In: Sprache und Datenverarbeitung. International Journal for Language Data Processing 31(1-2), 2007, 81-89.

Bernhard Jussen, Alexander Mehler, and Alexandra Ernst

A Corpus Management System for Historical Semantics

Abstract Englisch

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.

Abstract Deutsch

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.

1 Theoretical Background: Historical Semantics

This paper presents a corpus management system for historical semantics. It focuses on exploring a vast number of corpora of natural language texts ordered by a long-term time line. The central resources are text collections as, e.g., the Patrologia Latina [1, 9] which covers texts published in a period of more than 1,000 years. Obviously, the exploration of such chronologically ordered corpora demands a special research methodology which is well informed about history *and* corpus linguistics. In the present paper, we refer to *corpus-based historical semantics* as such a methodology.

Among historians, historical semantics is a well established research field based on a strong linguistic methodology. However, it does not yet realize an integrated view of the modes, media and social conditions under which meanings emerged and evolved in past societies [10]. According to this reorientation, historical semantics asks for the anthropological, psychological, sociological and technological prerequisites of sign systems by which past cultures organized their knowledge. At the same time, it views meaning constitution as a dynamic process which evolves into fluent equilibria of temporary stability. Accordingly, historical semantics explores semantic change as an access point to social change. It analyses linguistic means of enforcing, combatting, stabilizing,

marginalizing or transforming meanings in past cultures. The empirical basis of this research program is necessarily given by corpora of natural language texts.

The corpus management system to be presented supports this research program which aims at theoretically exploring, formally specifying and empirically investigating the notion of *linguistic change* as an indicator of social processes (e.g. of inclusion or exclusion). The main research question of this program runs as follows:

> **Q1:** *To what extent are long-term social processes manifested by linguistic change and, thus, accessible by a computational approach to usage-based semantics?*

The linguistic background of this question is twofold: On the one hand, it relies on the *weak contextual hypothesis* of Miller & Charles [14] which says that the contextual similarity of words, that is, their tendency to occur in similar contexts, contributes to their meaning. On the other hand, Q1 relies on the distributional hypothesis of Biber [3] who states that situational and functional demands are regularly manifested by preferred linguistic means which are accessible by corpus analysis (see also [4]). Q1 combines these two views with a strong emphasis on the temporal dynamics of social processes and language change. In other words: *it transforms Biber's synchronic view into a diachronic one*. In order to answer Q1, we need to consider long-term processes of language change as manifested by *chronologically ordered corpora* which inform about language use of several generations of agents of the same type of communities. This involves the following research questions:

> **Short-term meaning constitution in an ontogenetic perspective:** *To what extent is a social configuration (e.g. of inclusion or exclusion) of agents within the area under consideration indicated by linguistic means, that is, by these agents' language use?*
>
> **Long-term meaning constitution in a sociogenetic perspective:** *To what extent do social processes and language change co-evolve?*

In order to tackle these and related questions, we need to further develop the apparatus of corpus linguistics. The reason is that although there is much research on linguistic dynamics, computer-based analyses of chronologically ordered corpora documenting language use of consecutive generations are still at their beginning. Present-day approaches mainly focus on probabilistic extensions of well established methods (e.g. probabilistic grammars or statistical collocation analyses [5, 12]) in order to grasp this dynamics. These approaches disregard the specifics of exploring long-term processes of meaning constitution subject to social processes. In other words: We do not yet have computational models of how lexical meanings evolve during long-term periods of, e.g., several hundred years (cf. [2]).

Although the development of such models is out of reach of the present paper, it nevertheless defines the broader research goal under which the paper has to be subsumed. That is, we do not provide an extended methodology but concentrate on providing and

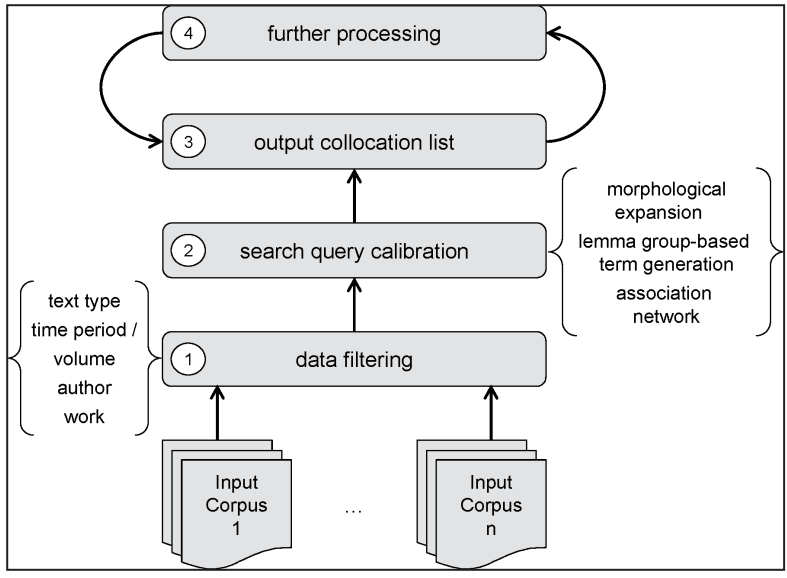

Figure 1: Outline of the query process based on the HSCM

preprocessing relevant corpora as a prerequisite of this developmental goal. Thus, our empirical basis are chronologically ordered corpora of natural language texts as an access point to regularities of linguistic change on three levels:

