

---

## Language Change Quantification Using Time-separated Parallel Translations

**Type** Journal Article

**Author** Kemal Altintas

**Author** Fazli Can

**Author** Jon M. Patton

**Abstract** We introduce a systematic approach to language change quantification by studying unconsciously used language features in time-separated parallel translations. For this purpose, we use objective style markers such as vocabulary richness and lengths of words, word stems and suffixes, and employ statistical methods to measure their changes over time. In this study, we focus on the change in Turkish in the second half of the twentieth century. To obtain word stems, we first introduce various stemming techniques and show that they are highly effective. Our statistical analyses show that over time, for both text and lexicon, the length of Turkish words has become significantly longer, and word stems have become significantly shorter. We also show that suffix lengths have become significantly longer for types and the vocabulary richness based on word stems has shrunk significantly. These observations indicate that in contemporary Turkish one would use more suffixes to compensate for the fewer stems to preserve the expressive power of the language at the same level. Our approach can be adapted for quantifying the change in other languages.

**Publication** Lit Linguist Computing

**Volume** 22

**Issue** 4

**Pages** 375-393

**Date** November 1, 2007

**DOI** 10.1093/lc/fqm026

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/22/4/375>

**Accessed** Sun Apr 5 04:35:10 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:35:10 2009

**Modified** Sun Apr 5 04:35:10 2009

### Tags:

computer\_science, linguistics, multi-institutional

### Notes:

## Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Corpus Linguistics and South Asian Languages: Corpus Creation and Tool Development

**Type** Journal Article

**Author** Paul Baker

**Author** Andrew Hardie

**Author** Tony McEnery

**Author** Richard Xiao

**Author** Kalina Bontcheva

**Author** Hamish Cunningham

**Author** Robert Gaizauskas

**Author** Oana Hamza

**Author** Diana Maynard

**Author** Valentin Tablan

**Author** Cristian Ursu

**Author** B. D. Jayaram

**Author** Mark Leisher

**Abstract** This paper describes the work carried out on the EMILLE Project (Enabling Minority Language Engineering), which was undertaken by the Universities of Lancaster and Sheffield. The primary resource developed by the project is the EMILLE Corpus, which consists of a series of monolingual corpora for fourteen South Asian languages, totalling more than 96 million words, and a parallel corpus of English and five of these languages. The EMILLE Corpus also includes an annotated component, namely, part-of-speech tagged Urdu data, together with twenty written Hindi corpus files annotated to show the nature of demonstrative use in Hindi. In addition, the project has had to address a number of issues related to establishing a language engineering (LE) environment for South Asian language processing, such as translating 8-bit language data into Unicode and producing a number of basic LE tools. The development of tools for EMILLE has contributed to the ongoing development of the LE architecture GATE, which has

been extended to make use of Unicode. GATE thus plugs some of the gaps for language processing R&D necessary for the exploitation of the EMILLE corpora.

**Publication** Lit Linguist Computing  
**Volume** 19  
**Issue** 4  
**Pages** 509-524  
**Date** November 1, 2004  
**DOI** 10.1093/llc/19.4.509  
**Short Title** Corpus Linguistics and South Asian Languages  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/19/4/509>  
**Accessed** Sun Apr 5 05:41:13 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:41:13 2009  
**Modified** Sun Apr 5 05:41:13 2009

### Tags:

linguistics, multi-country, multi-institutional, project\_team, tool\_development

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## The Computational Modern Greek Morphological Lexicon--An Efficient and Comprehensive System for Morphological Analysis and Synthesis

**Type** Journal Article  
**Author** S. D. Baldzis  
**Author** S. A. Kolalas  
**Author** E. Eumeridou  
**Abstract** In this paper, we present a system performing morphological analysis and synthesis for the Greek language. Its main features are the use of a single framework to describe and classify morphological data into well-defined classes characterized by their own set of properties and mechanisms and the structuring of the database into distinct levels corresponding to the different levels of morphological information present in the above framework. The resulting system is characterized by simple and flexible algorithms with 100% success in the recognition and generation of morphological forms of the language independently

of the complexity of the data they are handling.

**Publication** Lit Linguist Computing  
**Volume** 20  
**Issue** 2  
**Pages** 153-187  
**Date** June 1, 2005  
**DOI** 10.1093/lc/fqh032  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/20/2/153>  
**Accessed** Sun Apr 5 05:32:21 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:32:21 2009  
**Modified** Sun Apr 5 05:32:21 2009

### Tags:

informatics, linguistics, single\_institution, tool\_development

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Ordering Chaos: An Integrated Guide and Online Archive of Walt Whitman's Poetry Manuscripts

**Type** Journal Article  
**Author** Brett Barney  
**Author** Mary Ellen Ducey  
**Author** Andrew Jewell  
**Author** Kenneth M. Price  
**Author** Brian Pytlik Zillig  
**Author** Katherine L. Walter  
**Abstract** In order to organize the widely dispersed manuscripts of Walt Whitman, The Walt Whitman Archive, in partnership with the University of Nebraska-Lincoln Libraries, has utilized the power of Encoded Archival Description (EAD) to create a single, scholarly enhanced guide to Whitman's poetry manuscripts. This integrated finding guide to Whitman's poetry manuscripts includes item-level description, links to repository guides that provide both location information and

collection context, links to digital images of the manuscripts, and links to Text Encoding Initiative (TEI) transcriptions. In creating such a guide, we had to work cooperatively across disciplines and institutions, expand the use of EAD, and address how best to integrate description and transcription (EAD and TEI files). This essay describes our procedure as we created the integrated guide. From collecting finding aids and creating partnerships with other institutions, to developing a proper encoding standard and establishing good cross-department working relations, our project has embodied many of the benefits and challenges of digital work in the humanities. By identifying our procedures, and by laying out our future hurdles, we hope we can advance knowledge about Whitman and about how scholars and archivists can collaborate effectively to advance research, improve access, and realize the potential of EAD.

**Publication** Lit Linguist Computing  
**Volume** 20  
**Issue** 2  
**Pages** 205-217  
**Date** June 1, 2005  
**DOI** 10.1093/lc/fqi002  
**Short Title** Ordering Chaos  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/20/2/205>  
**Accessed** Sun Apr 5 05:32:25 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:32:25 2009  
**Modified** Sun Apr 5 05:32:25 2009

**Tags:**

english, libraries, markup, project\_team, single\_institution

**Notes:**

Brett Barney, Mary Ellen Ducey, Andrew Jewell,  
Kenneth M. Price, Brian Pytlik Zillig, and Katherine L. Walter  
University of Nebraska-Lincoln, USA

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text

**Type** Journal Article

**Author** Marco Baroni

**Author** Silvia Bernardini

**Abstract** In this article we describe an approach to the identification of translationese' based on monolingual comparable corpora and machine learning techniques for text categorization. The article reports on experiments in which support vector machines (SVMs) are employed to recognize translated text in a corpus of Italian articles from the geopolitical domain. An ensemble of SVMs reaches 86.7% accuracy with 89.3% precision and 83.3% recall on this task. A preliminary analysis of the features used by the SVMs suggests that the distribution of function words and morphosyntactic categories in general, and personal pronouns and adverbs in particular, are among the cues used by the SVMs to perform the discrimination task. A follow-up experiment shows that the performance attained by SVMs is well above the average performance of ten human subjects, including five professional translators, on the same task. Our results offer solid evidence supporting the translationese hypothesis, and our method seems to have promising applications in translation studies and in quantitative style analysis in general. Implications for the machine learning/text categorization community are equally important, both because this is a novel application and especially because we provide explicit evidence that a relatively knowledge-poor machine learning algorithm can outperform human beings in a text classification task.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 3

**Pages** 259-274

**Date** September 1, 2006

**DOI** 10.1093/lc/fqi039

**Short Title** A New Approach to the Study of Translationese

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/21/3/259>

**Accessed** Sun Apr 5 05:07:19 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:07:19 2009

**Modified** Sun Apr 5 05:07:19 2009

### Tags:

computer\_program, NLP, single\_institution

## Attachments

HighWire Full Text PDF

HighWire Snapshot

---

### Communities of Interest: Issues in Establishing a Digital Resource on Murrinh-patha song at Wadeye (Port Keats), NT

**Type** Journal Article

**Author** Linda Barwick

**Author** Allan Marett

**Author** Michael Walsh

**Author** Nicholas Reid

**Author** Lysbeth Ford

**Abstract** Linguistics and musicology, along with other fieldwork-based disciplines, have obligations to facilitate access to research results by the communities whose cultural heritage is recorded and analysed, especially when the languages and musics in question are otherwise little documented, have few speakers or performers, and are threatened by the global dominance of English. This article presents the early results of our planning for establishment of a digital resource to preserve and make accessible recordings and other documentation of Murrinh-patha public dance-songs at Wadeye, a remote Indigenous community in Australia's Northern Territory. With the recent establishment of the Wadeye Knowledge Centre, copies of recordings previously left in the community by researchers have been digitized and made available through computer workstations. Many of these digitized recordings, however, have poor or no documentation and thus are difficult to locate and access. The most urgent task is to work with elderly performers and composers to assemble metadata about the oldest recordings of songs and who composed and performed them. In order to maximise local accessibility and use, both elders and young people will be involved in planning and creation of a bilingual search interface to the collection. Planning must also consider sustainability issues through integration with other local initiatives, appropriate use of open standards and formats, locally sustainable technical platforms, and regular backup and maintenance.

**Publication** Lit Linguist Computing

**Volume** 20

**Issue** 4

**Pages** 383-397

**Date** November 1, 2005

**DOI** 10.1093/llc/fqi048  
**Short Title** Communities of Interest  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/20/4/383>  
**Accessed** Sun Apr 5 05:22:46 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:22:46 2009  
**Modified** Sun Apr 5 05:22:46 2009

**Tags:**

multi-institutional, project\_team

**Attachments**

HighWire Full Text PDF  
HighWire Snapshot

---

**Human Computing--Modelling with Meaning**

**Type** Journal Article  
**Author** Meurig Beynon  
**Author** Steve Russ  
**Author** Willard McCarty  
**Abstract** This article is based on a session given by the authors at the ACH/ALLC conference at the University of Victoria in June 2005. It discusses the prospects for partnership between the humanities and computing from the alternative perspective afforded by Empirical Modelling (EM). Perceived dualities that separate the two cultures of science and art are identified as the primary impediment to this partnership. A vision for human computing' that promises to dissolve these dualities is outlined. The key characteristics and potential for EM for the humanities are illustrated with reference to a modelling exercise on the theme of Schubert's Erlkonig. This highlights how each of the six varieties of modelling identified by McCarty can be represented within an EM model. The implications of EM are discussed with reference to McCarty's account of the key role for modelling in the humanities, in relation to James's philosophic attitude' of Radical Empiricism and to ideas from phenomenological sources.  
**Publication** Lit Linguist Computing  
**Volume** 21  
**Issue** 2  
**Pages** 141-157

**Date** June 1, 2006

**DOI** 10.1093/llc/fql015

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/2/141>

**Accessed** Sun Apr 5 05:12:08 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:12:08 2009

**Modified** Sun Apr 5 05:12:08 2009

### Tags:

computer\_science, humanities, modeling, multi-institutional

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Philology Meets Text Encoding in the New Scholarly Edition of Henrik Ibsen's Writings

**Type** Journal Article

**Author** Hilde Boe

**Author** Jon Gunnar Jorgensen

**Author** Stine Brenna Taugbol

**Abstract** In Norway, the project Henrik Ibsen's Writings is currently establishing a new historical-critical edition (both electronically and in print) of the complete writings of playwright Henrik Ibsen. In the years the project has existed, there has been a continuing internal discussion on the relationship between philology and text encoding. This paper outlines the philological principles of the project and describes its methods of establishing texts and ensuring quality. It also looks at and describes, in detail, the consequences of combining philology and text encoding through examples of problems solved in the encoding of complex changes in manuscripts as well as parallel structures in verse dramas. The paper concludes that it is very important, in a project such as Henrik Ibsen's Writings, to focus on the relationship between philology and text encoding because of the influence, even in the smallest details, of these on each other.

**Publication** Lit Linguist Computing

**Volume** 19

**Issue** 1

**Pages** 55-71

**Date** April 1, 2004

**DOI** 10.1093/llc/19.1.55

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/19/1/55>

**Accessed** Sun Apr 5 05:53:52 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:53:52 2009

**Modified** Sun Apr 5 05:53:52 2009

### Tags:

markup, project\_team, single\_institution

### Notes:

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Supporting Annotation as a Scholarly Tool--Experiences From the Online Chopin Variorum Edition

**Type** Journal Article

**Author** John Bradley

**Author** Paul Vetch

**Abstract** In a meeting at King's College London in May 2000, John Unsworth proposed a list of seven scholarly primitives' which he claimed were self-understood' functions forming the basis for higher-level scholarly projects, arguments, statements [and] interpretations' (Unsworth, 2000). He claimed that his list summarized activities that were basic to scholarship across eras and across media', and went on to say that an analysis of these scholarly primitives might result in a clearer sense of how computing tools could support the scholarly endeavour. Here we focus on the primitive that was second on Unsworth's list, after 'Discovering': 'Annotation'. Our work on annotation arises out of a developing awareness that established Humanities Computing (HC) areas of interest, do not seem always to

connect with the actual process of the research work being carried out by most humanists. We claimed in Bradley (2005) that a fundamentally different usage paradigm than those in operation in established HC was necessary to even notice, and then follow-up on, the potential of scholarly annotation as a computer-supported activity. This article presents our experiences, and the eventual outcomes, of the process of developing annotations tools for the Online Chopin Variorum Edition project (OCVE).<sup>1</sup> Beginning with a brief overview of activities related to annotation in Humanities Computing and Computing Science, we introduce the visible parts of the OCVE project, and address some discussion to the structures behind the scenes that support what it does, reporting what worked and what did not. We conclude by analysing the significance of our findings and describing the direction we think our annotation tool will take.

