SDV Vol.31.1-2-2007

SDV. Sprache und Datenverarbeitung
International Journal for Language Data Processing
Heft 1-2/2007

Inhalt

Haugen, Odd Einar
Medieval Unicode Font Initiative (MUFI). Coordinating Medieval characters in the Latin alphabet

Declerck, Thierry & Ide, Nancy & Trippel, Thorsten
Interoperable Language Resources

Larson, Martha & Jijkoun, Valentin & Löffler, Jobst & Tjong Kim Sang, Erik
Practical applications of stand-off annotation

Abstracts

Vorwort

Autor:
Gippert, Jost & Schmitz, Hans-Christian

Aufsatztitel:
Vorwort

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
5

Abstract:
Die diachrone Sprachwissenschaft ist auf nach Sprachräumen und Zeitstufen klassifizierte Korpusdaten angewiesen, welche idealerweise linguistisch annotiert sind und nach verschiedenen Kriterien effektiv durchsucht werden können. Dazu müssen Texte in großem Umfang digitalisiert, aufbereitet und zur Nutzung bereitgestellt werden. Für das Deutsche ist dies noch nicht in ausreichender Form geschehen. Es gibt allerdings zahlreiche Bemühungen und Aktivitäten um die Etablierung diachroner Textkorpora. Diese Bemühungen und Aktivitäten gilt es zu koordinieren, um eine gemeinsame Grundlage für die Integration der bereits erarbeiteten und noch zu erarbeitenden Daten zu schaffen.
Die Koordination derartiger Ansätze zu fördern, war das Ziel eines internationalen Workshops über diachrone Korpora, historische Syntax und Texttechnologie, der im April 2007 in Frankfurt am Main stattfand und vom Graduiertenkolleg Satzarten: Variation und Interpretation der Universität Frankfurt zusammen mit den Arbeitskreisen Historisch-Vergleichende Sprachwissenschaft und Korpuslinguistik der Gesellschaft für Linguistische Datenverarbeitung veranstaltet wurde. In sieben Präsentationen wurden bestehende bzw. geplante Korpusprojekte mit ihren spezifischen Ansätzen und zu erwartenden oder bereits

vorliegenden Ergebnissen vorgestellt; vier weitere Präsentationen waren allgemeinen Fragen der Korpusarchitektur, der Annotation und des Retrievals gewidmet.
Um die Diskussion einem breiteren Publikum zugänglich zu machen, haben wir neun der auf dem Workshop gehaltenen Vorträge im vorliegenden Band zusammengestellt. Außerdem haben wir vier weitere Beiträge aufgenommen, die ein zusätzliches Korpus und seine Nutzungsoberfläche beschreiben (Fisseni et al.), zentrale technische Fragestellungen eruieren (Haugen, Larson et al.) und die Schnittstelle zu anderen (Text-)Wissenschaften – hier: der Geschichtswissenschaft – beleuchten (Jussen et al.). Wir danken den Beiträgern für ihre stetige und zielführende Kooperation und den Herausgebern der SDV für die Bereitschaft, den vorliegenden Band als Einzelheft der Zeitschrift zu veröffentlichen.

Überlegungen zum Design und zur Architektur eines diachronen Korpus

Autor:
Lüdeling, Anke

Aufsatztitel:
Überlegungen zum Design und zur Architektur eines diachronen Korpus

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
7-14

Abstract:
This article sketches the plans, developed in the project initiative DeutschDiachronDigital, for a diachronic corpus of German. Many historical texts of German are digitized. They cannot, however, easily be used for diachronic research because they differ in source, diplomaticity, level and type of annotation and format. The project initiative developed a flexible multi-layer corpus architecture which makes it possible to integrate existing digitized texts and add texts as well as annotation layers when needed. It also developed a set of common digitization and annotation standards.

Der Artikel skizziert Pläne für ein integriertes diachrones Korpus des Deutschen, die in der Projektinitiative DeutschDiachronDigital entwickelt wurden. Die bisher digitalisierten Texte des historischen Deutsch können nicht durchgehend einheitlich untersucht werden, da sie sich in Quelle, Diplomatizität, Annotationsebenen und Format unterscheiden. Die Projektinitiative hat daher eine flexible Mehrebenenarchitektur entworfen (aufbauend auf dem ODAG-Modell), in der bereits digitalisierte Texte ebenso repräsentiert werden können wie neu aufgenommene. Gleichzeitig hat die Projektinitiative inhaltliche Standards für die Digitalisierung und die Annotation entwickelt, die eine einheitliche Bearbeitung ermöglichen.

