



storica **MENTE**
LABORATORIO DI STORIA



ALMA MATER STUDIORUM
Università di Bologna
Dipartimento di Storia Culture Civiltà



TECNO

STORIA

STORICAMENTE.ORG

Laboratorio di Storia

Roberta Cimino, Tim Geelhaar, Silke Schwandt

Digital Approaches to Historical Semantics: new research directions at Frankfurt University

Numero 11 - 2015

ISSN: 1825-411X

Art. 7

pp. 1-16

DOI: 10.12977/stor594

Editore: BraDypUS

Data di pubblicazione: 7/7/2015

Licenza: CC BY-NC-ND 4.0 International

Sezione: Tecnostoria

Digital Approaches to Historical Semantics: new research directions at Frankfurt University

ROBERTA CIMINO, TIM GEELHAAR, SILKE SCHWANDT

Univ. of Nottingham, Department of History
Goethe Universität Frankfurt am Main, Historisches Seminar
Bielefeld Universität, Geschichte des Hoch- und Spätmittelalters

In the past few years a multidisciplinary team of scholars based at Goethe Universität Frankfurt has been involved in the development of three projects: the research project “Political language in the Middle Ages: Semantic Approaches”, and two online platforms, “Computational Historical Semantics” and “eHumanities Desktop”. These are closely related to each other, as they bring together historical research on Latin medieval texts and Digital Humanities. This article will offer an overview of the projects, focusing particularly on the digital tools which have been developed by the team.

Political language in the Middle Ages: Semantic Approaches

The project “Political Language in the Middle Ages – Semantic Approaches” (<http://www.uni-frankfurt.de/47656167/politicallanguage?>), led by Professor Bernhard Jussen (Goethe Universität Frankfurt – Historisches Seminar), has been funded by the Gottfried Wilhelm Leibniz Award (2008-2014). It aims at contributing to the history of political ideas by introducing computer-assisted methods of textual analysis. The project employs corpus-linguistic methods developed by research in modern history and applies them to medieval texts, by making use of

the growing number of digital editions of Latin sources.

This approach is taking an established field of study, the *Begriffsgeschichte*, or history of concepts, to a new level. The last two generations of historians have been paying increasing attention to the meaning of words and concepts, and to the way in which they can help us to understand historical change. Developed in Germany, particularly at the University of Bielefeld, *Begriffsgeschichte* studies the way in which concepts were used, understood and evolved in past societies. Its most famous exponent, Reinhart Koselleck, has focused on the way in which concepts were semantised, that is to say how semantics were created and shared by society. His approach is based on the concept of repeatability, the importance of the meaning of a concept to be shared and understood by multiple individuals. Only when a collective group shares and agrees upon a meaning, does this become important. In other words, society creates semantics, and at the same time, semantics shape society in terms of the way in which a collective group defines itself and communicates among its members. As well as *Begriffsgeschichte*, Historical Semantics is a discipline based on a “word-centred” research approach, namely concentrating on the historical use of specific terms. Historical Semantics allows us to explore conditions and processes through which meanings were shaped in past societies. This discipline provides scholars with the opportunity of bringing together two approaches: a hermeneutic discipline as history, and, on the other hand, quantitative investigation, which can be enhanced by informative computational analysis. “Political language in the Middle Ages” aims to be a crossover between these two fields and to investigate the diachronic and synchronic evolution of political concepts and ideas in medieval Latin texts through the support of Digital Humanities.

However, a politics-specific vocabulary is missing throughout most of the Middle Ages, as concepts such as politics, political, policy etc. are absent from the medieval vocabulary. This means that a differentiation of societal systems is not possible, or at least very problematic, for medieval

societies. For this reason, when looking for political language, historians have, first, to define what can be understood as political. The project follows a non-essentialist, communicative concept of politics. That means that there are no specific issues that can be considered as political, but rather modes of communication creating a political dimension. It takes into account the work carried out by the Collaborative Research Centre “The Political as Communicative Space in History”, based at Bielefeld University (2000–2012). According to this approach, the “Political” can be identified as all practices, discourses and definitions of boundaries that: a) transcend individuals in effectiveness and work on an extensive basis; b) are not ephemeral, but permanent; c) aim at obligation¹. These conditions show the importance of language for defining the political in societies without their own concept of politics. Moreover, they also reveal that the study of political language means facing an extremely complex and varied corpus of texts which are not necessarily political according to our modern categorisation.