**Publication** Lit Linguist Computing  
**Volume** 22  
**Issue** 2  
**Pages** 225-241  
**Date** June 1, 2007  
**DOI** 10.1093/lc/fqm001  
**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/22/2/225>  
**Accessed** Sun Apr 5 04:45:25 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:45:25 2009  
**Modified** Sun Apr 5 04:45:25 2009

### Tags:

project\_team, single\_institution, tool\_development

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Expressing complex associations in medieval historical documents: the Henry III Fine Rolls Project

**Type** Journal Article  
**Author** Arianna Ciula  
**Author** Paul Spence

**Author** Jose Miguel Vieira

**Abstract** This article focuses on the use of technologies traditionally associated with knowledge representation to express complex associations between entities in historical texts that have been marked up in XML, according to the Text Encoding Initiative guidelines. In particular, we describe our exploration of the potential role of an ontology in facilitating the interpretation of implicit and hidden associations in the sources of interest, examining its use, and limits in a digital humanities project in connection with editing tools and delivery issues. We demonstrate our findings based on the Henry III Fine Rolls project, where an ontology--built using the RDF (Resource Description Framework)/OWL (Web Ontology Language) technologies--is being developed to make explicit information about person, place, and subject entities marked up as instances in the core texts themselves. For any historian, there is a natural tension between primary sources (as documentary records) and the analysis that produces a context for interpretation. We will argue that the combination of core mark-up (encoded in TEI) and an ontology (in RDF/OWL) provides a powerful model for representing the complexity of this tension and facilitates the necessarily dynamic process of scholarly interpretation.

**Publication** Lit Linguist Computing

**Volume** 23

**Issue** 3

**Pages** 311-325

**Date** September 1, 2008

**DOI** 10.1093/lc/fqn018

**Short Title** Expressing complex associations in medieval historical documents

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/23/3/311>

**Accessed** Sun Apr 5 04:16:35 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:16:35 2009

**Modified** Sun Apr 5 04:16:35 2009

### Tags:

markup, project\_team, single\_institution

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

North American English Vowels: A Factor-analytic Perspective

**Type** Journal Article

**Author** Cynthia G. Clopper

**Author** John C. Paolillo

**Abstract** Previous studies of American English have identified a number of robust patterns involving the vowel system, such as the Northern Cities Chain Shift and the Southern Vowel Shift. These studies primarily employ methods which treat separately the phonetic properties of specific vowels as produced by individual speakers which are later assembled into complete vowel systems. While this provides a useful picture of production, it is not adequate for comparison with dialect perception studies, where interpretation of the results often requires some understanding of the correlations among linguistic features and between those features and individual talkers. We conducted a factor analysis of the duration and first and second formant frequencies of each of the fourteen vowels produced by forty-eight speakers representing six regional varieties of American English and both genders. The data were submitted to factor analysis using maximum likelihood estimation and Varimax rotation. Results confirmed significant correlations between regional dialect and acoustic-phonetic properties of the vowel systems, although these patterns are complicated by interactions with gender. These results illustrate the utility of factor analytic methods in examining systematic variation across an entire linguistic system such as the vowels.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 4

**Pages** 445-462

**Date** November 1, 2006

**DOI** 10.1093/lc/fql039

**Short Title** North American English Vowels

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/21/4/445>

**Accessed** Sun Apr 5 05:01:28 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:01:28 2009

**Modified** Sun Apr 5 05:01:28 2009

### Tags:

linguistics, single\_institution

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## In the Philosophy Room: Australian Realism and Digital Content Development

**Type** Journal Article

**Author** Creagh Cole

**Author** Paul Scifleet

**Abstract** The text ontology debates inspired by the descriptive encoding practices of the Text Encoding Initiative (TEI) community have been conducted between literary theorists concerned with the adequacy of such encoding to capture the interpretative and playful aspects of literary appreciation on the one hand, and those who regard encoding as one of the formal sciences and seek greater disambiguation in the interests of more efficient machine processing. We argue for a practice-oriented view that has not been represented adequately by either of these poles. Our position has received unexpected support from the systematic realist philosophy of John Anderson which we encountered in digitizing his lecture notes held by the University of Sydney Archives. The process of encoding the lecture notes informed our understanding of the problems of encoding primary source materials, but Anderson's realism also located the space we sought to occupy in the TEI debates between the technical, formal model of encoding and the anti-realist preferences of many literary scholars. In this article, we argue on the basis of our reflections the need for further empirical studies of real world encoding practices as these new documentary forms are integrated into existing institutional and informational processes.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 2

**Pages** 159-167

**Date** June 1, 2006

**DOI** 10.1093/lc/fql017

**Short Title** In the Philosophy Room

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/21/2/159>

**Accessed** Sun Apr 5 05:12:10 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:12:10 2009

**Modified** Sun Apr 5 05:12:10 2009

### Tags:

faculty, markup, project\_staff, project\_team, single\_institution

**Notes:**

Creagh Cole  
Scholarly Electronic Text & Image Service,  
The University of Sydney  
Paul Scifleet  
Discipline of Information Systems, Faculty of Economics &  
Business, The University of Sydney

**Attachments**

HighWire Full Text PDF  
HighWire Snapshot

---

**Using Ancillary Text to Index Web-based Multimedia Objects**

**Type** Journal Article

**Author** Lyne Da Sylva

**Author** James M. Turner

**Abstract** PeriCulture is the name of a research project at the Universite de Montreal which is part of a larger project based at the Universite de Sherbrooke. The parent project aimed to form a research network for managing Canadian digital cultural content. The general research objective of PeriCulture was to study indexing methods for web-based non-textual cultural content, specifically still images. The research results reported here build on work in image indexing and automatic (text) indexing by studying properties of text associated with images in a networked environment to try to gain some understanding of how the ancillary text associated with images on web pages can be exploited to index the corresponding images. We studied this question in the context of selected web sites, i.e. that contained multimedia objects, that had text associated with these objects (broader than file names and captions), that were bilingual (English and French), and that housed Canadian digital cultural content. We identified keywords that were useful in indexing and studied their proximity to the object described. Potential indexing terms were identified in various HTML tags and full text (each considered a different source of ancillary text). Our study found that a large number of useful indexing terms are available in the ancillary text of many web sites with cultural content, and that ancillary text of different sources have variable usefulness in retrieval. Our results suggest that these terms can be manipulated in a number of ways in automated retrieval systems to improve search results.

**Publication** Lit Linguist Computing

**Volume** 21  
**Issue** 2  
**Pages** 219-228  
**Date** June 1, 2006  
**DOI** 10.1093/llc/fql018  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/2/219>  
**Accessed** Sun Apr 5 05:12:20 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:12:20 2009  
**Modified** Sun Apr 5 05:12:20 2009

**Tags:**

IR, project\_team, single\_institution

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Goldsmith and the Busy Body

**Type** Journal Article  
**Author** Peter Dixon  
**Author** David Mannion  
**Abstract** Assessments of Goldsmith's contribution to the Busy Body have fluctuated widely. We examine the ten possible attributions, gathering evidence from verbal parallels, selected linguistic features, and measures of sentence-length, together with idiosyncrasies of vocabulary and syntax in the doubtful' essays themselves. We conclude that apart from the essay on London clubs, which he later acknowledged, only one piece can be attributed to Goldsmith with any confidence.  
**Publication** Lit Linguist Computing  
**Volume** 22  
**Issue** 4  
**Pages** 435-446  
**Date** November 1, 2007  
**DOI** 10.1093/llc/fqm028  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/4/435>

**Accessed** Sun Apr 5 04:35:16 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:35:16 2009  
**Modified** Sun Apr 5 04:35:16 2009

**Tags:**

multi-institutional, stylistics

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Goldsmith's Contributions to the Weekly Magazine

**Type** Journal Article  
**Author** Peter Dixon  
**Author** David Mannion  
**Abstract** Fifteen items in the Weekly Magazine have been attributed to Goldsmith. Our study uses traditional kinds of internal evidence (mainly verbal parallels) together with evidence from selected linguistic features. A preliminary analysis identifies features which best distinguish Goldsmith samples from those of a number of contemporary authors. Using this selection of features, we calculate the distances of the fifteen Weekly items from the cluster of Goldsmith samples; an item at too large a distance is unlikely to be his. A parallel investigation is based on sentence-length statistics. We conclude that seven essays may plausibly be assigned to Goldsmith, that he probably co-authored two pieces, and that in three cases he merely made minor additions to material from other sources.  
**Publication** Lit Linguist Computing  
**Volume** 22  
**Issue** 4  
**Pages** 447-468  
**Date** November 1, 2007  
**DOI** 10.1093/lc/fqm019  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/4/447>  
**Accessed** Sun Apr 5 04:35:18 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:35:18 2009

**Modified** Sun Apr 5 04:35:18 2009

**Tags:**

multi-institutional, stylistics

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

**Morphological Analysis of the Qur'an**

**Type** Journal Article

**Author** Judith Dror

**Author** Dudu Shaharabani

**Author** Rafi Talmon

**Author** Shuly Wintner

**Abstract** We present a computational system for morphological analysis and annotation of the Qur'an, for research and teaching purposes. The system facilitates a variety of queries on the Qur'anic text that make reference not only to the words, but also to their linguistic attributes. The core of the system is a set of finite-state based rules which describe the morpho-phonological and morpho-syntactic phenomena of the Qur'anic language. Using a finite-state toolbox we apply the rules to the Qur'anic text and obtain full morphological analysis of its words. The results of the analysis are stored in an efficient database and are accessed through a graphical user interface which facilitates the presentation of complex queries. The system is currently being used for teaching and research purposes; we exemplify its usefulness for investigating several morphological, syntactic, semantic, and stylistic aspects of the Qur'anic text.

**Publication** Lit Linguist Computing

**Volume** 19

**Issue** 4

**Pages** 431-452

**Date** November 1, 2004

**DOI** 10.1093/lc/19.4.431

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/19/4/431>

**Accessed** Sun Apr 5 05:41:09 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:41:09 2009

**Modified** Sun Apr 5 05:41:09 2009

**Tags:**

single\_institution, tool\_development

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## A Default Inheritance Hierarchy for Computing Hebrew Verb Morphology

**Type** Journal Article

**Author** Raphael Finkel

**Author** Gregory Stump

**Abstract** We apply default inheritance hierarchies to generating the morphology of Hebrew verbs. This approach represents inflectional exponents as markings associated with the application of rules by which complex word forms are deduced from simpler roots or stems. The high degree of similarity among verbs of different conjugation classes allows us to formulate general rules; these general rules are, however, sometimes overridden by conjugation-specific rules. Similarly, a verb's form within a particular conjugation is determined both by default rules and by overriding rules specific to lexical stem peculiarities. Our result is a concise set of rules defining the morphology of Hebrew verbs in all conjugations. We express these rules in KATR, both a formalism for default inheritance hierarchies and associated software for generating the forms specified by those rules. As we describe the rules, we point out general strategies for expressing morphology in KATR. We conclude by discussing KATR's advantages over ordinary DATR for the representation of morphological systems and our plans for KATR's successor, LATR.

**Publication** Lit Linguist Computing

**Volume** 22

**Issue** 2

**Pages** 117-136

**Date** June 1, 2007

**DOI** 10.1093/lc/fqm004

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/22/2/117>

**Accessed** Sun Apr 5 04:45:15 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:45:15 2009

**Modified** Sun Apr 5 04:45:15 2009

**Tags:**

computer\_program, linguistics, single\_institution

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

**Digitizing a Dictionary of Medieval Irish: the eDIL Project**

**Type** Journal Article

**Author** Maxim Fomin

**Author** Gregory Toner

**Abstract** The Centre for Irish and Celtic Studies at the University of Ulster is currently producing a digital dictionary of medieval Irish (eDIL) based on the standard Dictionary of the Irish Language published by the Royal Irish Academy, Dublin. This paper addresses some of the problems encountered in the digitization process, including data capture, processing non-standard characters, modifications to the TEI guidelines, automatic generation of tags, and the establishment of a lexical view while preserving the original format of the paper dictionary.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 1

**Pages** 83-90

**Date** April 1, 2006

**DOI** 10.1093/lc/fqh050

**Short Title** Digitizing a Dictionary of Medieval Irish

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/21/1/83>

**Accessed** Sun Apr 5 05:17:39 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:17:39 2009

**Modified** Sun Apr 5 05:17:39 2009

**Tags:**

digitization, project\_team, single\_institution

## Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Function Words in Authorship Attribution Studies

**Type** Journal Article

**Author** Antonio Miranda Garcia

**Author** Javier Calle Martin

**Abstract** The search for a reliable expression to measure an author's lexical richness has constituted many statisticians' holy grail over the last decades in their attempt to solve some controversial authorship attributions. The greatest effort has been devoted to find a formula grounded on the computation of tokens, word-types, most-frequent-word(s), hapax legomena, hapax dislegomena, etc., such that it would characterize a text successfully, independent of its length. In this line, Yule's K and Zipf's Z seem to be generally accepted by scholars as reliable measures of lexical repetition and lexical richness by computing content and function words altogether.<sup>1</sup> Given the latter's higher frequency, they prove to be more reliable identifiers when isolatedly computed in p.c.a. and Delta-based attribution studies, and their rate to the former also measures the functional density of a text. In this paper, we aim to show that each constant serves to measure a specific feature and, as such, they are thought to complement one another since a supposedly rich text (in terms of its lemmas) does necessarily have to characterize by its low functional density, and vice versa. For this purpose, an annotated corpus of the West Saxon Gospels (WSG) and Apollonius of Tyre (AoT) has been used along with a huge raw corpus.