1. on the level of agents and social communities involved,
2. on the level of text types of these communities' communication practices,
3. and on the level of lexical units whose regularities evolve on the basis of the latter practices.

In order to support historical semantics along these reference points, corpora have to be built whose texts are annotated at least with respect to their *authorship*, *period or date of production* as well as regarding the *text type* instantiated by these texts. Such a corpus management system is described in this paper which is organized as follows: Section 2 describes the corpus system; Section 3 explains a sample analysis based thereon and Section 4 concludes and prospects future work.

2 A Corpus Management System for Historical Semantics

This section presents HSCM, a *Historical Semantics Corpus Management System*[1]. Functionally speaking, the system is settled around a number of core functions for the

1 The HSCM system is accessible via http://www.hscm.org. Note that access to special corpora such as the *Patrologia Latina* corpus is possible only for users who own a corresponding licence. The system can nevertheless be accessed by a guest account. Interested scientists may contact Bernhard Jussen (Bernhard.Jussen@uni-bielefeld.de) to get this account.

preprocessing *of* and the retrieval *from* historical corpora mainly in Latin. The HSCM system combines a list-oriented lemmatization of Latin with a corpus retrieval system based on a fine-grained user management system. We demonstrate the functionality of the HSCM system by example of Jacques Paul Migne's *Patrologia Latina* (MPL) corpus. An interesting feature of the MPL is that it integrates also words in Greek and Hebrew. This feature is a challenge for corpus management systems which, for the time being, the HSCM system masters by list-based preprocessing strategies (see below).

The corpus model of the HSCM system focuses on the Logical Document Structure (LDS) [15] of input texts. That is, texts are modelled as ordered hierarchies of non-overlapping content objects (e.g. sections, paragraphs, sentences) whose leafs denote lemmatized tokens [16] – sentence structure is not yet mapped. This model is, in turn, serialized by means of the HyGraphDB [6] which utilizes the *Graph eXchange Language* (GXL) [8] in order to map graph structures of a wide range of complexity. A central benefit of the HyGraphDB is the expressiveness of its underlying data model – it allows to capture web and text documents as well as multimodal documents. This is complemented by an efficient implementation based on the BerkeleyDB[2]. As the HyGraphDB offers an Application Programming Interface (API) – in C++ and in Java – its usage enables users to explore a wide range of linguistic data structures including the LDS model of the Patrologia Latina.[3] In the present paper, we concentrate on the part of the corpus model of the HSCM system which is already accessible via its web interface.

A central problem of segmenting the logical document structure in *diachronic corpora* [11] relates to time-dependent variations of logically identical reference units. This may include units which in "synchronic" corpora are far from variation problems and, thus, can be captured by predefined lists. In the case of the Patrologia Latina, there exist many spelling variants either due to the edition process or due to the change of author or the period of time. Obviously, this and related problems demand extending the notion of a word form as an instance of a certain type by grasping the fact that a type is possibly instantiated differently in different periods without necessarily mixing these alternatives within the same period. Generally speaking, a time-related notion of instantiation has to be included when dealing with type-token relations in diachronic corpora. The HSCM system manages this by its notion of a *lexeme group* which is open to defining abstract sets of lexemes (see below).

The next section outlines the basic query process based on the HSCM system.

2 Cf. http://www.oracle.com/database/berkeley-db/db/index.html.

3 The HyGraphDB [6, 7] has been developed as part of the *X1 project* of the Sonderforschungsbereich 673 *Alignment in Communication* and of the *A4 project* of the Research Group 437 *Text Technological Information Modeling*. Cf. http://www.sfb673.org/X1 for more information on the HyGraphDB. Please send an email to Alexander.Mehler@uni-bielefeld.de for more information on using the HyGraphDB.

2.1 Input Data Filtering: Restricting the Search Space

Figure 1 presents a diagrammatic view of the query process realized by the HSCM system so far. It includes selecting subsets of documents of specific volumes, authors or works and, thus, restricting the search space of documents to be processed. First of all, MPL documents can be selected according to their status as source or editorial material. This includes selecting documents of patristic and medieval authors and separating additions made by later editions. A second major selection criterion is the *text type* by which, e.g., sermons and aphorisms can be selected while instantiations of other text types are excluded.

