Institute of Information and Communication Technologies
Bulgarian Academy of Sciences

Proceedings

of

The Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3)

Editors:
Francesco Mambrini
Marco Passarotti
Caroline Sporleder

Supported by:

Ministry of Education and Science

Ontotext AD

12 December 2013
Sofia, Bulgaria

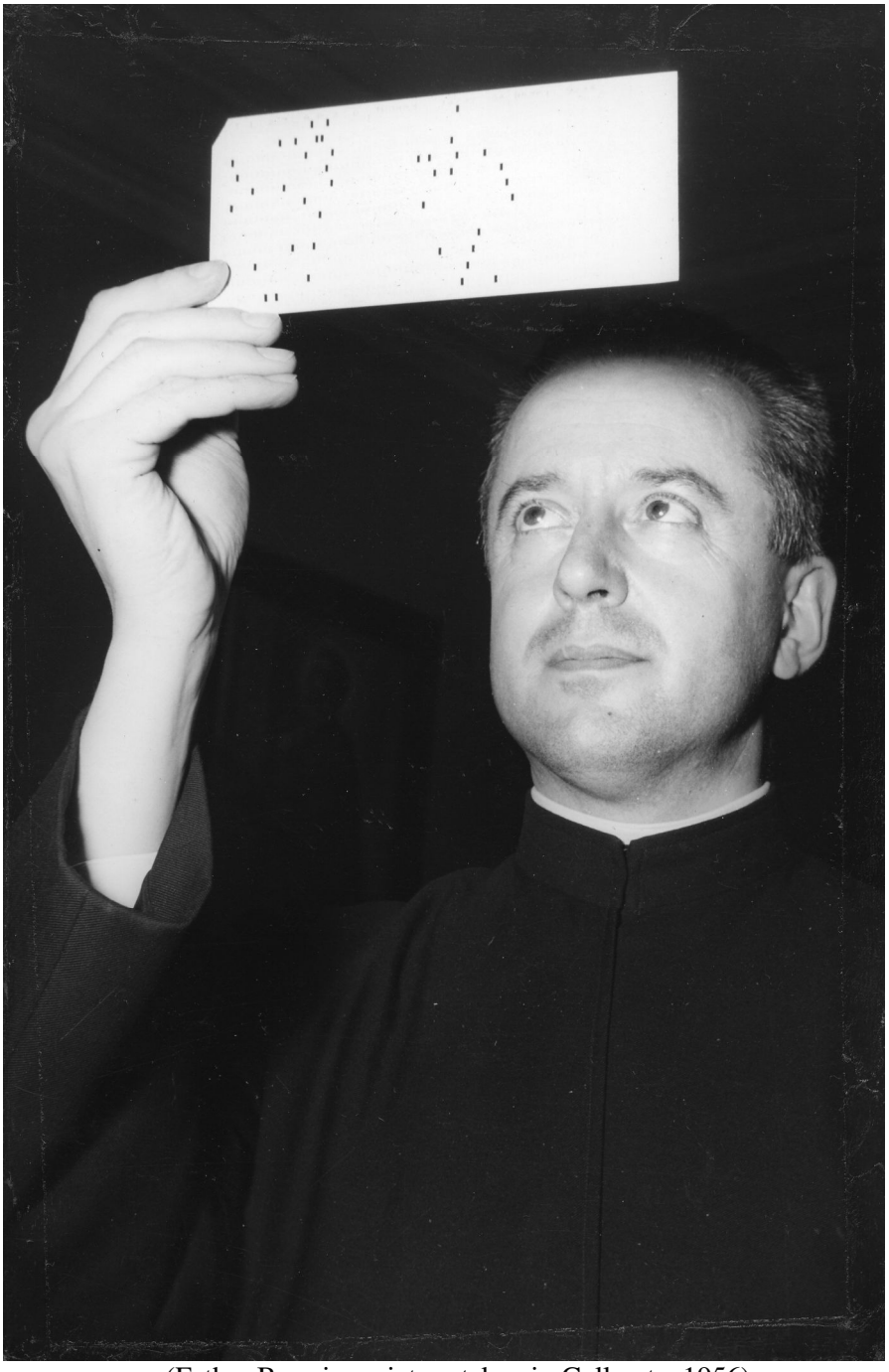
The workshop is supported by:

Ministry of Education and Science

Ontotext AD

ISBN: 978-954-91700-5-4

In loving memory of
Father Roberto Busa SJ
(1913-2011)



(Father Busa in a picture taken in Gallarate, 1956)

Preface

Over three consecutive years, the workshop on Annotation of Corpora for Research in the Humanitites (ACRH) has established itself as an occasion to foster cooperation between historical, philological and linguistic studies and current corpus and computational linguistics. On the one hand, we started from the impression that there is an undeniable similarity between how the form and meaning of the documents are examined by scholars in the Humanities and the task of corpus annotation. On the other hand, historical and literary documents are complex artifacts that require a multidisciplinary approach.

This aim has remained unaltered throughout the whole series of the ACRH workshops. Many of the accepted papers, in this and the previous editions, illustrate the point with perfect clarity, in relation to specific cases and corpora. In the present volume, Korkiakangas and Lassila discuss their approach to extend standard treebank annotation on morphology and syntax with philological and paleographic information; the work shows how crucial this layer of information can be for diachronic linguistic studies. Rehbein et al. discuss how existing POS tagsets should be augmented in order to allow researchers to annotate new content from the social media. A workflow for annotating texts that lays outside the canon of standardised modern corpora is also presented by Scrivner et al., who describe their work on the Old Occitan *Roman de Flamenca* (13th Century CE).

A set of questions that are addressed in some of the papers of ACRH-3 is how to annotate or retrieve complex semantic information (mostly from diachronic corpora) that can be of the greatest importance for research in the Humanities. Komen introduces and evaluates two models for predicting referential statuses from a manually annotated set of texts from Old to Modern English. Hendrickx et al. present an experiment on the application of unsupervised NLP methods to cluster texts dealing with the same event; their case study, which can provide interesting insights for historical corpus-based studies, focuses on retrieving strike events in a collection of newspaper articles. Kokkinakis, too, deals with historically relevant information in a diachronic corpus; his paper is dedicated to annotation of interpersonal relations in a collection of Swedish prose fiction.

In the proceedings of the first conference, we illustrated our aim by making explicit reference to the motivation that inspired a pioneer project in computational linguistics. Roberto Busa, we wrote, created the *Index Thomisticus* as a resource to help his and other scholars' investigation of the

philosophy of Thomas Aquinas[1]; to us, he represented a clear model of interaction between the work on (digitised) corpora and a research agenda in a discipline of the Humanities.

Father Roberto Busa (1913-2011) would have turned 100-year on November 28, only a few days before the beginning of our conference. Yet, it is not only to honor his person or the anniversary that we chose to dedicate the third edition of our workshop to his memory. After ACRH has already covered a significant stretch of road, it is only appropriate to look back and verify, if we may borrow one of Busa's most vivid metaphors that will be discussed in Passarotti's paper, how sound the foundations of our road are.

Two essays of ACRH-3 cover the legacy of Father Busa's work. In his key-note speech, Professor Willard McCarty, the recipient of the 2013 ADHO Roberto Busa Award, addresses the fundamental issue of the relations between philology and computer technologies. In his words, the stated aim of our workshop makes for "a fertile research *question*: what do these technological means have *fundamentally* to do with the humanities, and vice versa?"

The paper of Passarotti, on the other hand, bears a vivid witness of Father Busa's teaching, rooted in a long experience of common work on the *Index Thomisticus* and, later, on the *Index Thomisticus* Treebank. Of the many threads in Busa's legacy that will be presented to the audience there, one deserves to be singled out briefly in this preface. The stress that Father Busa laid on the manual work of annotation as a means to acquire familiarity with the language and structure of the corpus is a tangible representation of one form of the peculiar interaction between computational linguistics and philology that we have mentioned several times. The commandment to almost "love thy data" stems from the tradition in the Humanities, where each document is often considered as unique and special in its own way and subject to careful scrutiny. At the same time, it is this commitment to the texts that led Busa to the adoption of pioneering techniques and methods, as a means to open up new ways to acquire more knowledge and a better understanding of the corpus.

In this edition of ACRH, we received 7 submissions. Each of them was blindly reviewed by 3 members of the scientific committee. The program committee who was in charge of the review process was formed by 18 scholars from 12 countries in Europe and North America. Six of the proposed papers received positive reviews and were accepted for oral presentation. This very high acceptance rate points certainly to the strong motivation of the authors who submitted their work. It must be seen however whether the low number of submission (when compared to the 24 and 14 respectively for the first and second edition) means that we should try to address a broader

[1] Preface, in Mambrini Francesco, Passarotti Marco, Sporleder Caroline, editors, *Annotation of Corpora for Research in the Humanities: Proceedings of the ACRH Workshop, Heidelberg, 5 January 2012*, Journal for Language Technology and Computational Linguistics (JLCL), 26(2), pages 7-10, 2011.

audience in the Digital Humanities, lest our aim of fostering a dialogue between different disciplines is frustrated.

As with the previous editions, ACRH-3 is co-located with the Workshop on Treebanks and Linguistic Theories (TLT-12), which is hosted by the BulTreeBank Group in Sofia, Bulgaria. We wish to thank all the persons who cooperated with the organisation, and made the event possible. In particular, our special thanks go to Erhard Hinrichs, the local and non-local organisers of TLT-12, and especially Petya Osenova and Kiril Simov, the program committee, Willard McCarty and the authors who submitted papers.

The ACRH-3 Co-Chairs and Organisers
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Trier, Germany)

Program Committee

Chairs:
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Trier, Germany)

Members:
Stefanie Dipper (Germany)
Voula Giouli (Greece)
Iris Hendrickx (Portugal)
Erhard Hinrichs (Germany)
Cerstin Mahlow (Switzerland)
Alexander Mehler (Germany)
Jirí Mírovský (Czech Republic)
Michael Piotrowski (Germany)
Paul Rayson (UK)
Martin Reynaert (The Netherlands)
Jeff Rydberg Cox (USA)
Kiril Simov (Bulgaria)
Stefan Sinclair (Canada)
Mark Steedman (UK)
Frank Van Eynde (Belgium)

Organising Committee

Chairs:
Francesco Mambrini (Deutsches Archäologisches Institut, Berlin, Germany)
Marco Passarotti (Università Cattolica del Sacro Cuore, Milan, Italy)
Caroline Sporleder (University of Trier, Germany)

Local Committee:
Petya Osenova (University of Sofia)
Kiril Simov (IICT-BAS)
Stanislava Kancheva (University of Sofia)
Georgi Georgiev (Ontotext)
Borislav Popov (Ontotext)

Contents

What does Turing have to do with Busa?

Willard McCarty

Website: www.mccarty.org.uk/

> New media encounters are a proxy wrestle for the soul of the person and the civilization... We want a way of imagining our encounter with new media that surprises us out of the "us" we thought we knew.
>
> Alan Liu (2007) [28]

1 Digitizing humanities

In its statement of motivation and aims, this third ACRH Workshop conjures the ancient image of the scholar pouring over the written record of the past, brings us to the present by noting the addition or substitution of a digital machine for the codex and then comes to rest on a crux I want to consider in some detail: the problematic relation between technological means and hermeneutic ends. It notes that technological and hermeneutic work on the written record remain disjoint. It recommends a "tighter collaboration between people working in various areas of the Humanities... and the research community involved in developing, using and making accessible annotated corpora".

A laudable aim. But the central difficulty is not merely an inconvenience, inefficiency or stumbling block, nor is it merely to be overcome, say, by assembling individuals around a table or lab-bench, as desirable and appropriate as that may sometimes be. It is a fertile research *question*: what do these technological means have *fundamentally* to do with the humanities, and vice versa? In what sense, if any, are they other and more than resources to be exploited? If we can do no better than a utilitarian relationship between user and used, no matter how efficient, collaborative and harmonious it may be, there will be no digital humanities worth the candle, only digital services. Social scientists may study the impact of computing

on the humanities; computer scientists may discover problems worthy of their efforts; scholars may get further than they could have otherwise. But none of these, or all of them together, constitute a discipline *of* as well as *in* the humanities.

Why does that matter? If, as the Workshop statement suggests, digital humanities is defined (which is to say, confined) by digitization as this is usually understood, then it hardly matters at all. But if by "digitization" we mean *everything* involved in rendering cultural expressions and artefacts digital, including that which is currently beyond capture and that which might never be captured, including all reflection on and analysis of the attempt to render digitally, then at issue is a cornucopia for research, worthy of a presence among the other disciplines. Otherwise all we're being offered is infrastructure.

2 Busa and Turing

My role here is to help celebrate the centenary of Fr Roberto Busa's birth, and so my emphatic insistence on our reaching beyond the ordinary, reaching *ad astra per aspera*. I did not know Busa well and so cannot celebrate the man. To me, rather, he was and remains through his work an enlightening and kindred spirit, an intellectual father-figure, who was there, at that problematic cross-roads of the technological and the hermeneutic, from the beginning. There are other beginnings on offer. In crediting Fr Busa with the honour we digital humanists take as paradigmatic what he did, said and wrote; we take as paradigmatic (though not confining) the promise and challenge of his digital philology.

Hence the Busa Prize, given every three years by the Alliance of Digital Humanities Organizations, which now comprehends European, North American, Australasian and Japanese professional societies. The idea for the Busa Prize came from the great systematizer and theoretician of digital textual annotation, Michael Sperberg-McQueen. Michael's inspiration for it was the Turing Award, named (as you know) after Alan Turing for his invention of a fundamentally new kind of machine. This "prestigious technical award" has been given annually by the Association for Computing Machinery since 1966 for "major contributions of lasting importance to computing",[1] indeed, we cannot but say, of lasting importance *full stop*. In conversation Michael put the matter well: "if you want to know what computer science is all about," he said, "you go to the Turing Award lectures".

It's not for me to say how far we have come in realizing Michael's ideal for the Busa Prize lectures. But his implicit juxtaposition of Turing and Busa invites further thought. So I ask, what do the kinds of work which these two awards signify have to do with each other? I am asking from the perspective of the humanities, so I look at Turing's work and all that followed from it not as steps in the march of progress toward, say, Samsung's "Life Companion", but as scenes from a complex intellectual history in which our work is embedded. I want to know about the *shape*

[1] See `http://amturing.acm.org/` (2/9/13).

this history has other than merely the temporal, and what it can tell us about where and how digital humanities fits in, where and how it makes a lasting contribution.

3 The shape of technological history

Allow me to illustrate. Turing's abstract scheme, later known as the Turing Machine, was a byproduct of the 1936 negative mathematical proof that put David Hilbert's "decision problem" to rest [52; 13]. But Turing's scheme quickly diverged from its subservient role. By 1943 it had inspired the philosophical neurophysiologist Warren McCulloch and the mathematical logician Walter Pitts to design a model of the brain as a Turing Machine [37]. Two years later John von Neumann, who had read the McCulloch-Pitts paper [3: 40, 180-1; 36: 9], adopted their model in his "First Draft of a Report on the EDVAC" (1945), in which he sketched the architecture for digital hardware we still use today [56]. A modular notion of mind eventually followed. In 1948 von Neumann, who was deeply preoccupied with the physical realities of mind and machine, proposed that imitating natural intelligence might better be done "with a network that will fit into the actual volume of the human brain" [54: 34; 55]. Today precisely this is the goal of the DARPA SyNAPSE program,[2] which uses neuromorphic hardware that reflects current ideas of neurological plasticity [7].

What I want you to notice here is the historical back-and-forth, or around-and-around, of invention and human self-conception. I want to suggest that it is an instance of a "looping effect" between humans and their devices, each influencing the development of the other.[3] It has been studied, for example, in the relation between 19th and 20th-century electro-mechanical technologies and ideas of human physiology [34: 31-2]. It was given far reaching attention in the early years of computing by Douglas Engelbart, J. C. R. Licklider and others, though only for their present and future. Unlike them I want to set the human-computer relationship into the *longue durée* of human history so that we can see it as an instance of that which makes us human, and so provoke a serious rethink of digital humanities. I want to argue that if we can see our ongoing confrontation with computing as integral to that history, then there can be no doubt that digital humanities is *of* the disciplines in which we have situated it.

But I get ahead of myself. First I need to argue for the common ground of computing Turing and Busa shared and have passed on to us. Then I will set Turing's machine and its progeny into the larger history of the sciences from which they arose, specifically to connect them with the moral argument that became integral to the scientific programme in the Early Modern period. To borrow Gould's and Eldredge's evolutionary language [17], I want to use the moral dimension of this programme to argue that as a techno-scientific instrument, computing's central ef-

[2] See `www.artificialbrains.com/darpa-synapse-program` (5/5/13).

[3] I am borrowing Hacking's term, which he coined for the psychodynamic interrelation of individuals and ideas of "human kinds" [19].

fect is to punctuate our existential equilibrium and so to move us on to becoming differently human. In a sense this is nothing special: all the humanities do it. But that's my point.

4 Turing's man-machine

First, Turing. Consider his 1936 paper apart from the mathematical challenge Hilbert laid down. Consider it as a socio-cultural document as well as a mathematical proof. What do you see? What jumps out at me is the metaphor with which he begins, of the bureaucrat doing his sums. "We may compare a man in the process of computing a real number", he wrote, "to a machine which is only capable of a finite number of conditions…" [52: 59, 49]. Thus he reduces his imagined "computer" (as that man would then have been called) to a "computer" (as we would now call the corresponding machine), collapsing a familiar human role into an abstract set of exact procedures. He creates an actor-device purged of everything extraneous to those procedures and thereby, through a long and complex argument, demonstrates that no such computer-become-computer can decide whether in principle a mathematical assertion is true. But thereby he also implicitly shows the inexhaustible role of imagination in mathematics, and so in the life of the mind as a whole [33: 167-70].

We are apt to regard Turing as a rather odd, one-of-a-kind genius, but to isolate him like that covers up important connections [2; 22]. In particular is Jon Agar's demonstration of how Turing's actor-device is perfectly of its time, matching the then widespread notion of human society, government and industry as a machine [2]. This notion is found, for example, in Taylorian management [49], Fordist manufacturing practices [18] and Keynesian economics [24]. It is tragi-comically played out in Charlie Chaplin's *Modern Times*, which was released in 1936, just as Turing's paper was going into print. Man becoming machine in a machine world was in the air, so to speak.

When, almost immediately, Turing's abstract machine took on a life of its own, its implicit role in illumining the imagination became much harder to see. For many Turing's scheme supplied a model *for* mind, still visible in cognitive science, indeed, increasingly become a model for everything else. Busa, however, implicitly followed Turing's use of the machine to illumine what it could not do. In 1976 Busa, who by then had processed 15 million words for the *Index Thomisticus*, asked, "Why can a computer do so little?" [6]. Inadequate machinery could not be blamed, he wrote; human ignorance was (and is) the problem. Busa argued again and again against the emphasis on saving of labour, to which so many turned to justify what they were doing. This emphasis had been noted and attacked from the first publications on computing in the humanities, e.g. in 1962 by Cambridge linguist and philosopher Margaret Masterman, who condemned the notion of the computer "as a purely menial tool" [30: 38], and in 1966 by the American literary critic Louis Milic, who pointed out that a focus on alleviation of drudgery narrowed

research to problems involving drudgery rather than expanding its horizons. "We are still not thinking of the computer as anything but a myriad of clerks or assistants in one convenient console", he wrote [39: 4]. "In language processing", Busa wrote,

> the use of computers is not aimed towards less human effort, or for doing things faster and with less labour, but for more human work, more mental effort; we must strive to know, more systematically, deeper, and better, what is in our mouth at every moment, the mysterious world of our words. [6: 3]

Systems scientist Sir Charles Geoffrey Vickers had written a few years earlier that the powerful temptation to save human effort would bury the potential of computing to help resolve "the major epistemological problem of our time". He stated this in terms which take us back to Turing: "[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" [53].

5 Incunabular digital humanities

But practitioners seem not to have heeded the advice. Attempts to explore the implications of computing for the humanities appear either to have been censured (Masterman's playful experiments in poetry-writing by machine provoked the wrath of F. R. Leavis)[4] or to have been ignored. In 1989 literary critic Rosanne Potter wrote that "literary computing still remains outside the recognized mainstream of literary criticism. It has not been rejected, but rather neglected" [45: xvi]. In 1991 she surveyed the previous 25 years' work of *Computers and the Humanities*, identified a chorus of scholars who had written similarly about the problem in *CHum* and concluded with them that poverty of theory was to blame [46: 402-7]. The same year literary historian Mark Olsen recommended that close analytical work with computers be abandoned.[5] The obvious question here is not why literary computing (and by inference digital humanities as a whole) had failed to make an impact – poverty of theory in an age of critical theory is a sufficient explanation – rather, why practitioners remained isolated from the theoretical debates in the humanities and seemingly unaware of the exciting developments of computing in the sciences, if only to note them as irrelevant. But let me put those two historical questions on hold for a moment.

[4] Leavis [26] does not name Masterman (who had studied with Wittgenstein 1933-34 and founded the Cambridge Language Research Unit in 1955), but the details he gives suggest her strongly: "a philosopher, a lady and cultivated; her place and conditions of residence gave her access to a friendly computer laboratory"; see also [32] and [31].

[5] Olsen's MLA conference paper [43] caused such a furore among practitioners that a double issue of *CHum*, edited by Paul Fortier, resulted; Olsen's [44] was the lead piece in that issue.

The onset of the Web following its public release in 1991 (the year of Potter's review and Olsen's recommendation) seemed to seal off the first four decades of digital humanities – its incunabular period, as I call it – and mark the beginning of a new era. Some have argued that progress in the form of the Web marked a decisive turn away from a rather unimpressive past, rendering it irrelevant to present concerns. But once the dust settled ca. 2004-5 (when the first comprehensive survey [47] and theoretical treatment [33] were published), it became clear that the Web had not solved the fundamental problems, rather temporarily distracted attention from them. Thus, borrowing a term from the defensive rhetoric of government-funded services, digital humanists began talking about proving "evidence of value" – and so revealed the longevity of old anxieties [35: 118]. In 2012 a young scholar of American literature yet again asked if digital methods have had any significant effect on literary studies.[6] In late September of this year, members of the online seminar *Humanist*, which I moderate, likewise took up the old question, asking whether any great digital works of scholarship can be identified, and if so how.[7]

In other words, although digital humanities has changed with the technology, it remains on the trajectory of those formative, incunabular years: struggling with the relation of theorizing to making; uncertain of its position between the technosciences and the humanities; and, most serious of all, without a normal discourse of its own, and so without the criticism for which Alan Liu [27] and Fred Gibbs [14] have called.

At first glance it would seem simple to explain why those incunabular practitioners remained isolated, why the majority of scholars fled to the theoretical high-ground.[8]

An explanation might run something like this: attracted by technological progress, empirically minded scholars raised in the critical environment of I. A. Richards, John Crowe Ransom *et al*. were drawn to computing as soon as it became available, however unrealistic the promises on offer. But the computer was a formidable object then – a massive, noisy, sequestered, technically complex and expensive mainframe, access to which was only for the dedicated technical staff that managed it. "*The* computer" – note the definite article –was widely known to be complicit in bureaucratization of daily life, the industrialization of research and the frightening developments of the Cold War, which began with computing and ended almost exactly with the public release of the Web in 1991. This period saw the exponential growth of Jon Agar's "government machine" throughout Eisenhower's "military-industrial complex"; rampant paranoia, especially in the United States; and the threat of nuclear annihilation felt across the world. It can hardly be surprising that affiliation with computing was rare among humanist scholars. Computers had been

[6] See Ryan Cordell's posting to *Humanist* 26.257, `www.dhhumanist.org/cgi-bin/archive/archive.cgi?list=/Humanist.vol26.txt` (1/10/13).

[7] See *Humanist* 27.357, 358, 363, 369, 370, 374, `www.dhhumanist.org/cgi-bin/archive/archive.cgi?list=/Humanist.vol27.txt` (1/10/13).

[8] This is a complex historical question I do not have time to unpick here. For the asking of it, see [23], which cites [8]; see also, in the same volume, [40].

developed for numerical calculation and had only become widely available through a massive effort of salesmanship: they were not produced to meet a need, rather that need had to be found or, as often the case, created through advertising [29: 49f]. Computing thus came on the shoulders of hype, including claims of a better life for everyone, and by the very nature of Turing's scheme, with an inexhaustible future of technological progress guaranteed in principle – though manifested in emotionally dark uses.

We can, then, infer a strongly discouraging anxiety about the machine. Unfortunately (perhaps tellingly) evidence from the scholarly mainstream is sparse at best. Intriguing testimony from the mid 1950s suggests, however, that an "uneasy, half embarrassed... furtive rivalry between man and machine... [was] being fought underground because even to consider the existence of such a contest would be undignified" [50: 482-3]. If this is right (which I think it is) then perhaps we should regard the sparse, scattered signs of anxiety we do find in the professional literature most remarkable. There we come across anxieties over the distortions computing would work on the humanities if taken seriously, for which the words of those who did take it seriously provided evidence; anxieties over an immanent mechanization of scholarship, leaving scholars little to do; and anxieties over its revolutionary force, threatening to cast aside familiar ways of thinking. Curiously gratuitous reassurances that all would be well, that the scholar still had a function, suggest the very anxieties being allayed. I think what we witness here is fundamentally an existential angst, a "fear and trembling", as one scholar said [42], quoting Søren Kirkegaard: not so much "Will I have a job?" but "Who am I to become in a world defined by the computer?" There cannot be any doubt that like everyone else in Europe and North America, scholars of the incunabular period were exposed to the strongly polarized views of the machine that saturated popular media. It would be paranoid to regard these media as broadcasting an orchestrated message, but from the time of Edward Bernays' influential book *Propaganda* (1928) "[t]he conscious and intelligent manipulation of the organized habits and opinions of the masses" was doctrine in public relations and advertising [4: 9]. In the early days of the Cold War, when fear became a deliberate instrument of social control, Bernays exulted in "this enormous amplifying system" of media "which reach every corner... no matter how remote or isolated. Words hammer continually", he wrote, "at the eyes and ears" [5: 113]. The marriage of policy, commerce and propaganda then took hold.

Consider in this context the 1982 American dystopian science fiction thriller *Blade Runner* along with numerous other examples. Add the evidence of existential angst attested by the early digital humanists. Add also the many speculations about machines outstripping humans, especially in cognitive performance. Again that voice from the mid 1950s: "We have become used to machines that are more powerful, more durable, more accurate, and faster than we are, but machines that challenge our intelligence are hard to take" [50: 482]. Aren't they still?

6 The scientific programme as moral programme

Now I want to pull back from the incunabular period, first with the help of Sigmund Freud. Famously, twice in 1917, he declared that scientific research had precipitated three great crises in human self-conception, or as he put it, three "great outrages" ("große Kränkungen") [11;12]: first by Copernican cosmology, which de-centered humankind; then by Darwinian evolution, which de-throned us, setting in motion discoveries of how intimately we belong to life; and finally by his own psychoanalysis, which showed we are not even masters of own house. Less often noticed is his suggestion (implicit in the German *Kränkung*, from *krank*, "ill, sick, diseased") that these dis-easings of mind can be turned to therapeutic effect. We are apt to see only the physician here, but Freud was in fact showing his inheritance from the whole moral tradition of the physical sciences. At least from Bacon and Galileo in the 17th Century this tradition had identified the cognitively and morally curative function of science acting against fanciful or capricious knowledge – "the sciences as one would", Bacon called it in *Novum Organum*, (I.xlix). Science for them was a corrective, restorative force: "the moral enterprise of freedom for the enquiring mind", historian Alastair Crombie has written [9: 8]. We now know that in its origins science was not anti-religious; its aim was restoration of cognitively diseased humankind to prelapsarian Adamic intelligence [34: 33-4]. The religious language has gone from science, but the moral imperative remains. Freud's series of outrages is thus radically incomplete: they do not stop with him because the imperative to correct "the sciences as one would" is integral to the scientific programme.

