Extracting domain knowledge from tables of contents

Harald Lüngen
luengen@uni-giessen.de
Justus-Liebig-Universität Gießen

Henning Lobin
henning.lobin@germanistik.uni-giessen.de
Justus-Liebig-Universität Gießen

1. Introduction

Knowledge in textual form is always presented as visually and hierarchically structured units of text, which is particularly true in the case of academic texts. One research hypothesis of the ongoing project *Knowledge ordering in texts—text structure and structure visualisations as sources of natural ontologies*[1] is that the textual structure of academic texts effectively mirrors essential parts of the knowledge structure that is built up in the text. The structuring of a modern dissertation thesis (e.g. in the form of an automatically generated table of contents - *tocs*), for example, represents a compromise between requirements of the text type and the methodological and conceptual structure of its subject-matter. The aim of the project is to examine how visual-hierarchical structuring systems are constructed, how knowledge structures are encoded in them, and how they can be exploited to automatically derive ontological knowledge for navigation, archiving, or search tasks. The idea to extract domain concepts and semantic relations mainly from the structural and linguistic information gathered from tables of contents represents a novel approach to ontology learning.

2. Data and annotations

In the present phase, we examine German academic text books, in later phases, dissertations, research articles and historical scientific texts will also be taken into account. A corpus of digital versions of 32 text books from 12 different academic disciplines has been compiled,[2] the textual content and an XML document structure markup was subsequently extracted (e.g. using the Adobe Pro software). Using a series of XSLT style sheets, the initial XML was converted to XML encoding of the document structure according to the TEI P5 guidelines. At the same time, the texts were annotated with morphological analyses and phrase chunking markup using the Tree Tagger and Chunker software from Stuttgart University[3] and converted to a suitable XML representation, and also with dependency-syntactic analysis using the Machinese Syntax Parser by Connexor Oy,[4] resulting in XML markup as well. Further linguistic annotation levels (such as domain terms and lexical-semantic relations) will be added and combined in XStandoff documents representing multi-layer annotations that can be queried using XML standards and tools as described in Stührenberg & Jettka (2009).

Presently, all available annotation layers are stored in an eXist native XML database[5] and are queried using the Oxygen XML editor[6] as a database client.

The corpus infrastructure is used to explore the document applying the method of toc fragment analysis as described in the following section, and to implement functions for concept extraction and semantic relation analysis.

2.1. Analysis of toc fragments

Our method of analysing toc fragments consists of the following steps:

1. Identification of a toc fragment
2. Representation of the fragment meaning as a MultiNet
3. Identification of the configuration of elements on different structural levels that induce the fragment meaning
4. Hypothesis about the generalisation of a toc fragment, a structuring schema
5. Corpus research to verify the generalisation hypothesis

```
[booktitle] Einführung Pädagogik
[5.] Ausgewählte Subdisziplinen und
 Fachrichtungen
 [5.1.] Literatur
 [5.2.] Erlebnispädagogik
  [5.2.1.] Begrifflichkeit
   [5.2.1.1.] Erlebnis als prioritäre Kategorie
```

```
 [5.2.2.] Historie
 [5.2.3.] Theoretische Fundierungen und
Menschenbilder
 [5.2.4.] Ziele und Funktionen der
Erlebnispädagogik
  [5.2.4.1.] Ziele
  [5.2.4.2.] Subjektbezogene Funktionen und
mögliche Wirkungsweisen
  [5.2.4.3.] Gesellschaftliche Funktionen
 [5.2.5.] Merkmale und Modelle der
Erlebnispädagogik
 [5.2.6.] Beispiele erlebnispädagogischer
Angebote
  [5.2.6.1.] Outward Bound-Konzeption
  [5.2.6.2.] Outdoor Management Development
 [5.2.7.] Kritikpunkte
 [5.2.8.] Einführungsliteratur (zum
Weiterlesen)
[5.3.] Erwachsenenbildung
 [5.3.1.] Begriffsklärung
 [5.3.2.] Geschichtliche Entwicklung
 [5.3.3.] Struktur und Funktionsperspektiven in
der Erwachsenenbildung
 [5.3.4.] Theoretische Orientierungen der
Erwachsenenbildung
 [5.3.5.] Forschungsfelder
 [5.3.6.] Einführungsliteratur (zum
Weiterlesen)
[5.4.] Gesundheitspädagogik
```

