List of Topics

1 Introduction 1
2 Resource Management and Modeling 1
3 Integrated Representation of Lexica and Text Corpora 4
4 eLexicon Browser 4
   4.1 Browsing & Searching 4
   4.1.1 Working with query results 7
   4.1.2 Editors 7
   4.1.3 Ontologies 8
5 Feedback and Support 10
6 How to Cite this Work 10
7 Acknowledgement 10

1 Introduction

This report presents the latest additions and improvements of the eHumanities Desktop. The eHumanities Desktop is an extensible web-based system which enables users to organize, browse, process and analyse resources online. All functions are offered by means of application modules. The eHumanities Desktop is actively used as a working platform by several research projects, and has been recently extended with new applications. This concerns in particular the projects IV Multimodal Corpus Management of the LOEWE center and the Leibniz projects of Bernhard Jussen. Figure 1 shows a screenshot of the eHumanities Desktop with the Corpus Manager opened.

This report marks the release of version 2.0 which was published on 05/24/2012. This release offers a new and integrated representation of lexica and text corpora which is presented in this report. Note that the previous version 0.3.12 is still online and hosting the current release of the ImageDB. The next release of the eHumanities Desktop (2.2) will merge both distributions. This is due to major redesigns of the server software.

2 Resource Management and Modeling

Feedback from users working with the eHumanities Desktop as well as from developers designing new applications has led to a major redesign of the system. Version 2.0 marks the first public release which is based on the new system design. An important feature of the new design is the graph-driven representation of documents, repositories, annotations as well as users and groups and their
Figure 1: Screenshot of the eHumanities Desktop showing the Corpus Manager.
interrelations. Technically, we now use the graph database Neo4J as a persistance layer, which has proven to scale well on large amounts of data. Every aspect of the master database is now consequently modeled by means of a typed, attributed graph. Consequently, annotation schemas as well as annotations are now modeled as graphs, using RDF as the interchange format. Figure 2 shows a schematic example of how two annotations of different users on the same document are modeled in terms of a graph. Furthermore, the developers API of the eHumanities Desktop has received a general redesign. These improvements “under the hood” greatly simplify application development and ultimately give end users more powerful tools to work on their resources. Another major addition is the integration of Apache Hadoop which allows to store documents in a cluster. On the one hand this step further increases protection of data against hardware or system failure. On the other hand it will allow to implement distributed computation on data. As in the previous version the way how exactly a document is stored in the eHumanities Desktop is transparent to the user.

The application modules HSCM and eLexicon Browser are the first modules which benefit from the new system design. Both applications underwent a major redesign. In case of the HSCM the user interface has not changed much (in contrast to major works on the server side). The eLexicon Browser reflects best the new features. Therefore we put a focus on this module. Please note that the ImageDB, the annotation components and other modules are not yet available in the 2.0 release. They can be used in the 0.3.12 release. The next major release 2.2 will integrate all modules and data again.
3 Integrated Representation of Lexica and Text Corpora

With release 2.0 of the eHumanities Desktop, lexica and text documents represented in TEI P5 are stored in a graph database. Every word form (w-Tag in TEI P5) can explicitly be linked to a corresponding entry in a lexicon. Figure 3 exemplifies how a sentence and its constituents are linked with entries of a lexicon. This allows for a line of new features which are presented in the following section.

4 eLexicon Browser

The eLexicon Browser allows for browsing and editing lexica. With the new release, it benefits from the explicit linking of lexicon items with indexed text documents which are represented according to the TEI P5 standard. That is: modifications in lexica can directly be applied to documents in order to correct erroneous linguistic annotations. In the following sections we describe some of the important features of the eLexicon Browser.

4.1 Browsing & Searching

Queries: an important means to browsing and exploring lexica. Queries can aim at finding exact matches or they may include wildcards (* and ?). By default only lemmas are searched. The type of objects to be searched can be changed according to the ontology of the selected lexicon. Standard types are “Lemma”, “Word Form”, “Super Lemma”, “Lexeme Group” and “Synset”. Beyond that, the system allows users to define their own ontologies for lexica which can likewise be incorporated for queries. In some cases it may be relevant to know what the last activity on a given lexicon item was. Therefore queries can be limited to a specific activity (“create”, “change”) of a specific user. Furthermore, queries can be either case sensitive or case insensitive, and the part-of-speech (POS) for the query can be specified. In the current implementation the POS is global; that is, it applies to all lexica.

Optionally some statistical measures (namingly IDF, RIDF, corpus frequency, text frequency) can be computed for word forms, lemmas and super lemmas. The values are computed ad hoc and depend on the set of documents which are linked to the lexicon entries. Thus, adding a new document or altering the linking between a document and the lexicon will directly affect the values of the statistical measures. Figure 5 shows a screenshot of the eLexicon Browser which demonstrates how statistical data of word forms can be visualized by means of plots online.