**Publication** Lit Linguist Computing

**Volume** 22

**Issue** 1

**Pages** 49-66

**Date** April 1, 2007

**DOI** 10.1093/lc/fql048

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/22/1/49>

**Accessed** Sun Apr 5 04:55:34 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:55:34 2009

**Modified** Sun Apr 5 04:55:34 2009

**Tags:**

single\_institution, stylistics

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## The Relative Contribution of Pronunciational, Lexical, and Prosodic Differences to the Perceived Distances between Norwegian Dialects

**Type** Journal Article

**Author** Charlotte Gooskens

**Author** Wilbert Heeringa

**Abstract** In the period between 1999 and 2002, Jorn Almborg and Kristian Skarbo compiled a database which consists of recordings and phonetic transcriptions of translations of the fable 'The North Wind and the Sun' in about fifty Norwegian dialects. On the basis of fifteen of these recordings, Charlotte Gooskens carried out a perception experiment (Gooskens and Heeringa, 2004). In this experiment she investigated the distances between the fifteen dialects as perceived by the speakers themselves. On the basis of the phonetic transcriptions, Wilbert Heeringa (2004) measured computational linguistic distances between the fifteen Norwegian varieties (Gooskens and Heeringa, 2004). Distances were calculated by means of Levenshtein distance, which finds the minimum cost of changing one pronunciation into another by inserting, substituting or deleting phonetic segments. Gooskens and Heeringa (2004) correlated the perceptual distances with these computational distances and found a significant correlation of  $r = 0.67$ . In the computational distances, pronunciational, lexical, and morphological variation is processed, but these levels are not studied separately. The contribution of this article is that we measure pronunciational, lexical, and prosodic distances separately. Within pronunciational distances we distinguish between consonants and vowels on the one hand, and between substitutions and insertions/deletions on the other hand. When correlating the separate levels with perception and using multiple linear regression analyses we found that pronunciation is most important in perception and especially vowel substitutions play a major role.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 4

**Pages** 477-492

**Date** November 1, 2006

**DOI** 10.1093/llc/fq1038

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/4/477>

**Accessed** Sun Apr 5 05:01:30 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:01:30 2009

**Modified** Sun Apr 5 05:01:30 2009

### Tags:

humanities\_computing, linguistics, Literature, multi-discipline, single\_institution

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Mutual Comprehensibility of Written Afrikaans and Dutch: Symmetrical or Asymmetrical?

**Type** Journal Article

**Author** Charlotte Gooskens

**Author** Renee van Bezooijen

**Abstract** The two West-Germanic languages Dutch and Afrikaans are so closely related that they can be expected to be mutually intelligible to a large extent. The present investigation focuses on written language. Comprehension was established by means of cloze tests on the basis of two newspaper articles. Results suggest that it is easier for Dutch subjects to understand written Afrikaans than it is for South African subjects to understand written Dutch. In order to explain the results, attitudes as well as several types of linguistic distances were assessed. The relations between attitude scales and intelligibility scores were few and weak. Asymmetries in the linguistic relationships between the two languages are probably more important, especially the asymmetries in the number of noncognates and the opacity of the relatedness of cognates. These asymmetries are caused by historical developments in Dutch and Afrikaans, with respect to the lexicon, grammar, and spelling.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 4  
**Pages** 543-557  
**Date** November 1, 2006  
**DOI** 10.1093/llc/fql036  
**Short Title** Mutual Comprehensibility of Written Afrikaans and Dutch  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/4/543>  
**Accessed** Sun Apr 5 05:01:33 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:01:33 2009  
**Modified** Sun Apr 5 05:01:33 2009

**Tags:**

linguistics, multi-institutional

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## How to Handle Small Samples: Bootstrap and Bayesian Methods in the Analysis of Linguistic Change

**Type** Journal Article  
**Author** Alexander Hinneburg  
**Author** Heikki Mannila  
**Author** Samuli Kaislaniemi  
**Author** Terttu Nevalainen  
**Author** Helena Raumolin-Brunberg  
**Abstract** Estimating the relative frequencies of linguistic features is a fundamental task in linguistic computation. As the amount of text or speech that is available from a given user of the language typically varies greatly, and the sample sizes tend to be small, the most straightforward methods do not always give the most informative answers. Bootstrap and Bayesian methods provide techniques for handling the uncertainty in small samples. We describe these techniques for estimating frequencies from small samples, and show how they can be applied to the study of linguistic change. As a test case, we use the introduction of the pronoun *you* as subject in the data provided by the Corpus of Early English Correspondence (c. 1410-1681).

**Publication** Lit Linguist Computing  
**Volume** 22  
**Issue** 2  
**Pages** 137-150  
**Date** June 1, 2007  
**DOI** 10.1093/lc/fqm006  
**Short Title** How to Handle Small Samples  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/2/137>  
**Accessed** Sun Apr 5 04:45:17 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:45:17 2009  
**Modified** Sun Apr 5 04:45:17 2009

**Tags:**

linguistics, multi-country, multi-institutional

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Bigrams of Syntactic Labels for Authorship Discrimination of Short Texts

**Type** Journal Article  
**Author** Graeme Hirst  
**Author** Ol'ga Feiguina  
**Abstract** We present a method for authorship discrimination that is based on the frequency of bigrams of syntactic labels that arise from partial parsing of the text. We show that this method, alone or combined with other classification features, achieves a high accuracy on discrimination of the work of Anne and Charlotte Bronte, which is very difficult to do by traditional methods. Moreover, high accuracies are achieved even on fragments of text little more than 200 words long.

**Publication** Lit Linguist Computing  
**Volume** 22  
**Issue** 4  
**Pages** 405-417  
**Date** November 1, 2007

**DOI** 10.1093/llc/fqm023

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/4/405>

**Accessed** Sun Apr 5 04:35:12 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:35:12 2009

**Modified** Sun Apr 5 04:35:12 2009

### Tags:

single\_institution

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Syntactic Positions of Prepositional Phrases in the History of Chinese: Using the Developing Sheffield Corpus of Chinese for Diachronic Linguistic Studies

**Type** Journal Article

**Author** Xiaoling Hu

**Author** Jamie McLaughlin

**Author** Nigel Williamson

**Abstract** This paper reports the completion of the first expansion phase of the Sheffield Corpus of Chinese (SCC). We describe the major improvements we made in expanding the corpus. They involve the coverage of time periods, choice of text types and categories, and selection of individual texts; the mark up scheme and the integral search and analysis tool. We use the developing SCC to examine Li and Thompson's; (1974, 1975, 1976) controversial postverbal predominance hypothesis for prepositional phrases (PPs) in Archaic Chinese and their word order change hypothesis for PPs in general in the history of the Chinese language. Our study provides no evidence for the postverbal predominance hypothesis for PPs in Archaic Chinese and the word order change hypothesis for PPs in general from postverbal in Archaic Chinese to preverbal in Modern Chinese. Our findings show that postverbal and preverbal PPs have been in coexistence and there have always been more occurrences of preverbal PPs than postverbal PPs in all the time periods covered in the current SCC. Although use of some PPs declined in some time periods and use of others emerged in other time periods, there was never a predominant position for PPs in any time period in the history of Chinese. We

show differences in the distribution of PPs in different time periods and provide an account of the syntactic positions of PPs in those time periods.

**Publication** Lit Linguist Computing  
**Volume** 22  
**Issue** 4  
**Pages** 419-434  
**Date** November 1, 2007  
**DOI** 10.1093/llc/fqm017  
**Short Title** Syntactic Positions of Prepositional Phrases in the History of Chinese  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/4/419>  
**Accessed** Sun Apr 5 04:35:14 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:35:14 2009  
**Modified** Sun Apr 5 04:35:14 2009

### Tags:

linguistics, multi-institutional

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Sheffield Corpus of Chinese for Diachronic Linguistic Study1

**Type** Journal Article  
**Author** Xiaoling Hu  
**Author** Nigel Williamson  
**Author** Jamie McLaughlin  
**Abstract** The paper presents the outcome of the pilot phase of a major project which aims to build a digital resource for the study of historical Chinese texts with a view to facilitating linguistic analysis of the language, particularly from a diachronic point of view. The approach to general problems for a diachronic corpus is discussed. Details of the tag set and the tagging system devised are given. The development of a sophisticated automatic mark-up scheme for Chinese texts from widely different time periods and genres is indicated.  
**Publication** Lit Linguist Computing  
**Volume** 20

**Issue** 3  
**Pages** 281-293  
**Date** September 1, 2005  
**DOI** 10.1093/llc/fqi034  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/20/3/281>  
**Accessed** Sun Apr 5 05:28:59 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:28:59 2009  
**Modified** Sun Apr 5 05:28:59 2009

**Tags:**

markup, single\_institution

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

**What is transcription?**

**Type** Journal Article  
**Author** Claus Huitfeldt  
**Author** C. M. Sperberg-McQueen  
**Abstract** This paper describes preliminary sketches for a formal account of transcription as it is performed in scholarly editing and in the creation of digital resources. After a general outline of our approach, we present two formal models of transcription. The first addresses only the very simplest cases, the second addresses some but not all of the gaps in the first. Finally, we mention some less simple cases and discuss some elaborations of the model which we hope to develop in future work.  
**Publication** Lit Linguist Computing  
**Volume** 23  
**Issue** 3  
**Pages** 295-310  
**Date** September 1, 2008  
**DOI** 10.1093/llc/fqn013  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/23/3/295>  
**Accessed** Sun Apr 5 04:16:33 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:16:33 2009

**Modified** Sun Apr 5 04:16:33 2009

**Tags:**

markup, multi-country, multi-institutional

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Multivariate Analysis of Finnish Dialect Data An Overview of Lexical Variation

**Type** Journal Article

**Author** Saara Hyvonen

**Author** Antti Leino

**Author** Marko Salmenkivi

**Abstract** During the process of writing a comprehensive dictionary of Finnish dialects, a large set of maps describing the regional distribution of the dialect words have been compiled in electronic form. In this article, we set out to analyse this corpus of data in order to gain new insight on the variation of Finnish dialects. We use a wide range of multivariate data analysis methods, including principal components analysis, independent components analysis, clustering, and multidimensional scaling. We explain how to preprocess the data to overcome the problem of uneven sampling caused by the way the data has been collected. We discuss the results obtained by these methods and compare them to the traditional view of Finnish dialect groups.

**Publication** Lit Linguist Computing

**Volume** 22

**Issue** 3

**Pages** 271-290

**Date** September 1, 2007

**DOI** 10.1093/lc/fqm009

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/22/3/271>

**Accessed** Sun Apr 5 04:39:28 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:39:28 2009

**Modified** Sun Apr 5 04:39:28 2009

**Tags:**

linguistics, multi-institutional

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## A Small-Corpus-Based Approach to Alice's Roles

**Type** Journal Article

**Author** Akiko Inaki

**Author** Tomoko Okita

**Abstract** This paper is a case study for examining how a small-corpus-based approach can contribute to research in stylistics. Specifically, we have built small corpora of the two Alice books and retrieved, using WordSmith Tools suite, first, verbs of saying and their adverbials to elucidate how Alice speaks to others in the stories, and secondly, modifiers of Alice' to get the images of the main character. An analysis of these data reveals that Alice's role in each book is quite distinct: an unexpected visitor thrown into the passive state in Wonderland and an active explorer in Looking-Glass. These findings objectively serve to reinforce our argument over what Alice is called through the perusal of the texts. Alice's roles in the two books are thus interactively supported by the small-corpus-based approach and the non-corpus-based approach, which may explore the validity of the interfaced approach, the collaborative work of quantitative processing and qualitative speculation.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 3

**Pages** 283-294

**Date** September 1, 2006

**DOI** 10.1093/llc/fqi042

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/3/283>

**Accessed** Sun Apr 5 05:07:21 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:07:21 2009

**Modified** Sun Apr 5 05:07:21 2009

**Tags:**

multi-institutional, stylistics

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification

**Type** Journal Article

**Author** Matthew L. Jockers

**Author** Daniela M. Witten

**Author** Craig S. Criddle

**Abstract** Mormon prophet Joseph Smith (1805-44) claimed that more than two-dozen ancient individuals (Nephi, Mormon, Alma, etc.) living from around 2200 BC to 421 AD authored the Book of Mormon (1830), and that he translated their inscriptions into English. Later researchers who analyzed selections from the Book of Mormon concluded that differences between selections supported Smith's claim of multiple authorship and ancient origins. We offer a new approach that employs two classification techniques: delta' commonly used to determine probable authorship and nearest shrunken centroid' (NSC), a more generally applicable classifier. We use both methods to determine, on a chapter-by-chapter basis, the probability that each of seven potential authors wrote or contributed to the Book of Mormon. Five of the seven have known or alleged connections to the Book of Mormon, two do not, and were added as controls based on their thematic, linguistic, and historical similarity to the Book of Mormon. Our results indicate that likely nineteenth century contributors were Solomon Spalding, a writer of historical fantasies; Sidney Rigdon, an eloquent but perhaps unstable preacher; and Oliver Cowdery, a schoolteacher with editing experience. Our findings support the hypothesis that Rigdon was the main architect of the Book of Mormon and are consistent with historical evidence suggesting that he fabricated the book by adding theology to the unpublished writings of Spalding (then deceased).