Furthermore, “political language” can be conveyed not only through specific words, but also through the relationship between a word and a semantic context in which it is used. Therefore it is not only to be found by looking for single terms, but also by looking for recurring phrases, manners of speech and argumentative patterns. In order to understand these complex patterns and relationships, it is necessary to look at the linguistic structure of the text. The way in which we understand the meaning of a word can be profoundly influenced by the other terms it is associated with. To do this the relationship between words needs to be analysed on a multilayered level: not only in terms of meaning, but also syntactically. Through this kind of analysis it is possible to trace distinctive patterns of word uses, which are linked to different meanings, and therefore can reveal significant information on the argumentative structure of a text.

¹ Cited from: [http://www.uni-bielefeld.de/\(en\)/geschichte/forschung/sfb584](http://www.uni-bielefeld.de/(en)/geschichte/forschung/sfb584).

This is where Digital Humanities can help historical analysis. Even if historians are able to study in-depth the use of a specific word, they cannot necessarily do the same for combinations of words, or manually analyse a vast number of texts. Computational analysis allows one to carry out investigations of co-occurrences, that is to analyse the use of multiple terms through their relationships with other relevant words. The user will therefore easily obtain information on how often certain words are used in conjunction, and in the same sentence, within a text. Furthermore, the computer offers the opportunity to carry out this analysis on an extensive corpus of texts, much larger than what could be possibly analysed manually. Once accumulated, the numerical data on the occurrences and co-occurrences of certain words must be analysed. In order to interpret the use of these patterns, they need to be correlated with the historical context or the communicative situation that they represent, as well as the theoretical framework of the general relation between language and society. What are the moments in which a meaning becomes stable? When does it change? Analysing language patterns helps us to identify meanings and their changes within societies. In other words, language patterns can help uncover and follow the existence and evolution of mechanisms of sense production. By tracing phenomena that connect words and meanings – such as semantic interaction, syntactic correlation and argumentative interrelation – historical semantics illuminates the perception and interpretation of the world in past societies. The variations of these phenomena reflect or foster conflicts or changes within the society itself. In order to achieve these aims, “Political languages in the Middle Ages” combines qualitative analysis with quantitative data, using a corpus based approach, which is possible thanks to the opportunities offered by the Computational Historical Semantics Project.

Computational Historical Semantics: a project and a digital tool

Computational Historical Semantics (CHS) (<http://www.comphistsem.org/project.html>) is a cooperative project between four different universities and which combines three different disciplines. The project has been developed at the University of Frankfurt by an interdisciplinary team led by Bernhard Jussen and Alexander Mehler, head of Text Technology Lab at the Department of Computer Science and Mathematics (Fachbereich Informatik und Mathematik). They are joined by specialists in Romance studies at the Universities of Tübingen (the late Professor Peter Koch, now Professor Sarah Dessi Schmid), Regensburg (Maria Selig) and Bielefeld (Barbara Job). The project aims at defining historical-semantic spaces and developing tools for its analysis, by conducting computer based research on processes of linguistic change. In addition, the project strives to achieve two further goals: first, to create a comprehensive database, which will allow researchers to obtain new insights in processes of semantical changes, such as linguistic change, lexicalisation and grammaticalisation. Secondly, it aims at helping methodological development, as it brings together research methods from different disciplines that are currently separate, and as such has an impact on the way in which scholars analyse texts. Scholars interested in the synchronic and diachronic changes in language, grammar and syntax can make use of the CHS website. It contains a considerable database of about 4.000 Latin texts written between the second and the fifteen century A.D. The corpus has been built thanks to the support of several collections of digitised texts: so far, the *Patrologia Latina*, *Monumenta Germaniae Historica*, *Corpus Corporum* and the *Bibliotheca Augustana*. Since CHS is a work in progress, the database is continuously growing as new texts are being added. The database can be accessed and explored through several search tools, which analyse and compare texts through a computational-linguistic approach.