Figure 1: Section from the table of contents of Raithel et al. (2007)

Consider the section of the generated table of contents of the text book *Einführung Pädagogik* by Raithel et al. (2007) shown in Figure 1. By choosing the heading 5. *Ausgewählte Subdisziplinen und Fachbereiche* and its immediately superordinated heading (in this case the title of the book) as well as its immediately subordinated headings, we arrive at the toc fragment (or "window") shown in Figure 2. In the toc fragment, four terms from the domain are contained, *Pädagogik*, *Erlebnispädagogik*, *Erwachsenenbildung*, and *Gesundheitspädagogik*.[7] The terms identification component must distinguish such expressions denoting domain-specific concepts from relational nouns commonly found in academic and scientific discourse (such as *Einführung*, *Subdisziplin* and *Fachrichtung*) and from terms denoting text-type structural categories of academic texts such as *Literatur*).

We employ the semantic network approach *Multilayered Extended Semantic Networks* (acronym: MultiNets) by Helbig (2006) to represent the domain concepts and semantic relations between them

```
[booktitle] Einführung in die Pädagogik
[5.] Ausgewählte Subdisziplinen und
 Fachrichtungen
[5.1.]  Literatur
[5.2.]  Erlebnispädagogik
[5.3.]  Erwachsenenbildung
[5.4.]  Gesundheitspädagogik
```

Figure 2: Toc fragment

expressed in a toc fragment. The MultiNet approach is a fully-fledged semantic theory and provides a rich and consistent inventory of semantic entity types, features, relations and functions, and has been previously employed in the syntactic-semantic analysis components of QA systems (Hartrumpf 2005). Using the graphical MWR editor for designing MultiNets,[8] we represent the semantics of the above toc fragment as shown in Figure 3.

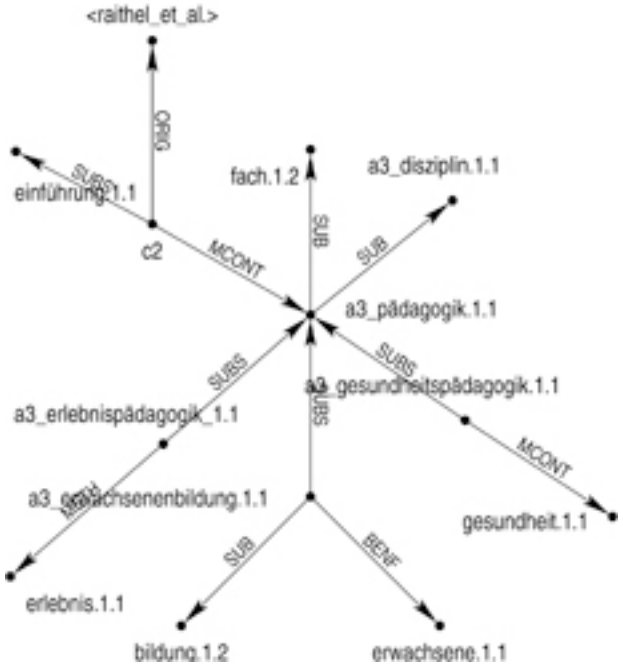

Figure 3: MultiNet analysis of toc fragment

In the semantic network in Figure 3, the concepts `a3_erlebnispädagogik.1.1`, `a3_erwachsenenbildung.1.1`, and `a3_gesundheitspädagogik.1.1` are related to `a3_pädagogik.1.1` by the SUBS relation denoting subordination of (abstract) situations; `a3_pädagogik.1.1` is in turn related to `fach.1.2` and `a3_disziplin.1.1` by the SUB relation denoting the subordination of concepts representing objects.[9] Furthermore, the semantic decompositions of the three compounds are analysed using the relations

MCONT (mental or informational content), BENF (beneficiary) and METH (method), and the relation between the concept `c2` representing the textbook as such and `a3_pädagogik.1.1` is specified as MCONT, the relation between `c2` and its authors as ORIG (mental of informational origin, cf. Helbig 2006).

