysis: the Greek chunker is based on a grammar of 186 rules [5] developed for the automatic recognition of non-recursive phrasal categories: adjectives, adverbs, prepositional phrases, nouns, verbs (chunks) [15].

Term extraction: a Greek Term Extractor was used for spotting terms and idiomatic words [8]. Term Extractor's method proceeds in three pipelined stages: (a) morphosyntactic annotation of the domain corpus, (b) corpus parsing based on a pattern grammar endowed with regular expressions and feature-structure unification, and (c) lemmatization. Candidate terms are then statistically evaluated with an aim to skim valid domain terms and lessen the overgeneration effect caused by pattern grammars (hybrid methodology).

Named Entity Recognition was then performed using MENER (Maximum Entropy Named Entity Recognizer), a system compatible with the ACE (Automatic Content Extraction) scheme, catering for the recognition and classification of the following types of NEs: person (PER), organization (ORG), location (LOC) and geopolitical entity (GPE) [9].

### *4.2 NLP Suite for Bulgarian*

In the processing of the Bulgarian part of the corpus we are using the language technology tools developed within BulTreeBank project.[7] Within this project we have to tune the tools to the specific language types – diachronic and area related text. Some of the literary works are written in 19th century or the beginning of 20th century and their language reflect the writing standards of the corresponding period. Also some words with higher distribution in the target regions appear in some of the works. In order to deal with them we had to extend the used lexicons, create a guesser for the unknown words and add new rules to the chunk grammar to cope with some specific word order within the texts. What follows is a list of tools that we have used. They are implemented within CLaRK System [19].

Tokenization: There is a hierarchy of tokenizers within the CLaRK system, which tokenize the texts in an appropriate way. Additionally, one can decide what the category of the token is and to assign it.

A Morphosyntactic analyzer for Bulgarian assigns then all possible analyses to the word tokens. It is based on a morphological lexicon which covers the grammatical information of about 100,000 lexemes (1,600,000 word forms); gazetteers of about 25,000 names and 1,500 abbreviations. The analyzer is compiled as a set of regular grammars within CLaRK. It assigns all the possible analyses to the tokens. The analyses are encoded as position based tags within the BulTreeBank tagset.[8] In the places where competing analyses arise between a common word and a name or an abbreviation, we try to use the token classification strategy and the prompts of the context. If there is no clear preference, we leave the decision to the human

[7] http://www.bultreebank.org

[8] See http://www.bultreebank.org/TechRep/BTB-TR03.pdf

annotator or to the morphosyntactic disambiguator described below. On the basis of the lexicons and the list of unknown word forms within the corpus we have created guesser for the unknown words. Because the inflection in Bulgarian could be homonymic in a great degree, the guesser in many cases assigns more that one tags. These ambiguous cases are resolved on the next stage of processing.

MorphoSyntactic Disambiguation: We have already implemented a rule-based morpho-syntactic disambiguator, encoded as a set of constraints within the CLaRK system. This rule-based disambiguator exploits context information like agreement between an adjective and a noun in a noun phrase, specific positions like a noun after a preposition, but it also deals with some fixed phrases. The disambiguator does not try to solve unsure cases, but leaves them for further processing. Its coverage is about 80%. For next step of disambiguation we have developed a neural-network-based disambiguator [19]. It achieves an accuracy of 95.25% for part-of-speech and tagging 93.17% for complete morpho-syntactic disambiguation.

Lemmatization: We have implemented a functional lemmatization based on the morphological lexicon mentioned above. The functions are defined via two operations on word forms: remove and concatenate. The rules have the following form:

**if** *tag* = Tag **then** {*remove* OldEnd; *concatenate* NewEnd}

where Tag is the tag of the word form, OldEnd is the string which has to de removed from the end of the word form and NewEnd is the string which has to be concatenated to the beginning of the word form in order to produce the lemma. Here is an example of such a rule:

if tag = Vpitf-o1s then remove ox; concatenate a

The application of the rule to the verb form четох (remove: ox; concatenate: a) gives the lemma чета.

In order facilitate the application of the rules we attach them to the word forms in the lexicon. In this way we gain two things: (1) we implement the lemmatization tool as a regular grammar in CLaRK and (2) the level of ambiguity is less than 2%. Additionally we encode rules for unknown words in the form of guesser word forms: #ox and tag=Vpitf-o1s.

Partial Grammars

We have constructed grammars for:

- *Sentence splitting.* At the moment it is fully automated and reliable only for the basic and clear cases. For solving complex and ambiguous cases this grammar is combined with supporting modules for abbreviation detection.
- *Named-entity recognition.* Identifying numerical expressions, names, abbreviations, special symbols. They are designed to work in cooperation with the morphosyntactic analyzer. If necessary, the grammars can overwrite the analysis of the morphosyntactic analyzer.
- *Chunking.* Generally speaking, the chunking process conforms to the following requirements: it deals with non-recursive constituents; relies on a clear-indicator strategy; delays the attachment decisions; ignores the semantic information; aims

at accuracy, not coverage. Besides an NP chunker, there are chunk grammars for APs, AdvPs, PPs and some non-problematic clauses.

This level annotation is used for facilitating the alignment of the corpus and glossary creation.

### *4.3 Sentence Alignment*

The comparative study of parallel documents is facilitated through the visualization of alignments at the sentence level. These were performed semi-automatically by means of an Alignment tool developed in ILSP [14, 17]. The tool which has been trained on a corpus of legal texts (CELEX) is language independent and uses surface linguistic information coupled with information about possible unit delimiters depending on the level at which the alignment is sought. The alignments acquired automatically were hand-validated by expert linguists and translators.

## 5 Tools Customization and Metadata Harmonization

Annotation is performed by several means. Half of the annotations were checked manually. After identification of the errors in this part of the corpus we have performed manual check in the second part of the corpus only for these cases which were recognized as errors during the validation of the first part.

As already stated, the tools that were deployed for the linguistic processing are generic ones that were initially developed for different text types/genres. Moreover, the data at hand posed another difficulty, that is, coping with older/obsolete language usage. In fact, some of the literary works were written in the 19th century or the beginning of 20th century, and their language reflects the writing standards of the corresponding period.

Therefore, as expected, the overall performance of the afore-mentioned tools was lower than the one reported for the texts these tools were initially trained for. Performance at POS-tagging level dropped from 97% to 77% for the Greek data since no normalization of the primary data was performed. Processing at the levels of chunks and NEs were even lower.

On the other hand, the BG morphological analyzer coverage, whose benchmark performance is 96% dropped to 92% on poems and folktales and to 94% on literary texts and legends. The reason was that the language of processed literary texts and legends came normalized from the sources, while the poems and folktales kept some percentage of archaic or dialect words. Thus, additionally to the guesser, a post POS processing was performed on the unknown words. Moreover, the accuracy of the neural network disambiguator and the rule-based one was 97%, i.e. the same as for other applications.

Within the project we had to tune the tools to the specific language types, such as diachronically remote texts and domain specific texts (folklore). Also, some words with higher distribution in the target regions appear in some of the works. In order to deal with them we had to extend the used lexicons, to create a guesser for the unknown words and add new rules to the chunk grammar to handle some specific word order within the texts.

To cater for shortcomings in the processing of Bulgarian we have written constraints in the CLaRK system which automatically find places where some rules are not met. After this check we estimate the level of error to be less than 3% in the automatic processed part of the corpus.

Additionally, the deployment of tools that are specific to each language and compatible with completely distinct annotation standards brought about the issue of metadata harmonization. To this end, although the Greek tools were developed to conform to the afore-mentioned annotation standards, this was not the case for Bulgarian. The first encoding scheme followed the BulTreeBank morphological and chunk annotation scheme. Afterwards, the information was transferred into the project scheme in order to be consistent with the Greek data and applicable for web representation. As a result, the morphosyntactic features of the BG tagset, which is a more specialized version of the Multext-East tagset were mapped onto the relative PAROLE tags.

## 6 Bilingual Glossaries

The bilingual corpus is accompanied by a suite of glossaries that were semi-automatically extracted from it, in order to assist not only in the comparative study of the texts but also in the cross-lingual retrieval. The following glossaries are developed: (a) location names semi-automatically identified in the texts; (b) terms extracted from the folklore texts; (c) words idiosyncratic to the writer(s), idiomatic and/or dialectical, that were selected manually from literary texts; (d) the most frequent lemmas in the corpus. A total of 3700 entries are included in the resulting resource.

Each entry is coupled with appropriate information encoded by expert lexicographers and language experts. Part-of-speech obtained from the tagger process and alternative forms are assigned to all entries. Additionally, on the basis of the alignments at the sentence level, translations from the source to the target language are also provided. It should be noted here that lemmas/words corresponding to different senses are listed as separate entries and are assigned the relative translational equivalents. Definitions in both languages are also retained for entries that are idiomatic or specific to the texts, language variety, etc.; semantic relations such as synonymy and hypernymy are also encoded to a small subset of the glossary entries that pertain to the folklore domain. Finally, dialectical entries are appropriately marked.

## 7 Content Management

All the data collected (being the primary literary or folklore texts or meta-documents, etc.) along with their translations, the multi-layered annotations, and the resulting glossaries were integrated in a database platform that was developed to serve as a content management system. Being the backbone of that platform, the metadata material facilitates the interlinking of similar documents, and the access to the primary data via the web. To this end, a specially designed web site was developed to satisfy the needs of end-users (the general public and the special groups of researchers and other scientists). The website features a trilingual interface (Greek, Bulgarian, English) as well as advanced search and retrieval mechanisms on the entire bilingual content or a user-specified part of it. The users can perform combined searches by author name, title, genre, etc. Furthermore, they can search for single keywords/wordforms or for two wordforms that can be a user-specified number of words apart from each other. Searches by lemma and/or by phrase have been also implemented. The latter rely on a matcher, which tries to link the query word(s) with the stored lemmas/wordforms. Additionally, a stemmer for Greek and Bulgarian has been used for the on-line stemming of queries, which will then be matched with the already stemmed corpus. When all the above fails, fuzzy matching techniques are being employed, facilitating, thus, effective query expansion functionality. Finally, apart from wordforms and lemmas, the collection can also be queried for morphosyntactic tags or any combination thereof; results, then, come in the form of concordances and statistics (frequency information), hence the relative document(s) can also be retrieved. Moreover, users can search the whole corpus or define a sub-corpus based on the classification and annotation parameters accompanying each text, thus creating sub-corpora of a specific author, or belonging to a specific genre, text type, domain, time period, etc.

In addition, the web interface allows the users to simultaneously view on screen both Greek and Bulgarian texts, aligned and in parallel, so that to become acquainted with the comparative aspects of the two languages or perform specific linguistic, lexicographic or translation tasks. Alternatively, the user can consult the bilingual glossary of terms and the aligned list of NEs. The latter is often very interesting, especially with respect to Location entities, since transliteration is usually not adequate.

The design of the web interface effectively blends simplicity and advanced functionality so that to fully support the intended usage scenarios (comparative study of literary and folklore texts equally by specialists, laymen or students, language and/or literary teaching and learning, lexicographic projects, etc.). Finally, the web interface has been enhanced by integrating last generation of synthetic speech technology for both Greek and Bulgarian. This speech-enhanced user interface [18] offers innovative web accessibility for blind and vision impaired Greek and Bulgarian users as well as for other users who use speech as their preferable modality to information access. The key feature of this web-speech technology is that it lets users to interact with the underlying system; so that they can hear only the portions of a specific web page they are interested in, being able

at the same time to navigate through the entire web site and visit only the web pages of their choice.

## 8 Conclusions

We have described work targeted at the promotion and study of the cultural heritage of the cross-border regions of Greece – Bulgaria, with a focus on literature, folklore and language of the two people, by means of modern and technologically advanced platforms. To this end, a digital collection of literary and folklore texts has been compiled along with accompanying material selected from various (online and printed sources), which is integrated into a platform with advanced search and retrieval mechanisms.

However, the cultural value of the bilingual cultural Greek-Bulgarian corpus goes beyond the border areas that it was intended for, because it shows the similarities and the differences between the two neighboring countries. More specifically, it can be used for supporting the study of the other language in both countries. Also, it can be explored for comparing the cultural and social attitudes in diachronic depth and genre variety. Apart from the usages from a humanities point of view, the corpus can become a good base for testing taggers, parsers and aligners. It would especially challenge the processing of the regional dialects, the language of poems, and the language of non-contemporary works.

Future work is being envisaged in the following directions: extending the corpus with more texts, and respectively the glossaries – with more terms, adding more layers of linguistic analysis (predicate-argument structure, etc.), and further enhance search and retrieval with the construction and deployment of an applicable thesaurus.

**Acknowledgements** The work presented here was conducted in the framework of a project funded under the Community Initiative Programme INTERREG III A / PHARE CBC Greece – Bulgaria. The project was implemented by the Institute for Language and Speech Processing (ILSP, http://www.ilsp.gr) and a group of researchers from the Bulgarian Academy of Sciences, (http://www.bultreebank.org/).

## References

1. Aarne, A.: The Types of the Folktale: A Classification and Bibliography., 2nd rev. ed. edn. Suomalainen Tiedeakatemia / FF Communications, Helsinki (1961). Translated and Enlarged by Stith Thompson.
2. Bontcheva, K., Maynard, D., Cunningham, H., Saggion, H.: Using human language technology for automatic annotation and indexing of digital library content. In: Proc. of the 6th European Conference on Research and Advanced Technology for Digital Libraries., *Lecture Notes In Computer Science*, vol. 2458, pp. 613–625 (2002)

3. Borin, L., Forsberg, M., Kokkinakis, D.: Diabase: Towards a diachronic BLARK in support of historical studies. In: Proc. of LREC (2010)
4. Borin, L., Kokkinakis, D., Olsson, L.J.: Naming the past: Named entity and animacy recognition in the 19th century swedish literature. In: Proc. of the ACL Workshop: Language Technology for Cultural Heritage Data (LaTeCH.)., pp. 1–8. ACL, Prague (2007)
5. Boutsis, S., Prokopidis, P., Giouli, V., Piperidis., S.: A robust parser for unrestricted greek text. In: Proc. of the 2nd Language and Resources Evaluation Conference, pp. 467–473. Athens, Greece (2000)
6. Brill, E.: A corpus-based approach to language learning. Ph.D. thesis, University of Pennsylvania (1997)
7. Crane, G.: Cultural heritage digital libraries: Needs and components. In: Proc. of the 6th European Conference on Research and Advanced Technology for Digital Libraries., *Lecture Notes In Computer Science*, vol. 2458, pp. 51–60 (2002)
8. Georgantopoulos, B., Piperidis, S.: Term-based identification of sentences for text summarization. In: Proceedings of LREC2000 (2000)
9. Giouli, V., Konstandinidis, A., Desypri, E., Papageorgiou., H.: Multi-domain multi-lingual named entity recognition: Revisiting & grounding the resources issue. In: Proceedings of LREC 2006 (2006)
10. IMDI: Metadata elements for session descriptions, version 2.1 (June 2001)
11. IMDI: Metadata elements for session descriptions, version 3.0.4 (Sept. 2003). http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf. Accessed 22.01.2007.
12. Liddy, E.D., Allen, E., Harwell, S., Corieri, S., Yilmazel, O., Ozgencil, N., Diekema, A., McCracken, N., Silverstein, J., Sutton, S.: Automatic metadata generation & evaluation. In: The 25th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002), pp. 401–402. Tampere, Finland (2002)
13. Nissim, M., Matheson, C., Reid, J.: Recognizing geographical entities in scottish historical documents. In: Proc. of the Workshop on Geographic Information Retrieval at SIGIR 2004 (2004)
14. Papageorgiou, H., Cranias, L., Piperidis., S.: Automatic alignment in parallel corpora. In: Proceedings of ACL 1994 (1994)
15. Papageorgiou, H., Prokopidis, P., Giouli, V., Demiros, I., Konstantinidis, A., Piperidis, S.: Multi-level XML-based corpus annotation. In: Proceedings of the 3nd Language and Resources Evaluation Conference (2002)
16. Papageorgiou, H., Prokopidis, P., Giouli, V., Piperidis, S.: A unified pos tagging architecture and its application to greek. In: Proceedings of the 2nd Language and Resources Evaluation Conference, pp. 1455–1462. Athens, Greece (2000)
17. Piperidis, S.: Interactive corpus based translation drafting tool. In: ASLIB Proceedings, vol. 47(3) (1995)
18. Raptis, S., Spais, I., Tsiakoulis., P.: A tool for enhancing web accessibility: Synthetic speech and content restructuring. In: Proc. HCII 2005: 11th International Conference on Human-Computer Interaction. Las Vegas, Nevada, USA (2005)
19. Simov, K., Osenova, P.: A hybrid system for MorphoSyntactic disambiguation in Bulgarian. In: Proc. of the RANLP 2001 Conference, pp. 288–290. Tzigov Chark, Bulgaria (2001)
20. Witte, R., Gitzinger, T., Kappler, T., Krestel, R.: A semantic Wiki approach to cultural heritage data management. In: Language Technology for Cultural Heritage Data (LaTeCH 2008), Workshop at LREC 2008. Marrakech, Morocco (2008)

# Part IV
# Personalisation

# Authoring Semantic and Linguistic Knowledge for the Dynamic Generation of Personalized Descriptions

Stasinos Konstantopoulos, Vangelis Karkaletsis, Dimitrios Vogiatzis and Dimitris Bilidas

**Abstract** We present the ELEON/NATURALOWL system, an application of Semantic Web and Natural Language Generation technologies that combines a conceptual representation of cultural heritage objects with linguistic and adaptation resources. This combined model is used to automatically generate multi-lingual and personalized textual descriptions of cultural heritage objects represented as instances of an OWL domain ontology annotated by RDF linguistic and adaptation resources. Metadata and annotations are created using an authoring environment, which considerably reduces the effort required to port the system to a new domain.

**Key words:** semantic web technologies, personalized natural language generation

## 1 Introduction

Cultural heritage organizations create and maintain repositories comprising extensive metadata about cultural objects, such as artifacts, artists, and locations. The repositories are typically used to catalogue, index, and classify the cultural content, for the purpose of providing semantic *searching* and *browsing* facilities to professional users as well as to the general public. A further opportunity, however, which we consider in this chapter, is to automatically generate *textual descriptions* of cultural objects from a repository, descriptions which are customized to a *variety of audiences*, in *multiple languages*, and which serve *different presentation objectives*.

In this chapter, we present the combined ELEON/NATURALOWL system and how it links a *conceptual representation* of cultural heritage objects (the *domain*) with the *linguistic and adaptation resources* necessary to realize elements of this

All authors are at the
Institute of Informatics & Telecommunications, NCSR 'Demokritos',
Ag. Paraskevi GR-15310, Athens, Greece
email: {konstant,vangelis,dimitrv,dbilid}@iit.demokritos.gr

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_7, 

representation as personalized text. ELEON is an authoring environment for creating domains in the form of OWL ontologies, OWL being the standard formalism to specify ontologies on the Semantic Web [4], as well as the linguistic and adaptation resources.[1] NATURALOWL is a *natural language generation* (NLG) engine that exploits the linguistic and adaptation resources authored via ELEON to dynamically produce texts from OWL ontologies.[2] The advantages of this approach are manifold:

- OWL ontologies constitute machine-readable and reusable models of the cultural repository's content. Besides supporting natural language generation, such models can be used for the semantic indexing and searching of the repository. This can also be seen from the reverse perspective: the natural language descriptions can be derived from existing conceptual models originally created for the purpose of semantic indexing and searching.
- The conceptual representations are realized as texts using reusable linguistic modules with clearly separated and configurable domain-dependent linguistic and profiling resources. By clearly separating the conceptual representations from the linguistic modules and their domain-dependent resources, the same conceptual descriptions can be realized in different languages and the same linguistic modules can be used to realize conceptual representations of different repositories.
- The dynamic generation of textual descriptions is driven by user adaptation resources that personalize the descriptions for different audiences, but also adapt them to different contexts and situations. Furthermore, ELEON invokes automatic reasoning systems that assist the author by automatically inferring missing profile parameters, alleviating the burden of explicitly providing all necessary parameters for large numbers of objects and audience types.

Although the system can be used in a variety of domains and human-computer interaction applications, it is particularly pertinent to cultural heritage content, especially when the content has to be presented in different situations and interaction contexts (e.g., via portable devices or remotely over the Web) and to audiences with wide ranges of age groups, levels of expertise, cultural and educational backgrounds.

In the rest of this chapter, we first focus our attention to the representation used by the ELEON/NATURALOWL system, including representing the domain (Sect. 2) and the linguistic and adaptivity annotations that complement it (Sect. 3). We then proceed to discuss how these are used by NATURALOWL to generate descriptions of the objects of the domain (Sect. 4) and how ELEON facilitates the creation of these resources (Sect. 5). The chapter closes with comparison to previous approaches (Sect. 6) and some concluding remarks (Sect. 7).[3]

[1] ELEON was developed at the Institute of Informatics and Telecommunications, NCSR 'Demokritos' and is publicly available as open-source software; see http://www.iit.demokritos.gr/~eleon for more information.

[2] NATURALOWL was developed by the Natural Language Processing Group, Department of Informatics, Athens University of Economics and Business. It is publicly available as open-source software; see http://nlp.cs.aueb.gr/software.html for more information.

[3] Section 4 is based on text by Ion Androutsopoulos and Gerasimos Lampouras, adapted and used with their permission. Please see also Acknowledgements.

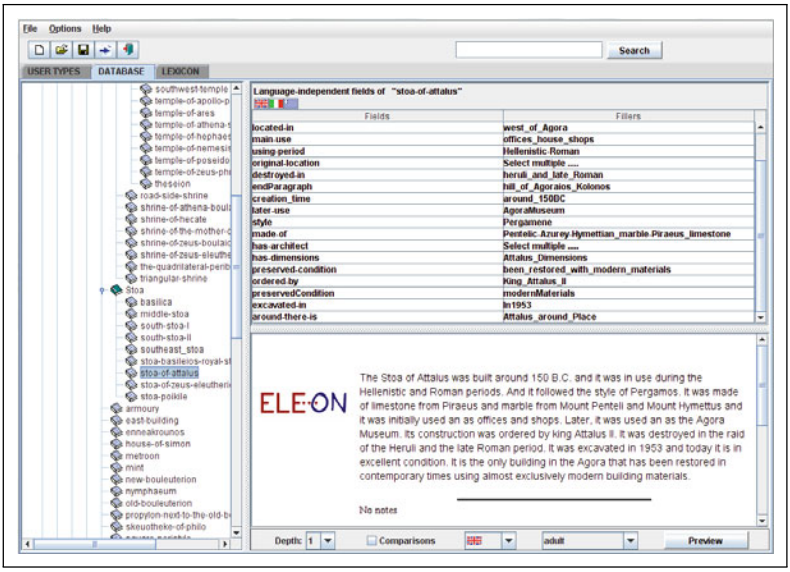


**Fig. 1** ELEON screenshot showing the concept hierarchy of the ontology and the instances of each concept (left), the fields of the currently selected instance (right top), and a preview of the textual description generated for the instance (right bottom). The preview language and profile can be seen on (and selected from) the bar at the bottom of the screen.

## 2 Authoring Domain Ontologies

ELEON assumes that its users, called *authors*, are persons who have domain expertise, but not expertise in knowledge representation and NLG. It helps them configure the system for a new application domain by defining the domain ontology, as well as the domain-dependent language and adaptation resources.

The language resources allow the NLG engine to turn facts of the ontology into coherent natural language texts, whereas the adaptation resources are consulted to adapt the generated texts to each visitor's preferences and presumed background knowledge, but also to the interaction goals set by the author. ELEON also enables the authors to generate text previews using the NLG engine, in order to examine the effect of their updates to the domain ontology, the language, and the adaptation resources. In this section, we focus on the functionality that ELEON provides to create or edit the domain ontology. The linguistic and adaptation resources are discussed in the following sections.

ELEON assumes that the domain ontology encodes knowledge in the form of *entity types* (concepts), *entities* (instances of concepts), *relations* between entities, and *attributes* that connect entities to datatype values; this assumption is compatible with OWL ontologies, one of the types of ontologies supported by ELEON, where the corresponding terms are *classes*, *individuals*, *object properties*, and *datatype properties*. Figure 1 illustrates part of a domain ontology that encodes knowledge

about the Ancient Agora of Athens. This ontology was used in the INDIGO project, where ELEON/NATURALOWL was embedded in a mobile robot acting as a guide in a cultural centre; in that setting, monuments of the Agora were displayed on wall-mounted screens.[4]

ELEON ontologies encode domain knowledge in the form of *entity types* (concepts), *entities* (instances), and relations between them. Figure 1 illustrates part of such an ontology that encodes knowledge about the Ancient Agora of Athens. This ontology was used in the INDIGO project to implement a use case where the system guides visitors through an exhibition on the Ancient Agora of Athens, introducing the buildings to them before they attend a virtual 3D tour of the Agora hosted at the Foundation of the Hellenic World. The examples used in this paper are drawn from this domain.

In the example of Fig. 1, stoa-of-attalus is an instance of the entity type Stoa, a sub-type of Building, which is in turn a sub-type of ArchitecturalConstruction, a sub-type of PhysicalObject. Attributes and relationships are expressed using fields. At any entity type, it is possible to introduce new fields, which then become available to all the entities that belong to that type and its subtypes. In Fig. 1, the field locatedIn was introduced at the ArchitecturalConstruction entity type, and it was defined (previously, not shown in the screenshot) as expressing a relationship between ArchitecturalConstruction entities and Place entities. Similarly, the using-period field introduced at the PhysicalObject entity type, as expressing a relationship between PhysicalObject entities and Period entities; consequently all entities of type PhysicalObject and its subtypes, including, e.g., ArchitecturalConstruction and ArtObject, inherit the using-period field.

OWL domains ontologies can be created in ELEON from scratch, or by importing and editing existing OWL ontologies; this facilitates the use of well-established conceptual models in the cultural heritage domain. The CIDOC Conceptual Reference Model (CRM), for example, is available as an OWL ontology.[5] Work is in progress to create linguistic resources and to, in general, extend support for CIDOC CRM beyond the current capability of defining a domain with the CIDOC CRM framework. Most other cultural heritage vocabularies, thesauri, and classification schemes that use XML or relational database data models are compatible with the Simple Knowledge Organization System (SKOS) and can be automatically converted to ontologies.[6] For the Ancient Agora ontology, we have adopted the *Generalized Upper Model* as ontological foundation, a general, task and domain-independent upper level ontology, geared towards defining ontological classes appropriate for flexible expression in natural language.[7]

[4] See http://www.ics.forth.gr/indigo for more information about INDIGO. A video of the robotic guide in action is available at http://www.youtube.com/watch?v=qCzBx4LzGak.

[5] See http://cidoc.ics.forth.gr/official_release_cidoc.html.

[6] See http://www.w3.org/2004/02/skos/ about SKOS. A variety of tools exist for converting SKOS data models to, or aligning them with, ontological models. See, for example, http://www.heppnetz.de/projects/skos2gentax/ and http://annocultor.sourceforge.net/.

[7] Please see http://www.fb10.uni-bremen.de/anglistik/langpro/webspace/jb/gum for more information.

## 3 Description Adaptation

Besides modelling the cultural heritage domain itself, ELEON supports annotating the objects, classes, and properties of the domain with adaptation and linguistic information. Such information is used by NLG engines to (a) plan the description that will be generated, adapting it to the current audience and circumstance, and (b) realize the planned description in a particular language.

*Realization* is based on clause plans (micro-plans) that specify how an ontological property can be expressed in each supported natural language. The author specifies the clause to be generated in abstract terms, by specifying, for example, the verb to be used, the voice and tense of the resulting clause, etc. Similar annotations for instances and classes specify how they should be realized as noun phrases that fill slots in the property-generated clauses. Micro-plan annotations also comprise several other language-specific parameters, such as whether the resulting clause can be aggregated into a longer sentence or not, its voice and tense, and so on, as described in more detail by [3, Sect. 3].

*Adaptive planning*, on the other hand, operates at the abstract level and does not involve specifics of the target language. It is rather aimed at reflecting a *synthetic personality* in the description, as well as *personalizing* it for a particular audience. Adaptation parameters are provided in the form of *profile attributes* that control aspects of the text plan such as how many and which of the facts known about an object should be used to describe it, as discussed in more detail below.

### *3.1 Personalization and Personality*

The system supports authoring the *adaptation profiles* that control the *dynamic adaptation* of the generated descriptions. Profiles permit the author to specify, for example, that technical vocabulary be used when generating for experts, or that shorter and simpler sentences are generated for children, but also to gear the system towards achieving different interaction goals, i.e., aiming at the assimilation by the visitor of certain (types of) facts are assimilated. Adaptivity is achieved by providing a variety of generation parameters though adaptation profiles, including a numerical *interest* attribute of the properties of the ontology. Isard et al. [5] describe how interest is used to impose a preference ordering of the properties of ontological entities, controlling which facts will be used when describing each entity.

In ELEON, we have extended interest models in two respects:

- by generalizing *interest* into arbitrary, author-defined *profile attributes*; and
- by permitting profile attributes to apply not only to ontological properties, but also to individuals and classes.

Using these extensions, authors can define *personality profiles* for generating text, managing dialogue, and simulating emotional variation in a way that reflects a certain *personality* on behalf of the system.

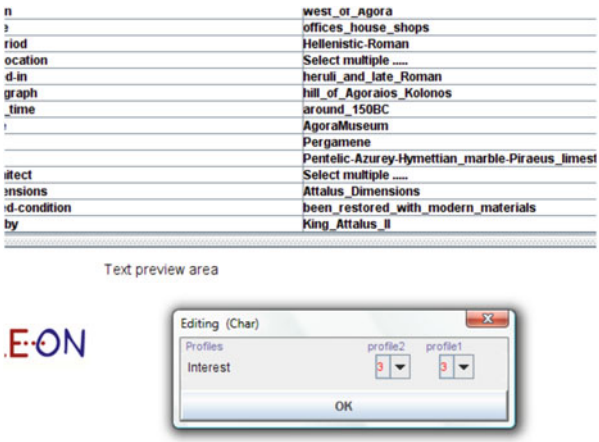


**Fig. 2** Screen fragment, showing the pop-up window for providing profile attribute values for an exhibit. Automatically inferred attribute values are displayed in red, to stand out from explicitly provided ones which are displayed in black.

In the INDIGO project we used these profiles in a human-robot interaction application, where a robotic tour guide that gives the impression of empathizing with the visitor is perceived as more natural and user-friendly. But the methodology is generally interesting in any context of generating descriptions of cultural heritage content, especially if the individual descriptions are aggregated in a tour of the collection. In such contexts, dialogue-management adaptivity can vary the exhibits included in personalized tours and emotional state variation can match the described content and make the tour more engaging and lively.

The way in which personality profiles are used in INDIGO to estimate the *preference* towards exhibits and their properties and parametrize dialogue management and simulated emotions are discussed in more detail elsewhere [10, 11], so we shall only briefly outline it here: In INDIGO preference is calculated based on a logic model of the robot's personality traits and also on ground facts regarding objective attributes of the content—such as the *importance* of an exhibit—but also subjective attributes that reflect the robot's perception of the content—such as how *interesting* an exhibit is. Importance, in this context, reflects the interaction goals set out by the author and how the system is to adapt to achieving them; interest, on the other hand, reflects adaptivity to the visitors' interests, possibly counter to the 'objective', or rather curator-defined, importance.

Emotional variation is achieved by using the personality profile to estimate the *emotional appraisal* of dialogue acts and update the mood and emotional state of artificial agents. Dialogue management is affected both directly, by taking exhibit preference into account when deliberating over dialogue acts, and indirectly, by being influenced by the artificial agent's current mood.

The detailed profiles required by INDIGO are, however, difficult to author and maintain. In the work described here, we alleviate the burden of manually providing all the ground parameters, exploiting the fact that these parameters are strongly inter-related and can, to a large extend, be automatically inferred. More specifically, ELEON backs the profile authoring process by reasoning over manually provided exhibit attributes in order to infer what the values of the missing attributes should be. The author can inspect the explicitly provided as well as the automatically inferred values and make corrections where necessary (Fig. 2). Manual corrections trigger a re-estimation of the missing values, so that after each round of corrections the overall model is a closer approximation of the author's intention.

## *3.2 Representation and Interoperability*

Linguistic and profile annotations are represented in the *Resource Description Framework* (RDF) [12], a knowledge representation technology built around the concept of using subject-predicate-object triples to describe abstract entities, *resources*. RDF triples assign to their subject resource the property of being related to the object through the predicate resource. Predicates can be *data properties*, in which case their objects are concrete values (numbers, strings, time periods, and so on), or *object properties*, in which case their objects are abstract resources.

Although OWL is not formally defined in RDF, it is defined in such a way that it can be represented within RDF. In fact, the OWL specification itself provides a serialization of OWL ontologies as RDF for transport and data interchange purposes. This motivates our usage of RDF to represent linguistic and profile annotations, since it allows us to directly represent those as RDF triples of extra-ontological properties of the OWL ontology. In this manner, standard RDF tools can directly access the link between domain instances and their respective annotations, while at the same time annotations remain 'invisible' to OWL inference tools.

The RDF vocabulary used defines a property that relates ontological entities (individuals, classes, and properties) with profile attribute nodes that involve:

- the profile to which they are pertinent, e.g., 'expert';
- the attribute, e.g., 'interest' or 'importance'; and
- the numerical value of the attribute for this entity in this profile.

When applied to ontology properties, profile attribute nodes can be further elaborated to apply only to properties of instances of a particular class. For example, one can express that users find it more interesting to know the architectural style when discussing temples than when discussing stoas.

## 4 Adaptive Natural Language Generation

NATURALOWL is based on ideas from ILEX [17] and M-PIRO [9]. It adopts a typical pipeline NLG architecture [18] to produce text in three sequential stages: *document planning*, *micro-planning*, and *surface realization*. In document planning, the system first selects the logical facts of the domain ontology that will be conveyed to the visitor and it specifies the document's structure. In micro-planning, it constructs abstract forms of sentences, it aggregates them into abstract forms of longer sentences, and it produces appropriate referring expressions (e.g., pronouns, proper names, noun phrases). Finally, in surface realization the abstract forms of the sentences are transformed into a real text.

The system can also opportunistically include in the generated texts comparisons to previously encountered objects (e.g., 'Unlike all the vessels that you saw, which were decorated with the black-figure technique, this amphora was decorated with the red-figure technique.'), as well as comparisons to similar objects of the entire collection (e.g., 'This is the only vessel of the collection that was decorated with the black-figure technique.'). We do not discuss comparisons here, but related methods can be found elsewhere [8, 13, 15].

### *4.1 Document Planning*

To produce a natural language description of an object (an entity of the domain ontology), NATURALOWL begins by selecting from the domain ontology, assumed to be in OWL, all the logical facts that are directly relevant to that object. For example, when describing an exhibit whose identifier is exhibit24, it might select the following facts, which associate exhibit24 with the class (concept) aryballos and the entities archaeological-delos (the Archaeological Museum of Delos), iraion-delos (an archaeological site), and archaic-period.

```
<exhibit24, rdf:type, aryballos>
<exhibit24, current-location, archaelogical-delos>
<exhibit24, location-found, iraion-delos>
<exhibit24, creation-period, archaic-period>
```

OWL facts can be represented as RDF triples [12], which is why facts are shown here as triples. The triples correspond to fields of ELEON; the first element of a triple is the owner of the field, the second element is the name of the field, and the third one is the field's filler.

NATURALOWL may also be instructed to include facts that are indirectly relevant to the described object, i.e., facts (triples) connected to the directly relevant facts. In that case, the selected facts of our example might also include facts like the following:

```
<archaic-period, covers, archaic-period-duration>
<aryballos, rdfs:subclassOf, vessel>
```

The set of selected facts is subsequently finalized by first removing already assimilated facts, as indicated by a user model maintained for each particular visitor. If a directly relevant fact is removed, then the indirectly relevant facts that depend on it are also removed. Furthermore, the user adaptation resources, to be discussed in the following sections, specify an interest score for each fact of the ontology. Among the remaining facts, those with the lowest interest scores are removed, until the remaining number of facts (direct and indirect) does not exceed the maximum allowed number of facts per text. The latter parameter is also provided by the user adaptation resources.

The directly relevant facts are then ordered by consulting a (partial) order (see below) of the ontology's properties (fields); the order may indicate, for example, that the current location of any exhibit should be mentioned first, followed by the location where it was found, and then the period when it was created. The indirectly relevant facts are placed right after the corresponding directly relevant facts, ordered again by consulting the property order. This arrangement produces texts like 'This is an aryballos. It was created during the Archaic Period. The Archaic Period lasted from...' In the application domains we have considered, this ordering scheme was adequate, although in other domains more elaborate text planning approaches may be needed.