**Publication** Lit Linguist Computing

**Volume** 23

**Issue** 4

**Pages** 465-491

**Date** December 1, 2008

**DOI** 10.1093/llc/fqn040

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/23/4/465>

**Accessed** Sun Apr 5 04:09:04 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:09:04 2009

**Modified** Sun Apr 5 04:09:04 2009

### Tags:

Civil and Environmental Engineering, english, single\_institution, statistics, stylistics

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Named Entity Recognition for the Mainland Scandinavian Languages

**Type** Journal Article

**Author** Janne Bondi Johannessen

**Author** Kristin Hagen

**Author** Asne Haaland

**Author** Andra Bjork Jonsdottir

**Author** Anders Noklestad

**Author** Dimitris Kokkinakis

**Author** Paul Meurer

**Author** Eckhard Bick

**Author** Dorte Haltrup

**Abstract** In this paper we discuss the results of the Nomen Nescio Named Entity Recognition project, a joint effort for the mainland Scandinavian languages-- Norwegian, Swedish, and Danish. Five research groups have been involved, and developed NE recognizers using rule-based as well as statistical methods. We focus particularly on the choice of semantic categories and the problems regarding metonymy and semantic polysemy. Furthermore, we discuss the extent to which different approaches to these problems have different effects on the different types of systems, and look at two strategies, which we call Function over Form, and Form over Function.

**Publication** Lit Linguist Computing  
**Volume** 20  
**Issue** 1  
**Pages** 91-102  
**Date** March 1, 2005  
**DOI** 10.1093/lc/fqh045  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/20/1/91>  
**Accessed** Sun Apr 5 05:36:14 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:36:13 2009  
**Modified** Sun Apr 5 05:36:13 2009

**Tags:**

linguistics, multi-country, multi-institutional, project\_team

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Developing Web Databases for Aboriginal Language Preservation

**Type** Journal Article  
**Author** Marie-Odile Junker  
**Author** Radu Luchian  
**Abstract** This article discusses the development of integrated multilingual Web databases to help the preservation of the Native American language East Cree. The creation of digital online resources for threatened aboriginal languages presents many technical, educational and ethical challenges. We focus here on the technical challenges in order to discuss both the problems encountered in this particular context, and the solutions we have considered and explored. We illustrate our discussion with examples from an Oral Stories Database we developed in collaboration with Cree education consultants and speakers in 2002-04. We advocate an approach that includes fast-prototyping, open-source development, and design for the database engine that balances speed, availability, features, and resources. We discuss the impact the combination of this technical approach and the participatory action research method is having on language maintenance.  
**Publication** Lit Linguist Computing

**Volume** 22  
**Issue** 2  
**Pages** 187-206  
**Date** June 1, 2007  
**DOI** 10.1093/llc/fq1049  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/2/187>  
**Accessed** Sun Apr 5 04:45:21 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:45:21 2009  
**Modified** Sun Apr 5 04:45:21 2009

**Tags:**

linguistics, project\_team, single\_institution

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## A Prototype for Authorship Attribution Studies

**Type** Journal Article  
**Author** Patrick Juola  
**Author** John Sofko  
**Author** Patrick Brennan  
**Abstract** Despite a century of research, statistical and computational methods for authorship attribution are neither reliable, well-regarded, widely used, or well-understood. This article presents a survey of the current state of the art as well as a framework for uniform and unified development of a tool to apply the state of the art, despite the wide variety of methods and techniques used. The usefulness of the framework is confirmed by the development of a tool using that framework that can be applied to authorship analysis by researchers without a computing specialization. Using this tool, it may be possible both to expand the pool of available researchers as well as to enhance the quality of the overall solutions [for example, by incorporating improved algorithms as discovered through empirical analysis (Juola, P. (2004a). Ad-hoc Authorship Attribution Competition. In Proceedings 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004), Goteborg, Sweden)].

**Publication** Lit Linguist Computing  
**Volume** 21  
**Issue** 2  
**Pages** 169-178  
**Date** June 1, 2006  
**DOI** 10.1093/lc/fql019  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/2/169>  
**Accessed** Sun Apr 5 05:12:12 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:12:12 2009  
**Modified** Sun Apr 5 05:12:12 2009

**Tags:**

computer\_program, computer\_science, single\_institution

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

**teipublisher : A Repository Management System for TEI Documents**

**Type** Journal Article  
**Author** Amit Kumar  
**Author** Susan Schreibman  
**Author** Stewart Arneil  
**Author** Martin Holmes  
**Author** Alejandro Bia  
**Author** John Walsh  
**Abstract** Digital Humanities (DH) and Digital Library (DL) projects are complex systems that require specialized programming skills. Many encoders cannot take their work to the next level by transforming their collections of structured XML texts into a web searchable and browsable database. Often teams of text encoders are able to encode their texts with a high degree of sophistication, but unless they have funds to hire a programmer, their collections far too often remain on local disk storage away from public access. aims to relieve some of this burden by providing the tools to manage an extensible, modular and configurable XML-based repository which will house, search, browse, and display documents

encoded in TEI-Lite on the world wide web. provides an administrative interface that allows DL and DH administrators to upload and delete documents from a web accessible repository; analyze XML documents to determine elements for searching and browsing; refine ontology development; select inter and intra document links; partition the repository into collections; create backups; generate search, browse, and display pages; customize the interface; and associate XSL transformation scripts and CSS stylesheets to obtain different target outputs (HTML, PDF, etc.).

**Publication** Lit Linguist Computing  
**Volume** 20  
**Issue** 1  
**Pages** 117-132  
**Date** March 1, 2005  
**DOI** 10.1093/lc/fqh047  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/20/1/117>  
**Accessed** Sun Apr 5 05:36:18 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:36:18 2009  
**Modified** Sat Apr 18 03:31:17 2009

### Tags:

multi-institutional, project\_team, tech\_development

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## A Tool for Literary Studies: Intertextual Distance and Tree Classification

**Type** Journal Article  
**Author** Cyril Labbe  
**Author** Dominique Labbe  
**Abstract** How to measure proximities and oppositions in large text corpora? Intertextual distance provides a simple and interesting solution. Its properties make it a good tool for text classification, and especially for tree-analysis which is fully presented and discussed here. In order to measure the quality of this classification, two indices are proposed. The method presented provides an accurate tool for literary studies--as is demonstrated by applying it to two areas of French literature,

Racine's tragedies and an authorship attribution experiment.

**Publication** Lit Linguist Computing  
**Volume** 21  
**Issue** 3  
**Pages** 311-326  
**Date** September 1, 2006  
**DOI** 10.1093/lc/fqi063  
**Short Title** A Tool for Literary Studies  
**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/21/3/311>  
**Accessed** Sun Apr 5 05:07:23 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:07:23 2009  
**Modified** Sun Apr 5 05:07:23 2009

### Tags:

multi-institutional, statistics

### Notes:

Grenoble I University, France

Grenoble II University, France

assuming different institutionals

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Trees and After: The Concept of Text Topology. Some Applications to Verb-Form Distributions in Language Corpora

**Type** Journal Article  
**Author** Xuan Luong  
**Author** Michel Juillard  
**Author** Sylvie Mellet

**Author** Dominique Longree

**Abstract** The model described here relies on the key concepts of topology, i.e. neighbourhood and equivalence of shape. A linguistic object L is studied in text T by means of one or several local questions Q. The set of successive local answers is processed so as to provide a global function characterizing the textual space under scrutiny. We begin with short sequences of tenses to illustrate the way in which to explore originally Emile Benveniste's concepts of history and discourse.<sup>1</sup> We then supply life-size examples of other objects selected for their heuristic value. We go on to demonstrate the model at work on the distribution of strings of finite (F) and non-finite (n) verbal forms in the LOB Corpus of English. A topological chart is produced as the synthetic image mirroring the locations of the relevant linguistic entities throughout the text. All the individual strings concatenating any number of F and n are classified in a table. Alternatively, individual full-text strings can be extracted. We then proceed to refine the notion of lexical distribution in 'rafales' in a lemmatized corpus of Latin texts, the purpose being to test the stability of the distributions in individual texts of selected verbs and assess whether a verb's behaviour is related to its semantic status. The final section is devoted to other Latin texts. The use of segments of equal length makes it possible to draw up the narrative profile of each author as revealed by his handling of tenses in main clauses.

**Publication** Lit Linguist Computing

**Volume** 22

**Issue** 2

**Pages** 167-186

**Date** June 1, 2007

**DOI** 10.1093/llc/fqm008

**Short Title** Trees and After

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/2/167>

**Accessed** Sun Apr 5 04:45:19 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:45:19 2009

**Modified** Sun Apr 5 04:45:19 2009

### Tags:

linguistics, multi-country, multi-institutional

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Cryogenics and Creativity: The Frankenstein Factor in Cultural Preservation

**Type** Journal Article

**Author** Eileen Maitland

**Author** Cordelia Hall

**Abstract** The emergence of new media technologies and their integration into the creative process has led to an explosion of hybrid works whose complexities (both conceptual and material) challenge received ideas about the nature of art and its relationship with the future. 'Variable media' has been coined as a descriptor for creative projects incorporating elements whose viability within future incarnations of the same works may be compromised. The flexible and imaginative approach taken by the innovative new media arts community has much to offer information professionals facing the conservation challenges presented by digital materials. By the same token, the potential application of principles emerging in the field of digital preservation extends well beyond digital resources to encompass works of art and creative enterprises whose material constitution paradoxically threatens them with extinction. The paramount role played by 'process' (as opposed to 'object') has led us to explore various metadata development initiatives, including an extensive online questionnaire developed by the Variable Media Initiative at the Guggenheim Museum in New York, and the IFLA Requirements of Bibliographic Standards, with a view to combining elements of each and developing a new model to accommodate the complex documentation of the life cycles of new media artworks and of digital objects.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 3

**Pages** 327-339

**Date** September 1, 2006

**DOI** 10.1093/lc/fqi064

**Short Title** Cryogenics and Creativity

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/3/327>

**Accessed** Sun Apr 5 05:07:25 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:07:25 2009

**Modified** Sun Apr 5 05:07:25 2009

### Tags:

computer\_science, cultural\_heritage, project\_team, single\_institution

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## To What Extent are Surnames Words? Comparing Geographic Patterns of Surname and Dialect Variation in the Netherlands

**Type** Journal Article

**Author** Franz Manni

**Author** Wilbert Heeringa

**Author** John Nerbonne

**Abstract** Since the early studies by Sokal (1988) and Cavalli-Sforza et al. (1989), there has been an increasing interest in depicting the history of human migrations by comparing genetic and linguistic differences that mirror different aspects of human history. Most of the literature concerns continental or macroregional patterns of variation, while regional and microregional scales were investigated less successfully. In this article we concentrate on the Netherlands, an area of only 40,000 km<sup>2</sup>. The focus of the article is on the analysis of surnames, which have been proven to be reliable genetic markers since in patrilineal systems they are transmitted--virtually unchanged--along generations, similar to a genetic locus on the Y-chromosome. We shall compare their distribution to that of dialect pronunciations, which are clearly culturally transmitted (children learn one of the linguistic varieties they are exposed to, normally that of their peers in the same area or that of their families). Since surnames, at the time of their introduction, were words subject to the same linguistic processes that otherwise result in dialect differences, one might expect the distribution of surnames to be correlated with dialect pronunciation differences. But we shall argue that once the collinear effects of geography on both genetics and cultural transmission are taken into account, there is in fact no statistically significant association between the two. We show that surnames cannot be taken as a proxy for dialect variation, even though they can be safely used as a proxy to Y-chromosome genetic variation. We work primarily with regression analyses, which show that both surname and dialect variation are strongly correlated with geographic distance. In view of this strong correlation, we focus on the residuals of the regression, which seeks to explain genetic and linguistic variation on the basis of geography (where geographic distance is the independent variable, and surname diversity or linguistic diversity is the dependent variable). We then seek a more detailed portrait of the geographic patterns of variation by identifying the barriers' (namely the areas where the residuals are greatest) by applying the Monmonier algorithm. We find the results historically and geographically insightful, hopefully leading to a deeper understanding of the role of the local migrations and cultural diffusion that are responsible for surname and dialect diversity.