The main step in data filtering is to select authors, volumes or works to be considered. By selecting volumes, the search is automatically restricted to the corresponding authors and, vice versa, a selected subset of authors restricts the range of volumes being considered. Nevertheless, authors, volumes and works can be selected independently from each other. At any time, texts fulfilling the selection criteria can be accessed via the web interface of the HSCM system. Further, in order to ease the query process, users can save search schemas in order to reuse and modify them later on.

2.2 Lexical Completion of Search Queries

Search queries are basically Boolean expressions of literals which denote word forms. A search query is answered by identifying all contexts within the selected subset of input documents containing the forms. A basic function of the HSCM system is to support the user in specifying search queries. This relates to providing word form-related morphological expansions of search terms and additionally proposing lemmata as term candidates based on lexeme groups – whether user-defined or built-in.

Within the HSCM system, lemmatization of Latin texts is list-based. Our starting point is a four-level model which distinguishes tokens as instances of word forms which in turn are manifestations of lemmata finally mapped onto lexeme groups. These groups exist as built-in groups of semantically related terms, but may also be defined by the user. The reason to do this is to enlarge the search capabilities of the HSCM system. The lemma list was used to lemmatize and, thus, to pre-process the complete MPL.

Based on the selected subset of documents and the possibly expanded search query, the HSCM system performs a collocation analysis. In accordance with classical approaches in this field [13], the HSCM system allows to restrict collocation contexts to units of the logical document structure. A user may, for example, decide to select varying left- and right-hand side search windows in terms of the number of sentences or in terms of the number of words.

2.3 Further Processing

The downloadable results of search queries are presented in a table format. The HSCM system supports the user in lemmatizing result lists further on in order to generate more

and more abstract output. Generally speaking, there are three methods of lemmatization to be utilized:

- The *system-centred* approach solely refers to built-in lexical groups as provided by the HSCM system.
- The *user-centred* approach solely refers to user-defined lemmatization rules and lemma groups.
- Finally, the *mixed* approach preferably utilizes user-defined lexical groups in addition to the ones provided by the system.

The ongoing lemmatization of result tables serves to define more abstract collocation classes. In historical semantics, this step is indispensable as it allows scientists to rely not only on directly observable word forms but also on interpretative units as an output of reflection processes of historians. Note that from a computational point of view one might expect that such abstract lexical classes are automatically generated by some latent semantic analysis as indicated in Figure 1 where on the level of search query calibration association networks are referred to as additional data resources – this is the reference point where computational linguistics can contribute to strengthen the methodology of historical semantics.

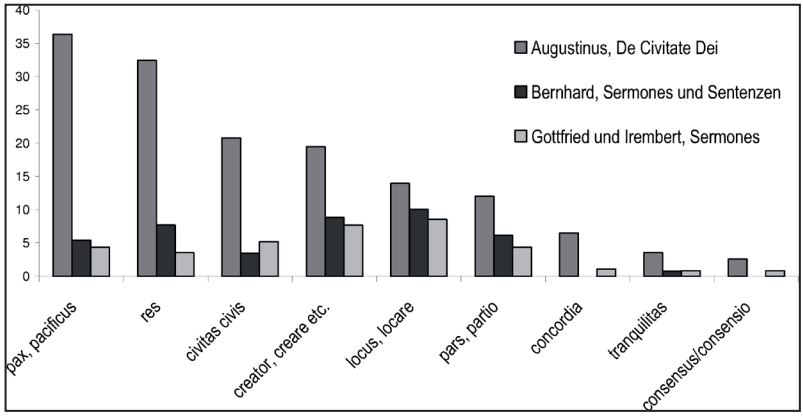

Figure 2: A sample comparative collocational analysis of three corpora regarding collocations of ordo within these texts (axis of ordinates are percentages of collocations of ordo with the respective target term)

2.4 User Administration

The HSCM system includes a user administration tool. It allows mapping access rights of users which are sensitive to all parameters of corpus selection discussed so far. This makes it possible to smoothly define authorizations which restrict corpus access even to specific parts of the corpora managed by the HSCM system. As the MPL and other corpora have restricted access, such a user administration is indispensable.

2.5 Prospect

In the near future, the addition of a full-text search facility which is sensitive to the logical document structure of the input texts is planned in conjunction with providing a range of corpora related to the MPL. This includes, amongst others, the *Acta Sanctorum* which is a collection of hagiographic texts about martyrs and saints of the Catholic Church.