On account of this analysis, the following hypothesis is formed:

Given a potential structuring schema, consisting of an initial expression N, and an expression N-1 related to N by a *heading_of* relation on the document structure level, and an expression N+1 to which N is related by the *heading_of* relation on the document structure level (cf. Figure 4), if

```
N-1
 N
  N+1
```

Figure 4: Toc Schema

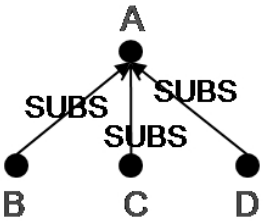

Figure 5: MultiNet Schema

- N contains the lexeme Subdisziplin or a synonym on the lexical level
- and N-1 contains the domain concept A
- and N+1 contains the domain concept B,

then, by multiple application, construct a MultiNet-Schema as represented by the graph in Figure 5.

```
<result doc="schruender-
lenzen_schriftspracherwerb_2007" docID="i72">
     <head level="n-1">9. Schwierigkeiten des
 Schriftspracherwerbs rechtzeitig erkennen und
 gezielt helfen</head>
     <head level="n">9.2 Zentrale
 Wahrnehmungsbereiche und ihr Risikopotential</
head>
     <head level="n+1">9.2.1 Visuelle
 Wahrnehmung</head>
     <head level="n+1">9.2.2 Auditive
 Wahrnehmung</head>
</result>
<result
 doc="brosius_kommunikationsforschung_2008"
 docID="i137">
     <head level="n-1">8. Kapitel:
 Inhaltsanalyse I: Grundlagen</head>
     <head level="n">8.4 Anwendungsgebiete und
 typische Fragestellungen</head>
     <head level="n+1">8.4.1 Inhaltsanalysen auf
 dem Feld der politischen Kommunikation</head>
     <head level="n+1">8.4.2 Inhaltsanalysen in
 der Gewaltforschung</head>
     <head level="n+1">8.4.3 Inhaltsanalysen in
 der Minderheitenforschung</head>
</result>
<result doc="raithel_paedagogik_2007"
 docID="i180">
     <head level="n-1">BOOKTITLE: Einführung
 Pädagogik</head>
     <head level="n">D Ausgewählte
 Subdisziplinen und Fachrichtungen</head>
     <head level="n+1"> Literatur</head>
     <head level="n+1">Erlebnispädagogik</head>
     <head level="n+1">Erwachsenenbildung</
head>
 ...
```

Figure 6: Query Result Document

The Hypothesis is verified by formulating the potential structuring schema as a query to the corpus using the XQuery query language. The query result document then contains a set of toc fragments that can now be inspected as to whether their semantics conform to the hypothesis or not, leading to a small statistic about the validity of the hypothesis. Sometimes the inspection may also lead to a modification of the original query. In the first result fragment in Figure 6, for instance, the superordinate concept *Wahrnehmung* is not contained in N-1, but as the compound modifier of *Bereich* (a synonym of *Subdisziplin*).[10]

In this example it becomes clear that analyses on the morphological and lexical-semantic level interact with the analyses of the structuring information in that both levels provide conditions or constraints when building the semantic analysis of a toc fragment. Our corpus infrastructure is designed such that information from multiple linguistic and structural levels can be taken into account.