**Publication** Lit Linguist Computing  
**Volume** 21  
**Issue** 4  
**Pages** 507-527  
**Date** November 1, 2006  
**DOI** 10.1093/lc/fql040  
**Short Title** To What Extent are Surnames Words?  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/4/507>  
**Accessed** Sun Apr 5 05:01:32 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:01:31 2009  
**Modified** Sun Apr 5 05:01:31 2009

**Tags:**

linguistics, literary\_dna, multi-country, multi-institutional

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Sentence-length and Authorship Attribution: the Case of Oliver Goldsmith

**Type** Journal Article  
**Author** David Mannion  
**Author** Peter Dixon  
**Abstract** The sentence-lengths of sixteen essays by Goldsmith are examined in relation to data from ten essays (we call these doubtfuls') which have been attributed to him. Comparisons between the doubtfuls' and the known Goldsmiths are made with reference to the  $\chi^2$  goodness-of-fit test, and the method of reciprocal averaging. The Goldsmith essays form a close group, with four of the doubtful essays well outside, two less remote and four within the Goldsmith cluster. Comparison with fifty essays by nine of Goldsmith's contemporaries reveals the distinctiveness of his sentence-length patterns, and strengthens the probability that the four least doubtful essays are his. In the case of Goldsmith, then, sentence-length may be considered a reliable stylistic marker.  
**Publication** Lit Linguist Computing  
**Volume** 19

**Issue** 4  
**Pages** 497-508  
**Date** November 1, 2004  
**DOI** 10.1093/llc/19.4.497  
**Short Title** Sentence-length and Authorship Attribution  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/19/4/497>  
**Accessed** Sun Apr 5 05:41:11 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:41:11 2009  
**Modified** Sun Apr 5 05:41:11 2009

**Tags:**

multi-institutional, stylistics

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Tools for Searching, Annotation and Analysis of Speech, Music, Film and Video A Survey

**Type** Journal Article  
**Author** Alan Marsden  
**Author** Adrian Mackenzie  
**Author** Adam Lindsay  
**Author** Harriet Nock  
**Author** John Coleman  
**Author** Greg Kochanski  
**Abstract** This article examines the actual and potential use of software tools in research in the arts and humanities focusing on audiovisual (AV) materials such as recorded speech, music, video and film. The quantity of such materials available to researchers is massive and rapidly expanding. Researchers need to locate the material of interest in the vast quantity available, and to organize and process the material once collected. Locating and organizing often depend on metadata and tags to describe the actual content, but standards for metadata for AV materials are not widely adopted. Content-based search is becoming possible for speech, but is still beyond the horizon for music, and even more distant for video. Copyright

protection hampers research with AV materials, and Digital Rights Management (DRM) systems threaten to prevent research altogether. Once material has been located and accessed, much research proceeds by annotation, for which many tools exist. Many researchers make some kind of transcription of materials, and would value tools to automate this process. Such tools exist for speech, though with important limits to their accuracy and applicability. For music and video, researchers can make use of visualizations. A better understanding (in general terms) by researchers of the processes carried out by computer software and of the limitations of its results would lead to more effective use of Information and Communications Technology (ICT).

**Publication** Lit Linguist Computing  
**Volume** 22  
**Issue** 4  
**Pages** 469-488  
**Date** November 1, 2007  
**DOI** 10.1093/lc/fqm021  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/4/469>  
**Accessed** Sun Apr 5 04:35:20 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:35:20 2009  
**Modified** Sun Apr 5 04:35:20 2009

**Tags:**

multi-institutional, project\_team, tool\_development

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Word Sense Disambiguation Using Target Language Corpus in a Machine Translation System

**Type** Journal Article  
**Author** Tayebeh Mosavi Miangah  
**Author** Ali Delavar Khalafi  
**Abstract** This article studies different aspects of a new approach to word sense disambiguation using statistical information gained from a monolingual corpus of

the target language. Here, the source language is English and the target is Persian, and the disambiguation method can be directly applied in the system of English-to-Persian machine translation for solving lexical ambiguity problems in this system. Unlike other disambiguation approaches, using corpora for handling the problem, which use the Most Likelihood Model in their statistical works, this article proposes the Random Numbers Model. We believe that this model is more reasonable from the scientific point of view and find that it offers the most precise and accurate results. This method has been tested for a selected set of English texts containing multiple-meaning words with respect to Persian language and the results are encouraging.

**Publication** Lit Linguist Computing  
**Volume** 20  
**Issue** 2  
**Pages** 237-249  
**Date** June 1, 2005  
**DOI** 10.1093/lc/fqi029  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/20/2/237>  
**Accessed** Sun Apr 5 05:32:27 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:32:27 2009  
**Modified** Sun Apr 5 05:32:27 2009

**Tags:**

NLP, single\_institution

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## A Statistical Analysis of Editorial Influence and Author Character Similarities in 1990s New Yorker Fiction

**Type** Journal Article  
**Author** Katherine L. Milkman  
**Author** Rene Carmona  
**Author** William Gleason

**Abstract** We present a quantitative analysis of 442 pieces of fiction published between 5 October 1992 and 17 September 2001 in the New Yorker magazine. We address two independent questions using the same data set. First, we examine whether changes in the Executive Editor or Fiction Editor are associated with significant changes in the type of fiction published at the New Yorker. Second, we examine whether New Yorker authors write fiction more often than not about characters with whom they share demographic traits. We find that changes in Fiction Editor at the New Yorker are associated with numerous significant, quantifiable changes in the magazine's fiction and that these effects are greater than those associated with a change in the New Yorker's Executive Editor. We also find that authors of New Yorker fiction write significantly more often than not about protagonists who share their race, gender, and country of origin and who are within or below their age range. The same is true of secondary characters except in the case of gender.

**Publication** Lit Linguist Computing  
**Volume** 22  
**Issue** 3  
**Pages** 305-328  
**Date** September 1, 2007  
**DOI** 10.1093/lc/fqm011  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/3/305>  
**Accessed** Sun Apr 5 04:39:30 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:39:30 2009  
**Modified** Sun Apr 5 04:39:30 2009

### Tags:

multi-institutional, statistics

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## The Elements of Drawing

**Type** Journal Article  
**Author** Jonathan Miller  
**Author** Rupert Shepherd

**Abstract** This paper describes the work of The Elements of Drawing, a project to digitize the teaching collection assembled by John Ruskin at the University of Oxford. It outlines John Ruskin's links with Oxford, his reasons for creating the collection as an aid to his teaching of drawing at the University, and the ways in which he organized and catalogued the collection. It then discusses the particular benefits which digitization brings to the collection, and outlines the methods being used to digitize the collection as a series of images, texts, catalogue data, and associated metadata, and how it is being made available over the world wide web.

**Publication** Lit Linguist Computing

**Volume** 19

**Issue** 3

**Pages** 385-396

**Date** September 1, 2004

**DOI** 10.1093/lc/19.3.385

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/19/3/385>

**Accessed** Sun Apr 5 05:46:18 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:46:18 2009

**Modified** Sun Apr 5 05:46:18 2009

### Tags:

digitization, multi-discipline, multi-institutional, project\_team

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Applying the Semantic Web: The VICODI Experience in Creating Visual Contextualization for History

**Type** Journal Article

**Author** Gabor Nagypal

**Author** Richard Deswarte

**Author** Jan Oosthoek

**Abstract** Semantic Web applications in the humanities that visualize knowledge are still few and far between. The Visual Contextualization of Digital Content (VICODI) project brought together Semantic Web technologies with the concepts of

contextualization and visualization of knowledge, an approach which we term visual contextualization. The goal was to enhance users' understanding of digital content in the domain of history. It succeeded in doing this by creating an ontology-based web portal of European history where extra historical knowledge or context' is added to resources and visualized through textual hyperlinks and interactive Scalable Vector Graphics historical maps. VICODI also created a history-specific ontology. In this article the novel approach of visual contextualization is introduced in conjunction with a detailed explanation of the core elements of the VICODI portal. The article also addresses several of the problems encountered in developing a Semantic Web application for a humanities domain.

**Publication** Lit Linguist Computing  
**Volume** 20  
**Issue** 3  
**Pages** 327-349  
**Date** September 1, 2005  
**DOI** 10.1093/llc/fqi037  
**Short Title** Applying the Semantic Web  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/20/3/327>  
**Accessed** Sun Apr 5 05:29:01 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:29:01 2009  
**Modified** Sun Apr 5 05:29:01 2009

### Tags:

multi-country, multi-institutional, project\_team, tool\_development

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Progress in Dialectometry: Toward Explanation

**Type** Journal Article  
**Author** John Nerbonne  
**Author** William Kretschmar

**Abstract** Dialectometric techniques analyze linguistic variation quantitatively, allowing one to aggregate over what are frequently rebarbative geographic patterns of individual linguistic variants, such as which word is used for a particular concept in a language area, or which sounds are used in particular words. This leads to general formulations of the relation between linguistic variation and explanatory factors. Dialectometric techniques are maturing continuously, paving the way to genuinely new opportunities for the explanation of linguistic variation. These include, most prominently, techniques for analyzing syntactic variation, techniques for comparing the relative importance of different individual linguistic variables, techniques for comparing the relative importance of linguistic levels such as pronunciation, vocabulary, and/or prosody, and many more. This article serves as an introduction to a special issue of *Literary and Linguistic Computing* devoted to presenting a new work constituting *Progress in Dialectometry: Toward Explanation*.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 4

**Pages** 387-397

**Date** November 1, 2006

**DOI** 10.1093/lc/fql034

**Short Title** Progress in Dialectometry

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/21/4/387>

**Accessed** Sun Apr 5 05:01:26 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:01:26 2009

**Modified** Sun Apr 5 05:01:26 2009

### Tags:

linguistics, multi-country, multi-institutional

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Use of the Chi-Squared Test to Examine Vocabulary Differences in English Language Corpora Representing Seven Different Countries

**Type** Journal Article

**Author** Michael P. Oakes

**Author** Malcolm Farrow

**Abstract** The chi-squared test is used to find the vocabulary most typical of seven different ICAME corpora, each representing the English used in a particular country. In a closely related study, Leech and Fallon (1992, Computer corpora - what do they tell us about culture? ICAME Journal, 16: 29-50) found differences in the vocabulary used in the Brown Corpus of American English and that the Lancaster-Oslo-Bergen Corpus of British English. They were mainly interested in those vocabulary differences which they assumed to be due to cultural differences between the United States and Britain, but we are equally interested in vocabulary differences which reveal linguistic preferences in the various countries in which English is spoken. Whether vocabulary differences are cultural or linguistic in nature, they can be used for the automatic classification according to variety of English of texts of unknown provenance. The extent to which the vocabulary differences between the corpora represent vocabulary differences between the varieties of English as a whole depends on the extent to which the corpora represent the full range of topics typical of their associated cultures, and thus there is a need for corpora designed to represent the topics and vocabulary of cultures or dialects, rather than stratified across a set range of topics and genres. This will require methods to determine the range of topics addressed in each culture, then methods to sample adequately from each topical domain.

**Publication** Lit Linguist Computing

**Volume** 22

**Issue** 1

**Pages** 85-99

**Date** April 1, 2007

**DOI** 10.1093/lc/fq1044

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/22/1/85>

**Accessed** Sun Apr 5 04:55:36 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:55:36 2009

**Modified** Sun Apr 5 04:55:36 2009

### Tags:

linguistics, multi-institutional

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Mining millions of metaphors

**Type** Journal Article

**Author** Brad Pasanek

**Author** D. Sculley

**Abstract** One of the first decisions made in any research concerns the selection of an appropriate scale of analysis--are we looking out into the heavens, or down into atoms? To conceive a digital library as a collection of a million books may restrict analysis to only one level of granularity. In this article, we examine the consequences and opportunities resulting from a shift in scale, where the desired unit of interpretation is something smaller than a text: it is a keyword, a motif, or a metaphor. A million books distilled into a billion meaningful components become raw material for a history of language, literature, and thought that has never before been possible. While books herded into genres and organized by period remain irregular, idiosyncratic, and meaningful in only the most shifting and context-dependent ways, keywords or metaphors are lowest common denominators. At the semantic level--the level of words, images, and metaphors--long-term regularity and patterns emerge in collection, analysis, and taxonomy. This article follows the foregoing course of thought through three stages: first, the manual curation of a high quality database of metaphors; second, the expansion of this database through automated and human-assisted techniques; finally, the description of future experiments and opportunities for the application of machine learning, data mining, and natural language processing techniques to help find patterns and meaning concealed at this important level of granularity.

**Publication** Lit Linguist Computing

**Volume** 23

**Issue** 3

**Pages** 345-360

**Date** September 1, 2008

**DOI** 10.1093/lc/fqn010

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/23/3/345>

**Accessed** Sun Apr 5 04:16:37 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:16:37 2009

**Modified** Sun Apr 5 04:16:37 2009

### Tags:

computer\_science, english, multi-institutional, text mining

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Discovery of Language Resources on the Web: Information Extraction from Heterogeneous Documents

**Type** Journal Article**Author** Viktor Pekar**Author** Richard Evans

**Abstract** The present article is concerned with the problem of automatic database population via information extraction (IE) from web pages obtained from heterogeneous sources, such as those retrieved by a domain crawler. Specifically, we address the task of filling single multi-field templates from individual documents, a common scenario that involves free-format documents with the same communicative goal such as job adverts, CVs, or meeting/seminar announcements. We discuss challenges that arise in this scenario and propose solutions to them at different levels of the processing of web page content. Our main focus is on the issue of information extraction, which we address with a two-step machine learning approach that first aims to determine segments of a page that are likely to contain relevant facts and then delimits specific natural language expressions with which to fill template fields. We also present a range of techniques for the enrichment of web pages with semantic annotations, such as recognition of named entities, domain terminology and coreference resolution, and examine their effect on the information extraction method. We evaluate the developed IE system on the task of automatically populating a database with information on language resources available on the web.

**Publication** Lit Linguist Computing**Volume** 22**Issue** 3**Pages** 329-343**Date** September 1, 2007**DOI** 10.1093/lc/fqm010**Short Title** Discovery of Language Resources on the Web**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/3/329>**Accessed** Sun Apr 5 04:39:32 2009**Repository** HighWire**Date Added** Sun Apr 5 04:39:32 2009**Modified** Sun Apr 5 04:39:32 2009

**Tags:**

single\_institution

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

**Rule-based Search in Text Databases with Nonstandard Orthography****Type** Journal Article**Author** Thomas Pilz**Author** Wolfram Luther**Author** Norbert Fuhr**Author** Ulrich Ammon

**Abstract** In this article, we describe our interdisciplinary project 'Rule-based search in text databases with nonstandard orthography (RSNSR)' in support of the conservation of cultural heritage, especially for the German reception of the philosopher Nietzsche. We present a rule-based fuzzy search engine that allows users to retrieve text data independently of its orthographical realization. The rules used are derived from statistical analyses, historical publications, linguistic principles, and expert knowledge. Our Web-based tool is intended for experts as well as interested amateurs. Along with its present features, further functions are currently worked out. Among them are automatic rule derivation and finer result classification through a generalized Levenshtein similarity measure. Our work is associated with the recently launched project Deutsch Diachron Digital (DDD) to build a complete diachronic corpus of German for the first time with texts from the ninth century (Old High German) to the present (Modern German).