The advance of this programme, in recent decades thanks to the computer, is impressive by anyone's measure. Consider, for example, philosopher Paul Humphreys argument that because of computing "scientific epistemology is no longer human epistemology" [20: 8]. He concludes in language reminiscent of Milton's *Paradise Lost*: "The Copernican Revolution first removed humans from their position at the center of the physical universe, and science has now driven humans from the center of the epistemological universe" [20: 156].

The odd echo of Adam and Eve's expulsion from Paradise, with implicit appeal to our foundational mythology, gives us a deeply ironic clue. It is, if you will, a reach for certainty impelled by the success of the very scheme Turing used to show there could be none. So we are on sensitive ground. Humphreys implication is that all we imagine can only be narcissistic, since consciousness of anything that cannot be effectively computed from external input has to be a self-reflection. He is not alone. Consider, for example, cosmologist and Nobel laureate Steven Weinberg, who like Freud takes aim at this narcissism, proclaiming that we live in "an overwhelmingly hostile universe" [58: 148] whose laws are "as impersonal and free of human values as the laws of arithmetic" [57: 43], "that human life is... a more-or-less farcical outcome of a chain of accidents reaching back to the first three minutes" after the Big Bang [58: 148]. Consider also the words of geneticist and Nobel laureate Jacques Monod, who aims at the same target, proclaiming "that,

like a gypsy, [man] lives on the boundary of an alien world that is deaf to his music, and as indifferent to his hopes as it is to his suffering or his crimes" [41: 160].

Grow up and face facts! we are told. But however extreme these two distinguished scientists may be, they are indicative of a much broader sense of a mounting attack of ourselves as scientists upon ourselves as humans. The case is summed up by biological anthropologist Melvin Konner: "It would seem", he concludes, "that we are sorted to a pulp, caught in a vise made, on the one side, of the increasing power of evolutionary biology. . . and, on the other, of the relentless duplication of human mental faculties by increasingly subtle and complex machines." He asks, "So what is left of us?" [25: 120].

This, I would argue, is one of those punctuations of the equilibrium that force us to rethink ourselves. Ah, the postmodern condition, you may say. Yes, but in the *longue durée* of becoming human, this is one among many punctuations. The story told for example by Roger Smith [48] and by Giorgio Agamben [1], who cites Carolus Linnaeus' 18th-century classification of the human as that species which is perpetually coming to know itself, *homo nosce te ipsum*. And, at the other end of the scale, it is the story of our every moment's "going on being" in the anxious construction of self that Anthony Giddens brilliantly describes [15]. It is legible in the attempts, such as René Descartes' in 1637, to counteract the most corrosive discovery of his age, the Great Apes, so physiologically similar to humans, physician Nicolaes Tulp wrote in 1641, *ut vix ovum videris similius*, "that it would be difficult to find one egg more like another" [51: 274]. Recall now Alan Turing's paper of 1950, in which he argues playfully that once we can no longer *tell* the difference between ourselves and our computers, there won't *be* any [52: 433-64].

7 Present and future digital humanities

Now it is time to make the connection with digital humanities and so to conclude.

In a quietly brilliant article Julia Flanders writes of the "productive unease" in textual encoding that foregrounds "issues of how we model the sources we study, in such a way that [these issues] cannot be sidestepped" [10: 22]. She argues that an irresolvable struggle is the point of it all. Literary critic and editor Jerome McGann agrees; he argues that the aim is to fail so well that all you can see is what he calls the "hem of a quantum garment" [38: 201] – the anomalous exception which once taken seriously transforms everything. Now recall Melvin Konner's agonized question, "So what is left of us?" once we face what we now know and have built, or are about to build. Isn't it formally the same question that Flanders' encoder constantly asks, mindful of the "productive unease" from which she struggles to learn? Isn't it the same question McGann has illumined by that reach for the "hem of a quantum garment" when all else but the inexplicable anomaly has been nailed down? Here is a signal of a world outgrown, and a transformed one in the offing, a catastrophe which punctuates the old equilibrium, precipitating a new order of

things, a new idea of the human.

Research in human-computer interaction has given us many fine things, including essentials of the machine I used to write these words. I would not easily give it up or surrender its most companionable interface for something less friendly. I treasure my Android phone and sense it becoming a "life companion". But for digital humanities as an intellectual pursuit I am arguing for a different kind of value, existential and cognitive, which comes from internalizing Flanders' "productive unease" in digitizing the humanities. Consider textual encoding further. If I am told by my inability to fit what I think to be an instance of personification into an ontology I have devised, then yes, of course, the ontology needs rejigging. But as my effort to render metaphor computationally tractable continues, the struggle becomes more meaningful, more and more about, as Busa said, "what is in our mouth at every moment, the mysterious world of our words", which is to say, our mysterious self. Consider another example: computational stylistics. The Australian scholar John Burrows after decades of work has amassed mounting evidence that literary style is probabilistic, in other words, that working within author and reader alike is a process identical in important respects to how the physical world as a whole operates. Most attempts to show that a computer can be creative seem quite silly to me. But this is something quite different, something that calls for one of those existential rethinks.

So I end by asking: how is such work not fundamentally of the humanities? If we claim Roberto Busa, we must also claim Alan Turing for one of our own – and pay attention to them both. Happy birthday Father Busa!

References

[1] Giorgio Agamben. *The Open: Man and Animal*. Trans. Kevin Attell. Stanford University Press, Stanford CA, 2004/2002.

[2] Jon Agar. *The Government Machine: A Revolutionary History of the Computer*. MIT Press, Cambridge MA, 2003.

[3] William Aspray. *John von Neumann and The Origins of Modern Computing*. MIT Press, Cambridge MA, 1990.

[4] Edward L Bernays. *Propaganda*. Horace Liveright, New York, 1928.

[5] Edward L Bernays. The engineering of consent. *Annals of the American Academy of Political and Social Science*, 250:113–20, 1947.

[6] R. Busa, S.J. Why can a computer do so little? *ALLC Bulletin*, 4(1):1–3, 1976.

[7] Suparna Choudhury and Jan Slaby, editors. *Critical Neuroscience: A Handbook of the Social and Cultural Contexts of Neuroscience*. Wiley-Blackwell, Chichester, 2012.

[8] W. R Connor. Scholarship and technology in Classical Studies. In Katzen [21], pages 52–62.

[9] A. C Crombie. *Styles of scientific thinking in the European tradition. The history of argument and explanation especially in the mathematical and biomedical sciences and arts*. Duckworth, London, 1994.

[10] Julia Flanders. The productive unease of 21st-century digital scholarship. *Digital Humanities Quarterly*, 3(3), 2009.

[11] Sigmund Freud. *A General Introduction to Psychoanalysis*. Trans. G. Stanley Hall. Boni and Liveright, New York, 1920/1917.

[12] Sigmund Freud. One of the difficulties of psycho-analysis. Trans. Joan Riviere. *International Journal of Psychoanalysis*, 1:17–23, 1920/1917.

[13] Robin Gandy. The confluence of ideas in 1936. In Rolf Herken, editor, *The Universal Turing Machine: A Half-Century Survey*, pages 51–102, Wien, 1994/1995. Springer.

[14] Fred Gibbs. Critical discourse in digital humanities. *Journal of Digital Humanities*, 1(1):34–42, 2011.

[15] Anthony Giddens. *Modernity and Self-Identity: Self and Society in the Late Modern Age.* Polity, London, 1991.

[16] Matthew K Gold, editor. *Debates in the Digital Humanities*. University of Minnesota Press, Minneapolis MN, 2012.

[17] Stephen Jay Gould. *The Structure of Evolutionary Theory*. Harvard University Press, Cambridge MA, 2002.

[18] Antonio Gramsci. Americanism and fordism. In *Selections from the Prison Notebooks*, Ed. and trans. Quintin Hoare and Nowell Smith, pages 277–318, New York, 1971/1934. International Publishers.

[19] Ian Hacking. The looping effect of human kinds. In David Premack Dan Sperber and Ann James Premack, editors, *Causal Cognition: A Multi-Disciplinary Debate*, pages 351–83, Oxford, 1995. Clarendon Press.

[20] Paul Humphreys. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford University Press, Oxford, 2004.

[21] May Katzen, editor. *Scholarship and Technology in the Humanities*. Proceedings of a Conference held at Elvetham Hall, Hampshire, UK, 9th-12th May 1990. Bowker, London, 1991.

[22] Hugh Kenner. *The Counterfeiters: An Historical Comedy*. Indiana University Press, Indianapolis IN, 1968.

[23] Anthony Kenny. *Computers and the Humanities*. The Ninth British Library Research Lecture. The British Library, London, 1992.

[24] John Maynard Keynes. *The General Theory of Employment, Interest and Money*. Macmillan and Company, London, 1936.

[25] Melvin Konner. Human nature and culture: Biology and the residue of uniqueness. In James J. Sheehan and Morton Sosna, editors, *The Boundaries of Humanity: Humans, Animals, Machines*, pages 103–24, Berkeley, 1991. University of California Press.

[26] F. R. Leavis. 'Literarism' versus 'Scientism': The misconception and the menace: A public lecture given in the University of Bristol. *Times Literary Supplement*, 23 April:441–4, 1970.

[27] Alan Liu. Imagining the new media encounter. In Ray Siemens and Susan Schreibman, editors, *A Companion to Digital Literaty Studies*, pages 3–25, Oxford, 2007. Blackwell Publishing.

[28] Alan Liu. Where is the cultural criticism in the Digital Humanities? In Gold [16], pages 490–509.

[29] Michael Sean Mahoney. *Histories of Computing*. Ed. Thomas Haigh. Harvard University Press, Cambridge MA, 2011.

[30] Margaret Masterman. The intellect's new eye. In *Freeing the Mind: Articles and Letters from The Times Literary Supplement during March-June, 1962*, pages 38–44, London, 1962. The Times Publishing Company.

[31] Margaret Masterman. Computerized haiku. In Jasia Reichardt, editor, *Cybernetics, Art and Ideas*, pages 175–83, London, 1971. Studio Vista.

[32] Margaret Masterman and Robin McKinnon Wood. The poet and the computer. *Times Literary Supplement*, 18 June:667–8, 1970.

[33] Willard McCarty. *Humanities Computing*. Palgrave, Basingstoke, 2005.

[34] Willard McCarty. The residue of uniqueness. *Historical Social Research / Historische Sozialforschung*, 37(3):24–45, 2012.

[35] Willard McCarty. A telescope for the mind? In Gold [16], pages 113–23.

[36] Warren S McCulloch. *Embodiments of Mind*. MIT Press, Cambridge MA, 1989.

[37] Warren S. McCulloch and Walter H Pitts. A logical calculus of the ideas imminent in nervous activity. In *Embodiments of Mind* [36], pages 19–39.

[38] Jerome McGann. Marking texts of many dimension. In Schreibman et al. [47], pages 198–217.

[39] Louis Milic. The next step. *Computers and the Humanities*, 1(1):3–6, 1966.

[40] J. Hillis Miller. Literary theory, telecommunications, and the making of history. In Katzen [21], pages 11–20.

[41] Jacques Monod. *Chance and Necessity: An Essay on the Natural Philosophy of Modern Biology*. Trans. Austryn Wainhouse. Collins, London, 1972/1970.

[42] Ellen W Nold. Fear and trembling: The humanist approaches the computer. *College Composition and Communication*, 26(3):269–73, 1975.

[43] Mark Olsen. What can and cannot be done with electronic text in historical and literary research. Paper for the Modern Language Association Annual Meeting, San Francisco, December 1991.

[44] Mark Olsen. Signs, symbols and discourses: A new direction for computer-aided literature studies. *Computers and the Humanities*, 27:309–14, 1993.

[45] Rosanne Potter. Preface. In *Literary Computing and Literary Criticism: Theoretical and Practical Essays on Theme and Rhetoric*, pages xv–xxix, Philadelphia PA, 1989. University of Pennsylvania Press.

[46] Rosanne Potter. Statistical analysis of literature: A retrospective on computers and the humanities, 1966-1990. *Computers and the Humanities*, 25:401–29, 1991.

[47] Susan Schreibman, Ray Siemens, and John Unsworth, editors. *A Companion to Digital Humanities*. Blackwell, Oxford, 2004.

[48] Roger Smith. *Being Human: Historical Knowledge and the Creation of Human Nature*. Columbia University Press, New York, 2011.

[49] Frederick Winslow Taylor. *The Principles of Scientific Management*. Harper and Brothers, New York, 1919/1911.

[50] John H. Troll. The thinking of men and machines. In Cleanth Brooks and Robert Penn Warren, editors, *Modern Rhetoric*, pages 62–5, New York, 1958. Harcourt, Brace and World. [Rpt. from *Atlantic Monthly*, (July):62–5, 1954].

[51] Nicolaes Tulp. *Observationum Medicarum. Libri Tres. Cum aeneis figuris*. Ludovicus Elzevirium, Amsterdam, 1641.

[52] A. M Turing. *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life plus The Secrets of Enigma*. B. Jack Copeland. Clarendon Press, Oxford, 2004.

[53] Sir Charles Geoffrey Vickers. Keepers of rules versus players of roles. *Times Literary Supplement*, 21 May:585, 1971.

[54] John von Neumann. The general and logical theory of automata. In Lloyd A. Jeffries, editor, *General Mechanisms in Behavior: The Hixon Symposium*, pages 1–41, New York, 1951. John Wiley & Sons.

[55] John von Neumann. *The Computer and the Brain*. Mrs Hepsa Ely Silliman Memorial Lectures. Yale University Press, New Haven, 1958.

[56] John von Neumann. First draft of a report on the edvac. *IEEE Annals of the History of Computing*, 15(4):27–43, 1993/1945.

[57] Steven Weinberg. Reflections of a working scientist. *Daedalus*, 103(3):33–45, 1974.

[58] Steven Weinberg. *The First Three Minutes: A modern view of the origin of the universe*. Flamingo, London, 1983/1977.

One Hundred Years Ago. In Memory of Father Roberto Busa SJ

Marco Passarotti
Università Cattolica del Sacro Cuore, Milan, Italy
marco.passarotti@unicatt.it

1 Introduction

I was fortunate enough to meet Father Roberto Busa even before I was born. Father Busa was a close family friend and someone I'd always seen, at least up until my last year of high school, as an old Jesuit who, they said, used computers to study Thomas Aquinas.

When the time came for me to choose which university faculty to enrol in, my parents encouraged me to study medicine. This represented a solid qualification guaranteeing a fairly smooth career (this was the early '90s and our present economic crisis lay in the distant future).

But, like many teenagers, I thought little about career opportunities and focused more on personal passions - in my case, Greek and Latin. When I informed my parents of my interest in Classics and thus my wish to enrol in the Faculty of Arts at the University of Milan, I met with their disapproval, motivated primarily by justifiable concern for my future.

In order to convince me not to go down the Humanities route, I was sent to Father Busa, whose obvious wisdom would, according to my parents, guarantee my orientation towards anatomy rather than philology.

Things didn't work out that way. To my question, «Father, I'm thinking of studying Classics. What do you think?» Busa replied, «Good idea. It's always best to choose the least travelled road». My mother, present at the interview, realised immediately that she shouldn't have put the future of her son in the hands of this eighty-year-old Jesuit, and saw the physician's white-coat evaporate in favour of a Latin, or Greek, dictionary.

For my part, I was delighted at the priest's words of approval. The opinion of such an authoritative figure cemented my decision. Just as I was leaving, Busa added, «Marco, studying Classics is the right thing to do. But only on one condition. You must buy a computer immediately and learn how to use it to analyse texts».

I admit it - I thought it was a crazy idea, just the sort of thing you'd expect from an ageing clergyman who's not quite all there. Among my reasons for choosing Classics was, in fact, not having anything to do with all that computer stuff. And

this man in the black cassock had just told me not only to buy one, but also to learn to use it to analyse Cicero and Sophocles!

Twenty years later, I am reminded every day just how right Father Busa was in putting that condition on my study of ancient texts: the use of computational machines and methods. That eighty-year-old turned out to be much younger than I was.

Readers will forgive me if a personal anecdote opens this volume dedicated to the memory of Father Roberto Busa on the centenary of his birth. My intention is not to indulge in smudgeness, but rather to hold up a real example of what Father Busa considered to be the unbreakable link between the study of language (modern or ancient) and the application of computational methods.

2 Biography. The *Index Thomisticus*

Roberto Busa was born in Vicenza on November 28, 1913. His family was originally from Lusiana, one of the seven municipalities of the Asiago plateau, where there is a small district called "Busa", now regularly inhabited by just one person.

The second of five children, the future Father Busa spent his childhood wandering between the different railway stations to which his father, a station master, was posted. Among these were Genoa, Bolzano, Verona and Belluno. And it was in Belluno that, in 1928, the young Roberto entered the Seminary to continue his high school studies. Among his classmates was Albino Luciani, «the only Pope I could address as a friend», he said. Father Busa had fond memories of those years spent in study, prayer, and the cheerful companionship of the other seminarians. Their moments of recreation also figure in those memories - such as the many football games, a sport at which the young priest was so bad that his friends insisted he play for the other side!

In 1933, at the age of twenty, Busa decided to join the Society of Jesus, motivated by both vocation and a desire to be a missionary. In 1940, shortly after being ordained, he came up before his superior, who had the task of assigning him to an area of expertise within the Order. Father Busa's recollection of that moment was always very clear, and he often recounted it in the form of a dialogue:

- Superior: «Father Busa, would you like to be a teacher?»

- Busa: «Well, not really, no»

- Superior (with a huge smile): «All right. But, you'll do it, just the same»

And so Father Busa was «shipped off» (his words) to the Pontifical Gregorian University in Rome where, in 1946, he was awarded a degree in Philosophy with a thesis entitled *La terminologia tomistica dell'interiorità* (which would later be published as a monograph in 1949). The subject of the thesis represents a focal point in the life of Father Busa. Indeed, while studying the vocabulary used by Thomas Aquinas to express the topic of interiority, Father Busa sensed that, in the texts of Thomas, the idea was conveyed by the word *in* and words beginning with the prefix *in-*, where the latter does not have the function of negation. In order to

support this hypothesis with objective evidence, Busa began to organise the words of the texts of Thomas Aquinas. Whenever a text contained an occurrence of *in*, or any of the above-mentioned words formed with the prefix *in-*, he wrote it on a card, adding a small part of its left and right context. Finally, he added the reference (work, section, etc.).

This analytical work convinced Busa that «I would be doing myself and others a huge favour» if he could find a secure and automatic way to organise and manage the words in a large set of texts. And it was with this idea in mind that, in New York, in 1949, Father Busa sat at the desk of Thomas J. Watson Sr., founder of IBM. At this meeting, Father Busa asked Watson if IBM would use its own computers to work on the texts of St. Thomas. After hearing him out, Watson asked Busa to submit his proposal as a written draft, so it could be presented to the software engineers of the company. A few days later, Busa was again summoned to Watson's office to receive project feedback from the IBM technicians. The verdict was negative - what the priest asked was impossible. And Father Busa was considered "more American than the Americans" - a kind way to say "crazy like a fox".

However: "Never take no for an answer". Father Busa was not discouraged. In fact, he challenged Watson further, focusing first on the company's methodology («Does it seem reasonable to you to say that something is impossible without even trying?») and then on its pride - in the reception area of Watson's office, Father Busa had picked up and pocketed a piece of paper on which was printed a trendy slogan, popular at IBM at the time. It read: "The difficult we do right away, the impossible takes a little longer". Father Busa pulled out the piece of paper and put it in Watson's hand, saying, «You can have this back, then. What it claims is not true». Struck by this act, Watson decided to place his trust in the visionary Italian Jesuit by giving the go-ahead to initial funding. There was, however, one condition - Father Busa must always remember that IBM stood for "International Business Machines" and not "International Busa Machines"!

The initial results obtained with that funding was an archive of 12 million punch cards, which filled a row of cabinets 90 metres long and weighed 500 tonnes. The archive was initially put together and located in Gallarate where Father Busa had been sent by his Order. He was staying at the *Aloisianum* Philosophical Institute.

IBM would finance the *Index Thomisticus* project for the next thirty years, publishing its 54 volumes in the late '70s and early '80s. In the early '90s it was published on CD-ROM, and finally appeared on the Internet in 2005[1].

In November of 2010 Father Busa sealed his relationship with IBM by gifting them his own copy of the *Index*, on the occasion of the company's centenary (1911-2011); it ranks among the greatest of IBM's achievements.

Together with the Brown Corpus, the *Index Thomisticus* is considered to be the first textual corpus recorded in a machine-readable format (today we would say

[1] `http://www.corpusthomisticum.org/it/`.

"digital"). It contains the complete works of Thomas Aquinas, a total of 118 texts, to which Father Busa added a further 61 (by as many authors, all connected to Thomas Aquinas in some way). The total number of words is around 11 million.

Father Busa passed away on August 9th, 2011, at the *Aloisianum* in Gallarate. He led the *Index Thomisticus* project right up until the final weeks of his life. To my specific questions on various decisions which needed to be taken with regard to the corpus, he always replied with a wisdom deriving from his experience of more than half a century of computer text management. He had been thinking for some years about organising an event for his 100th birthday. Not for the purpose of honouring himself, certainly; but rather as a way to use the occasion to raise funds for the continuation of the *Index Thomisticus* project.

3 Father Busa in a Nutshell

One of the most distinctive elements of Father Busa's personality was his versatility. Though far from exhaustive, this section will briefly highlight some of the outstanding features which have marked the life and scientific work of Father Busa.

3.1 His relationship with the data

Father Busa had an intimate relationship with the linguistic data. He had a deep respect for them, to the point of being skeptical about the application of natural language processing tools (NLP) to produce annotated corpora. He believed that delegating a machine to process the data would both belittle their role and risk losing control of the process.

Father Busa managed the data with absolute rigour (the same rigour he required of his collaborators), as he was convinced that the quality of data input was essential for ensuring the quality of data output. He often quoted the saying: "garbage in, garbage out". This attention to detail also comes through clearly in the motto, "aut omnia aut nihil" ("all or nothing"), which characterised the entire scientific production of Father Busa. During its compilation, one of the flagship features of the *Index Thomisticus* was precisely the fact that it contained the concordances of all the words from every text by Thomas Aquinas, «including *et* (*and*)» as Busa used to say. Today, this would not be such a big deal; in the Fifties, it was a real innovation. It was the firm conviction of Busa that, in matters of language, usable conclusions can only be achieved via complete classifications of large amounts of data.

From his frequent and rigorous consultation of scientific journals, Father Busa would often remark that most research in the Humanities consisted of a mile of algorithms based on a mere inch of foundation. He contrasted this with the methodology he employed throughout his career. As was his habit, he explained it with a metaphor. On a foundation a mile long, he would raise the research by an inch along the whole mile length. He would then proceed to raise the level by a further

inch along the whole length of the mile, and so on. All the evidence provided by each level of analysis was taken into consideration before moving on to the next level – this one slightly more advanced than the preceding one. According to Father Busa, only in this way was it possible to provide a solid basis for research conclusions.

Added to the fundamental role attributed to the empirical objectivity of the data was Busa's highly critical approach to what he called «impressionistic wisdom», i.e. conclusions, often replete with "difficult words", based on little or no evidence. It was not uncommon for Busa to ask speakers at conferences about the empirical data on which their speculations were based.

I still have a fond memory of Father Busa, now old but still combative and true to his beliefs, asking for a simple program to be developed that would allow him to disambiguate (by hand, and from his hospital bed) each of the 262,331 occurrences of the word *quod* (conjunction or relative pronoun in Latin) in the *Index Thomisticus*. This was yet another example of his tenacious willpower and staunch attachment to the source text.

3.2 The ability to communicate and engage

Anyone who ever had the opportunity to speak to Father Busa or attend one of his lectures, will remember his incredible communication skills. He often said to me that every topic, however technical or difficult, could be explained in simple terms and therefore understood by all, provided that it was clear in the mind of the speaker, «which is less frequent than you might think».

For Busa, this rule applied not only to communicating at scientific conferences, but also from the pulpit. He would frequently joke that God's existence was easy to prove - many continue to believe, despite two millennia of sermons.

This great talent for communication lent him an extraordinary presence at many different levels. The *Index Thomisticus* was put together not only by linguists and computer scientists, but also (and perhaps above all) by ordinary individuals, often on a voluntary basis. One of the major initial objectives of the *Index Thomisticus* project was to convert the texts of Thomas Aquinas from their paper format into a machine-readable version, using punch cards. Father Busa's wish was for this work to be carried out by young people who had no experience of punching cards and no knowledge of Latin. There were two reasons for this: (a) pedagogical - the experience would give them a professionally transferable and documented skill attested to by Father Busa himself, and (b) operational - the ignorant scribe is often more faithful to the text he reproduces precisely because he is not able to interpret it. Father Busa had noticed that text typed by operators who did not know Latin was of a higher quality than that produced by those who did.

But those contributing to the *Index Thomisticus* project also included elderly pensioners, who enthusiastically helped to transport huge computers and boxes of tapes and punch cards. There were also entrepreneurs who provided rooms big enough to accommodate those computers, as well as the trucks to transport

them. All of them were, in their own way, carried along by the irresistible charm of this Venetian Jesuit who had been able to convince Thomas Watson to finance a computer-based analysis of the texts of a Catholic Saint from more than seven centuries ago.

3.3 Faith

How many times did I hear Father Busa remark that he had had long discussions with his «boss», who seemed to amuse himself by continually presenting the priest with new difficulties and challenges.

One of his worst «practical jokes» took place during one of the many moves of the *Index Thomisticus* (this was the '60s, and the Internet was as long way off). The corpus, recorded at that time on magnetic tape, was being moved from Pisa to Venice. Father Busa had organised the move. Everything was loaded onto two trucks, one of which accidentally caught fire on the motorway, shortly before arriving in Venice. Several years of work went up in smoke in just a few minutes. Father Busa told me that at first he was annoyed with his Boss, but then realised that the accident had been nothing more than a warning for him to be a bit more clever. The tapes of the two copies of the *Index Thomisticus* had not been divided equally between the two trucks, but rather mixed, partly in one and partly in the other. This episode is an example of the unassailable strength of Father Busa's faith, who saw every event in a divine light, able to provide a higher reason for the events of life.

He used to say that the computer is the son of man, and therefore grandson of God. Consequently, he believed that you had to commit to using it as efficiently as possible. In keeping with his role as a priest, Busa believed that one of the noblest computer applications was to support the production of objects that would be useful to mankind. For this reason, to those who were surprised that a Jesuit would be involved with computational linguistics, Busa would reply, «I am a linguist, not *despite* being a priest, but precisely because I *am* a priest». In this sentence lies the very concept of computational linguistics according to Father Busa - a servile discipline, in the sense that it serves other disciplines (both linguistic and non-linguistic), providing them with objective documentation, reliable data and reproducible results on which to base firm conclusions.

3.4 Language, computers and knowledge

According to Busa, programming a computer is an act of wisdom, because "sapientis est ordinare" ("setting things in order is the wise man's business")[2] - «knowledge means organising, because specific to knowledge is the way in which it is

[2]This sentence is the Latin translation of a line taken from Aristotle's *Metaphysics* [I 3]. The sentence occurs 13 times in the *Index Thomisticus*, 12 of which in texts of Thomas Aquinas and 1 in the *Continuatio S. Thomae De regno* of Ptolomaeus de Lucca. Among the 12 occurrences of Thomas, 2 are reported with altered ordering, namely "ordinare sapientis est" and "ordinare est sapientis".

produced, i.e. the manner in which it is brought into being – its organisation»[3]. The *Index Thomisticus* was created precisely out of a need to meticulously organise a huge mass of words, in order to be able to manage them as well as possible and, ultimately, to have detailed knowledge of them. If programming is an act of wisdom, using the computer to analyse natural language requires wisdom.