2.2. Conclusion

We presently inventorise sets of complex conditions connecting a structuring schema with a MultiNet Schema as *constructions* in the sense of Construction Grammar (CxG). Construction Grammar (Kay 1995, Östman & Fried 2004) is a theory of grammar which is not based on phrase structure rules operating on lexical elements, but as combinations of constructions in which form schemata are

associated with meaning schemata and is therefore appropriate for the description task at hand. The inventory of constructions will then be employed in ontology learning, particularly for the task of automatically extracting domain concepts and semantic relations between them. Constructions describing document structuring schemata as described above play a role similar to the lexico-syntactic "Hearst Patterns" described in Hearst (1992), which have been employed for extracting semantic relations from running text.

References

Brosius, Hans-Bernd, Koschel, Friederike, Haas, Alexander (2008). *Soziologie. Methoden der empirischen Kommunikationsforschung. 4. Aufl.* Wiesbaden.

Glöckner, Ingo, Hartrumpf, Sven, Helbig, Hermann, Leveling, Johannes, Osswald, Rainer (2007). 'Automatic semantic analysis for NLP applications'. *Zeitschrift für Sprachwissenschaft.* **Jg. 26, H. 2**: 241–266.

Hartrumpf, Sven (2005). 'University of Hagen at QA@CLEF 2005: Extending knowledge and deepening linguistic processing for question answering.'. *Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop*. Peters, Carol (ed.). Vienna: Centromedia.

Hearst, Matti A. (1992). 'Automatic acquisition of hyponyms from large text corpora'. *Proceedings of the 14th International Conference on Computational Linguistics*.

Helbig, Hermann (2006). *Knowledge Representation and the Semantics of Natural Language*. Cognitive Technologies. Heidelberg: Springer

Kay, Paul (1995). 'Construction Grammar'. *Handbook of Pragmatics Manual*. Verschueren, Jef, Östman, Jan-Ola, Blommaert, Jan (eds.). Amsterdam: John Benjamins, pp. 171–177.

Östman, Jan-Ola, Fried, Mirjam (eds.) (2004). *Construction Grammars: Cognitive grounding and theoretical extensions*. Amsterdam: John Benjamins.

Raithel, Jürgen, Dollinger, Bernd, Hörmann, Georg (2007). *Einführung in die Pädagogik. Begriff - Strömungen - Klassiker - Fachrichtungen. 2. Aufl.*. VS Verlag für Sozialwissenschaften. Wiesbaden: Springer

Schmid, Helmut (1994). 'Probabilistic Part-of-Speech Tagging using Decision Trees'. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK.

Schründer-Lenzen, Agi (2007). *Schriftspracherwerb. Bausteine professionellen Handlungswissens. 2. Aufl.* VS Verlag für Sozialwissenschaften. Wiesbaden: Springer

Stührenberg, Maik, Jettka, Daniel (2009). 'A tookit for multi-dimensional markup: the development of SGF to XStandoff'. *Proceedings of Balisage: The Markup Conference 2009*. Balisage Series on Markup Technologies. 3 vols.

Tapanainen, Pasi, Järvinen, Timo (1997). 'A non-projective dependency parser'. *Proceedings of the fifth conference on Applied natural language processing*. Washington D.C.

Notes

1. funded within the framework of LOEWE, the excellence initiative of the state of Hesse, as part of the *LOEWE-Schwerpunkt Kulturtechniken und ihre Medialisierung*, cf. `http://www.zmi.uni-giessen.de/projekte/projekt-36.html`.
2. We would like to thank the publishers *Facultas*, *Haupt*, *Narr/Francke/Attempto*, *Springer*, *UTB*, *Vandenhoek & Ruprecht*, and *Wissenschaftliche Buchgesellschaft* for kindly making available digital versions of textbooks for us.
3. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
4. `http://www.connexor.eu/`
5. `http://exist-db.org/`
6. `http://www.oxygenxml.com/`
7. We consider terms to be linguistic expressions that refer to domain concepts.
8. which was kindly made available for us by Professor Helbig's group in Hagen.
9. An `a3_` prefix in a concept name indicates that the concept was not found in the required reading in the semantic lexicon HagenLex (Glöckner et al. 2007) which is consulted by the MWR tool.
10. Other titles from our corpus cited in Figure 6 are Brosius (2008), and Schründer-Lenzen (2008).