**Publication** Lit Linguist Computing**Volume** 21**Issue** 2**Pages** 179-186**Date** June 1, 2006**DOI** 10.1093/lc/fq1020**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/21/2/179>**Accessed** Sun Apr 5 05:12:14 2009**Repository** HighWire**Date Added** Sun Apr 5 05:12:14 2009

**Modified** Sun Apr 5 05:12:14 2009

**Tags:**

computer\_science, german, multi-discipline, project\_team, single\_institution, tool\_development

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

The Identification of Spelling Variants in English and German Historical Texts: Manual or Automatic?

**Type** Journal Article

**Author** Thomas Pilz

**Author** Andrea Ernst-Gerlach

**Author** Sebastian Kempken

**Author** Paul Rayson

**Author** Dawn Archer

**Abstract** In this article, we describe the respective approaches we have taken when addressing issues of spelling variation in German and English historical texts. More specifically, we describe an experiment to evaluate automatic techniques for the development of letter replacement heuristics against manually created gold standards of known letter replacements rules. As will become clear, the motivation for the research differs according to the team of researchers: the German researchers are seeking to develop a search engine for historical texts; the English researchers want to improve the results obtained when applying corpus linguistic techniques (developed for modern language) to historical data. However, the respective teams do share a longer term goal of assessing whether it is possible to develop a generic spelling detection tool for Indo-European languages.

**Publication** Lit Linguist Computing

**Volume** 23

**Issue** 1

**Pages** 65-72

**Date** April 1, 2008

**DOI** 10.1093/lc/fqm044

**Short Title** The Identification of Spelling Variants in English and German Historical Texts

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/23/1/65>

**Accessed** Sun Apr 5 04:29:42 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:29:42 2009  
**Modified** Sun Apr 5 04:29:42 2009

**Tags:**

computer\_science, linguistics, multi-country, multi-discipline, multi-institutional, project\_team

**Notes:**

Thomas Pilz, Andrea Ernst-Gerlach and Sebastian Kempken  
Department of Computer Science and Applied Cognitive Science,  
Faculty of Engineering, University of Duisburg-Essen, D-47048  
Duisburg, Lotharstr. 65, Germany  
Paul Rayson  
Computing Department, Infolab21, Lancaster University, Lancaster  
LA1 4WA, UK  
Dawn Archer  
Department of Humanities, University of Central Lancashire,  
Preston PR1 2HE, UK

.....

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Functional Disambiguation Based on Syntactic Structures

**Type** Journal Article  
**Author** Octavio Santana Suarez  
**Author** Jose Rafael Perez Aguiar  
**Author** Luis Losada Garcia  
**Author** Francisco Javier Carreras Riudavets  
**Abstract** This article presents a disambiguation method which diminishes the functional combinations of the words of a sentence taking into account the context in which they appear. This process is built in two phases: the first phase is based on the local syntactic structures of the Spanish language and reaches an average yield of 87%. The second one is supported by syntactic tree representation and pushes the

results up to an approximate high end of 96%. This process constitutes the starting point towards an automated syntactic analysis.

**Publication** Lit Linguist Computing  
**Volume** 21  
**Issue** 2  
**Pages** 187-197  
**Date** June 1, 2006  
**DOI** 10.1093/llc/fql016  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/2/187>  
**Accessed** Sun Apr 5 05:12:16 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:12:16 2009  
**Modified** Sun Apr 5 05:12:16 2009

### Tags:

single\_institution, stylistics

### Attachments

HighWire Full Text PDF  
HighWire Snapshot

---

## Cross-collection Searching: A Pandora's Box or the Holy Grail?

**Type** Journal Article  
**Author** Susan Schreibman  
**Author** Jennifer O'Brien Roper  
**Author** Gretchen Gueguen  
**Abstract** As digital libraries have expanded to absorb existing collections as well as to create new ones, it has become clear that cross collection discovery is not simply desirable, but is increasingly a necessity demanded by users. Similarly, in the digital humanities community, thematic research collections once distinct from one another now would seem to benefit from interoperability. However, efforts to aggregate disparate resources are often stymied by differing metadata schema and controlled vocabulary. Using the lessons learned from the Thomas MacGreevy Archive, The University of Maryland Libraries designed its digital repository to provide for discovery across object types and collections using Fedora as the underlying architecture. To facilitate access to multiple collections within one

repository, University of Maryland developed a flexible metadata standard. This metadata schema is used to describe varying types of materials at varying levels of granularity, while allowing for controlled vocabularies appropriate to specific collections.

**Publication** Lit Linguist Computing  
**Volume** 23  
**Issue** 1  
**Pages** 13-25  
**Date** April 1, 2008  
**DOI** 10.1093/lc/fqm039  
**Short Title** Cross-collection Searching  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/23/1/13>  
**Accessed** Sun Apr 5 04:29:40 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:29:40 2009  
**Modified** Sun Apr 5 04:29:40 2009

### Tags:

Metadata, project\_team, single\_institution

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Meaning and mining: the impact of implicit assumptions in data mining for the humanities

**Type** Journal Article  
**Author** D. Sculley  
**Author** Bradley M. Pasanek  
**Abstract** As the use of data mining and machine learning methods in the humanities becomes more common, it will be increasingly important to examine implicit biases, assumptions, and limitations these methods bring with them. This article makes explicit some of the foundational assumptions of machine learning methods, and presents a series of experiments as a case study and object lesson in the potential pitfalls in the use of data mining methods for hypothesis testing in literary scholarship. The worst dangers may lie in the humanist's; ability to

interpret nearly any result, projecting his or her own biases into the outcome of an experiment--perhaps all the more unwittingly due to the superficial objectivity of computational methods. We argue that in the digital humanities, the standards for the initial production of evidence should be even more rigorous than in the empirical sciences because of the subjective nature of the work that follows. Thus, we conclude with a discussion of recommended best practices for making results from data mining in the humanities domain as meaningful as possible. These include methods for keeping the the boundary between computational results and subsequent interpretation as clearly delineated as possible.

**Publication** Lit Linguist Computing  
**Volume** 23  
**Issue** 4  
**Pages** 409-424  
**Date** December 1, 2008  
**DOI** 10.1093/llc/fqn019  
**Short Title** Meaning and mining  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/23/4/409>  
**Accessed** Sun Apr 5 04:08:59 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:08:59 2009  
**Modified** Sun Apr 5 04:08:59 2009

**Tags:**

computer\_science, english, multi-institutional, text mining

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

**Callimachus--Avoiding the Pitfalls of XML for Collaborative Text Analysis**

**Type** Journal Article  
**Author** Jeff Smith  
**Author** Joel Deshaye  
**Author** Peter Stoicheff  
**Abstract** We present our experience in developing an on-line infrastructure to support collaborative analysis of text, which we distinguish from existing, well-explored

efforts to create annotative electronic editions. Using Faulkner's *The Sound and the Fury* as our primary text case, we outline the features and rationale of our collaborative framework, called Callimachus. We present our findings concerning that text and explore how these findings only became possible after breaking with the received wisdom concerning the application of XML and the Text Encoding Initiative to such analytical projects.

**Publication** Lit Linguist Computing  
**Volume** 21  
**Issue** 2  
**Pages** 199-218  
**Date** June 1, 2006  
**DOI** 10.1093/lc/fql021  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/2/199>  
**Accessed** Sun Apr 5 05:12:18 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:12:18 2009  
**Modified** Sun Apr 5 05:12:18 2009

**Tags:**

computer\_science, single\_institution, tool\_development

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Corpus Tools and Methods, Today and Tomorrow: Incorporating Linguists' Manual Annotations

**Type** Journal Article  
**Author** Nicholas Smith  
**Author** Sebastian Hoffmann  
**Author** Paul Rayson  
**Abstract** Today's corpus tools offer the user a wide range of features that greatly facilitate the linguistic analysis of large amounts of authentic language data (e.g. frequency distributions, collocations, keywords, etc.). However, these tools typically fail to address the fundamental need of the linguist to add interpretive information to a concordance or query result, by coding individual concordance lines for structural,

functional, discursal, and other features in a flexible way. The ability to add such qualitative data is indispensable to a fuller understanding of the phenomenon under investigation as it allows the linguist to produce more rigorous descriptions--and theories--about language in use. Our article has two aims: first, to assess the merits and drawbacks of existing solutions, by surveying what can be achieved using state-of-the-art corpus tools and generic database software; second, we draw up a set of desiderata and recommendations for the incorporation of flexible encoding features into future corpus tools. We describe an initial step in this direction, with a recent enhancement to the BNCweb corpus analysis software. More generally, we hope our suggestions will lead to linguists and software developers working together more closely to ensure that the needs of the former are provided for by the available technology.

**Publication** Lit Linguist Computing  
**Volume** 23  
**Issue** 2  
**Pages** 163-180  
**Date** June 1, 2008  
**DOI** 10.1093/lc/fqn004  
**Short Title** Corpus Tools and Methods, Today and Tomorrow  
**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/23/2/163>  
**Accessed** Sun Apr 5 04:23:52 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:23:52 2009  
**Modified** Sun Apr 5 04:23:52 2009

**Tags:**

computer\_science, english, linguistics, multi-institutional, project\_team, tool\_development

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

**Optimal Strategies for Accurate Transcription**

**Type** Journal Article  
**Author** Matthew Spencer  
**Author** Christopher J. Howe

**Abstract** Accurate transcription is difficult and time-consuming. It is therefore worth choosing the transcription strategy that will yield the smallest number of errors for a given total effort. We use a mathematical model of the transcription process to compare two basic strategies: a single transcription with repeated checking, and a pair of transcriptions with a smaller amount of checking. Our model for checking is an adequate description of the rate of error detection in a real transcription. We show how to optimize the proportion of effort allocated to checking locations where the two transcriptions disagree, and discuss the factors that favour either of the strategies. We suggest how one might design an optimal transcription strategy in practice.

**Publication** Lit Linguist Computing

**Volume** 21

**Issue** 3

**Pages** 353-362

**Date** September 1, 2006

**DOI** 10.1093/lc/fqi030

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/3/353>

**Accessed** Sun Apr 5 05:07:27 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:07:27 2009

**Modified** Sun Apr 5 05:07:27 2009

### Tags:

single\_institution, statistics

### Notes:

Matthew Spencer and Christopher J. Howe  
University of Cambridge, UK

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

Employing Thematic Variables for Enhancing Classification Accuracy  
Within Author Discrimination Experiments

**Type** Journal Article

**Author** George Tambouratzis

**Author** Marina Vassiliou

**Abstract** This article reports on experiments performed with a large corpus, aiming at separating texts according to the author style. The study initially focusses on whether the classification accuracy regarding the author identity may be improved, if the text topic is known in advance. The experimental results indicate that this kind of information contributes to more accurate author recognition. Furthermore, as the diversity of a topic set increases, the classification accuracy is reduced. In general, the experimental results indicate that taking into account knowledge regarding the text topic can lead to the construction of specialized models for each author with higher classification accuracy. For example, by focussing on a specific topic, the accuracy with which the author identity is determined increases, the exact amount depending on the specific topic. This also applies when the topic of the text is more broadly determined, as a set of topic categories. In an associated task, the most salient parameters within an 85-parameter vector are studied, for a number of subsets of the corpus, where each subset contains speeches from a single topic. These studies indicate that the salient parameters are the same for the different subsets. Two fixed data vectors have been defined, using 16 and 25 parameters, respectively. The classification accuracy obtained, even with the smallest data vector, is only 5% less than with the complete vector. This indicates that the parameters retained in the reduced vectors bear a large amount of discriminatory information and suffice for an accurate classification of the corpus.

**Publication** Lit Linguist Computing

**Volume** 22

**Issue** 2

**Pages** 207-224

**Date** June 1, 2007

**DOI** 10.1093/lc/fqm003

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/2/207>

**Accessed** Sun Apr 5 04:45:23 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:45:23 2009

**Modified** Sun Apr 5 04:45:23 2009

### Tags:

single\_institution, stylistics

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Discriminating the Registers and Styles in the Modern Greek Language-Part 1: Diglossia in Stylistic Analysis

**Type** Journal Article**Author** George Tambouratzis**Author** Stella Markantonatou**Author** Nikolaos Hairetakis**Author** Marina Vassiliou**Author** George Carayannis**Author** Dimitrios Tambouratzis

**Abstract** This article describes a method for discriminating among registers of Modern Greek and among authors within a given register. Two issues have been investigated: (a) whether register discrimination can successfully exploit linguistic information reflecting the evolution of a language (such as diglossia features of the Modern Greek language) and (b) what kind of linguistic information and which statistical techniques may be applied to author discrimination within one register. Using clustering techniques and variables reflecting the diglossia situation, we have successfully discriminated registers in Modern Greek. However, diglossia information on its own has not been shown sufficient for author discrimination within one register. Instead, other linguistic features, including PoS distribution and discourse tendencies, have been combined with methods such as discriminant analysis in order to obtain a high degree of accuracy.