One of the (common) views on computational linguistics which most annoyed Father Busa was the idea that computers would make things faster. Busa responded to this by emphasising the quality of the results of the computational work rather than the speed with which they could be obtained. Indeed, he argued that preparing textual data for computer analysis requires the researcher to dedicate more time (and effort) than that required for non-computer-aided research. He was convinced that striving to formalise language, even a native language, in front of a computer screen represents an extraordinary method to arrive at a detailed knowledge of it; the computer forces you to deal with all the linguistic facts you encounter, and come up with a solution to each. According to Busa, lemmatising or morphologically annotating a text (tasks now largely delegated to machines) requires and allows a penetration of language issues which «trains us for an exploration of our own inner logic, which is the spiritual centre of the personal dignity and consistency of each of us»[4]. This is a continuous "know thyself" activity, and derives from the work needed for a computer to be able to analyse language, a distinctive and essential feature of being human. This, the central aspect of computational linguistics, was for Busa an inexhaustible source of questions and a constant challenge to find the answers.

For this reason, Busa insisted that his students confront the data initially without the aid of any NLP tool. Before applying a POS tagger, the student had to morphologically annotate several thousand words by hand. Before developing an automatic lemmatiser for a particular language, the student had to have lemmatised by hand a complete medium-sized text in that language. Only direct experience with the data and the effort required to find maximum coherence in a computer analysis would permit the student to learn and grow, due to direct contact with the problems which also arise in seemingly simple operations such as morphological analysis and lemmatisation.

3.5 Computers and the Humanities

Father Busa argued that «today, the gigantic strides of microscience must be laid upon the eternal roots of our culture».

In one of his final conference speeches (XV Colloquium on Latin Linguistics, Innsbruck, 2009), Father Busa sent his best wishes and affectionate encouragement to all those who applied computational methods to Classical languages, with a recommendation to implement the saying «if at first you don't succeed. . . » - «because

[3]Roberto Busa (2000). *Dal computer agli angeli*. Itacalibri, p. 77. My translation.

[4]Roberto Busa (2007). *Libro dei metodi - vol. 20*. CAEL, p. 59. My translation.

making established culture accept a new idea is like trying to insert a new brick into a wall that's already been built».

He loved to remark that computational linguistics emerged from his work on Latin texts, but could not avoid the fact that the world of classicists and, more generally, of the Humanities is often so conservative as to reject out of hand computational tools, preferring traditional methods, as if they were incompatible with the computer. Father Busa considered this defensive position of superficial conservatism to be an elegant way to conceal fear of the unknown and a laziness towards questioning the methods you have learned.

Despite his awareness of this resistance, his attitude was always optimistic. He was sure that in a few decades, even the Humanities would find itself before a crossroads and would have to decide whether to turn and continue its journey, or go straight ahead, and condemn itself to oblivion.

4 The Importance of Father Busa

Universally recognised as one of the pioneers of computational linguistics, Father Busa was a visionary with his feet on the ground. Capable of aiming at huge long-term goals, he was intelligent enough to understand and rigorously implement the individual (sometimes tedious) steps which the facts indicated were necessary for the overall project.

The significance of Father Busa and the *Index Thomisticus* is both considerable and widespread, transcending the disciplinary boundaries of corpus linguistics, and of medieval philosophy. This is demonstrated by the fact that, even without knowing it, people benefit daily from the results obtained from the project which led to the *Index Thomisticus*. Today's ubiquitous relationship with digital texts allows us to forget how many seemingly trivial procedures are in fact the applied result of years of basic research, from the encoding of alphanumeric characters to the operations of search engines. In this sense, the *Index Thomisticus*, created with the aim of "analysing the texts of St. Thomas with a computer", was an extraordinary project of fundamental research whose design, development and testing would ultimately enter all our homes. And that's the reason IBM decided to invest in Father Busa and in that project that was more American than the Americans. Watson was far-sighted and realised at the very beginning of the information society, what huge benefits there would be for a company which could provide automatic management services for large sets of texts.

Insisting on the reproducibility of results even in the Humanities, Father Busa was able to (re)connect the Natural Sciences with the Humanities, two areas which are still too often kept separate in many academic organisations. In this sense, one of the most important contributions of Busa consisted of bringing scientific rigour to research in the Humanities, putting in place a process which survives to this day.

When asked how he saw the future of computational linguistics, Father Busa replied that the discipline would experience a «big boom» thanks to increasingly

powerful computers, the widespread diffusion of digital technology and the ease of transfer of information across the Internet. This boom would lead to excellent results, but it would also be necessary to deal with the risk of upsetting the identity of the discipline, which he considered to be closely linked to the data. He foresaw that the wide availability of NLP tools, annotated corpora, lexicons and ontologies would run the risk of being incorrectly exploited. Busa believed the greatest danger lay in considering computational linguistics not as a discipline aimed at doing things better, but rather as a tool to do things increasingly faster. He feared that the computational linguists of the third millennium would become picky about dealing with the data (which should be their bread and butter) and lose the humility to check each analysis, preferring to process huge masses of texts quickly and approximatively without even reading a line.

I believe this is the true legacy of Father Busa - a rigorous, objective and, in a word, scientific approach to linguistic data, which must remain at the centre of computational linguistic research, instead of allowing the field to merely become a hunting ground for the best-performing tool, or the largest treebank.

5 Conclusion

The last conference speech of Father Busa was at the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8, Milan, 2009). His presentation, entitled *From Punch Cards to Treebanks: 60 Years of Computational Linguistics*, gave participants an overview of his exciting research experience, beginning with the story of his first meeting with Watson.

I have a fond memory of Busa autographing volumes of the proceedings of TLT8 which participants brought up to him. And I remember the happiness in the eyes of Professor Erhard Hinrichs, who had wanted Busa as an invited speaker at TLT8 and had finally fulfilled his wish to meet a man who was truly a living legend.

So many times at various conferences, I was asked if Busa was still alive and I smile at the memory of the awe of computational linguists on hearing my positive response, which was usually, «Yes. Alive... and kickin'». But one incident narrated to me by Busa himself reveals that there were some who went further. One day, on a train taking him to Milan, Busa met a person who, having studied his face for a few minutes, asked, «Excuse me, but weren't you dead?». Busa, true to character, thought this was hilarious, and often said that such events actually increased his longevity.

Today, the work of Father Busa continues with the project of the *Index Thomisticus* Treebank, which I have the honour to head up at the CIRCSE research centre (Università Cattolica del Sacro Cuore, Milan)[5]. The project, aimed at producing the syntactic annotation of the entire *Index Thomisticus*, began in 2006 and has inserted the *Index* into the cutting edge of annotated corpora and linguistic resources

[5] `http://centridiricerca.unicatt.it/circse_index.html`.

of modern languages, making Latin, the mother-tongue of computational linguistics, a language which is no longer less-resourced. To this end there is also the close collaboration that the *Index Thomisticus* Treebank enjoys with other treebanks of ancient languages (above all the Ancient Greek and Latin Dependency Treebanks[6] and the PROIEL corpus[7]) as well as its recent integration into the CLARIN infrastructure of language resources[8].

Father Busa also donated his own (huge) archive to the Università Cattolica. Organised according to an almost maniacal order, the archive recounts a large part of the history of computational linguistics through letters, conference brochures, handouts, project reports and records of meetings. When sifting through the archive, one comes up against many historic figures from the discipline and the unique opportunity of glimpsing the most personal aspects of their relationship with Father Busa. But the archive also talks about Busa, the Jesuit priest, exuding the ardent faith that accompanied him always. The archive is now in the process of being catalogued in the library of the Università Cattolica in order to make it accessible to those students and scholars who wish to physically consult its contents, enhancing it and making it a lively and useful historical record for future generations of computational linguists, as well as others.

It is an honour and a pleasure for me to be able to contribute these few lines in memory of Father Busa, my scientific and spiritual mentor. That they appear as an opening to the proceedings of the third edition of the ACRH workshop (again co-located with TLT) would, I believe, have given Busa a great deal of pleasure; he was always a supporter of meetings between scholars in the Humanities, computer scientists and computational linguists. And this is also the goal of ACRH which, in its own modest way, wishes to continue to cultivate the seed which was planted so well by Father Busa.

[6] http://nlp.perseus.tufts.edu/syntax/treebank/latin.html.

[7] http://www.hf.uio.no/ifikk/english/research/projects/proiel/.

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

Searching and Finding Strikes in the New York Times

Iris Hendrickx, Marten Düring, Kalliopi Zervanou and Antal van den Bosch

Centre for Language Studies, Radboud University Nijmegen, The Netherlands
E-mail: {I.Hendrickx, M.During, K.Zervanou, {A.vandenBosch}@let.ru.nl

Abstract

The huge digitization step that archives, publishers, and libraries are currently undertaking enables access to a vast amount of information for historians. Yet, this does not necessarily make life easier for historians, as the main problem remains how to find relevant sources in this sea of information. We present a case study demonstrating how automatic text analysis can aid historians in finding relevant primary sources. We focus on strike events in the 1980s in the USA. In earlier work on strikes, researchers did not have at their disposal a full and comprehensive list of major strike events. Existing databases of this kind (e.g [19, 22]) are the result of intensive manual work and took years to build. Natural language processing (NLP) tools allow for faster assembly of datasets of this kind on the basis of collections of free texts that contain the information that should be in the database. We aim to construct a database of events using a digital newspaper archive and unsupervised NLP methods such as Latent Dirichlet Allocation (LDA) and clustering techniques to group together newspaper articles that describe the same strike. We study the effect of different feature representations, such as simple bag-of-words features, named entities, and time stamp information. We evaluate our results on a manually labeled sample of news articles describing a small set of strikes.

1 Introduction

The time period of the 1980s is a interesting period to study social unrest as this period saw great social, economic, and general change. The early 1980s were marked by a global recession in the industrialized countries. Many multinational corporations migrated their factories to emerging countries in Asia and Mexico leaving behind many unemployed workers. In the United States, the Reagan administration strived for free market economy and tax cuts to stimulate economic growth [15].

In this study we focus on strike events as proxy indicators of social unrest. We view a strike event as a labor action by a (significantly large) group of workers at a certain company, governmental body or specific sector that halted their work for a clear motivation, such as better working conditions, or a higher salary.

We use automatic unsupervised techniques and an online available newspaper archive as our source to find out which individual strikes took place in the 1980's. In particular, our objective is to cluster news articles based on the strike they refer to in order to seed a database of strike events. In our case study, we work with the online archive of the New York Times (NYT)[1], one of the largest and most influential daily newspapers in USA. The NYT offers an online search interface[2] to their archive that covers newspapers from 1981 to now. To facilitate the search, the articles have been manually labeled by NYT with metadata and controlled keywords (facets) such as locations, topics, and normalized names. One of these keywords is 'STRIKES', which trivially allows to retrieve all strike-related news articles from the 1980s (assuming the keyword is assigned correctly and completely). In total there are 5,987 articles on strikes in this period which form the basic set in our study. Still, the facet does not give information on which articles talk about the same strike. To identify this is a challenging task, as all strike articles are covering the same topic and therefore are similar to each other. We can expect many of these articles to cover sub-events such as picketing, negotiations between worker unions and company directors, or demonstrations. To discover how many different individual strike events took place, we need to focus on who was involved in these strikes, where they took place, and in which time period.

Our approach operates at the document level. Our task can be seen as a type of text clustering where the goal is to detect a latent group structure [18]. Note that our approach is different from a subcategorisation frame-filling approach [1] or a template-filling approach [13] to event detection, where the aim is to find an event trigger, usually a verb, and some slot or template fillers, such as agent, goal, and location at the sentence level. In our text clustering definition of an event at document level, the particular variables (who, what, when) are not explicit in the event representation. However, these variables are crucial in distinguishing one strike from another and are therefore important features in the document representation. Even though we cannot know beforehand which company names or sectors to expect, we do know that names of organizations, such as companies and labor unions will play an important role in establishing that two articles are describing the same event. Time also plays an important role, as we can expect to find news articles on the same strike when they are close in time, and when they mention the same start date of the strike, its duration and possibly an end date.

In this investigation, we use off-the-shelf NLP tools that operate on relatively simple document representations to automatically cluster documents describing the same strike. We compare two unsupervised techniques to achieve this goal. First,

[1] `http://www.nyt.com`

[2] NYT API, version 1: `http://api.nytimes.com/svc/search/v1/article`

we use *sequential Information Bottleneck* (sIB) [20], a clustering method that efficiently casts bottom-up agglomerative clustering as a sequential clustering process. This clustering algorithm has been shown to be successful in document clustering. Second, LDA [2] has become a popular method for event clustering (e.g. [24, 5]). In LDA, a document is seen as a probability distribution over a set of topics, where a topic is a probabilistic distribution over a set of words. We assume that documents about the same strike event share a similar probability distribution and the same topic will be assigned to the articles about the same event.

Note that applying a supervised technique is not feasible for this problem. Manually labeling a subset of the strike articles is not likely to be representative for detecting other individual strikes as the documents are so much alike and the unique features that will distinguish one strike from another are not expected to re-occur.

In the rest of this paper, we first present related work in Section 2. We then detail the experimental setup of our investigation in Section 3 and discuss our results (Section 4). Finally, we conclude in Section 5 with the principal findings of this study and our plans for future work.

2 Related work

A study similar to ours was carried out by Yang et al.[24]. They also aim to support historical research by applying topic modeling on a collection of historical newspaper articles. They however take an explorative and open approach and they study to what extent automatically generated topics (represented as a set of keywords) reflect historical trends. They use an expert historian to interpret the found topics. In our approach, we rather focus on a specific type of event and use the topics as document labels to find documents describing the same strike event.

De Smet and Moens [5] studied the representation of news articles for event clustering. They focus on short term events from Wikinews, and their data set had only a few relevant articles per event. They show in their work that a complex methods such as LSA or LDA did not outperform the classic tf*idf-weighted [17] bag-of-words approach on the task of event detection.

In the work of [9] LDA is applied to clustering texts into general categories such as books, business, fashion, etc. They focus on dealing with short messages as they work with tweets. [23] also apply their own proposed algorithm based on Wavelet-based signals and LDA to event detection in tweets. In their manual evaluation of the results, they conclude that LDA topic models are difficult to interpret and evaluate.

3 Experimental setup

In our experimental setup, we follow the approach taken by [5] and compare a tf*idf-weighted bag-of-words approach to LDA topic modelling. We first perform

a baseline experiment with a feature representation based on content words only. In a second experiment, we use additional features to explicate and emphasize aspects that are likely to distinguish one strike event from another: persons, locations, organizations and time stamps.

Our data set consists of the 5,987 articles that were labeled as relevant to 'STRIKE' in the NYT online archive in the the period 1980–1989. Each article is automatically tokenized, part-of-speech tagged and labeled with named entity (NE) information and time information using the Stanford CoreNLP tools [21, 7].

The Stanford POS-tagger achieved an accuracy of 97.2% on the well-known WSJ test set [21]. The Stanford NE tool obtains an F-score of 87% on the CoNLL 2003 shared task material [7]. Both test sets consist of newspaper text similar to the data that we are working with in our experiment.

We retain only content words (i.e. nouns, adjectives, verbs, adverbs) as the basis for feature representation. For the clustering, we use a bag-of-words feature vector representation. Each word receives an importance weight by computing its tf*idf score [17]. We represent each article as a weighted word vector using the 20,000 most frequently occurring content words in the data set.

For topic modeling the sequential order of the words in the article is kept, as topic modeling is based on word co-occurences. We assume that articles describing the same strike event share a similar probability distribution and the same topic will be assigned to the articles about the same event. Therefore, we assign the topic label with the highest score to each article, using the topics as class labels. We do not implement a frequency cut-off for the LDA feature representation. We apply the LDA algorithms as implemented in the Mallet toolkit [11]. For LDA, we use Gibbs sampling and asymmetric Dirichlet priors on document-topic distributions, as it was found to lead to better results [12]. In their study, MacCallum et al. [12] also showed that it is generally better to set the number of topics too large than too small. With a good topic model, the superfluous topics will have few entries, and the overall distribution of topics will still be good. In our experiments, we tested a range of 100, 150, 200, and 300 for the number of topics.

For the siB clustering algorithm, we use the Weka toolkit implementation [8] and 25 optimization iterations. As we do not know how many individual strike events actually occur in our full data set, we tested a same range of numbers of clusters as for LDA. As both unsupervised methods take random initializations, we repeated each experiment ten times with different random seeds and report the average over these ten runs.

In a second round of experiments, we add the named entities and time expressions (dates and duration) as predicted by the Stanford NLP tools as extra features to the feature vector representation. For example, the named entity string *Transport_Workers_Union* was added to the vector as a new feature, in addition to the separate words *Transport*, *Workers*, and *Union* that were already present in the feature vector. Numbers were excluded from the bag-of-words vector in the first experiment, but are now kept when part of a time expression such as *1984* or *January_1980*. Replacing individual tokens by longer strings instead of adding them

would lead to more sparsity. This way we hope to profit from the more complex and meaningful named entities in combination with the robustness of simple word frequencies.

3.1 Evaluation

For the evaluation of our approach, we manually labeled a sample of NYT articles with strike event information. Many recent studies on event detection either used a fully labeled data set (for example [5, 9, 4]), or they manually evaluated their results afterwards (for example [24, 23]). Here we chose to label a small sample beforehand in the following way. First, one of the authors, a historian, created a small list of eight strike events that occurred in the 1980s using different sources (for example [6, 19, 14, 3]). Next, he verified that all articles in the NYT archive describing these eight events were found. In total, 299 articles were linked to the eight strikes. The manual sample exhibits unevenly balanced amounts of articles per strike. The '1981 Major League Baseball' strike alone is linked to 100 relevant articles, while for the 'Chicago Tribune newspaper' strike in 1986 we only found 2 articles in the NYT. This also exemplifies a weak point in our current study: we aim to create a list of strikes in the USA in the 1980s only based on one source, the NYT. We do assume that the NYT will report on all major strikes in the country, but it is likely that a strike in Chicago is documented more extensively by regional newspapers. We run the unsupervised techniques on the full set of almost six thousand articles, and we compute recall, precision and F-scores on the manually labeled subset of 299 articles.

Both unsupervised techniques that we have applied (i.e. LDA topic modeling and siB clustering), assign an arbitrary label to each cluster, and we need to match these labels to the true manual individual strike event clusters. For every true event cluster, we check which arbitrary label was predicted most times, and we mark this label as corresponding to the true label. Once a true label is chosen as matching to a given cluster, it cannot be re-assigned to any other cluster. We computed recall, precision and F-scores on the found correspondences.

4 Results

In this section, we present the results of our experiments starting with the baseline experiments with single word features as input for LDA and the tf*idf weighted word vectors for siB clustering. We present micro-averages over the 299 documents that cover 8 strike events, averaged over 10 random initializations. The results of both algorithms with the varied number of topics/clusters can be seen in table 1 and is depicted in the left-side figures 1a and 2a. The siB algorithm performs better than LDA. In alternating the desired number of topics/clusters, we observe a clear trade-off between recall and precision. As illustrated in the figure for LDA (2a), there is an increase in precision and a decrease in recall, when we

	sIB			LDA		
# topics	recall	prec	Fscore	recall	prec	Fscore
100	70.2	51.8	59.6	64.6	40.6	49.9
150	59.7	56.2	57.9	63.8	45.6	53.1
200	49.7	61.9	55.3	61.6	48.7	54.3
300	38.7	70.3	49.9	59.4	52.1	55.5

Table 1: Baseline experiments. Micro averages of siB clustering and LDA, averaged over 10 randomly initializations. Scores were computed on the labeled sub set of 299 articles belonging to 8 different strike events.

increase the number of topics. A similar trend with steeper curves is also indicated for the siB clustering experiments (figure 1a). It is interesting to observe that while for both algorithms show this same trade-off, the overall F-score of siB decreases when increasing the number of clusters, while for LDA the F-score increases.

In table 2 and 3 we zoom in on the results for the individual strike events for those cases that have shown the highest F-score in table 1 (100 clusters for sIB and 300 topics for LDA).

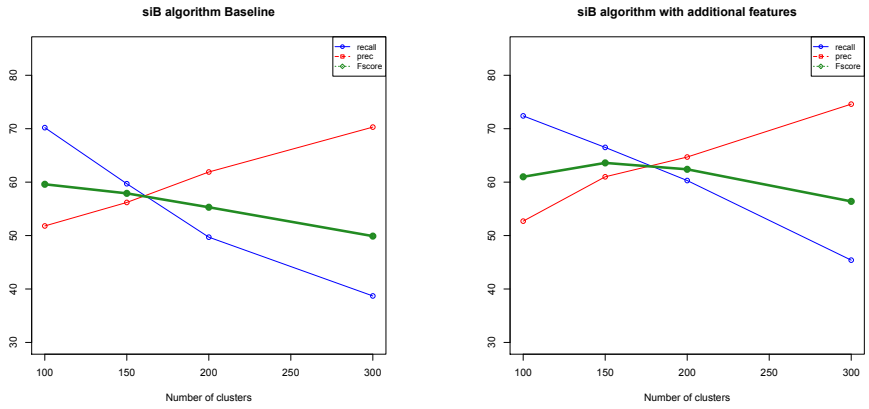
# Art	Strike event	Recall	Prec	F-score
100	Baseball League	46.1	38.3	41.8
68	Austin Hormel	92.5	88.4	90.4
52	Pittston Mine	86.4	68.4	76.2
41	Writers 1981	88.3	41.6	56.5
17	Writers 1988	5.9	1.9	2.8
14	T. W.A.	89.29	21.1	34.0
5	Arizona Copper	100.0	10.1	18.3
2	Chicago Tribune	60.0	2.5	4.8

Table 2: Results for each of the individual strike events with clustering algorithm siB with 100 clusters and 25 iterations (averaged over 10 random initializations).

In Table 2 we list the recall, precision, and F-scores for each of the eight strike events for the results of the clustering algorithm siB with 100 clusters and 25 iterations (averaged over ten random initializations).The rows in the table are ordered on event cluster size. We observe large differences between the scores on the different strike events. On the largest cluster about the Baseball League strike an F-score of only 41.8% is achieved, while the Austin Hormel meat packer strike achieved the best F-score of 90.4%. We see that low scores are obtained for the strikes represented by only a few relevant documents. We observe a very low F-score of 2.8% for the Writers strike in 1988. The explanation of this is that the algorithm did not

# Art	Strike event	Recall	Prec	F-score
100	Baseball League	26.7	39.9	31.4
68	Austin Hormel	90.3	77.3	83.3
52	Pittston Mine	72.1	67.5	66.5
41	Writers 1981	79.0	47.3	57.1
17	Writers 1988	17.1	14.1	13.6
14	T.W.A	76.4	36.3	47.7
5	Arizona Copper	98.0	26.5	40.7
2	Chicago Tribune	50.0	4.1	7.3

Table 3: Results for each of the individual strike events with LDA with 300 topics and 10 optimization iterations (averaged over 10 random initializations).



(a) (b)

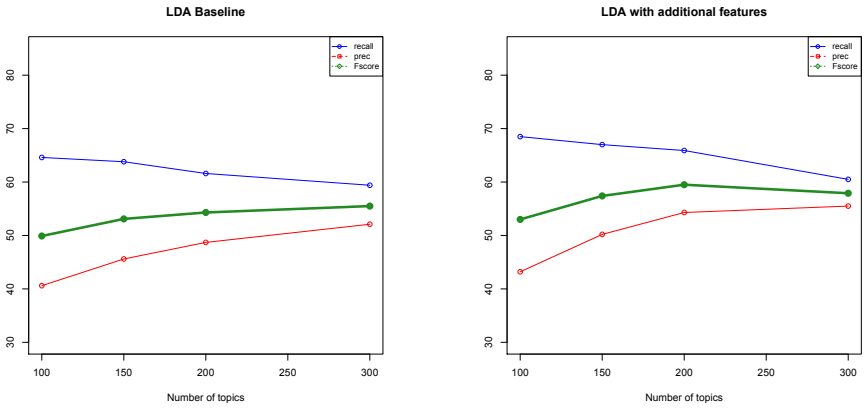
Figure 1: Micro averages of siB clustering. The baseline experiments are shown in (a) and experiments with the additional named entities and time information in (b).

succeed to distinguish the two strikes about writers: they were clustered together and were assigned the same cluster label.

In Table 3 we list the results for the same strike events obtained with LDA topic modeling with 300 topics. In general, we observe lower F-scores than with sIB for the larger event clusters and higher F-scores for the strike events covered by only a few news articles due to an increase in precision.

In the second round of experiments, we exploit additional features for named entities and timestamps in the articles. These features indeed help, as we observe an increase in the overall performance of both LDA and siB clustering as shown in figures 1 and 2.

For siB (figure 1) a clear improvement in precision and a slight decrease in recall can be observed when adding the new features. In figure (1b) sIB attains a



(a) (b)

Figure 2: Micro averages of LDA. The baseline experiments are shown in (a) and experiments with the additional named entities and time information in (b).

maximal F-score of 63.6% with 150 clusters and combines a precision of 61% and a recall 63.6%.

For LDA in figure 2, both recall and precision improve with this new feature set. The run with 200 clusters in figure (2b) has the highest F-score of 59.5%, a recall of 65.9% and 54.3% precision.

4.1 Qualitative analysis

In this section, we look at some examples to see which words are chosen by the LDA method to represent the individual clusters. This gives us some insight to what extent the techniques were indeed taking the important words as their most important features.

In Table 4 we show an example of the top words for each topic produced by LDA for three individual strikes [3]. As shown in table 3, the average results for the meatpackers strike at Hormel were rather accurate: an F-score of 82% with a recall of 91%. When we look at the top words of the most assigned topic label, we can see that all words in this topic are indeed relevant and specific. The strike took place in Austin, Minnesota and Rogers and Guyettte were union members and spokesmen in this strike against Geo A. Hormel & Company in 1986.

Most of the top words for the topic assigned to the articles about the strike event at Trans World Airlines in 1986 are indeed relevant, but this topic does not exclusively describe the T.W.A. strike, because the terms referring to Pan Am and Continental point to other companies not involved in this particular strike. This mix-up of several strikes related to airline companies explains the low precision (9.8%) and higher recall (64.3%) for this particular strike cluster.

[3] we show one example from the 10 random initializations with 200 topics.

Strike of meatpackers at Geo A. Hormel & Company in 1986
hormel plant company austin workers local p strikers union rogers meatpackers food minn parent commercial united plants geo guyette a

Strike at Trans World Airlines in 1986 of flight attendants and machinists
pilots airline flight attendants united airlines pan am company association flights continental american line international hired machinists mechanics carrier percent

The 1981 Major League Baseball
owners players miller grebey baseball kuhn committee moffett relations league commissioner negotiations bargaining owner player steinbrenner negotiating donovan ray labor
players owners free compensation player agent team agents club clubs pool baseball association proposal miller league teams grebey signing agency
baseball players league mets game yankees season stadium play team year yankee home major ball manager club pitcher sox games

Table 4: Examples of the topic models for three manually labeled strikes

For the Major League Baseball event we see the opposite happening: this strike is not covered by one main topic, but by three topics (topics 149,37,194), each covering about 25% of the articles. These topics are clearly all relevant and closely related to each other, because the terms *owners players baseball league* occur in all three topics.