The query result is presented in the form of a table. The number and type of columns as well as the list of menu items depend on the type of lexicon objects which have been retrieved. For example a query on word forms will offer an option to get to the instances within all linked documents(!). A query for lemmas on the other hand will give access to the respective list of word forms. Finally, results of a query for super lemmas give access to the associated lemmas (which may, for example, be spelling variants). The screenshot shown in Figure 4 exemplifies how query results for word forms are displayed.
Figure 3: This graph gives an example of how a sentence is represented and linked to a lexion.
Figure 4: The eLexicon Browser of the eHumanities Desktop
Note that a given lemma is always assigned to a super lemma. In order to avoid lemmas without a superior super lemma, lemma creation is only possible in the context of an existing super lemma. Existing lemmas can also be moved to other super lemmas when they appear to be assigned incorrectly.

Query parameters can be saved and restored later.

4.1.1 Working with query results

Word forms, lemmas, super lemmas and lexeme groups when shown as results of a query can be linked to types of a user defined ontology. Likewise, such lexicon entries can be interrelated based on a user-defined relation ontology. This functionality requires that the user has access to this feature of the eHumanities Desktop and that a proper node and/or relation ontology has been created for the lexicon. A double-click on an entry in the query results opens a graphical view of the element as well as an editor for some specific properties. Some options of the context menu of query results are specific to the type of the query results:

**Word Form** gives access to the word form editor which allows users to delete the linking between the selected word form and all documents.

**Lemma** gives access to the lemma editor which allows users to edit the word forms of the selected lemma.

**Super Lemma** gives access to the lemmas of a selected super lemma. It also allows users to create a new super lemma.

**Lexeme Group** offers a means to browse and edit super lemmas for a given lexeme group. The editor also allows users to create new lemma groups. The system distinguishes global lexeme groups which are visible to all users from custom groups which are only visible to the current user.

4.1.2 Editors

As already mentioned, word forms, lemmas, super lemmas and lexeme groups can be edited when the user has the proper permissions. The editors can be accessed via the context menu in query results. Data is presented in tables and can be grouped and sorted by any given column. Most columns offer a filter which allows users to limit the entries of the table according to specific properties and values.
Word Form Editor   The word form editor gives the user a fine grained control over the linking between word forms in the lexicon and their instances in TEI P5 documents (w-Tags). A word form is always presented with its context to help judge whether a specific linking is correct.

Linked word forms are highlighted by means of a color code. Green represents an unambiguous assignment. Yellow indicates that multiple wordforms exist, but that they all belong to the same lemma. Orange means that the assignment is ambiguous and that the ambiguity spans several lemmas.

Links to instances can be established or deleted (for example to resolve an ambiguity). This function can also be accessed via the HSCM when a word form is clicked which is linked to the lexicon.

Lemma Editor   The lemma editor allows users to edit the word forms of a lemma. For every word form the morphological information as well as the corpus- and text-frequency is displayed. By opening the context menu on a word form the word form editor can be started. Furthermore the context menu allows users to edit the linking between the selected word form and the document collection.

When a new word form has been created a query is triggered which looks for potential instances of the new wordform in TEI P5 documents. Then the new word form is automatically linked with the instances found.

A double-click on an entry in the table shows a graphical view of the word form which also offers users the option to edit additional properties of the entry.

Super Lemma Editor   The super lemma editor supports, among other things, the creation of new super lemmas and adding or deleting lemmas. By opening the context menu on one of the lemma entries the word form editor can be opened. Furthermore the menu allows users to automatically create links to TEI P5 documents.

When creating a new lemma the eLexicon Browser offers an automatic expansion of possible word forms based on specified morphological informations. The result of such an expansion is displayed in a table which can then be corrected and refined as required. Finally the newly created word forms can be linked to instances in the document collection.

A double-click on a lemma opens its graphical representation where properties can also be edited.

Lexeme Group Editor   The lexeme group editor allows users to group super lemmas. Only existing super lemmas can be added. A combobox which supports auto-completion allows users to search for super lemmas. Super lemmas which already have been added to a lexeme group are highlighted.

4.1.3 Ontologies

For a given lexicon an arbitrary number of ontologies can be defined by the user. The system distinguishes between ontologies for nodes and relations. Deleting an ontology or removing parts of it directly affect annotations of lexicon entries accordingly. Lexicon elements are not deleted. For this purpose a tree editor is being offered.
Figure 5: Screenshot of the eLexicon Browser showing plots of statistical data of wordforms.
Node Ontology  The node ontology editor allows users to freely create hierarchies of types which can later be used to type lexicon entries. Lexicon entries can be typed via the context menu in the query result tables.

Relation Ontology  The relation ontology editor allows users to define custom types of relations between lexicon entries. The creation of custom typed relations can be done via context menus on entries in query results. Relations can be deleted by opening the Element Relations-view (which is accessible via a double-click on an entry).

5 Feedback and Support

If you wish to get an account for the eHumanities Desktop or if you wish the give feedback or report a bug you are welcome to contact us via support@hudesktop.org.

6 How to Cite this Work

If you are using the eHumanities Desktop in one of your publications, please cite one of the following references:

- Gleim:Warner:Mehler:2010

7 Acknowledgement

This work has been carried out by the Text Technology Lab at the Goethe-University Frankfurt within the LOEWE center Digital Humanities (http://www.dhhe.de/) and the the Leibniz project Political Language in the Middle Ages (https://semantik.geschichte.uni-frankfurt.de/). Financial support is gratefully acknowledged.
Notes