**Publication** Lit Linguist Computing**Volume** 19**Issue** 2**Pages** 197-220**Date** June 1, 2004**DOI** 10.1093/lc/19.2.197**Short Title** Discriminating the Registers and Styles in the Modern Greek Language-Part 1**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/19/2/197>**Accessed** Sat Apr 18 04:31:18 2009**Repository** HighWire**Date Added** Sat Apr 18 04:31:18 2009**Modified** Sat Apr 18 04:31:18 2009

**Tags:**

linguistics, multi-institutional

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Discriminating the Registers and Styles in the Modern Greek Language-Part 2: Extending the Feature Vector to Optimize Author Discrimination

**Type** Journal Article**Author** George Tambouratzis**Author** Stella Markantonatou**Author** Nikolaos Hairetakis**Author** Marina Vassiliou**Author** George Carayannis**Author** Dimitrios Tambouratzis

**Abstract** This article describes a method for discriminating among authors within a given register of Modern Greek. The focus here is to determine to what extent the stylistic differences among authors can be detected with a high degree of accuracy for a set of texts belonging to a well-defined register. To that end, the chosen register is characterized by a well-defined sub-language, from which a corpus of more than 1,000 documents has been created. To discriminate the texts according to author style, a series of experiments have been performed using statistical techniques. Each text has been represented by a vector covering several linguistic aspects, in an effort to determine the most effective style markers. The experimental results indicate that the proposed approach can successfully separate the author styles for a given register. An extensive study of the effectiveness of the different variable categories has been performed. For instance, diglossia information on its own is not sufficient for author discrimination. Instead, a systematic evaluation process indicates that part-of-speech, structural and algorithmically derived lemma-frequency variables are the most important style markers, their use leading to an author discrimination accuracy exceeding 90%.

**Publication** Lit Linguist Computing**Volume** 19**Issue** 2**Pages** 221-242

**Date** June 1, 2004  
**DOI** 10.1093/llc/19.2.221  
**Short Title** Discriminating the Registers and Styles in the Modern Greek Language-Part 2  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/19/2/221>  
**Accessed** Sun Apr 5 05:50:58 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:50:58 2009  
**Modified** Sun Apr 5 05:50:58 2009

**Tags:**

multi-institutional, stylistics

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## An algorithm for automated authorship attribution using neural networks

**Type** Journal Article  
**Author** Matt Tearle  
**Author** Kye Taylor  
**Author** Howard Demuth  
**Abstract** We present an algorithm as evidence of the possibility of a truly automated stylometric authorship attribution tool, based on committees of artificial neural networks. Neural networks have an advantage over traditional statistical stylometry in that they are inherently nonlinear, and therefore can consider nonlinear interactions between stylometric variables. The algorithm presented (1) is intended to demonstrate the feasibility of an automated approach using neural networks and (2) highlights important areas for further research. We present results of two separate test experiments--Shakespeare and Marlowe, and the Federalist Papers--as a demonstration of the method's generality. In both cases, our algorithm produces committees that correctly predict the test works, without requiring the usual precursory statistical study to determine efficacious stylometric measures.  
**Publication** Lit Linguist Computing  
**Volume** 23  
**Issue** 4

**Pages** 425-442

**Date** December 1, 2008

**DOI** 10.1093/llc/fqn022

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/23/4/425>

**Accessed** Sun Apr 5 04:09:01 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:09:00 2009

**Modified** Sun Apr 5 04:09:00 2009

### Tags:

single\_institution, stylistics

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Downs and Acrosses: Textual Markup on a Stroke Level

**Type** Journal Article

**Author** Melissa Terras

**Author** Paul Robertson

**Abstract** Textual encoding is one of the main focuses of Humanities Computing. However, existing encoding schemes and initiatives focus on text' from the character level upwards, and are of little use to scholars, such as papyrologists and palaeographers, who study the constituent strokes of individual characters. This paper discusses the development of a markup system used to annotate a corpus of images of Roman texts, resulting in an XML representation of each character on a stroke by stroke basis. The XML data generated allows further interrogation of the palaeographic data, increasing the knowledge available regarding the palaeography of the documentation produced by the Roman Army. Additionally, the corpus was used to train an Artificial Intelligence system to effectively read' in stroke data of unknown text and output possible, reliable, interpretations of that text: the next step in aiding historians in the reading of ancient texts. The development and implementation of the markup scheme is introduced, the results of our initial encoding effort are presented, and it is demonstrated that textual markup on a stroke level can extend the remit of marked-up digital texts in the humanities.

**Publication** Lit Linguist Computing

**Volume** 19  
**Issue** 3  
**Pages** 397-414  
**Date** September 1, 2004  
**DOI** 10.1093/llc/19.3.397  
**Short Title** Downs and Acrosses  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/19/3/397>  
**Accessed** Sun Apr 5 05:46:20 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:46:20 2009  
**Modified** Sun Apr 5 05:46:20 2009

**Tags:**

markup, multi-country, multi-institutional

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Visual Knowledge: Textual Iconography of the Quixote, a Hypertextual Archive

**Type** Journal Article  
**Author** Eduardo Urbina  
**Author** Richard Furuta  
**Author** Steven Escar Smith  
**Author** Neal Audenaert  
**Author** Jie Deng  
**Author** Carlos Monroy  
**Abstract** Ever since its initial publication four hundred years ago, thousands of editions, most often illustrated, have been published of Cervantes' masterpiece, Don Quixote. Imagery has become an integral part of the reception and interpretation of the text. To date, a comprehensive collection of these images, the textual iconography of the Quixote, has not been published. We report in this paper on overcoming two key obstacles: limitations on the availability of materials and limitations due to the technical and financial characteristics of print-based dissemination. Our digital iconography makes a rich artistic tradition accessible to

readers for the first time, and reveals a wealth of information about the historical, cultural, and literary contexts into which the Quixote has been placed.

**Publication** Lit Linguist Computing  
**Volume** 21  
**Issue** 2  
**Pages** 247-258  
**Date** June 1, 2006  
**DOI** 10.1093/lc/fql023  
**Short Title** Visual Knowledge  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/21/2/247>  
**Accessed** Sun Apr 5 05:12:22 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:12:22 2009  
**Modified** Sun Apr 5 05:12:22 2009

### Tags:

computer\_science, Literature, multi-discipline, project\_team, single\_institution, tool\_development

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Editorial Theory and Practice in Flanders and the Centre for Scholarly Editing and Document Studies

**Type** Journal Article  
**Author** Bert Van Raemdonck  
**Author** Edward Vanhoutte  
**Abstract** After Professor Marcel De Smedt of the University of Leuven introduced scholarly editing of modern texts as a discipline in Flanders in the 1980s, the worrying fact emerged during the last decade of the twentieth century that Flemish universities and scholarly research groups were falling well behind in the field of scholarly editing. As a reaction, the inter-university task force Genese was founded in 1993 with its main goal to promote and coordinate the theories and practice of scholarly editing in Flanders. The next decisive step was taken by the Royal Academy of Dutch Language and Literature (Koninklijke Academie voor Nederlandse Taal- en Letterkunde--KANTL) when they decided to make scholarly

editing their primary objective as of January 1998. That decision paved the way for the founding of the Centre for Scholarly Editing and Document Studies (Centrum voor Teksteditie en Bronnenstudie--CTB), which started on 1 August 2000 as a research institute of the Academy, and which has become the centre of expertise in the field of (electronic) scholarly editing in the Low Countries.

**Publication** Lit Linguist Computing  
**Volume** 19  
**Issue** 1  
**Pages** 119-127  
**Date** April 1, 2004  
**DOI** 10.1093/lc/19.1.119  
**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/19/1/119>  
**Accessed** Sun Apr 5 05:53:54 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:53:54 2009  
**Modified** Sun Apr 5 05:53:54 2009

**Tags:**

project\_team, single\_institution

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Presentational and Representational Issues in Correspondence Reconstruction and Sorting

**Type** Journal Article  
**Author** Edward Vanhoutte  
**Author** Ron Van den Branden  
**Abstract** When theorizing about correspondence reconstruction and sorting, we first need both a definition of a letter' and of correspondence'. In this article, we propose such definitions and investigate the impact of our understanding of what correspondence reconstruction is on the production of the DALF formal framework for the transcription of epistolary material. The focus of the paper is on the instruments built into DALF for allowing all types of sorting and classification, hence correspondence reconstruction.

**Publication** Lit Linguist Computing  
**Volume** 19  
**Issue** 1  
**Pages** 45-54  
**Date** April 1, 2004  
**DOI** 10.1093/ljc/19.1.45  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/19/1/45>  
**Accessed** Sun Apr 5 05:53:50 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:53:50 2009  
**Modified** Sun Apr 5 05:53:50 2009

**Tags:**

markup, project\_team, single\_institution

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Delta for Middle Dutch Author and Copyist Distinction in Walewein

**Type** Journal Article  
**Author** Karina van Dalen-Oskam  
**Author** Joris van Zundert  
**Abstract** The Middle Dutch Arthurian romance Roman van Walewein ( Romance of Gawain') is attributed in the text itself to two authors, Penninc and Vostaert. Very little quantitative research into this dual authorship has been done. This article describes our progress in applying different non-traditional authorship attribution methods to the text of Walewein. After providing an introduction to the romance and an overview of earlier research, we evaluate previous statements on authorship and stylistics by applying both Yule's measure of lexical richness and Burrows's Delta. To find out whether these new methods would confirm or even enhance our present knowledge about the differences between the two authors, we applied an adapted version of John Burrows's Delta procedure. The adapted version seems to be able to distinguish the double authorship of the romance. It also helps us to confirm some and to reject other earlier statements about the position in the text where the second author started his work.  
**Publication** Lit Linguist Computing

**Volume** 22  
**Issue** 3  
**Pages** 345-362  
**Date** September 1, 2007  
**DOI** 10.1093/llc/fqm012  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/22/3/345>  
**Accessed** Sun Apr 5 04:39:33 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:39:33 2009  
**Modified** Sun Apr 5 04:39:33 2009

**Tags:**

single\_institution, stylistics

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## Modelling Features of Characters: Some Digital Ways to Look at Names in Literary Texts

**Type** Journal Article  
**Author** Karina van-Oskam  
**Author** Joris van Zundert  
**Abstract** In the context of ongoing research into new methods and techniques for literary research we describe a primary implementation of a web application called Autonom, intended to be developed into a framework for textual parsing algorithms that may be used by literary researchers to trace literary phenomena in texts. We describe the technical parsing fundamentals and good practices the development of the framework is based upon, we clarify different design considerations and choices and we present an overview of the current state of implementation and functionality. We also demonstrate the application of a proper name parsing algorithm implemented within the framework, meant to be the first step in a new method for the research of names in literary texts. The algorithm is tested on Karel Glastra van Loon's novel *Lisa's adem* (*Lisa's Breath*, 2001). We go into the results that this test has yielded so far and summarily describe some of the consequences for the analysis of the names in the novel. We conclude with a short description of the directions new developments could take.

**Publication** Lit Linguist Computing  
**Volume** 19  
**Issue** 3  
**Pages** 289-301  
**Date** September 1, 2004  
**DOI** 10.1093/lc/19.3.289  
**Short Title** Modelling Features of Characters  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/19/3/289>  
**Accessed** Sun Apr 5 05:46:16 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 05:46:16 2009  
**Modified** Sun Apr 5 05:46:16 2009

**Tags:**

single\_institution, tool\_development

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities through Statistical Analysis of User Log Data

**Type** Journal Article  
**Author** Claire Warwick  
**Author** Melissa Terras  
**Author** Paul Huntington  
**Author** Nikoleta Pappa  
**Abstract** There are now many online, digital resources in the humanities, and their creation is funded by various governmental, academic, and philanthropic sources. What happens to these resources after completion is very poorly understood. No systematic survey of digital resource usage in the humanities has ever been undertaken--and the factors for use and non-use of digital resources are unknown. The LAIRAH (Log Analysis of Internet Resources in the Arts and Humanities) Project is a 15-month long study into the factors which determine long-term use

and neglect of digital resources in the Arts and Humanities. Using quantitative Deep Log Analysis techniques to understand real-time user behaviour and qualitative user workshops to gain an understanding of user approaches to digital resources in the arts and humanities, the study identifies factors that may predispose a digital resource to become used or neglected in the long-term. This article provides an overview of the techniques used in the LAIRAH project, and presents some preliminary results that may be of use to both the creators of digital resources in the humanities, and the funders of these projects, to ensure that significant intellectual effort and time, and financial resources, are not wasted in the creation of projects that are then neglected by the user community.

**Publication** Lit Linguist Computing  
**Volume** 23  
**Issue** 1  
**Pages** 85-102  
**Date** April 1, 2008  
**DOI** 10.1093/llc/fqm045  
**Short Title** If You Build It Will They Come?  
**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/23/1/85>  
**Accessed** Sun Apr 5 04:29:44 2009  
**Repository** HighWire  
**Date Added** Sun Apr 5 04:29:44 2009  
**Modified** Sun Apr 5 04:29:44 2009

**Tags:**

project\_team, single\_institution, user\_evaluation

**Attachments**

HighWire Full Text PDF

HighWire Snapshot

---

## The master builders: LAIRAH research on good practice in the construction of digital humanities projects

**Type** Journal Article  
**Author** Claire Warwick  
**Author** Isabel Galina  
**Author** Melissa Terras

**Author** Paul Huntington

**Author** Nikoleta Pappa

**Abstract** Although many digital humanities resources are being developed for online use, there is little understanding of why some become popular, whilst others are neglected. Through log analysis techniques, the LAIRAH project identified twenty-one popular and well-used digital humanities projects, and in order to ascertain the factors they had in common, which predisposed them to be well used, conducted in-depth interviews with the creators of these resources. This article presents the findings of the study, highlighting areas that developers should be aware of, and providing a set of recommendations for both funders and creators, which should ensure that a digital humanities resource will have the best possible chance of being used in the long term.