LDA is clearly capable of finding topics that model the individual strikes. The top words in the topics point to the relevant sectors, worker groups (e.g. meatpackers, players, mechanics), companies, organizations, spokespersons and actions (e.g. negotiating, signing) that were parts of these strikes. However, an absolute one-to-one correspondence with the individual clusters was not found for all of the clusters in the small manually labeled sample.

5 Conclusions

In these experiments, we addressed the challenging task of clustering strike-related news articles into unique strike event clusters using unsupervised techniques and shallow word and named-entity features. Our experiments have shown that both LDA and the siB algorithm are capable of detecting those specific and relevant word features that distinguish individual strike events. For events with a smaller media coverage, we obtained lower results due to a loss in precision as multiple events are assigned to the same cluster.

The results in this case study indicate that these unsupervised techniques cannot detect individual strike events fully automatically. The results are not good enough for that. Still, it can be expected that the automatic clustering techniques will

reduce the work of the historian. Large events can to a large extent be detected automatically, and many articles can be automatically assigned correctly to single events, so that manual work can concentrate more on the long tail of incorrectly clustered articles.

As a next step we would also like to repeat this study for older historical documents. We refer to the work of [16]. They compare four named-entity extraction systems including the Stanford NE tool [7] applied to historical text material collected by applying optical character recognition to images of typed holocaust testimonies. The Standford NE tool performs best but with low scores between 44% and 60% F-score. As our method relies heavily on the performance of the named entity detector, this gives us an indication of what we can expect if we change to older newspapers. In such case retraining the NE tool will be necessary.

In our current study we treated LDA's topics as clusters. In future work, we would like to experiment with the topics that LDA produced as features instead of cluster labels, similar to how Lin and Hovy use topic signatures for summarisation [10]. Since LDA assigns multiple topic labels to each document, this rich topical representation could be the input of a subsequent unsupervised clustering approach, e.g. with sIB. This way we may combine the different strengths of the two methods.

6 Acknowledgements

We are grateful to the New York Times for sharing their articles for via the Times Developer Network. This work has been carried out within the framework of the Digging into Data project ISHER[4].

References

[1] Roberto Basili, Cristina Giannone, and Diego De Cao. Learning domain-specific framenets from texts. In *Proceedings of the ECAI Workshop on Ontology Learning and Population.*, Patras, Greece, 2008.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] Aaron Brenner, Benjamin Day, and Immanuel Ness. *Encyclopedia of Strikes in America*. ME Sharpe, 2011.

[4] Yan Chen, Hadi Amiri, Zhoujun Li, and Tat-Seng Chua. Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 43–52. ACM, 2013.

[4]Integrated Social History Environment for Research – Digging into Social Unrest: `http://www.diggingintodata.org/Home/AwardRecipientsRound22011/ISHER/tabid/196/Default.aspx`

[5] Wim De Smet and Marie-Francine Moens. Representations for multi-document event clustering. *Data Mining and Knowledge Discovery*, 26(3):533–558, 2013.

[6] Steve Early. Strike lessons from the last twenty-five years: Walking out and winning. In *Against The Current*, volume 124. 2006. http://www.solidarity-us.org/current/node/113.

[7] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[8] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

[9] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[10] Chin-Yew Lin and Eduard Hovy. The automated acquisition of topic signatures for text summarization. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 495–501. Association for Computational Linguistics, 2000.

[11] Andrew McCallum. MALLET: A machine learning for language toolkit. 2002. http://mallet.cs.umass.edu.

[12] Andrew Mccallum, David M. Mimno, and Hanna M. Wallach. Rethinking LDA: Why Priors Matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981, 2009.

[13] MUC-7. Muc-7 coreference task definition. version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998.

[14] Bradley Nash Jr. *The Crises of Unions in the 1980s*. PhD thesis, Virginia Polytechnic Institute and State University, 2000.

[15] William A. Niskanen. Reaganomics. concise encyclopedia of economics. *Library of Economics and Liberty*, 1992.

[16] Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. Comparison of named entity recognition tools for raw OCR text. In *Proceedings of KONVENS 2012*, pages 410–414. ÖGAI, 2012.

[17] Gerard Salton and Michael McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, 1983.

[18] Fabrizio Sebastiani. *The Encyclopedia of Database Technologies and Applications*, chapter Text categorization. Idea Group Publishing, 2005.

[19] Beverly J. Silver. *Forces of labor: workers' movements and globalization since 1870*. Cambridge University Press, 2003.

[20] Noam Slonim, Nir Friedman, and Naftali Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 129–136, 2002.

[21] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[22] Sjaak Van der Velden. *Stakingen in Nederland, Arbeidersstrijd 1830-1995*. Stichting beheer IISG/NIWI, 2009. Available online at: http://www.onvoltooidverleden.nl/fileadmin/redactie/Velden/Stakingen_in_Nederland.pdf.

[23] Jianshu Weng and Bu-Sung Lee. Event detection in twitter. In *Proceedings of the fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

[24] Tze-I Yang, Andrew J. Torget, and Rada Mihalcea. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104. Association for Computational Linguistics, 2011.

Annotation of interpersonal relations in Swedish prose fiction

Dimitrios Kokkinakis

Språkbanken, Department of Swedish Language
University of Gothenburg, Sweden
E-mail: dimitrios.kokkinakis@gu.se

Abstract

This paper describes the manual annotation of a small sample of Swedish 19th and 20th century prose fiction with interpersonal relations between characters in six literary works. An interpersonal relationship is an association between two or more people that may range in duration from brief to enduring. The annotation is guided by a named entity recognition step. Our goal is to get an in-depth understanding of the difficulties of such a task and elaborate a model that can be applied for similar annotation on a larger scale, both manually as well as automatically. The identification of interpersonal relations can, hopefully, aid the reader of a Swedish literary work to better understand its content and plot, and get a bird's eye view on the landscape of the core story. Our aim is to use such annotations in a hybrid context, i.e., using machine learning and rule-based methods, which, in conjunction with named entity recognition, can provide the necessary infrastructure for creating detailed biographical sketches and extracting facts for various named entities which can be exploited in various possible ways by Natural Language Processing (NLP) technologies such as summarization, question answering, as well as visual analytic techniques.

1 Introduction

The aim of this paper is to provide a description of the development of a manually annotated sample of 19th and 20th century Swedish prose fiction with semantic, personal relationships that (may) exist between the main characters of a literary work. Our goal is to get an in-depth understanding of the difficulties of such task and elaborate a model that can be applied for annotation in a larger scale, both manually (for providing a gold standard for NLP experimentation) as well as automatically (for learning such relations automatically in new novels). Binary interpersonal relation annotation is based on the context between two relevant, identified person named entities. For the task, we first applied a hybrid named entity recognition system [3,4] that identified and marked person names which were manually reviewed. Thereafter, we utilized the content of on-line available lexical (semantic) resources. These resources were various types of suitable vocabularies and ontologies with pre-defined interpersonal relationships, and relevant instances, which guided the human annotators to choose a suitable semantic label for the relation that might exist between the identified persons. Named entity recognition, in combination with vocabulary modeling, provided the basic mechanism for manually annotating the wealth of interpersonal relations in these novels.

The implications of the results can contribute to: i) the creation of a suitable material for empirical NLP in the cultural heritage domain and digital humanities (e.g., information extraction, text mining); ii) to build the first important steps for future (navigational) systems that will aid the reader of a literary work to better understand its content and plot, and therefore, iii) to support future users of digitized literature collections with tools that enable semantic search and browsing, and thus aid the computer-assisted literary analysis using more semantically oriented techniques. A long term goal of the presented work is to have the appropriate tools (annotated data, models etc.) to generate a complete profile, an exhaustive list of any kind of semantic relations, i.e. interpersonal relationships, such as *Friend-Of* and *Antagonist-Of* that can be encountered between the main characters in any literary work. In general, we are interested in macro analytic techniques that may "reduce the text to a few elements, and abstract them from the narrative flow, and construct a new, artificial object" [13] that can be useful to scholars in the humanities and social sciences. This way we can enhance the exploratory analysis by going beyond text search and navigation which depend on co-occurrences of words or entities.

Interpersonal and other types of semantic relations can aid the reader of e.g., a literary work to better understand its content and plot, and get a bird's eye view on the landscape of the core story. Despite the risks of spoiling the enjoyment that some readers of the narrative would otherwise have experienced without revealing of any plot elements, we still believe that such supporting aid can be used for an in-depth story understanding (for the human reader). Moreover, creating biographical sketches (e.g., birthplace) and extracting facts for entities (e.g., individuals) can be easily exploited in various possible ways in information science and NLP technologies such as summarization and question answering [5,11].

Naturally, there can be a large number of different relationship types between individuals in the volumes of literary archives, but as a starting point we look into interpersonal ones. In the long run it is also desirable to extract more than merely interdependency relations of individuals (e.g., birth place, workplace etc.). This work is related to the extraction of social networks [6,9].

2 Interpersonal Relationships

Sociologists, psychologists and researchers in Social Network Analysis (SNA) and communication studies have long being interested in interpersonal relationships for various reasons. Social roles form a critical component in understanding human relationships [17,18]. According to Wikipedia[1] an interpersonal relationship is: "an association between two or more people that may range in duration from brief to enduring". This association may be based on inference, love, solidarity, regular business interactions, or some other type of social commitment. Interpersonal relationships are formed in the context of

[1] <http://en.wikipedia.org/wiki/Interpersonal_relationship>; visited July, 2013.

social, cultural and other influences. The context can vary from family or kinship relations, friendship, marriage, relations with associates, work, clubs, neighborhoods, and places of worship. Furthermore, they may be regulated by law, custom, or mutual agreement, and are the basis of social groups and society as a whole. Interpersonal relationships are formed in the context of social, cultural and other influences, people can enact different social roles across the plot of a novel as they interact with different people.

3 Data: Textual and Lexical Resources

The texts we have selected for the annotation come from the Swedish Literature Bank[2], which is a co-operation between the Swedish Academy, the Royal Library of Sweden, the Royal Swedish Academy of Letters, History and Antiquities, the Language Bank of the University of Gothenburg, the Swedish Society for Belles Lettres, and the Society of Swedish Literature in Finland. While publishing canonical works by Swedish authors, the Swedish Literature Bank also focuses on neglected authors and genres, effectively establishing a set of 'minor classics' alongside the canonical works. So far, mainly texts in Swedish are available, but over time, selected works will be offered in translation as well. The texts are available free of charge. The digital texts are based on printed first editions or on later scholarly editions. They are carefully proof-read, thus establishing a basis for scholarly work. The texts in the Swedish Literature Bank (with minor exceptions) have no copyright restrictions. Table 1 shows some statistics for the annotated sample which comprises six randomly chosen books.

Title	Author	Publ. Year	Tokens	Pers. Entities	Rela-tions
Stockholms detektiven	Prins Pierre	1893	105941	2752	225
Vi Bookar Krokar och Rothar	Hjalmar Bergman	1912	98718	4004	241
Blå Spåret	Julius Regis	1916	76302	3052	154
Godnatt, Jord	Ivar-Lo Johansson	1933	158300	4191	171
Kungsgatan	Ivar-Lo Johansson	1935	182004	4296	260
Det Hemliga Namnet	Inger Edelfeldt	1999	94169	2374	181

Table 1. Characteristics of the annotated sample.

3.1 Lexical Input

A small number of suitable to our task, freely available resources, were exploited in order to find the most appropriate one for aiding and simplifying the manual annotation procedure. Three such resources were identified, namely: the *RELATIONSHIP*[3] vocabulary and two Swedish lexical semantic resources, namely the *SweFN++*[4] and the *Swesaurus*[5]. These resources have relevant features, and were thus useful in providing the appropriate machinery

[2] <http://litteraturbanken.se/#!om/inenglish>.
[3] <http://vocab.org/relationship/.html>.
[4] <http://spraakbanken.gu.se/eng/swefn>.
[5] <http://spraakbanken.gu.se/swe/forskning/swefn/swesaurus>.

for our goals, namely the appropriate interpersonal relationship labels and even instances of such relationships.

We decided to use the RELATIONSHIP vocabulary defined by Davis and Vitiello [7] which was a good starting point since it provides a description of over thirty possible familial and social relationship types that can occur between individuals[6]. This resource was designed to refine the semantics of the property `knows` in the Friend-of-a-Friend vocabulary (FOAF[7]). The description is not unproblematic though[8]. Some of these relationships may be partially overlapping or even tautological depending on context, such as `ChildOf` vs. `AncestorOf` / `DescedentOf`, `friendOf` vs. `CloseFriendOf` and `worksWith` vs. `collaboratesWith`. Moreover, there is not much available information with respect to the theoretical background of this resource which also has an approximative level of granularity. The two other resources, namely the Swedish Swesaurus [2], that is fuzzy synsets in a WordNet-like resource under active development, and the Swedish Swedish FrameNet (SweFn++) [1] provide a large and constantly growing number of synonyms and related word instances that are important for future automated relation extraction, not only between person named entities but also other words, e.g. animate nouns. In the Swedish SweFN++ such instances are called *lexical units* and are described by a number of *frames*. A frame is a script-like structure of concepts, which are linked to the meanings of linguistic units and associated with a specific event or state. A number of frames, and the lexical units encoded therein, are relevant for interpersonal relationship extraction. Such frames are for instance the *Personal_Relationship* (with lexical units: *flickvän* 'girlfriend' and *make* 'husband', etc.) and the *Kinship* (with lexical units: *barnbarn* 'grandchild', *bror* 'brother', and *dotter* 'daughter', etc.). Other frames, such as the *Forming_Relationship* (with lexical units: *förlova_sig* 'become engaged with', *gifta_sig* 'marry with', etc.) are also good candidates but this time for modeling the context between identified person entities in future hybrid (automated) annotation systems. Similarly, we have experimented with the Swedish *Swesaurus* in order to identify synonyms for the lexical units in SweFN++. This way we can increase the amount of word instances that can be part of various relevant relation types. Thus, for the word *kollega* 'colleague' we can get a set of near synonyms such as *arbetskamrat* 'co-worker'. All these instances, appropriately categorized, are good candidates for the actual lexical manifestation of the personal relationships' bearers that are not named entities, or the personal relationships' context which we intend to exploit in the future. An alphabetical list of all relations defined in the RELATIONSHIP vocabulary is given in the Appendix.

[6] More coarse grained categorization (typically between: family, work, and social) is too limited for our future goals [14], as those were stated in the introductory section.

[7] <http://www.foaf-project.org/>.

[8] For a criticism of the vocabulary, see: <http://many.corante.com/archives/2004/03/16/relationship_a_vocabulary_for_describing_relationships_between_people.php> by Clay Shirky at the "Corante blog", visited October, 2013.

4 Annotation Process

The sample was automatically annotated with named entities, including their gender (e.g., male, female or unknown). Semi-automatic review of the labels was performed in order to identify possible omissions and inconsistencies (e.g. an entity labeled sometime as male and sometime as unknown). Although the automatic name assignment was manually scrutinized, a few problems remained. For instance, collective names, that is names of groups of individuals, rather than a single individual entity, such as *[…] syskonen Robeira* 'the Robeira siblings' in which 'Robeira' was annotated as a single person while it is obviously a named entity that refers to a group. Details of the name entity recognition procedure are described in [3,4].

In contrast to the above, any type of manual semantic annotation, as in our case with the semantic relations, tends to be labor-intensive and in many cases the amount of data that can be annotated is usually restricted by the resources at hand, both financial and technological. For instance, it is desirable to maximize the amount of annotated data that can subsequently be used for automated experiments (e.g., machine learning). We created some guidelines for the annotation, which were refined a couple of times during the annotation process so that the annotation work could be come as simple and as consistent as possible. Particular attention was given to coherence, in the sense that the category boundaries should be clear and categories should describe a coherent concept, e.g., overlapping annotations were prohibited. However, the annotators were encouraged to use their 'own' labels for an interpersonal relation in case none of the provided could be easily applied. Moreover, we could not know in advance whether the distribution of the relations was going to be balanced or not. The guidelines contained mostly general principles that are not specific to a single relation, e.g., the annotators were encouraged to annotate the relations according to their sentential context provided and not introspection or other means [16].

The most serious and pervasive problem encountered was that a large number of context between two person entities could be assigned multiple relations. This problematic "ambiguity" lies in the question of which category best fits an interaction [8]. Apparently, a major cause of such cases could be attributed both to the vagueness of the controlled vocabulary and also to the related issue that there are a number of hyperonymic/hyponymic and tautological relations between the defined labels. Therefore, we decided to permit the assignation of multiple labels to ambiguous relations, but strive as much as possible for a minimalistic approach. Moreover, items that at a first glance appeared to be assigned as `Unknown` or `ParticipantInRelation` (which is a highly generic relation type that was used 326 times or 26.4% of all the relations found) were instructed to be re-checked and as much as possible minimized. Nevertheless, we included the `Unknown` category for relations in the guidelines which was intended to be used when the annotator was unable to interpret a candidate relation, since either the provided sentential context was insufficient to deduce the appropriate meaning or simply because there were cases that there was no obvious relationship

between two mentioned characters (based on the near context). Finally, if a sentence contained more than two person entities, all relations between these entities were produced.

4.1 Agreement and "Disagreement"

Two annotators were used for the task; both were native speakers of Swedish. The raw agreement and the Cohen's kappa score on the test set of 1232 potential interpersonal contexts was 783 instances or 63.5% corresponding to a Cohen's kappa score of 0.61. The results underline both the difficulty of the relation's annotation task and the need for an even more rigorous annotation scheme development when working with semantic annotations. In our case, each context between two person entities was presented alongside the sentence in which it was found in the corpus and a window of three sentences before and after that candidate relation-containing sentence. Each annotator labeled the relation[9] with the appropriate semantic category or categories. The example below is a representative example in which the relation assigned was not from the vocabulary but 'invented' by the annotator, namely `Identical` which implies that the two person entities refer to the same individual. The `Identical` relation was added in the guidelines and the RELATIONSHIP's list. For example: `Identical('Hesselman', 'Inez Robeira'):` *<relation type="identical">»Häktningsorder på madame <person g="female">Hesselman</person> alias <person g="female">Inez Robeira </person>, Götgatan 14!»* ('»Arrest Order for Madame Hesselman alias Inez Robeira, Götgatan 14!»'). Another addition required to the vocabulary was the modifier `ex-` 'previous', as in e.g. `ex-SpouseOf('József Imre', 'Györgyi'):` *Brev från <person>József Imre</person> till hans frånskilda hustru <person>Györgyi</person>, daterat i december […]* 'Letter from József Imre to his divorced wife Györgyi, dated December […]'.

It is interesting to investigate which categories caused the most disagreement between the annotators, and which inter-category boundaries were least clear. One simple way of identifying category-specific differences between the annotators is to compare the number of items each annotator assigned to each category; this may indicate whether one annotator has a stronger preference for a given category than the other annotator has, but it does not tell us about actual agreement. A simple comparison confirms that one annotator had preference for the generic `ParticipantInRelation` while the other one was more keen to introduce new labels, although it was clearly stated that this should be avoided as much as possible, e.g. `girlFriendOf` instead of `friendOf` (*<relation type="friendOf"> … medan Jans vän <person g="male">Feri</person> kör, med sin mörklockiga flickvän <person g="female">Magda</person> sittande bredvid</relation>* '…while Jan's friend Feri drives, with his dark haired girlfriend Magda sitting beside') or `walksWith` instead of `worksWith` (*<relation*

[9] The TEI-5 element "<relation type="..." .../>" was used for the relation annotation: <http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-relation.html>.

type="worksWith"> <person g="male">Wallion</person> tog <person g="male">Beyler</person> under armen och förde honom in i sitt privatrum</relation> 'Wallion took Beyler under his arm and took him into his private room').

Although the number of disagreements could probably be reduced by providing more context, that is more than the three sentences before and after a candidate relation between a named entity pair, disagreement can never be really avoided considering the difficulty of the task. Perhaps, the provided guidelines were too vague. This can be a further reason that made annotators decide differently, particularly in conjunction with some of the labels in the vocabulary that were overlapping. Moreover, some of the potential mistakes contradict the guidelines without offering a reasonable explanation by the annotator(s). For entities in the genitive form we decided to allow two named entity annotations and also add a relationship if there was a relevant context. For instance, *Feris vän András* 'Feris friend András' was annotated with the `friendOf` relationship as `FriendOf('Feris', 'András')`, with the NER annotations *<person>Feris vän</person> <person>András</person>* rather than adapting the whole annotation as a single entity, e.g. *<person>Feris vän András</person>*. Moreover, since `ParticipantInRelation` is a generic relation type, all cases where this relation name was found together with another more specific one, such as *<relation type="participantInRelation + siblingOf">...</relation>* were reduced to the specific one, in this case *<relation type="siblingOf"> ... </relation>*.

5 Conclusions and Reflections

Raw electronic corpora are of limited use to humanities researchers. If such corpora can be enriched with various layers of linguistic annotation then the true potential of such resources is unlocked. In this paper we have provided a description of a small corpus manually annotated with interpersonal relations. Our future goal is to use such data, in conjunction with suitable techniques (e.g., supervised learning), in order to automatically create a very detailed character profiling of all characters in 19th and 20th century Swedish prose fiction. Other, complementary methods, such as clustering, could be also used [10,12], since the context similarity between extracted pairs of entities can be measured in various ways. Therefore, profile implies both intra-sentential relationship discovery between person entities as well as any kind of descriptive information about the entities in the plot.

The aim is to support the users of digitized literature collections with tools that enable semantic search and browsing. In this sense, we can offer new ways for exploring the volumes of literary texts being made available through cultural heritage digitization projects. In the future we also intend to even elaborate on relationships not only between main characters, but also other categories driven by named entities, such as between persons and locations and improve both the quantity and quality of the results. This way we can also

extract significant properties of the characters and not only interpersonal relationships. It should be fairly straightforward since named entities can be reliably identified and a similar methodology can be applied to find their relationships. Applying other types of named entity types will eventually detect more relations about the characters and this will make the profiling more comprehensive which may reveal a clearer picture of the main characters' activities, friendships and associations.

During the manual annotation process, the annotators were specifically looking at *explicit* relationships supported by textual evidence and did not include relations that dependent on the reader's understanding of the document's meaning and/or her world knowledge. Although a number of implicit relations could be inferred, (e.g. `X ChildOf Y` implies `Y ParentOf X`), this issue wasn't taken into consideration at this point. Moreover we would like to explore co-reference (pronominal references) since it plays an important role for profiling (biographical) extraction and for recognizing a much larger set of relations between characters.

Computational systems often have little or no understanding of the many roles a person might be engaged in. Our future goal is to develop methods and richer computational models of semantic relationships at a finer level of granularity that would be useful for a number of applications such as semantic search or question answering, by e.g., extracting various interpersonal relation features such as communication intensity and regularity and temporal tendency. Moreover, person-related information can also be used as an important data representation possibility, where metadata sharing and reuse between various resources is an intermediate goal [15]. The goal of this research is to generate a complete profile for all main characters in each arbitrary volume in a literature collection of 19th and 20th century fiction. We also aim at a methodology that should be easily transferable to any other piece of literary work. A complete profile implies an exhaustive list of any kind of interpersonal relationships and improve previous results based on unsupervised methods [10] in the same domain and for similar tasks [12].

Acknowledgments

This work was partially supported by the Centre for Language Technology <http://clt.gu.se/> and the "A small CLT project" initiative. The author would also like to express his gratitude to the two human annotators and the anonymous reviewers for useful comments.

References

[1] Borin, Lars Dannélls, Dana, Forsberg, Markus, Toporowska Gronostaj, Maria and Kokkinakis, Dimitrios (2009). Thinking Green: Toward Swedish FrameNet++. *In Proceedings of the FrameNet Masterclass and Workshop*. Milan, Italy.

[2] Borin, Lars and Forsberg, Markus. (2010). Beyond the synset: Swesaurus – a fuzzy Swedish wordnet. *In Proceedings of the symposium: Re-thinking synonymy: semantic sameness and similarity in languages and their description*. Helsinki, Finland.

[3] Borin, Lars, Kokkinakis, Dimitrios, and Olsson, Leif-Jöran. (2007). Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature. *In Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*. Pages 1–8. Prague.

[4] Borin, Lars and Kokkinakis, Dimitrios. (2010). Literary Onomastics and Language Technology. In *Literary Education and Digital Learning. Methods and Technologies for Humanities Studies*. van Peer W., Zyngier S., Viana V. (eds). Pp. 53-78. IGI Global.

[5] Chambers, Nathanael and Jurafsky, Dan. (2008). Unsupervised learning of narrative event chains. *In Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics (ACL-08)*. Pages 789–797, Columbus, Ohio.

[6] Culotta, Aron, Bekkerman, Ron and McCallum, Andrew. (2004). Extracting social networks and contact information from email and the web. *In Proceedings of the First Conference on Email and Anti-Spam* (*CEAS)*. Mountain View, CA.

[7] Davis, Ian and Vitiello Jr, Erik. *RELATIONSHIP: A vocabulary for describing relationships between people*. <http://vocab.org/relationship/.html>; Visited July 2013.

[8] Devereux, Barry and Costello, Fintan. (2005). Investigating the relations used in conceptual combination. *Artificial Intelligence Review*. 24 (3-4): 489–515.

[9] Elson, K. David, Dames Nicholas and McKeown, R. Kathleen. (2010). Extracting social networks from literary fiction. *In Proceeding of the 48th Annual Meeting of the Association for Computational Linguistics*. Pages 138-147. PA, USA.

[10] Hasegawa, Takaaki Sekine, Satoshi and Grishman, Ralph. (2004). Discovering relations among named entities from large corpora. *In Proceeding of the 42nd Annual Meeting on Association for Computational Linguistics*. Barcelona, Spain

[11] Jing, Hongyan, Kambhatla, Nanda and Roukos Salim. (2007). Extracting social networks and biographical facts from conversational speech transcripts. *In Proceedings of the 45th Meeting of the Assoc. of Computational Linguistics*. Prague, Czech Rep.

[12] Kokkinakis, Dimitrios and Malm, Mats. (2011). Character Profiling in 19th Century Fiction. *In Proceedings of the Language Technologies for Digital Humanities and Cultural Heritage Workshop: in conjunction with the Recent Advances in Natural Language Processing (RANLP)*. Pp. 70-77. Hissar, Bulgaria.

[13] Moretti, Franco. (2005). *Graphs, maps, trees: abstract models for a literary history*. R. R. Donnelley and Sons.

[14] Ozenc, Fatih Kursat and Farnham, Shelly. (2011) Life "modes" in social media. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI).* 561-570.

[15] Pattuelli, Cristina. (2011). *Mapping Subjectivity: Performing People-Centered Vocabulary Alignment*. In Smiraglia, Richard P., ed. Proceedings from North American Symposium on Knowledge Organization, Vol. 3. Toronto, Canada, pp. 174-184. (Also published as Mapping People-centered Properties for Linked Open data. *Knowledge Organization* 38:352-59).

[16] Pustejovsky, James and Stubbs, Amber. (2012). *Natural Language Annotation for Machine Learning. A Guide to Corpus-Building for Applications*. O'Reilly.

[17] Stanley Wasserman and Katherine Faust. (2009). *Social Network Analysis – Methods and Applications*. 19th printing. Cambridge University Press.

[18] Tajfel, Henri. (1978). *Differentiation Between Social Groups: Studies in the Social Psychology of Intergroup Relations*. London Academic Press.

Appendix

A list of the relations defined in the RELATIONSHIP vocabulary.