## *4.2 Micro-Planning*

Each selected fact is of the form $\langle S, P, O \rangle$, where $S$ corresponds to the subject of a sentence, $P$ is a property of the domain ontology that is typically expressed by using a verb, and $O$ is the value of the property, typically expressed as the verb's object. For each property $P$ of the domain ontology, one or more micro-plans need to be specified per language to indicate how to express as a sentence any fact that involves that property. Each micro-plan can be thought of as a sequence of slots, along with instructions specifying how to fill the slots in. Each slot can be filled in by:

- An expression referring to the $S$ of the fact (to exhibit24 in the case of `<exhibit24, current-location, archaeological-delos>`).
- An expression referring to the $O$ of the fact (to archaeological-delos in the case of `<exhibit24, current-location, archaeological-delos>`).
- A fixed string. If the string is a verb, it is also tagged with tense and voice; these tags are used in sentence aggregation.

Micro-plans are specified as RDF annotations of the corresponding properties of the domain ontology. The following RDF triples, for example, provide an English micro-plan for facts involving the current-location property; they also set the property's order score to 1, i.e., the resulting sentence should be placed before any other sentence that expresses a fact involving a property with a larger order score. The micro-plan of the example has four slots. The first one must be filled

```
<owlnl:property rdf:about="...#current-location">
  <owlnl:order>1</owlnl:order>
  <owlnl:EnglishMicroplans ...>
    <owlnl:microplan ...>
    <owlnl:aggrAllowed>true</owlnl:aggrAllowed>
    <owlnl:slots ...>
      <owlnl:owner>
        <owlnl:case>nominative</owlnl:case>
        <owlnl:retype>re_auto</owlnl:retype>
      </owlnl:owner>
      <owlnl:verb>
        <owlnl:voice>active</owlnl:voice>
        <owlnl:tense>present</owlnl:tense>
        <owlnl:val>is located</owlnl:val>
      </owlnl:verb>
      <owlnl:text>
        <owlnl:val>in</owlnl:Val>
      </owlnl:text>
      <owlnl:filler>
        <owlnl:case>accusative</owlnl:case>
        <owlnl:retype>re_auto</owlnl:retype>
      </owlnl:filler>
    </owlnl:slots>
  </owlnl:microplan>
  </owlnl:EnglishMicroplans>
  <owlnl:GreekMicroplans ...>
  ...
</owlnl:Property>
```

**Fig. 3** Micro-plan example

in by a referring expression for *S* (the owner of the corresponding field in ELEON terminology) in nominative case. The `re_auto` in `owlnl:retype` lets the system select automatically among using *S*'s name in natural language (if there is one in the lexicon, see below), a noun phrase (e.g., 'this aryballos'), or a pronoun to refer to *S*, depending on the context. The second slot is to be filled in by the string 'is located', which is marked up as being a verb form in present tense and active voice. The third slot will be filled in by the string 'in', and the fourth slot by an accusative case automatically selected referring expression for *O* (the filler of the corresponding field). The micro-plan in Fig. 3 produces sentences like 'It is located in the Archaeological Museum of Delos.'

NATURALOWL currently employs a very simple algorithm for generating referring expressions: once the object being described has been introduced by mentioning its class (e.g., 'This is an aryballos'), it uses pronouns to refer to that object (e.g., '*It* was decorated with the red-figure technique. *It* was created during the Archaic Period.') until the focus moves to another entity via an indirect fact. The new focus is first referred to by its name and then by pronouns ('*The Archaic Period* lasted from 700 till 480 B.C. *It* was when the Greek city-states…'). Then, when

```
<owlnl:owlClass rdf:about="...#aryballos">
  <owlnl:hasLexEntry rdf:resource="#aryballos-lexicon"/>
</owlnl:owlClass>

<owlnl:lexEntry rdf:ID="aryballos-lexicon">
  <owlnl:LanguagesLexEntry ...>
  <owlnl:EnglishLexEntry>
    <owlnl:gender>nonpersonal</owlnl:gender>
    <owlnl:singular ...>aryballos</owlnl:singular>
    <owlnl:plural ...>aryballoi</owlnl:plural>
  </owlnl:EnglishLexEntry>
  <owlnl:GreekLexEntry>
    <owlnl:gender>masculine</owlnl:gender>
    <owlnl:singularForms>
      <owlnl:nominative ...>...</owlnl:nominative>
      <owlnl:genitive ...>...</owlnl:genitive>
      <owlnl:accusative ...>...</owlnl:accusative>
    </owlnl:singularForms>
    ...
</owlnl:lexEntry>
```

**Fig. 4** Lexicon entry example, listing the various forms of the noun *aryballos* and providing gender information

the focus returns to the original object, a demonstrative is used ('*This aryballos* is made of…'). Some property values (*O*) may contain long canned strings (e.g., anecdotes about exhibits), and there are special annotations to flag canned strings that change the focus, so as to avoid using a pronoun in the next sentence. More elaborate referring expression generation algorithms could in principle be added in future versions of the system.

To generate referring noun phrases, like 'this aryballos', NATURALOWL requires OWL classes and entities to be associated with lexicon entries of nouns or proper names. This is again achieved by using RDF annotations. In the RDF triples of Fig. 4, the class aryballos is associated with a noun lexicon entry `aryballos-lexicon`. The lexicon entry contains the various forms of the noun, provides information on gender etc. In practice, all the RDF annotations, including lexicon entries and micro-plans, are constructed by using ELEON, instead of directly editing RDF statements.

## *4.3 Surface Realization*

In surface realization, the system simply concatenates the slot values of the filled-in micro-plans to produce actual sentences. Each micro-plan gives rise to a single sentence (e.g., 'This is an aryballos. It was decorated with the red-figure technique.') These sentences are then aggregated to form longer ones (e.g., 'This is an aryballos

decorated with the red-figure technique.') using domain-independent aggregation rules based on those of M-PIRO [14]. Space does not permit a more detailed description of the aggregation stage, which is actually part of micro-planning and operates on (filled-in) micro-plans, before surface realization. NATURALOWL's surface realizer can also add syntactic or semantic markup. For example, each sentence may be marked up with the corresponding OWL triples, leading to texts readable by both humans and computer applications; the latter would rely on the semantic markup.

## 5 Intelligent Authoring Support

As already discussed above, the detailed profiles required to adapt NLG are difficult to manually author and maintain. In this section we describe two automations that support the authoring process, an *inference* approach exploiting that these parameters are strongly inter-related and can, to a large extend, be inferred from each other; and a *data mining* approach that constructs user models from interaction logs.

### *5.1 Profile Completion*

We have previously discussed how profile attributes are represented as RDF annotations. While is advantageous from the perspective of containing the ontology within the Description Logics complexity fragment, this choice leaves profile attributes outside the scope of reasoning tools.

In order to be able to efficiently reason over and draw inferences about profile attributes themselves, we have chosen to interpret profile attributes within *many-valued description logics*. Using description logics has the advantage of direct access to the domain ontology; using many-valued valuations has the advantage of providing a means to represent and reason over numerical values.

Profile attributes of individuals are captured by normalizing in the $[0,1]$ range and then using the normalized value as a class membership degree. So, for example, if `interesting` is such an attribute of individual exhibits, then an exhibit with a (normalized) interest level of 0.7 is an instance of the Interesting class at a degree of 0.7.

Attributes of classes are reduced to attributes of the members of the class, expressed by a class subsumption assertion at the degree of the attribute. So, if the class of stoas is interesting at a degree of 0.6, this is expressed by asserting that being a member of Stoa implies being a member of Interesting. The implication is asserted at a degree of 0.6, which, under Łukasiewicz-Tarski semantics, means that being a stoa implies being interesting at a loss of 0.4 of a degree. Thus individuals that are members of the Stoa class at a degree of 1.0, are implicitly interesting at a degree of 0.6. Although this is not identical to saying that the class itself is interesting, it clearly captures the intention behind the original RDF annotation.

| Resource | Property | Value | Interest |
|---|---|---|---|
| Stoa of Attalus | style | Doric | 0.8 |
| Stoa of Attalus | style | Ionic | 0.7 |
| Stoa of Attalus | style | Pergamene | 0.3 |
| Stoa of Attalus | orderedBy | Attalus | 0.9 |

**Table 1** Ontology and profile fragment, showing the interest factors of the fillers of properties of the Stoa of Attalus

Profile attributes can also characterize properties, such as style, orderedBy, or creationEra, encoding the information that it might, for example, be more interesting to describe the artistic style of an exhibit rather than provide historical data about it. This is interpreted as the strength of the connection between how interesting an exhibit is, and how interesting its properties are. In other words, if having an interesting filler for style also makes the exhibit interesting, this is taken to mean that the style relation itself is an interesting one. Formulated in logical terms, having interesting relation fillers implies being interesting, and the implication holds at a degree provided by the interest level of the relation itself.

For example, let us assume that the style property has an interest factor of 0.8 and the orderedBy property an interest factor of 0.4. We interpret this as:

$$\mathsf{Interesting} \sqsubseteq \exists \mathsf{style}.\mathsf{Interesting} : 0.8$$
$$\mathsf{Interesting} \sqsubseteq \exists \mathsf{orderedBy}.\mathsf{Interesting} : 0.4$$

That is to say, the class of things that are related to at least one Interesting instance with either style or orderedBy property, are themselves Interesting; however the level of interest that is 'transferred' from the filler to the resource that has the property is higher for the style property than it is for orderedBy.

Given an ontology and profile fragment like the one in Table 1, Stoa of Attalus has an interesting style at a degree of 0.8, which is the maximum among the three architectural styles found in the stoa (Doric, Ionic, and Pergamene). Since style fillers transfer interest at a loss of 0.2, style contributes 0.6 to the stoa's Interestingness. By contrast, the filler of orderedBy (which is more interesting in this profile than any of the architectural styles) only contributes 0.3 of a degree, because orderedBy is annotated as uninteresting and interest transfers across it at a heavy loss.

We have so far discussed how to infer profile attribute values for the individuals of the domain. Classes and relations receive the value of the *minimal instance* of the class (or relation). That is to say, the individual (or pair of individuals) for which nothing else is known, except that it is a member of the class (or relation).

As an example, consider a DoricBuilding class which is a subclass of Building that only admits instances that have a style relation with Doric. The minimal instance of this class is a supposed and unnamed member of Interesting through having an interesting property as discussed above, even though nothing else is known about it. This membership degree in Interesting is taken to be an attribute of the class itself

rather than any one of its members, and is used as the attribute value for the class itself.

For relations, two minimal instances of the relation's domain and range are created. The attribute value for the property is the degree of the implication that having this property makes the domain individual have the attribute. For example, in order to infer how interesting the property devotedTo is, we first observe that it relates Temple instances with MythicalPerson instances, and create bare instances of these two classes. The implication that having a devotedTo relation to an Interesting individual leads to being member of Interesting holds to a degree that can be calculated, given the Interesting degrees of the Temple and MythicalPerson instances involved in the relation. The degree of the implication is then used as the value of the `interesting` attribute.

## *5.2 Interaction Log Mining*

The NCSR Personalization Server (PSERVER) is a domain and application independent system for storing and processing user profiles. User profiles comprise static properties, numeric or symbolic, like age, gender, level or expertise, etc. as well as dynamic features. Features are name-value pairs and different features can be defined as required by each applications, with feature values representing an estimation of the user's affinity to the corresponding feature.

More specifically, one can define a set of features, which 'describe' the application, with a default value for each feature, based on user properties. Whenever a visitor reacts to a feature (by, for example, accessing a exhibit that is described by the feature), PSERVER increases the value of that feature for the specific user. Whenever queried by the application, PCSERVER predicts the interest that the user has on a given exhibit, based on the values of the user model for the features that describe the exhibit.

PSERVER 'sketches' user interests without any prior information about the user, but only relying on the interaction between features of exhibits and the interest level for each feature found in the user model. PSERVER implements clustering methods for creating *user communities* (groups of users with common profiles) as well as *feature groups* (groups of features which describe common domains).

In the case of INDIGO visitors, features relate to the INDIGO domain ontology. More specifically, PSERVER features correspond to the ontological properties of exhibits—creator, place of origin, current location, historic period, artistic style, etc. Each interaction between the visitor and INDIGO (i.e. the user's querying the robot about an exhibit) increases the value of the feature or features the object of the interaction (the exhibit) relates to.

## 6 Related Work

ELEON/NATURALOWL is based on ideas from ILEX [17] and M-PIRO [9]. The ILEX project developed an NLG system that was demonstrated mostly with conceptual representations of museum exhibits; it did not support, however, OWL.[8] In subsequent work, the M-PIRO project produced a multilingual extension of the ILEX system, which was tested in several domains, as well as a precursor to ELEON [3]. However, attempts to add support for OWL in the M-PIRO generation system ran into problems, because of incompatibilities between OWL and the ontological model used in M-PIRO [2]. By contrast, NATURALOWL was especially developed for OWL ontologies, which are also supported by ELEON. Both systems inherit from their precursors the core idea of separating the domain ontology from the linguistic and user adaptation resources, which has significant advantages, as already discussed.

Previous versions of ELEON also featured authoring facilities such as using an external description logic reasoner to catch logical errors by checking the consistency of the authored ontology [5]. In the work presented here, the intelligence behind the tool is substantially extended by using logical inference to *predict* values that have not been explicitly entered by the user, alleviating the need to manually provide large volumes of numerical data.

## 7 Conclusion

In this chapter we have presented ELEON/NATURALOWL, an integrated authoring and NLG system that can be used to create domain ontologies in OWL, annotate them with linguistic and user adaptation resources, and generate textual descriptions of the ontology's entities. We have also discussed the use of ELEON/NATURALOWL to generate descriptions of cultural heritage objects.

The advantages of using ELEON instead of generic knowledge authoring tools, such as Protégé,[9] stem from the ability to couple ELEON with external engines that provide important conveniences to the author, such as the semantic profile completion and usage log mining facilities discussed in Section 5.

These facilities considerably reduce the effort required to create a fully functional model as the author can start out with initial profiles which can be refined by iterating through cycles of providing information, previewing the generated text, and only elaborating the model where the text is unsatisfactory. This iterative process converges to satisfactory descriptions much faster than having to manually enter all adaptation parameters, especially for large and complex domains.

Our future plans include enhancing PSERVER with a recommender module which will suggest items of possible interest to the current user. It can provide content based recommendations, that is suggestions that are similar to past user preferences.

[8] Dale et al. [6] describe a similar museum system.

[9] See http://protege.stanford.edu

In doing so, the system tries to discover commonalities between positively-rated exhibits seen in the past. In particular, some of following features of exhibits can be used to detect commonalities: author, historical period, type of artifact, etc. Also, a collaborative recommender system could suggest interesting exhibits to a user based on items judged as interesting by a group of similar minded users [1]. For instance, it might be discovered that visitors interested in architecture are also interested in sculpture. Thus, a new visitor, after having received information about an architectural exhibit, might be offered advice for seeing the collection of sculptures.

To enhance the system's acceptance by humans, it would be desirable to justify the exhibits it suggests by revealing its reasoning. Such justifications or explanations are classified as *opaque* or *transparent*. *Opaque explanations* [7] are neighbour-based and essentially provide a statistics of the similar users' preferences. This type is more pertinent in collaborative based recommendations, exemplified by the following dialogue excerpt:

*User:* Why did you guide me to the Tholos?
*Robot:* 75% of visitors similar to you chose to visit it.

*Transparent explanations* [16], on the other hand, are based on reciting the features of the suggested exhibit. In case there is a follow-up question with the user requesting further explanation, then some of the features of the current exhibit can be associated with features of a previously visited exhibit:

*User:* Why did you suggest to view the temple of Hephestos?
*Robot:* It was build in the Ancient Agora and was constructed by Phedias
*User:* How did you guess that I am interested in Phedias?
*Robot:* Considering you have expressed interest in the Acropolis, which was also built by Phedias.

Given the nature of the descriptions generated by ELEON/NATURALOWL, collaborative filtering and opaque explanations are expected to be less relevant; it is rather transparent explanations that would mostly contribute to the persuasiveness of the system.

Finally, NATURALOWL is also being further developed to generate descriptions of both classes and entities, engaging richer language resources and more complex aggregation mechanisms. Extensions are also being made to support ontologies in OWL 2 [19], the most recent OWL specification, and the various complexities of event-based ontologies, such as CIDOC CRM. Lastly, the domain-specific language resources are being reviewed and redefined as OWL ontologies instead of RDF annotations.

**Acknowledgements** Much of the work reported here was carried out in the context of the Greek project XENIOS and the subsequent European project INDIGO, where human-robot interaction technology was developed.[10]

We are grateful to Ion Androutsopoulos[11] and Gerasimos Lampouras[12] for allowing us to include their text on NATURALOWL, and on Natural Language Generation in general, as well as for their corrections and contributions throughout this chapter. Naturally, all errors and omissions are our own.

Finally, we wish to acknowledge the help of the Foundation of the Hellenic World, whose staff used the ELEON/NATURALOWL system to create the initial version of the Ancient Agora of Athens ontology.

## References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering **17**(6) (2005)
2. Androutsopoulos, I., Kallonis, S., Karkaletsis, V.: Exploiting OWL ontologies in the multilingual generation of object descriptions. In: Proceedings of the 10th European Workshop on Natural Language Generation (ENLG-05), Aberdeen, August 2005 (2005)
3. Androutsopoulos, I., Oberlander, J., Karkaletsis, V.: Source authoring for multilingual generation of personalised object descriptions. Journal of Natural Language Engineering **13**(3), 191–233 (2007)
4. Antoniou, G., van Harmelen, F.: A Semantic Web Primer, 2nd edn. MIT Press (2008)
5. Bilidas, D., Theologou, M., Karkaletsis, V.: Enriching OWL ontologies with linguistic and user-related annotations: the ELEON system. In: Proc. 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-2007), Patras, Greece, October 29–31, 2007, vol. 2. IEEE Computer Society (2007)
6. Dale, R., Green, S.J., Milosavljevic, M., Paris, C., Verspoor, C., Williams, S.: Dynamic document delivery: generating natural language texts on demand. In: Proceedings of the 9th International Conference and Workshop on Database and Expert Systems Applications, pp. 131–136. Vienna, Austria (1998)
7. Herlocker, J.L., Konstan, J.A., Riedl, J.: Explaining collaborative filtering recommendations. In: Proceedings of the ACM conference on Computer supported cooperative work (CSCW 2000), pp. 241–250. ACM, New York, USA (2000)
8. Isard, A.: Choosing the best comparison under the circumstances. In: Proceedings of the International Workshop on Personalization Enhanced Access to Cultural Heritage (PATCH07), 11th International Conference on User Modeling (UM07), June 2007, Corfu, Greece (2007)
9. Isard, A., Oberlander, J., Androutsopoulos, I., Matheson, C.: Speaking the users' languages. IEEE Intelligent Systems **18**(1), 40–45 (2003)
10. Konstantopoulos, S.: An embodied dialogue system with personality and emotions. In: M. Danieli, B. Gambäck, Y. Wilks (eds.) Proc. of Workshop on Companionable Dialogue Systems, ACL 2010. Association for Computational Linguistics (ACL), Uppsala, Sweden (2010)
11. Konstantopoulos, S., Karkaletsis, V., Matheson, C.: Robot personality: Representation and externalization. In: Proceedings of International Workshop on Computational Aspects of

[10] Please see http://www.ics.forth.gr/xenios (in Greek) and http://www.ics.forth.gr/indigo for more information.

[11] Department of Informatics, Athens University of Economics and Business, and Digital Curation Unit - IMIS, Research Centre 'Athena'

[12] Department of Informatics, Athens University of Economics and Business

Affective and Emotional Interaction (CAFFEi 08), Patras, Greece, July 21st, 2008, pp. 5–13. ECCAI (2008). URL http://www.iit.demokritos.gr/~konstant/dload/Pubs/caffei08.pdf
12. Manola, F., Miller, E.: RDF primer. W3C Recommendation, 10 February 2004 (2004). URL http://www.w3.org/TR/rdf-primer/
13. Marge, M., Isard, A., Moore, J.: Creation of a new domain and evaluation of comparison generation in a natural language generation system. In: Proceedings of the Fifth International Language Generation Conference (INLG08), June 2008, Salt Fork, Ohio, USA (2008)
14. Melengoglou, A.: Multilingual aggregation in the M-PIRO system. Master's thesis, School of Informatics, University of Edinburgh (2002)
15. Milosavljevic, M.: The automatic generation of comparison in descriptions of entities. Ph.D. thesis, Department of Computing, Macquarie University, Australia (1999)
16. Mooney, R.J., Roy, L.: Content-based book recommending using learning for text categorization. In: Proceedings of the 5th ACM conference on Digital libraries (DL 2000), San Antonio, Texas, USA, pp. 195–204. ACM, New York, USA (2000). DOI http://doi.acm.org/10.1145/336597.336662
17. Ó Donnell, M., Mellish, C., Oberlander, J., Knott, A.: ILEX: an architecture for a dynamic hypertext generation system. Natural Language Engineering **7**(3), 225–250 (2001)
18. Reiter, E., Dale, R.: Building natural language generation systems. Cambridge University Press (2000)
19. W3C OWL Working Group: OWL 2 web ontology language. W3C Recommendation, 27 October 2009 (2009). URL http://www.w3.org/TR/owl2-overview/

# Part V
# Structural and Narrative Analysis

# Automatic Pragmatic Text Segmentation of Historical Letters

Iris Hendrickx, Michel Généreux and Rita Marquilhas

**Abstract** In this investigation we aim to reduce the manual workload by automatic processing of a corpus of historical letters for pragmatic research. We focus on two consecutive sub tasks: the first task is automatic text segmentation of the letters in formal/informal parts using a statistical n-gram based technique. As a second task we perform semantic labeling of the formal parts of the letters using supervised machine learning. The main stumbling block in our investigation is data sparsity due to the small size of the data set and enlarged by the spelling variation present in the historical letters. We try to address the latter problem with a dictionary look up and edit distance text normalization step. We achieve results of 86% micro-averaged F-score for the text segmentation task and 66.3% for the semantic labeling task. Even though these scores are not high enough to completely replace the manual annotation with automatic annotation, our results are promising and demonstrate that an automatic approach based on such a small data set is feasible.

**Key words:** historical text, text segmentation, semantic labeling, text normalization

## 1 Introduction

In the CARDS project, private Portuguese letters from 16th to 19th century are manually transcribed into a digital format[1]. The aim of the project is to produce an electronic critical edition and historical-linguistic treatment of the text of the letters. The project main linguistic focus is on discourse analysis and historical pragmatics; private letters seem to be the best possible trace left by the linguistic behavior of speakers of the past in their role of social agents. A small part of the collection is

Centro de Linguística da Universidade de Lisboa, Av. Prof. Gama Pinto, 2 1649-003 Lisboa, Portugal e-mail: iris@clul.ul.pt,genereux@clul.ul.pt,rmarquilhas@fl.ul.pt

[1] http://alfclul.clul.ul.pt/cards-fly

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing, DOI 10.1007/978-3-642-20227-8_8, © Springer-Verlag Berlin Heidelberg 2011

annotated with textual and semantic labels to serve as the basis for corpus-based pragmatic studies.

In the study presented here, the aim is to reduce the manual tasks by automatic processing of the corpus of historical letters. We use the small manually labelled sub-set of letters to develop automatic systems in order to label the other part of the collection automatically. We present our first attempt to apply standard statistical and machine learning methods on the labeled data.

We focus on two consecutive sub tasks within the field of historical pragmatics that are important for the research in the CARDS project. The first task is automatic text segmentation of the letters. Letters are a specific genre of text and most of them follow a rather strict division. In general, a letter has an opening part, a formal introduction, the body of the letter that conveys the actual message, and a more formal closing part. To study the language usage in the letters in depth, it is important to discriminate between these formal and informal parts. This type of task is closely related to topic-based text segmentation [21] or paragraph detection [37]. Here we evaluate a method to discover the formal and informal text segments automatically using a statistical method based on n-grams similar to [4] and a smoothing algorithm to reduce discontinuities.

The second task is to investigate the formal parts in more detail and study their pragmatic relevance. For this purpose, the formal parts of the letters were further annotated with semantic labels from the Lancaster semantic taxonomy [35]. This semantic labeling system has already been applied successfully to historical English texts [2]. We try to predict the semantic labels automatically. As this type of semantic labeling can be viewed as a kind of coarse-grained word sense disambiguation (WSD), we take a supervised machine learning approach as previous research in this field has shown that this is a successful approach [13, 23].

This study of automatic methods for processing historical letters faces several challenging tasks. First of all, the data set is small, as we have only 502 manually annotated historical letters to work with. This is a small set to serve as training material for automatic methods. Secondly, these letters are handwritten in a time period well before spelling standardization so the texts contain many spelling variations. The letters cover a time period of four centuries aggravating the problem of word variation. Furthermore, there are not many preprocessing tools available for this type of historical text, such as a lemmatizer which would have produced useful and valuable information sources for our automatic approach.

Spelling variation is a huge hindrance for automatic processing of historical texts, for example in automatic text retrieval [14, 24] and corpus linguistics [3]. Hence the first step was to normalize the spelling of the texts. We used dictionary lookup and edit distance as is further detailed in Sect. 3 on experimental methods. In Sect. 4 we present the automatic text segmentation. Section 5 details the semantic labeling of the formal parts of the letters. We conclude in Sect. 6. In the next section we give more details about the CARDS project and the data set we used in our experiments.

## 2 Corpus of Historical Letters

Historical research builds on a kind of empirical data that is not easy to process automatically. The sources, as the data are traditionally called, lose much information as they are ripped from their original context. In order to re-build those contexts, historians have to double their attention to the object's original material features, which can vary immensely. When we think of written documents as historical sources, the only viable approach is based on Textual Criticism methods[2] [5]. As a consequence, the documents have to be read in the original, or in an original-like support, and the more idiosyncratic features are registered, the stronger the historical interpretation can become. This means to work physically in the archives and to transcribe the documents by hand into a new, digitalized, highly annotated format. It is an expensive and time-consuming process which would highly benefit from a proper automatic operation designed for it.

In the CARDS project, private letters from 16th to 19th century Portugal are being manually transcribed; its aim is to produce an electronic critical edition and historical-linguistic interpretation of the letters. The CARDS sources are unpublished epistolary manuscripts that were kept within criminal law-suits by the Portuguese Inquisition and the Royal Appeal Court as evidence either of guilty parties, or of innocent ones. The scanned manuscripts from Inquisition law-suits are from the 16th, 17th and 18th century, and the Royal Appeal Court law-suits are from the three first decades of the 19th century. In such a textual collection, the discourse of the lower classes is widely represented since the social elites were openly protected in those *Ancien Régime* times, both by the Church and the Royal powers. To complete the social representativeness of this rather 'popular' letters collection, a small sample of the elite's discourse in letters is also being gathered, coming from powerful families' archives. The resulting corpus contains, for now, a sum of 450K words belonging to 1700 letters. The project will stop at 2000 letters (1850 letters from law-suits, 150 letters from elite family archives), involving more than a 1000 identified participants (plus the anonymous ones). However, for the investigation presented here, we only use a small part of the full corpus, 502 letters which have been fully annotated. In Table 1 we list the chronological distribution of the full set of 2000 letters and the part used here as our data set.

The manuscripts are almost always originals and they are transcribed by the quasi-diplomatic method to XML files (this means that it is only allowed to normalize for word boundaries and i/j, u/v variation). Lacunae, difficult deciphering, abbreviations, diacritics and non-orthographic uses are all kept in the transcription by means of the tags which were specifically developed for primary sources editions

[2] Textual Criticism is the discipline that combines several methods (paleography, diplomatics, codicology, bibliography, textual recension and editorial emendation) to publish selected texts in the closest possible way to their authors' final intentions.

**Table 1** Distribution of letters in the different centuries in the complete CARDS corpus and the data set used in this study

| century | # total | # data set |
|---|---|---|
| 16th | 50 | 11 |
| 17th | 250 | 165 |
| 18th | 700 | 215 |
| 19th | 1000 | 111 |
| total | 2000 | 502 |

by TEI (Text Encoding Initiative[3]) and for epistolary data by DALF (Digital Archive of Letters in Flanders[4]).

The information on social backgrounds and situational contexts, being largely recoverable, is also encoded within a database whose entries are anchored within the letters' XML files. By following Textual Criticism methodologies while keeping in mind social analyses preoccupations, the CARDS project manages to connect such diverse information as manuscripts' physical layout, original writing, authorial emendations, editorial conjectures, information on variants (when copies or comments also survived), information on the context of the letter's composition (an event within a social context), and information on the participants' biographies.

In recent literature on historical pragmatics and discourse analysis, letters are extremely valuable sources for the study of the past social uses of language [12,30]. Nevertheless, being highly subject to genre constraints, not all textual parts within letters can be considered as similar candidates for an idealized representation of face-to-face interaction. Latitude has to be allowed for the presence of strict social rules applying to the epistolary practice only. As Susan Fitzmaurice [17, page 77], puts it:

> Individual writer's letters are as distinctive as their signatures, yet, at the same time, the writers appear to share practices that amount to recognizable conventions of letter writing.

This is true, above all, for opening and closing formulae. Well before approaching the probability of the oral-written parallelism, two questions arise concerning the pragmatic value of these textual conventions: Are they always valid for all social classes, for men as well as for women and for all communicative situations? Or are the conventions culture-dependent and subject to change?

Literary critics of the letter genre seem to have strong views about the rigidity of its conventions. Claudio Guillén, when speaking of the "profoundly primitive quality" of the familiar letter in prose, whose origins, in Western culture, go back a great many centuries, to the Mediterranean Antiquity, commented this: "Hence also its highly conventional character, visible since antiquity in the worn and yet indispensable formulas of salutation, apology, recommendation, farewell, and the

[3] http://www.tei-c.org/Guidelines/P5/

[4] http://www.kantl.be/ctb/project/dalf/index.htm

like, on which both the humble letter writer and the sophisticated poet depend [18, page 81]". Since the CARDS sources are the kind of data that represent the typical "humble letter writer" in a substantial and measurable way, we are in the ideal position to test this kind of assumption. We believe that three hypotheses take form concerning people writing letters in the 16th to 19th century time span: 1. they were just celebrating, over and over, the same rituals when beginning and ending the interaction, regardless of their communicative intention, their social origin, their genre, or 2. they were celebrating the same rituals, regardless of their communicative intention, but social origin and genre were consistent factors of ritual variation, or 3. they chose to celebrate or not the same rituals, depending on their communicative intentions. Current discussion on that pragmatic topic is divided between the *universal* approach, and the *politic* approach. According to the first one, there is a 'pan-cultural interpretability of politeness phenomena' [8, page 283] such as deferential language. But authors taking the *politic* approach choose to disagree: they state that 'deferential language such as terms of address, greetings and leave-taking are weapons in a struggle to exercise power' [39, page 156]. The testing of the three hypotheses can become important empirical evidence for the study of ritualized forms of language.

## *2.1 Annotated Data Set*

In the CARDS project we want to see whether different social agents of the past, wishing to act through their letters (as lovers, as deferential servants, as tyrants, as vindictive judgers, as friends, etc.) used to write or not the same things in the greeting and leave-taking part of their letters. Textual and semantic labeling is then necessary. The textual labeling covers chunks of the CARDS' letter-texts that seem to adopt formulaic beginnings and endings in the light of traditional rhetorics. As for the semantic labeling, on the one hand it is designed to measure the distance between the discursive topics in those formulaic parts, and on the other hand, to unveil the main topics developed in the letter as a whole textual unit. Accordingly, the CARDS corpus is to be tagged for conventions in the *opener* part of the letters (consisting of nomination, date, location, salutations) a formal introduction part ( *harengue*), conclusion ( *peroration*) and the *closer* part (including signature, nomination, location). In Table 2 we list the number of formal parts in the data set of 502 letters and 177,082 tokens.[5]

These parts are manually annotated with semantic labels. Semantic classification of historical texts can be very biased, so we followed the same choice made by Archer and Culpeper [1] in their study of language usage in a close-to-spoken corpus similar to the CARDS one, namely dialogues in plays and trial proceedings of the 17th and 18th centuries; this means we also applied the UCREL Semantic Analysis System [35]. This taxonomy, abbreviated as USAS, is a conceptually driven schema and has 21 broad categories such as 'body and the individual','time' and 'social

[5] Punctuation was not counted.

**Table 2** Statistics on the data set of historical letters, frequencies of formal segment types

| segment | frequency |
|---|---|
| opener | 351 |
| harengue | 160 |
| peroration | 342 |
| closer | 434 |
| letters | 502 |
| tokens | 177,082 |

actions, states and processes'. Each of the broad categories is subdivided in more fine-grained categories. A sub-set of the CARDS corpus has been manually labeled with a subset of 15 fine-grained labels from the USAS scheme presented in Table 3. Four annotators annotated independently from each other the data according to annotation guidelines and a manual revision was made in the end by one annotator only.

In example (1) we show a sentence from a peroration of a letter written in 1636 by a merchant to his cousin. The author of the letter apologizes to the receiver for bothering him (labeled as politeness S.1.2.4), expresses respect by addressing him with a formal title (respect is S7.2) and wishes that God (labeled as religion S9) will bless him with a long life (a reference to health, labeled as B2)[6].

*Example 0.1.*
Comtudo eu <S1.2.4>não deixarei enfadar a Vm</S1.2.4> Com Couzas minhas pois <S7.2> tenho Vm por meu Sro </S7.2><S9>a quem elle gde</S9> Com <B2>larguos annos de vida</B2>

To validate the manual annotations of the USAS labeling, we performed a small experiment. We selected a random sample of 20 letters from the CARDS data set and had three annotators label the tokens in the formal parts independently of each other. Next we computed the inter-annotator agreement on the three different annotations. Although we used just a very small sample, this gave us some indication of the difficulty of the task and the consistency of the labeling of the annotators. We computed the kappa statistics [9], which resulted in a kappa value of .63 for the average on the pairwise comparisons of the three labeled sets. The scores are not very high, but this in line with other semantic labeling tasks such as WSD [31].

We did observe that there is also some room for improvement of the manual labeling as several disagreements are caused by placing boundaries at different points. Refining the annotation guidelines on what concerns fixed expressions might lead to somewhat higher agreement scores.

[6] *English translation: However, I won't bother Your Honor with things of mine since I have Your Honor as my Lord whom he (God) guards with many years of life.*

**Table 3** Sub-set of USAS labels used in our investigation.

| Semantic Group | Sub-classification |
|---|---|
| A: General names | A9: Possession |
| B: Body | B2: Health and diseases |
| E: Emotional | E2: Liking |
| | E4: Happy/sad |
| I: Money and Commerce | I1: Money |
| Q: Linguistic | Q1.1.: Communication |
| | Q2.2.: Speech acts |
| S: Social | S1.1.1 General (goodbye) |
| | S1.2.4.:Politeness |
| | S3.1.:Relationship in general |
| | S4: Kin |
| | S7.1: Power, organizing |
| | S7.2.: Respect |
| | S9: Religion |
| X: Psychological | X1: General |

## 3 Experimental Setup

This section presents our experimental setup, the preprocessing step of text normalization and the resources used in our experiments. To perform our experiments, we randomly divided the manually annotated data set of 502 letters in 100 for testing and 402 for training. Some of the letters in this data set are written by the same person. To prevent unfair overlap between training and testing data, we made sure that none of the test letters was written by the same person as a letter in the training data.

One particular feature of our data is the lack of sentence boundary information. The detection of sentence boundaries is a basic preprocessing step when automatically processing text documents. Unfortunately, our corpus of historical letters has such characteristics that this becomes a difficult task. First of all, the letters do not contain many punctuation marks at all. They date from the time before spelling or writing conventions were widely used, and several of the letters are written by people who hardly had learned how to write. Secondly, most of the letters do not signal the beginning of a sentence with a capital letter. Therefore, we can not use the standard sentence boundary detection techniques that rely on punctuation and capitalization clues such as the work of [36] or [28]. We do split off punctuation marks from words when they are present and otherwise the letters are considered as a long string of words.

Typical for the text genre of letters is a large amount of proper names. Names are more informative as a category than as a word, since the proper noun itself does not tend to reappear across texts. We have made an attempt to tag proper nouns on the basis of an ad hoc list of 6923 Portuguese names selected from diverse sources. Each proper noun was replaced by the tag *PROPERNOUN*.

To normalize the text further, we took the following steps. The main resource for the text normalization step was a historical Portuguese dictionary [7] of more than 161,000 entries. Rafael Bluteau's Vocabulario is a large Portuguese-Latin encyclopedic dictionary occupying ten in-folio volumes. It was written in the late 17th century and published in the early 18th century by an English born French priest. The dictionary importance, today, comes mostly from its author being a careful producer of metalinguistic, comparative judgements on the Portuguese language of the Modern period. Besides the dictionary we also made an additional abbreviation list. In the manual annotation, abbreviated words were labeled as such. We extracted a list of these abbreviations from the training set, in total 1224, excluding adverbial abbreviations as these can still have much variance in them (e.g *principalmte, principlmte* for *principalmente*).

For each word in the letters we checked whether it was either an abbreviation or present in the dictionary. Each word that was not, was considered being a spelling variation. These words were replaced with one of the top 5000 most frequent entries in the dictionary having the smallest edit distance, also known as the Levenshtein distance [26]. We limited the maximum edit distance to 5, otherwise we maintained the original word form.