3 A Sample Analysis

This section presents a sample analysis of three works based on the HSCM system. Our starting point is the broadly shared hypothesis that the notion of *ordo* is central to the medieval conception of world and society in the way that there are seminal works which strongly influenced its conceptualization in subsequent epochs. Note that this paper is *not* the place to explain the background of this hypothesis in terms of historical semantics – see [10] for a detailed analysis of this and related issues. Rather, this section concentrates on exemplifying how to deal with such questions by means of the HSCM system.

In order to operationalise the given hypothesis we start from Augustine's *De Civitate Dei* which is seen to be seminal in the sense that it had the kind of influence just claimed. Supposed that the collocational regularities of this work are not retained over a longer period of time and across text types, that is, if the predicted linguistic influence cannot be observed, the hypothesis is falsified in its present form. In order to check this, we select two works of Bernard of Clairvaux (sermons and aphorisms) and of Geoffrey of Admont (sermons). The reason is to provide a basis of comparison which varies the text type. Figure 2, shows that – other than predicted by our reference hypothesis – central *ordo*-related collocations in Augustine's city of god are not retained within these two reference works. Although this result has to be confirmed by further analyses of other corpora, it nevertheless shows how corpus management systems can be utilized to falsify the given type of hypotheses.

The HSCM system is developed as a system which enables this kind of time-related collocation analysis. It includes a statistics tool which supports distribution analysis and time-related tabulations of collocation data. By making the HSCM sensitive to the manifold parameters of the situational context of text production and its temporal dynamics, we aim at integrating methods in historical semantics, corpus and quantitative linguistics. This will be a major effort of the further development of the HSCM in the near future.

4 Conclusion

This paper presented a corpus management system which serves to investigate *linguistic* processes of meaning constitution as an indicator of *social* processes. According to its interdisciplinary stance, the paper integrates the hermeneutic tradition of interpretation of historians with the empirical, quantitative approach to text-technology. The simplicity

of the corpus management system being presented is indispensable in order to guarantee a low barrier to use by historians. Future work aims at extending this corpus management system in order to implement a wide range of corpus analytic methods with a focus on language change.

References

[1] Patrologia Latina database. *Informationsmittel für Bibliotheken (IFB)*, 3(1), 1995.

[2] M. V. Arapov and M. M. Cherc. *Mathematische Methoden in der historischen Linguistik*. Brockmeyer, Bochum, 1983.

[3] D. Biber. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge, 1995.

[4] D. Biber, S. Conrad, and R. Reppen. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press, Cambridge, 1998.

[5] R. Bod. *Beyond Grammar. An Experience-Based Theory of Language*. CSLI Publications, Stanford, 1998.

[6] R. Gleim, A. Mehler, and H.-J. Eikmeyer. Representing and maintaining large corpora. In *Proceedings of the Corpus Linguistics 2007 Conference, Birmingham (UK)*, 2007.

[7] R. Gleim, A. Mehler, H.-J. Eikmeyer, and H. Rieser. Ein Ansatz zur Repräsentation und Verarbeitung großer Korpora multimodaler Daten. In *Proceedings of the Biennial GLDV Conference 2007, 11.-13. April, Universität Tübingen*, Tübingen, 2007. Narr.

[8] R. C. Holt, A. Schurr, S. Elliott Sim, and A. Winter. GXL: A graph-based standard exchange format for reengineering. *Science of Computer Programming*, 60(2):149–170, 2006.

[9] M. D. Jordan, editor. *Patrologia Latina database*. Chadwyck-Healey, Cambridge, 1995.

[10] B. Jussen. Ordo zwischen Ideengeschichte und Lexikometrie. Vorarbeiten an einem Hilfsmittel mediävistischer Begriffsgeschichte. In B. S. und Stefan Weinfurter, editor, *Ordnungskonfigurationen im Hohen Mittelalter*, volume 64 of *Vorträge und Forschungen*, pages 227–256. Sigmaringen, 2006.

[11] A. Lüdeling, T. Poschenrieder, and L. C. Faulstich. DeutschDiachronDigital — Ein diachrones Korpus des Deutschen. *Jahrbuch fur Computerphilologie*, pages 119–136, 2005.

[12] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts, 1999.

[13] O. Mason. Parameters of collocation: The word in the centre of gravity. In J. M. Kirk, editor, *Corpora Galore: Analyses and Techniques in Describing English*, pages 267–280. Rodopoi, Amsterdam, 1999.

[14] G. A. Miller and W. G. Charles. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28, 1991.

[15] R. Power, D. Scott, and N. Bouayad-Agha. Document structure. *Computational Linguistics*, 29(2):211–260, 2003.

[16] A. Renear. Out of praxis: Three (meta)theories of textuality. In K. Sutherland, editor, *Electronic Text. Investigations in Method and Theory*, pages 107–126. Clarendon Press, Oxford, 1997.