Zur Dimensionierung historischer Textkorpora

Autor:
Klein, Thomas

Aufsatztitel:
Zur Dimensionierung historischer Textkorpora

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
15-22

Abstract:
Annotated historical linguistic corpora need to be restricted in order to create a corpus large enough for the intended analysis on the one hand but not too large to hinder the realisation of the project on the other hand. The following paper will show that data from a smaller comparable corpus is required in order to optimize the scope of the new corpus.

Der Umfang annotierter historischer Sprachkorpora muss so beschränkt werden, dass das Korpus einerseits für die intendierten Untersuchungen groß genug ist und andererseits nur so groß, dass Erstellung und Annotation des Korpus realisierbar sind. Der folgende Aufsatz will zeigen, dass man für die Festlegung des optimalen Korpusumfangs schon über die Daten aus einem (kleineren) Vergleichskorpus verfügen muss.

Technological and linguistic issues in the construction of parsed corpora

Autor:
Santorini, Beatrice

Aufsatztitel:
Technological and linguistic issues in the construction of parsed corpora

Jahrgang:
31
Heft:
01/02 (2007)

Abstract:
This paper discusses issues concerning parsed corpora, focusing on three main questions: Why do we need parsed corpora? How do we annotate them? How do we search them? The experience on which the paper is based comes mainly from English historical corpora constructed in the Linguistics Departments at the University of Pennsylvania and the University of York, but also from corpora in other languages, including French, Portuguese, and Sumerian.

Diachronic corpora and historical syntax

Autor:
Kroch, Anthony

Aufsatztitel:
Diachronic corpora and historical syntax

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
31-38

Abstract:
Work at the University of Pennsylvania in the United States and at the University of York in the United Kingdom has produced a set of parsed corpora, comprising nearly seven million words of running text. These corpora are intended to serve as resources for the investigation of the quantitative historical syntax of English and, in combination with similar corpora of other languages, for the development of a science of quantitative comparative syntax. In this communication, we briefly present two case studies in English historical syntax to illustrate the utility of parsed corpora for research. The first is a study of dialect variation in the form and evolution of the verb-second constraint in Middle English. The second is an account of the decline in the frequency of topicalization in the course of Middle and Early Modern English.

Zur informationsstrukturellen Annotation sprachhistorischer Texte

Autor:
Donhauser, Karin

Aufsatztitel:
Zur informationsstrukturellen Annotation sprachhistorischer Texte

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
39-45

Abstract:
The article demonstrates a procedure for annotating information structure in texts from historical corpora which was developed by Collaborate Research Group SFB 632 Information Structure (SFB) at Humboldt University, Berlin. Here, the annotation is designed in a way allowing to retrace each decision step made throughout the annotation process. Therefore, annotation is conducted incrementally distinguishing the following three layers of Information Structure: (1) cognitive status, (2) predication structure and (3) informational relevance. This cumulative approach has delivered optimal results in practice, among all in working out an information-structural cartography of the left and right sentence periphery in Old High German. This is exemplified here with respect to focus positions in subordinate clauses.

Der Artikel stellt ein Verfahren zur informationsstrukturellen Annotation von sprachhistorischen Texten vor, das im Rahmen des SFB 63 2 Informationsstruktur an der Humboldt-Universität zu Berlin erarbeitet wurde. Dabei werden Annotationen so angelegt, dass Entscheidungswege immer nachvollziehbar bleiben. Die Annotation erfolgt deshalb sequentiell unter strikter Trennung folgender informationsstruktureller Ebenen: (1) kognitiver Status, (2) prädikative Strukturierung und (3) informationsstrukturelle Gewichtung. Dieses kumulative Vorgehen hat sich im Projekt hervorragend bewährt, u.a. bei der informationsstrukturellen Kartographie des linken und rechten Satzrandes im Althochdeutschen, deren Ergebnisse hier an einem Beispiel (Fokuspositionen im Nebensatz) illustriert werden.