Relation	Description
Acquaintance Of	A person having more than slight or superficial knowledge of this person but short of friendship.
Ambivalent Of	A person towards whom this person has mixed feelings or emotions.
Ancestor Of	A person who is a descendant of this person.
Antagonist Of	A person who opposes and contends against this person.
Apprentice To	A person to whom this person serves as a trusted counselor or teacher.
Child Of	A person who was given birth to or nurtured and raised by this person.
Close Friend Of	Aperson who shares a close mutual friendship with this person.
Collaborates With	A person who works towards a common goal with this person.
Colleague Of	A person who is a member of the same profession as this person.
Descendant Of	A person from whom this person is descended.
Employed By	A person for whom this person's services have been engaged.
Employer Of	A person who engages the services of this person.
Enemy Of	A person towards whom this person feels hatred, intends injury to, or opposes the interests of.
Engaged To	A person to whom this person is betrothed.
Friend Of	A person who shares mutual friendship with this person.
Grandchild Of	A person who is a child of any of this person's children.
Grandparent Of	A person who is the parent of any of this person's parents.
Has Met	A person who has met this person whether in passing or longer.
Influenced By	A person who has influenced this person.
Knows By Reputation	A person known by this person primarily for a particular action, position or field of endeavour.
Knows In Passing	A person whom this person has slight or superficial knowledge of.
Knows Of	A person who has come to be known to this person through their actions or position.
Life Partner of	A person who has made a long-term commitment to this person's.
Lives With	A person who shares a residence with this person.
Lost Contact With	A person who was once known by this person but has subsequently become uncontactable.
Mentor Of	A person who serves as a trusted counselor or teacher to this person.
Neighbor Of	A person who lives in the same locality as this person.
Parent Of	A person who has given birth to or nurtured and raised this person.
Participant In Relationship	A class whose members are a particular type of connection existing between people related to or having dealings with each other.
Sibling Of	A person having one or both parents in common with this person.
Spouse Of	A person who is married to this person.
Works With	A person who works for the same employer as this person.
Would Like To Know	A person whom this person would desire to know more closely.

Predicting referential states using enriched texts

Erwin R. Komen

Radboud University Nijmegen and SIL-International
E-mail: E.Komen@Let.ru.nl

Abstract

Information structure research that makes use of diachronic corpora would be greatly facilitated by having texts that are not only syntactically parsed, but for which the referential state (an indicator of newness) of each noun phrase is available as well. Manual or semi-automatic annotation of the referential state is a tedious and time-consuming job, but the availability of 18 texts that have been annotated by our research group, allows training a statistical predictor and evaluating its performance. The statistical predictor discussed in this paper makes use of TiMBL, outperforms a hard-coded deterministic predictor, and reaches an overall precision of 83% to 87%. While this is not good enough for fully automatic annotation of texts, the predictors are usable in the kind of diachronic information structure research that requires only a rough estimate of the referential state of noun phrases.

1 Introduction

Diachronic corpora provide an ideal platform for research into the relation between syntax and information structure, since changes in the syntax of a language may lead to changes in the information structure system, as has been shown for English [10, 16]. Yet, while there is a growing number of historical texts that have been parsed syntactically, which makes them suitable for finding and tracking syntactic changes, texts that contain adequate information structure annotation are scarse. Komen [9] argues that the information structure of a sentence can be calculated from the combination of referential and syntactic information, so that syntactically annotated texts only need to be enriched with referential annotation. It is because of its crucial role in diachronic research that referential enrichment is currently being undertaken by projects like PROIEL [6], ISWOC [1] and the Nijmegen group [8], to which the author belongs.

Referential annotation consists of two parts: (a) a label for the referential status of each relevant constituent, and (b) a link to the antecedent of the constituent, if the constituent has one.[1] The referential status annotation task, if done manually, is a tedious one, since high accuracy is pursued, so that the annotated texts can serve information structure research in the most reliable way.[2] Fully automatic annotation of the five different possible referential states (see 2.1) has not yet been reported, but much work has been done on automatic coreference resolution (part "b" of the referential annotation

[1] The three projects mentioned differ in the particular labels that are used for the referential states (see 2.1).

[2] The Nijmegen group reports a Cohen's kappa above 0.8 for the Pentaset annotation. The Proiel group reports a kappa of 0.89 in distinguishing "New", "Old" and "Accessible", which was reached after several trials and discussions of inconsistently tagged situations [6].

discussed above), even though this generally aims at only finding noun phrases with a "Given" status as well as a link to their antecedent (one exception is [15]). The results of coreference resolution are reported to reach a precision of 70%-81% for the fully automatic task [12, 13]. Information structure research is more interested in part "a": getting the labels for the referential states of NPs. The "Cesax" program handles both referential status and antecedent link, yields high enough precision [8], but its semi-automatic nature requires substantial user-input.

The Nijmegen group has annotated a small body of some 18 texts (appr. 100 kWords) with referential status and antecedent location, and the availability of this material opens the doors to new directions in tackling the annotation task: the annotated texts can be used as training and test data for statistical approaches to the tasks of determining the referential status of constituents and determining their antecedents.[3]

This paper seeks to establish whether a statistical approach to referential status prediction is feasible and how it compares to a deterministic approach. Referential status prediction has applications in those information structure research topics that do not need detailed access to antecedents, but only require coarse-grained referential status distinctions.[4] A full-fledged referential status prediction will also have an application in computational linguistics as a logical preprocessing step before coreference resolution takes place, witness the research conducted by Uryupina [15] and Gegg-Harrison et al. [5].

2 Developing predictors

The way a predictor works depends on the required output (2.1) and the information that is fed into it (2.2). Both the deterministic predictor (2.3) and the statistical predictor (2.4) use the same syntactically annotated texts as input, but their capabilities differ due to their nature: the deterministic one requires detailed prior information about the interaction between syntax and referential states, whereas the statistical one does not.

2.1 The output of the predictor: referential states

The experiments described in this paper use predictors that specify referential state in a number of ways. The coarsest distinction that is made is the one between "Link" constituents and "NoLink" constituents: those marked "Link" have an antecedent in or outside the text, while those that are "NoLink" do not have an antecedent. The finest distinctions that are made

[3] While the principle of using manually annotated material as training for statistical prediction has been used in coreference resolution in general, it has not, as far as I am aware, been used in the finer-grained task of referential annotation.

[4] Research into Old English V2 behaviour and the transition from OV to VO for instance, has, until now, only made use of such coarse information [10, 11, 16].

boil down to five different primitives, that can be referred to as the "Pentaset" [8]. The different referential states can be discerned in the following way:

(1) *Referential states included in "Link" and "NoLink"*

Link

a. Identity (Proiel: OLD) The constituent has an antecedent in the text, and the referents of both are identical.
b. Inferred (Proiel: ACC-inf) The constituent has an antecedent in the text, but the referents of the current constituent and its antecedent are not the same (they can be in a part-whole relation, for instance). The mention of the first noun phrase must already have implied the existence of the second noun phrase, which infers from it.
c. Assumed (Proiel: ACC-sit + ACC-gen) The constituent has an antecedent, but it is outside the text. The referents of the current constituent and this antecedent must be equal.

NoLink

a. New (Proiel: NEW) The constituent does not have an antecedent inside or outside the text, and it can be referred to later on.
b. Inert (Proiel: not labelled) The constituent does not have an antecedent inside or outside the text, and it cannot be referred to in the following context.

Section 3 describes experiments that differ in terms of the output distinctions in referential states that are made. The different schemes are in (2).

(2) *Output schemes for a predictor*

a. "Link-NoLink" – Predict whether the constituent has an antecedent (Link) or not (NoLink).
b. "Link-New-Inert" – Make a three-way distinction. Constituents are first checked on whether they have an antecedent (Link) or not (NoLink). The last category "NoLink" is then further divided in constituents that can function as antecedents (New) and those that cannot (Inert).
c. "Pentaset" – Make the five-way distinction as in (1), so that the output is one of the states of the "Pentaset": Identity, Inferred, Assumed, New or Inert.

The output scheme in (2.a) is chosen, since it seems to be the easiest distinction that can be made and that is still useful for information structure research. The three-way scheme in (2.b) and the five-wey scheme in (2.c) are included to find out in what way increasing the number of referential states to be distinguished influences the performance of the predictors.

2.2 The input to the predictor

The knowledge of which the predictor described in this paper can make use consists of the syntactic, morphological and functional information available in a number of syntactically parsed texts that have been taken of four historical corpora containing excerpts from English literature from roughly

1000 A.D until 1914 (see [14] for the oldest corpus). References to these sources and to the 18 text subset of them that have been enriched with referential status annotation are listed in Komen [9]. The texts are mostly narratives, history and sermons; the average number of words per text is 5000. The referential state annotation that has been added to the noun phrases in these texts uses the five "Pentaset" states discussed in section 2.1. The annotation used in these English corpora differs in detail from the wider-known Treebank II annotation, but the principle is comparable [2, 14].

A referential state predictor needs to be able to determine the status of all noun phrases in a text, except for those that are lexically empty (their antecedents are predictable and including them would skew the data).[5] The information a referential state predictor is able to use, then, consists of all the morphological, syntactic and functional information available in the parsed English corpora.

2.3 The deterministic predictor

The Deterministic predictor is hard-coded as a built-in Xquery function `ru:RefState` within the program "CorpusStudio" [7]. It is within this program that it has easy access to all kinds of information that can be gleaned from the syntactically parsed texts. The output of the deterministic predictor is the two-way Link-NoLink division. A description of the algorithm behind `ru:RefState` follows here in (3).

(3) REFSTATE(*ndThis*)
1 *hd* ← HEAD(*ndThis*)
2 *npt* ← *hd.NPtype*
3 **if** *npt* = 'Proper' **then**
4 **return** (**if** OCCURSBEFORE(*hd*) **then** 'Link' **else** 'NoLink')
5 **end if**
6 *pm* ← POSTMODIFIER(*ndThis, hd*)
7 **if** EXISTS(*pm*) **then**
8 **if** *hd.Label* **in** Adj, Adv, N, NS, Num, PP, Q **return** 'NoLink'
9 **else if** *hd.Label* **in** Pro, D **return** 'Link'
10 **else return** 'Link'
11 **end if**
12 **end if**
13 **if** *npt* **in** Dem, DemNP, Pro, PossPro, PossDet, … **return** 'Link'
14 **else return** 'NoLink'
15 **end if**

The algorithm first of all in (3.1) attempts to find the head of the NP. If the head is a proper noun (3.3), a previous occurrence determines whether the referential state prediction is "Link" or "NoLink" respectively. Step (3.6) of the algorithm checks the presence of a post-modifier, and if it finds one, it

[5] Lexically empty NPs in the parsed English corpora include subjects that are elided under coordination, traces for *wh*-movement, empty expletives and some more categories [14].

determines the referential status just by looking at the syntactic tag of the head in (3.8-11). If there was no post-modifier, then steps (3.13-15) of the `ru:RefState` procedure evaluate the "NPtype", which comes in the form of a feature that has been automatically added to every noun phrase previously, solely on the basis of the available syntactic information.[6] Certain types of noun phrases translate directly into a matching referential states: noun phrase types of DEM (independent demonstrative pronouns), DEMNP (noun phrases headed by a demonstrative), PRO (pronouns), POSSPRO (possessive pronouns) and POSSDET (noun phrases that start with a determiner in the form of a possessive noun or proper noun) all result in a state of "Link", and noun phrases of type QUANTNP (quantifier noun phrases), INDEFNP (indefinite NPs), BARE (bare nouns) and EXPL (explitives) are all marked as "NoLink".

2.4 The statistical approach

The statistical predictor that has been used in this research makes use of memory-based learning (see [3] for an introduction in memory-based language processing). The reason to opt for a memory-based approach instead of for a more generalizing approach such as a maximum entropy one or a naive Bayesian, is that it seems quite likely that there are some idiosyncratic combinations of features determining a particular referential state, and we also expect there to be lexical dependencies. A statistical approach that defines classes based on generalizing over samples will, necessarily, miss out on idiosyncratic outcomes, whereas a memory-based approach should not.

The memory-based approach, being statistical in nature, needs to start with a training phase: it needs to have a collection of feature-value combinations with their corresponding classification. The input for the *training* consists of a list that gives the features of each noun phrase and the referential status that has been assigned to it. The referential status of a newly encountered noun phrase is, after training has taken place, determined by comparing the features of this noun phrase with the features of *all* the noun phrases in the training set. The referential category of the nearest neighbour in the feature space is assigned to this new noun phrase. Table 1 lists the features that have been used in the three experiments described in this paper.

The information available in the five features used in experiment 1 is comparable to what the function `ru:RefState` in the deterministic approach described in 2.2 uses: the information contained in the label of the NP (feature "NP_Label"), the information available from the head (feature "Head_Label"), the presence of an anchor (feature "Ch_Anchor"), the post-modifiers information (feature "Ch_PostMod") and the value of the NPtype

[6] The "NPtype" can have the following values: DEM, DEMNP, PRO, POSSPRO, PRONP, POSSDET, QUANTNP, INDEFNP, BARE, BAREWITHPP, EXPL, ANCHOREDNP, PROPER, DEFNP, FULLNP or UNKNOWN. The exact definition of these categories is less important for the purpose of this paper.

feature (feature "NP_Type"). The difference between the two methods is that the memory-based approach does not depend upon hard-coding, which means that it is not prone to oversight on the part of the programmer, and that it is easily extendable to other languages, without requiring language-specific knowledge.

Name	Description	Exp 1	Exp 2	Exp 3
Period	Time-period of text		+	+
NP_Label	Phrase label of NP	+	+	+
NP_Type	NP feature	+	+	+
NP_GrRole	NP feature		+	+
NP_PGN	NP feature		+	+
NP_words	Number of words in NP		+	+
Ch_FreeRel	NP is a free relative		+	+
Ch_Rel	NP has an RC child		+	+
Ch_Neg	NP has a negator		+	+
Ch_PreMod	Phrase label of pre-modifier		+	+
Ch_PostMod	Phrase label of post-modifier	+	+	+
Ch_Anchor	NP has a possessive pronoun	+		
Ch1_Label	Phrase label of NP-child #1		+	+
Ch1_WrdType	Word type of NP-child #1		+	+
Ch1_WrdText	Text of NP-child #1		+	+
Ch2_Label	Phrase label of NP-child #2		+	+
Ch3_Label	Phrase label of NP-child #3		+	+
Ch4_Label	Phrase label of NP-child #4		+	+
Head_Text	NP-head text		+	+
Head_Before	NP-head occurred earlier		+	+
Head_Label	NP-head phrase or POS label	+		
SisterBE	NP has *be*-verb sister		+	+
SisterSBJ	NP has subject as sister		+	+
SisterV	NP has verbal sister		+	+
SisterCP	NP has any CP as sister		+	+
Sbj_NPtype	NPtype feature of subject		+	+
Sbj_Text	Text of the subject		+	+
Cls_Mood	Mood of the clause		+	+
Cls_Speech	Clause is direct speech		+	+

Table 1 Features used in the memory-based approach[7]

Experiments 2 and 3 make use of a larger feature set, as shown in Table 1, but they do away with the Ch_Anchor feature (which is replaced by the more general Ch1_Label feature) and the Head_Label feature.

[7] There is a lot of 'implicit' information in the features. NPs that are non-linking since they are part of a presentational constructions of type "there is/are NP", for instance, can be recognized by the combination of "SisterSBJ" having value "1" and the feature "Sbj_NPtype" having the value "Expletive".

3 Results

Both the deterministic as well as the statistical predictor are evaluated by making use of the "CorpusStudio" program, which provides an environment for running Xquery on the annotated English texts [7].

3.1 Testing the performance of the predictors

The deterministic referential state predictor does not require learning, since it is "hard-wired" as the Xquery function `ru:RefState` within the program "CorpusStudio". The Xquery code that uses this function and returns a list outlining how many instances of each of the predicted states have been found for each of the actual states follows the algorithm sketched in (4).

(4) TESTPREDICTOR(*list_of_texts*, *pred_type*, *scheme*)

```
1  for each np in list_of_texts
2      if not(empty(np)) then
3          ref ← np.RefState
4          actual ← SCHEMESTATE(ref, scheme)
5          if pred_type = 'Deterministic' then
6              pred ← RU:REFSTATE(np)
7              ADDTOOUTPUT(actual, pred)
8          else
9              feat_vec ← GETFEATUREVECTOR(np)
10             TIMBLPREP(feat_vec, actual, 70)
11     end if
12 next np
```

The algorithm in (4), when called with *pred_type* set to 'Deterministic', serves to test the "Link-NoLink" output scheme described in (2) in section 2.1. The deterministic predictor (see section 2.2 and the algorithm in 3) in the form of the function `ru:RefState` does not distinguish all five referential states from the Pentaset, which means that we cannot test its performance for the second and third predictor output scheme (2.b) and (2.c).

Since the memory-based predictor is a statistical one, testing is done by dividing the available data into a training set and a test set. The procedure to test the memory-based predictor retrieves the actual referential state of each noun phrase, it determines the values of the features that are going to be used in the prediction, and it divides the noun phrases of the 18 texts used for this experiment over a training and a test set.

The variable *scheme* that is passed to the procedure in (4) determines which of the three output schemes is required: (a) the two-way "Link-NoLink" division, (b) the three-way "Link-New-Inert" division, or (c) the five-way Pentaset division "Identity-Inferred-Assumed-New-Inert". Step (4.4) makes sure the output state matches the scheme that is being used. Step (4.9) calls a user-built Xquery function that determines the values of the features used for the predictor, and step (4.10) makes sure that a training set

(70%) and a test set (30%) with feature vectors and outcomes is prepared for further processing by TiMBL, the memory-based engine that is used here [4].[8] It is TiMBL that performs the actual prediction.

3.2 Performance of the two predictors

So far, I have discussed two predictors, a deterministic one (2.2) and a statistical one (2.4), and I have shown that there are slightly different ways to test the performance of these predictors (3.1), although they both can be tested by using the program CorpusStudio. This section discusses the outcome of the tests that have been done to determine the performance of the predictors, and it follows the predictor outcome schemes described in (2).

3.2.1 Performance for the "Link-NoLink" output

The first comparison of the two different predictor types takes the "Link-NoLink" output scheme defined in (2.a). The deterministic predictor is tested on all of the available data, while the statistical one is trained on 70% of the data and then tested on the remaining 30%. Table 2 contains the confusion matrices that show which kinds of mistakes are being made by the predictors, as well as the overall performance in terms of precision and F-score (abbreviated as "P, F").

	Deterministic predictor				**Memory-based (timbl) predictor**			
Actual	**Link**		**NoLink**		**Link**		**NoLink**	
Link	**9977**	41,60%	2867	11,90%	**3320**	47,00%	472	6,70%
NoLink	1089	4,50%	**10049**	41,90%	445	6,30%	**2825**	40,00%
P, F	83,5%, 91,0%				87,0%, 93,1%			

Table 2 Predictor performance for the "Link-NoLink" output scheme

The statistical memory-based predictor outperforms the deterministic one for this test: the precision is higher and the F-score is slightly higher too. The performance of the predictors for individual referential outcome categories can best be observed by looking at their individual precision, recall and F-score values.

Method	**Feature**	**TP**	**FP**	**FN**	**TN**	**Precision**	**Recall**	**F-Score**
Deterministic	Link	9977	1089	2869	10057	90,2%	77,7%	83,4%
Deterministic	NoLink	10049	2868	1096	9979	77,8%	90,2%	83,5%
Statistical	Link	3320	445	472	2825	88,2%	87,6%	87,9%
Statistical	NoLink	2825	472	445	3320	85,7%	86,4%	86,0%

Table 3 Performance per referential state for the "Link-NoLink" scheme

The statistical memory-based predictor shows a more balanced behaviour when it comes to the performance on individual referential states. The

[8] The memory-based predictor uses the default settings of TiMBL: "IB1" algorithm, "Overlap" metric, "GainRatio" weighing. Future work will include experimenting with different settings, in order to see how much the predictor can be improved.

deterministic one is particularly bad at predicting the state "NoLink": there are over twice as much false-positives (marked "FP") than when predicting the "Link" state. The deterministic predictor gains a higher precision when it comes to predicting the state "Link", but it does so at the cost of an increased number of false-negatives (marked "FN"), which leads to a smaller recall.

Apparently the deterministic predictor is too conservative in recognizing the state "NoLink", while it is too optimistic when assigning the state "Link". A detailed analysis of the data would be needed to find out exactly why this is so, and what could be done to remedy this.

3.2.2 Performance for the "Link-New-Inert" output

While the deterministic predictor is limited to discerning whether an NP has a "Link" or a "NoLink" state, the statistical predictor can be used to get a more detailed output. The second experiment for the referential state prediction is done only with the statistical predictor, it uses the 27 features shown in Table 1, and its outcome distinguishes three states: (a) "Link" (which combines the Pentaset states of "Identity", "Inferred" and "Assumed"), (b) "New" and (c) "Inert". Table 4 shows the confusion matrix that results for this output scheme.

Actual	**New**		**Link**		**Inert**	
New	**2387**	33,1%	404	5,6%	127	1,8%
Link	380	5,3%	**4065**	56,4%	43	0,6%
Inert	114	1,6%	58	0,8%	**488**	6,8%
P, F	86,0%, 92,5%					

Table 4 Predictor performance for the "Link-New-Inert" output scheme

Comparing the confusion matrix in Table 4 with the one in Table 2, we can see that the overall performance of the statistical predictor does not change radically when we turn from a rough two-way distinction (87% precision) to a finer three-way one (86%). The performance of the three different states is shown in Table 5.

Feature	**TP**	**FP**	**FN**	**TN**	**Precision**	**Recall**	**F-Score**
New	2387	494	531	5148	82,9%	81,8%	82,3%
Link	4065	462	423	3578	89,8%	90,6%	90,2%
Inert	488	170	172	7406	74,2%	73,9%	74,1%

Table 5 Performance per referential state for the "Link-New-Inert" scheme

The precision of the state "Link" in Table 5 (89,8%) is actually a little better than the one for "Link" in the two-way experiment in Table 3 (88,2%), which is what we would expect, given the higher number of features (27 instead of 5) taken into account. Making the distinction between the referential states "New" and "Inert" proves to be possible with less precision, and future work will need to find out whether crucial features are missing from the set that make the prediction of these states more effective.

3.2.3 Performance for the "Pentaset" output

The third experiment aims at predicting the full range of Pentaet states (2.c), which is currently only possible with the statistical predictor. Table 6 shows that the overall precision does decrease for this task (it changes from 86% in the three-way distinction to 81% in the five-way Pentaset distinction), and Table 7 shows the performance of the predictor for each of the individual referential states.

Actual	**Assumed**		**New**		**Identity**		**Inferred**		**Inert**	
Assumed	**63**	0,9%	63	0,9%	82	1,1%	20	0,3%	0	0,0%
New	42	0,6%	**2360**	32,7%	228	3,2%	121	1,7%	116	1,6%
Identity	57	0,8%	210	2,9%	**3474**	48,2%	87	1,2%	32	0,4%
Inferred	13	0,2%	105	1,5%	158	2,2%	**87**	1,2%	11	0,2%
Inert	3	0,0%	97	1,3%	44	0,6%	12	0,2%	**433**	6,0%
P, F	81,0%, 89,5%									

Table 6 Predictor performance for the "Pentaset" output scheme

Feature	**TP**	**FP**	**FN**	**TN**	**Precision**	**Recall**	**F-Score**
Assumed	63	115	165	7575	35,4%	27,6%	31,0%
New	2360	475	507	4576	83,2%	82,3%	82,8%
Identity	3474	512	386	3546	87,2%	90,0%	88,6%
Inferred	87	240	287	6871	26,6%	23,3%	24,8%
Inert	433	159	156	7170	73,1%	73,5%	73,3%

Table 7 Performance per referential state for the "Pentaset" output scheme

The precision for determining the referential states "Assumed" and "Inferred" are both quite low: 35,4% and 26,6% respectively. Only 87 out of a total of 374 noun phrases that should be labelled "Inferred" are recognized as such. The reason for these mis-classifications may be quite obvious: the form of a noun phrase alone (e.g. *the table*) is just not enough to be able to say whether it has the referential state "Identity", "Assumed" or "Inferred": in all three situations definite NPs may occur. The feature "Head_Before" makes it clear whether the current NP has in some context been mentioned previously in the text, and in this way helps recognizing clear "Identity" cases. But it is quite obvious that more research needs to be done to find relevant features that help distinguish "Assumed" and "Inferred" NPs.

4 Conclusions and discussion

In this paper I have described and evaluated the feasibility of a statistical approach to referential status prediction and I have compared it with a deterministic approach. The deterministic predictor does not need training, but is very much language (and corpus) specific. The statistical predictor discussed here makes use of memory-based learning, needs training, but outperforms the deterministic one. The evaluation has consisted of three experiments, and the first one was a direct comparison between the two

predictors, where both were fed with more or less the same information, and where the outcome was a two-way referential state division: “Link” versus “NoLink”. The deterministic predictor reached an overall precision of 83,5%, while the statistical one reached 87%. The second and third experiments concentrated on establishing the limits of the statistical predictor, which was now fed with 27 features. The outcome of the second experiment was a three-way referential state division, Link-New-Inert, and the precision reached was 86%. The precision with which the three states were predicted did not differ very much. The outcome of the third experiment was a full-fledged five-way referential state division, and the precision reached was 81%. This experiment revealed the current limit in referential state prediction: “Assumed” and “Inferred” were not predicted with an acceptable precision and recall. Future work will aim for a detailed investigation of the circumstances under which these referential states occur, in order to improve the precision of their prediction. Referential state prediction is a logical complementary task to coreference resolution, and its potential as a preprocessing step deserves more attention.

To sum up, referential state prediction is possible, and ready to use for coarse-grained diachronic research, there is room for improvements and it seems likely to serve as sparring partner for coreference resolution.

5 References

[1] Bech, Kristin, and Eide, Kristine Gunn (2011) The annotation of morphology, syntax and information structure in a multilayered diachronic corpus. *Journal for language technology and computational linguistics* 26, (2), 13-24.

[2] Bies, Ann, Ferguson, Mark, Katz, Karen, and MacIntyre, Robert (1995) Bracketing guidelines for Treebank II style Penn Treebank project

[3] Daelemans, Walter, and Bosch, Antal van den (2005) Memory-based language processing (Cambridge University Press, 2005)

[4] Daelemans, Walter, Zavrel, Jakub, van der Sloot, Ko, and Bosch, Antal van den (2009) TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide ILK Technical Report 10-01.

[5] Gegg-Harrison, Whitney, and Byron, Donna K. 2004 Eliminating non-referring noun phrases from coreference resolution. In Proc. Proceedings of DAARC, pp. 21-26

[6] Haug, Dag T. T., Jøhndal, Marius L., Eckhoff, Hanne M., Welo, Eirik, Hertzenberg, Mari J. B., and Müth, Angelika (2009) Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages. *TAL* 50, (2), 17-45.

[7] Komen, Erwin R. (2009) CorpusStudio. Nijmegen: Radboud University Nijmegen.

[8] Komen, Erwin R. (2012) Coreferenced corpora for information structure research. In Tyrkkö, Jukka, Kilpiö, Matti, Nevalainen, Terttu, and Rissanen, Matti (eds.) *Outposts of Historical Corpus Linguistics: From*

the Helsinki Corpus to a Proliferation of Resources. (Studies in Variation, Contacts and Change in English 10). Helsinki, Finland: Research Unit for Variation, Contacts, and Change in English.

[9] Komen, Erwin R. (2013): Finding focus: a study of the historical development of focus in English. PhD dissertation, Radboud University Nijmegen

[10] Los, Bettelou, López-Couso, María José, and Meurman-Solin, Anneli (2012) On the interplay of syntax and information structure. In Meurman-Solin, Anneli, López-Couso, María José, and Los, Bettelou (eds.) *Information structure and syntactic change in the history of English*, pp. 3-18. New York: Oxford University Press.