CHS is based on two databases, the lexicon and the text database, both

of which can be accessed from the homepage. The “Lexicon” (<http://www.comphistsem.org/lexicon.html>) allows the user to search for specific words, and therefore to acquire quantitative data on the occurrence of said words. Through this search, the user will obtain information on the grammatical, linguistic and lexical use of the word. Additionally, the “Text” section (<http://www.comphistsem.org/texts.html>) allows the user to carry out a text based search of the whole corpus, and as such to conduct sophisticated searches for word occurrences and co-occurrences in individual texts or groups of texts.

1. Lexicon

By clicking on the section “Lexicon” the user has the opportunity to carry out a multilayered word based search. The three categories through which the search can be carried out are Word Form, Lemma and Super Lemma. The Word Form category indicates a specific declined form of a word (for example *ecclesiam*, accusative singular of *ecclesia*). By searching for a specific Word Form, the user will obtain quantitative information: the number of times in which the Word Form occurs throughout the CHS corpus. For example, the results report that *ecclesiam* appears 60075 throughout the text corpus (Fig. 1). Furthermore, the user will also obtain morphological and lexical information: the results indicate which Part of Speech (PoS) the Word Form belongs to, as well as its Number (in case of *ecclesiam*: singular), Case (accusative) and Genus (feminine). When searching for a verbal Word Form, the user will also obtain information on Person, Tense, Mood and Voice.

Secondly, it is possible to search for the Lemma, which corresponds to a specific spelling form of a word. Following up the example used above, the corresponding Lemma of *ecclesiam* is *ecclesia*. By typing *ecclesia* into the Lemma mask, the user will learn that it appears 312662 in the database, and that it is composed of 71 Word Forms. Finally, the search takes the user to the highest level of this pyramidal structure: the Super Lemma, which represents the normalised spelling of a word and serves



Fig. 1. The lexicon search mask on www.compshistsem.org: searching a Word Form.

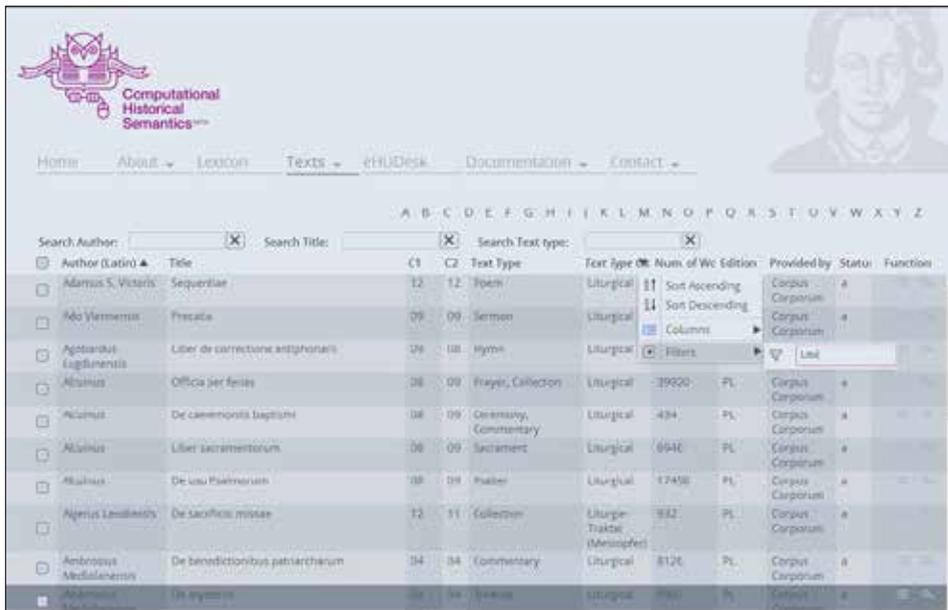
to collect all possible spellings of the same term.

By searching for the Super Lemma *ecclesia*, the user will learn that it corresponds to 12 Lemmata (that is 12 ways of spelling the same word: *aecclesia*, *ecclesia*, *eclesia*, etc.) and obtain numerical data on the occurrences of each Lemma. This structure and organisation is crucially important to operate effective lexical searches on medieval Latin texts. Medieval Latin is extremely fluid, and continuously changing, and therefore many words are spelt in different ways, according to who the author was as well as where and when he was writing. The organisation of Super Lemma – Lemma – Word Form allows users to understand the use of words on multiple levels – grammatical, linguistic and lexical – and to avoid being misled by the frequent spelling variations which can be found in medieval texts.