We performed a small manual validation of the text normalization. We randomly selected 20 letters and 4354 tokens from the data set and compared the original texts against the normalized versions. On this small sample, the normalization procedure (including proper name detection) achieved an accuracy of 89.7%, recall of 97.5% and a precision of 91.8%. The majority of the precision errors is due to mismatches between the proposed normalized form of a word and the actual correct form, and 9% of these were (capitalized) words mislabeled as names. Two-third of the errors in the recall (false negatives) are due to unrecognized names.

As an additional resource in our experiments, we used the Tycho Brahe Parsed Corpus of Historical Portuguese[7] (TBCHP), an electronic corpus of historical texts written in Portuguese. The corpus contains prose from different text genres and covers a time period from the Middle Ages to the Late Modern era. TBCHP contains 52 works source texts but not all of them are annotated in the same way. Some of the texts maintain original spelling variation, while other texts, intended for part-of-speech and syntactic annotation, were standardized for spelling. Both modern Brazilian and European Portuguese orthographies are represented in this standardization, depending on the chosen printed sources, but the difference between these two writing standards involves only a few cases of vowel accentuation and a few digraphs within classical loan words. From the whole TBCHP corpus of 52 titles, we discarded the texts dated before 1550, given the need to avoid anachronic generalizations. We also discarded most of the texts with non-modernized spelling, thus arriving at a sample that contains 33 texts (12 of which are letter collections) and 1,516,801 tokens.

We also trained a part-of-speech (POS) tagger on the subpart of TBCHP using MBT [10, 11]. As not all texts in our sample are annotated with part-of-speech, we

[7] http://www.tycho.iel.unicamp.br/∼tycho/corpus/en/index.html

used here a sample of 23 texts (11 of which are letter collections) and 1,137,344 tokens. We only used 45 coarse-grained POS tags and left out specifications of verb tenses or gender/number distinctions. To estimate the performance of the POS tagger, we arbitrarily split the TBCHP data set in a training set of 917,467 tokens and a test set of 219,877 tokens. The measured accuracy on the test set is 95.1%. We cannot estimate the performance of the tagger on the CARDS data as we do not have gold standard labeling of POS tags available here. We expect the accuracy of the POS tagger to be somewhat lower as we switch to another corpus of a specific text genre.

## 4 Text Segmentation

The segmentation task is to assign each word (1-gram) of the historical letters a single of five classes: four formal segment types, (*opener*, *closer*, *harengue* or *peroration*) and one class for words that do not belong to any formal classes (*free*). Our approach is slightly counter-intuitive, as we rely on lexical models for each class we are trying to identify. There are two compelling reasons for this:

1. Albeit small, we have at our disposal texts that are explicitly marked with segment boundaries, so we are in a position to exploit a fully supervised approach, a position very few researchers were in to tackle the segmentation task.
2. Our lexical models will provide us with a useful source of information to characterize each class.

A more intuitive approach would be to look for lexical patterns allowing the identification of the topic boundaries [16, 21], possibly including a variety of knowledge bases (textual indicators, information linked to syntax or discourse) [37]. Argumentative zoning is another approach [19, 38] making use of lexical cues to identify boundaries. Given the complexities of our corpus, an interesting option is a hybrid approach exploiting *a priori* rules concerning the beginning and ending of the formal parts of the letters. Our approach, similar to [4], is to assign each n-gram ($n \leq 3$) in the training data a score representing its salience for the class in which it appears. These scores are used to compute the best class for each word. We use the log odds ratio as a statistical measure of salience or prominence of a n-gram. The log odds ratio measure [15] compares the frequency of occurrence of each n-gram in a given specialized corpus with its frequency of occurrence in a reference corpus as given in Eq. (1) where $a$ is the frequency of a word in the specialized corpus, $b$ is the size of the specialized corpus minus $a$, $c$ is the frequency of the word in the general corpus and $d$ is the size of the general corpus minus $c$.

$$ln(ad/cb) = ln(a) + ln(d) - ln(c) - ln(b) \tag{1}$$

High positive log odds scores indicate strong salience, while high negative log odds scores indicate words irrelevant for the class. We constructed five specialized corpora, one for each of the five classes: *opener*, *harengue*, *peroration*, *closer* and

*free*. We adopted the TBCHP (described in Sect. 3) as our reference corpus. The TBCHP is a good reference corpus for two main reasons:

1. It is quite diversified in terms of genres while CARDS only contains private letters.
2. It is largely standardized for orthography.

The training set of 402 letters was used to create the specialized corpora for each text segment class by concatenating, for each letter, all words belonging to one particular class. This produces the specialized training corpora, whose number of generated texts, relative n-gram distribution and average frequencies (per text) are detailed in Table 4. Not every letter contains all formal parts but most of them contain at least one formal element. There are 45 letters consisting only of free text.

**Table 4** Statistics on n-grams in the specialized corpora for the 5 segment classes. $\overline{fq}$ = average frequency of the n-grams per text

| corpus | # texts | 1-gram ($\overline{fq}$) | 2-gram ($\overline{fq}$) | 3-gram ($\overline{fq}$) |
|---|---|---|---|---|
| opener | 275 | 1.1% (3.1) | 0.9% (1.5) | 0.7% (1.3) |
| harengue | 121 | 2.3% (3.6) | 2.2% (1.3) | 2.2% (1.1) |
| peroration | 231 | 2.6% (4.0) | 2.5% (1.4) | 2.4% (1.1) |
| closer | 343 | 3.3% (3.9) | 3.1% (1.6) | 2.9% (1.2) |
| free | 402 | 90.7% (11.9) | 91.3% (1.7) | 91.8% (1.1) |
| total | 1372 | 100% | 100% | 100% |

The salience for each n-gram was then computed and sorted from the highest to the lowest. In table 5 we show for each class the most salient 3-grams and its log odds ratio.

**Table 5** Most salient 3-grams for each of the four formal segment types

| segment | 3-gram | | log-odds |
|---|---|---|---|
| opener | *Illmo e Exmo* | 'most illustrious and excellent' | 12.9 |
| harengue | *da q me* | 'of the (health) that I' | 8.6 |
| peroration | *Ds gde a* | 'God guards to' | 11.6 |
| closer | *De V Exa* | 'of your excellency' | 10.8 |

### *4.1 Classifying Each Word*

The lists of n-grams with salience values for each class constitute our language models for our classifiers. To derive one particular class tag for each word, the classifier adopts the following two-step strategy:

1. Each 1-gram is assigned salience values for each class as found in the model, zero otherwise;
2. Each word of a n-gram (n $\in$ {2,3}) has its salience values augmented by the corresponding salience values for the n-gram in the model, if they exist.

The above procedure can be restricted to a subset of one (1-grams, 2-grams or 3-grams only), two (1 and 2-grams only, 1 and 3-grams only or 2 and 3-grams only) or three (1, 2 and 3-grams) models. Therefore, each word from the letter has a salience value for each class, possibly taking into account contextual information (if models above 1-gram were included in the computation process). At this point, one could simply select the class with the highest salience value, but we decided to include a further step in the computation in order to give a fairer evaluation of the salience. The highest value was decreased by the value of salience from other classes. For example, if a word has the following salience values: opener:2, closer:5, harengue:-4, peroration:1 and free:0, it is classified as a *closer* (the largest value) with:

$$salience = 5 - 2 - (-4) - 1 - 0) = 6$$

We evaluate the performance of our classifier on the test set of 100 letters. We present F-scores and overall accuracy computed at the word level. We also computed the micro-average F-score over the five classes. Results are shown in Table 6. The 3-gram model alone clearly achieves the best results.

**Table 6** F-scores and overall accuracy for text segmentation

| *n*-gram used | F-scores% | | | | | | Overall accuracy |
|---|---|---|---|---|---|---|---|
| | op | har | per | clo | free | Microaver. | |
| {1} | 4.3 | 13.5 | 5.3 | 24.2 | 66.3 | 61.2 | 50.1 |
| {2} | 8.2 | 18.0 | 7.7 | 32.5 | 74.5 | 69.1 | 60.8 |
| {3} | 22.2 | 15.3 | 13.8 | 24.2 | 91.7 | **84.6** | **87.4** |
| {12} | 5.4 | 14.2 | 7.0 | 26.0 | 58.6 | 54.5 | 43.1 |
| {13} | 5.0 | 17.5 | 7.2 | 29.2 | 65.2 | 60.6 | 49.8 |
| {23} | 8.5 | 19.9 | 8.1 | 34.6 | 74.9 | 69.7 | 61.5 |
| {123} | 5.6 | 15.1 | 7.7 | 26.5 | 58.8 | 54.7 | 43.4 |

### *4.2 Segment Production (Smoothing)*

The previous approach to classify words from a letter on an individual basis can produce rather sketchy classification in which *holes* exist that could otherwise be used as connectors between distant members of the same class. Let's look at the following hypothetical example where subscripts[8] indicate each word tag (as computed by the classifier) and bracketed tags indicate the true tag (as annotated by humans)[9]:

<opener> Meo$_o$ amo$_o$ e$_c$ Sr$_o$ </opener> <harengue>Ainda$_c$ q$_h$ VM$_f$ me$_h$ não$_h$ quer$_h$ dar$_h$ o$_c$ alivio$_h$ de$_h$ suas$_h$ novas$_h$ a$_h$ minha$_h$ amizade$_h$ não$_h$ pide$_h$ tal$_h$ discuido$_c$ e$_h$ assi$_h$ lembresse$_h$ VM$_h$ de$_o$ mim$_f$ q$_h$ com$_h$ novas$_h$ suas$_h$ q$_h$ bem$_h$ sabe$_h$ q$_f$ não$_h$ tem$_h$ qm$_p$ lhas$_h$ dezeje$_h$ com$_h$ mais$_h$ veras$_p$ . </harengue> Sabado$_f$ nove$_f$ deste$_f$ mes$_f$ Domingos$_f$ ... por$_f$ não$_f$ ficar$_f$ com$_f$ escrupello$_f$ <peroration> aqui$_p$ fico$_p$ ás$_h$ ordens$_p$ de$_p$ VM$_p$ pa$_p$ o$_f$ q$_p$ me$_p$ quizer$_p$ mandar$_p$ com$_f$ gde$_p$ vontade$_p$ Ds$_p$ gde$_p$ a$_p$ VM$_p$ </peroration> <closer> Prada$_f$ 10$_f$ de$_f$ Julho$_c$ de$_c$ 1712$_f$ Mayor$_f$ Amo$_c$ e$_c$ Servidor$_c$ de$_f$ VM$_c$ Frando$_f$ de$_c$ SáMzes$_f$ </closer>

Although the patterns of computed tags follow roughly the true annotation, a smoothing technique could be applied to attempt to fill the gaps and create boundaries approaching those created by human annotators, as in sequence modeling [27]. Our approach to text segmentation is more simple yet very effective. We aim at identifying words that appear "lost" among words that belong to a same class (or a majority class). For example, in the sequence

... não$_h$ tem$_h$ qm$_p$ lhas$_h$ dezeje$_h$ ...

the word "qm" appears to break a "harengue" sequence because in the middle of a sequence with an overwhelming majority of words belonging to the "harengue" class. Therefore, our approach is to move a window of fixed size (in words) along the text to identify and change the class of a middle word that is at odds with the majority class of the neighboring words. What should be the size of the window? First, we need to have an idea of the average length and standard deviation of the formal text segments (in words). We obtained those statistics from the training data, shown in table 7.

The combined values for mean and standard deviation will give us an idea of the size of the segments for each class we should be aiming at, on average. We choose an interval for the length of each class so that 95% of the segment size falls within the interval. This is given in normal distribution by computing (mean $\mp 2\times$standard deviation). This formula does not hold when distribution are skewed, but we will

[8] o=opener, c=closer, h=harengue, p=peroration and f=free

[9] English translation: <**opener**> My friend and Lord </**opener**> <**harengue**> Although Your Honor does not want to relieve me with news from Your Honor, my friendship does not call for such a lack of attention, so, remember to give me some news, because you know well that there is nobody else that desires them more than I do, really </**harengue**> Saturday 9 of this month ... of not having scruples <**peroration**> here I remain at the orders of Your Honor for all that Your Honor chooses to command, with all my good will, God keeps Your Honor </**peroration**> <**closer**>Prada, 10th of July 1712 The greatest friend and servant of Your Honor Fernando de Sá Menezes </**closer**>

**Table 7** Distribution of segment lengths of each class

| Class | Mean | St. Deviation |
|---|---|---|
| **opener** | 5.28 | 4.5 |
| **harengue** | 25.18 | 17.48 |
| **peroration** | 15.35 | 14.33 |
| **closer** | 13.69 | 6.91 |

neglect skewness for our purpose. Computing intervals for each class, we have: [1,15] for *opener*, [1,28] for *closer*, [1,60] for *harengue* and [1,43] for *peroration*. Therefore we have experimented with window sizes ranging from 15 to 60 words, and defined two possible values for a majority: 60% or 80%. For example, a window of size 10 will require at least 6 or 8 words of the same class to warrant switching the class of a middle-word.

The best results were obtained using a window size of 20 words and a majority of 60%. We evaluated the performance of this smoothed classifier (also at word level) on the same 100 letters. As can be observed in table 8, the scores are higher than the results without smoothing. In particular, accuracies are clearly above what could be expected from a random baseline (five classes → 20%), at the same time they are slightly above a majority baseline (*free* → 91% ) or an "average" baseline (89%)[10]. The competitive figures for F-scores from tables 6 and 8 with regards to the four classes of interest are very encouraging but somewhat surprising, given that we only have a small data set to work with and we do not use automatic pre-processing of the data. We conjecture that using POS information may further improve the system's performance.

**Table 8** F-scores and overall accuracy for text segmentation after smoothing

| *n*-gram | F-scores% | | | | | | Overall |
|---|---|---|---|---|---|---|---|
| used | op | har | per | clo | free | Microaver. | accuracy |
| {1} | 18.3 | 13.0 | 3.3 | 28.9 | 92.3 | 84.5 | 86.9 |
| {2} | 26.3 | 19.4 | 9.1 | 32.9 | 91.7 | 84.7 | 86.6 |
| {3} | 25.1 | 7.1 | 6.8 | 17.9 | 94.3 | **86.0** | **92.4** |
| {12} | 8.8 | 17.6 | 8.6 | 30.7 | 79.1 | 73.2 | 65.8 |
| {13} | 15.9 | 19.9 | 7.8 | 34.8 | 90.6 | 83.7 | 83.8 |
| {23} | 26.9 | 20.3 | 9.9 | 34.7 | 91.9 | 85.1 | 86.9 |
| {123} | 9.1 | 20.2 | 9.4 | 31.1 | 79.1 | 73.4 | 66.0 |

[10] The average baseline is obtained by segmenting the text in the canonical order: *opener*, *harengue*, *free*, *peroration* and *closer*. Then the size of each segment is computed in accordance with the average size of each segment in a text: 1%, 8%, 83%, 4% and 4% respectively.

Even though our approach does not produce truly realistic segments with no discontinuities, it represents indeed a good starting point. Results from tables 6 and 8 also suggest the following interesting observations:

- In general, larger n-grams can make a better discrimination between the five classes.
- *Free* and *closer* are the classes which can be discriminated best, while *opener* and *peroration* are most difficult.
- Smoothing clearly improves classification performance for all classes. In particular, it improves the F-measure micro average by 26% and the overall accuracy by 43%.

An evaluation of the approach at the segment level would give a clearer picture of the results. However, to our knowledge there is no evaluation metric for segmentation using multiple classes, so future work should look at how existing metrics such as WindowDiff [33] or those proposed in [16] could be adapted for this task.

## 5 Semantic Tagging

The second task we investigated was to do a further semantic labeling of the letters' formulaic parts. Approximately 56% of the words in the training set has been labeled with a one of the 15 semantic classes listed in table 3. The other part of the words, not assigned a semantic label, are labeled 'O' (outside). Only 515 words in the training set are ambiguous and occur with more than one possible label, on average they have 3.3 different labels. 155 of these ambiguous words occur at least 10 times in the training set, this low number again confirms the small size of the data set we are working with.

In a general supervised WSD approach, contextual features are the most important information source [22]. In such a standard WSD approach, the words left and right from the target words, their POS and/or lemma information are represented as a feature vector for training a machine learning algorithm. We adopted this strategy and used local context information of neighboring words as our main information source in a ML approach. We used the POS tagger described in Sect. 3 to predict POS tags. We created a feature vector representing the context of each word in the text: three neighboring words and predicted POS tags left and right of the target word (denoted as *313*). As our data is sparse, we tried to grasp the general orthographic properties of the target words in set of standard orthographic binary features: starts-with-capital, all-capitals, all-lower-case, contains-number. We also used the feature word length and prefix and suffix information of the target word (strings of 2 and 3 characters) as features. As an additional information source we know for each word in which segment type it occurs; whether it occurs in a *opener, harengue, peroration* or *closer* segment. We investigated whether the segment type is a useful feature for the semantic labeling task. For each target word and for each neighboring word in the local context, we added an additional feature specifying

the segment type. We used the manually annotated segment type information in our experiments.

We evaluated the different features in the following way. We ran 10-fold cross validation experiments on the training set with three different classifiers and four feature set combinations and measured micro-average F-scores. We chose the following classifiers: Naive Bayes (NB), Decision trees (J48) and Instance-based learning with a search scope of 3 inverse-distance-weighted nearest neighbors (IB3) as implemented in the Weka toolkit [20]. We tried the following feature combinations: the target word and POS and local context (*313*), the addition of the orthographic (*313.orth*), and suffix/prefix information of the target word (*313.orth.fix*), and all these features combined together with the formal segment class features (*313.orth.fix.seg*).

The type of semantic labeling is related to coarse-grained word sense disambiguation but also has ties to multi-word expression recognition. Each word is assigned a semantic tag, but often these words are part of situational fixed expressions as the formal part of the letters consist mostly of ritual utterances [29]. The average length of the labeled semantic expressions is 3.4 tokens (measured on the training set). As the semantic labeling task also has this sequential aspect, we decided to exploit a more sequential oriented approach that is commonly used for more syntactic tasks such as POS tagging or chunking [34], namely using a specific algorithm that produces structured predictions. We tested the following algorithm designed for sequence modeling: conditional random fields (CRF) [25][11]. We use CRF with a feature set of 3 words and POS tags left and right of the target word and ran 10-fold cross validation experiments on the training set[12].

The results of the experiments are presented in Table 9. The structured predicting algorithm, CRF, performs slightly better than the other three classifiers when using the same local context information. For the other three classifiers we also varied the feature sets which had different effects for the different classifiers. Adding the orthographic features only has a marginal positive effect. The prefix and suffix information improved performance for IB3 and Naive Bayes, but not for Decision Trees. Adding segment class features leads to better scores for IB3, but decreases results for Naive Bayes. Overall, IB3 with all features achieves the highest result of 68.4 F-score.

For the experiments on the test set we computed two baselines shown in Table 10. For computation of the *majority baseline* we assigned all words in the test set the same label that has the highest frequency on the training set, the label 'O'. The second baseline is a more challenging version of the majority baseline (*Majority per word label*): we assigned each word in the test set its individual most frequent label as it occurred in the training set.

We evaluated both the CFR algorithm and IB3 with all features on the test set. The detailed precision, recall and F-scores for each of the semantic classes separately are shown in Table 11 and 12 respectively. On the test set, the scores

[11] We used CRF++, version 0.54 written by Taku Kudo.

[12] We experimented with more complex features such as bigrams and larger dependencies, but this did not lead to better results.

**Table 9** Average F-scores of 10 fold cross validation experiments on the training set with Naive Bayes (NB), Decision trees (J48) and instance-based learning with k=3 (IB3) with four different feature set variants and conditional random fields (CRF)

| feature set | NB | J48 | IB3 | CRF |
|---|---|---|---|---|
| 313 | 62.3 | 64.5 | 61.2 | 65.9 |
| 313.orth | 62.7 | 64.7 | 62.6 | - |
| 13.orth.fix | 65.9 | 64.7 | 67.4 | - |
| 313.orth.fix.seg | 64.2 | 64.8 | 68.4 | - |

**Table 10** Baseline recall, precision, F-scores and accuracies on the test set

| baseline | precision | recall | F-score | accuracy |
|---|---|---|---|---|
| Majority label | 22.4 | 47.3 | 30.4 | 47.2 |
| Majority per word label | 59.9 | 61.6 | 60.7 | 61.6 |

for IB3 are slightly lower, an accuracy of 64.36% and a micro-averaged F-score of 64.4%. The CRF achieves a slightly higher performance with an accuracy score of 66.7% and a F-score of 66.3%. We compared these results against the baseline scores in Table 10. To check whether the classifiers perform significantly better than the baselines, we computed significance intervals for the F-scores with bootstrap resampling [32] using 250 random samples. When the interval scores of the classifiers do not overlap with the interval of the baselines, the scores can be regarded to be significantly different ($p<0.05$). We found a significant difference between CFR and both of the baselines. IB3 however, only has a significant difference with the lower Majority baseline.

The individual semantic class label with the highest score is the most frequent class 'O'. Some semantic classes such as *I1* or *Q2.2* hardly occur at all in our data set, and the scores are consequently zero. For low frequent classes we can expect low scores due to data sparsity. The class *E4* has a remarkably low score considering that it occurs quite frequently. When we look at the confusion matrix, we can observe that this class is confused with almost all other classes; many other class examples have been wrongly predicted as *E4*, and also the other way around. Also *E2* has a low F-score. Also the baseline scores for these two classes are low, around 13% F-score. As both of these belong to the same broad semantic category 'Emotion', our results suggest that this class is particularly difficult to classify.

## 6 Conclusions

We presented our first results on automatic text segmentation and semantic labeling for historical letters. We produced an automatic text segmenter that distinguishes between informal free text and four formal parts of the letters. We also created

**Table 11** Recall, precision and F-scores on the test set with CRF for the USAS classes. The second row presents the label frequency in the test set

| label | O | a9 | b2 | e2 | e4 | i1 | q11 | q22 | s111 | s124 | s31 | s4 | s71 | s72 | s9 | x1 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| freq | 1686 | 1 | 174 | 57 | 102 | 1 | 145 | 9 | 43 | 50 | 75 | 174 | 50 | 393 | 479 | 127 | 3566 |
| rec | 86.0 | 0 | 55.2 | 10.5 | 4.9 | 0 | 30.3 | 0 | 20.9 | 72.0 | 34.7 | 42.0 | 30.0 | 67.9 | 63.7 | 37.8 | 66.7 |
| prec | 71.0 | 0 | 64.9 | 33.3 | 20.8 | 0 | 63.8 | 0 | 81.8 | 51.4 | 86.7 | 72.3 | 71.4 | 50.7 | 71.6 | 61.5 | 65.8 |
| F | 77.8 | 0 | 59.6 | 16.0 | 7.9 | 0 | 41.1 | 0 | 33.3 | 60.0 | 49.5 | 53.1 | 42.3 | 58.0 | 67.4 | 46.8 | 66.3 |

**Table 12** Recall, precision and F-scores on the test set with IB3 for the USAS classes

| label | O | a9 | b2 | e2 | e4 | i1 | q11 | q22 | s111 | s124 | s31 | s4 | s71 | s72 | s9 | x1 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| rec | 77.8 | 0 | 57.5 | 19.3 | 8.8 | 0 | 55.9 | 0 | 27.9 | 70.0 | 40.0 | 42.5 | 48.0 | 64.6 | 62.8 | 40.9 | 64.4 |
| prec | 77.9 | 0 | 54.9 | 35.5 | 17.6 | 0 | 50.3 | 0 | 48.0 | 49.3 | 51.7 | 54.8 | 61.5 | 44.6 | 66.4 | 54.2 | 64.3 |
| F | 77.8 | 0 | 56.2 | 25.0 | 11.8 | 0 | 52.9 | 0 | 35.3 | 57.9 | 45.1 | 47.9 | 53.9 | 52.8 | 64.6 | 46.6 | 64.4 |

a classifier that can predict semantic labels for words and fixed expressions in the formal parts. We tried to overcome the main stumbling block of the data set: data sparsity due to the small size of the data set and enlarged by the spelling variation present in the data. We applied a dictionary-based text normalization and replaced names with a placeholder. We achieved 86% micro av. F-score on text segmentation and 66.3% on the semantic labeling task. Our current results are not good enough that it can replace the need for manual annotation completely, but they are a promising result.

It is worth mentioning that the n-gram extraction as described in Sect. 4, combined with semantic labeling information, already gave us a valuable resource for further pragmatic studies. An analysis of the most salient n-grams allows us to make the following general comments on the pragmatic function of the formal parts:

- *opener*: salience of the semantics of social respect expressed by nominal addressing forms of deference (for example *Your Excellency*)
- *harengue*: salience of the semantics of health, combined with psychological verbs and phatic expressions, also typical of wishes of good health in the beginning of spoken dialogues (for example *I hope you are in good health*)
- *peroration*: salience of the semantics of religion, combined with phatic expressions, also typical of the God-invoking behavior in the ending of spoken dialogues (for example *May God be with you*)
- *closer*: once again, salience of the semantics of social respect, expressed here by adjectival and nominal forms of auto-derision (for example *I am your humble servant*).

There is much room for improvement. We applied a text normalization step to reduce word variation. A next step will be to develop a lemmatizer or stemming algorithm that is suited for historical Portuguese text. As Portuguese is a highly inflectional language, lemma information can reduce data sparsity. For the text

segmentation task we would like to further investigate alternative smoothing techniques and more advanced methods to compute the final segmentation, including lexical models of the shifts between segments. Also, a future version of the system should take into account constraints drawn from the inspection of the training data: there should be no more than one *opener*, one *closer*, maximally two *harengues* and maximally two *perorations*, and the sequence ordering must be *opener*, *harengue*, *peroration* and *closer*. For semantic labeling we presented a first exploration with a sequence modeling algorithm, conditional random fields. For both tasks we would like to explore sequence modeling algorithms to a deeper extent and integrate both tasks into a genuine sequential process. As we have such a small annotated data set, applying techniques to incorporate unlabeled data as additional training material such as co-training [6] could be an effective solution.

**Acknowledgements** We would like to thank Mariana Gomes, Ana Rita Guilherme and Leonor Tavares for the manual annotation. We are grateful to João Paulo Silvestre for sharing his electronic version of the Bluteau Dictionary and frequency counts. This work is funded by the Portuguese Science Foundation, FCT (Fundação para a Ciência e a Tecnologia).

## References

1. Archer, D., Culpeper, J.: Identifying key sociophilological usage in plays and trial proceedings): An empirical approach via corpus annotation. Journal of Historical Pragmatics **10**(2), 286–309 (2009)
2. Archer, D., McEnery, T., Rayson, P., Hardie, A.: Developing an automated semantic analysis system for early modern english. In: Proceedings of the Corpus Linguistics 2003 conference, pp. 22 – 31 (2003)
3. Baron, A., Rayson, P.: VARD2: A tool for dealing with spelling variation in historical corpora. In: Proceedings of the Postgraduate Conference in Corpus Linguistics (2008)
4. Baroni, M., Bernardini, S.: Bootcat: Bootstrapping corpora and terms from the web. In: Proceedings of Language Resources and Evaluation (LREC) 2004, pp. 1313–1316 (2004)
5. Blecua, A.: Manual de Crítica Textual. Castalia, Madrid (1983)
6. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers (1998)
7. Bluteau, R.: Vocabulario portuguez, e latino [followed by] supplemento ao vocabulario portuguez. vols. 1-8, I-II. Coimbra-Lisboa. (1712–1728)
8. Brown, P., Levinson, S.C.: Politeness: some universals in language usage. Cambridge University Press, Cambridge (1987)
9. Cohen, J.: A coefficient of agreement for nominal scales. Education and Psychological Measuremen **20**, 37–46 (1960)
10. Daelemans, W., A.Van den Bosch: Memory-Based Language Processing. Cambridge University Press, Cambridge, UK (2005)
11. Daelemans, W., Zavrel, J., Van den Bosch, A., Van der Sloot, K.: Mbt: Memory-based tagger, version 3.1, reference guide. Tech. rep., ILK Technical Report Series 07-08 (2007)
12. Dossena, M., van Ostade, I.T.B. (eds.): Studies in Late Modern English Correspondence. Peter Lang, Bern (2008)
13. Edmonds, P., Kilgarriff, A.: Introduction to the special issue on evaluating word sense disambiguation systems. Natural Language Engineerin **8**(4), 279–291 (2002)

14. Ernst-Gerlach, A., Fuhr, N.: Retrieval in text collections with historic spelling using linguistic and spelling variants. In: Proceedings of the ACM/IEEE-CS conference on Digital libraries, pp. 333–341 (2007)
15. Everitt, B.: The Analysis of Contingency Tables, 2nd edn. Chapman and Hall (1992)
16. Ferret, O.: Segmenter et structurer thématiquement des textes par l'utilisation conjointe de collocations et de la récurrence lexicale. In: TALN 2002. Nancy (2002)
17. Fitzmaurice, S.M.: Epistolary identity: convention and idiosyncrasy in late modern english letters. In: Studies in Late Modern English Correspondence, pp. 77–112. Peter Lang (2008)
18. Guillén, C.: Renaissance Genres: Essays on Theory, History and Interpretation, chap. Notes towards the study of the Renaissance letter, pp. 70–101. Harvard University Press (1986)
19. Hachey, B., Grover, C.: Extractive summarisation of legal texts. Artificial Intelligence and Law: Special Issue on E-government **14**, 305–345 (2007)
20. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explorations **11**(1) (2009)
21. Hearst, M.A.: Texttiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics **23**(1), 33–64 (1997)
22. Jurafsky, D., Martin, J.H.: Speech and Language Processing. 2nd edition. Prentice-Hall (2009)
23. Kilgarriff, A., Palmer, M.: Introduction to the special issue on senseval. Computers in the Humanities **34**(1-2), 1–13. (2000)
24. Koolen, M., Adriaans, F., Kamps, J., de Rijke, M.: A cross-language approach to historic document retrieval. In: Advances in Information Retrieval: 28th European Conference on IR Research (ECIR 2006), *LNCS*, vol. 3936, pp. 407–419. Springer Verlag, Heidelberg (2006)
25. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conf. on Machine Learning, pp. 282–289. Morgan Kaufmann, San Francisco, CA (2001)
26. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. Sovjet Physics Doklady **10**, 707–710 (1966)
27. Merity, S., Murphy, T., Curran, J.R.: Accurate argumentative zoning with maximum entropy models. In: NLPIR4DL '09: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, pp. 19–26. Association for Computational Linguistics, Morristown, NJ, USA (2009)
28. Mikheev, A.: Periods, capitalized words, etc. Computational Linguistics **28**, 289–318 (1999)
29. Moon, R.: Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford University Press, Oxford (1998)
30. Nevalainen, T., Tanskanen, S.K. (eds.): Letter Writing. John Benjamins Publishing Company, Amsterdam/Philadelphia (2007)
31. Ng, H.T., Lim, C.Y., Foo, S.K.: A case study on inter-annotator agreement for word sense disambiguation. In: Proceedings of the SIGLEX Workshop On Standardizing Lexical Resources (1999)
32. Noreen, E.W.: Computer-Intensive Methods for Testing Hypotheses. John Wiley & Sons (1989)
33. Pevzner, L., Hearst, M.A.: A critique and improvement of an evaluation metric for text segmentation. Comp. Linguistics **28**, 1–19 (2002)
34. Ramshaw, L., Marcus, M.: Text chunking using transformation-based learning. In: Proceedings of the Third Workshop on Very Large Corpora, pp. 82–94 (1995)
35. Rayson, P., Archer, D., Piao, S.L., McEnery, T.: The UCREL semantic analysis system. In: Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks (LREC 2004), pp. 7–12 (2004)
36. Reynar, J.C., Ratnaparkhi, A.: A maximum entropy approach to identifying sentence boundaries. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 16–19 (1997)
37. Sporleder, C., Lapata, M.: Broad coverage paragraph segmentation across languages and domains. ACM Transactions on Speech and Language Processing **3**(2), 1–35 (2006)
38. Teufel, S., Moens, M.: What's yours and what's mine: Determining intellectual attribution in scientific text. In: In EMNLP-VLC (2000)
39. Watts, R.: Politeness. Cambridge University Press, Cambridge (2003)

# Proppian Content Descriptors in an Integrated Annotation Schema for Fairy Tales

Thierry Declerck, Antonia Scheidel and Piroska Lendvai

**Abstract** This chapter describes the actual state of APftML (Augmented Proppian fairy tale Markup Language), which is a schema combining linguistic and domain specific annotation for supporting Cultural Heritage and Digital Humanities research, exemplified in the fairy tale domain. APftML should in particular guide automated text analysis to detect and mark up fairy tale characters and the typical actions they are involved in, which can be subsequently queried in a corpus by both linguists and specialists in the field. The characters and actions are defined with the help of Propp's formal analysis of folktales, which we aim to implement in a fully fledged way, contrary to existing computational resources based on his theory. In order to respond to current formalisation requirements APfML abstracts away from some aspects of the theory of Propp, and we also discuss the integration of Proppian elements within modern semantic annotation approaches. The chapter focuses on the resulting revised and extended set of narrative elements APftML is dealing with.

**Key words:** linguistic and semantic annotation schema for narratives, Propp

Thierry Declerck
DFKI GmbH, Language Technology Lab, Stuhlsatzenhausweg, 3, D66123 Saarbrücken, Germany. e-mail: declerck@dfki.de

Antonia Scheidel
Universität des Saarlandes, Departement of Computational Linguistics and Phonetics, Postfach 15 11 50, D66041 Saarbrücken, Germany. e-mail: ascheidel@coli.uni-sb.de

Piroska Lendvai
Research Institute for Linguistics, Hungarian Academy of Sciences, Benczúr u. 33., H-1068 Budapest, Hungary. e-mail: piroska@nytud.hu

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing, DOI 10.1007/978-3-642-20227-8_9, © Springer-Verlag Berlin Heidelberg 2011

## 1 Introduction

The work described in this chapter is motivated by the long-term objective of devising a means for linguistic processing of folktale texts in order to support their automated semantic annotation in terms of Propp's theory (see [7]), using an appropriate encoding schema which is presumably also adaptable to other theories of fairy tales or literary genres. As a starting point, we analysed the utility of available computational resources targeting the Proppian analysis, like [5] and [6]. In [3] and [4] we discuss some limitations of those approaches, in which we mainly notice the lack of linguistic considerations, and we suggest an annotation approach for linguistic information that is based on TEI[1] and on the stand-off multilayer annotation strategy of ISO TC37/SC4 on language resources management[2], as the basis for an annotation schema that integrates linguistic knowledge, Proppian elements and also recent semantic annotation approaches.

The resulting Augmented Proppian fairy tale Markup Language (APftML) is a flexible multi-layer annotation scheme for fairy tales that allows for fine-grained annotation, aligning the textual structure of the fairy tale with its "Proppian interpretation", i.e. properties of a tale's characters, Proppian functions, and narration assigned to textual segments. As current formalisation requirements necessitate, APfML abstracts away from some aspects of the theory of Propp: in order to make the markup language as flexible and generally applicable as possible, a number of rules concerning the composition of Russian fairy tales are excluded. On the other hand, APftML includes Proppian concepts which have not been considered in most of the previous approaches to *Morphology of the Folktale*. In this chapter we present this revised set of narrative elements we deal with in APftML, including a discussion on the integration of narrative interpretation in recent semantic annotation frameworks.

## 2 Summary of Propp's Analysis

Propp introduces the following concepts for the interpretation of Russian fairy tales:

Characters

Main characters (or *dramatis personae*) occuring in a fairy tale may be[3]:

1. Hero: a character that seeks something;
2. Villain: opposes or actively blocks the hero's quest;

[1] TEI stands for Text Encoding Initiative. See http://www.tei-c.org/

[2] See http://www.tc37sc4.org/

[3] Source: http://www.adamranson.plus.com/Propp.htm

3. Donor: provides the hero with an object of magical properties;
4. Dispatcher: sends the hero on his/her quest via a message;
5. False Hero: disrupts the hero's success by making false claims;
6. Helper: aids the hero;
7. Princess: acts as the reward for the hero and the object of the villain's plots;
8. Her Father: acts to reward the hero for his effort.

## 31 Functions

At the heart of the *Morphology of the Folktale* lies the description of actions that can be performed by the dramatis personae of a folktale. These so-called *functions* are the prototypical invariant features of fairy tales such as "Conflict", "Call for Help", "Kidnapping", "Test of Hero", and so on. Propp considers those functions as being ordered, as Fig. 1 shows.

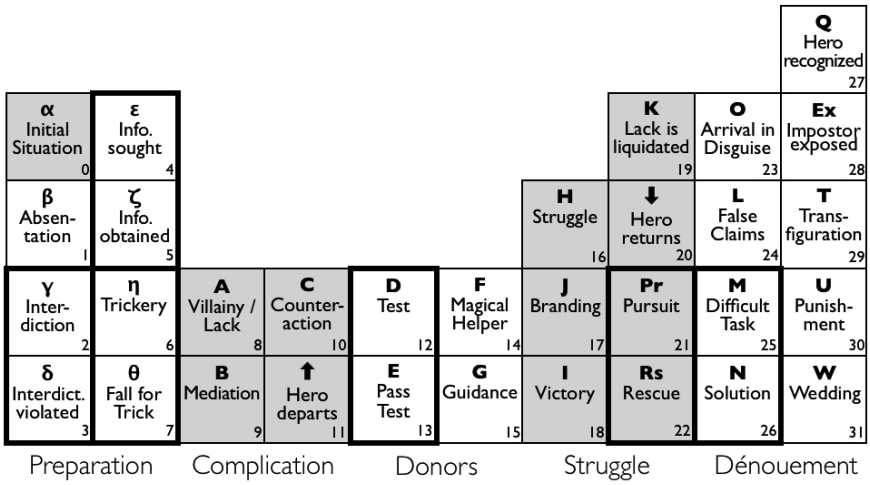


**Fig. 1** The ordered list of Proppian functions

Functions are frequently divided into sub-functions: in the case of function *A: Villainy*, they range from $A^1$*: The villain abducts a person* to $A^{19}$*: The villain declares war*. Functions and subfunctions are described in detail and illustrated with examples from Russian folktales in [7].