[11] Pintzuk, Susan, and Taylor, Ann (2006) The loss of OV order in the history of English. In van Kemenade, Ans, and Los, Bettelou (eds.) *The Blackwell Handbook of the History of English*. Oxford: Blackwell.

[12] Poon, Hoifung, and Domingos, Pedro 2008 Joint unsupervised coreference resolution with Markov logic. In Proc. EMNLP '08 Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 650-659

[13] Raghunathan, Karthik, Lee, Heeyoung, Rangarajan, Sudarshan, Chambers, Nathanael, Surdeanu, Mihai, Jurafsky, Dan, and Manning, Christopher 2010 A multi-pass sieve for coreference resolution. In Proc. EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 492-501

[14] Taylor, Ann (2003) The York-Toronto-Helsinki Parsed Corpus of Old English Prose *Syntactic Annotation Reference Manual* University of York.

[15] Uryupina, Olga (2009) Detecting anaphoricity and antecedenthood for coreference resolution *Procesamiento del lenguaje natural. N. 42 (marzo 2009)*, pp. 113-120. Jaén: Sociedad Española para el Procesamiento del Lenguaje Natural.

[16] van Kemenade, Ans, and Milicev, Tanja (2012) Syntax and discourse in Old English and Middle English word order. In Jonas, Dianne, Garrett, Andrew, and Whitman, John (eds.) *Grammatical Change: Origins, Nature, Outcomes*, pp. 239-254. Oxford: Oxford University Press.

Abbreviations, fragmentary words, formulaic language: treebanking medieval charter material

Timo Korkiakangas – Matti Lassila

University of Helsinki – University of Tampere

Abstract

This article proposes a method that makes possible the linguistic study of textually difficult hand-written materials which are imperfectly preserved. These materials include medieval manuscripts, letters, and legal as well as private documents. With these, the normal treebanking procedure is not sufficient. We present the case of medieval Latin charter texts, i.e., private documents, that 1) are partly fragmentary and 2) exhibit massive use of abbreviations, e.g., *chartul* for *chartulam* 'charter'. In addition, 3) charter texts are highly formulaic and display passages that differ from each other in their language use. It is not possible to ascertain the inflexional endings of most of the fragmentary and abbreviated words, so a method of excluding them from morphological (but not from syntactic) analysis is needed. Moreover, due to the varying degree of formulaicity in certain parts of charter texts, the language of these parts must be studied separately. Therefore, a method of merging two XML layers is introduced. One layer that contains lemmatic, morphological, and syntactic analysis according to the Perseus Latin Dependency Treebank standard is aligned with the other layer that contains textual information (abbreviations, fragmentary words, diplomatic segmentation).

1 Encoding textual data in treebanks

Before the invention of printing, all texts were written by hand. In the Middle Ages, the period through which, for example, all the Classical Latin literature was transmitted, it became more and more customary to abbreviate certain common Latin words or inflexional endings. The scribes wrote, for example, *dns* for *dominus* 'lord' or *chartulā*, with a small horizontal stroke over the final *a*, for *chartulam* 'charter', where the abbreviated *-m* stands for an accusative ending. The practice of abbreviating poses limitations to how the abbreviated words can be used in linguistic analysis. If correctly applied, linguistic analysis of medieval Latin texts can tell us, for example, how well the scribes managed their Latin, which traits of spoken language infiltrated the written code, and whether there was regional variation.

Along with the abbreviations, the state of preservation of the physical object, on which the text is written, may affect linguistic study. Both the medieval literary texts and the texts that were written for practical purpose,

such as charters and other legal documents, survive in parchment manuscripts (codices or independent sheets) that have often suffered various damages. The writing may have been deteriorated by attrition or by humidity. The illegible letters, words, or lines set important restrictions to any linguistic study concerning spelling, morphology, and, in many cases, also syntax. Especially morphological study is impossible with words that lack their case-endings either due to a hole in the parchment or due to the ending being abbreviated by an ambiguous abbreviation (e.g. *not* for *notarius* (nominative) or *notarium* (accusative) or *notarii* (genitive) 'notary', etc.). It is obvious that treebanking should be able to manage these phenomena. How fragmentation and ambiguous abbreviations can be dealt with, is described in section 3.

In addition to abbreviations and fragmentary parts, it is often important for linguistic study to treat different discursive units separately. For example, prefaces of books are often written in a style and language different from those used in the other parts of the book, or a medieval prayer book may contain practical instructions between different types of prayers. Similarly, medieval charters have various diplomatic parts (e.g. beginning formula, contract text proper, subscriptions) that display different linguistic realities. Treebanking has to be able to cope with this internal linguistic division of texts. The role of the diplomatic parts in charters is discussed in section 4.

The aim of the present paper is to introduce one practical method for individual scholars or projects of managing this kind of material with the above-described variation within one well-known treebanking framework, namely that of the Perseus Latin Dependency Treebank (LDT).[1] The method presented in this paper is meant for charter Latin, but it is possible to apply the idea to the linguistic study of whichever material that displays similar textual features. We, indeed, suggest that our method ought to be adopted, with necessary adaptations, by all the future treebanking projects concerning 1) texts that were originally hand-written (mainly in the Middle Ages), 2) texts that contain various text types, and 3) edited texts that, in general, have been transmitted through a textual history and, consequently, represent a compromise between the readings of several manuscripts.

What makes this suggestion even more noteworthy is the immense amount of this kind of material that will become available to linguistic analysis in the near future, thanks to the numerous ongoing digital humanities

[1] The Perseus Latin and Ancient Greek Dependency Treebanks are a project aimed at treebanking texts in standard Latin and Greek. The Perseus Project is hosted at Tufts University, Medford, USA (http://nlp.perseus.tufts.edu/syntax/treebank/index.html). Also other workstations, such as Anastasia (http://anastasia.sourceforge.net/index.html), are available for digital philology nowadays, but we found the one provided by LDT the best suited for our purposes because of its clarity and flexibility.

projects. In the long run, serious linguistic study cannot be based on texts with no textual information provided. For example, LDT will probably want to add to its treebanks of edited Latin and Greek texts the textual information on various readings of manuscript witnesses that is provided in the critical apparatuses of those editions. An important initiative of annotating text re-uses of fragmentary authors with CTS and CITE URN syntax has, indeed, been launched by the Perseus Project (Almas and Berti 2013 [1]).

2 The LLCT treebank

In our study of medieval charter Latin, it is necessary to take into account all the above-mentioned textual factors (abbreviations, fragmentary words, and diplomatic parts) when constructing and using the treebank. Our treebank, the Late Latin Charter Treebank (LLCT), is built on a corpus of 519 Tuscan private charters from the 8th and 9th centuries.[2] The corpus consists of 198,700 words. Because of the elevated number of ambiguously abbreviated (13,134) and fragmentary words (1,523), the actual number of linguistically analyzable words falls to 184,043. In the following sections, we explain how the abbreviated and fragmentary words are defined and sorted out.

In our corpus, the linguistic annotation is realized using the tools provided by the Perseus and Alpheios Projects. The annotation follows the *Guidelines for the Syntactic Annotation of Latin Treebanks* launched by LDT and by the Index Thomisticus Treebank (Bamman et al. 2007 [2]).[3] The LDT standard is based on Dependency Grammar, successfully applied to the analytical layer of linguistic annotation in the pioneering Prague Dependency Treebank of Czech language (Hajič 1998 [4]).

In the Perseus annotation environment, the lemmatic and morphological analyses are entered in a table editor where each word form is given an appropriate lemma and a nine-place morphological tagset, which encodes part of speech proper, person, number, tense, mood, voice, gender, case, and degree. The syntactic annotation comprises syntactic tags (e.g. PRED, SBJ, OBJ, ATR, ADV) and head-dependent relations that are defined in the Alpheios Treebank Editor (Bamman et al. 2007 [2]). Both the table editor and the Alpheios Treebank Editor have a graphical user interface and both data

[2] For the composition of the corpus, see Korkiakangas and Passarotti [5] 2011.

[3] The Alpheios Project maintains a set of digital humanities tools intended for learning and reading classical languages (http://alpheios.net/). The Index Thomisticus Treebank is a project aimed at the syntactic annotation of the Index Thomisticus, a morphologically annotated corpus of the texts of St. Thomas Aquinas. The project is hosted at the Catholic University of the Sacred Heart, Milan, Italy (http://itreebank.marginalia.it/).

outputs are saved as attributes of a single *word* element for each word in treebank XML files.

In LLCT, the linguistic annotation of the treebank XML files is linked as standoff markup with the TEI[4] XML files that contain the edited source text provided with the above-described textual information as inline annotations. How all this is done is explained in section 5. Section 6 presents two case studies that illustrate the relevance of the proposed method.

3 Abbreviations and fragmentary words

In the following, we explain which abbreviated and fragmentary words or passages have to be eliminated from the morphological analysis and how we carry this out. This specific procedure is required by our charter corpus, but the researchers of, say, Classical Latin texts can apply the same principles in order to exclude corrupt or dubious readings from the analysis. The first of the following excerpts presents a passage of plain text from a donation charter written in Populonia, Tuscany, in 778. We have expanded the abbreviations and restored the fragmentary words.

(a) *In nomine Domini Dei et Salvatori nostri Iesu Christi regnante domno nostro Carulo rex Francorum seo et Langubardorum*, *anno regni eius in Etalia quinto*, *Kalendis Septembre*, *in natale sancti Reguli*, *indictione prima feliciter*. (MED 172)

"In the name of Lord God and our Saviour, Jesus Christ, under the reign of our lord Charles, king of the Franks and Lombards, in the fifth year of his reign in Italy, on the Kalends of September, in the feast of Saint Regulus, during the first indiction, under good auspices."

In the second version, the abbreviated letters are marked with round brackets and the fragmentary letters with square brackets:

(b) *In n(omine) D(omi)ni D(e)i et Salvatori n(ostr)i Iesu Christi regnante d(om)n(o) n(ostr)o Carulo rex Francor(um) seo et Langubard[orum]*, *anno regni eius in Etalia quinto*, *Kal(endis) Septembre*, *in natale s(an)c(t)i Reguli*, *ind(ictione) prima f(e)l(iciter)*. (MED 172)

[4] TEI stands for the Text Encoding Initiative, a consortium that delivers a set of guidelines which specify encoding methods for machine-readable texts in XML format, chiefly in the humanities, social sciences, and linguistics (http://www.tei-c.org/index.xml). LDT employs the P4 release of the TEI Guidelines.

It turns out that version (a) conceals a considerable amount of editorial activity behind it. All the bracketed words in version (b) appear frequently in our charters and their function is well-known. Therefore, expanding the abbreviations and reconstructing the fragmentary words is justified. In all the above cases, there is no doubt about the lemma: e.g., *n* is known to represent the word *nomen* 'name' in this context and *Langubard* is most probably a residue of the genitive plural form *Langobardorum* of the lemma *Langobardus* 'Lombard'. Some of these words are, however, unreliable in regard to their endings. These words are found underlined in version (b).

Since Latin encodes the morphological information in the final syllables of words, the abbreviated words whose inflexional morphemes are not written out and cannot be reliably expanded (e.g. *n*(*omine*) or *ind*(*ictione*)) cannot be linguistically analyzed in a reliable way: *n* can also stand, for example, for *nomen*, *dn* for *domni* 'of lord', *Kal* for *Kalendas* 'Kalends', and *ind* for *indictionem* 'indiction'. In contrast, certain other abbreviations can be reliably expanded: one can be sure that the frequent abbreviation *dni* with the genitive ending *-i* stands for the genitive form *Domini* 'of Lord' (or *domni* 'of lord', depending on the context); similarly, it is evident that the specific word-final ligature consisting of the letter *R* and a small curve stands for the genitive plural ending *-rum*, as in *Francor*(*um*) 'of the Franks' (Bischoff 1990 [3], 150–154). To sum up, certain abbreviations can be reliably expanded while certain others cannot. This is because certain abbreviations are unambiguous by interpretation while the others can abbreviate several inflexional forms of a single word. From the perspective of standard Latin grammar, the context might often seem to justify reconstructing even the ambiguously abbreviated endings but, in the case of early medieval Latin, one cannot be sure which ending the scribe really had in mind.

When we edited the TEI XML text for the treebank, we started by expanding all the words that appeared abbreviated in the original charters. This was possible because we did not aspire to create a diplomatic edition. We only needed to know which abbreviations were ambiguous and which not.[5] Therefore, we decided to annotate with *expan* tags those words where the inflexional ending was involved in the abbreviation and the expansion is, thus, unreliable. For example, abbreviation *Kal* can be expanded in several ways depending on the context, e.g., *Kalendis* or *Kalendas*. In contexts, such as *Kal Februariis*, where the ablative plural ending *-is* is found in the adjective *Februariis*, we expand *Kalendis Februariis* 'on the Kalends of

[5] Approximately 45% of the abbreviations that involve the case ending cannot be reliably expanded. Respectively, some 16,000 of all the 184,043 linguistically analyzable words included in the TEI XML edition derive from unambiguous abbreviations, e.g. *dni* for *domni*.

February'. Nevertheless, since the interpretation of the case ending of *Kal* is always unreliable, we exclude the whole word *Kalendis* from the morphological analysis by *expan* tag. In version (c), the abbreviated and fragmentary words to be excluded from the morphological analysis are marked with the TEI tags *expan* and *damage*, respectively:

> (c) *In* <expan>*nomine*</expan> *Domini Dei et Salvatori nostri Iesu Christi regnante* <expan>*domno*</expan> *nostro Carulo rex Francorum seo et* <damage>*Langubardorum*</damage>, *anno regni eius in Etalia quinto*, <expan>*Kalendis*</expan> *Septembre*, *in natale sancti Reguli*, <expan>*indictione*</expan> *prima feliciter.* (MED 172)

In spite of their exclusion from the morphological analysis, the expanded abbreviations of this type can usually be fully acknowledged in lemmatic and syntactic analyses, provided that there is no possibility of mistaking the lemma and/or syntactic function. As stated above, in the present context *n* cannot be but an instance of a word form derived from lemma *nomen* and, similarly, *Kal* comes surely from lemma *Kalendae*. In the major part of cases, there is no doubt about the syntactic function and the head-dependent relation of the expanded words: in phrase *in n*(*omine*) *D*(*omi*)*ni*, word *n*(*omine*) is a complement to the preposition *in* for sure, as well as, in the sample passage of MED 172, *Kal* is an adverb dependent on the governing participle *regnante* of the ablative absolute construction. Thus, both these words occupy their positions in the dependency structure although they are excluded from morphological analysis.

As for fragmentary words, all the words with an illegible inflexional ending are, of course, excluded from morphological analysis by *damage* tags even though the beginning of the word would be completely preserved and the ending deducible with a good probability (e.g. *Langubard*[*orum*]). Those fragmentary words that preserve their inflexional endings have been included in morphological analysis on the condition that their lemmas and syntactic functions are known. This is often the case, as the predictability of the formulaic charter language usually helps to deduce the missing letters.

If only the lemma and the syntactic function of a fragmentary word, whether its ending be legible or not, are known, the word in question can serve as an element in the syntactic dependency structure of the sentence. This practice of including both fully and partly restorable fragmentary words into dependency structures is adopted in order to ensure the maximal usability of the dependency network of sentences, even in fragmentarily preserved charters.

4 Diplomatic parts

A distinctive feature of legal documents has always been their tendency towards standardization of both contents and language. This leads to the stabilization of the regular, repetitive parts of documents into formulae. In the 8th and 9th century Italy, the formulae were produced by memory or by comparison with earlier charters at hand, instead of copying them formularies (Schiaparelli 1933 [8], 3). Writing from memory, of course, resulted in a considerable amount of variation which is of linguistic interest. Even more interesting, from the linguistic point of view, are the non-formulaic or so-called free parts of charters that contain the case-specific details of the legal act that made the composition of the charter necessary.

Diplomatics, i.e., textual analysis of historical documents, calls the most important free part *dispositio*. The *dispositio* states the transferred property, the measures and boundaries of the plot of land, or the sum of money paid. Generally speaking, the free parts, such as *dispositio*, are the only ones in which the scribe could not rely on formulae but had to improvise. As a consequence, the free parts are usually written in a relatively simple language that reflects the developments of spoken language, which, at the time, had diverged far from standard Latin. By standard Latin, we mean an established written variety based on the morphology and vocabulary of Classical Latin. The formulaic parts, instead, abound with errors and contaminations arising from poor command of standard and from misinterpretations of the complex juridical phraseology (Larson 2000 [6], 152–153).

It is important not to treat equally the formulaic and the less formulaic parts in linguistic analysis. They represent different linguistic realities and, thus, display different types of errors. To take this difference into account, we complement the textual annotation, presented in section 3, with diplomatic text type segmentation. This annotation indicates the diplomatic division of the charter by marking the non-formulaic parts of the document by *seg type*="*free*" tags while the formulaic parts are left without tagging.

To further sophisticate the analysis of text types, we tag by *seg type*="*subs*" the autograph subscriptions of witnesses who knew how to write, but did not necessarily have sufficient experience in it. This latter category is, however, of minor importance, as the subscriptions reproduce almost verbatim the few allowed subscription formulae. With this procedure of diplomatic annotation, it is possible to analyze the language of the non-formulaic and formulaic parts separately from each other and, respectively, the language of the subscriptions and the charter text proper separately from

each other. These are the two most important text type dichotomies that appear in our charters.

The same method can be used in several other cases as well. For example, a student of Cicero's rhetorical techniques might be interested in investigating how the author's morphological and syntactic choices vary in different rhetorical parts of his speeches, such as *exordium*, *narratio*, *partitio*, *confirmatio*, *refutatio*, and *peroratio*. The researcher would then segment the speeches by those parts he or she finds the most revealing concerning the analysis.

The next section presents a practical method of aligning the above-discussed textual markup categories with the linguistic annotation.

5 Technical realization

The following steps describe the workflow required for merging the LDT treebank markup (standoff annotation of lemma, morphology, and syntactic dependencies) with the LLCT-specific markup (inline annotation of abbreviations, damaged words, and selected diplomatic parts) of TEI XML files. A more detailed documentation of the process is available in the transformation script files. The complete scripts can be found online at https://github.com/mjlassila/linguistic-annotation-merger. First, both the treebank and TEI XML files are saved in a BaseX XML database (http://basex.org).

1) Run the alignannotations-latin XQuery transformation scenario to merge the two annotation layers through XPointers (by courtesy of Bridget Almas, Perseus Digital Library Project, Tufts University, USA). The XPointers link the sentence and word identifier numbers of the treebank file with the corresponding words in the TEI XML file.
2) Convert the segmentation annotation (tags <seg type="subs"> and <seg type="free">) and the status annotation (<damage>, <expan>) into XML attributes for each *word* element inside the XPointer file (02-convert-elements-to-attributes.xq).
3) Merge the newly created segmentation and status attributes of the XPointer file with each *word* element of the treebank file (03-merge-attributes-with-treebank-data.xq).
4) Transform the merged treebank file into PML format by using the XSL transformation style sheet (by courtesy of Francesco Mambrini, Harvard University, USA) (04-run-aldt-to-pml-transformation). The style sheet is modified in order to include the custom attributes

(segmentation and status information) in the resulting file. The custom attributes are also included in the TrEd Tree Editor schema.

The Prague Markup Language (PML) is a data format developed for the Prague Dependency Treebank. To query our treebank, we use the PML Tree Query, which is a query language and search engine that can be used as a plug-in to the TrEd Tree Editor (http://ufal.mff.cuni.cz/tred/). The PML Tree Query is aimed at querying multi-layer annotated treebanks (Štěpánek and Pajas 2010 [9], 1828–1830).

The first advantage of our approach is that it is based on a set of pre-existing open-source tools and well-established services that can be flexibly modified and combined. The second advantage is that our method can be applied to handling an infinite number of new, study-specific annotations with minor modifications of the transformation scripts. The method can be, thus, adapted to a wide range of projects on the field of textual studies, concerning both linguistics and the humanities in general.

6 Two case studies

In this section we show, through two case studies, how important it is to combine textual data with linguistic annotation in treebanks.

The first study is about distinguishing the diplomatic parts of the charters. In Classical Latin, adnominal possession was most often encoded by the genitive case, e.g., *actor reginae* 'procurator of the queen'. Also other related functions, such as being part of something, e.g., *portio terrae* 'portion of land', were expressed by genitive. Prepositional constructions, such as *unus de illis* 'one of them', were sometimes used in partitive functions. In the course of time, the prepositional construction with *de* 'of, from' began to compete with the genitive case in all possessive functions. Along with the loss of case inflexion, *de* finally ousted the original genitival construction. The modern Romance languages, except Romanian, have only the prepositional construction based on *de*: e.g., It. *rettore della chiesa* 'rector of the church' or Fr. *pièce de terre* 'piece of land' (Väänänen 1981 [10], 114).

In the 8th and 9th century Italy, this evolutionary process must have been well on the way in spoken language, which is the locus of linguistic change. As written texts are usually conservative, the innovations of spoken language surface only accidentally in them. From this perspective, it is interesting to observe the figures in table 1.

		N	%	total of all words
all parts together		977	0.53	184,043
diplomatic division	free parts	608	1.28	47,353
	formulaic parts	365	0.29	125,889
	subscriptions	4	–	10,801

Table 1: Nouns governing PP's headed by preposition *de*

Nouns governing a prepositional phrase with *de*, e.g., *portio de terra* 'portion of land', are 977, which are equivalent to 0.53% of the total number of words in our treebank. If the different diplomatic parts are taken into account, it becomes immediately clear that the 0.53% does not tell the whole truth. The percentage of the *de* constructions is more than four times higher in the free parts (1.28%) than in the formulaic parts (0.29%). The percentages are small because the possessive construction is relatively infrequent. However, if the treebank is divided in smaller subsections, the same pattern is attested in each subsection (formulaic parts 0.24–0.80%, free parts 0.91–1.36%), which corroborates the above results. Without the help of diplomatic segmentation, it would have remained unnoticed that the prepositional possession construction was gaining ground mainly in the free parts, which are thought to reflect spoken language better than the formulaic parts. This is in line with the hypothesis that the *de* construction was already well established in the spoken language of the time.

In the second study, we examine the genitive case that continued to be used in charter Latin along with the ever increasing prepositional possessive constructions. If we count the 1st declension feminine genitive singular nouns, e.g., *terre* 'of land', in the normal way, with no attention to the abbreviated words that appear expanded in the edition, they total 1,733 occurrences in the whole treebank. When we rule out those abbreviations that cannot be reliably expanded, the number of occurrences falls to 804. Thus, the portion of ambiguously abbreviated genitive singulars (929) is 53.6% of the alleged total of occurrences. This is due to certain common abbreviations that are inflexionally ambiguous, e.g., *eccl*(*esie*) 'of church' or *b*(*one*) *m*(*emorie*) 'of blessed memory', an epithet of the deceased.

This case study clearly demonstrates what amount of information is lost and what amount of ambiguity introduced into linguistic analysis if no attention is paid to the abbreviations. Any morphological study involving the 1st declension feminine genitive singular risks going astray, unless the great percentage of ambiguous abbreviations is excluded from the calculation. The strength of our method is that it enables the exclusion of these ambiguous

abbreviations from morphological analysis though they can still be exploited as heads and dependents in lemmatic and syntactic analyses.

7 Conclusion

Linguists often come across historical material that does not allow normal treebanking, but requires a special treatment because of its particular textual properties. Most typically, these particularities are holes in parchment, abbreviations, and different discursive units featuring within a single text. This is the case also in our charter treebank.

The textually problematic words cannot be used in linguistic analysis in the same way as unambiguous words. Therefore, we have developed a method that manages these phenomena by using separated layers of textual inline annotation and linguistic standoff annotation, aligned with each other through XPointers. The method makes it possible

- to utilize well-supported open-source tools and services by combining them with each other;
- to eliminate fragmentary words and unreliably restorable abbreviations from morphological analysis, in which they would skew the results;
- to utilize the same fragmentary and abbreviated words in syntactic analysis, i.e., as nodes in the dependency structure, on condition that they are not mutilated to the extent that their lemma and syntactic function cannot be reliably deduced;
- to submit to linguistic study all the digital humanities material that is becoming available online with increasing speed, but that does not allow traditional corpus linguistic study because of various textual restrictions.

We believe that, in the future, providing treebanks with a vast range of textual information will be an everyday practice. This will greatly enhance the possibilities of linguistic study that nowadays has to content itself with texts that lack even the text-critical data available in printed editions. In this process, our method can be seen as a small but essential step forward.

Aknowledgements

We would like to thank the anonymous reviewers for their insightful comments and suggestions to improve the paper, as well as Bridget Almas and Francesco Mambrini, who gave us invaluable support with several technical issues, and Leena Enqvist, who proof-read the final version.

References

[1] Almas, Bridget and Berti, Monica (2013) The Linked Fragment: TEI and the Encoding of Text Re-uses of Lost Authors. In *The Linked TEI: Text Encoding in the Web*. TEI Conference and Members Meeting 2013.

[2] Bamman, David, Passarotti, Marco, Crane, Gregory and Raynaud, Savine (2007) *Guidelines for the Syntactic Annotation of Latin Treebanks* (v. 1.3) (URL: http://nlp.perseus.tufts.edu/syntax/treebank/ ldt/1.5/docs/ guidelines.pdf).

[3] Bischoff, Bernhard (1990) *Latin palaeography: antiquity and the Middle Ages*. Translated by Dáibhí Ó Cróinín and David Ganz. Cambridge, New York: Cambridge University Press.

[4] Hajič, Jan (1998) Building a syntactically annotated corpus: The Prague Dependency Treebank. In Hajičová, Eva (ed.) *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*, pp. 12–19. Prague: Charles University Press.

[5] Korkiakangas, Timo, Passarotti, Marco (2011) Challenges in Annotating Medieval Latin Charters. In Francesco Mambrini et al. (eds.) *Proceedings of the ACRH Workshop*, Heidelberg, 2012, 103–114.

[6] Larson, Pär (2000) Tra linguistica e fonti diplomatiche: quello che le carte dicono e non dicono. In Herman, József and Marinetti, Anna (eds.) *La preistoria dell'italiano*, pp. 151–166. Tübingen: Niemeyer.

[7] MED = Barsocchini, Domenico (1837) *Memorie e documenti per servire all'istoria del Ducato di Lucca*. Tomus 5, volume 2. Lucca.

[8] Schiaparelli, Luigi (1933) Note diplomatiche sulle carte longobarde, II: Tracce di antichi formulari nelle carte longobarde. In *Archivio storico italiano* 19, pp. 3–34.

[9] Štěpánek, Jan and Pajas, Petr (2010) Querying Diverse Treebanks in a Uniform Way. In Nicoletta Calzolari et al. (eds.) *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2010, pp. 1828–35. European Language Resources Association.

[10] Väänänen, Veikko (1981) *Introduction au latin vulgaire*. Paris: Éditions Klincksieck.

Discussing best practices for the annotation of Twitter microtext

Ines Rehbein, Emiel Visser
and Nadine Lestmann
E-mail: irehbein|visser|lestmann@uni-potsdam.de

Abstract

This paper contributes to the discussion on best practices for the syntactic analysis of non-canonical language, focusing on Twitter microtext. We present an annotation experiment where we test an existing POS tagset, the Stuttgart-Tübingen Tagset (STTS), with respect to its applicability for annotating new text from the social media, in particular from Twitter microblogs. We discuss different tagset extensions proposed in the literature and test our extended tagset on a set of 506 tweets (7.418 tokens) where we achieve an inter-annotator agreement for two human annotators in the range of 92.7 to 94.4 (κ). Our error analysis shows that especially the annotation of Twitter-specific phenomena such as hashtags and at-mentions causes disagreements between the human annotators. Following up on this, we provide a discussion of the different uses of the @- and #-marker in Twitter and argue against analysing both on the POS level by means of an at-mention or hashtag label. Instead, we sketch a syntactic analysis which describes these phenomena by means of syntactic categories and grammatical functions.