2. Texts

The second option is to operate a search in the Text database. This allows one to focus on a specific text, author, or on a group of texts selected according to period or genre. Users have the possibility to explore CHS by isolating individual texts or groups of texts according to their research interests. For example, they will be able to search for texts written by the same author, in the same period, and which belong to the same text genre. This can be done by applying filters by clicking on the

heading of each column (Fig. 2). If, for example, the user is interested in liturgical texts written in the twelfth century, he can select the filter “Liturgical” on the text type column, and the filter “12” in the C1 and C2 column (C1 and C2 offer information on the life span of the author of a text. They differ in cases in which the author lived over the turn of a century). As the database is chronologically and geographically very broad, this will provide users with a good tool to familiarise themselves with the relevant content of the corpus. The user will also be able to analyse co-occurrences, in order to understand how often two words are used in association with each other within a predefined length of text, namely in the same sentence. Finally, the user can also compare two or



Computational Historical Semantics

Home About Lexicon Texts ehUDesk Documentation Contact

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Search Author: [X] Search Title: [X] Search Text type: [X]

<input type="checkbox"/>	Author (Latin) ▲	Title	C1	C2	Text Type	Text type	Num. of Wo. Edition	Provided by	Status	Function
<input type="checkbox"/>	Adamus S. Victoris	Sequentiae	12	12	Poem	Liturgical	11	Corpus Corporum	+	
<input type="checkbox"/>	Ado Vitervensis	Preceae	09	09	Sermon	Liturgical	14	Corpus Corporum	+	
<input type="checkbox"/>	Agobardus Lugdunensis	Liber de correctione antiphonarum	09	08	Hymn	Liturgical	1	Corpus Corporum	+	
<input type="checkbox"/>	Alcuinus	Officia per feries	08	09	Prayer, Collection	Liturgical	39920	PL	Corpus Corporum	+
<input type="checkbox"/>	Alcuinus	De caeremonia baptismi	08	09	Ceremony, Commentary	Liturgical	494	PL	Corpus Corporum	+
<input type="checkbox"/>	Alcuinus	Liber sacramentorum	08	09	Sacrament	Liturgical	6940	PL	Corpus Corporum	+
<input type="checkbox"/>	Alcuinus	De usu Paschensium	08	08	Prayer	Liturgical	17458	PL	Corpus Corporum	+
<input type="checkbox"/>	Agerius Laudunensis	De sacrificio missae	12	11	Collection	Liturgical (Missopfer)	932	PL	Corpus Corporum	+
<input type="checkbox"/>	Ambrosius Mediolanensis	De benedictionibus patriarcharum	04	04	Commentary	Liturgical	8126	PL	Corpus Corporum	+
<input type="checkbox"/>	Ambrosius	De ingenuis	04	04	Treatise	Liturgical	960	PL	Corpus Corporum	+

Fig. 2. Filtering the text database.

more texts according to their linguistic and lexical features. By accessing the section “Texts” (<http://www.comphistsem.org/texts.html>), the user will see the full list of the texts contained in the database – which are ordered alphabetically according to the author’s name –

scholars who are focusing on a specific work or author. Most importantly, the numerical data provided by the search results will offer the user a starting point to formulate questions and to determine answers on the grammatical and lexical choices of an author.

2b. Analysing specific words in corresponding texts

The second possibility is the search for co-occurrences. This provides a very useful tool in order to understand how words are combined with each other. Having found the text of interest, the user will click on the third icon, “Analyse specific terms”, represented by the magnifying glass symbol, located in the far right column of the table (Icon 3 in Fig. 3). At this point a search mask will appear, called “Lemma Query”. Here the user will type in the search word. The results will then be a tabulated listing all the Lemmata which appear in the same sentence as the search word. The first line of the table always contains the search Lemma: the table reports the number and percentage of co-occurrences (in this case, obviously, 100%) and a list of the Word Forms of which the Lemma is composed. Underneath the first line, the user will find the co-occurring Lemmata, in decreasing order of frequency. The data indicate the amount of times in which each Lemma is used in the same sentence as the search word throughout the entire text.