A sequence of all the functions from one folktale is called a *scheme* which can be used as a formal representation of the tale (see Fig. 2 for an example).

$$\gamma^1\beta^1\delta^1A^1C\uparrow\{[DE^neg.Fneg.]^3d^7E^7F^9\}G^4K^1\downarrow[Pr^1D^1E^1F^9=Rs^4]^3$$

**Fig. 2** Functional scheme for *The Magic Swan-Geese*

150 Elements

In Appendix I of *Morphology of the Folktale*, Propp provides what he calls a "list of all the elements of the fairy tale". The list contains 150 elements, distributed over six tables:

1. Initial Situation
2. Preparatory Section
3. Complication
4. Donors
5. From the Entry of the Helper to the End of the First Move
6. Beginning of the Second Move

Some of the 150 elements appear alone, others are grouped under a descriptive heading. If these "element clusters" (as shown in Fig. 3) are counted as one, the appendix contains 56 - as they shall tentatively be called in the following - narratemes. About a third of the narratemes can be mapped directly to functions, such as *30-32. Violation of an interdiction*, which is also the content of Fig. 3. Other narratemes can be combined to form an equivalent to a function (together, narratemes *71-77: Donors* and *78: Preparation for the transmission of a magical agent* can presumably be considered as a superset to the information expressed by function *D: First Function of the donor*.

```
30-32. Violation of an interdiction
     30. person performing
     31. form of violation
     32. motivation
```

**Fig. 3** Example for a narrateme

Another group of narratemes, however, goes beyond the 31 functions: *70. Journey from home to the donor*, for example, can be seen as filling the gap between the functions ↑*: Departure* and *D: First function of the donor*. The first table (*Initial Situation*[4]) contains a multitude of narratemes dedicated to the circumstances of the hero's birth and other events/situations which precede the actual adventure.

Furthermore, Table 1 (*Initial Situation*) includes two "element-clusters" describing the hero and false hero, respectively, in term of 'future hero' (see Fig. 4).

```
10-15. The future hero
      10. nomenclature; sex
      11. rapid growth
      12. connection with hearth, ashes
      13. spiritual qualities
      14. mischievousness
      15. other qualities
```

**Fig. 4** Example for an element cluster serving as profile

A closer examination of the appendix of *Morphology of the Folktale* reveals such 'profiles' for each of the dramatis personae, although sometimes spread over two clusters or narratemes.

## 3 Preprocessing Propp

Before introducing the components of APftML, we need to address a number of general design questions. In the process of creating the APftML schema, we analysed a number of "rules" and directions Propp gives in *Morphology of the Folktale*, including some in the schema and discarding others. These decisions shall be documented in the following.

### *3.1 Relaxing the "Fairy Tale Grammar"*

One of the major findings Propp introduces in [7] is that "The sequence of functions is always identical". This is reflected in the numbering of functions (with $\alpha$*: Absentation* as the first and *W: Wedding* as the last function (see Fig. 1)), which is to be interpreted as the order of functions within each move of a fairy tale.

Function pairs usually have adjacent function numbers (and / or symbols), like $\gamma$: *Interdiction* (Function number 2) – $\delta$: *Violation* (number 3), Pr: *Pursuit* (21) – Rs: *Rescue* (22), etc. In the case of function pairs, the first half entails the second half of the pair. Propp stresses that an interdiction which is not violated is of no importance for the development of the tale; it is the combination of an authority figure making and the hero breaking a rule which sets the adventure in motion.

This is why the APftML schema strongly encourages linking between function pairs: More precisely, we want each second half of a function to point back to its first half, as in

```
<tns:Violation type="violated" id="delta0" ref="gamma0">
```

[4] Propp makes use of the symbol $\alpha$*: Initial Situation* to refer to everything that happens before the hero's parents announce their departure, but it is not a function as such.

where *gamma0* indicates the interdiction being violated.

However, APftML does not enforce the canonical order of functions suggested by Fig. 1[5].

It must be noted that the Proppian order of functions does not necessarily follow the actual temporal progression of events in the fairy tale, but the narration: The most prominent example is function $\alpha$*: Absentation*. In many fairy tales, the departure of the parents is preceded by both an announcement of their intention to leave and an interdiction (e.g. not to leave the house). Propp seems to assign function $\beta$*: Absentation* to the announcement of the absentation rather than the actual departure, resulting in the sequence $\beta, \gamma, \delta$, even though the event of the parents absenting themselves naturally happens in between the interdiction and its violation.

### *3.2 Functions and Moves*

The **move** is an important concept in Propp's study of fairy tale structure. A move is defined as a development leading from a villainy or lack (the eighth function, A) through intermediary functions to a terminal function; typically marriage (W), but possibly also a reward (F), the liquidation of a lack (K) or the escape from pursuit (Rs).

As we distance ourselves from the Proppian order of functions, the concept of **moves** also becomes subject to discussion. Moves do not necessarily ooccuring ne after the other; in the contrary, we notice that there are as much as six ways in which moves may be nested, interrupted and resumed at a later stageof the narrative or simply interweaved, making it very hard to determine which function belongs to which move. As a consequence, we have decided not to include moves in the APftML schema for the time being.

## 4 Functions and Frames

### *4.1 Proppian "Frames" and FrameNet*

We encountered in the appendix of [7] many elements that refer to the semantic roles of lexical units, with some element clusters bearing a distinct resemblance to (FrameNet) *frames*. The element cluster linked to function $\gamma$*: Interdiction*, for example, has its counterpart in FrameNet, Frame "Deny permission"[6] (see Table 1).

While we did briefly consider building a "Propp-compliant" subset of FrameNet for use in APftML, discarding the highly specific frames available in FrameNet and

[5] And looking in details at the appendix of *Morphology of the Folktale*, where Propp presents functional schemes for 49 fairy tales, we see that it already contains a number of deviations from

**Table 1** Comparison of a Proppian "element cluster" and FrameNet frame in regard to the respective definitions of typical semantic roles.

| | Proppian "Frame" | FrameNet Frame |
|---|---|---|
| Name | Interdiction | Deny_permission |
| Agent role | person performing | Authority |
| Patient role | receiver of the interdiction (inferred) | Protagonist |
| Theme role | contents | Action |

settling on a small set of abstract semantic roles (Agent, Patient, Theme etc., cf. [2]) proved to be more rewarding.

A more abstract approach seems to suit the nature of some Proppian functions especially well: They are, typically, the second parts of function pairs: While the interdiction from function $\gamma$ always corresponds to the frame for denying permission, the frame associated with the violation of the interdiction cannot be predefined. It depends entirely on the action at the "theme" (or "Action") position of the interdiction frame. Furthermore, we are not content with simply "pointing" to the action that is being forbidden – we want to be able to express connections like the one in Fig. 5.

In order to achieve this capability, we introduce a set of "APftML Frames":

## *4.2 APftML Frame Elements*

Every function in APftML was given a **Frame** element in an attempt to align functions to the actual structure of their corresponding sentences and paragraphs. The alignment was accomplished by the use of three different types of Frames:

Frames

The most basic type of Frame serves as the representation of a simple sentence. They contain elements for each relevant semantic role, i.e., Agent, Patient, etc., and an element tentatively named "CentralInformation". This element contains a reference to the lexical unit the Frame belongs to, e.g., "kidnapped" in the context "The swan-geese kidnapped the little brother".

Frames can be divided into two types with regard to the spatial information they contain, "motion frames" and "stationary frames": Motion frames have two slots, **Source** and **Goal**, and can be used to represent actions which include movement

the order. Considering that we want to use APftML for a larger group of narratives, we deemed it advisable to allow as much flexibility in regard to the combination of functions as possible.

[6] http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=118&frame=Deny_permission

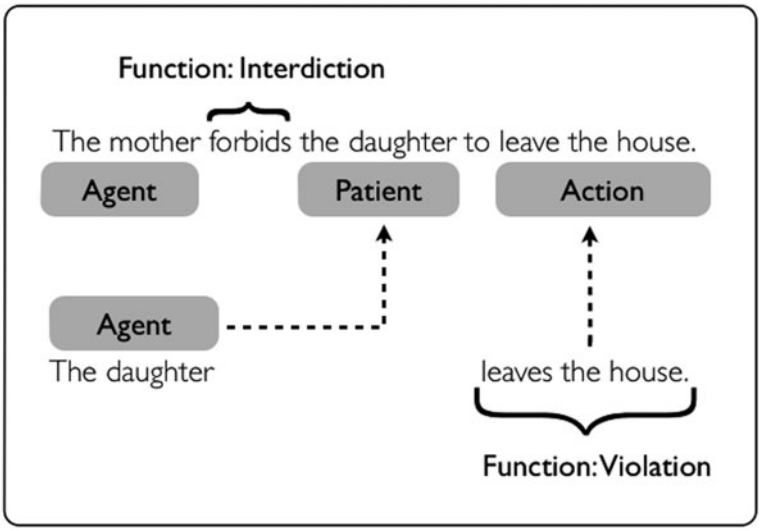


**Fig. 5** When an interdiction is violated, the patient of the interdiction becomes the agent of an action which corresponds to the theme of the interdiction

from one place to the other[7]. Stationary frames have one **Location** slot and can describe actions confined to one place.
Every frame must either contain a reference to a time interval in the form of a **Time** element or be part of a complex frame (see ConditionalFrames and InfluenceFrames below) where a time interval is specified.

ConditionalFrames

ConditionalFrames contain two simple frames, with the first indicating the action that must be carried out as a condition for the action in the second frame (the consequence). This frame is used in Donor sequences, where the hero must often accomplish a task before the donor returns a favour.

InfluenceFrames

When an interdiction is addressed to the hero, our focus is not so much on the utterance of the interdiction but on the *action* the hero should refrain from.

[7] e.g. flying, running, but also kidnapping – which is not a movement per se, but includes the spatial transference of the victim

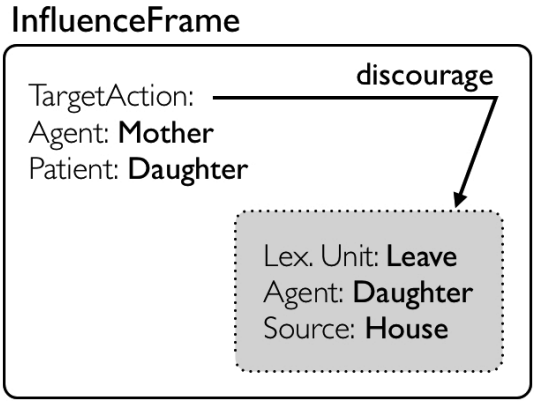


**Fig. 6** A simplified representation of the InfluenceFrame used to express the interdiction not to leave the house, addressed by the mother to the daughter. Other relations to the TargetAction could be "encourage" (amounting in a request or an order instead of an interdiction), "learnAbout" (e.g. for the situation where a sister discovers that her brother has been kidnapped) and "informAbout" (e.g. the hero naively revealing his plans to the villain)

Aiming to find a way to refer to that forbidden action at a later point, we introduced InfluenceFrames. Figure 6 is an example illustrating the concept behind an InfluenceFrame.

### *4.3 Functional Annotation*

Apart from Frames, functions in APftML (see Fig. 7 for a more detailed view on the XSD design of the function element) contain a number of other components, which shall be introduced in the following.

Context – the Span Element

Clear borders between functions are rare: There are, naturally, a number of hints in the content of the fairy tale (such as changes in location or the introduction of a new character) as well as in the syntactic structure (paragraph or sentence borders). However, an excessively strict interpretation of these hints will lead to a functional annotation too coarse-grained to cover cases like single sentences that contain two functions, such as this excerpt from *The Magic Swan-Geese* (Fig. 8):

The first half of the sentence can be classified as function *K: Liquidation of Lack* (incidentally also implying function *↓: Return*), the second as function *Pr: Pursuit*. In APftML, we ask annotators to indicate the **context** of a function (The corresponding element in the schema is named "Span".) – which may include one or more sentences or even a whole paragraph. While the context or span of a function

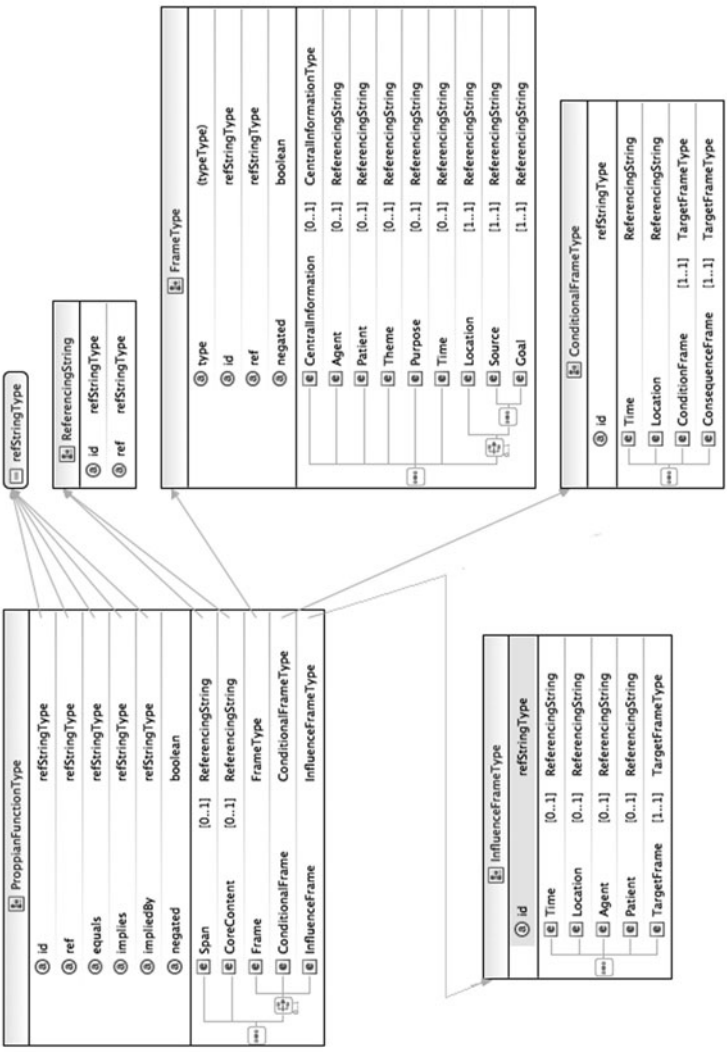


**Fig. 7** XSD design of the generic type for Proppian functions

must be given at least at sentence level (as opposed to clauses or phrases), smaller units may be used in the next APftML component:

Pinpointing Function "Cores" – the CoreContent Element

In the excerpt above, the liquidation of lack (i.e., the retrieval of the little brother) is brought about by the sister seizing him. Consequently, the APftML annotation of

His sister saw [the little brother], crept near him, seized him and carried him away, and the geese flew after her in pursuit; the evil-doers were overtaking them; where was there to hide?

**Fig. 8** Excerpt from *The Magic Swan-Geese*: The little girl has found her little brother in Baba Yaga's hut. She attempts to rescue him but is pursued by the swan-geese.

```
<tns:Relations>
  <tns:FamilyRelation type="wife" ref="ch1" />
  <tns:FamilyRelation type="daughter" ref="ch2" />
  <tns:FamilyRelation type="son" ref="ch3" />
  </tns:Relations>
```

**Fig. 9** A "Relations" element

the "Liquidation of Lack" function would have the elliptical clause "seized him" as its **CoreContent** element. CoreContent for the pursuit function is "the geese flew after her in pursuit".

## 5 Fairy Tale Characters

The discovery of the seven dramatis personae is one of the major findings in [7] – and Propp devotes a large part of the appendix to the description of character attributes. We have therefore decided to dedicate an annotation layer to the characters appearing in a fairy tale. For each character, we compile a profile, loosely based on the character profile Propp outlines in chapter VIII of [7]:

- **References**: A list of referring expressions to the character in question ("an old man", "the daughter", etc.). Includes references to groups of which the character is a part, e.g., "the parents" for a father.
- **First Appearance**: Propp puts special emphasis and attention to the manner in which a character is introduced. We store the first mention of a character in this element.
- **Attributes**: Baba Yaga's bony leg, the old age of a father, the beauty of a princess, etc. This element serves as a container for all the more specific attributes Propp included in the "list of fairy tale elements", such as "negative qualities" of a false hero or "spiritual qualities" of a hero.
- **Relations**: Propp predicts a hero to have a family and a villain to (possibly) have a retinue. We therefore allow characters to be connected to others through **familyRelation**s or **hierarchyRelation**s (cf. Fig. 9).
- **Functions**: Finally, we create a list of all functions the character took part in and store the information on the role they played, i.e., whether they were, for example, in the Agent or the Patient position of the function.

## *5.1 Characters vs. Dramatis Personae*

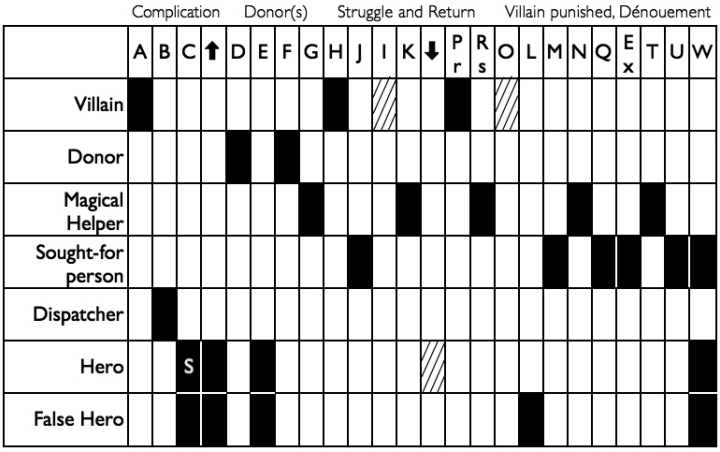


**Fig. 10** The spheres of action of the dramatis personae, modelled after chapter VI of [7]. It must be noted that three post-preparation functions are missing from this chapter although the attribution of characters seems obvious: *I: The villain is defeated*, ↓*: The hero returns* and *O: The hero, unrecognised, arrives home*. We have taken the liberty to add these functions to their respective spheres of action (see hatched areas).The "S" in the hero's first function, (*C: Beginning counteraction*) indicates that this function is only part of a **seeker** hero's sphere of action.

It is important to note at this point that we are, in fact, annotating information on "simple" characters without assigning them to a dramatis personae category (yet). This has several reasons:

Proppian roles like hero and villain are not attributed on the basis of character traits. A hero is neither courageous or noble by definition, nor is the villain necessarily an "evil" figure. On the contrary, folktales often feature a number of protagonists (i.e., heros) distinguished by disobedience, recklessness or even cruelty. Not even characters' actions reliably inform the reader about their role: Kidnapping is a villain's action – to which a hero may react with the very same crime, a "counter-kidnapping" ([7]). It is the **functions** a character takes part in and the functions alone which determine the character's role: The character performing a kidnapping interpreted as function $A^1$, a realisation of Villainy, is a villain. If the kidnapping is a "Liquidation of Lack" (Function *K*), the kidnapper is a hero.

The table in Fig. 10 was compiled from chapter VI of [7], *The Distribution of Functions Among Dramatis Personae.* It maps each function to the role associated with it (or vice versa – Propp speaks of the "spheres of action" of the dramatis personae). Equipped with this table and the aforementioned list of every function a character plays a role in, we can attempt to infer for each of the characters which dramatis personae category they belong to.

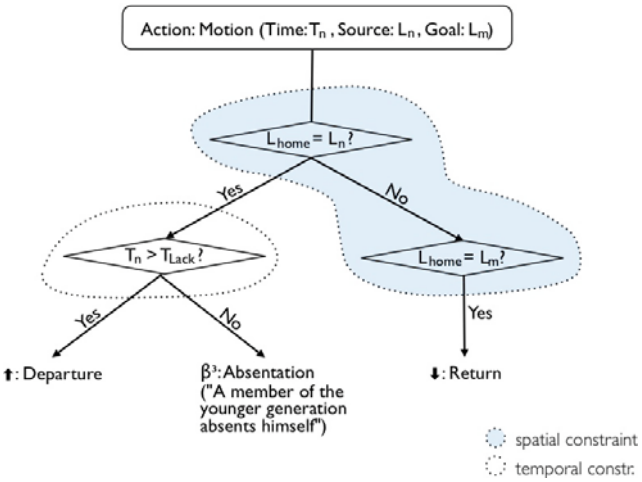


**Fig. 11** The hero moving from one location to the other can be classified at least as three different functions, depending on the spatial and temporal constraints associated with each individual function.

## 6 Temporal and Spatial Structure

Vagueness in regard to places and time is one of the characteristics of fairy tales. But even though exact references to points in time or actual locations are rare, it is certainly feasible to determine changes in location or the time intervals that make up a tale.

It is not uncommon for the cast of a fairy tale to be distributed over a number of locations (e.g., the hero is subjected to a test by the donor in the forest (location 1) while the princess is waiting for him, chained up in the dragon's lair (location 2)). However, due to the linear nature of fairy tale narration, there is no "Meanwhile, in the king's palace...": At any given point in time $t_n$, the focus is on **one** location only, even though other locations (and the characters there) can be assumed as background knowledge of both the "focus character" and the reader.

We chose to exploit this fairy tale feature for our annotation scheme and include a Location element (or Source-Goal pair) in every function's Frame element.

Furthermore, The APftML schema contains an element **TemporalStructure**, with a number of **TimeInterval** child elements. This serves as a preliminary (and largely interpretation-free) segmentation of the fairy tale.

Frame elements include a "Time" element with a reference to one of the time intervals introduced in the TemporalStructure element. Every two functions with the same "timestamp" can be interpreted as equal to each other – or, conversely, if one function $f_n$ is marked as equal to another function $f_m$, it will receive $f_m$'s timestamp. Notable exceptions from this rule are function pairs with a strong link to dialogues, e.g., most realisations of the *Information Reconnaissance – Delivery*

pair or a number of *Donor Function – Hero's Reaction* realisations, where it can be assumed that, for example, the villain's question and the hero's answer occur in the same time interval. This is one of the reasons why dialogue is annotated separately in APftML, see also 7.

The timestamps constitute our answer to the problem addressed in 3.1: Instead of attempting to replicate the order inferred from Propp's canonical numbering of the functions, the functions in APftML-annotated fairy tales can be ordered chronologically.

## 7 Dialogue and Narration

When analysing, for example, the cues obtained from verb tenses, care must be taken to distinguish narration and dialogue (or direct speech). While verb tenses in narration can usually be exploited to extract the temporal structure from a text, tenses used in dialogue cannot be interpreted in the same way. Furthermore, direct speech reflects the speaker's perception of a scene rather than the perspective given by the narrator in the rest of the tale. To account for this, the APftML schema includes an element **Dialogues**, a collection of all the dialogue that appears in the annotated fairy tale.

We have annotated dialogues in a fine-grained fashion, using the terminology introduced in [1], who describe the development of an annotation scheme for dialogue acts, DiaML[8] We display just below a dialogue between the heroine and a donor figure, taken from our annotation example, *The Magic Swan-Geese*:

```
<tns:Dialogue id="dia3">
   <tns:Span>
      "River of milk, and shores of pudding, where have the
      geese flown?"
      "If you eat my simple pudding with milk, I will tell you."
      "Oh, in my father's house we do not even eat cream."
   </tns:Span>
   <tns:Participant ref="ch7" />
   <tns:Participant ref="ch2" />
   <tns:Turn id="turn7" sender="ch2" receiver="ch7">
         <tns:Quotative type="implied" />
         <tns:Utterance id="utt11" type="salutation">
            River of milk, and shores of pudding
         </tns:Utterance>
         <tns:Utterance id="utt12" type="question">
            where have the geese flown?
         </tns:Utterance>
   </tns:Turn>
   <tns:Turn id="turn8" sender="ch7" receiver="ch2">
         <tns:Quotative type="implied" />
```

[8] The "Dialogues" element in APftML is not entirely compliant with DiaML as of now (as we want to include more information about the speakers) but could presumably be translated into that scheme if necessary.

```
        <tns:Utterance id="utt13" type="addressRequest">
               If you eat my simple pudding with milk, I will tell you.
        </tns:Utterance>
   </tns:Turn>
   <tns:Turn id="turn9" sender="ch2" receiver="ch7">
        <tns:Quotative type="implied" />
        <tns:Utterance id="utt14" type="declineRequest">
           Oh, in my father's house we do not even eat cream.
        </tns:Utterance>
   </tns:Turn>
</tns:Dialogue>
```

## 8 Conclusion

We described ongoing work in developing an integration annotation schema for fairy tales, focusing here on the revision and extension of Proppian elements and their integration in recent semantic annotation approaches. The integration of linguistic information, discussed in [4], is aiming as supported the semi-automatic semantic annotation of folktales, also in term of narrative elements. Generalizing the results will shed light on the computational applicability of Humanities resources Digital Humanities research. If we can detect genre-specific narrative units on evidence based on a fine-grained annotated corpus, we plan to take this research further and analyse higher level motifs (such as narratemes), as well as other types of narratives.

**Acknowledgements** The ongoing work described in this paper has been partially supported by the European FP7 Project MONNET (Multilingual Ontologies for Networked Knowledge), with Grant 248458, and by the BMBF project D-SPIN. Investigating higher-order content units such as motifs is the focus of the AMICUS project, which is supported by The Netherlands Organisation for Scientific Research (NWO).

## References

1. Bunt, H., Alexandersson, J., Carletta, J., Choe, J.W., Fang, A., Hasida, K., Lee, K., Petukhova, V., Popescu-Belis, A., Romary, L., Soria, C., Traum, D.: Towards a standard for dialogue act annotation. In: 7th International Conference on Language Resources and Evaluation (2010). URL http://www.lrec-conf.org/proceedings/lrec2010/summaries/560.html
2. Fillmore, C.: The case for case. In: E. Bach, R. Harms (eds.) Universals in Linguistic Theory. Holt, Rinehart, and Winston, New York (1968)
3. Lendvai, P., Declerck, T., Darányi, S., Gervás, P., Hervás, R., Malec, S., Peinado, F.: Integration of linguistic markup into semantic models of folk narratives: The fairy tale use case. In: 7th International Conference on Language Resources and Evaluation (2010)
4. Lendvai, P., Declerck, T., Darányi, S., Malec, S.: Propp revisited: Integration of linguistic markup into structured content descriptors of tales. In: Digital Humanities 2010. Oxford University Press (2010)
5. Malec, S.A.: Proppian structural analysis and XML modeling. In: Proceedings of Computers, Literature and Philology (CLiP 2001) (2001)

6. Peinado, F., Gervás, P., Díaz-Agudo, B.: A description logic ontology for fairy tale generation. In: Language Resources for Linguistic Creativity Workshop, 4th LREC Conference, pp. 56–61 (2004)
7. Propp, V.A.: Morphology of the Folktale, 2nd edn. Publications of the American Folklore Society. University of Texas Press (1968)

# Adapting NLP Tools and Frame-Semantic Resources for the Semantic Analysis of Ritual Descriptions

Nils Reiter, Oliver Hellwig, Anette Frank, Irina Gossmann, Borayin Maitreya Larios, Julio Rodrigues and Britta Zeller

**Abstract** In this paper we investigate the use of standard natural language processing (NLP) tools and annotation methods for processing linguistic data from ritual science, which is concerned with the study of structure and variance of rituals. The work is embedded in an interdisciplinary project that addresses this study by applying empirical and quantitative computational linguistic analysis techniques to ritual descriptions from Indian rituals. We present motivation and prospects of such a computational approach to ritual structure research and sketch the overall project research plan. In particular, we motivate the choice of frame semantics as a theoretical framework for the semantic analysis of rituals. We discuss the special characteristics of the textual data and examine several domain adaptation strategies in order to use standard NLP resources and tools on the ritual domain. We also report on our workflows and methods for semi-automatic semantic annotation, which is used as a basis for the extraction of event chains. We close with some preliminary investigations on how to uncover regularities and differences of rituals.

**Key words:** ritual structure, semantic analysis, event chains, domain adaptation

## 1 Introduction

Led by the observation of similarities and variances in rituals across times and cultures, ritual scientists are discussing the existence of a "ritual grammar", an abstract underlying – and possibly universal – structure of rituals, which nevertheless is subject to variation. It is controversial whether such structures exist, and if so, whether they are culture-independent or not.

Our interdisciplinary project[1] addresses this issue in a novel empirical fashion. Using computational linguistics methods, we aim at obtaining quantitative analyses of similarities and variances in ritual descriptions, thereby offering ritual scientists new views on their data.

Nils Reiter, Anette Frank, Irina Gossmann, Julio Rodrigues and Britta Zeller
Department of Computational Linguistics, Heidelberg University, Germany,
e-mail: reiter@cl.uni-heidelberg.de

Oliver Hellwig and Borayin Maitreya Larios
South Asia Institute, Heidelberg University, Germany, e-mail: hellwig7@gmx.de

[1] The project is part of a collaborative research center (Sonderforschungsbereich, SFB) on "Ritual Dynamics". Over 90 researchers from 21 scientific fields examine the structure and dynamics within and across different cultures. http://www.ritualdynamik.de

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_10, 

Ritual researchers analyze rituals as complex event sequences, involving designated participants, objects, places and times. Such sequences are usually encoded in natural language descriptions. However, the knowledge of recurrent structures in ritual event sequences is often private among researchers devoted to particular cultures or scientific fields, because an all-encompassing theoretical framework for the analysis of rituals across different cultures does not yet exist. In our work, we attempt to make characteristic properties and structures in rituals overt. For this sake, we apply formal and quantitative computational linguistic analysis techniques on textual ritual descriptions. We will investigate data-driven approaches to detect regularities and variations of rituals, based on semi-automatic semantic annotation of ritual descriptions, thereby addressing this research issue in a novel empirical fashion.

Since a ritual can be divided into complex event sequences, the computational linguistic analysis of ritual descriptions needs to focus on discourse semantic aspects: the recognition and analysis of events and roles, temporal relations between events and coreference and anaphora resolution regarding participants of these events, to name just a few. In order to capture variations and similarities across rituals, it is important to analyze and quantify variations in event successions (e.g., is a specific action accompanied by another one, or strictly followed by it?), as well as variance regarding the ontological type of participants (what kinds of materials or living beings are subject to or involved in specific actions in different roles?).

**Computational Linguistics Resources and Tools for the Analysis of Ritual Structure.** Computational Linguistics has developed a variety of resources and processing tools for semantic and discourse processing that can be put to use for such a task. The community has developed semantic lexica and processing tools for the formal analysis of events and their predicate-argument structure, in terms of semantic roles [15, 20, 26], temporal relation recognition [39], and anaphora and coreference resolution [25, 35, 40]. Using these resources and processing tools, we aim at computing structured and normalized semantic representations of event sequences from textual descriptions of rituals, and on their basis identify recurrent patterns and variations across rituals by quantitative analysis. Frame semantics [15], with its concept of scenario frames connected by frame relations and role inheritance, offers a particularly powerful framework for the modeling of complex event sequences. It can be used to structure event sequences into more abstract concepts that may subsume different kinds of initial, transitional or closing events of rituals. Through the annotation of word senses, using lexical ontologies such as WordNet [14], we can observe and analyze variations in the selectional characteristics of specific events and their roles across rituals. The integration of semantically annotated corpora and explicit linking of annotated concepts to reference ontologies in a multi-layered annotation scheme offers possibilities to reason over corpora and external knowledge resources [6].

**Processing Ritual Descriptions with Standard NLP Tools.** The semantic annotations, however, need to be built upon linguistically preprocessed data. This preprocessing consists of several layers, starting with tokenization, part of speech tagging, and shallow or full syntactic analysis. Semantic analysis tasks, such as semantic role labeling or coreference resolution typically build on these pre-processing levels. As a basis for semantic annotation we use existing open-source systems for tokenizing, part of speech tagging, chunking or parsing. Automatic anaphora and coreference resolution provide important information for a coherent textual representation based on semantic role analysis. The systems we use for this preprocessing are data-driven, and have proven to obtain high performance scores, as they are typically trained on large corpora. In fact, such statistical systems often outperform rule-based systems.

However, there is one caveat: Most currently available training (and testing) data is taken from the news domain or encyclopedias like Wikipedia, which represent one or more particular domain(s). The assumption that data-driven approaches can be applied to an arbitrary new domain relies on the availability of training data for this domain. This, however, is rarely the case, especially if we move to “small” domains featuring special linguistic phenomena combined with restricted textual sources and a complete lack of annotated textual material. In this paper, we will investigate the effects of domain dependence of standard computational linguistic analysis tools, and explore methods to adapt them to the special type of language encountered in ritual texts.

**Detecting Similarities and Variance of Ritual Event Sequences.** The semantic annotation of rituals produces an abstract representation of sequences of events and their involved participants as they are found in ritual descriptions. The annotations are designed in such a way that the annotated texts can be compared at varying degrees of abstractness. Frames are embedded in a hierarchical network of inheritance (and other) relations (e.g. OFFERING – GIVING – LOSE-POSSESSION), while specific OFFERING events are still lexically distinguished (e.g. *sacrifice* – *offer*). Similarly, the semantic classes of participants that realise a frame's semantic role (e.g. the THEME role of a GIVING frame) can be represented at varying degrees of abstractness, by labeling them with senses defined in a hierarchical lexical network such as WordNet: from the annotation of the most specific sense, such as BOWL, we can generalize to its hypernyms such as VESSEL, CONTAINER. This allows us to detect events that are similar with respect to the action performed (i.e. regarding their frame), but may exhibit variations in the inventory of ritual objects used. On the other hand, we might observe event sequences that involve the same type of object as participant, yet in slightly different events. Comparing full sequences of events in this way is a non-trivial task. It requires efficient algorithms and flexibly configurable comparison functions. Here, we plan to build on and adapt algorithms for sequence alignment, as established in Bioinformatics and recently employed in NLP [2, 30].

**Outline.** In this paper, we report on first steps to provide a proof of concept for using computational linguistic resources and analysis methods for the study of ritual structures, based on small collections of data, analyzed at the intended levels of representation. We present initial studies that assess (i) the performance of standard NLP tools and resources for processing linguistic data from the ritual domain and (ii) the need and basic methods for domain adaptation for preprocessing and semantic annotation. We further (iii) present an outlook on the type of analyses we expect to conduct to enable empirical studies on the structure and variance of rituals.

Section 2 presents the project research plan and related work. In Sect. 3, we discuss special linguistic characteristics of the textual data that have an impact for automatic processing. Section 4 presents experiments that measure the performance of standard NLP processing tools on various linguistic depths and assess basic domain adaptation techniques. Section 5 presents our methodology for performing the semantic annotation of ritual descriptions. We describe our annotation framework, domain issues, as well as a principled work flow to enable increasing levels of automation. Finally, in Sect. 6, we present initial investigations into how to exploit semantic annotations of rituals to detect similarities and shared structures across ritual descriptions using sequence alignment techniques. Section 7 describes plans for future work and concludes.

## 2 Computational Linguistics for Ritual Structure Research

### *2.1 Project Research Plan*

The project is divided into two consecutive stages of research, which concentrate on corpus creation and annotation and on the analysis and exploitation of the data, respectively.

**Corpus Creation and Annotation.** In the first stage, a comprehensive corpus of linguistically and semantically annotated rituals from different cultures will be created from natural language descriptions of rituals that are procured by experts. The semantic annotation follows the frame semantics paradigm [15] and comprises both general linguistic and ritual-specific annotations.

To provide an empirical basis for the conceptualization of the domain, we automatically identify relevant domain terms on the basis of scientific publications on ritual research which in turn can serve to establish a base vocabulary for the annotation with ritual-specific concepts [31].

**Analyzing the Structure of Rituals.** Based on the semantic annotation of ritual descriptions, logical and statistical methods will be deployed to detect recurring structures in ritual descriptions, as well as systematic variances. In close cooperation with the ritual researchers, we will provide tools for the exploration of our data-driven, quantitative analyses of rituals.

## *2.2 Related Work*

Central to the structure of rituals are sequences of events and participants involved in these events. Hence, an important research topic is the **detection and analysis of event chains** in texts. The use of frame semantics as a useful abstraction layer for analyzing event chains has been investigated in [5]. A case study demonstrated how relations between instances of frames and roles can be inferred in context, using frame relations as well as contextual information, such as coreference or syntactic association. A related shared task on "linking roles in discourse" [33] has been organized as part of SemEval 2010. Recently, a statistical approach has been proposed for unsupervised detection of event chains, using co-occurrence of a single discourse entity as argument of different verbs as well as coreference information as criteria for extracting event chains [7, 8]. [17] applies similar linguistic and computational techniques to a collection of Sanskrit texts. Here, the computed event chains are used to establish the chronology of a group of alchemical texts, which cannot be dated using traditional historical approaches.