Korpusgestützte Analyse von Präpositionen und Präpositionalphrasen im Mittelhochdeutschen

Autor:
Waldenberger, Sandra

Aufsatztitel:
Korpusgestützte Analyse von Präpositionen und Präpositionalphrasen im Mittelhochdeutschen

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
47-57

Abstract:
The objective of this paper is to demonstrate how the ‘Bochumer Mittelhochdeutsches Korpus‘ is being utilised for the analysis of prepositions and prepositional phrases within the scope of the DFG research project ‘Mittelhochdeutsche Grammatik’. The corpus is comprised of annotated Middle High German texts (part of speech tagging and morphological annotation) and its specific structure allows the analysis of the obtained data with respect to dialect, text type or diachronic developments within the period researched. The samples of the analyses of Middle High German prepositions and prepositional phrases are intended to exemplify working with the annotated language data at hand and a structured corpus.

In diesem Beitrag soll illustriert werden, wie das Bochumer Mittelhochdeutschkorpus zur Analyse von Präpositionen und Präpositionalphrasen im Mittelhochdeutschen im Rahmen des DFG-Projekts Mittelhochdeutsche Grammatik genutzt wird. Das Bochumer Mittelhochdeutschkorpus ist (part of speech und morphologisch) annotiert und ermöglicht durch seine Strukturmerkmale die Untersuchung der Daten auf diachrone, diatopische oder textartabhängige Ausprägungen. Die beispielhaft vorgestellten Teilaspekte der Analyse von mittelhochdeutschen Präpositionen und Präpositionalphrasen veranschaulichen die konkrete Arbeit mit den annotierten Sprachdaten und mit einem strukturierten Korpus.

Stand und Perspektiven des Mittelhochdeutschen Textarchivs

Autor:
Gärtner, Kurt & Plate, Ralf

Aufsatztitel:
Stand und Perspektiven des Mittelhochdeutschen Textarchivs

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
59-65

Abstract:
Das Mittelhochdeutsche Textarchiv (MTA), über das hier berichtet wird, ist eingebettet in ein Artikelredaktionssystem, das für die Ausarbeitung des neuen Mittelhochdeutschen Wörterbuchs (MWB) entwickelt wurde. Das Textarchiv umfasst derzeit rund 235 E-Texte, die philologisch gesichert und verlässlich zitierbar sind. Es handelt sich um digitale Volltexte, die auf edierten Texten von in der Regel überlieferungsnahen kritischen Editionen beruhen und als TEI-konforme XML-Dokumente aufbereitet sind. Den Kern des Archivs bilden ca. 16 0 lemmatisierte Texte, aus denen das elektronische Belegarchiv des MWB generiert wurde, das eine Komponete des Artikelredaktionssystems bildet und dessen Funktionalität im einzelnen beschrieben wird. Die E-Texte des Belegarchivs sind mit den im MWB zitierten Belegen verknüpft. Zum Archiv gehören rund 75 weitere, z. T. sehr umfangreiche E-Texte aus dem Corpus des ‚Findebuchs', dessen Quellen in einem DFG/NSF-geförderten Kooperationsprojekt digitalisiert wurden. Zu den Perspektiven der künftigen Arbeiten am MTA gehören der Ausbau des Kernbereichs, d. h. die Lemmatisierung weiterer Texte, und die Erweiterung des Archivs durch die Digitalisierung möglichst vieler neuer Texte.

This contribution focuses on the Middle High German Text Archive (MTA) which is embedded in a lexicographical toolkit for the editing and publication of the new Middle High German Dictionary (Mittelhochdeutsches Wörterbuch, MWB). At present, the MTA consists of app. 235 philologically reliable e-texts. These e-texts are digital full texts, they are based on critical editions, preferably on those closely representing the transmission, and they are TEI-compliant XML documents. The core of the MTA consists of app. 16 0 lemmatized e-texts from which the MWB's electronic archive of evidences had been generated. The functionality of this archive of evidences, one of the lexicographer's key resources, is described in detail. These e-texts are interlinked with the entries of the electronic version of the MWB. The MTA contains in addition app. 75 non-lemmatized e-texts of the ‚Findebuch'-corpus, the sources of which have been digitized in a DFG/NSF-funded collaborative project. We expect that we will be able in near future to increase the number of lemmatised texts and add as many new e-texts to our archive as we can.