1 Introduction

Through the emergence of new technologies, human communication practices have undergone radical changes (examples are communication by email, chat, text messages or Twitter microblogs). New text types from the web, in particular from the social media, challenge traditional views on the distinction between orality and literacy [9, 8] by combining features of both, oral and written communication. Our interest is in understanding these changes and in investigating the properties of these newly emerging text types. For this undertaking, linguistically annotated corpora would be of great help.

However, most annotation schemes for annotating part-of-speech (POS) tags and syntax have been developed for canonical written text (often from the newspaper domain), and it is not clear whether they also allow us to adequately describe the properties of user-generated content from the web.

Furthermore, recent work on POS tagging Twitter data has shown a low agreement of human annotators on tweets, yielding inter-annotator agreement (IAA)

scores in the range of 92-93% while the same scores on canonical, written text are in the high nineties [5]. This is highly problematic for the development of automatic, supervised methods for POS annotation of CMC, as those rely on the quality of the manually annotated data for training and evaluation, which will provide an upper bound on the performance of automatic methods.[1] Thus, to improve the quality of automatic POS tagging of CMC, we need linguistically sound annotation schemes which can be applied with high reliability by the human coders, and which provide a meaningful analysis of the phenomena of CMC.

In the paper, we present an annotation experiment where we assign parts-of-speech from the Stuttgart-Tübingen Tag Set (STTS) [15] to German microtext from Twitter, asking the following questions:

- What are the main problems for analysing computer-mediated communication (CMC) on the POS level, using an annotation scheme developed for canonical written language?
- How reliable are human annotations of POS on social media text?

We report on inter-annotator agreement results obtained for POS tagging German tweets and discuss the reasons for the lower agreement obtained on Twitter microtext as compared to, e.g., newspaper text. Based on our annotation study, we address the main problems encountered during the annotation and propose a different approach, which, in our opinion, is more promising to yield a reliable and adequate analysis of social media data. In particular, we focus on Twitter-specific phenomena like the @- and #-marker.[2] We illustrate that both have multiple functions and argue against an analysis of references and linking information by means of an at-mention and hashtag label on the POS level, as proposed in the literature [14, 6, 3]. Instead, we advocate for encoding this type of information on the syntactic level and sketch a possible solution.

The paper is structured as follows. In section 2, we review related work on extending POS tagsets for the annotation of Twitter microtext. Section 3 presents our annotation experiment and reports on our inter-annotator agreement for human annotation of POS on German tweets. In section 4 we illustrate the problems we encountered during the annotation and discuss different solutions. We conclude in section 5.

2 Related work

There is some recent work on developing or extending POS tagsets for annotating Twitter microtext. Ritter et al. [14] expand the Penn treebank POS tagset by adding

[1] The relatively low IAA scores for POS-tagging microtext also put into question results for semi-supervised and unsupervised POS tagging as those are evaluated against a (less accurate) handcrafted goldstandard.

[2] The at-mentions (@) identify the addressee of the tweet and provide a link to the users' Twitter profile while the hashtags (#) function as semantic tags or keywords.

four new, unambiguous tags for hyperlinks, user names, hashtags and retweets and manually annotate a testset of 800 English tweets.[3] They do not give numbers for inter-annotator agreement of the annotations.

Gimpel et al. [6] annotate English tweets using a coarse-grained tagset (25 tags) with five new tags for CMC-specific phenomena. These include emoticons, hyperlinks, hashtags (linking the tweet to a semantic category), at-mentions (indicating the recipient of the tweet) and a tag for annotating *'RT'* and *':'* in retweet constructions. In contrast to [14], Gimpel et al. [6] annotate tokens as hashtags only when they are not integrated in the tweet message. Syntactically integrated instances of hashtags are annotated according to their distribution. They report an inter-annotator agreement of 0.914 on a small testset of 72 tweets.

Avontuur et al. [1] annotate Dutch tweets, using a hierarchical morpho-syntactic tagset with 320 tags, based on a tagset developed for written text, and the five new, Twitter-specific tags from [6]. They obtain an inter-annotator agreement in the range of 0.912 to 0.933 (Cohen's κ) on a testset of 1,056 tweets.

In all three studies, the agreement for human annotations on Twitter is in the same range, and substantially lower than the one obtained on canonical, written text (see, e.g., Brants [5] for IAA on German newspaper text).

3 Reliability of manual annotations on German tweets

In our annotation experiment, we use the 54 tags of the Stuttgart-Tübingen Tag Set (STTS) [15] to annotate German tweets. We follow the proposals above and also introduce new tags for annotating emoticons, hashtags, at-mentions and hyperlinks. Similar to [6], we only annotate tokens as hashtags when they are not integrated in the tweet. In contrast to [6], we do the same with at-mentions and hyperlinks (also see section 4). As we are interested in investigating conceptual orality in a written register [8], we also add new tags for discourse phenomena such as filled pauses, question tags or backchannel signals from an extension of the STTS developed for annotating spoken language [12] (for details see section 4.1).

A prominent feature of CMC taken from spoken language is the contraction of individual lexical units into a new form, inspired by their pronunciation in spoken discourse. We do not correct these non-canonical tokenisations but follow the approach of Gimpel et al. [6] and use combinations of POS tags to annotate the contracted word forms, as shown in Example (1).[4]

In the experiment, we annotated German tweets which we collected from Twitter over a time period from July 2012 to February 2013, using the Python Tweepy module[5] as an interface to the Twitter Search API[6]. Our test set includes 506 tweets

[3] By "unambiguous" we mean that all tokens starting with an @ or # as well as all hyperlinks and emoticons are labelled with the corresponding tag, regardless of their distribution.

[4] The same approach is taken in the STTS for annotating merged prepositions (APPR) and definite determiners (ART) as APPRART, e.g. *in/APPR dem/ART* (in the) vs. *im/APPRART* (in_the).

[5] `http://pythonhosted.org/tweepy/html`

[6] `https://dev.twitter.com/docs/api/1/get/search`

(7,425 tokens) which were annotated independently by two human annotators. Table 1 shows our inter-annotator agreement on the data.

	# Tagset	# Testset	κ
this work	72	506 tweets	0.92.6 - 0.94.4
Gimpel et al. (2011)	25	72 tweets	0.914
Avontuur et al. (2012)	325	1056 tweets	0.912 - 0.933

Table 1: Inter-annotator agreement on German, English and Dutch tweets.

Our results are in line with other studies on inter-annotator agreement of English and Dutch Twitter data [6, 1]. Most interestingly, the size of the tagset does not seem to have a huge impact on the results. All three studies show an agreement well above 0.9 (κ), despite the different sizes of the three tagsets.

We now come to the question why IAA on Twitter microtext is so much lower than the one on canonical written text. Our error analysis shows that the most difficult decisions during the annotation concern the distinction between proper names (NE) and nouns (12% of all disagreements), NE and foreign language material (6.3%), NE and at-mentions (5.1%) and NE and hashtags (3.0%).

We take this as a starting point to have a closer look at the disagreements on CMC-specific phenomena and discuss these in more detail. We argue that the POS tags for at-mentions and hashtags do not provide an adequate description of the different functions of these markers and that their linking function should not be encoded on the POS level.

4 Twitter – the data

Communication on Twitter is shaped by a liberal use of orthographic rules where spelling conventions are often ignored and the capitalisation of German nouns is not done in a systematic way (1). In addition, German compound words are often split up into their components while, at the same time, individual lexical units are contracted into a new form, inspired by their pronunciation in spoken language.

(1) der **briten regierung hamse** doch ins gehirn geschissen und vergessen umzurühren
the British governm. have_they but in_the brain shat and forgotten to stir
"The British government got shit for brains"

4.1 Features from spoken language

Besides contractions, we find many other features imitating informal spoken language in a written medium. We annotate those using an extended version of the STTS developed for the annotation of spoken language phenomena [12].

One phenomenon is the use of disfluencies like repairs and filled pauses in Twitter which, considering that the communication is not subject to time-pressure caused by online processing and that the users have the possibility to revise and

edit their messages, is at least unexpected.[7] We assign filled pauses in Twitter the PTKFILL tag (2).

(2) On the road **äh**$_{PTKFILL}$ train **äh**$_{PTKFILL}$ also Ihr wisst schon :)
On the road uh train uh well you know already :)
"On the road uh train uh, well, you know :)

Private communication on Twitter is highly informal, which is shown by the high number of interjections, discourse markers and verbless sentences. Tweets are also highly interactive, as indicated by the frequent use of backchannel signals and question tags which we assign the labels PTKREZ (3) and PTKQU (4).

(3) @userA: yaa dann mach das soo @userB: **hmm**$_{PTKREZ}$ .... muss noch nachdenken !
@userA: yaa then do it like that @userB: hmm .... have to still think !
"@userA: Yeah, then do it like this @userB: hmm ... still have to think¡'

(4) geil , **wa**$_{PTKQU}$!? xD
cool , what !? xD
"Cool, isn't it?"

Other extensions cover the use of discourse-structuring particles, onomatopoeia and forms of echoism, unfinished words, and a new punctuation sign for marking abandoned utterances. These extensions to the STTS for annotating spoken language differ from the original STTS by way of being defined by discourse-pragmatic criteria instead of morpho-syntactic ones. It could be argued that these distinctions are hard to operationalise and thus should not be encoded on the POS level. For many NLP applications such as Information Retrieval or Named Entity Recognition, discourse particles do not seem to be relevant. For the comparative study of orality in spoken and written discourse, however, these particles can augment the corpus with useful information.[8]

4.2 CMC-specific features

4.2.1 Emoticons

A major drawback of the written medium is the lack of important channels of non-verbal communication such as mimics, prosody and stress. To make up for this, Twitter users adopt different techniques to express themselves. In addition to a frequent use of interjections and exclamative constructions, we observe the duplication of characters (5) to express emphasis, and the use of uppercased words to indicate shouting (6). Emoticons are another way of expressing emotion in CMC. We follow [14, 6] and introduce a new tag for the annotation of emoticons (EMO) (5),(6).

(5) **Awww** wie **süüüß *o***$_{EMO}$
Aw how sweet *o*

(6) Peinlich , aber **JA ! :-)**$_{EMO}$
embarrassing , but yes ! :-)

[7] See [13] for an analysis of the different functions of filled pauses in Twitter.

[8] Another area where this type of information might be useful is Sentiment Analysis/Opinion Mining.

4.2.2 Hyperlinks

Hyperlinks in tweets can be positioned either at the beginning or at the end of the tweet, linking the tweet to additional, external information (7), or can be syntactically integrated in the tweet (8). While [14, 6] use a new, unambiguous tag to annotate all hyperlinks, regardless of their distribution, we distinguish between external links (annotated as URL) and syntactically integrated ones (annotated as proper names).

(7) bei dem Wetter... **http://t.co/ywjSHuhK**
with the weather... http://t.co/ywjSHuhk
"In weather like this..."

(8) Hast du eventuell mal mit **http://t.co/EsNtqGku** verglichen ?
Have you maybe PTCL with http://t.co/EsNtqGku compared ?
"Have you compared it with http://t.co/EsNtqGku?"

The annotators' agreement on the distinction between integrated and non-integrated instances in our annotation study was quite low. This is mostly due to the non-systematic use of punctuation and capitalisation in Twitter which makes sentence segmentation difficult. While examples (7) and (8) are straightforward, in examples (9) and (10) it is less clear whether the hyperlinks are integrated or not.

(9) Neue monatliche Umfrage jetzt online auf unserer Homepage http://t.co/cvwiJTLA .
New monthly poll now online on our homepage http://t.co/cvwiJTLA .

(10) Und ich dachte schon, #Siri hätte mich nicht mehr lieb: http://t.co/Xclx -Siri ist toll
And I thought already, #Siri was REFL not still fond: http://t.co/Xclx -Siri is great
"And I already thought that #Siri wasn't fond of me any more: http://t.co... -Siri is great"

Given that hyperlinks are identifiers referring to objects in the world, we argue that it is appropriate to annotate all hyperlinks as proper names on the POS level. While we acknowledge that the linking information might be useful for some applications, we do not think that they justify the introduction of a new part of speech category but would rather shift this type of information to a different level, e.g. including it as a new Named Entity type and encoding it as part of the syntactic annotation, similar to the approach in the TüBa-D/Z [7] (release 8).

4.2.3 At-mentions

Originally, the @-sign has been used as an address marker to refer to the addressee of a tweet (or a chat message) (11), but is now also used in a number of other contexts and with different functions.

(11) **@Schebacca** ok warum ist das wichtig ???
@Schebacca ok why is that important ???
"@Schebacca Ok, why is that important?"

In (12), the @ occurs in isolation, separated by whitespaces, and is used as a local preposition.

(12) Rest des Tages dann Home-Office , vielleicht im Garten ? (@ Bahnhof Ansbach)
Rest of the day then home office , maybe in the garden ? (@ Bahnhof Ansbach)
”Home office for the rest of the day, maybe in the garden? (at Ansbach train station)

In contrast, the @ in (13) is not a token of its own but is contracted with a location name, Bad Hersfeld. There are two possible analyses here. First, we could assume that the @ again functions as a preposition and should be separated from the location name by the tokeniser. The second analysis opposes the first one by assuming that *Bad Hersfeld* is a post-modifying NP, and that the sole function of the @-sign is to provide a link to the profile of *Bad Hersfeld* (without having the explicit semantics of a local preposition).

(13) Danke an die ehem. Medusa Bar **@Bad Hersfeld**, top Leute und super Stimmung !
Thanks to the former Medusa Bar @Bad Hersfeld, great people and super atmosphere !

The second analysis is backed up by cases like (14), where the @ was merged with a proper name but does not license the reading as a preposition. The attempt to replace the @ with a preposition would even result in an ungrammatical utterance.

(14) ich folge ja nun der **@GrinseDame** ..
I follow PTCL now the @GrinseDame ..
“I now follow the @GrinningLady ..”

Examples (15) and (16) support our analysis by showing that the users do not conceptualise the @ as a preposition, but combine user names marked by @ with additional prepositions, which - if the first analysis for (13) was correct - should be redundant.

(15) Warum wird der scheiss tweet **an @mondmiri** nicht gesendet ???
Why is the shitty tweet to @mondmiri not sent ???
“Why hasn’t the shitty tweet to @mondmiri been sent ?”

(16) Wenn ich **bei @lidl** eine Stunde am Pfandautomaten warten muss geht es
When I at @lidl one hour at the deposit redemption machine wait must goes it
immer noch schneller als **im @kaufland**
always still quicker as in the @kaufland
“Even if I have to wait at @lidl for one hour in the queue for the deposit redemption machine, it’ll still be faster than at @kaufland”

In conclusion, we argue that the @ is used as an address marker or preposition only in some cases but has lost its original meaning in many others. We thus refrain from separating the @ from the following token and annotating it as a preposition or an address marker. Our main reason for being rather conservative about changing the tokenisation is that separating all @-signs from user or location names would result in a substantial increase in token numbers for CMC corpora, thus leading to an artificially higher type-token ratio (TTR) for CMC as compared to other types of text. This would lead to skewed results for comparative corpus studies of register variation using corpus-linguistic measures like the TTR, sentence length or measures of syntactic complexity (which are often based

on sentence length).[9]

4.2.4 Hashtags

Similar to hyperlinks and at-mentions, hashtags can be syntactically integrated in the tweet message (17), or can be positioned at the beginning or at the end of the tweet, as in (18). Hashtags can be used as keywords or semantic tags to categorise the tweet and thus allow users to search for other tweets of the same category.

(17) Jetzt **#Stromanbieter #vergleichen**
Now suppliers of electric energy compare
"Compare suppliers of electric energy now "

(18) " Spül [spiel] mir das Lied vom Tod " **#Spülwitze**
" Wash [play] me the song of death " #washing jokes

The function of hashtags, however, cannot be reduced to semantic tagging. They are frequently used to add an evaluation to the (otherwise neutral) tweet, as in (19).

(19) Laut meiner #wetterapp hat es 7 grad **#toocold**
As per my weather app has it 7 degrees #toocold
"According to my weather app we have 7 degrees #toocold"

They can also add relevant context information needed for understanding the message of a (highly underspecified) tweet, as in (20).

(20) Hey drückt @ich_seh_weiss um 12 die daumen :-) **#fahrprüfung**
Hey press @ich_seh_weiss at 12 the thumbs :-) #drivingtest
"Hey, fingers crossed for @I_see_white at 12 :-) #drivingtest"

Some tweets include nothing but a hashtag (21). These often serve as a statement about the general (emotional) state of mind of the Twitter user, in a highly compressed format.

(21) **#übermüdeteresistzufrühmeckertweet**
#overtired-it-is-too-early-rant-tweet

Twitter users are also highly register-aware. Sometimes hashtags are simply used because of this, as stated in the self-ironic tweet in (22).

(22) da fehlt noch #tweet #hashtag
there lacks still #tweet #hashtag
#wortedieichsowiesoschongeschriebenhabeimzweifelnochmalaufenglischalshashtaghinterher
#words-which-I-anyway-already-written-have-in-doubt-again-on-English-as-hashtag-afterwards
"The #tweet #hashtag is still missing here. #words-which-I've- already-written-anyway-when-in-doubt-then-I'll-add-them-in-English-to-the-end-of-the-tweet"

[9] For the English OCT27 data set [10], separating the @ would result in a seemingly higher token number of 27,896 as compared to 26,594 tokens in the original data set, and an additional segmentation of the # would further increase the number of tokens to 28,316.

Some hashtags include complex inflective constructions (23).[10]

(23) #mitfreu #superfreu #keinenekrophilenwitzemach
#with-you-rejoice$_{noninflected}$ #super-rejoice$_{noninfl}$ #no-necrophile-jokes-make$_{noninfl}$

Inflective constructions in CMC are often enclosed by asterisks or inequality signs, but users also encode nouns, verb phrases or whole sentences in that way (24). In most cases, the so-marked constructions are not syntactically integrated in the tweet but function as meta-comments, adding information on the emotional state of the user and her environment, or set the stage (25), similar to stage directions in a screenplay.

(24) *Vorfreude* / *kaffeetasseheb* / *hat schokolade gefunden*
anticipation / *coffee_cup_lift$_{noninflected}$* / *has chocolate found*

(25) *Trommelwirbel* / *an dieser Stelle bitte fröhliches Pfeifen einblenden*
drum roll / *at this point please jolly whizzling fade in*

Less frequent, but nonetheless existent, are instances of noninflected forms which are syntactically integrated, as shown in (26), (27). These challenge the analysis of noninflected verbs as independent interactive units [2, 3], classified in the same category as interjections, answer particles, emoticons and user names, and support an analysis which integrates non-inflected verb forms in the verbal paradigm.

(26) Jetzt mal erst so *tür aufmach* und dann *rausgeh*
now PTCL first so *door open$_{noninflected}$* and then *step$_{noninflected}$_out*
"*opening door* now and *stepping out*"

(27) dafür *zitter* und *dick einmumm*
instead shiver$_{noninflected}$ and *thick wrap_up_well$_{noninflected}$*
"instead shivering and wrapping myself up well"

Complex inflective constructions, on the other hand, pose a major challenge for automatic POS tagging, as they are often written as one token, sometimes (but not always) separated by space or by hyphens. Depending on the way they have been transcribed, they will either be split up into individual tokens or will be treated as single unit by the tokeniser.

In previous work [11], we have annotated complex inflective constructions using the COMMENT label. This, however, is not sufficient to encode the rich information expressed by these units. Here we expand our analysis and argue that the components of these constructions should be tokenised and annotated as individual units on the sub-token level. Figure 1 displays the syntactic structure of the inflective construction and the complex hashtag in (28), neither of which is syntactically integrated in the tweet. On the token level, the inflective construction and the hashtag are both treated as one unit. A more detailed analysis of the internal

[10] Inflectives (non-inflected verb forms) are a frequent stylistic means in German comics and computer-mediated communication [16].

structure of the two constructions is given on a sub-token level, where the complex inflective construction and the hashtag are split up and analysed individually.

(28) @pillenknick Moin Hendryk , schon unterwegs ? ***Kaffeetasseheb***
@username Morning Hendryk , already on your way ? *coffee-cup-lift*
#nochimmerschläfrig
#still-tired

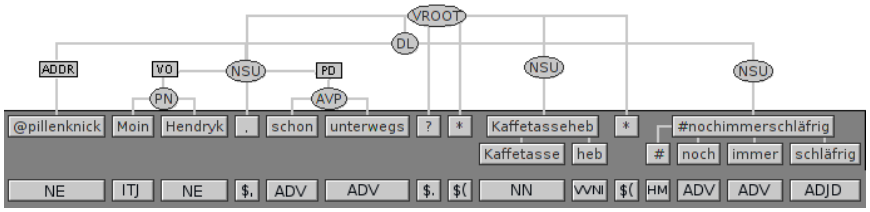


Figure 1: Analysis of complex inflectives and hashtags on the sub-token level

The syntactic analysis follows the annotation scheme of the TIGER treebank [4] as closely as possible. The DL (discourse level) node is the top node of the tweet. The user name (@pillenknick), referencing the addressee of the tweet, is marked by the new grammatical function label ADDR.[11] The actual tweet message (Moin Hendryk, schon unterwegs?) is governed by a NSU (non-sentential unit) node,[12] as are the inflective construction and the hashtag. We do not include the asterisks as part of the token, as they are not a necessary component of the inflective contruction, as opposed to the # for hashtags.

Figure 2 illustrates the representation of integrated inflectives on the POS level and in the syntax.

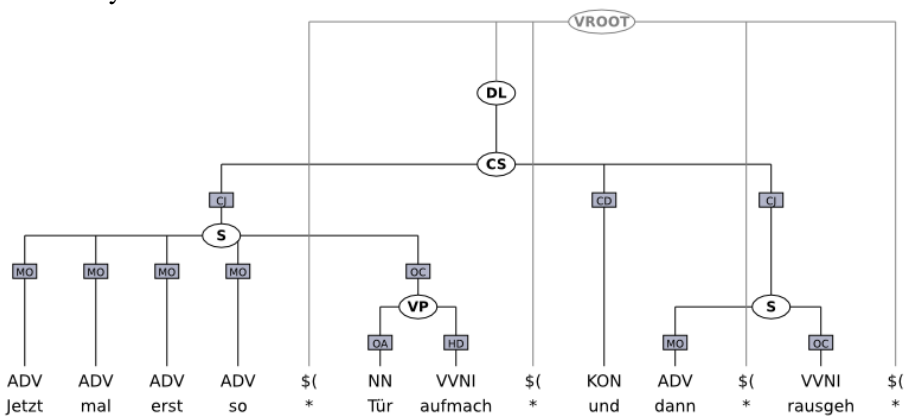


Figure 2: Analysis of integrated non-inflected verb forms

[11] Legend for figures 1 and 2: *Syntactic categories:* DL: discourse level, S: clause, CS: coordinated clause, NSU: non-sentential unit, PN: proper name, AVP: adverbial phrase, VP: verb phrase; *Grammatical functions:* ADDR: addressee, VO: vocative, PD: predicate, OP: prepositional object, CJ: conjunct, CD: coordinating conjunction, MO: modifier, OC: clausal object; *POS:* NE: proper name, ITJ: interjection, ADV: adverb, NN: noun, VVNI: non-inflected verb, HM: hashtag marker, $,: comma, $.: sentence-final punctuation, $(: sentence-internal punctuation.

[12] We distinguish between sentential and non-sentential units. Sentential units do include a finite verb while NSU nodes don't.

To sum up, hashtags not only provide a semantic classification of the tweets but also allow the users to express their emotions or comment on their physical condition or the state of the world in general. They do not correspond to one particular part of speech but can take the form of any arbitrary word or construction, of sentences even, depending on the creativity of the users. We thus argue that hashtags should not be annotated with a special hashtag label but should be analysed according to their distributional properties and internal structure.

5 Conclusions

This paper contributes to the discussion on best practices for the syntactic analysis of non-canonical language, focusing on Twitter microtext. We first presented an annotation experiment where we tested proposals from the literature for POS annotation of tweets and compared our inter-annotator agreement to related work. While our overall inter-annotator agreement was in line with, or even higher than, what has been reported in comparable studies, our error analysis showed that especially the annotation of Twitter-specific phenomena such as hashtags and at-mentions causes disagreements between human annotators. We argued that the new POS tags introduced to label user names and hyperlinks do not correspond to new grammatical part-of-speech categories. Accordingly, we advocate the annotation of user names and hyperlinks as proper names.

Furthermore, we discussed the multiple functions of the @- and #-sign in Twitter, showing that POS tags like AT-MENTION or HASHTAG fall short of capturing the information encoded by these phenomena. Instead, we sketched a possible way of annotating complex non-inflected constructions and hashtags in the syntax tree, providing a coarse-grained analysis on the token level and a more detailed one on the sub-token level.

References

[1] T. Avontuur, I. Balemans, L. Elshof, N. van Noord, and M. van Zaanen. Developing a part-of-speech tagger for Dutch tweets. *Computational Linguistics in the Netherlands Journal*, 2:34–51, 2012.

[2] M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, and A. Storrer. A TEI schema for the representation of computer-mediated communication. *Journal of the Text Encoding Initiative*, (3):1 – 31, 2012.

[3] M. Beißwenger, M. Ermakova, A. Geyken, L. Lemnitzer, and A. Storrer. DeRiK: A German reference corpus of computer-mediated communication. *Literary and Linguistic Computing*, 4(28):531–537, 2013.

[4] S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. The TIGER treebank. In *Proceedings of TLT*, Sozopol, Bulgaria, 2002.

[5] T. Brants. Inter-annotator agreement for a german newspaper corpus. In *Proceedings of LREC*, Athens, Greece, 2000.

[6] K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N.A. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of ACL*, Portland, Oregon, 2011.

[7] E. Hinrichs, S. Kübler, K. Naumann, H. Telljohann, and J. Trushkina. Recent developments in linguistic annotations of the TüBa-D/Z treebank. In *Proceedings of TLT*, Tübingen, Germany, 2004.

[8] P. Koch and W. Oesterreicher. Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36(85):15–43, 1985.

[9] W. Ong. *Orality and literacy: The technologizing of the word.* London; New York: Methuen, 1982.

[10] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel, N. Schneider, and N.A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*, Atlanta, Georgia, 2013.

[11] I. Rehbein. Fine-grained POS tagging of German tweets. In *Proceedings of GSCL*, Darmstadt, Germany, 2013.

[12] I. Rehbein and S. Schalowski. Extending the STTS for the annotation of spoken language. In *Proceedings of KONVENS 2012*, Vienna, Austria, 2012.

[13] I. Rehbein, S. Schalowski, N. Reinhold, and E. Visser. Ähm, äh... filled pauses in computer-mediated communication. Potsdam, Germany, 2012. Talk presented at the DGfS Workshop on "Modelling Non-Standardized Writing".

[14] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: an experimental study. In *Proceedings of EMNLP*, Edinburgh, United Kingdom, 2011.

[15] A. Schiller, S. Teufel, and C. Thielen. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report, IMS-CL, University Stuttgart, Germany, 1995.

[16] P. Schlobinski. *knuddel – zurueckknuddel – dich ganzdollknuddel*. Inflektive und Inflektivkonstruktionen im Deutschen. *Zeitschrift für germanistische Linguistik*, 29(2):192–218, 2001.