Another central issue related to our work is **domain adaptation**. As most state-of-the-art NLP tools are trained on news corpora, they do in general not generalize well to novel and special domains. A principled approach for addressing the domain adaptation problem is augmenting the feature space to model both domain and general, domain-independent characteristics [10]. A very similar approach employs a hierarchical bayesian prior to encourage features to take similar weights across domains, unless the differences of the data demand otherwise [16]. Both methods make use of labelled data. In a contrastive approach [19] make use of an instance weighting framework, where unlabeled instances of the target domain contribute to the model estimations.

# 3 Ritual Descriptions

We collect ritual descriptions from different sources. The collection process has been started with Hindu rituals from Nepal and rituals from the Middle East, but we plan to extend it to rituals from Ancient Egypt and the Middle Ages in central Europe. All our methods and techniques are culture-independent and can be adapted to other, non-English, languages.

We decided to concentrate on translated ritual descriptions that have already been published in scientific literature in order to quickly collect larger amounts of data that is relevant and trustworthy. All ritual descriptions are entered by a ritual researcher. We use a trac Wiki[2] as an interface, because it (i) allows easy structuring rules, (ii) is readable by every project member on the Web without knowledge of XML or other markup languages and (iii) is designed for automatic processing.

In the following, we discuss specific properties of the ritual descriptions in our corpus that are relevant from a computational linguistics point of view.

[2] http://trac.edgewall.org

## 3.1 Textual Sources

We use two types of textual sources. The first comprises studies by ritual researchers that deal with the religious, ethnologic and social background of rituals and are strongly theory-oriented. These texts will serve as a basis for building a ritual specific ontology, starting from a common terminology [31]. The second type of texts are descriptions of rituals. These sources form the basis of the ritual corpus and are, therefore, of special importance for the project. Two subtypes of ritual descriptions can be distinguished.

**Ethnographic observations** are an important source for our knowledge of how rituals are performed in modern times. These texts are written in English, though not always by native speakers. Some scholars tend to intersperse descriptive passages with theoretical interpretations of what was observed, making it hard to clearly separate the actual course of the ritual from interpretations (see Sect. 3.2).

Translations of indigenous **ritual manuals** that may date back several centuries are the second subtype of the ritual descriptions. Originally, the manuals are written in non-English languages (e.g., Sanskrit), but English translations of them have been published in ethnographic literature. Contrary to the ethnographic observations, these sources are mainly prescriptive in character. Since many of these manuals are intended as a kind of memory aid for ritual practitioners, they often record only the most important or extraordinary steps of a ritual, while typical, recurrent elements are omitted. This selective choice of content complicates the alignment of such manuals with the exhaustive descriptions of modern observers.

The subtype of ritual descriptions is stored as meta data attached to the source text, along with the bibliographic source of the descriptions, original language and related types of information.

## 3.2 Text Characteristics

Dealing with ritual descriptions requires handling of special phenomena on the lexical, syntactical and discourse-level. We describe these challenges in the following.

**Foreign Terms.** A ritual description produced by a ritual expert (be it a researcher or a practitioner) often contains terminology specific to the cultural context of the ritual. English counterparts for these terms often do not exist. Therefore, they often remain untranslated in the texts (although transliterated into Latin characters).

*Example 1.* He sweeps the place for the sacrificial fire with *kuśa*.

*Kuśa* is a Sanskrit term for a kind of grass that is very important in Vedic rituals. For this ritual, it is necessary to sweep with *kuśa* and not any other grass.

The term *kuśa* has never been seen by a common, newspaper trained part of speech tagger nor is it contained in a lexicon of a rule-based grammar. We therefore decided to annotate such terms with English paraphrases as in Example 2. For automatic processing, the original terms are replaced by the paraphrases and are later re-inserted.

*Example 2.* He sweeps the place for the sacrificial fire with <grass * kuśa>.

**Fixed Expressions.** Most rituals contain fixed expressions consisting of multiple words or sentences. These expressions are often prescribed pieces of text which have to be spoken or chanted while a ritual is performed (e.g., *Our father* in Christian church service).

*Example 3.* Salutation to Kubera reciting the mantra *arddha-māsāḥ [. . . ]*;

There is no common way in handbooks or scientific literature to refer to such fixed expressions. Sometimes, prayers or chants have a title or name; sometimes, first words or the refrain are given and the expert knows the exact expression.

As most fixed expressions cannot be translated literally, we adopt them as unanalyzed expressions in a foreign language. We ask the ritual experts to mark them as such, so that we can replace them with indexed placeholders during processing and re-insert them later.

**Imperatives.** As ritual manuals are often written by and for ritual practicioners, they contain a high amount of imperative sentences. In a randomly selected sample of ritual descriptions, we found 20% of the sentences realized in an imperative construction. The ritual description with the highest amount of imperatives contains over 70% of sentences with imperative constructions. By contrast, in the British National Corpus, only about 2% of the sentences contain imperatives.

**PP-attachments and Nested Sentences.** Prepositional phrases (PPs) are quite common in the data, as becomes apparent in Example 1. This introduces ambiguities that are hard to resolve. Deeply embedded PPs (as in Example 4) are difficult to attach correctly, but appear in the texts regularly.

*Example 4.* . . . worship of the doors of the house of the worshipper.

The frequency of syntactic coordination and nested sentence structures is varying between languages and text types. In Sanskrit, which is the source language of most of our data, long and nested sentences are very common. This characteristic is also reflected in the texts' translations into English, as the translators (i) try to preserve the original text character as much as possible and (ii) do not aim at producing well-to-read English sentences.

The joint occurrence of PP attachment, coordinations and sentence embedding are a challenge for syntactic processing. Example 5 illustrates the interaction of coordination (*italic*) and PP attachments (underlined) in a long sentence.

*Example 5.* Beyond the members of the lineage, these visits lead to the paternal aunts of three generations which includes father's *and* grandfather's paternal aunts *and* their daughters *and* granddaughters, the maternal uncles *and* maternal aunts of their grandmother *as well as* their maternal uncles of three generations.

This leads to a combinatorial explosion of possible analyses and – in case of statistical disambiguation – a parser is deemed to make wrong guesses. Therefore, since full-fledged syntactic analyses are not necessarily needed for role semantic labeling (see e.g. [13]), we opted for a flat syntactic analysis based on chunks.

**Interpretations.** Ritual descriptions that have been published in scientific literature often do not contain "clean" descriptions restricted to the ritual performance only. Instead, the description of a ritual performance is interwoven with comments or interpretations that help the reader understand the ritual.

*Example 6.* The involvement of the nephews can be understood as a symbolic action to address those of the following generation who do not belong to the lineage of the deceased.

Example 6 is clearly not an event that occurs during the ritual, but a scientific interpretation. Although it is in principle possible to annotate such sentences with frames and frame elements, they represent a different level of information that does not belong to the ritual itself. As we want to automatically extract common event sequences from the ritual descriptions, such interpretations need to be clearly separated from descriptions of factual events. In order to systematically address this issue, we divided the sentences into three classes:

1. Sentences clearly indicating events during the ritual performance (Example 3)
2. Clear interpretations, citations or comments (Example 6)

3. Sentences that are ambiguous with respect to these classes, or sentences that contain elements of both classes (Example 7)

*Example 7.* The wife of the chief mourner [. . . ] will carry a symbolic mat that represents the bed of the deceased [. . . ].

We performed an annotation study on a randomly selected ritual description (40 sentences) and found that 15% of the sentences contain both interpretative and factual elements or are ambiguous (clear interpretations: 17.5%, clear factual statements: 67.5%). We did not yet experiment with automatic tagging of sentences according to their class. One possibility, however, could be the application of methods used for the automatic detection of hedges. Academic writers tend to use a high amount of hedges [18]. From the examples in our ritual descriptions, hedges indeed appear quite often. Following the definitions given in [23] and [22], 42.9% of our sentences with a clear interpretative character contain linguistic hedges. There is existing work on the automatic detection of hedges [23, 36] which may be adapted to our specific concerns.

As a first partial solution to the problem, we decided to annotate the clear interpretative sentences as such. They will be ignored for the frame annotation, but remain in the texts.

## 4 Automatic Linguistic Processing

As a basis for semantic annotation and processing, the ritual descriptions are preprocessed with standard NLP tools. We use UIMA (Unstructured Information Management Architecture)[3] as a pipeline framework, in which we have integrated a rule-based tokenizer, the OpenNLP[4] part of speech tagger, the Stanford Lemmatizer [38] and the OpenNLP chunker.

### *4.1 Tokenizing*

Many of our texts contain special, non-English characters (*ś*) or complete tokens (*Gaṇeśa*). Therefore, we employ a rule-based tokenizer that uses Unicode character ranges in conjunction with an abbreviation lexicon to detect common abbreviations such as *etc.* or *i.e.*

### *4.2 Part of Speech Tagging and Chunking*

Using standard models for part of speech tagging and chunking produces rather poor results. This is due to the fact that our data contains a lot of unseen tokens and a high amount of rare and uncommon constructions. We experimented with different scenarios for the domain adaptation of an existing part of speech tagger and chunker.

As we aim at a culture- and source-language independent framework, we decided to use a statistical part of speech tagger and chunker, that can be trained on specific corpora. Large amounts of training material for both labeling tasks are available from other domains, and the annotation of small amounts of data from the ritual domain is feasible. This corresponds to the scenario of fully supervised techniques for domain adaptation discussed in the literature [10]. We experimented with different combination techniques, which are outlined in the following section.

[3] http://incubator.apache.org/uima/

[4] http://opennlp.sf.net

### 4.2.1 Data Sets

Our training data comes from two different sources. We manually annotated 532 sentences of our ritual descriptions with part of speech tags and chunks, using the Penn Treebank tagset.

As a second domain corpus we chose the Wall Street Journal, which features compatible part of speech and chunk annotations. For the extraction of chunks from the Penn Treeebank we made use of the CoNLL 2000 scripts. They were also used for the evaluation of the chunker.

For the marking of chunks, we used a modified version of the CoNLL 2000 style of marking chunks [34]: The beginning of PP-chunks is marked with `B-PP` as usual. All tokens covered by the PP that are contained in a further embedded NP are marked with complex chunk tag, for example: `B-NP/I-PP`. This way, we can encode embedded structures in chunks to a certain extent.

We used 10-fold cross-validation to evaluate the data. In cases of training on mixed corpus types (see below), we "folded" the ritual corpus before mixing it with the Wall Street Journal data. This way, we make sure that our test data did not include any non-ritual data.

**Table 1** Training sets for part of speech tagger and chunker

| Name | Description | Sentences (one fold) | Tokens/Sentence |
|---|---|---|---|
| WSJ | The Wall Street Journal | 47,861 | 24 |
| RIT | Ritual Descriptions | 470 | 19 |
| WSJ + RIT | Union | 48,331 | |
| WSJ + RIT ↑ | oversampling RIT | 106,955 | |
| WSJ ↓ + RIT | undersampling WSJ | 939 | |
| WSJ × RIT | Combined feature space [10] | 48,331 | |
| WSJ × RIT ↑ | oversampling RIT | 106,955 | |
| WSJ ↓ × RIT | undersampling WSJ | 939 | |

Table 1 shows the different data sets and the sizes of one (average) training fold. WSJ + RIT is a simple union of the two sets. As the sizes of the two data sets differ vastly, we also experimented with equally sized corpora, by use of over- and undersampling. WSJ + RIT ↑ represents the union of the WSJ with the oversampled RIT corpus, WSJ ↓ + RIT stands for the union of the undersampled WSJ corpus with the RIT corpus.

The data set WSJ × RIT was produced by augmenting the feature space along the lines of the work in [10]. Let $\mathbf{v}_i =< f_1, f_2, ..., f_n >$ be the original feature vector for item $i$ and $d$ be a function returning an identifier for a domain. $d(0)$ is then a string representing the general domain, $d(1)$ the domain of rituals and $d(2)$ the domain of news articles. $f_k^{d(x)}$ is the same feature value as $f_k$, but prefixed with $d(x)$, a domain identifier. The augmented feature vector is then $\mathbf{v}'_i =< f_1^{d(0)}, f_2^{d(0)}, \ldots, f_n^{d(0)}, f_1^{d(i)}, f_2^{d(i)}, \ldots, f_n^{d(i)} >$, with $i = 1$ or 2. This way, each training example is annotated with a general domain feature vector and a domain-specific feature vector. The learner then can learn whether to use the general domain feature set (for which it has massive training data) or the domain-specific feature set (with small training data). Again, we used the same over- and undersampling techniques as before.

### 4.2.2 Evaluation

**Part of speech tagging.** Table 2 lists the results obtained by training the POS-tagger on different data sets. We use the model trained on the WSJ data set only, i.e., without any domain adaptation, as a baseline. Its performance is 90.9% accuracy.

**Table 2** Part of speech tagging results with different models

| Training data | Accuracy (%) |
|---|---|
| WSJ | 90.90 |
| RIT | 94.82 |
| WSJ + RIT | 95.72 |
| WSJ + RIT ↑ | **96.23** |
| WSJ ↓ + RIT | 95.25 |
| WSJ × RIT | **96.86** |
| WSJ × RIT ↑ | **96.85** |
| WSJ ↓ × RIT | 95.92 |

If RIT is used as (small) training set, the POS tagger achieves a performance of 94.82%. Training on the union of RIT and WSJ yields an increase in performance (95.72%) compared to RIT. Balancing the training sets again increases the performance if the ritual data is oversampled (resulting in a very large training set). If the WSJ data is undersampled, performance decreases compared to the unbalanced union. Augmenting the feature space yields minor improvements, even if the training data is unbalanced. The best performing model is trained on WSJ × RIT, while WSJ × RIT ↑ performs similarly. The small data set, WSJ ↓ × RIT, achieves less performance than a large and balanced, but un-augmented data set (WSJ + RIT ↑).

**Table 3** Chunking results with different models

| Training data | Precision | Recall | $F_{\beta=1}$ |
|---|---|---|---|
| WSJ | 86.3 | 87.0 | 86.6 |
| RIT | 85.5 | 86.0 | 85.7 |
| WSJ + RIT | 86.3 | 87.0 | 86.6 |
| WSJ + RIT ↑ | 87.7 | 88.5 | 88.1 |
| WSJ ↓ + RIT | 86.9 | 79.7 | 83.1 |
| WSJ × RIT | 74.0 | 74.9 | 74.4 |
| WSJ × RIT ↑ | 81.0 | 81.5 | 81.3 |
| WSJ ↓ × RIT | 74.8 | 71.8 | 73.3 |

**Chunking.** Table 3 shows the results of the chunking models trained on the different data sets. Again, we use a model trained on the Wall Street Journal as baseline (WSJ). This model achieves an f-score of 86.6. The model trained on the ritual data (RIT) performs slightly lower, achieving an f-score of 85.7. Training the model on the simple union (WSJ + RIT), does not increase the performance compared to the baseline. However, if we oversample the ritual data and thus balance the training data (WSJ + RIT ↑), we achieve a minor improvement in f-score. Undersampling the WSJ data decreases the performance. The augmentation of the feature space decreases the

performance on all data sets. This is in contrast with the results for part of speech-tagging (above) and the results presented in [32]. Within the augmented feature space models, we can observe similar tendencies as in the other models: Oversampling improves the performance compared to unbalanced data, while undersampling decreases it. Currently, we can only speculate on possible reasons why feature augmentation does not work well with embedded PP chunks.

## *4.3 Anaphora and Coreference Resolution*

In order to be able to extract continuous and consistent event chains (cf. Sect. 6), it is necessary to compute entity chains from anaphoric and coreferent expressions involved in different events. Anaphoric expressions are (typically) pronouns that receive their interpretation by linking to an antecedent (see Example 8). Also full noun phrases may be co-referent with antecedent noun phrases (see Example 9).

*Example 8.* The father should touch the girl [...]. Let him give a golden coin as ritual fee [...].

*Example 9.* Let the girl sit on the seat [...]. Let the girl wash her face [...].

In order to study overall performance and potential out-of-domain effects, we tested several anaphora and coreference resolution (ACR) systems on a single ritual description and evaluated their performance. In [32] we performed a first, broad comparison of several ACR systems. In this chapter, we concentrate on BART [40] and JavaRAP, two complementary systems that yielded the best results in [32].

**Candidate Systems.** BART [40] is a modular machine learning toolkit that computes coreference chains using features inspired by [35] and a maximum entropy learner. In order to extract these features, the data need to be parsed or at least chunked. In our experiment, we configured BART to work on the basis of chunks instead of parses. We did not exploit any of BART's tuning possibilities but used the standard classifier.

In contrast to BART, JavaRAP is a rule-based anaphora resolution system that implements the Lappin & Leass algorithm for pronominal anaphora resolution [21]. It exclusively treats third person pronouns and lexical anaphors like reflexives and reciprocals and recognizes pleonastic pronouns. While BART computes full coreference chains, JavaRAP only generates pairs of anaphors and antecedents. JavaRAP uses the Charniak parser [9] for preprocessing. Sentence splitting for parsing was done manually, apart from that we used JavaRAP with the default configuration.

The following subsections present a detailed assessment of the systems' ACR performance, as well as an in-depth manual error analysis, carried out on a sample ritual description of 37 sentences.

### 4.3.1 Anaphora Resolution Task

For examining BART on the anaphora resolution task, we only evaluated coreferent mention pairs (as predicted by BART) that include a pronoun. We compared the results of automatic annotation against a manually constructed gold standard consisting of 26 anaphor-antecedent pairs. The 26 pairs correspond to 10 entity coreference chains. If both elements of a predicted pair refer to the same entity in the text, i.e., they are part of the correct entity chain in the gold standard annotation, the pair is considered correct. We evaluated JavaRAP in the same way and on the same data. The results are given in Table 4.[5]

[5] In Table 4 'wrong pairs' refers to cases of anaphoric pronouns resolved incorrectly, i.e. to an entity belonging to another coreference chain; 'overgeneration' refers to cases of expletive

**Table 4** Anaphora resolution results

| | | BART | | JavaRAP | |
|---|---|---|---|---|---|
| | | abs | percent | abs | percent |
| **Coverage** | Gold standard pairs | 26 | 100% | 26 | 100% |
| | System pairs | 23 | 88.5% | 31 | 119.2% |
| **Errors** | Wrong pairs | 10 | 38.5% | 11 | 42.3% |
| | Overgeneration | 5 | 19.2% | 4 | 15.9% |
| | Missed anaphors | 8 | 30.7% | 0 | 0% |
| **Performance** | Correct pairs | 8 | 30.7% | 16 | 61.5% |
| | Precision | | 34.8% | | 45.2% |
| | Recall | | 30.8% | | 53.8% |
| | $F_1$-measure | | 32.7% | | 49.1% |

JavaRAP generates more anaphor-antecedent pairs than what is contained in the gold standard, while BART misses some pairs. Typically, systems do not overgenerate in the anaphora resolution task. However, overgeneration may occur by considering expletive pronouns, such as *it*, or demonstrative pronouns as anaphoric to entities. Both BART and JavaRAP handle expletive pronouns, with JavaRAP resolving 33% of the expletives correctly to NULL, while BART does not recognize them as expletives, and incorrectly links them to pronominal antecedents. We further note that eight pronouns are not recognized by BART. This means that BART's pronoun detection is both incomplete and erroneous, while JavaRAP's analysis is complete, with remaining errors concerning, i.a., the treatment of expletive pronouns.

A closer analysis shows that the distance between anaphor and antecedent is restricted (maximally one sentence) and does not influence system performance. Good performance is obtained on the basis of simple pattern match methods, as in chains containing several pronouns (e.g. *the girl – she – she*). Similarly, different pronouns referring to an entity of the same number and gender (e.g. *his – he*) are frequently detected correctly, except for one gender-related mismatch of BART, shown in Example 10. All but one of BART's correct anaphor-antecedent pairs are pronouns referring to pronouns. This leads to the assumption that the pattern match and gender comparison strategy works well; yet, this method fails for both systems in sentences that contain references to several entities of the same gender, as in Example 11.

*Example 10.* He$_1$ received a ritual mark [...] from his$_1$ wife [...]. She$_1$ took [...].

*Example 11.* She$_1$ is known for her$_1$ healing powers, so Kalpana$_2$ visited her$_1$ to gain protection for her$_2$ husband.

JavaRAP occasionally resolves anaphors (incorrectly) to inanimate common nouns like "occasion" or "wellbeing" but also detects proper name antecedents, while BART never resolves pronouns to inanimate nouns, and sometimes does not link to the correct antecedent proper name.

Overall, JavaRAP performs better than BART in the anaphora resolution task. Although JavaRAP's performance is a reasonably good outcome, the system does not build coreference chains, hence only delivers partial information.

---

or demonstrative pronouns incorrectly linked to an antecedent; 'missed anaphors' lists cases of anaphoric pronouns not resolved to any antecedent.

### 4.3.2 Coreference Resolution Task

For evaluting BART on the coreference resolution task, we considered exclusively the coreference chains which contain at least one personal or possessive pronoun in third person such as *it* or *him*.[6] We used the scorer implemented for the SemEval-2010 coreference resolution task [29] and measured the standard precision and recall for mention identification, and the MUC [41] and B-CUBED [1] precision and recall metric for coreference resolution. The MUC metric emphasizes the correctness of links between mentions within coreference chains, and only weakly penalizes incorrect links between different entity chains. It therefore shows a strong bias to prefer longer chains. In contrast, the B-CUBED metric emphasizes the presence or absence of mentions in entity chains [11].[7]

Table 5 first analyzes BART's performance in the mention detection task. A mention is identified as strictly correct if the system returns exactly the token of the gold standard. If the system returns a substring of a gold mention, it is counted as partially correct. The sum of strictly and 0.5 times partially correct identified mentions is used as number of true positives. As we can see, BART correctly identifies more than 80% of the mentions (see low number of false negatives), however, it tends to overgenerate, with a high number of 'invented' mentions (false positives). For complex (embedded) noun phrases, BART repeatedly fails to recognize the encompassing noun phrase mention, and only delivers the embedded noun phrase. These cases are recorded as correct (embedded) and missed mentions (for the higher level NP). But also very simple mentions with standard NP structure are not detected completely (e.g. *grandmother* instead of *the grandmother*). A surprising discovery (see Sect. 4.3.1) is that BART misses to detect simple pronouns like *she*.

**Table 5** Evaluation results for mention identification

| Measure | BART | |
|---|---|---|
| | abs | percent |
| Gold standard mentions | 53 | 100% |
| Total found | 81 | 152.8% |
| Strictly correct | 44 | 83% |
| Partially correct | 1 | 1.9% |
| False positives | 36 | 67.9% |
| False negatives | 8 | 15.1% |
| Precision | | 54.93% |
| Recall | | 83.96% |
| $F_1$-measure | | 66.41% |

Table 6 shows precision, recall and f-measure using the MUC and B-CUBED metrics for joint anaphora and coreference resolution, as standardly performed by BART, and a separate evaluation that only considers coreference resolution between nominal NPs. As a baseline, we applied the simple heuristic of resolving a pronominal/nominal NP to the nearest preceding NP. In the joint

[6] In this way we exclude chains such as *vessel – vessel – vessel*, which are often not coreferential, and not of primary interest in our scenario. Hence, we focus on entity chains for protagonists that are taken up using pronominal references.

[7] More specifically, MUC precision is calculated by dividing the number of links in the system output that match the manual annotations by the total number of links in the system output. MUC recall is the ratio between the number of links common to the manual annotation and the system output and the total number of manually annotated links. B-CUBED precision and recall are basically calculated in the same way, but computed over the number of mentions.

anaphora and coreference setup, the baseline considers the closest preceding NP (pronominal or nominal), while in the coreference only setup, it chooses the closest nominal NP.

**Table 6** Evaluation results for joint anaphora and corefernce and isolated coreference resolution

| Measure | | Anaphora & Coreference | | Coreference only | |
|---|---|---|---|---|---|
| | | Baseline | BART | Baseline | BART |
| MUC | Precision | 12.1 | 39.72 | 6.2 | 23.5 |
| | Recall | 12.1 | 70.73 | 6.6 | 66.6 |
| | $F_1$-measure | 12.1 | 50.87 | 6.4 | 34.4 |
| B-CUBED | Precision | 30.4 | 11.38 | 29.6 | 7.2 |
| | Recall | 19.3 | 65.95 | 25.9 | 64.9 |
| | $F_1$-measure | 23.6 | 19.41 | 27.7 | 13.0 |

For the standard joint task, BART outperforms the baseline with high recall and modest precision according to MUC. In contrast, according to B-CUBED measures, the baseline performs considerably better, while BART's precision suffers severely. Its performance drops below the simple baseline heuristics for precision and f-measure. This suggests that BART tends to build long chains and inaccurately assigned single mentions. We observe a parallel, but more pronounced effect when evaluating coreference resolution in isolation. This suggests that BART's performance is severely impacted by coreference resolution. Note in particular that in contrast to BART, the baseline tends to perform better in the isolated coreference resolution evaluation, as opposed to the joint task.[8]

Thus, on our sample data from ritual descriptions, BART does not achieve state-of-the-art performance. In order to detect the main sources of the problems, we performed a detailed analysis of BART's results, including the inspection of intermediate results: the pairs of mentions judged as coreferring. In this way we hope to develop strategies for addressing the coreference resolution problem for ritual texts.

### 4.3.3 Detailed Evaluation

Our investigation showed that the proposed coreference chains often contain correct links. However, these chains are extended by incorrect links, due to BART's general tendency to overgeneration.

As to the identification of coreferent NPs, we observed two major issues: strongly divergent surface realizations of co-referring entities, e.g. *the boy or girl - the child*, and nouns with lexically marked gender like *the grandmother* or *her husband* which are linked to antecedents with conflicting gender. Even pronouns with explicit gender are found to be incorrectly linked, as in Example 10 above. This points to a general weakness of BART in exploiting gender information and a strong reliance on surface form. Also, some domain-specific text characteristics seem to complicate the coreference resolution task. One of these is the alternating usage of generic and specific denominations of participiants (Example 12). This leads to grammatical changes which might affect BART's syntactic features and thus leads to inappropriate linkages. The aggregation

[8] This is despite the fact that in our evaluation the baseline proposes a single antecedent per pair, while for BART, we build candidate pairs from the full coreference chains it proposes. While BART generates a higher number of correct candidate pairs, in absolute numbers, as compared to the baseline, its performance suffers from overgeneration.

of participants into groups and their subsequent separation into sub-groups as in Example 13 has similar effects.

*Example 12.* Now at the auspicious time bring the girls$_1$ holding their$_1$ hands reciting the mantra. [...] Let the girl$_1$ sit on the seat [...].

*Example 13.* The grandmother paints a red mark on the forehead of the boy or girl$_1$. In addition, boys$_2$ receive a stroke of yoghurt on their right temple, girls$_3$ on their left temple.

A domain-specific reason for BART's overgeneration is the frequent mention of named entities – mostly names of gods such as *Śiva* or *Viṣṇu*.

Examining the overall discourse characteristics of ritual descriptions, we observe that the number of individuals who participate in a ritual as "protagonists", and thus being potentially part of a coreference chain, is rather limited, with a tendency for (important) protagonists to be contained in longer entity chains.[9]

Our observations suggest a number of strategies that could import **domain knowledge** into the ACR task. Given that rituals feature a small number of participants, the list of mention candidates could be restricted by establishing an inventory of typical protagonist entity types and roles (such as *child*, *father*, *priest*, *mourner*, etc.). This could contribute to reduce the massive overgeneration we oberve for BART. Given the frequent use of relational nouns, they could be explicitly modeled in terms of disjointness constraints, to be used as highly ranked features in judging individual candidate pairs and in the computation of entity chains. Occasionally, non-animate ritual objects are contained in entity chains. These are typically referred to by identical or synonymous noun phrases, hypernyms, or else neuter pronouns. This suggests more explicit modeling of gender information.

Overall, our study suggests several strategies for improving the currently poor performance of standard ACR systems for the processing of ritual descriptions. (i) Given that the two systems under consideration, JavaRAP and BART, as well as the baseline heuristics, display complementary strengths and weaknesses on the tasks of anaphora and coreference resolution,[10] an ensemble learning architecture should be considered [3]. (ii) The special discourse characteristics of ritual texts – with a restricted number of "protagonists" involved in (a small number of) entity coreference chains per text – could be exploited by modeling the entity domain of rituals regarding coreference vs. disjointness conditions, lexical gender marking, and by restricting the set of potentially coreferent entities, i.e., the set of mentions to be considered by the system, to a constrained set of semantic entity types.

These strategies are in line with the project's aims to develop an ontology of relevant concepts of the ritual domain, and will be supported by the lexical semantic annotation of ritual texts, which is discussed in the following sections.

## 5 Semantic Annotation of Ritual Descriptions

We use frame semantics [15] as a theoretical framework to encode the ritual sequences such that each separable action mentioned in the ritual corpus is represented by its own frame. The actors that perform the ritual actions as well as objects, times and places are annotated as frame roles (cf. Fig. 1, page 186).

[9] In our sample text we observe an average chain length of 5.3 (median: 4, maximal length: 13).

[10] The sets of correct pairs proposed by BART and JavaRAP, and BART and the baseline in the two base tasks are only partially overlapping.

## *5.1 Adaptation of Existing Resources*

To guarantee a consistent encoding of ritual frames, the FrameNet lexical database is used to deliver a base inventory of frames. We try to map the ritual actions to frames that are already defined in FrameNet. For this sake, verbs found in the ritual descriptions are extracted automatically from the chunked ritual descriptions. They are ordered in semantic groups and subsequently searched for in the FrameNet database. This way we can make use of a well structured inventory of frames.

**Coverage.** According to a first estimation reported in [31], over 80% of the verbs mentioned in the ritual corpus are contained as lexical units in FrameNet. However, a closer inspection of the ritual data reveals that numerous terms are only identical at the lexical level, but occur in different senses. Moreover, a large number of concepts that are important in ritual descriptions are not dealt with in FrameNet. At the current state of annotation, it is difficult to give comprehensive figures about the coverage of FrameNet on ritual corpora. However, areas not (or only scarcely) covered by FrameNet include, for example, the fields of preparing and serving food.

**Granularity.** Frequently, the frames contained in FrameNet represent concepts that are too abstract for the annotation of rituals. In these cases, FrameNet groups several lexical units into one frame that would not correspond to a single concept in a genuine ritual frame ontology. Examples are the verbs "to cover", "to anoint" and "to fill". These verbs are assigned to a single frame FILLING in FrameNet because they express the same idea of "filling containers and covering areas with some thing, things or substance". Although we use frames to generalize from the literal level of ritual descriptions, annotating "to fill" and "to anoint" by a single frame in a ritual context would certainly lead to over-generalization and, therefore, to a significant loss of information. In such cases, new frames have to be designed. For the given example, we decided to further specify the frame Filling with three new frames: FILLING_CONTAINER (filling a container), BESMEARING_SURFACE (covering a surface with a liquid) and WRAPPING_OBJECT (wrapping an object with another object).

On the other hand, the granularity of FrameNet frames can also be higher than what is needed for our purposes. This case occurs, for instance, in the area of legal concepts, which are covered in great detail by FrameNet. Such cases are easier to resolve than those resulting from coarse-grainedness of frames discussed above, due to the FrameNet inheritance hierarchy. That is, we can use predefined, more abstract frames from higher levels in the hierarchy.

The existence of semantic fields that are covered in FrameNet in great detail clearly demonstrates that it has been successful in modeling specific domains. Thus, for the present project, domain adaptation will mainly consist in modeling finer-grained frame structures for semantic fields that are relevant for the ritual domain.

**Annotation Process and Automation.** In a first phase, we started with an initial corpus of manual annotations, concentrating on developing a suitable frame inventory for the ritual domain. The manual annotations of two annotators are checked for inter-annotator agreement and differences are harmonized. Having established an initial frame inventory and corpus of annotations, we currently train a frame-semantic labeler to assign frames to target words automatically. In a second manual annotation phase, the automatically assigned frames are checked and further annotated with frame semantic roles. This results in an increasing database of text data labeled with frames and frame roles, which are used to improve the model for automatic frame labeling. In the next phase of annotation, we are planning to apply and possibly retrain frame-semantic labelers like [13, 27] to explore automatic or semi-automatic annotation of frames, frame roles and frame scenarios.

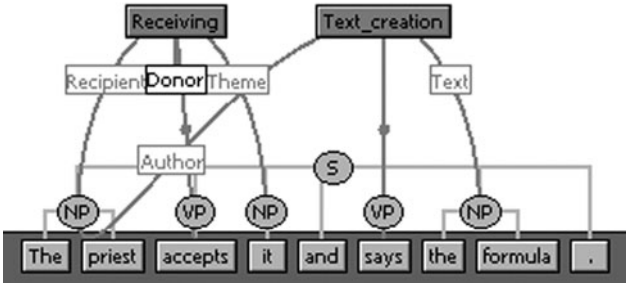


**Fig. 1** Annotated sentence "The priest accepts it and says the formula."

## Manual Annotation.

We are using the Salsa tool [12] (Figure 1) to support manual annotation, as well as manual correction of automatic annotations. The frame-semantic information is stored as a separate layer along with the source text and additional layers of linguistic annotation. When the corpus text or the linguistic preprocessing layers are updated, this layering mechanism makes it possible to reassign the frame-semantic annotation, thus avoiding manual re-annotation.

## Automatic Frame Assignment.

Five manually annotated ritual descriptions are used to train a classifier for automatic frame assignment. The classifier uses triples of the form $\{t, f, b\}$ as feature vectors. $t$ is a target word in an annotated sentence $s$, $f$ is the name of the frame assigned to $t$, and $b$ is a bag of words vector that encodes the vocabulary of the sentence $s$. To build the bag of words vectors, we compile a dictionary $D$ that contains all words occurring at least twice in the five ritual descriptions. Position $i$ of the bag of words vector is set to 1 if word $i$ from $D$ is contained in the sentence $s$; else to 0. The triples from the training and test set are combined into the sets $A_{Train}$ and $A_{Test}$, respectively.

The automatic annotation proceeds by single frames $f_s$. For each target word $t_x$ that has been annotated at least once with the frame $f_s$ in $A_{Train}$, the triples $T_i = \{t_i, f_i, b_i\}$ with $t_i = t_x$ are extracted from $A_{Train}$, which results in the training subset $A_{Train'} = \{T_i \in A_{Test} | t_i = t_x\}$. If all frame names in $A_{Train'}$ are identical, the frame names $f$ in $A_{Test'} = \{T_i \in A_{Test} | t_i = t_x\}$ are set to $f_s$. Else, the feature vectors in $A_{Train'}$ are used to train a classifier for assigning frame names to the triples in $A_{Test'}$. If $|A_{Test'}| > \Theta$, we apply support vector machines, which are supplied by the `R` library `kernlab` [28]. If $|A_{Test'}| \leq \Theta$, classification is performed with a simple majority vote that selects the most frequent frame from $A_{Train'}$ for a target word in $A_{Test'}$. To estimate $\Theta$, we repeat the classification process with different values of $\Theta$. The results recorded in Table 7 show that $\Theta = 10$ offers an acceptable tradeoff. This threshold divides the whole training set into two groups that contain the low and high frequency target words, respectively. More exact estimations of $\Theta$ are only possible with a larger training set.

**Table 7** Threshold $\Theta$ and overall accuracy

| $\Theta$ | 5 | 10 | 20 | 30 | 40 | 60 |
|---|---|---|---|---|---|---|
| Overall accuracy | 0.6606 | **0.6788** | 0.6485 | 0.6485 | 0.7212 | 0.6242 |

In general, the approach yields promising classification performance during training and testing. Table 8 gives the accuracy rates for target words that are associated with more than one frame. The classification accuracy is significantly correlated with the number of training samples (Kendall's $\tau = 0.5810$, $p = 0.012$*), while the number of frames associated with a target word does not influence the accuracy at an $\alpha$ error level of 10% (Kendall's $\tau = 0.1796$, $p = 0.4801$). Therefore, we expect increasing classification accuracy with a growing amount of training data.

**Table 8** Accuracy rates for targets with more than one frame

| Target | # Frames | # Training Samples | Classifier | Accuracy |
|---|---|---|---|---|
| blessing | 2 | 4 | Maj | 0.7500 |
| decorate | 2 | 4 | Maj | 0.0000 |
| feed | 2 | 3 | Maj | 0.0000 |
| give | 3 | 36 | SVM | 0.8611 |
| go | 2 | 2 | Maj | 0.0000 |
| make | 4 | 11 | SVM | 0.4545 |
| offer | 3 | 58 | SVM | 0.7414 |
| place | 2 | 18 | SVM | 0.9444 |
| prepare | 2 | 2 | Maj | 0.0000 |
| put | 2 | 4 | Maj | 0.7500 |
| seat | 2 | 2 | Maj | 0.0000 |
| take | 2 | 17 | SVM | 1.0000 |

The classification is less reliable when the annotation of the training data is not consistent or when targets are annotated consistently with more than one frame. Table 9 shows the classification results for the target "offer".