2c. Difference and intersection

This function allows users to compare the occurrence of the same term in two different texts. In order to use this function, it is first necessary to obtain general term lists for two or more texts (see 2a. *Lexical analysis of an individual text*).

Through the icon “Compare statistics” on the right top side of the term list, users will be able to select the texts that they intend to compare. At this point the user will obtain a more elaborate table (Fig. 5). The table will list, in order of decreasing frequency, the Lemmata which appear

in the first selected text (on the left side), and compare this frequency with the second text (on the right side). The user may select two texts written by Pope Gregor I, such as the *Moralia* and the *Regula Pastoralis* (as shown in Figure 5), and compare the different lexical choices of the same author in two different works. Similarly, the user could compare, for instance, two texts written by different authors in the same period.

The screenshot shows a search interface with a list of texts and a detailed comparison of lemmas between two selected texts: *Moralia* and *Regula Pastoralis*.

Search Results Table:

Search Author:	Search Title:	C1	C2	Text Type	Text Type C1	Num. of We	Edition	Provided by	State	Function
Gregorius I	Homiliae in Ezechielum	06	07	Bible Exegesis: Old Testament; Sermon	Theological; Paraenetic	121883	FL	Corpus Corporum	a	
Gregorius I	Moralia	06	07	Bible Exegesis: Old Testament; Treatise (Thematic)	Theological; Paraenetic	317974	FL	Corpus Corporum	a	
Gregorius I	Moralia	06	07	Bible Exegesis: Old Testament; Treatise (Thematic)	Theological; Paraenetic	267710	FL	Corpus Corporum	a	
Gregorius I	Regula Pastoralis	04	05	Treatise (Thematic)	Paraenetic	46845	FL	Corpus Corporum	a	
Gregorius I	Homiliae in Evangelia	06	07	Bible Exegesis: New	Theological	100678	FL	Corpus	a	

Lemma Comparison Table:

Lemma	POS	Dubium	Moralia #	Moralia %	Moralia Word Form(s)	Regula Pastora	Regula Pastora	Regula Pastoralis Word Form(s)
deus	Noun	<input type="checkbox"/>	1815	1.045	deus(18) deus(2) deus(104)	146	0.368	de(58) deo(57) deum(15) deus(22)
sua	Pronoun	<input checked="" type="checkbox"/>	1097	0.632	sua(41) sua(49) suae(1) suarum(1) suarum(45) suar(3) sub(que)3) suat(33) suae(1) suae(1)	138	0.509	sua(32) suae(1) suarum(2) suae(4) sub(que)1) suae(10)
omnis	Adject.	<input type="checkbox"/>	1034	0.595	omne(19) omnes(38) omnesque(2) omnes(234) omnesque(6) omne(4) omnia(190) omniaque(3) omniaque(2) omnia(120) omniaque(1) omnia(70)	81	0.362	omne(18) omnes(1) omnes(11) omne(12) omnia(14) omniaque(14) omnia(14) omniaque(7)
homo	Noun	<input type="checkbox"/>	804	0.568	homin(50) homin(29)	83	0.323	homin(8) homin(8)

Fig. 5. Intersection for Gregory I's *Moralia* and *Regula Pastoralis*.

Managing text corpora: the eHumanities Desktop (eHU Desk).

The opportunities offered by the CHS can be taken to a more sophisticated level by another platform, the eHumanities desktop (eHU Desk) (<http://www.hudesktop.hucompute.org>), developed by the “Text Technology Lab” (<http://www.hucompute.org/ressourcen/ehumanities-desktop>), and headed by Prof. Mehler at Goethe University in Frankfurt. Unlike CHS, access to the eHU Desk is limited, and users need to register before they can work with the “Historical Semantics