Le Roman de Flamenca: An Annotated Corpus of Old Occitan

Olga Scrivner, Sandra Kübler, Barbara Vance, Eric Beuerlein

Indiana University
{obscrivn,skuebler,bvance,ebeuerle}@indiana.edu

Abstract

This paper describes an ongoing effort to digitize and annotate the corpus of *Le Roman de Flamenca*, a 13th-century romance written in Old Occitan. The goal of this project is twofold: The first objective is to digitize one of the earliest editions of the text and to create an interactive online database that will allow parallel access to a glossary, to translations of verses, and to comments from Paul Meyer's edition. The second objective is to lemmatize and syntactically annotate the corpus and make it accessible using the ANNIS online-search engine.

1 Introduction

Le Roman de Flamenca holds a unique position in Provençal literature. "Flamenca est la création d'un homme d'esprit qui a voulu faire une oeuvre agréable où fût représentée dans ce qu'elle avait de plus brillant la vie des cours au XII[I] siècle. C'était un roman de moeurs contemporaines[1]" [14]. In the past, the 13th-century manuscript of Flamenca has been extensively studied in its raw text format. The potential value of this historical resource, however, is limited by the lack of an accessible digital format and linguistic annotation.

This paper focuses on the creation of an annotated corpus of Old Occitan that preserves one of the earliest editions of the manuscript [15]. While it was succeeded by many editions and translations, "no student of the manuscript can afford to overlook Meyer's editions" [2]. Thus, the purpose of this corpus is twofold. It is intended not only as material for linguistic research, but also to aid in broader studies. The first objective is to provide access to *Le Roman de Flamenca* through an interactive online database that will allow parallel access to a glossary, to translations of verses, and to comments from Paul Meyer's edition [15]. The second

[1] "Flamenca is the creation of a man of talent who wished to write an agreeable work representing the most brilliant aspects of courtly life in the twelfth century. It is a novel of manners" [3]. Note that elsewhere Meyer places Flamenca in the 13th century, a date which is also universally accepted today, and so the reference here to the 12th century must be in error.

objective is to lemmatize and syntactically annotate the corpus and to make it accessible using the ANNIS web-search engine.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the language Old Occitan and the romance of Flamenca. Section 3 outlines the structure, content, and annotation process of the corpus. Section 4 describes a range of applications of the corpus interfaces. Section 5 concludes and presents directions for future work.

2 Old Occitan and *Le Roman de Flamenca*

Old Occitan, formerly referred to as Old Provençal after one of its major dialects, is the ancestor of the endangered language spoken today in southern France by an undetermined number of bilingual individuals. This language constitutes an important element of the literary, linguistic, and cultural heritage of the Romance languages; it was known throughout the western medieval world through the lyric poetry of the Troubadours. While the historical importance of this language is indisputable, Occitan, as a language, remains linguistically understudied. In the past decade, a number of annotated corpora have been developed for other Medieval Romance languages, for example, Old Spanish [5], Old Portuguese [6], and Old French [13, 19]. However, annotated data for Old Occitan are still sparse. There exist (to our knowledge) two electronic databases, "The Concordance of Medieval Occitan"[2] [17] and "Provençal poetry"[3] [1], but users of those corpora are limited to lexical search.

This project focuses on the 13th-century Old Occitan romance *Le Roman de Flamenca*. The anonymous manuscript of *Le Roman de Flamenca* was accidentally discovered in Carcassonne (France) by Raynouard and was first fully edited and translated by P. Meyer in 1865. This romance has been variously characterized as a comedy of manners, "the first modern novel", and a psychological romance, among other characterizations [2, 3, 14]. This prose in verse played an influential role in the development of French literature [11]. Apart from a very intriguing love story between beautiful Flamenca, who is imprisoned in a tower by her jealous husband Archambaut, and the sharp-witted knight Guillem, this 8095-line story is a very interesting linguistic document and is the "universally acknowledged masterpiece of Old Occitan narrative" [8]. This romance features multiple literary styles, such as internal monologues, dialogues, and narratives, and offers a rich lexical, morphological, and syntactic representation of the language spoken in medieval southern France.

[2]The database includes Gschwind's edition [9] of *Le Roman de Flamenca*

[3]http://artfl-project.uchicago.edu/content/provençal

3 Structure of the Corpus *Le Roman de Flamenca*

3.1 Data

We are convinced that a digitized and annotated corpus of Old Occitan will be a valuable resource, not only for corpus linguistics studies, but also for a more general audience that wishes to become acquainted with Occitan literature. Therefore, one important goal is public accessibility. As a consequence, the corpus comprises only the editions that are not subject to copyright restrictions. Thus, we have selected the second edition of the manuscript by Meyer [15], supplemented with a glossary, translation of the manuscript into French, and a plot summary from the first edition by Meyer [14].

3.2 Workflow

The Romance of Flamenca [14, 15] is available in a scanned format, digitized by Google[4]. The manuscript consists of 8095 lines, footnotes, and 110 pages of glossary. As an initial step of text processing, digital images of the book are first sent through OCR and then manually corrected, using TESSERACT.[5] As a baseline, TESSERACT's Old French language model was used. With Old French, the OCR engine was then trained using sample images from *Flamenca*. However, it quickly became apparent that the glossary and character selection of Old French, although the closest match, is not ideal. Specifically, the Old French model contains numerous diacritics (e.g., é, û, ë) that are not used in Old Occitan. For better initial OCR results, these characters were disallowed during text recognition. In addition, some individual text strings were disallowed as they caused systematic problems. For example, without additional commands, the Old French search engine regularly recognized the Old Occitan word "anc" as "ane". Therefore, we disallowed "ane" and at the same time added "anc" to a list of additional words.

Figure 1 shows a scanned page from the Meyer's edited manuscript, which was then converted into a text format. We have created a TEI XML document, preserving verse lines, footnotes, paging, glossary, and comments by the author. We have also included the French translation provided by Paul Meyer in his first edition of the manuscript [14].

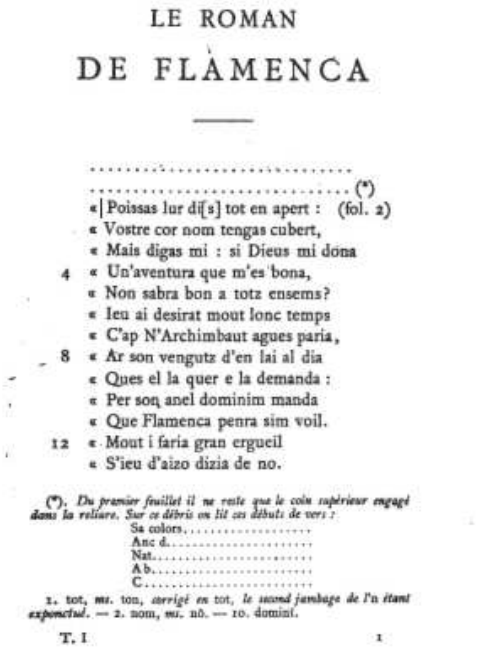
The second step of processing includes the following steps: 1) text tokenization, 2) POS tagging, 3) lemmatizing, 4) parsing, 5) converting the text into the format used by ANNIS [6], and 6) uploading it to the ANNIS database, for use with its graphical interface [20].

First, we have segmented our corpus into 52 200 lexically relevant tokens. In cases where the Meyer edition had a pronoun attached to a verb, negation, or other pronoun, we have detached them. For example, in (1a) the clitic *m* 'me' is attached

[4] http://www.archive.org/details/leromandeflamen00meyegoog

[5] http://tesseract-ocr.googlecode.com/svn/trunk/doc/tesseract.1.html

[6] http://korpling.german.hu-berlin.de/saltnpepper/



LE ROMAN

DE FLAMENCA

.....

..... (*)

« Poissas lur di[s] tot en apert : (fol. 2)
« Vostre cor nom tengas cubert,
« Mais digas mi : si Dieus mi dona
« Un'aventura que m'es bona,
« Non sabra bon a totz ensems?
« Ieu ai desirat mout lonc temps
« C'ap N'Archimbaut agues paria,
« Ar son vengutz d'en lai al dia
« Ques el la quer e la demanda :
« Per son anel dominim manda
« Que Flamenca penra sim voil.
« Mout i faria gran ergueil
« S'ieu d'aizo dizia de no.

(*). *Du premier feuillet il ne reste que le coin supérieur engagé dans la reliure. Sur ce débris on lit ces débuts de vers :*
Sa colors.....
Anc d.....
Nat.....
Ab.....
C.....

1. tot, *ms.* ton, *corrigé en* tot, *le second jambage de l'n étant exponctué.* — 2. nom, *ms.* nō. — 10. domini.

T. I 1

Figure 1: Scanned image of the Meyer's edition of *Flamenca*

to the preceding noun *domini* 'lordship' and to the conjunction *si* 'if'. The detached clitics are shown in (1b). In our syntactic interpretation of such cases, we follow more recent editions of the manuscript [2, 9, 10].

(1) a. *Per son anel domini**m*** *manda Que Flamenca penra si**m***
for his ring lordship-me sends that Flamenca takes if-me
voil.
want
'With his own ring he has let me know that he wants to marry Flamenca if I permit.' (line 10, [2])

b. Per son anel domini **-m** manda Que Flamenca penra si **-m** voil .

Since creating linguistic corpus annotation can be a labor-intensive and time-consuming process, we have addressed this issue with a resource-light method that exploits existing resources. Several studies have shown that linguistic information available for a resource-rich language can be transferred to a closely related resource-poor language [7, 18]. Given that Old Occitan and Old French share many lexical, morphological, and syntactic characteristics, we have selected the MCVF

Tag	Definition	Tag	Definition
ADJ	adjective	MDJ	present of modal verb
ADJNUM	cardinal adjective	MDPP	past participle of modal verb
ADJR	comparative form of adjective	MDX	infinitive of modal verb
ADV	adverb	NCPL	noun common plural
ADVR	comparative form of adverb	NCS	noun common singular
AG	gerundive of auxiliary 'to have'	NEG	negation
AJ	present of auxiliary 'to have'	NPRPL	noun proper plural
APP	past participle of auxiliary 'to have'	NPRS	noun proper singular
AX	infinitive of auxiliary 'to have'	NUM	numeral
CMP	prep. comparison	P	preposition
CONJO	coordinative conjunction	PON	punctuation inside the clause
CONJS	subordinate conjunction	PONFP	the end of the sentence
COMP	comparative adverb	PRO	pronoun
D	determiner (indefinite, definite, demonstrative)	Q	quantifier
DAT	dative	QR	quantifier (more, less)
DF	partitive article	VG	gerundive of the main verb
DZ	possessive determiner	VJ	present of the main verb
EG	gerundive of auxiliary 'to be'	VPP	past participle of the main verb
EJ	present of auxiliary 'to be'	VX	infinitive of the main verb
EPP	past participle of auxiliary 'to be'	WADV	interr., rel. or excl. adverb
EX	infinitive of auxiliary 'to be'	WD	interr., rel. or excl. determiner
ITJ	interjection	WPRO	interr., rel. or excl. pronoun
MDG	gerundive of modal verb		

Table 1: Occitan Part-of-Speech tagset (adapted from Martineau et al. [12])

corpus of Old French [12] annotated with part-of-speech (POS) tags and syntactic constituency labels. We have adopted the POS tagset from the MCVF (see Table 1) and have trained the TnT tagger [4] on 28 265 sentences from the Medieval French section of the MCVF. This trained model was used to POS tag Old Occitan. An evaluation of the accuracy of this approach can be found in our previous work [18]. The tagger output was further manually corrected. In addition, we have augmented tokens with lemmas from the Occitan dictionary. This dictionary is based on the glossary to *Le Roman de Flamenca* [15] and consists of 2 800 entries.

For syntactic parsing, we have trained the Berkeley parser [16] using a constituency treebank from the MCVF Medieval corpus [12]. The trained model was used to parse our corpus. We have adopted syntactic labeling from the MCVF corpus. However, we have added an additional label V for verbs in order to facilitate queries. Table 2 lists the syntactic labels used in the annotation of the Old Occitan corpus.

The parsed trees were manually corrected. As a reference guide we have used the MCVF manual[7]. However, for lack of resources, and in contrast to the MCVF corpus, we did not manually add traces and empty categories. An example of the syntactic structure in our corpus is shown in Figure 2.

In addition we manually added a discourse layer that identifies different speakers in the dialogues. The labels correspond to the main characters names, namely

[7] http://gtrc.voies.uottawa.ca/manuel/syntax-manual-fr/index.htm

Labels	Definition	Label	Definition
ADJP	Adjectival Phrase	IP-IMP	Imperative Proposition
ADVP	Adverbial Phrase	IP-INF	Infinitival Proposition
ADVP-LOC	Adverbial Locative Phrase	IP-MAT	Main Proposition
ADVP-TMP	Adverbial Temporal Phrase	IP-PPL	Participial Proposition
CONJP	Conjunction	IP-SUB	Subordinate Proposition
CP-ADV	Adverbial Clause	NP-ACC	Direct Object
CP-ADV-TMP	Temporal Clause	NP-COM	NP Complement
CP-CAR	Prepositional Clause	NP-DTV	Indirect Object
CP-CMP	Comparative Clause	NP-PRD	Predicative NP
CP-DEG	Degree Clause	NP-RFL	Reflexive NP
CP-EXL	Exclamative Clause	NP-SBJ	Subject NP
CP-FRL	Small Clause	NP-TMP	Temporal NP
CP-OPT	Optative Clause	PP	Prepositional Phrase
CP-QUE	Interrogative Clause	PP-DIR	Prepositional Directional Phrase
CP-REL	Relative Clause	PP-LOC	Prepositional Locative Phrase
CP-THT	Complement Clause	QP	Quantifier Phrase
INTJ	Interjection	V	Verb
-LFD	Left Dislocated Phrase	-PRN	Adjunct
-SPE	Direct Speech		

Table 2: Occitan Syntactic Labels (adapted from Martineau et al. [12])

Flamenca, Archambaut, Guillem, Father. Less important characters are marked as FemaleSpeakers and MaleSpeakers. This additional information could enhance studies focusing on social variation.

Finally, the tagged and parsed texts were merged and converted to the ANNIS[8] format. ANNIS is a web-based corpus application that allows for visualization and querying of the corpus at multiple levels [21].

4 Corpus Applications

Since we are targeting two different types of users, linguists and non-linguists, with different needs, the corpus is made available in two different modes. In the first mode, the user can mainly browse the text and look up translations and glosses, in an intuitive interface (see section 4.1). In the second mode, users interested in the linguistic annotation can query the corpus for linguistic phenomena. This requires a more complex query language and thus a more complex query tool. We show such queries in section 4.2.

4.1 Textual Representation

In order to provide access to a general audience, the TEI XML data are transformed into an interactive web database[9], which allows the user to read the romance without looking at the annotations, but with access to supplemental information in-

[8] http://www.sfb632.uni-potsdam.de/annis/

[9] http://nlp.indiana.edu/~obscrivn/Introduction.html

```
((IP-MAT-SPE
   (QP (Q assaz))
   (V (MDJ podes))
   (IP-INF
      (V
          (V (VX donar))
          (CONJO e)
          (V (VX metre))))
   (PONFP ;)))
```

Eng.: “you can bestow and spend a great deal ” (line 117, [2])

Figure 2: An example for the syntactic annotation.

tended to facilitate the access to the text. The corpus is divided into Meyer’s introduction, sections of the text, glossary, and the annotated text of the romance itself. Each section of the text provides access to its French translation by Meyer.[10] Glossary definitions, comments, and footnotes are linked to tokens and are made visible when the user hovers over a marked word, as shown in Figure 3 for a glossary entry for the word *acapte*.

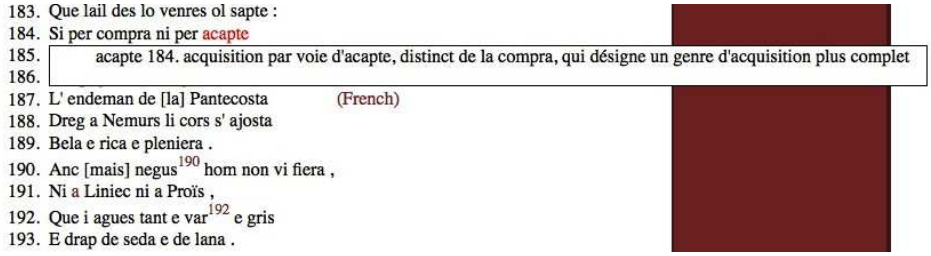


Figure 3: Glossary definition for the word *acapte*

4.2 Querying the Annotations

In order to allow queries that go beyond searching for (sequences of) words and to allow access to the annotations, we imported all the annotations into ANNIS.[11] Our web search based on ANNIS allows for basic queries, to search for a word or

[10] Although this translation (from Meyer’s first edition [14]) is sufficient to give a good idea of the content of the text, corrections to the reading of the manuscript made in later editions (including Meyer [15]) will obviously not be reflected.

[11] http://nlp.indiana.edu:8080/annis-gui-3.0.0/

phrase, and more complex queries for syntactic and morphosyntactic annotation. For example, to find all the occurrences of the word *cor* 'heart', the user can submit the query "cor" in the Search Window. The total number of occurrences will be shown in the Status Window, and the results will be displayed in the Query Result window, as shown in Figure 4.

Figure 4: Results for lexical query of the word *cor* 'heart'

The default view is in KWIC format, which displays only tokens and POS tags per line. But it is also possible to access a grid view with lemmas or a constituency tree, as shown in Figure 5.

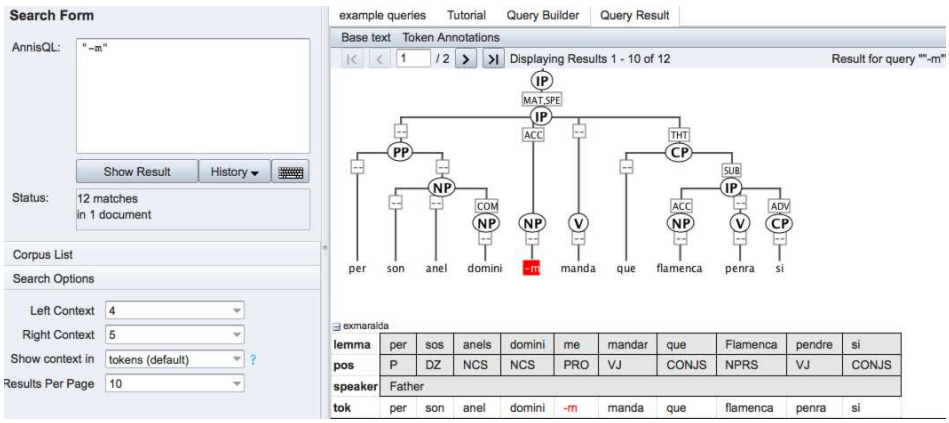


Figure 5: Grid and constituency tree visualization

ANNIS can also handle more complex queries. For instance, the query illustrated in Figure 6 shows how to search for all the cases of subjects that directly precede verbs. In this query, we describe the partial tree structure that corresponds to the phenomenon in which we are interested: We select an IP node (cat = IP) and specify a grammatical relation subject (func = SBJ) between the IP node and its NP daughter (cat = NP). To define a precedence relation between the NP and a verb (cat = V), we connect the two nodes by means of operator "." (direct precedence).

The results of queries can be exported and downloaded in plain text format with or without POS tags. The example of a result without POS tags is shown in (2a), and the example with POS tags is shown in (2b):

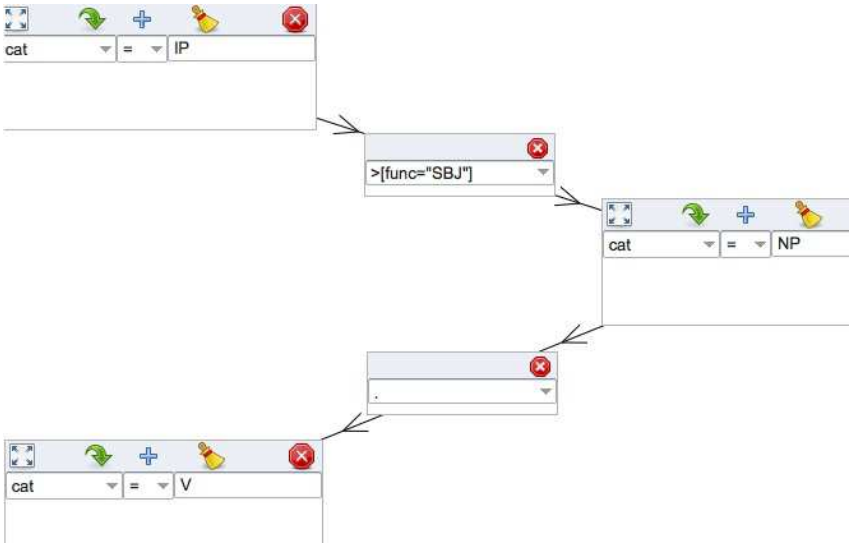


Figure 6: Example of syntactic query for a subject (NP-SBJ) preceding a verb (V)

(2) a. *le reis a dih a totz em bala :*
the king has said to all in assembly :
'The king said to the whole assembly' (line 714 [2])

b. le/D reis/NCS a/AJ dih/VPP a/DAT totz/Q em/P bala/NCS :/PONFP

In the following query, we show that it is possible to integrate different types of annotation into one query. Thus we restrict the previous search to only subject pronouns by adding a new condition, namely POS tag information. In this case, the syntactic category NP is connected to the POS tag PRO (pronoun) by means of the equality operator "_=_". The query is illustrated in Figure 7.

As a result, we find only 10 subject pronouns followed by a verb in the lines 1-798, compared to 80 in the previous query.

5 Conclusion

This paper describes an ongoing effort to digitize and annotate the corpus of *Le Roman de Flamenca*, a 13th-century romance, written in Old Occitan. In contrast to traditional corpora, this corpus is structured to fulfill two objectives. First, the web design facilitates the reading and understanding of *The Romance of Flamenca*. Words are interactively linked to the glossary, comments, and translations. Second, the corpus search design via its ANNIS interface allows for a visualization and for complex queries of the morpho-syntactic and syntactic annotations. Finally, the

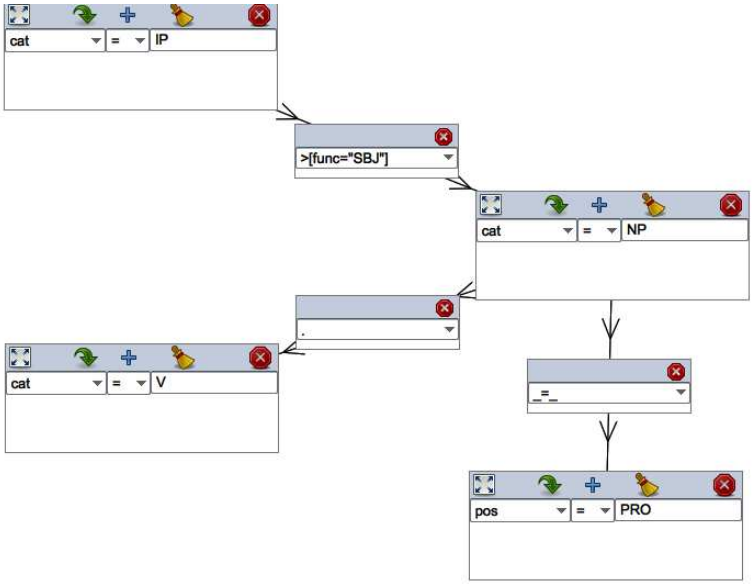


Figure 7: Example of syntactic query for a pronominal subject preceding a verb

manual correction of automatically processed text guarantees the high accuracy of the results.

In the future, we plan to augment our project with a parallel English translation [2] and to build an Old Occitan-English dictionary. In addition, we plan to add empty categories and traces to our constituency treebank, following the guidelines from the MCVF corpus [12].

6 Acknowledgements

The authors would like to thank Professor France Martineau for permission to use the MCVF corpus as a training model for our tagger and parser. Also we would like to thank Thomas Krause and Amir Zeldes for helping with the installation and configuration of ANNIS, and Michael McGuire for providing help with OCR.

References

[1] ARTFL Project. *Provençal Poetry database (American and French Research on the Treasury of the French Language), Robert Morrissey, director, with F.R. Akehurst*, 1998.

[2] E.D. Blodgett. *The Romance of Flamenca*. Garland, New York, 1995.

[3] W.A. Bradley. *The Story of Flamenca*. Harcourt Brace, New York, 1922.

[4] Thorsten Brants. TnT – a statistical part-of-speech tagger. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics and the 6th Conference on Applied Natural Language Processing (ANLP/NAACL)*, pages 224–231, Seattle, WA, 2000.

[5] Mark Davies. Corpus del Español: 100 million words, 1200s-1900s. Available online at `http://www.corpusdelespanol.org`, 2002.

[6] Mark Davies and Michael Ferreira. Corpus do Portugues: 45 million words, 1300s-1900s. Available online at `http://www.corpusdoportugues.org`, 2006.

[7] Anna Feldman and Jirka Hana. *A Resource-Light Approach to Morpho-Syntactic Tagging*. Rodopi, 2010.

[8] Suzanne Fleischmann. The non-lyric texts. In F.R.P. Akehurst and Judith M. Davis, editors, *A Handbook of the Troubadours*, pages 176–184. University of California Press, 1995.

[9] Ulrich Gschwind. *Le Roman de Flamenca. Nouvelle occitane du 13e siècle*, volume 2. Francke, Berne, 1976.

[10] Jean-Charles Huchet. *Flamenca. Roman Occitan du XIII siècle.* Union Générale d'Editions, Paris, 1988.

[11] René Lavaud and René Nelli. *Les Troubadours*. Paris: Desclée de Brouwer, 1960.

[12] France Martineau, Constanta Diaconescu, and Paul Hirschbühler. Le corpus 'voies du français': De l'élaboration à l'annotation. In Pierre Kunstmann and Achim Stein, editors, *Le Nouveau Corpus d'Amsterdam*, pages 121–142. Steiner, 2007.

[13] France Martineau, Paul Hirschbühler, Anthony Kroch, and Yves Charles Morin. Corpus MCVF (parsed corpus), modéliser le changement: les voies du français, Départment de Français, University of Ottawa. CD-ROM, first edition, `http://www.arts.uottawa.ca/voies/voies_fr.html`, 2010.

[14] Paul Meyer. *Le Roman de Flamenca*. Béziers, 1865.

[15] Paul Meyer. *Le Roman de Flamenca*. Librairie Emile Bouillon, 2nd edition, 1901.

[16] Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual*

Meeting of the Association for Computational Linguistics, pages 433–440, Sydney, Australia, 2006.

[17] Peter T. Ricketts and Alan Reed. *Concordance de l'Occitan Médiéval. COM 2: Les Troubadours, Les Textes Narratifs en vers*. Brepols, Turnhout, 2005.

[18] Olga Scrivner and Sandra Kübler. Building an Old Occitan corpus via cross-language transfer. In *Proceedings of the First International Workshop on Language Technology for Historical Text(s)*, Vienna, Austria, 2012.

[19] Achim Stein. Syntactic annotation of Old French text corpora. *Corpus*, 7:157–161, 2008.

[20] Amir Zeldes. *ANNIS: User Guide - Version 3.0.0*. SFB 632 Information Structure / D1 Linguistic Database, Humboldt-Universität zu Berlin & Universität Potsdam, June 2013.

[21] Amir Zeldes, J. Ritz, Anke Lüdeling, and Christian Chiarcos. ANNIS: A search tool for multi-layer annotated corpora. In *Proceedings of Corpus Linguistics*, Liverpool, UK, 2009.