**Table 9** Classification results for the target word "offer"

| True class | Classified as | | |
|---|---|---|---|
| | CREATE_PH._ARTWORK | GIVING | RITE |
| CREATE_PH._ARTWORK | 0 | 0 | 1 |
| GIVING | 0 | 36 | 2 |
| RITE | 12 | 0 | 7 |

Although with $p = 0.00089$ the classification rate is clearly above the baseline for a random assignment of classes [4, 625], we can observe systematic problems with the assignments of the frames RITE and CREATE_PHYSICAL_ARTWORK. These problems are caused by double manual annotations of (rare) expressions such as "offer a mark on the forehead (of the participant of a ritual)." In its literal meaning, this expression describes how paint is applied to a surface (the forehead) to create a mark and is, therefore, manually annotated with the frame CREATE_PHYSICAL_ARTWORK. In the context of a rite, however, drawing such a mark has a strong ritualistic connotation. Thus, the target is also annotated with RITE in the training data. Such conflicts that arise from multiple annotations in the training data will be resolved by allowing multiple assignments of frame names in future versions of the classifier.

Automatic Assignment of Semantic Roles.

Currently, we are performing experiments with SVM-based role annotation and with the system ASSERT [27], which is trained on FrameNet data. The data collected from manual role annotation will be used to adapt the role assignment model to frames specific to the ritual domain. The role annotations will be complemented by sense annotations, using as a first step, the English WordNet [14], with adaptations as required for our special domain.

## 6 Detecting Ritual Structure

With increasing density of the semantic annotations of our ritual text corpus, we will be in a position to make use of them to detect recurrent structures in the ritual descriptions. By aligning the frame-semantic annotations, we aim at producing more compact descriptions of the ritual sequences that can be used to detect common subsequences and, finally, ritual structures. Manual alignment of ritual sequences becomes unfeasible with large numbers of rituals or even few, but complex rituals. Therefore, automating the alignment of sequences is one of the most important areas of future research in our project.

One possible approach is to apply algorithms for multiple sequence alignment (MSA) to the sequences of frames that constitute the ritual descriptions. In such an approach, we interpret each frame and its roles as an atomic symbol. The aim of the algorithm is to find a global alignment between $n$ sequences describing the same ritual that minimizes a cost-based criterion function. Many implementations of MSA use the Needleman-Wunsch algorithm [24] or another variant of the Levenshtein algorithm to perform exact pairwise comparison of sequences. They calculate a distance matrix from the pairwise distances and apply a greedy clustering algorithm to find groups in this distance matrix. The final order of sequence alignment is determined from the clustering tree [37]. Similar techniques of MSA have been applied successfully to the problem of parallelizing groups of prescriptions found in Indian alchemical literature [17]. One of the central challenges of such an approach is the design of an appropriate function for determining the dissimilarity of symbols. Implementations of MSA that are used in bioinformatics frequently work with very small alphabets and apply binary dissimilarity functions because two symbols are either identical or not. Contrary to this, the "alphabet" of frames in rituals is virtually unlimited, and distances between symbol pairs must be calculated dynamically during run-time. In the study of Indian alchemy, the binary cost function has been replaced by a fuzzy cost function whose values range from 0 (two statements are completely dissimilar) to 1 (fully identical). This value is calculated by inspecting the similarity of role fillers as deduced from their arrangement in a domain-specific ontology of alchemical concepts. Currently, we are studying the application of MSA to the problem of parallelizing sequences of frames in ritual descriptions.

As a proof of concept for the types of analyses we can offer to ritual scientists on the basis of aligned semantic annotations, we constructed representations for a number of close variations of rituals. Four of these ritual descriptions are shown below with assigned frames printed as subscripts.

1. Offer$_{\text{GIVING}}$ a lamp (with a burning wick and the mantra$_{\text{TEXT_CREATION}}$) *tejo 'si*.
2. Worship$_{\text{RITE}}$ of the lamp, the wooden measuring vessel and the key (reciting$_{\text{TEXT_CREATION}}$) *agnir mūrdhā divaḥ* (and$_{\text{TEXT_CREATION}}$) *trātāram indram*. One should wave$_{\text{CAUSE_TO_MOVE_IN_PLACE}}$ with lamp, wooden measuring vessel and key.
3. Shower$_{\text{CAUSE_FLUIDIC_MOTION}}$ pieces of fruits (etc.) from the measuring vessel (on the head of the boy with$_{\text{TEXT_CREATION}}$) *yāḥ phalinī*. (Make this) three (times).
   Show$_{\text{CAUSE_TO_PERCEIVE}}$ (and offer$_{\text{GIVING}}$) the lamp (to the boy with$_{\text{TEXT_CREATION}}$) *tejo 'si*.
4. (Wave$_{\text{CAUSE_TO_MOVE_IN_PLACE}}$) light (with a burning wick). Now fragrant materials etc. Worship$_{\text{RITE}}$ the lamp, the wooden measuring vessel and the key (reciting$_{\text{TEXT_CREATION}}$) *agnir mūrdha* (and$_{\text{TEXT_CREATION}}$) *trātāram indram*.

Wave$_{\text{CAUSE_TO_MOVE_IN_PLACE}}$ the lamp, the (wooden) measuring vessel and the (iron) key (over the head of the boy reciting$_{\text{TEXT_CREATION}}$ ) *asuraghnam*.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | GIVING lamp | TEXT _ CREATION „tejo'si" |
| | RITE lamp, vessel, key | | TEXT _ CREATION „agnir mūrdhā" | TEXT _ CREATION „trātāram indram" | CAUSE _ TO_MOVE _ IN_PLACE lamp, vessel, key | | | |
| | | CAUSE _ FLUIDIC _ MOTION fruit Vessel Boy | TEXT _ CREATION „yāḥ phalini" | | | CAUSE _TO_ PERCEIVE lamp boy | GIVING lamp boy | TEXT _ CREATION „tejo'si" |
| CAUSE _TO_ MOVE _IN_ PLACE Lamp | RITE lamp, vessel, key | | TEXT _ CREATION „agnir mūrdhā" | TEXT _ CREATION „trātāram indram" | CAUSE _ TO_MOVE _ IN_PLACE lamp, vessel, key | | | TEXT _ CREATION „asura-ghnam" |

**Fig. 2** Four aligned ritual descriptions

We extract the event sequences from each description and align them using an MSA algorithm. This results in the alignment shown in Fig. 2. In this representation of the alignment, identical frames – which may still have different roles and role fillers – are aligned to the same columns in a table (refer to the highlighted column in Fig. 2).

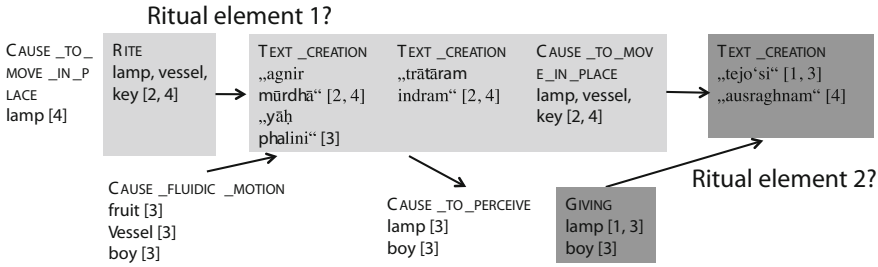


**Fig. 3** Compact representation of four aligned rituals and possible subsequences

In the next step, these frames are merged into a more compact graph representation, as displayed in Fig. 3. Such a representation can serve as a starting point for further research in ritual structures. The frequency information associated with each node of the graph makes it possible to detect frequent substructures of the aligned ritual sequences. In Fig. 3, for example, the highlighted subsequences can be identified as substructures (and potential *ritual elements*) shared by rituals 1 and 2, and 1 and 3, respectively, as indicated in square brackets. In a further step it will be possible

to inspect the inner structure of the nodes, to detect similarity and differences of frame roles and the conceptual classes of their fillers.

# 7 Future Work and Conclusions

## *7.1 Future Work*

As we have seen, anaphora and coreference resolution is currently an unsolved issue. We have performed a detailed error analysis of the available systems and identified a number of strategies using domain knowledge that we plan to explore and integrate with the available state-of-the-art ACR systems, in order to achieve reasonable performance with respect to the overall task.

A number of annotation tasks are still to be addressed in future work. Word sense and named entity annotations are needed as a basis for semantic annotation and the structural analysis of rituals. As we established in a pre-study, many ritual-specific concepts are not included in WordNet. Also, named entities occurring in ritual descriptions can often not be classified into the standard classes or do not appear in gazetteer lists. Thus, we expect that both word sense disambiguation and named entity recognition systems and resources need to be adapted to the ritual domain.

Using the types of annotations discussed in this paper, we will create structured and normalized semantic representations for ritual descriptions that are linked to an ontology comprising general and ritual-specific concepts and relations. This allows us to offer powerful querying functionalities for ritual researchers, so that they can test and validate their hypotheses against a corpus of structurally analyzed ritual descriptions. A well-defined and populated ontology can also be used to automatically identify event sequences in the data, and to inform coreference resolution.

Sequence analysis and the automatic detection of structure in rituals are the second focus of our future research. As soon as enough data has been semantically annotated, we plan to develop computational methods that support ritual researchers in finding recurrent patterns and variations in the ritual descriptions. Methods that will be adapted for this purpose include modeling of selectional preferences, as well as algorithms for detecting frequent item sets and statistical tests of significance.

## *7.2 Conclusions*

In this paper, we presented a detailed investigation of the performance of standard NLP tools and resources for the computational linguistic analysis of ritual descriptions. As standard “out of the box” tools perform poorly and lexical resources are lacking coverage and the appropriate granularity, the adaptation of tools and resources to different domains emerges as an important focus of our work. We observed that standard NLP tools behave poorly on our domain, but we have shown that we can improve the results significantly with rather small effort. This finding supports our basic tenet, that it is possible to make use of computational linguistics methods for the semantic and quantitative analysis of ritual texts. Further work will have to establish whether the representations we compute will allow us to help ritual researchers establish novel insights on the structure(s) of rituals.

Our work also explores to what degree methods of computational linguistics can be adapted to the needs of the Humanities. By using a rarely applied combination of computational and traditional scholarship, we are optimistic to achieve results that extend the knowledge in the field

of ritual research to a considerable degree. Moreover, we hope to open up new, more formal data-oriented ways for research in the Humanities.

**Acknowledgements** This research has been funded by the German Research Foundation (DFG) through the collaborative research center on ritual dynamics (Sonderforschungsbereich SFB-619, Ritualdynamik).

## References

1. Bagga, A., Baldwin, B.: Algorithms for Scoring Coreference Chains. In: Proceedings of the LREC 1998 Linguistic Coreference Workshop, pp. 536–566. Granada, Spain (1998)
2. Barzilay, R., Lee, L.: Learning to Paraphrase: An Unsupervised Approach Using Multiple-Sequence Alignment. In: Proceedings of the 2003 Human Language Technology Conference of the NAACL (HLT-NAACL '03), pp. 16–23. Edmonton (2003)
3. Bögel, T., Funk, L., Kull, A.: ELAC: Ensemble Learning for ACR-Systems. Software project, University of Heidelberg, http://dakhma.net/elac (2010)
4. Bortz, J.: Statistik für Human- und Sozialwissenschaftler, 6. edn. Springer Medizin Verlag, Heidelberg (2005)
5. Burchardt, A., Frank, A., Pinkal, M.: Building Text Meaning Representations from Contextually Related Frames – A Case Study. In: Proceedings of the 6th International Workshop on Computational Semantics (IWCS '05) (2005)
6. Burchardt, A., Pado, S., Spohr, D., Frank, A., Heid, U.: Constructing Integrated Corpus and Lexicon Models for Multi-Layer Annotations in OWL DL. Linguistic Issues in Language Technology **1**(1), 1–33 (2008)
7. Chambers, N., Jurafsky, D.: Unsupervised Learning of Narrative Event Chains. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT '08), pp. 789–797 (2008). URL http://www.aclweb.org/anthology/P/P08/P08-1090
8. Chambers, N., Jurafsky, D.: Unsupervised Learning of Narrative Schemas and their Participants. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP '09), pp. 602–610 (2009). URL http://www.aclweb.org/anthology/P/P09/P09-1068
9. Charniak, E.: A Maximum-Entropy-Inspired Parser. In: Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics (NAACL '00) (2000)
10. Daumé III, H.: Frustratingly Easy Domain Adaptation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07), pp. 256–263 (2007). URL http://www.aclweb.org/anthology/P07-1033
11. Denis, P.: New Learning Models for Robust Reference Resolution. Ph.D. thesis, University of Texas at Austin, Austin, TX, USA (2007)
12. Erk, K., Kowalski, A., Padó, S.: The SALSA Annotation Tool. In: Proceedings of the Workshop on Prospects and Advances in the Syntax/Semantics Interface (2003)
13. Erk, K., Padó, S.: Shalmaneser – a Toolchain for Shallow Semantic Parsing. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06) (2006)
14. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
15. Fillmore, C.J., Johnson, C.R., Petruck, M.R.: Background to FrameNet. International Journal of Lexicography **16**(3), 235–250 (2003)
16. Finkel, J.R., Manning, C.D.: Hierarchical Bayesian Domain Adaptation. In: Proceedings of the 2009 Human Language Technologies Conference of the NAACL (HLT-NAACL '09), pp. 602–610 (2009). URL http://www.aclweb.org/anthology/N/N09/N09-1068

17. Hellwig, O.: A Chronometric Approach to Indian Alchemical Literature. Literary and Linguistic Computing **24**(4), 373–383 (2009)
18. Hyland, K.: Hedging in Academic Writing and EAP Textbooks. English for Specific Purposes **13**(3), 239–256 (1994)
19. Jiang, J., Zhai, C.: Instance Weighting for Domain Adaptation in NLP. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07), pp. 264–271 (2007). URL http://www.aclweb.org/anthology/P07-1034
20. Kipper, K., Korhonen, A., Ryant, N., Palmer, M.: A Large-Scale Classification of English Verbs. Journal of Language Resources and Evaluation **42**(1), 21–40 (2008)
21. Lappin, S., Leass, H.J.: An Algorithm for Pronominal Anaphora Resolution. Computational Linguistics **20**(4), 535–561 (1994)
22. Light, M., Qiu, X.Y., Srinivasan, P.: The Language of Bioscience: Facts, Speculations, and Statements in Between. In: Proceedings of HLT-NAACL 2004 Workshop on Linking Biological Literature, Ontologies and Databases (BioLINK '04), pp. 17–24 (2004)
23. Medlock, B., Briscoe, T.: Weakly Supervised Learning for Hedge Classification in Scientific Literature. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL '07), pp. 992–999 (2007). URL http://www.aclweb.org/anthology/P/P07/P07-1125
24. Needleman, S.B., Wunsch, C.D.: A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. Journal of Molecular Biology **48**, 443–453 (1970)
25. Poesio, M., Kabadjov, M.A.: A General-Purpose, Off-the-Shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC '04) (2004)
26. Pradhan, S., Ward, W., Martin, J.H.: Towards Robust Semantic Role Labeling. Computational Linguistics, Special Issue on Semantic Role Labeling **34**(2), 289–310 (2008)
27. Pradhan, S.S., Ward, W., Hacioglu, K., Martin, J.H., Jurafsky, D.: Shallow Semantic Parsing using Support Vector Machines. In: Proceedings of the 2004 Human Language Technology Conference of the NAACL (HLT-NAACL '04) (2004)
28. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Wien (2007)
29. Recasens, M., Martí, T., Taulé, M., Màrquez, L., Sapena, E.: SemEval-2010 Task 1: Coreference Resolution in Multiple Languages. In: Proceedings of the HLT-NAACL 2009 Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW '09), pp. 70–75 (2009)
30. Regneri, M., Koller, A., Pinkal, M.: Learning script knowledge with web experiments. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10), pp. 979–988. Uppsala, Sweden (2010). URL http://www.aclweb.org/anthology/P10-1100
31. Reiter, N., Hellwig, O., Mishra, A., Frank, A., Burkhardt, J.: Using NLP methods for the Analysis of Rituals. In: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10) (2010)
32. Reiter, N., Hellwig, O., Mishra, A., Gossmann, I., Larios, B.M., Rodrigues, J., Zeller, B., Frank, A.: Adapting Standard NLP Tools and Resources to the Processing of Ritual Descriptions. In: Proceedings of ECAI 2010 Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH '10) (2010). URL http://www.cl.uni-heidelberg.de/~reiter/publications/Reiter2010b.pdf
33. Ruppenhofer, J., Sporleder, C., Morante, R., Baker, C., Palmer, M.: SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW '09), pp. 106–111 (2009)
34. Sang, E.F.T.K., Buchholz, S.: Introduction to the CoNLL-2000 Shared Task: Chunking. In: Proceedings of the 2nd Workshop on Learning Language in Logic and the 4th conference on Computational Natural Language Learning (CoNLL '00 and LLL '00) (2000)

35. Soon, W.M., Lim, D.C.Y., Ng, H.T.: A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics **27**(4), 521–544 (2001)
36. Szarvas, G., Vincze, V., Farkas, R., Csirik, J.: The BioScope Corpus: Annotation for Negation, Uncertainty and their Scope in Biomedical Texts. In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP '08), pp. 38–45 (2008). URL http://www.aclweb.org/anthology/W/W08/W08-0606
37. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix cxshoice. Nucleic Acids Research **22**(22), 4673–4680 (1994)
38. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Proceedings of the 2003 Human Language Technologies Conference of the NAACL (HLT-NAACL '03), pp. 252–259 (2003)
39. Verhagen, M., Pustejovsky, J.: Temporal Processing with the TARSQI Toolkit. In: Proceedings of the 22nd International Conference on Computational Linguistics: Demonstration Papers (COLING '08), pp. 189–192. Manchester, UK (2008). URL http://www.aclweb.org/anthology/C08-3012
40. Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A.: BART: A Modular Toolkit for Coreference Resolution. In: Companion Volume of the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL '08), pp. 9–12 (2008)
41. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A Model-Theoretic Coreference Scoring Scheme. In: Proceedings of the 6th Conference on Message Understanding (MUC6 '95), pp. 45–52. Morristown, NJ, USA (1995)

# Part VI
# Data Management, Visualisation and Retrieval

# Information Retrieval and Visualization for the Historical Domain

Yevgeni Berzak, Michal Richter, Carsten Ehrler and Todd Shore

**Abstract** Working with large and unstructured collections of historical documents is a challenging task for historians. Despite the recent growth in the volume of digitized historical data, available collections are rarely accompanied by computational tools that significantly facilitate this task. We address this shortage by proposing a visualization method for document collections that focuses on the graphical representation of similarities between documents. The strength of the similarities is measured according to the overlap of historically significant information such as named entities, or the overlap of general vocabulary. Similarity strengths are then encoded in the edges of a graph. The graph provides visual structure, revealing interpretable clusters and links between documents that are otherwise difficult to establish. We implement the idea of similarity graphs within an information retrieval system supported by an interactive graphical user interface. The system allows querying the database, visualizing the results and browsing the collection in an effective and intuitive way. Our approach can be easily adapted and extended to collections of documents in other domains.

**Key words:** historical collections, information retrieval, graph visualization, clustering, recommender system

## 1 Introduction

The availability of historical documents in digital form has been constantly increasing in recent years. Digitization of sources is extremely valuable for historians, as it contributes to preservation, facilitates accessibility and enables exploiting computational methods. Despite the growth in the

Yevgeni Berzak
Saarland University, 66041 Saarbrücken, Germany, e-mail: berzak@coli.uni-sb.de

Michal Richter
Saarland University, 66041 Saarbrücken, Germany, e-mail: mrichter@mpi-inf.mpg.de

Carsten Ehrler
Saarland University, 66041 Saarbrücken, Germany, e-mail: cehrler@mpi-inf.mpg.de

Todd Shore
Saarland University, 66041 Saarbrücken, Germany, e-mail: tshore@coli.uni-saarland.de

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_11, © Springer-Verlag Berlin Heidelberg 2011

volume of digitized historical data, available collections are rarely accompanied by supporting tools which significantly facilitate the work of historians.

Existing interfaces typically include a simple keyword or metadata search, providing users with a list of documents that match their query. Although such tools can indeed spare valuable time dedicated to manual work, they are far from exploiting the full power and flexibility that state of the art Natural Language Processing (NLP) and Information Retrieval (IR) technology has to offer. Moreover, the systems provided are usually generic, and rarely address the specific needs of researchers in the historical domain. This situation calls for the development and adaptation of NLP techniques for historical data, as well as for the creation of user interfaces that would enable historians to use this technology effectively, in a way that would meet their needs.

This work addresses both aspects of the current shortage. We take up the NLP domain adaptation challenge by applying and tailoring NLP tools that extract information relevant for historians and create links between documents according to similarity with regard to this information. The need for intuitive user interfaces is addressed by providing an interactive graphical tool that enables historians without computational background to benefit from NLP technology. This twofold approach aims at improving the chances of historians working with digitized sources to find information that is relevant for their research goals.

The core idea of our approach is to extract information with special importance for the historical domain, such as Named Entities (NEs) of the types PERSONS, LOCATIONS and ORGANIZATIONS mentioned in each document, and then use this information to determine similarity between documents. Our working hypothesis is that the higher the similarity between two documents is according to the NEs each contains, the more probable it is that they are related to each other in an interesting way. In addition, we also use a generic, domain independent similarity measure based on the remaining vocabulary of the collection. These different similarity measures can be used separately or in combination with one another. The measured similarity rates can be interpreted as strength of potential connections between documents and visualized as edges of a graph.

We incorporate this idea in an IR system wrapped with an interactive Graphical User Interface (GUI) that provides powerful search operations and allows for visual navigation through collections of historical documents. We exemplify our approach on a collection of speeches and other oratory materials by Fidel Castro. Our system includes keyword and metadata search, and retrieved documents are presented in a table synchronized with an interactive scalable graphical network that connects related documents. This representation is designed to support effective visual navigation in the collection, allowing the identification of documents which revolve around similar topics, distinguishing relevant from irrelevant documents, and also enabling the exploration of specific relations between any subset of the retrieved documents. Our system can also be viewed as a recommender tool: given an interest in a specific document, the user can easily identify documents that are most similar, and hence potentially related to it.

To the best of our knowledge, this is the first system of its kind. Althought the current implemetation focuses on the historical domain, our approach can be easily adapted to other types of databases, and is intended as a general alternative to current information retrieval systems.

This chapter is structured as follows: Section 2 presents related work and background. Section 3 describes the dataset we use and the information extraction process. In Sect. 4, we elaborate on the similarity measurements and the visualization of the collection according to these measurements. Section 5 describes the GUI that realizes our visualization approach, and in Sect. 6, we illustrate its use for historians and exemplify its advantages. Finally, we discuss future research perspectives in Sect. 7.

## 2 Background

This work can be located within the field of Domain Knowledge Visualization. This field is centered around visualization techniques of domain structures, in particular of scientific domains.

Its notable applications are mapping structures of domains, and supporting IR and information classification.

The general process flow of visualizing domain knowledge as described in [4] is the following:

1. Collection of data (i.e. extraction of documents belonging to the domain)
2. Definition of unit of analysis (e.g. documents, authors, terms)
3. Selection of similarity measures and calculation of similarity between units
4. Ordination or assignment of coordinates to units
5. Use of resulting visualization for analysis and interpretation

Our workflow conforms to these categories. We work with a historical database and use documents as the units of analysis. The cosine measure is used to determine similarity between pairs of documents. Finally, we visualize subsets of documents, selected according to a specified query, as graphs ordinated with a force-based layout algorithm.

While NLP and semantic web technologies are now commonplace in IR [9], the success of sophisticated visualization techniques is less widespread, with many users preferring traditional list-based results [10]. In this work, we attempt to design an interface that is easy to use for experts and laymen alike, taking inspiration from other interactive graph-based systems like iOPENER [5] and VisuWords [28].

We use tools and algorithms for NE Recognition and string similarity in order to extract the information according to which inter-document similarity is measured. The Vector Space Model (VSM) [20] serves as our framework for representation of documents in the collection. The VSM enables straightforward modeling of inter-document similarity as proximity of vectors in a multidimensional vector space. Furthermore, this representation is a standard approach in IR, allowing ranked retrieval of parts of the collection that match the user's query.

Our graphical model responsible for ordination of the documents is reminiscent of Pathfinder Networks (PN) [23]. Modeled as PNs, collections of documents may be presented as graphs which capture the relative proximities of objects to each other. Objects are represented as nodes and proximities between objects are represented as links between nodes, where only the strongest links according to an adjustable threshold are taken into consideration. Proximity can have several interpretations, one of them being similarity. In our model, the degree of similarity is reflected in the thickness of the edges. This approach provides an aesthetic, intuitive and transparent representation that conforms to the similarity relations between the documents in the collection.

## 3 Information Extraction from a Historical Collection

Our approach for detecting historically useful relations between documents focuses on exploiting domain relevant information for similarity measurements. One important type of such information in the historical domain is NEs. In this work, we address three types of NEs: PERSONS, LOCATIONS and ORGANIZATIONS. All three play an important role in historical documents in general, and collections of modern political history in particular. Moreover, these are well studied classes of NEs [18] and existing tools for their identification and classification perform well. In the following section, we describe our data and elaborate on the information extraction procedure.

### *3.1 Dataset*

Our case study for a collection of historical documents is the Castro Speech Database (CSDB) [13], maintained by the Latin American Network Information Center (LANIC) at the University of Texas at Austin. This collection contains 1492 English translations of speeches, interviews, press

conferences and other oratory materials by Fidel Castro from 1959 to 1996. Most of the documents in the database are annotated with metadata, including document type (e.g. SPEECH, INTERVIEW, MEETING), date, place, author, source and headline.

The documents are manual translations from modern Spanish into modern English. This characteristic, along with a considerable amount of newswire content, allows a relatively straightforward application of tools and models already deployed and currently used in NLP and IR.

However, the collection is very heterogeneous and comprises different genres and styles. Additionally, many documents in the corpus are based on spoken language, featuring heavily rhetorical content and vague structure. Given these characteristics, identification of useful relations between documents is a particularly challenging task.

## 3.2 Extraction of Named Entities

In order to recognize NEs in the documents of the CSDB, we use the Stanford Named Entity Recognizer [7]. This tool is based on Conditional Random Fields (CRF) using a wide range of features. It is available with pre-trained models and robust across domains, a property that is highly desirable for our diverse database.

For the purpose of computing the similarities, the recognized NEs can be regarded as reduced forms of the original documents. Following the extraction, the documents of the collection $D = \{d_1, d_2, \ldots, d_N\}$ are indexed as vectors in four distinct term spaces $\mathscr{T} = \{T_{PER}, T_{LOC}, T_{ORG}, T_{VOC}\}$ corresponding to the three types of named entities extracted from the collection and an additional term space for the general vocabulary. The general vocabulary term space contains all the content words in the collection that do not belong to the NE terms spaces. We consider two standard weighting schemes for measuring the importance of a term $t \in T$ for a document $d \in D$, namely TF and TF/IDF. An index matrix $I_T = \mathbb{R}^{|D| \times |T|}$ is computed for each term space $T \in \mathscr{T}$ and a weighting scheme $w \in \{\text{TF}, \text{TF/IDF}\}$. Each position in the matrix $I_{i,j}$ contains the score $w$ of term $t_j \in T$ in document $d_i \in D$.

## 3.3 Aliasing

Relying on the raw output of a NE recognizer is not sufficient for obtaining reliable counts of NEs, as the same NE can be manifested in a variety of phrases. For example, *Fidel Castro* and *Dr. Fidel Castro* are identified as distinct entities by the NE recognizer, while referring expressions to the same entity are ignored completely. In order to determine how many times each NE appears within and across documents, we should be able to identify all the phrases that refer to each real-world entity, mapping multiple linguistic variations to a single referent — a task known as Coreference Resolution.

Coreference Resolution is a complex problem, where state of the art tools achieve only moderate accuracy [27]. Due to the limitations of existing technology and the nature of our task, we restrict ourselves to multi-document aliasing, i.e. recognizing linguistic variants of names in a collection of documents. Aliasing of NEs (or in general strings) is usually done by employing a *string-similarity* measure, e.g. the *edit distance*, possibly accompanied by additional structural information that translates the subject string into a tree or graph based structure [26, 27].

### 3.3.1 Kernel Functions

The aliasing approach implemented in this work relies solely on string similarity. For each of the four term spaces $T_{\text{PER}}, T_{\text{LOC}}, T_{\text{ORG}}, T_{\text{VOC}}$, the similarity between NEs within that class is measured using a string kernel function [24].

In general, kernel functions are a way to compute pairwise similarities between objects from a category $C$ (e.g. graphs or strings). The key idea is, given a function $\Psi : C \rightarrow \mathbb{R}^n$ mapping elements $c \in C$ from the category into a vector space $\mathbb{R}^n$, a natural similarity measure is provided by the usual dot product

$$\textbf{sim}(c_1, c_2) := \langle \Psi(c_1), \Psi(c_2) \rangle.$$

Although the construction of such a mapping $\Psi$ into a vector space $\mathbb{R}^n$ is not straightforward for many domains including graphs, strings and trees, a commonly-used shortcut known as a 'kernel trick' [22] obviates the construction of this mapping. Instead, one can directly define a function $K : C \times C \rightarrow \mathbb{R}$ with the following property: if for all $c_1, ..., c_m \in C$ and constants $a_1, ..., a_m \in \mathbb{R}$, $K$ is a symmetric function such that

$$\sum_{i,j=1}^{n} a_i a_j K(c_i, c_j) \geq 0$$

holds, then $K$ is a positive-definite kernel [24]. Now, for every such positive definite kernel $K$, theory asserts the existence of a corresponding vector space together with a mapping $\Psi$ such that

$$K(c_1, c_2) := \langle \Psi(c_1), \Psi(c_2) \rangle.$$

However, one does not have to construct either the mapping or the space explicitly.

### 3.3.2 String Kernels

For the domain of strings, a large number of kernels have been proposed, e.g. *subsequence–* or *spectrum* kernels [15], and successfully applied to a variety of problems ranging from homology detection in bioinformatics to text classification in IR [14, 17].

In this work, a $p$-spectrum string kernel was used for aliasing. Given strings over an alphabet $\Sigma$, the $p$-spectrum kernel is defined by Eq. (1). Where $\phi_u^p(x)$ counts the number of occurrences of a substring $u$ of length $p$ in string $x$.

$$sk(s,t) = \sum_{u \in \Sigma^p} \phi_u^p(s) \phi_u^p(t) \tag{1}$$

Intuitively, the sum ranges over all $p$-grams that can be built from the alphabet $\Sigma$. The kernel then sums up the number of p-grams both $s$ and $t$ agree upon. The associated vector space for this kernel is $\mathbb{R}^{|\Sigma^p|}$, the space of all possible strings over $\Sigma$ of length $p$. Note that although the vector space is typically very large, an efficient implementation of the $p$-spectrum string kernel can be achieved by using trie data structures over $n$-gram models.

The advantage of this method over more traditional string similarity metrics such as edit distance is its flexibility, especially with regard to word order and variations. For example, the strings *Public Health Ministry* and *Ministry of Public Health* which have a large Edit Distance, are highly similar according to the $p$-spectrum kernel. The disadvantage of the method is that its flexibility may lead to over-generation by also considering string pairs such as *Polish People's Republic* and *Lao People's Republic* as highly similar. The over-generation can be reduced to some extent by setting a high similarity threshold for considering two NEs as aliases of each other.

The $p$-spectrum kernel was used to compute a kernel matrix $\mathbf{K}^T$ for each of the four term spaces $T \in \mathscr{T}$ such that $\mathbf{K}^T_{i,j}$ is the similarity between terms $t_i$ and $t_j$:

$$K^T_{i,j} = sk(t_i, t_j) \quad \forall t_i, t_j \in T$$

As a post-processing step, similarity values below a predefined threshold $t_{min}$ are discarded. Accordingly, aliased document vectors $\tilde{t}$ are computed by multiplying the kernel matrix $\mathbf{K}^T$ with the original document vectors $t$.

$$\tilde{t} = \mathbf{K}^T t$$

Through this expansion, aliases of names that appear in the documents term space $T$ receive additional weight. A similar expansion is used for queries that contain NEs. The aliasing method is exemplified in Fig. 1.

$$\underbrace{\begin{pmatrix} 1 & 0.8 & 0 \\ 0.8 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{K} \underbrace{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}}_{t} = \underbrace{\begin{pmatrix} 1 \\ 0.8 \\ 0 \end{pmatrix}}_{\tilde{t}}$$

**Fig. 1** Aliasing of a document with a string kernel similarity measure. The matrix $K$ is the kernel similarity matrix of the three terms $T = \{Fidel\ Castro, Dr.\ Fidel\ Castro, Raul\}$. A document vector $t = (1,0,0)^{\mathrm{T}}$ that contains only the term *Fidel Castro* is expanded into an aliased form $\tilde{t} = Kt$, in which the alias *Dr. Fidel Castro* has a non-zero weight.

Note that in this method, the importance of a NE in an aliased document depends on the number of its aliases. Given the presence of two NEs $NE_1$ and $NE_2$ with an equal number of appearances in a given document, if $NE_1$ has more aliases than $NE_2$ in the document collection, its importance in the document at hand will be greater. To prevent this effect, it is possible to perform normalization of the kernel matrix or the aliased documents.

The aliased document representation serves two purposes. First, we receive more reliable similarity measures between documents. Secondly, the flexibility of the querying mechanism is increased by expanding NE terms in the query to all their aliases, allowing retrieval of name variations for query keywords that are NEs.

# 4 Visualization of Document Similarities

Given the vector representations of documents, we can obtain their pairwise similarities and present them graphically according to this information. However, it is impossible to visualize multidimensional vectors directly, and therefore such representation must be reduced to two or three dimensions, a process often referred to as *ordination* [4]. In this section, we elaborate on our ordination approach for the presentation of multidimensional vectors in a 2D space which adheres to their similarities.

## *4.1 Similarity measurement*

Using the constructed indexes described in Sect. 3, we measure and store the similarity of each pair of documents in the collection. To determine their similarity, we use the standard cosine measure as defined in (2).

$$cosim(v, v') = \frac{v \cdot v'}{vv'} \tag{2}$$

It expresses the cosine of the angle between the document vectors $v$ and $v'$.

Separate similarity matrices are computed for each combination of term space, indexing scheme and aliasing setting. The similarity matrices of the different term spaces can be combined. For this purpose, a weight vector $w = [w_{\text{PER}}, w_{\text{LOC}}, w_{\text{ORG}}, w_{\text{VOC}}]$ is defined. The final similarity matrix is constructed as a weighted combination of the similarity matrices of the different term spaces according to the specified weight vector. By setting $w = [0.5, 0.25, 0.25, 0]$, we express that *persons* are twice as important as *organizations* and *locations* and that the general vocabulary is ignored.

## 4.2 Visualization of similarities

In order to visualize the documents as a 2D graph, we represent documents as nodes and encode the similarity rates between pairs of documents as weights for the edges connecting them. An edge is established only if the similarity between a pair of nodes exceeds a certain threshold. The thickness of each edge corresponds to the similarity measurement between the pair of documents it connects: the stronger the similarity, the thicker the edge. This gives an indication for the nodes similarity regardless of their positions in a graph. The graph is layouted on a rectangular surface using Force Directed Placement (FDP), resulting in a layout that facilitates the visual recognition of the similarity relations between nodes.

### 4.2.1 Force directed placement

FDP algorithms [6, 8] are a class of algorithms that is based on physical modeling of elements. In particular, nodes are modeled as physical objects that are electrically charged with the same sign, and the edges are modeled as springs. In this work, the default Jung [11] FDP implementation was used. In this implementation, there is a repulsion force $F_r$ between each pair of nodes such that:

$$F_r(v_1, v_2) = \begin{cases} 0 & \text{if } d(v_1, v_2) > d_{max} \\ \frac{C_r}{d(v_1, v_2)} & \text{otherwise} \end{cases} \tag{3}$$

where $C_r$ is a repulsion constant, $d(v_1, v_2)$ denotes the distance between nodes $v_1$ and $v_2$, and $d_{max}$ denotes the maximum distance on which the repulsion force takes effect.

An edge between two nodes $v_1$ and $v_2$ creates a spring force $F_s$ such that:

$$F_s(v_1, v_2) = C_s(d_{opt} - d(v_1, v_2))C_{stretch}^{deg(v_1)+deg(v_2)-2} \tag{4}$$

where $d_{opt}$ is the optimal edge length, $deg(v)$ stands for the number of edges adjacent to node $v$, $C_s$ is a spring constant and $C_{stretch}$ is a stretching constant. $C_{stretch}$ is set to a value between 0 and 1, which has the effect that the magnitude of the spring force decreases as deg$v_1$ and deg$v_2$ increase. The spring force affects both $v_1$ and $v_2$. In our implementation, $C_s$ is defined with regard to the measured similarity rate between the nodes conntected by the spring.