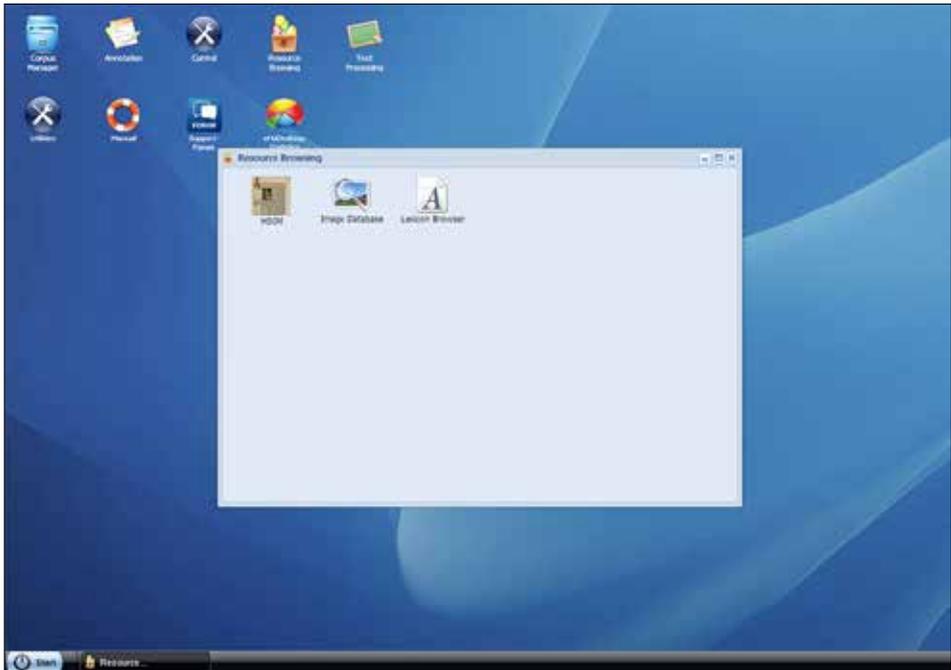


Fig. 6. HSCM and the Lexicon Browser on eHU-Desk.

Corpus Management” (HSCM) tool and the “Lexicon Browser” for accessing the Latin Lexicon. In order to obtain an account, interested users can contact the team via email (support@hudesktop.org). HSCM offers further interesting opportunities: the creation and management of corpora, and the possibility to carry out sophisticated search queries. Users can even upload their own texts via the “Corpus Browser” and analyse them through HSCM. They can also annotate their own texts with meta-information (such as author, date, title, providence, place, text type, etc.). Furthermore, users can eliminate mistakes in the text in order to get a clean and correctly tagged text that can be downloaded. As the figure below shows (lower part of the table in Fig. 7), there are several query options, which allow the user to obtain different types of data. The “Possible Hits” query shows all mentions of the search word in the texts of a chosen corpus (in the example below, the searched corpus is “Liturgical Texts”, which so far contains 256 texts). Through this query one obtains an overview on how many texts contain the search

cide whether the list shall be computed on the Super Lemma-Level, the Lemma-Level or the Word Form-Level. These options offer valuable insights into the syntactical structure of a text. Finally, it is possible to search specific parts of a text, such as the Header or the Main Body, as shown in the example below.

Further options are being tested at the moment, and more texts will be added in the future. A manual with full explanations, as well as the already visible help-function, are on their way, so that scholars will soon be able to explore the program and contribute with suggestions, critiques and perhaps also with their commitment to develop this tool.

While the work on these projects continues by enhancing the usability, creating new analysing functions and enlarging the text database, scholars have the opportunity to familiarise themselves with the innovative tools offered by CHS. This growing database will offer scholars new opportunities in textual analysis, providing fast and reliable numerical data that can then be subjected to hermeneutical interpretation. The