**Publication** Lit Linguist Computing

**Volume** 23

**Issue** 3

**Pages** 383-396

**Date** September 1, 2008

**DOI** 10.1093/llc/fqn017

**Short Title** The master builders

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/23/3/383>

**Accessed** Sun Apr 5 04:16:39 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:16:39 2009

**Modified** Sun Apr 5 04:16:39 2009

### Tags:

project\_team, single\_institution, user\_evaluation

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## The Identification of Exemplar Change in the Wife of Bath's Prologue Using the Maximum Chi-Squared Method

**Type** Journal Article

**Author** Heather F. Windram

**Author** Christopher J. Howe

**Author** Matthew Spencer

**Abstract** Chaucer's Wife of Bath's Prologue survives in both hand-written and early print witnesses dating from the 15th century. The introduction of material from more than one exemplar into a new copy results in contamination of a textual tradition. This contamination causes problems in standard phylogenetic analysis. We use an application of the maximum  $\chi^2$  method (developed for the detection of recombination in DNA sequences) to identify locations where scribes may have changed their exemplar whilst copying the tale. Our results are largely in agreement with other published sources, indicating that this method may prove useful in the analysis of a contaminated tradition.

**Publication** Lit Linguist Computing

**Volume** 20

**Issue** 2

**Pages** 189-204

**Date** June 1, 2005

**DOI** 10.1093/lc/fqi001

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/20/2/189>

**Accessed** Sun Apr 5 05:32:23 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:32:23 2009

**Modified** Sun Apr 5 05:32:23 2009

### Tags:

biochemistry, literary\_dna, multi-country, multi-discipline, multi-institutional, statistics, textual\_studies

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Dante's Monarchia as a test case for the use of phylogenetic methods in stemmatic analysis

**Type** Journal Article

**Author** Heather F. Windram

**Author** Prue Shaw

**Author** Peter Robinson

**Author** Christopher J. Howe

**Abstract** Dante's *Monarchia*, a fourteenth century treatise on political theory which survives in 20 manuscripts and the *editio princeps*, has been studied extensively by scholars using traditional analytical methods to establish textual transmission. It was selected as a suitable tradition for a blind study to test the application of computer-based phylogenetic methods to the stemmatic analysis of manuscript relationships. Our results show that these methods--maximum parsimony, NeighborNet and the Supernet algorithm--are capable of producing stemmata in very close agreement with those produced by traditional stemmatic analysis, including the identification of texts that change exemplar in the course of copying. The phylogenetic methods can correctly indicate the affiliations both before and after the point of exemplar change. The maximum chi-squared method (developed to detect recombination in DNA sequences) is able to indicate the region of exemplar change, allowing the precise location to be ascertained by textual analysis.

**Publication** Lit Linguist Computing

**Volume** 23

**Issue** 4

**Pages** 443-463

**Date** December 1, 2008

**DOI** 10.1093/lc/fqn023

**URL** <http://llc.oxfordjournals.org/cgi/content/abstract/23/4/443>

**Accessed** Sun Apr 5 04:09:02 2009

**Repository** HighWire

**Date Added** Sun Apr 5 04:09:02 2009

**Modified** Sun Apr 5 04:09:02 2009

### Tags:

biochemistry, multi-institutional, stylistics, textual\_studies

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

## Unification of XML Documents with Concurrent Markup

**Type** Journal Article

**Author** Andreas Witt

**Author** Daniela Goecke

**Author** Felix Sasaki

**Author** Harald Lungen

**Abstract** An approach to the unification of XML (Extensible Markup Language) documents with identical textual content and concurrent markup in the framework of XML-based multi-layer annotation is introduced. A Prolog program allows the possible relationships between element instances on two annotation layers that share PCDATA to be explored and also the computing of a target node hierarchy for a well-formed, merged XML document. Special attention is paid to identity conflicts between element instances, for which a default solution that takes into account metarelations that hold between element types on the different annotation layers is provided. In addition, rules can be specified by a user to prescribe how identity conflicts should be solved for certain element types.

**Publication** Lit Linguist Computing

**Volume** 20

**Issue** 1

**Pages** 103-116

**Date** March 1, 2005

**DOI** 10.1093/lc/fqh046

**URL** <http://lc.oxfordjournals.org/cgi/content/abstract/20/1/103>

**Accessed** Sun Apr 5 05:36:16 2009

**Repository** HighWire

**Date Added** Sun Apr 5 05:36:16 2009

**Modified** Sun Apr 5 05:36:16 2009

### Tags:

markup, multi-institutional

### Attachments

HighWire Full Text PDF

HighWire Snapshot

---

**Type** Note

**Date Added** Sun Apr 5 04:56:20 2009

**Modified** Sun Apr 5 04:58:11 2009

We may see more collaboratively authored articles b/c of this journal focuses on linguistics as well as literary scholarship. I assume that collaboratively written works are more common on linguistics.

What I'd like to know:

- \* why did the authors collaborate?
- \* what was the nature of the collaboration? who did what?
- \* how was the collaboration managed--schedules, tools, etc
- \* what was the result of the collaboration? was it a positive or negative experience?

---

### 2007.1 Lit Linguist Computing -- Table of Contents (April 2007, 22 [1])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol22/issue1/>

**Accessed** Sun Apr 5 04:54:41 2009

**Date Added** Sun Apr 5 04:54:41 2009

**Modified** Fri Apr 17 05:29:34 2009

#### Notes:

5 articles, 2 collab

#### Attachments

Lit Linguist Computing -- Table of Contents (April 2007, 22 [1])

---

### 2004.3 Lit Linguist Computing -- Table of Contents (2004, 19 [3])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol19/issue3/>

**Accessed** Sun Apr 5 05:42:53 2009

**Date Added** Sun Apr 5 05:42:53 2009

**Modified** Fri Apr 17 05:27:49 2009

#### Notes:

12 articles, 3 collab

## Attachments

Lit Linguist Computing -- Table of Contents (2004, 19 [3])

---

### 2008.2 Lit Linguist Computing -- Table of Contents (June 2008, 23 [2])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol23/issue2/>

**Accessed** Sun Apr 5 04:21:04 2009

**Date Added** Sun Apr 5 04:21:04 2009

**Modified** Fri Apr 17 05:27:05 2009

#### Notes:

1 of 5 original articles

## Attachments

Lit Linguist Computing -- Table of Contents (June 2008, 23 [2])

---

### 2008.4: Lit Linguist Computing -- Table of Contents (December 2008, 23 [4])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol23/issue4/index.dtl>

**Accessed** Sun Apr 5 04:06:18 2009

**Date Added** Sun Apr 5 04:06:17 2009

**Modified** Fri Apr 17 05:26:38 2009

#### Tags:

projectteam

#### Notes:

5 original articles, 4 jointly authored

## Attachments

Lit Linguist Computing -- Table of Contents (December 2008, 23 [4])

---

2006.3 Lit Linguist Computing -- Table of Contents (September 2006, 21 [3])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol21/issue3/>

**Accessed** Sun Apr 5 05:05:05 2009

**Date Added** Sun Apr 5 05:05:05 2009

**Modified** Fri Apr 17 05:29:18 2009

### Notes:

9 articles, 5 collab

## Attachments

Lit Linguist Computing -- Table of Contents (September 2006, 21 [3])

---

2006.2 Lit Linguist Computing -- Table of Contents (June 2006, 21 [2])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol21/issue2/>

**Accessed** Sun Apr 5 05:09:33 2009

**Date Added** Sun Apr 5 05:09:33 2009

**Modified** Fri Apr 17 05:29:09 2009

### Notes:

9 articles, 8 collab

## Attachments

Lit Linguist Computing -- Table of Contents (June 2006, 21 [2])

---

## 2008.1, Lit Linguist Computing -- Table of Contents (April 2008, 23 [1])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol23/issue1/>

**Accessed** Sun Apr 5 04:26:31 2009

**Date Added** Sun Apr 5 04:26:31 2009

**Modified** Fri Apr 17 05:26:58 2009

### Notes:

9 articles, 3 collab

### Attachments

Lit Linguist Computing -- Table of Contents (April 2008, 23 [1])

---

## 2007.3 Lit Linguist Computing -- Table of Contents (September 2007, 22 [3])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol22/issue3/>

**Accessed** Sun Apr 5 04:37:55 2009

**Date Added** Sun Apr 5 04:37:55 2009

**Modified** Fri Apr 17 05:29:51 2009

### Notes:

6 articles, 4 collab

### Attachments

Lit Linguist Computing -- Table of Contents (September 2007, 22 [3])

---

## 2006.4 Lit Linguist Computing -- Table of Contents (November 2006, 21 [4])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol21/issue4/>

**Accessed** Sun Apr 5 04:58:42 2009

**Date Added** Sun Apr 5 04:58:42 2009

**Modified** Fri Apr 17 05:28:52 2009

**Notes:**

11 articles, 5 collab

**Attachments**

Lit Linguist Computing -- Table of Contents (November 2006, 21 [4])

---

2007.4 Lit Linguist Computing -- Table of Contents (November 2007, 22 [4])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol22/issue4/>

**Accessed** Sun Apr 5 04:31:50 2009

**Date Added** Sun Apr 5 04:31:50 2009

**Modified** Fri Apr 17 05:29:25 2009

**Notes:**

7 articles, 6 collab

**Attachments**

Lit Linguist Computing -- Table of Contents (November 2007, 22 [4])

---

2005.4 Lit Linguist Computing -- Table of Contents (November 2005, 20 [4])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol20/issue4/>

**Accessed** Sun Apr 5 05:21:26 2009

**Date Added** Sun Apr 5 05:21:26 2009

**Modified** Fri Apr 17 05:28:34 2009

**Notes:**

6 articles, 1 collab

**Attachments**

Lit Linguist Computing -- Table of Contents (November 2005, 20 [4])

---

2008.3: Lit Linguist Computing -- Table of Contents (September 2008, 23 [3])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol23/issue3/>

**Accessed** Sun Apr 5 04:13:26 2009

**Date Added** Sun Apr 5 04:13:26 2009

**Modified** Fri Apr 17 05:27:18 2009

**Notes:**

9 original articles: 4 jointly authored, others (Clement, Bei Yu) on collaborative projects

**Attachments**

Lit Linguist Computing -- Table of Contents (September 2008, 23 [3])

---

2007.2 Lit Linguist Computing -- Table of Contents (June 2007, 22 [2])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol22/issue2/>

**Accessed** Sun Apr 5 04:42:46 2009

**Date Added** Sun Apr 5 04:42:46 2009

**Modified** Fri Apr 17 05:29:40 2009

**Notes:**

7 articles, 6 collab

### Attachments

Lit Linguist Computing -- Table of Contents (June 2007, 22 [2])

---

### 2004.4 Lit Linguist Computing -- Table of Contents (2004, 19 [4])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol19/issue4/>

**Accessed** Sun Apr 5 05:38:48 2009

**Date Added** Sun Apr 5 05:38:48 2009

**Modified** Fri Apr 17 05:27:57 2009

### Notes:

5 articles, 3 collab

### Attachments

Lit Linguist Computing -- Table of Contents (2004, 19 [4])

---

### 2006.1 Lit Linguist Computing -- Table of Contents (April 2006, 21 [1])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol21/issue1/>

**Accessed** Sun Apr 5 05:15:28 2009

**Date Added** Sun Apr 5 05:15:28 2009

**Modified** Fri Apr 17 05:29:03 2009

### Notes:

1 of 5 orig articles collab

### Attachments

Lit Linguist Computing -- Table of Contents (April 2006, 21 [1])

---

### 2005.3 Lit Linguist Computing -- Table of Contents (September 2005, 20 [3])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol20/issue3/>

**Accessed** Sun Apr 5 05:26:52 2009

**Date Added** Sun Apr 5 05:26:52 2009

**Modified** Fri Apr 17 05:28:27 2009

#### Notes:

7 articles, 2 collab

#### Attachments

Lit Linguist Computing -- Table of Contents (September 2005, 20 [3])

---

### 2005.2 Lit Linguist Computing -- Table of Contents (June 2005, 20 [2])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol20/issue2/>

**Accessed** Sun Apr 5 05:31:12 2009

**Date Added** Sun Apr 5 05:31:12 2009

**Modified** Fri Apr 17 05:28:18 2009

#### Notes:

5 articles, 4 collab

#### Attachments

Lit Linguist Computing -- Table of Contents (June 2005, 20 [2])

---

### 2004.1 Lit Linguist Computing -- Table of Contents (2004, 19 [1])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol19/issue1/>

**Accessed** Sun Apr 5 05:51:58 2009

**Date Added** Sun Apr 5 05:51:58 2009

**Modified** Fri Apr 17 05:27:29 2009

### Notes:

3 of 9

### Attachments

Lit Linguist Computing -- Table of Contents (2004, 19 [1])

---

## 2004.2 Lit Linguist Computing -- Table of Contents (2004, 19 [2])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol19/issue2/>

**Accessed** Sun Apr 5 05:49:20 2009

**Date Added** Sun Apr 5 05:49:20 2009

**Modified** Fri Apr 17 05:27:35 2009

### Notes:

2 of 5

### Attachments

Lit Linguist Computing -- Table of Contents (2004, 19 [2])

---

## 2005.1 Lit Linguist Computing -- Table of Contents (March 2005, 20 [1])

**Type** Web Page

**URL** <http://llc.oxfordjournals.org/content/vol20/issue1/>

**Accessed** Sun Apr 5 05:33:55 2009

**Date Added** Sun Apr 5 05:33:55 2009

**Modified** Fri Apr 17 05:28:10 2009

## **Notes:**

9 articles, 3 collab

## **Attachments**

Lit Linguist Computing -- Table of Contents (March 2005, 20 [1])