In each iteration of the algorithm, each node is moved according to the resultant force that acts on it. After several iterations, the position of the nodes becomes (quasi-)stationary and the system stabilizes in a local equilibrium state.

Since our framework contains features that allow dynamic changes to the graph, such as changing the edge density and node filtering, the layouting procedure should be capable of responding to small graph configuration changes by small changes in the layout. The iterative

nature of the algorithm can be utilized to present the successive layouts to the user, which results in smooth transitions between layouts when the graph configuration is changed.

### 4.2.2 Graph Clustering

One of the major advantages of our visualization approach is the ability to identify groups of documents that are highly similar to each other. Such groups may be informative for inferring common topics or help filtering out documents less relevant for the interests of the user.

Given the nature of our graph representation, it is often possible to identify groups of strongly interconnected documents without any manipulation of the graph. However, in many instances, an automatic identification of dense regions and re-arranging of the graph according to these regions might be useful.

In order to enable this functionality, Chinese Whispers clustering (CWC) [1] is used: CWC is a non-parametric algorithm that is applied on the nodes of weighted undirected graphs, such as our data structure. Since we do not know in advance the number of clusters that will emerge from our data, it is in fact desirable that CWC is non-parametric. This algorithm is also very efficient: its time complexity is linear in the number of edges.

The CWC algorithm works in a bottom-up fashion. During initialization, each node is assigned with its own cluster. The algorithm then performs a number of iterations in which each node is sequentially assigned to the class that has the highest sum of weighted edges to the node. After few iterations, a mostly stable clustering is achieved, with at most few nodes for which the algorithm might continue changing the class assignment.

The clustering output can be further used as an initial setup for the FDP algorithm. Node positions are initialized in such a way that nodes belonging to the same cluster are closer to each other which enables FDP to converge to a better layout.

Clustering is a powerful and efficient way of enhancing the informativeness of our visualization approach. In the following section, we demonstrate how this approach can be integrated in a GUI that enables users to utilize it for their needs.

## 5 Graphical User Interface

We incorporate our approach for visualizing collections of historical documents into a GUI. The GUI allows the user to query keywords and present the outcome of the query as an interactive graph, based on the description in Sect. 4. The GUI implements additional features which enable customization of the graph and aim at maximizing the flexibility and benefit that historians can gain from using our approach. A historian at Saarland University was consulted during the design process in order to ensure that our design indeed addresses the practical needs of our intended end-users. Figure 2 shows a screenshot of the entire system. We present the main characteristics and features of the GUI in this section.

Our system provides an IR mechanism, with which the user can specify both query terms (feature 1 in Fig. 2) and constraints on metadata, such as dates and type of document (feature 2). Query terms are transformed to a vector model representation of the query. Metadata constraints are translated into database queries. Documents that do not fulfill the metadata constraints are filtered out. The relevant documents are then sorted according to their cosine similarity to the query vector. The specified number of most similar documents is retrieved as the search results. The queries can be expanded using the aliasing method described in Sect. 3.3 to contain aliases of the NE query terms.

The top results of the query are displayed both in a table (top panel) and as a graph (central panel). The table (which is perhaps the more traditional way of receiving search results) lists the results of the query ranked according to their relevance to the query. The table contains all the metadata information of the documents and allows the user to identify them easily. The graph visualizes the connections between the documents in the table. The table and the graph are synchronized, i.e. when a document is chosen in the table, it is marked in the graph, and vice versa.

The graph is scalable (allows zooming in and out) and draggable. The positions of the nodes can be rearranged by dragging them, allowing a manual adjustment of the layout. We also enable automated layouting through the FDP algorithm (feature 6), as described in Sect. 4.2.

The size of the nodes corresponds to their relevance to the query: the bigger the node, the more relevant it is. The shape of the nodes corresponds to the type of document they represent, (e.g. *star* stands for SPEECH, *circle* for REPORT, *pentagon* for MESSAGE). Selecting a node presents the NEs it contains in a separate panel (feature 3), and selecting several nodes at once provides the intersection of the NEs contained in the selected documents.

The representation of the edge thickness is quantized: edges are presented as *dotted*, *normal* or *thick*. A separate menu (feature 5) allows adjustment of parameters related to the edges. For instance, the user can specify the similarity threshold for presenting an edge, as well as thresholds for each edge type. The edge thresholds can be set on a relative scale with respect to the similarities of the presented graph, or on an absolute scale.

Another menu (feature 4) allows filtering of the nodes such that only nodes that are connected to a selected node up to a specified depth are presented. Figure 3 illustrates this feature, where only neighbors and their neighbors (depth 2) of the selected node are presented.

Additional functionalities for each node are available via a context menu, which has options for viewing the text of the document with color-coded NEs and highlighted query terms, viewing metadata related to the document and launching a new search for the documents most similar to it in the collection. With this search option, the documents are ranked according to their similarity to the focused document and the sizes of the nodes in the resulting graph are proportional to these similarities. The constraints on the metadata can be applied in the same manner as for the normal query term search.

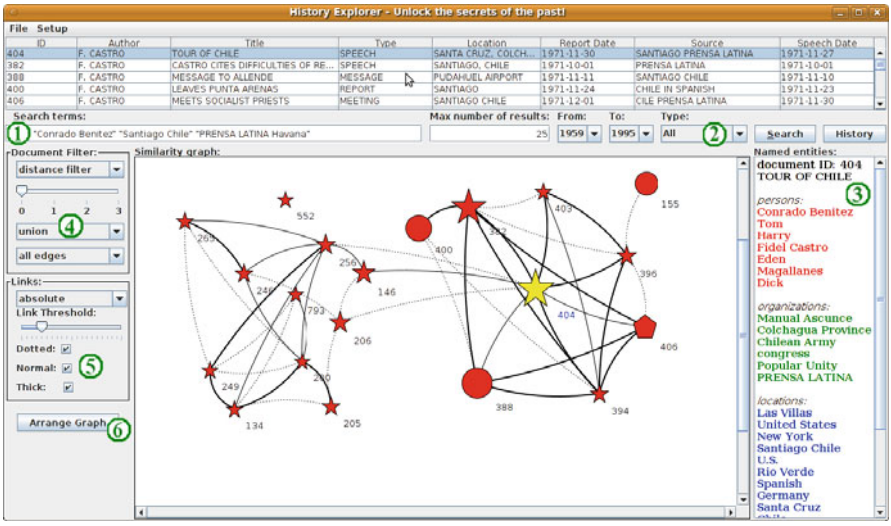


**Fig. 2** The "History Explorer" GUI

In order to provide an additional indication for the similarities strength to a particular document, the node context menu also has an option to automatically shade directly connected nodes in a range of colors (e.g. from black to white), where darker color indicates stronger similarity (see Fig. 3).

The CWC, described in Sect. 4.2.2, can be applied to a given graph. After the clustering is performed, the graph is redrawn with a layout that separates the clusters and uses a different node color for each cluster. Although the CWC is a non-parametric algorithm, we provide an option to specify the maximum number of presented clusters. If the algorithm constructs more clusters than the maximum, only the largest clusters are highlighted in the resulting graph: For example, if the algorithm constructs 10 clusters, but the maximum is set to 3, the nodes of the 1st-3rd (sorted in descending order according to the number of nodes each cluster comprises) will be highlighted, while the nodes of the 4-10th clusters will be assigned the default red color. It is also possible to define a required minimal size of the cluster in order to be highlighted. Figure 4 shows a graph after applying the clustering algorithm.

The GUI also contains advanced options for setting the weighted combination of the term spaces for the similarity measurements, choice between weighting schemes, enabling and disabling aliasing for search and for similarity measurements, as well as other parameters related to the retrieval and presentation of the graph.

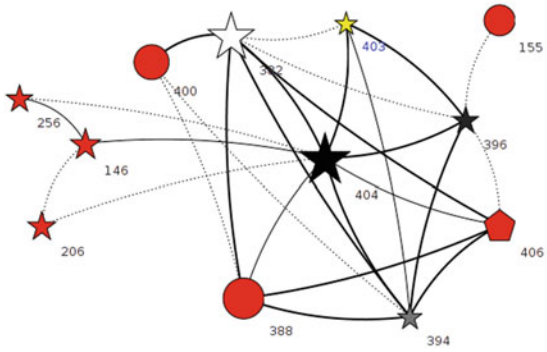


**Fig. 3** A graph that shows a chosen node (marked in yellow) after coloring the immediate neighbors of the node in a graded color scale according to similarity, and applying depth 2 filtering. The complete graph is presented in Fig. 2

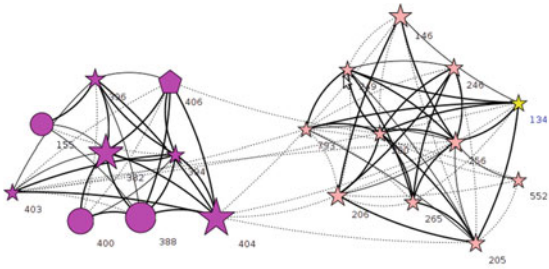


**Fig. 4** A graph with two distinct clusters that were identified using CWC

## 6 The Benefit for Historical Research

Many of the tasks historians face involve working with sources that often come in the form of historical documents. Such documents have numerous usages in the various branches of historical research. Some of these are related to the task of revealing general trends and patterns about a historic period or personality while others focus on discovering specific details and pieces of information. Historical documents are used both for the formulation of historical hypotheses and for their validation and rejection efforts.

While the range of written accounts of ancient history is well-known and relatively limited, researchers of modern times often have to face an abundance of written materials. Dealing with large collections of documents introduces additional challenges for historians. In particular, the focus is often shifted to the identification and retrieval of documents that might be of relevance. Furthermore, identifying trends as well as implicit and explicit connections between documents becomes an extremely difficult task if performed manually.

Our system is designed to support historians' work with electronic sources, specifically with large collections of documents such as the CSDB. The system presents the user with a visual *structure* that helps to discover *new knowledge* by highlighting interesting inter-document connections that are otherwise hidden.

The basic idea according to which the graphs are constructed is easily understandable. The produced visualizations allow for an intuitive interpretation which does not require the user to have any computational background. At the same time, they exploit a range of advanced techniques of text processing, IR and Data Visualization, which are utilized to produce the desired results.

A particular strength of our system is the ability to combine interactive search with visualization. While the former can be used to express a specific historical question, the latter helps in finding answers and formulating new questions.

Without any undesirable over-simplification regarding the original data, the types of connections or any other information, users receive a representation of the data that can considerably improve their ability to locate, infer and discover relevant information. In this sense, the presented approach is applicable to real research problems in the historical domain. Following are concrete aspects of our work that are likely to be appealing to historians.

We focus on automatic markup of NEs of types that are of potential interest for historians, both in terms of discovery of new entities and identification of known entities.

The IR system includes a query expansion mechanism that allows the retrieval of documents containing form variations of the entities appearing in the query. Our retrieval mechanism and graphical interface incorporates metadata, if such is provided with the documents in the collection. Using metadata further extends the flexibility of the search and allows a more informative graphical presentation of query results.

We present links between retrieved documents based on the overlap of NEs or general lexical overlap. These links can be valuable in many scenarios. For instance, a user interested in a specific document can easily identify which other documents contain similar NEs or similar vocabulary. The linking of documents in this way also supports the historian in inferring global statements about the collection or one of its subsets. In particular, one can identify groups of highly interconnected documents. Identified dense regions are likely to reflect different kinds of topics and might be correlated to various parameters, such as events, time and location. Absence of links and identification of "stand-alone" documents can also be highly informative, as they can indicate unique content.

Besides providing structure according to NEs, the application can also help to discover NEs that play an important role in a specified topic or affair that is expressed in the query or emerged as a graph cluster. This is achieved by listing shared NEs in a set of documents. Such information would be very difficult to infer in a big collection of documents that is accompanied only by a keyword search mechanism.

A query example demonstrating some of the advantages of our approach is presented in Fig. 5, showing the resulting graph for the query `health`: The graph includes two groups of interconnected documents. While the bigger group contains documents concerning or strongly

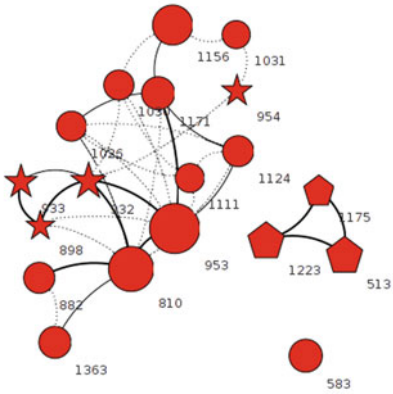


**Fig. 5** The resulting graph for the query "health", limited to 20 top documents

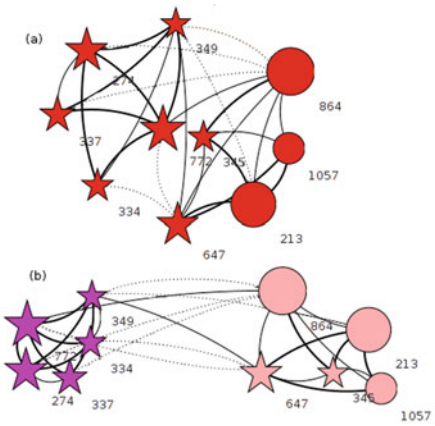


**Fig. 6** The resulting graphs for the query "Giron Kennedy" before clustering (a), and after clustering (b)

related to healthcare and health reforms in Cuba in the '80s, the smaller group has three documents that deal with a completely different topic. All three are descriptions of meetings of Castro with officials of the Romanian government. They were retrieved simply because the participants of the meetings exchange greetings, wishing each other "good health", and thus the phrase occurred in each of the documents. Grouping and separation between relevant and irrelevant documents is a clear benefit of connecting similar documents that cannot be seen using only standard keyword search. Furthermore, within the group that contains the relevant results, documents that discuss similar specific issues such as hygiene or elderly care tended to be connected more strongly than other documents. Such connections are visible in the thickness of the edges or by marking neighbors of specific documents for graded similarity.

In some graphs that contain large bundles of interconnected documents, a further insight can be gained by applying clustering. For example, a query consisting of `Giron` and `Kennedy` produces a highly interconnected graph. This graph can be split into two interpretable clusters using the CWC. Both layouts are shown in Fig. 6. In the clustered layout, one cluster contains documents directly related to the 1961 Bay of Pigs Invasion, e.g. speeches on the anniversary of the event or

victory speeches. The other cluster is composed of documents in which the invasion is mentioned but the document is not directly related to the event. These documents deal with different political and economic aspects of the US-Cuban relations and are thus related to each other. Throughout the different decades during which those documents were composed, the Bay of Pigs Invasion remained a symbol for the hostile nature of these relations.

The provided query examples demonstrate some of the benefits historians can gain from our approach. Nevertheless, it is important to note that in some cases, informative interpretation of the additional information provided by the system is rather challenging.

Overall, our GUI aims to be both intuitive and simple while allowing a considerable amount of flexibility. Finally, our system is designed to make retrieval and navigation through collections easy and enjoyable. We provide text viewing possibilities, and graph manipulation operations such that the user would be able to explore the collection effectively and with minimal effort.

# 7 Conclusion and Outlook

The presented visualization approach supports historians in their work with collections of historical documents. This is achieved by extracting historically relevant information about the documents and by exploiting this information in order to determine potential relations between documents. We organize the data and present it graphically, in a way that can reveal patterns and connections between the documents which are otherwise difficult to spot. In the remainder of this paper, we elaborate on future research directions aimed to improve our visualization approach further.

## *7.1 Topic Models*

Our current approach relies on the standard VSM. This framework is rather simplistic, and is limited in its ability to capture meaningful relationships between documents. Essentially, the information that can be provided is based only on lexical overlap. Furthermore, induction of topics in the collection is indirect, and relies on the user's interpretation of the provided clustering. A significant improvement of this approach might be achieved using more sophisticated forms of document representation. To this aim we intend to use Topic Models [2, 3].

Topic models are generative probabilistic models, which allow a collection of documents to be structured according to topics. The topics are probability distributions over lexical items. As documents are generated by drawing topics and subsequently topic words, one may discover a set of topics for a document collection using posterior statistical inference, relying on the words appearing in the documents. The topic modeling approach assumes that documents are mixtures of topics. This assumption is appealing in general, and particularly suited for our data, in which documents clearly contain multiple topics scattered throughout the text.

Topic models may enhance the presented IR and visualization approach in several manners. First, they provide the possibility to present the prevalent topics in the collection by listing their most probable words. Furthermore, each document can be displayed according to its topics. Topics can also be presented chronologically, allowing content evolution within the collection over time to be tracked.

More importantly, topic models provide an alternative and more sophisticated way of measuring similarities between documents. Instead of a simple comparison of terms, the similarity measurement reflects similarity of topics, and documents might be similar even if they do not share much vocabulary. Furthermore, one may allow the user to modularize the similarity measurements according to particular topics. This approach can produce better recommendations for the most similar documents and more transparent and informative clustering of the documents. Similarly to

our approach with the VSM, the vocabulary of topic models may be restricted to or combined with domain important terms such as NEs.

## 7.2 Clustering and Layouting

Various alternative algorithms could be used for clustering. In particular, Spectral Clustering algorithms [16] are among the most used. They have been proven to achieve results competitive with other clustering algorithms [25]. The advantage of these algorithms is that they impose rather weak assumptions on the shape of the clusters. Nevertheless, the running time of these algorithms is longer than the running time of the CWC algorithm and their implementation is much more involved. With traditional spectral clustering algorithms, the number of clusters to be constructed must be specified in advance. However, newer Spectral Clustering algorithms determine the number of clusters automatically [21].

With regard to layouting, other approaches may yield representations that better reflect the inter-document similarities. A particularly promising possibility is Spectral Graph Layouting [19]. This approach is based on the computation of eigenvalues of a Graph Laplacian. Its advantage over FDP algorithms is that a globally optimal layout according to the model specifications can be computed efficiently [12], while FDP algorithms do not guarantee convergence to a global optimum.

In the current graph representation, the similarity graph often becomes illegible when a large amount of nodes is presented. This issue can be alleviated using a hierarchical collection representation based on the clustering output: on the top level, a single node would represent each discovered cluster. The edges between the nodes/clusters would be constructed according to the edge density and the thicknesses of the edges connecting the nodes in different clusters. Detailed information can then be obtained by expanding selected clusters such that each document would have its own node. This technique can improve the scalability of the visual representation.

## 7.3 Evaluation

We have currently carried out a relatively small scale and mostly qualitative evaluation of the usefulness of our visualization model. Designing a valid evaluation scheme in our setup is very challenging: The notions of similarity or relatedness are difficult to translate into clear evaluation schemes due to the fact that our objects of analysis are full-length documents, and in particular are hightly rhetorical speeches which encompass a multitude of topics. Nevertheless, we plan to evaluate our setup in a more systematic manner in order to find out whether the similarity measurements according to NEs and general lexicon indeed correlate with human judgments of similarity and relatedness.

As our approach is user-oriented, one of the central aspects of our future evaluation scheme will be user evaluation and feedback. We plan to introduce our GUI tool to students and researchers in university history departments, apply our tool on historical text databases they use and receive feedback on their experiences with the system.

## *7.4 Adaptation to Other Domains*

It is important to stress that even though our system is dedicated to the historical domain, it can be relatively easily adapted to other domains. This can be done by selecting appropriate features that are specifically important for expressing inter-document similarity within these domains. The visualization method can be considered to be domain independent, as it represents objects according to similarity measurements regardless of the way these measurements were obtained.

Moreover, it is possible to think of other ways of determining domain important information. For example, one could utilize domain specific lexicons, and consider document terms that appear in such lexicons to be more important for the similarity measurements. One possible domain of application is law documents, where a lexicon of legal terms could be used for obtaining inter-document similarities.

Enhancement of our system in the above proposed directions and its application to additional types of document collections can further establish the relevance of our approach as a powerful alternative to standard information retrieval systems.

**Acknowledgements** We would like to thank the following people at Saarland University:

- Caroline Sporleder, for her dedicated guidance and valuable advice on the project.
- Martin Schreiber, for his feedback on the system from the user perspective.

Michal Richter has been supported by grant ME838 of the Czech Republic Ministry of Education, Youth and Sport.

# References

1. Biemann, C.: Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In: TextGraphs '06: Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing, pp. 73–80. Association for Computational Linguistics, Morristown, NJ, USA (2006)
2. Blei, D., Lafferty, J.: Topic models. Text mining: classification, clustering, and applications pp. 71–93 (2009)
3. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research **3**, 993–1022 (2003)
4. Börner, K., Chen, C., Boyack, K.: Visualizing knowledge domains. Annual review of information science and technology **37**(1), 179–255 (2003)
5. Dunne, C., Shneiderman, B., Dorr, B., Klavans, J.: iOPENER workbench: Tools for rapid understanding of scientific literature. In: Human-Computer Interaction Lab 27th Annual Symposium (2010)
6. Eades, P.: Graph drawing methods. In: P. Eklund, G. Ellis, G. Mann (eds.) Conceptual Structures: Knowledge Representation as Interlingua, *Lecture Notes in Computer Science*, vol. 1115, pp. 40–49. Springer Berlin / Heidelberg (1996)
7. Finkel, J., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics, pp. 363–370. Association for Computational Linguistics, Ann Arbor, MI, USA (2005)
8. Fruchterman, T.M.J., Reingold, E.M.: Graph drawing by force-directed placement (1991)
9. Greaves, M.: The growing semantic web. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases: Part I, ECML PKDD '09, p. 3. Springer, Berlin and Heidelberg, Germany (2009)

10. Hearst, M.A.: Search User Interfaces, chap. Information Visualization for Search Interfaces. Cambridge University Press, Cambridge, England (2009)
11. Java Universal Network/Graph framework (JUNG). Retrieved 19 Nov 2010, from http://jung.sourceforge.net/
12. Koren, Y.: On spectral graph drawing. In: Proc. 9th Inter. Computing and Combinatorics Conference (COCOON'03), LNCS 2697, pp. 496–508. Springer-Verlag (2002)
13. Castro Speech Database. Retrieved 17 Nov 2010, from http://lanic.utexas.edu/la/cb/cuba/castro.html
14. Leslie, C., Kuang, R.: Fast string kernels using inexact matching for protein sequences. The Journal of Machine Learning Research **5**, 1435–1455 (2004)
15. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. The Journal of Machine Learning Research **2**, 419–444 (2002)
16. Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**, 395–416 (2007)
17. Manning, C.D., Raghavan, P., Schütze, H.: An introduction to information retrieval. Cambridge University Press, Cambridge, England (2008)
18. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. In: D. Nadeau, S. Sekine (eds.) Named Entities: Recognition, classification and use. John Benjamins, Amsterdam, the Netherlands and New York, NY, USA (2009)
19. Puppe, T.: Spectral Graph Drawing: A Survey. VDM Verlag (2008)
20. Salton, G., Wong, A., Yang, C.: A vector space model for automatic indexing. Communications of the ACM **18**(11), 613–620 (1975)
21. Sanguinetti, G., Laidler, J., Lawrence, N.D.: Automatic determination of the number of clusters using spectral algorithms.in. In: IEEE Machine Learning for Signal Processing. 28-30 Sept 2005, pp. 28–30 (2005)
22. Scholkopf, B.: The kernel trick for distances. Advances in Neural Information Processing Systems **13**, 301–307 (2001)
23. Schvaneveldt, R.: Pathfinder associative networks: studies in knowledge organization. Ablex Series In Computational Science (1990)
24. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge University Press, Cambridge, England (2004)
25. Verma, D., Meila, M.: A comparison of spectral clustering algorithms. Tech. rep., Department of CSE, University of Washington (2003)
26. Versley, Y., Moschitti, A., Poesio, M., Yang, X.: Coreference systems based on kernels methods. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Computational Linguistics, Manchester, England (2008)
27. Versley, Y., Ponzetto, S., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., Moschitti, A.: BART: A modular toolkit for coreference resolution. In: Proceedings of the 6th International Conference on Language Resources and Evaluation. LREC, Marrakech, Morocco (2008)
28. Visuwords: online graphical dictionary. Retrieved 21 Oct 2010, from http://www.visuwords.com/

# Integrating Wiki Systems, Natural Language Processing, and Semantic Technologies for Cultural Heritage Data Management

René Witte, Thomas Kappler, Ralf Krestel, and Peter C. Lockemann

**Abstract** Modern documents can easily be structured and augmented to have the characteristics of a semantic knowledge base. Many older documents may also hold a trove of knowledge that would deserve to be organized as such a knowledge base. In this chapter, we show that modern semantic technologies offer the means to make these heritage documents accessible by transforming them into a semantic knowledge base. Using techniques from natural language processing and Semantic Computing, we automatically populate an ontology. Additionally, all content is made accessible in a user-friendly Wiki interface, combining original text with NLP-derived metadata and adding annotation capabilities for collaborative use. All these functions are combined into a single, cohesive system architecture that addresses the different requirements from end users, software engineering aspects, and knowledge discovery paradigms. The ideas were implemented and tested with a volume from the historic Encyclopedia of Architecture and a number of different user groups.

**Key words:** semantic wikis, ontology population, ontology queries, automatic summarization, index generation, web services, Handbuch der Architektur

## 1 Introduction

Modern documents can be turned into veritable knowledge bases by linking the text or parts thereof to a multitude of supportive data such as explication of semantics, user notes, discussion panels, background information, current news, quotes and citations, illustrative material, or more detailed presentations. Older documents such as books dealing with our cultural heritage often deserve the same kind of support. However, they exist only in analog form, and one has to search for and find related material by inspecting an often huge number of textual sources. If such documents are to

René Witte
Concordia University, Montréal, Canada, e-mail: witte@semanticsoftware.info

Thomas Kappler
Swiss Institute of Bioinformatics, Geneva, Switzerland, e-mail: tkappler@googlemail.com

Ralf Krestel
L3S Research Center, Hannover, Germany, e-mail: krestel@l3s.de

Peter C. Lockemann
Karlsruhe Institute of Technology, Germany, e-mail: Lockemann@kit.edu

C. Sporleder et al. (eds.), *Language Technology for Cultural Heritage: Selected Papers from the LaTeCH Workshop Series*, Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-20227-8_12, 

be repeatedly accessed by a group of peers it might be worthwhile to structure them along the lines of modern documents.

Since doing so by hand is a cumbersome and lengthy affair, one should find ways to build an initial structure by automatically extracting relevant information from the analog document. Our thesis is that extraction should be based on an understanding of the semantics contained in the document with its text, tables, figures, etc. We studied the issue in the context of a project for developing enhanced semantic support for users of textual cultural heritage data, more specifically on the historic Encyclopedia of Architecture, written in German between 1880–1943. Our aim was to apply modern semantic technologies to make these heritage documents more flexibly accessible by transforming them into a semantic knowledge base. More specifically, by using techniques from natural language processing and Semantic Computing, we automatically populate an ontology that allows building historians to navigate and query the encyclopedia, while architects can directly integrate it into contemporary construction tools. Additionally, all content is made accessible in a user-friendly Wiki interface, combining original text with NLP-derived metadata and adding annotation capabilities for collaborative use.

A particular result of our approach is the integration of different concerns into a single, cohesive system architecture that addresses requirements from end users, software engineering aspects, and knowledge discovery paradigms. The ideas were implemented and tested with one volume of the historic encyclopedia of architecture and a number of different user groups, including building historians, architects, and NLP system developers.

We discuss the user groups and their requirements in Sect. 2, examine in Sect. 3 the related work, and then develop our solution in Sect. 4. Section 5 closes the chapter.

## 2 User Groups and Requirements

Nowadays, the baseline for cultural heritage data management of book-type publications is the production of a scanned (digitized) version that can be viewed and distributed online, typically with some kind of Web interface. Before we can deliver more advanced access methods, we have to be more precise about the targeted end users. Who needs access to heritage data, and for what purpose?

### *2.1 User Groups*

Within our approach, we consider the requirements from four different user groups; each of them having a different background and expectations concerning the management of historical textual data.

**(1) Historians:** Within this group, we target users that deal with historical material from a scientific motivation, namely, historians. They require an electronic presentation that provides for a direct mapping to the printed original, e.g., for citation purposes. Additionally, semantic analysis tools should support their work through the formulation and verification of hypotheses.

**(2) Practitioners:** Under this group, we are concerned with users that need access to the historical material for their contemporary work. In our example scenario, the handbook on architecture, these are today's architects that need information on the building processes and materials used, e.g., within a restoration project of an old building. Here, the historical material contains knowledge that is not readily accessible in modern sources. Another example for such a user group are musicians dealing with old music scores and their descriptions, or lexicographers analyzing documents for the development of dictionary entries.

**(3) Laypersons:** Historical materials are a fascinating source of knowledge, as they preserve information over centuries. Providing widespread online access to materials that are otherwise only available in a controlled environment to scientists due to their fragile nature is perhaps one of the greatest benefits of digitization projects.

**(4) Computational Linguists:** Similarly to practitioners, linguists are often interested in historical documents from a functional point of view. However, their domain focuses on the properties of the language and its development over time rather than the underlying domain of discourse. They also have particular requirements for corpus construction, access, and annotation to support automated NLP analysis workflows.

## *2.2 Detected Requirements*

We can now derive a number of explicit requirements a system needs to fulfill, based on the user groups defined above:

**Web Interface.** To make the historical data available over the Internet, and to provide easy access within a familiar metaphor, the system needs to support a Web interface. This concerns all user groups to various degrees, but in particular the historians and laypersons.

**Annotation Support.** Users working with the historical data from a scientific point of view—in particular group (1)—often need to comment, add, and collaborate on the historical data. This should be supported within the same interface as the primary (historical) data, to avoid unnecessary context and application switches for the end users. At the same time, these annotations must be maintained by the architecture on clearly separated layers, to keep the integrity of the historical data intact.

**Corpus Generation.** While a Web interface is helpful for a human user, automated analyses using NLP tools and frameworks (user group (4)) can be better supported with a corpus in a standard (XML-based) markup, since HTML pages generated through Web frameworks typically mix content and layout information (menus, navigation bars, etc.). Thus, the architecture should provide a separate corpus that is automatically derived from the historical data and contains appropriate markup (for headlines, footnotes, figure captions, etc.). Ideally, it should allow to cross-link entities with the Web interface.

**NLP Services.** For large collections of (historical) documents, manual inspection of all content or even a subset obtained through information retrieval (IR) is not feasible. Here, NLP analyses can deliver additional benefit to end users, in particular groups (1)–(3), by integrating NLP analysis services (and their results) into the overall architecture. It should allow the execution of any service, developed by user group (4), and also deliver the results back to the clients. Examples for such NLP services are summarization, index generation, or named entity detection.

**Metadata Generation.** While NLP results can be useful for a human user, we also need to support further automated analysis workflows. User group (2) in particular requires access to the historical data, as well as its metadata, from external tools and applications relevant for their domain. To support external access to metadata from many different clients, the architecture should be capable of generating standards-compliant data formats, such as RDF (open linked data) and OWL (Semantic Web).

**Application Integration.** As pointed out in the last requirement, external applications should be provided with automated access to the historical data and its metadata. Generally speaking, this requires the introduction of a client/server model, where the communication, like the metadata format, should use open, established standards.

## 3 Related Work

Before we describe our approach in detail, we discuss related work relevant for the detected requirements.

The Cultural Heritage Language Technologies (CHLT) project [13,14] describes the use of NLP methods to help students and scholars to work with classic Greek and Latin corpora. Similar to our approach, collaboration is an important goal of the project. Not only for sharing metadata about the text itself, but also to offer users the possibility to annotate, comment, or correct the results of automated analyses. This metadata can also contain hyperlinks to connect related texts with each other. The importance of correct morphological analysis is stressed as a baseline technology for users in the humanities, a statement which is also reflected in our work by integrating a self-learning lemmatizer for the German language [12] for accurate index generation. Further processing in the CHLT project includes information retrieval and data visualization. Identifying keywords, clustering subsets of the data, and visualizing the resulting groups supports the users in grasping concepts or performing search. In contrast, our approach uses open, standardized data formats like an automatically populated ontology to facilitate searching and browsing through the corpus and a Wiki system to share information between users.

As outlined by Mavrikas et al. [10], access to cultural heritage data available in natural language can be facilitated using various NLP techniques. In the context of the Semantic Web, the proposed system extracts cultural heritage data from different sources in the Internet and processes the data afterwards. An ontology [3] is used to organize the mined data. Templates are used to extract relevant information, and the use of multi-document summarization is also proposed, as a way to present relevant information in a condensed way to the user. Here, we present an actual implementation of a system addressing these problems and extend the use of ontologies to allow easy browsing and querying of the document content for different user groups. Fujisawa [4] proposes to facilitate the access to images of cultural heritage by extracting metadata from the accompanying natural language text. Paraphrasing of the descriptions and the metadata based on the knowledge and experience of the user is proposed as a second step.

Another approach based on the CIDOC-CRM[1] ontology is presented by Généreux [5]. The system described there consists of two parts, one for extracting cultural heritage knowledge from natural language texts and saving the information in the ontology format, and one for using natural language to query the database. The natural language is reformatted to a SPARQL query using WordNet. This approach, in contrast to our system, stresses more the search aspect to find relevant data and offers no further possibilities for collaboration or processing of the data.

Sinclair et al. [17] present a system that enables the user to explore, navigate, link, and annotate digitized cultural heritage artifacts, such as videos, photos, or documents. The system also supports user-generated descriptions and content. The focus in this project lies on the integration of the different metadata formats of the source content, whereas we additionally focus on the processing and collaboration part.

From a technical perspective, semantic extensions to Wiki systems, based on *Semantic Web* technologies, such as OWL ontologies and RDF, are similar in that they provide the means for content structuring beyond the syntactical level. In these systems, the properties of and relations between objects can be made explicit, with the Wiki system "knowing" about them. This allows for automated processing of Wiki content, e.g., through software agents. Current implementations of these ideas can be found in systems such as Semantic MediaWiki (SMW) [7] or IkeWiki [15]. It is important to note that these tools are such as different from and complementary to our approach: While in our context, the content of a Wiki is subject to semantic analysis via NLP methods (with the Wiki engine itself not needing to have semantic capabilities), semantic Wikis like SMW have explicit notational and internal semantic capabilities. The next version of our Wiki/NLP integration architecture, currently under development, will support the SMW extension for storing results of NLP pipelines.

[1] CIDOC Conceptual Reference Model, http://www.cidoc-crm.org/

# 4 Semantic Heritage Data Management

In this section, we present our approach to cultural heritage data management, which integrates a number of different technologies in order to satisfy the requirements of the various user groups: (i) A Wiki user interface, (ii) text mining support using an NLP framework, (iii) Semantic Web ontologies based on OWL and RDF for metadata management, and (iv) W3C Web Services for application integration. We first present an overview of our system in the next subsection. The various subsystems are illustrated using examples from a productive, freely accessible[2] Web resource built around the German *Handbuch der Architektur* (handbook on architecture) from the $19^{th}$ century, described in detail in Sect. 4.2. The digitization process is described in Sect. 4.3. Necessary format conversions for the digital version are covered in Sect. 4.4. To support our user groups, we integrated several NLP analysis services, which are covered in Sect. 4.5. Finally, our semantic extensions for generating OWL/RDF metadata and application integration are covered in Sect. 4.6.

## *4.1 Architectural Overview*

As stated above, our goal is the development of a unified architecture that fulfills the requirements (Sect. 2.2) of the different user groups defined in Sect. 2.1, by integrating means for content access, analysis, and annotation.

One of the central pieces of our architecture is the introduction of a *Wiki* system [8]. Wiki systems provide the Web interface stipulated in our first requirement, while also allowing users to add meta-content in form of separate *discussion* or *annotation* pages. This capability directly addresses our second requirement, by allowing users to discuss and collaborate on heritage data, using an online tool and a single interface, while keeping the original data intact.[3]

Other clients, NLP services, and the actual content have to be integrated into this model. Fig. 1 shows how these and the remaining components are systematically assembled to form the overall *Semantic Assistants* architecture of our system [23].

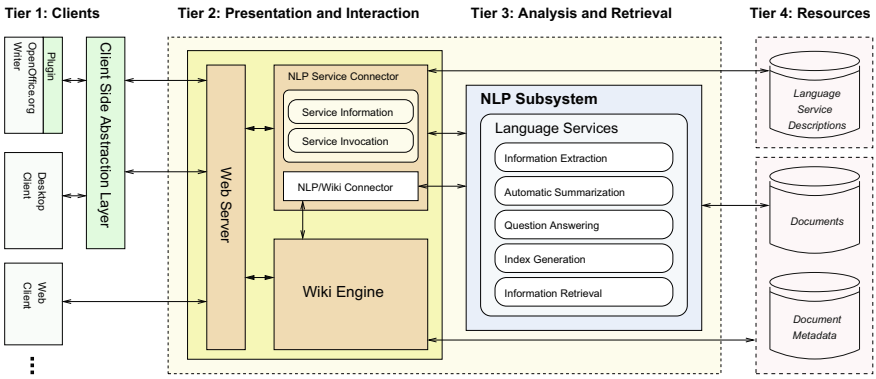


**Fig. 1** System architecture overview

[2] See http://durm.semanticsoftware.info

[3] Assuming the Wiki has been properly configured for this scenario; the technical details depend on the concrete Wiki system.