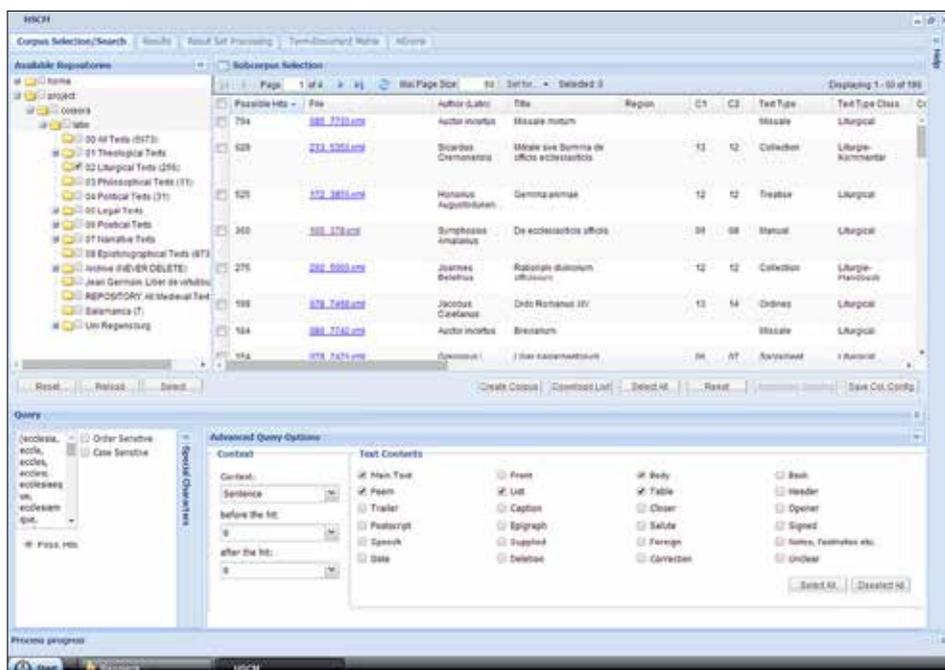


Fig. 8. The Advanced Query Options on HSCM.

program also offers the opportunity to interact with the working team, by signalling errors or ambiguous results and providing comments. Moreover, users can also contribute to the database by providing their own texts (for example transcriptions of manuscripts). CHS is aimed at all scholars who come from different disciplines, research approaches and methodologies, but are equally interested in linguistic changes in medieval Latin texts. It offers exciting opportunities to uncover diachronic and synchronic evolutions in lexical and syntactical structures, and as such will contribute to our understanding of the way in which concepts were understood, shared and transmitted in the medieval world.

Reference list

- Brunner O., Conze W., Koselleck R. (eds.) 1972, *Geschichtliche Grundbegriffe: historisches Lexikon zur politisch-sozialen Sprache in Deutschland*, 8 Voll. Stuttgart: Klett-Cotta.
- Jussen B. 2011, *Historische Semantik aus der Sicht der Geschichtswissenschaft*, in Schmid H.U., Ziegler, A. (eds.), *Jahrbuch für Germanistische Sprachgeschichte*, Vol. 2, Berlin: De Gruyter, 51-61.
- Jussen B., Mehler, A., Ernst, A. 2007, *A Corpus Management System for Historical Semantics*, «Sprache und Datenverarbeitung. International Journal for Language Data Processing», 31 (1-2): 81-89.
- Koselleck R., Spree U., Steinmetz W. 2006, *Begriffsgeschichten: Studien zur Semantik und Pragmatik der politischen und sozialen Sprache*, Frankfurt am Main: Suhrkamp.
- Mehler A., Schwandt S., Gleim R., Jussen B. 2011, *Der eHumanities Desktop als Werkzeug in der historischen Semantik: Funktionsspektrum und Einsatzszenarien*, «Journal for Language Technology and Computational Linguistics», 26 (1): 97-117, http://www.jlcl.org/2011_Heft1/8.pdf.
- Meier U., Papenheim M., Steinmetz W. 2012, *Semantiken des Politischen Vom Mittelalter bis ins 20. Jahrhundert*, Göttingen: Wallstein Verlag.
- Reichardt R. 1998, *Historische Semantik zwischen lexicométrie und New Cultural History. Einführende Bemerkungen zur Standortbestimmung*, in Reichardt, R. (ed.), *Aufklärung und Historische Semantik: interdisziplinäre Beiträge zur westeuropäischen Kulturgeschichte*, Berlin: Duncker & Humblot, 7-28.
- Steinmetz, W. 2007, *Neue Wege einer historischen Semantik des Politischen*, in Steinmetz, W. (ed.), «Politik»: *Situationen eines Wortgebrauchs im Europa der Neuzeit*, Frankfurt am Main-New York: Campus, 9-40.