FndhC/HTML und FnhdC/S

Autor:
Fisseni, Bernhard & Schmitz, Hans-Christian & Schröder, Bernhard

Aufsatztitel:
FndhC/HTML und FnhdC/S

Jahrgang:
31
Heft:
01/02 (20007)
Seiten:
67-69

Abstract:
Wir beschreiben kurz den explorativen (Fnhd/HTML) und analytischen Zugang (Fnhd/S) zum Bonner Frühneuhochdeutsch-Korpus, auf das über http://www.ikp.uni-bonn.de/fnhd/ zugegriffen werden kann.

We shortly describe an explorative (Fnhd/HTML) and an analytic interface (Fnhd/S) to the Bonn Corpus of Early New High German that can be accessed via http://www.ikp.uni-bonn.de/fnhd/

GerManC: A historical corpus of German 1650-1800

Autor:
Durrell, Martin & Ensslin, Astrid & Bennett, Paul

Aufsatztitel:
GerManC: A historical corpus of German 1650-1800

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
71-80

Abstract:
This paper gives an account of the GerManC corpus currently being compiled at theUniversity of Manchester (UK). The first pilot project for this representative historical corpus of German for the years 16 50-1800 is now complete and comprises some 100,000 words from newspapers. The corpus has been fully annotated according to TEI guidelines, and programs have been developed to tag and lemmatize it. As an illustration of the potential of the corpus as a resource for research into the history and standardization of German, short accounts are given of the developments of the variants for beide 'both', the gender forms for zwei 'two', and the weak adjective declension.

Dieser Artikel bietet einen kurzen Überblick des GerManC Korpus, das gegenwärtig an der Universität Manchester (Großbritannien) zusammengestellt wird. Der erste Teil dieses repräsentativen historischen Korpus der deutschen Sprache in den Jahren 1650-1800 is jetzt fertig und umfasst etwa 100 000 Wörter aus Zeitungstexten. Das Korpus ist nach den Richtlinien der TEI vollständig annotiert worden, und Programme wurden entwickelt, mit denen das Korpus getaggt und lemmatisiert werden kann. Das Potential des Korpus für die Erforschung der deutschen Sprache und ihrer Standardisierung in dieser Epoche wird durch drei kurze Beispiele aufgezeigt, und zwar zur Geschichte der Varianten für beide, zu den Genusformen des Zahlworts zwei, und zur Entwicklung der schwachen Adjektivflexion.

A Corpus Management System for Historical Semantics

Autor:
Jussen, Bernhard & Mehler, Alexander & Ernst, Alexandra

Aufsatztitel:
A Corpus Management System for Historical Semantics

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
81-89

Abstract:
This paper presents a corpus management system for historical semantics. Its background is a notion of semantics which relies on corpus analyses of diachronic corpora. These corpora are analyzed to explore semantic change as an access point to the understanding of social change. The system to be presented supports this kind of corpus-based historical semantics.

Der Beitrag beschreibt ein Korpusmanagementsystem für die historische Semantik. Die Grundlage hierfür bildet ein Bedeutungsbegriff, der – methodologisch gesprochen – auf der Analyse diachroner Korpora beruht. Das Ziel der Analyse dieser Korpora besteht darin, Bedeutungswandel als eine Bezugsgröße für den Wandel sozialer Systeme zu untersuchen. Das vorgestellte Korpusmanagementsystem unterstützt diese Art der korpusbasierten historischen Semantik.

Medieval Unicode Font Initiative (MUFI). Coordinating Medieval characters in the Latin alphabet

Autor:
Haugen, Odd Einar

Aufsatztitel:
Medieval Unicode Font Initiative (MUFI). Coordinating Medieval characters in the Latin alphabet

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
91-99

Abstract:
The Medieval Unicode Font Initiative (MUFI) is an informal network which aims to coordinate the usage of the Private Use Area (PUA) of the Unicode Standard among medievalists. The PUA can be used for characters not yet part of the Unicode Standard, but unless there is a coordination of code points in this area, any file exchange will entail an extensive search and replace process of PUA characters. MUFI now coordinates code point allocation in several font projects, e.g. TITUS (Germany), Junicode (US), Alphabetum (Spain) and LeedsUni (UK). In addition to coordinating code point allocation in the PUA through the MUFI character recommendation, MUFI participants work with other medievalists in writing proposals to Unicode for the inclusion of new characters. So far, two proposals of more than 100 characters have been met with approval by Unicode, and the characters in these proposals are likely to become part of an upcoming version of the standard. Moreover, participants in the MUFI network are active in developing fonts including medieval characters, many of which can be downloaded free of charge from the MUFI website, http://www.mufi.info. These fonts are compatible with all major computer

platforms (Linux, Mac, Windows) and may be used for the display of documents encoded in XML and other open formats, whether in print or on the web.