The architecture comprises four tiers. Tier 1 consists of clients that the users employ to access the system. Plug-in capable existing clients, such as the OpenOffice.org application suite, can also be extended to be integrated with our architecture [6]. New applications can have that functionality built in, such as the "Desktop Client" depicted in the diagram. The "Client-Side Abstraction Layer" (CSAL) facilitates connecting clients by providing common communication and data conversion functionality.

The clients communicate with a Web server on Tier 2, behind which we find the Wiki engine and a software module labeled "NLP Service Connector." The functionality of this module is offered as a SOAP Web service, as standardized by the W3C.[4] This means that there is a publicly accessible interface definition, written in the Web Service Description Language (WSDL), from which clients know how to use the offered functionality. The functionality itself is used through a Web service endpoint, to which the client sends and from where it receives messages. The main task of the NLP Service Connector is to receive input documents and have the NLP subsystem (Tier 3) perform various text analysis procedures on them. A sub-module of the NLP Service Connector, labeled "NLP/Wiki Connector," allows for the automatic retrieval, creation, and modification of Wiki content [22].

Finally, on Tier 4, we have metadata on the employed text analysis services (top), which the NLP Service Connector requires in order to operate these services. The bottom rectangle contains the documents maintained by the Wiki system as well as their metadata, which might have been provided by hand, or generated through automatic analysis methods. The latest version of the architecture, as well as some example NLP services, is available as open source software.[5]

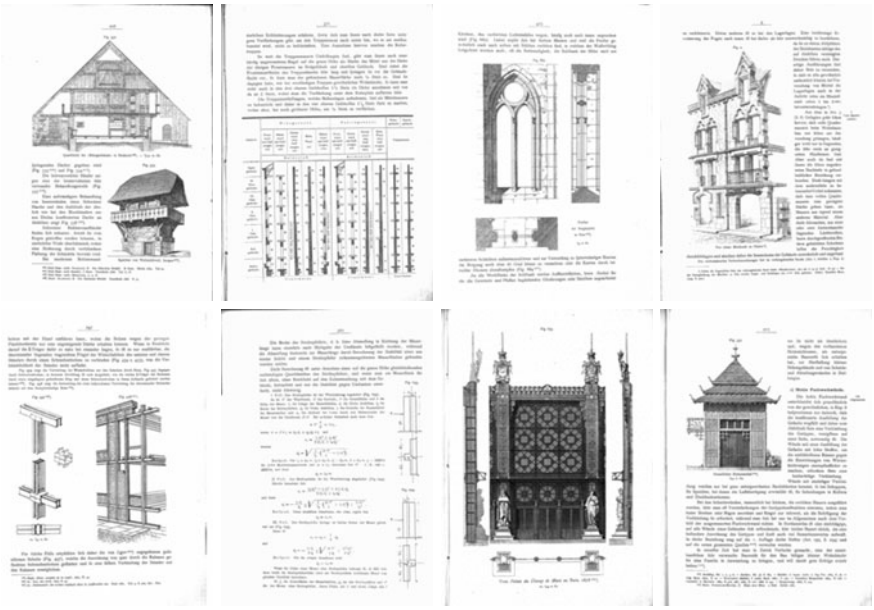

**Fig. 2** Source material examples: Scanned pages from *Handbuch der Architektur* (1900)

[4] Web Services Architecture, http://www.w3.org/TR/ws-arch/

[5] Semantic Assistants, http://www.semanticsoftware.info/semantic-assistants-project

## 4.2 Source Material

We implemented and evaluated the ideas described here for a particular set of historical documents: the German *Handbuch der Architektur*, a comprehensive multi-volume encyclopedia of architecture.[6] The full encyclopedia was written between the late $19^{th}$ and early $20^{th}$ century. It aimed to include all architectural knowledge at the time, both past and present, within the fields of architectural history, architectural styles, construction, statics, building equipment, physics, design, building conception, and town planning. The full encyclopedia comprises more than 140 individual publications and contains at least 25 000 pages.

Due to the ambitious scope, the long publication process, and the limitations of the technologies available at that time, it is extremely difficult to gain an overview of a single topic. Information is typically distributed over several parts containing a number of volumes, which in turn are split into books. Most of these do not contain any kind of index. In addition, some of the volumes were edited and reprinted and a supplement part was added.

Due to funding limitations, we only dealt with a single volume[7] within the project described in this chapter. However, the concepts and technologies have been designed with the complete dataset in mind.

## 4.3 Digitization and Error Correction

The source material was first digitized using specialized book scanners, producing a TIFF file for each physical page; in our case, with a grayscale resolution of 600dpi.

In a second step, the image files needed to be converted to machine-readable text to support, amongst others, NLP analysis and metadata generation. We initially planned to automate this process using OCR software. However, due to the complex layout of the original material (see Fig. 2), which contains an abundance of figures, graphs, photos, tables, diagrams, formulas, sketches, footnotes, margin notes, and mixed font sizes, as well as the varying quality of the 100-year old source material, this proved to be too unreliable. As the focus of this project was on developing enhanced semantic support for end users, not basic OCR research, we decided to manually convert the source material into an electronic document. This provided for not only a faster and more reliable conversion, but also accurately captured layout formation in explicit markup, such as footnotes, chapter titles, figure captions, and margin notes. This task was outsourced to a Chinese company for cost reasons; Manual conversion was performed twice to allow an automatic cross-check for error detection. The final, merged version contained only a very small amount of errors, which were eventually hand-corrected during the project. It is freely available online under an open content license.[8]

---

[6] Edited by Joseph Durm (⋆14.2.1837 Karlsruhe, Germany, †3.4.1919 ibidem) and three other architects since 1881.

[7] E. Marx: *Wände und Wandöffnungen* (Walls and Wall Openings). In "Handbuch der Architektur," Part III, Volume 2, Number I, Second edition, Stuttgart, Germany, 1900. Contains 506 pages with 956 figures.

[8] Durm Corpus, http://www.semanticsoftware.info/durm-corpus

## 4.4 Format Transformation and Wiki Upload

The digitized content was delivered in the *TUSTEP*[9] format. This content was first converted to XML, and finally to Wiki markup. In the following, we briefly describe the conversion process.

### 4.4.1 TUSTEP Format

TUSTEP is a toolkit for the "scientific work with textual data" [18], consisting of a document markup standard along with tools for text processing operations on TUSTEP documents. The markup is completely focused on layout, so that the visual structure of printed documents can be captured well. Structurally, it consists both of XML-like elements with an opening and closing tag, such as `<Z>` and `</Z>` for centered passages; and elements serving as control statements, such as `#H:` for starting text in superscript. The control statements remain in effect until another markup element cancels them out, such as `#G:` for adjusting the following text on the baseline.

TUSTEP predates XML, and while it is still in use at many universities, we found it makes automatic processing difficult. The control statements, for instance, make it hard to determine the range of text they affect, because their effect can be canceled by different elements. In addition, in the manual digitization process, markup was applied inconsistently. Therefore, we chose to first convert the data to a custom XML format, designed to closely match the given TUSTEP markup. This also enabled easier structural analysis and transformation of the text due to the uniform tree structure of XML and the availability of high-quality libraries for XML processing.

### 4.4.2 Custom XML

We developed a custom tool to transform TUSTEP data into XML. The generated XML data is intended to be as semantically close to the original markup as possible; as such, it contains mostly layout information such as line and page breaks and font changes. Except for the exact placement of figures and tables, all such information from the original book is retained.

Parsing the XML into a DOM[10] representation provides for easy and flexible data transformation, e.g., changing an element node of the document tree such as `<page no="12">` to a text node containing the appropriate Wiki markup in the next step. The resulting XML format can be directly used for NLP corpus generation, which is then loaded into an NLP framework, such as GATE [2]. This XML corpus is also freely available online.[11]

### 4.4.3 Wiki Markup

To make the historical data accessible via a Wiki, we have to further transform it into the data format used by a concrete Wiki engine. Since we were dealing with an encyclopedic original, we chose the *MediaWiki*[12] system, which is best known for its use within the *Wikipedia*[13] projects. MediaWiki stores the textual content in a MySQL database, the image files are stored as plain files

[9] TUebingen System of TExt processing Programs (TUSTEP), http://www.zdv.uni-tuebingen.de/tustep/tustep_eng.html

[10] Document Object Model (DOM), http://www.w3.org/DOM/

[11] Durm Corpus, http://www.semanticsoftware.info/durm-corpus

[12] MediaWiki, http://en.wikipedia.org/wiki/MediaWiki

[13] Wikipedia, http://www.wikipedia.org


**Fig. 3** The Wiki interface integrating digitized text, scanned originals, and separate "Discussion" pages

on the server. It provides a PHP-based dynamic web interface for browsing, searching, and manual editing of the content.

A challenging question was how to perform the concrete conversion from content presented in physical book layout to Wiki pages. Obviously, translating a single book page does not translate well into a single web page. We first attempted to translate each book chapter into a single page (with its topic as the Wiki entry). However, with only 15 chapters in a 500-page book, the resulting Web pages were too long to be used comfortably in the MediaWiki interface. Together with our end users, we finally decided to convert each sub-chapter (section) into a single Wiki page, with additional internal structuring derived from the margin notes preserved by the manual conversion.

MediaWiki uses the markup language *Wikitext*, which was designed as a "simplified alternative to HTML,"[14] and as such offers both semantic markup, like headings with different levels, as well as visual markup, like italic or bold text. Its expressiveness is largely equal to that of HTML, despite the simplified approach, because it lets users insert HTML if Wikitext does not suffice.

**Example: Footnote conversion.** Footnotes were delivered in TUSTEP in the form `#H:n#G:`) for each footnote *n*. The markup indicates text being set to superscript (`#H:`), then back to the standard baseline (`#G:`). The footnote reference in the text and the anchor in the footnote section of a page have the same markup, as they look the same. The tool converting to XML locates footnotes using a regular expression, and creates `<footnote to="n" />` resp. `<footnote from="n">...</footnote>` tags. Finally, the conversion to Wikitext trans-

[14] Wikitext, http://en.wikipedia.org/wiki/Wikitext

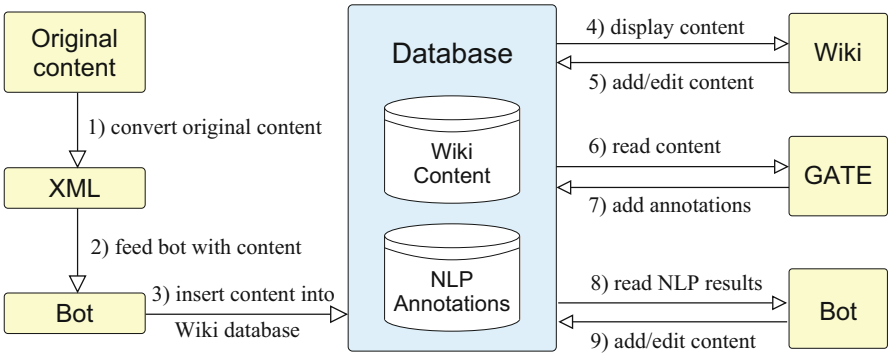


**Fig. 4** Workflow between document storage, retrieval, and NLP analysis

forms the references to the format `<span id="fn8ref" /><sup>[[#fn8|8)]]</sup>`. The HTML "sup" tag sets the text as superscript, and its content is a link to the anchor "fn8" on the same page, with the link text simply being "8". The footnote itself is represented by `<span id="fn8"/>''8)'' ... [[#fn8ref|^]]`. We see the anchor linked to from the reference, and vice versa a link to jump back upwards to the reference.

### 4.4.4 Wiki Interface Features

The conversion to Wikitext inserts further information for the Wiki users, such as links to scans of the original pages, and link/anchor combinations to emulate the page-based navigation of the book (see Fig. 3). For instance, the beginning of page 211, which is indicated in TUSTEP by `@@1@<S211><`, looks as follows in the resulting Wikitext:

```
<span id="page10" />
'''Seite 211 ([[Media:S211_large.gif|Scan]])'''
[[Image:S211_large.gif|thumb|200px|Scan der Originalseite 211]]
```

### 4.4.5 Wiki Data Upload

The workflow between the Wiki and the NLP subsystems is shown in Fig. 4. The individual subcomponents are loosely coupled through XML-based data exchange. Basically, three steps are necessary to populate the Wiki with both the encyclopedia text and the additional data generated by the NLP subsystem. Firstly (Step 1 in Fig. 4), the original Tustep markup of the digitized version of the encyclopedia is converted to XML as described above. In Step 2, the XML data is converted to the text markup used by MediaWiki. And finally (Step 3), the created Wiki markup is added to the MediaWiki system using parts of the Python Wikipedia Robot Framework,[15] a library offering routines for tasks such as adding, deleting, and modifying pages of a Wiki or changing the time stamps of pages.

[15] Python Wikipedia Robot Framework, http://pywikipediabot.sf.net

## 4.5 Integrating Natural Language Processing

One of the main goals of our work is to support the end users—groups (1) to (3)—with semantic analysis tools based on NLP. To make our architecture independent from the application domain (architecture, biology, music, ...) and their custom NLP analysis pipelines, we developed a general integration framework that allows us to deploy any kind of language service. The management, parameterization, and execution of these NLP services is handled in our framework (see Fig. 1, Tier 3, "NLP Subsystem") by GATE, the *General Architecture for Text Engineering* [2]. To allow a dynamic discovery of newly deployed language services, we added service descriptions written in OWL to our architecture (see Section 4.1).

Language services should help the users to find, understand, relate, share, and analyze the stored historical documents. In the following subsections, we describe some of the services we deployed in our implementation to support users of the historic encyclopedia, including index generation, automatic summarization, and ontology population.

### 4.5.1 Index Generation

Many documents, similarly to the architectural encyclopedia under discussion, do not come with a classical back-of-the-book index. Of course, in the absence of an index, full-text search can help to locate the various occurrences of a single term, but only if the user already knows what he is looking for. An index listing all nouns with their modifiers (adjectives), with links to their locations of occurrence, can help a user finding useful information he was not expecting, which is especially important for historical documents, which often contain terminology no longer in use.

For our automatic index generation, shown in Fig. 5, we first determine the part-of-speech for each word (noun, verb, adjective, etc.) using the TreeTagger [16].[16] Based on this information, the open source chunker MuNPEx[17] groups words into *noun phrases* (NPs), which consist of a head noun, a (possibly empty) list of adjectives, and an optional determiner. For each noun phrase, we compute the lemma of the head noun and keep track of its modifiers, page number, and corresponding Wiki page. To deal with the problem of correctly lemmatizing historical terminology no longer in use, we developed a self-learning lemmatizer for German [12], which is freely available online.[18] Nouns that have the same lemma are merged together with all their information. Then, we create an inverted index with the lemma as the main column and its modifiers as sub-indices, as shown in Fig. 5. The generated index is then uploaded from the NLP subsystem into the Wiki through a connector ("NLP/Wiki Connector" in Fig. 1).

### 4.5.2 Automatic Summarization

Large text corpora make it impossible for single users to deal with the whole document set. The sheer amount of information encoded in natural language in huge text collections poses a non-trivial challenge to information systems in order to adequately support the user. To find certain information, to get an overview of a document, or just to browse a text collection, automatic *summarization* [9] offers various methods of condensing texts.[19]

---

[16] TreeTagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[17] Multi-Lingual Noun Phrase Extractor (MuNPEx), http://www.semanticsoftware.info/munpex

[18] Durm German Lemmatizer, http://www.semanticsoftware.info/durm-german-lemmatizer

[19] See, e.g., the *Text Analysis Conference* (TAC), http://www.nist.gov/tac

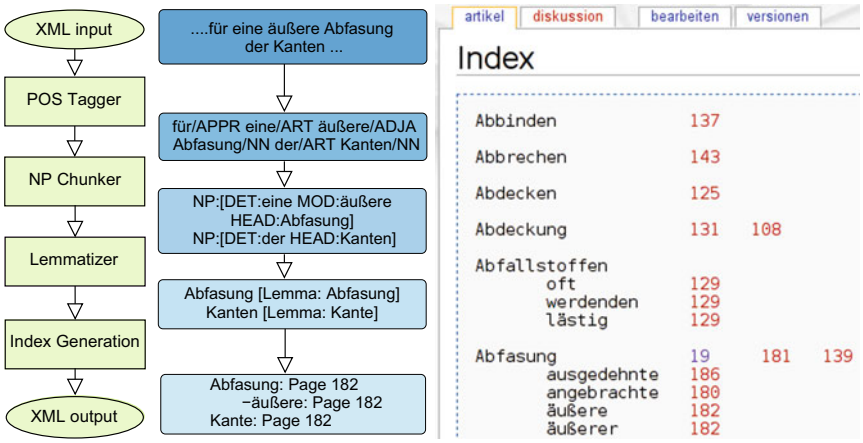


**Fig. 5** NLP-generated full text index, integrated into the Wiki interface (page numbers are hyperlinks to Wiki pages)

| **"Welche Art von Putz bietet Schutz vor Witterung?"** |
|---|
| Ist das Dichten der Fugen für die Erhaltung der Mauerwerke, namentlich an den der Witterung ausgesetzten Stellen, von Wichtigkeit, so ist es nicht minder die Beschaffenheit der Steine selbst. Bei der früher allgemein üblichen Art der gleichzeitigen Ausführung von Verblendung und Hintermauerung war allerdings mannigfach Gelegenheit zur Beschmutzung und Beschädigung der Verblendsteine geboten. Will man einen dauerhaften Putz erzielen, so gilt für alle Arten von Mauerwerk die Regel, daß die zu putzenden Flächen frei von Staub sein müssen, da dieser trennend zwischen Mauer und Putz wirken und das feste Anhaften des letzteren verhindern würde. … |

**Fig. 6** Excerpt from a focused summary generated based on a question (shown on top)

Short, headline-like summaries (around 10 words) that incorporate the most important concepts of a document or a Wiki page facilitate the search for particular information by giving a user an overview of the content at a glance. In addition, full-text summaries can be created for each page, e.g., with a length of 100 words or more. These summaries in free-text form can be read much more quickly than a full-length article, thereby helping a user to decide which Wiki pages he wants to read in full.

More advanced types of summaries can support users during both content creation and analysis. *Multi-document summaries* can combine knowledge from several pages within a Wiki or even across Wiki systems. *Update summaries* keep track of a user's reading history and only present information he has not read before, thereby further reducing the problem of information overload. *Contrastive Summaries* [20] can support a user in highlighting differences across a number of articles (or article versions) on the same topic, thereby showing both commonalities and differences. In our project, we contrasted modern building standards (DIN/SIN) with content from the historic encyclopedia.

*Focused summaries* [19] enable the user to formulate a query (natural language questions) the generated summary focuses on. This is especially useful to get a first impression of the available information about a certain topic in a collection. An example for such a summary is shown in Fig. 6: This summary provides a series of relevant sentences in answer to the user's question, *"Welche Art von Putz bietet Schutz vor Witterung?"* (Which kind of plaster would be suitable to protect brickwork against weather influences?). In [21], we further discuss the usefulness of focused summaries for a particular architectural scenario.

### 4.5.3 Integrating further NLP Web Services

The examples presented so far are by no means exhaustive. Depending on the type of data under investigation and the demands of the users concerned with their analysis (groups (1) and (2)), additional NLP services will need to be introduced. Due to our service-oriented approach (cf. Section 4.1), new services can be added at any time, as they are automatically detected by all connected clients through the metadata repository, without any changes on the client side. Likewise, new user clients can be added dynamically to the architecture, without requiring any changes to the NLP server.

## *4.6 Semantic Extensions*

The NLP analysis services introduced so far are aimed at supporting the user groups (1) and (3): Summaries, full-text indices, and question-answering all produce new natural language texts, which are convenient for humans. But they are less useful for providing further automated access to the historical data, e.g., through desktop tools targeted at user group (2). In our example scenario, the architects need to integrate the historical knowledge "stored" in the encyclopedia within contemporary architectural design tools: While viewing a certain construction element, the relevant content from the handbook should be extracted and presented alongside other project information. This requires the generation of metadata in a machine-processable format. In our architecture, this is provided through the NLP-driven population of formal (OWL-DL) ontologies. We discuss our ontology model in the next subsection, followed by a description of the automatic population process and the querying of the result format.

### 4.6.1 Ontology Model

Our NLP-generated metadata is formally represented using the *Web Ontology Language* (OWL),[20] which is a standard defined by the World Wide Web Consortium (W3C). Specifically, we use the sub-format OWL-DL, which is based on description logics (DL). DLs describe domains in terms of TBox (also known as concepts or classes), roles (also known as relationships or properties) and ABox (also known as individuals or instances). OWL is also the foundation of the Semantic Web initiative, which allows us to immediately make use of a large variety of tools and resources developed for OWL-based information processing (editors, triplestores, query languages, reasoners, visualization tools, etc.).

Our ontology has two parts: a *document* ontology describing the domain of NLP (documents, sentences, NPs, coreference chains, etc.) and a *domain* ontology. While the document ontology is independent of the content in the historical documents, the domain ontology has to be developed specifically for their discourse domain. In our example, this ontology needs to contain architectural concepts, such as doors, walls, or windows. By combining both ontologies, we can run semantic *queries* against the ontology, e.g., asking for all sentences where a certain concept appears.

**Document Ontology Model.** Our document ontology models a number of concepts relevant for the domain of NLP. One of the main concepts is *document*, representing an individual text processed by an NLP pipeline, containing: the *title* of the document; its *source* address (typically a URL or URI); and a relation *containsSentence* between a document and all its *sentences*.

Likewise, sentences are also represented by an ontology class, with: the start and end position (*beginLocation*, *endLocation*) within the document, given as character offset; the sentence's

[20] OWL, http://www.w3.org/2004/OWL/

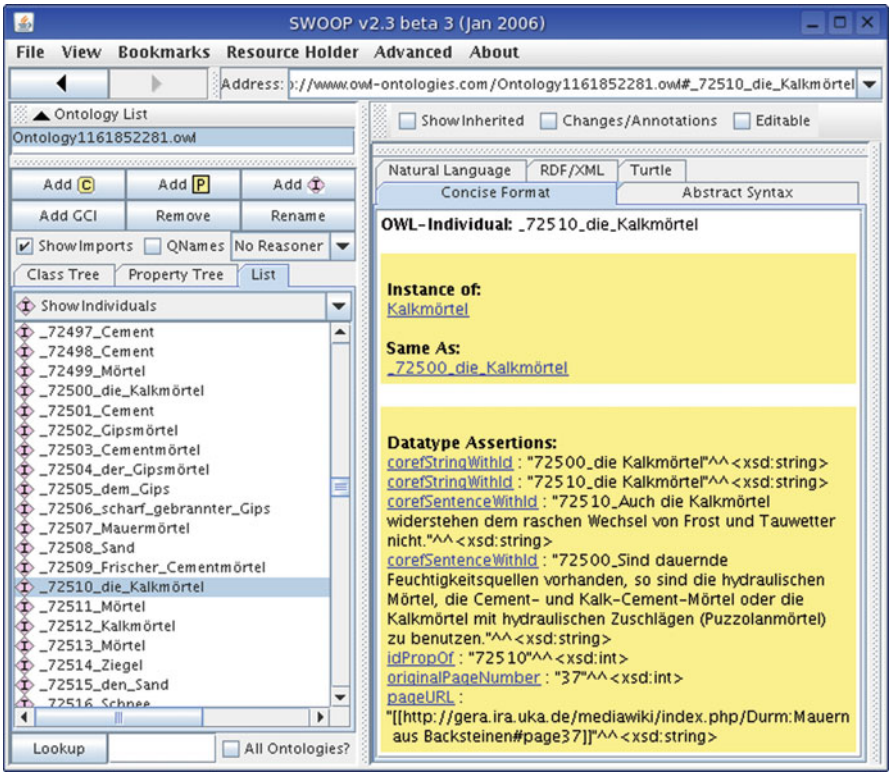


**Fig. 7** An ontology instance created through NLP

*content*, stored as plain text, i.e., without additional markup; and a relation *contains* between a sentence and all *named entities* that have been detected in it.

Each of the named entities has, in addition to its ontology class, a number of additional properties: a unique id (*idPropOf*) generated for this instance; the page number (*originalPageNumber*), where the instance can be found in the (printed) source; and the full URL (*pageURL*) for direct access to the instance in the Wiki system.

Additionally, we can represent the result of the coreference resolution algorithm using the OWL language feature *sameAs:* If two instances appear in the same coreference chain, two separate ontology instances are created (containing different ids and possibly different page/URL numbers), but both instances are included in such a *sameAs* relation. This allows ontology reasoners to interpret the syntactically different instances as semantically equivalent. Additionally, a relation *corefStringWithId* is created for every entity in the coreference chain, refering to its unique id stored in the *idPropOf* property; and the content of the sentence containing the co-refering entity is stored in *corefSentenceWithId*.

**Domain Ontology Model.** In addition to the generic NLP ontology, a domain-specific ontology can be plugged into the system to allow further structuring of the NLP results. If such an ontology is developed, it can also be used to further facilitate named entity detection as described below.

In our approach, we rely on a hand-constructed ontology of the domain. This could be enhanced with (semi-)automatic *ontology enrichment* or *ontology learning*. In general, the design

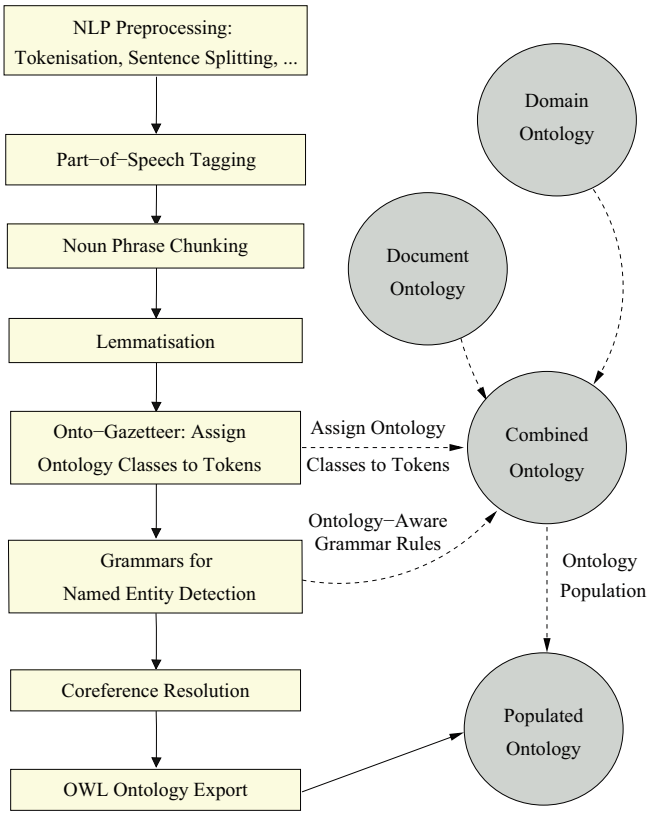


**Fig. 8** NLP pipeline for ontology population

of the domain ontology needs to take the requirements of the downstream applications using the populated ontology into account.

### 4.6.2 Automatic Ontology Population

We developed an *ontology population* NLP pipeline to automatically create OWL instances (individuals, see Fig. 7) for the ontology described above. An overview of the workflow is shown in Fig. 8.

The pipeline runs on the XML-based corpus described in Sect. 4.4. After a number of standard preprocessing steps, including tokenization, POS tagging, and NP chunking, named entities (NEs) are detected using a two-step process. First, an *OntoGazetteer* [1] labels each token in the text with all ontology classes it can belong to. And secondly, ontology-aware grammar rules written in the JAPE[21] language are used to find named entities (NEs). Evaluation of the correctness of the generated instances can be conducted using precision and recall measures [11].

[21] Java Annotations Pattern Engine, a regular expression-based language for writing grammars over document annotation graphs.

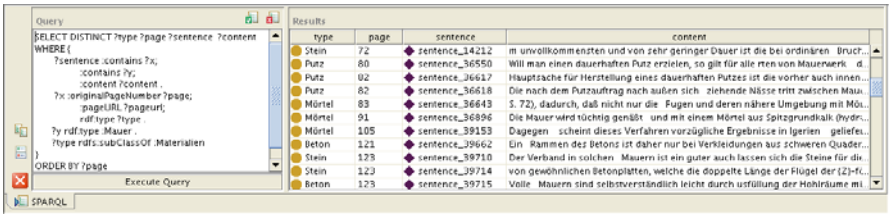


**Fig. 9** Posing a question to the historical knowledge base through a SPARQL query against the NLP-populated ontology

Finally, the created instances are exported into the result ontology, combining a number of domain and document features. An example instance, of the ontology class *Kalkmörtel* (lime mortar), is shown in Fig. 7. This ontology population process is facilitated by an application-independent GATE component, the *OwlExporter* [24], which we made available as open source software.[22]

### 4.6.3 Ontology Queries

The automatically populated ontology represents a machine-readable metadata format that can be *queried* through a number of standardized ontology query languages, such as SPARQL.[23] Queries are a much more expressive paradigm for analyzing text mining results than simple information retrieval (IR); in particular, if a domain model is available, they allow queries over the analyzed documents on a semantic level.

An example SPARQL query is shown in Fig. 9. The query shown in the left box represents the question *"Which building materials are mentioned in the handbook together with the concept 'Mauer' (wall), and on which page?"* The result of this query (executed using Protégé[24]), is shown on the right. The first column ("type") shows what kind of entity (stone, plaster, concrete, ...) was found, i.e., a sub-class of "material" in the domain ontology. The results can now be directly inspected by the user or used for further automatic processing by another application.

More abstractly speaking, ontology queries support automated problem-solving using a knowledge base. A user of our system, like a historian, might want to formulate hypotheses concerning the source material. Translated into an OWL query, the result can be used to confirm or refute the hypothesis. And as a standardized NLP result format, it also facilitates direct integration into an end-user application or a larger automated knowledge discovery workflow.

### 4.6.4 Application Integration

The populated ontology also serves as the basis for our final requirement, application integration. With "application" we mean any end-user accessible system for integrating the historical data within a different context. For example, in a museum setting, such an application might allow a visitor to access content directly relevant to an artifact. A lexicographer might want to query, navigate, and read content from historical documents while developing a lexical entry. And in our application example, an architect needs access to the knowledge stored in the handbook

[22] OwlExporter, http://www.semanticsoftware.info/owlexporter

[23] SPARQL, http://www.w3.org/TR/rdf-sparql-query/

[24] Protégé, http://protege.stanford.edu/

while planning a particular building restoration task. Here, construction elements displayed in a design tool (such as *window* or *window sill*) can be directly connected with the ontological entities contained in the NLP-populated knowledge. This allows an architect to view relevant content down to the level of an individual construction element using the named entities, while retaining the option to visit the full text through the provided Wiki link.

## 5 Summary and Conclusions

The thesis underlying our work was that by understanding the semantics contained in a document one can transform older documents into an initial semantic knowledge base. We demonstrated for an encyclopedia from the cultural heritage domain that this can indeed be done. We developed a methodology for organizing the transformation process, and we identified the necessary tools for implementing the methodology. To support users in the cultural heritage domain, a precise analysis of the different user groups and their particular requirements is essential. The challenge was to find a holistic approach based on a unified system architecture that highlights the many inter-dependencies in supporting different groups with particular features, aimed at different use cases: *Historians* have the support of NLP analysis tools and a user-friendly Web-based access and collaboration tool build around a standard Wiki system. *Laypersons* also benefit from these user-friendly features, while *practitioners*—in our scenario building architects—can additionally use NLP-generated ontology metadata for direct application integration. Finally, our approach also supports computational linguists through corpus construction and querying tools.

The experience from the implemented system using the example of a historical encyclopedia of architecture demonstrates the usefulness of these ideas. Indeed, providing a machine-readable knowledge base that integrates textual instances and domain-specific entities is consistent with the vision of the Semantic Web. The data for the encyclopedia, as well as a number of tools, are publicly accessible under open source licenses.

We believe that our methodology and tools are general enough to be applied to other knowledge domains, and hence have the potential to further enhance knowledge discovery for cultural heritage data.

**Acknowledgements** Praharshana Perera contributed to the automatic index generation and the Durm lemmatizer. Qiangqiang Li contributed to the ontology population pipeline. Thomas Gitzinger contributed to the index generation.

## References

1. Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H.: Evolving GATE to Meet New Challenges in Language Engineering. Natural Language Engineering (2004)
2. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proc. of the 40th Anniversary Meeting of the ACL (2002). http://gate.ac.uk
3. Doerr, M.: The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. AI Mag. **24**(3), 75–92 (2003)
4. Fujisawa, S.: Automatic creation and enhancement of metadata for cultural heritage. Bulletin of the IEEE Technical Committee on Digital Libraries (TCDL) **3**(3) (2007)
5. Généreux, M.: Cultural Heritage Digital Resources: From Extraction to Querying. In: Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pp. 41–48. ACL, Prague, Czech Republic (2007)

6. Gitzinger, T., Witte, R.: Enhancing the OpenOffice.org Word Processor with Natural Language Processing Capabilities. In: Natural Language Processing resources, algorithms and tools for authoring aids. Marrakech, Morocco (2008)
7. Krötzsch, M., Vrandečić, D., Völkel, M.: Semantic MediaWiki. In: I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, L. Aroyo (eds.) The Semantic Web – ISWC 2006, *LNCS*, vol. 4273, pp. 935–942. Springer (2006)
8. Leuf, B., Cunningham, W.: The Wiki Way, Quick Collaboration on the Web. Addison-Wesley (2001)
9. Mani, I.: Automatic Summarization. John Benjamins B.V. (2001)
10. Mavrikas, E.C., Nicoloyannis, N., Kavakli, E.: Cultural Heritage Information on the Semantic Web. In: E. Motta, N. Shadbolt, A. Stutt, N. Gibbins (eds.) EKAW, *Lecture Notes in Computer Science*, vol. 3257, pp. 477–478. Springer (2004)
11. Maynard, D., Peters, W., Li, Y.: Metrics for Evaluation of Ontology-based Information Extraction. In: Proceedings of the 4th International Workshop on Evaluation of Ontologies on the Web (EON 2006). Edingburgh, UK (2006)
12. Perera, P., Witte, R.: A Self-Learning Context-Aware Lemmatizer for German. In: Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), pp. 636–643. Vancouver, BC, Canada (2005)
13. Rydberg-Cox, J.A.: Cultural Heritage Language Technologies: Building an Infrastructure for Collaborative Digital Libraries in the Humanities. Ariadne **34** (2002)
14. Rydberg-Cox, J.A.: The Cultural Heritage Language Technologies Consortium. D-Lib Magazine **11**(5) (2005)
15. Schaffert, S.: IkeWiki: A Semantic Wiki for Collaborative Knowledge Management. In: WETICE, pp. 388–396 (2006). URL http://dblp.uni-trier.de/db/conf/wetice/wetice2006.html#Schaffert06
16. Schmid, H.: Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT-Workshop (1995)
17. Sinclair, P., Lewis, P., Martinez, K., Addis, M., Pillinger, A., Prideaux, D.: eCHASE: Exploiting Cultural Heritage using the Semantic Web. In: 4th International Semantic Web Conference (ISWC 2005). Galway, Ireland (2005)
18. Universität Tübingen – Zentrum für Datenverarbeitung: TUSTEP: Handbuch und Referenz (2008). Version 2008
19. Witte, R., Bergler, S.: Fuzzy Clustering for Topic Analysis and Summarization of Document Collections. In: Z. Kobti, D. Wu (eds.) Proc. of the 20th Canadian Conference on Artificial Intelligence (Canadian A.I. 2007), LNAI 4509, pp. 476–488. Springer, Montréal, Québec, Canada (2007)
20. Witte, R., Bergler, S.: Next-Generation Summarization: Contrastive, Focused, and Update Summaries. In: International Conference on Recent Advances in Natural Language Processing (RANLP 2007). Borovets, Bulgaria (2007). URL http://rene-witte.net/next-generation-summarization
21. Witte, R., Gerlach, P., Joachim, M., Kappler, T., Krestel, R., Perera, P.: Engineering a Semantic Desktop for Building Historians and Architects. In: Proc. of the Semantic Desktop Workshop at the ISWC 2005, *CEUR*, vol. 175, pp. 138–152. Galway, Ireland (2005)
22. Witte, R., Gitzinger, T.: Connecting Wikis and Natural Language Processing Systems. In: WikiSym '07: Proceedings of the 2007 International Symposium on Wikis, pp. 165–176. ACM, New York, NY, USA (2007). DOI http://doi.acm.org/10.1145/1296951.1296969. URL http://rene-witte.net/connecting-wikis-and-nlp
23. Witte, R., Gitzinger, T.: Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients. In: 3rd Asian Semantic Web Conference (ASWC 2008), *LNCS*, vol. 5367, pp. 360–374. Springer, Bangkok, Thailand (2009). URL http://rene-witte.net/semantic-assistants-aswc08
24. Witte, R., Khamis, N., Rilling, J.: Flexible Ontology Population from Text: The OwlExporter. In: The Seventh International Conference on Language Resources and Evaluation (LREC 2010), pp. 3845–3850. ELRA, Valletta, Malta (2010)