Interoperable Language Resources

Autor:
Declerck, Thierry & Ide, Nancy & Trippel, Thorsten

Aufsatztitel:
Interoperable Language Resources

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
101-113

Abstract:
In this contribution we present some work of the R&D European project “LIRICS” and of the ISO/TC 37/SC 4 committee related to the topic of interoperability and re-use of language resources. We introduce some basic mechanisms of the standardization work in ISO and describe in more details the general approach on how to cope with the annotation of language data within ISO.

Unser Beitrag beschreibt aktuelle Arbeiten des europäischen Projekts „LIRICS“ und des ISO-Ausschusses ISO/TC 37 / SC 4 zum Thema Interoperabilität und Wiederverwendbarkeit von Sprachresourcen. Neben einer allgemeinen Einführung zu den Mechanismen der Standardisierung bei ISO präsentieren wir einige ISO-Vorschläge für die Standardisierung der Annotierung von linguistischen Daten.

Practical applications of stand-off annotation

Autor:
Larson, Martha & Jijkoun, Valentin & Löffler, Jobst & Tjong Kim Sang, Erik

Aufsatztitel:
Practical applications of stand-off annotation

Jahrgang:
31
Heft:
01/02 (2007)
Seiten:
115-129

Abstract:
An information system that makes use of stand-off annotation stores metadata separately from the data they describe. System architectures separate metadata from data in order to cope with heterogeneous annotations or with multimedia formats. This paper discusses some of the practical aspects of implementing an information system with a stand-off architecture. Two systems that use stand-off annotations are described. The first is a prototype radio archive that provides users with content-based access to archived radio broadcasts. This system uses stand-off annotation to store structural metadata describing the broadcasts, which is used for interactive presentation, as well as speech recognition transcripts, which are used for search. The second system is a question answering system that searches a large text corpus in order to identify spans of text that provide answers to user questions. This system uses stand-off annotation to store metadata generated by a series of different linguistic analysis tools. The final section of the paper treats practical aspects of implementing a retrieval system for a diachronic language corpus. Similarities and differences with the prototype radio archive and the question answering system are discussed.

Ein Informationssystem, das stand-off-Annotation verwendet, speichert Metadaten getrennt von den eigentlichen Daten, die durch die Metadaten beschrieben werden. ystemarchitekturen trennen Metadaten von Daten, um die Handhabung von heterogenen Annotationen oder multimedialen Datenformaten zu ermöglichen. Dieser Beitrag diskutiert einige praktische Aspekte der Implementierung von Informationssystemen mit einer stand-off-Architektur. Zwei Anwendungssysteme, die stand-off-Annotationen einsetzen, werden beschrieben. Das erste ist der Prototype eines Radioarchivs, das dem Benutzer den inhaltsbasierten Zugang zu archivierten Radiosendungen ermöglicht. Das System benutzt stand-off-Annotation einerseits zur Speicherung struktureller Metadaten, die zur interaktiven Darstellung der Radiobeiträge am Benutzerarbeitsplatz eingesetzt werden. Zum anderen wird stand-off-Annotation hier verwendet, um Spracherkennungstransskripte zu verwalten, die vom Benutzer für die inhaltsbasierte Suche im Radioarchiv genutzt werden. Das zweite System ist ein Frage-Antwort-System, das ein großes Textkorpus durchsucht. Das Ziel ist die Identifizierung von Textbereichen, die Antworten auf die vom Benutzer gestellten Fragen geben. Dieses System setzt stand-off-Annotation für die Speicherung von Metadaten ein, die von einer Reihe von verschiedenen linguistischen Analysewerkzeugen erzeugt werden. Der abschließende Abschnitt dieses Beitrags diskutiert praktische Gesichtspunkte der Umsetzung eines Retrievalsystems für einen diachronischen Sprachkorpus. Ähnlichkeiten und Unterschiede der beiden besprochenen Anwendungssysteme, Radioarchiv und Frage-Antwort-System,werden erläutert.