



Modelling the logical structure of books and journals using augmented transition network grammars

Logical structure
of books and
journals

69

Birgit Stehno and Gregor Retti
University of Innsbruck, Innsbruck, Austria

Received 19 November
2001

Revised 18 July 2002

Accepted 13 August
2002

Keywords *Documentation, Analysis, Libraries, Internet, Networks*

Abstract *This paper presents a grammar for books and journals using augmented transition networks in automated document analysis. The approach takes the structure of layout elements in books and journals to be part of a semiotic system, which therefore can be described using methods developed for the description of other semiotic systems, e.g. languages. It differs from previous research in the domain of document analysis and understanding as it deals in an exhaustive way with rather generic classes of multi-page printed objects, i.e. books (monographs) and journals. To achieve this aim, abstract relations instead of document specific formatting rules are taken into account.*

1. Introduction

The reformatting of scanned images into structured documents – a central step in the retrospective conversion of printed material – has become an important field of research. A concise overview of previous research is given by Dori *et al.* (2000, p. 424f). As digital images cannot be accessed, searched, and manipulated like texts – let alone structurally encoded texts – the conversion of scanned documents into (Web-)accessible formats is an essential step in the digitisation process. In order to reduce the labour- and cost-intensive human effort, automatic recognition of document structures is indispensable. Automatic recognition and reformatting is required when dealing with large amounts of few and quite standardised documents like business letters, invoices, etc. as well as when managing a variety of different document types like books, journals, archival material, etc. as it is the case in digital libraries and archives.

The METAe project focuses on the latter field aiming to develop an application software (the metadata engine) which extensively automates and improves the digitisation and conversion process. The functionality of the software package will cover the image creation and pre-processing, the import of descriptive metadata like MARC21 (Library of Congress, Network Development and MARC Standards Office, 2001) from electronic library



This study is part of the METAe Project, a research and development project co-funded by the European Commission (5th Framework, IST Programme, Action Lines "Digital heritage and cultural content"; IST-number IST-1999-20021) and the Austrian Federal Ministry of Education, Science and Culture. See <http://meta-e.uibk.ac.at/>

Journal of Documentation
Vol. 59 No. 1, 2003
pp. 69-83

© MCB UP Limited
0022-0418

DOI 10.1108/00220410310458019

catalogues, OCR processing of the scanned or imported images (also of old type faces), the generation of technical and administrative metadata as well as the automated extraction of structural metadata for the digitised object. A permanent quality control on all levels and stages is applied. At the end of digitisation a highly standardised archival information package will be produced – ready for the further processing and integration into a digital library. One focal point of the METAe project is the layout analysis and document understanding component which is responsible for the extraction of structural and partly descriptive metadata.

2. Previous research in document analysis and understanding

A clear distinction must be drawn between a document's physical structure and its logical structure. While the physical structure is bound to the real world object, the logical structure is related to its content. Typical elements of the physical structure are pages or text blocks. Typical elements of the logical structure are chapters of a book or paragraphs of a chapter. While the physical structure does not offer any information to the reader to help him grasp the meaning of a document, the logical structure does. And while the physical structure may vary between different manifestations of a document, the logical structure will stay the same, if it is not altered deliberately. Although the logical structure is related to the content of a document it must not be confused with the semantic organisation of the text. The latter is not the subject of document analysis but addressed by discourse analysis.

The document analysis and understanding process can be divided into three different steps:

- (1) Layout analysis, which tries to extract and describe the physical structure of the document, i.e. it focuses on the text and graphic blocks of the pages, aims to identify them and figure out their hierarchical relations.
- (2) Document analysis and document understanding, which tries to classify and interpret the layout components of documents in order to derive their logical structure. This approach may include the detection of the reading order (Todoran *et al.*, 2001) as well as the identification of logical linkages (e.g. between a picture and the related caption).
- (3) Content interpretation, which tries to understand the result of optical character recognition, i.e. texts, tables, mathematical formulas, notes, etc. and may even apply object identification or scene understanding to interpret the content of pictures.

Whereas optical character recognition has been a research topic since the 1960s and has achieved a high degree of maturity, layout and document analysis, and especially document understanding, still present a number of unresolved problems and are attracting a lot of research activities. Most of the approaches engaged in the automatic capturing of document structures make intensive use

of knowledge-bases or models to guide the analysis process (Brugger, 1998, Logical structure of books and journals ch. 3.1.1).

A model is a generic description of a document class, i.e. a group of documents that share the same logical components or show similar layout structures. There are three main categories of models: rule-based models, grammar-based models, and models using statistical or probabilistic methods, though some hybrid models combine grammar- or rule-based approaches with statistical methods.

Rule-based models are well-known from the field of expert systems. Generally spoken, a rule is an ordered pair of situation or premiss and response or action. Rule-based models are used in the following approaches: Lee *et al.* (2000) apply a three-level rule-base in order to analyse the geometric structure of technical journals. Sainz Palmero and Dimitriadis (1999) represent the knowledge needed for the logical labelling in structured documents by a recurrent neuro-fuzzy system called "RfasArt". Altamura *et al.* (1998) encode the knowledge representation needed for the analysis of scientific articles into prolog rules. Prolog rules or predicates are further used by Niyogi and Srihari (1995) ("Cedar"-project) who try to identify the logical structure of newspaper images. Bayer (1993) and Bayer *et al.* (1994) extract the logical structure of business letters on the base of a semantic network. Rule-based models can further be found in the "OfficeMAID" system, which aims at capturing the logical structure of business letters (Dengel *et al.*, 1994), in "Graphein", a system for logical structure identification tested on library catalogue cards (Chenevoy and Belaïd, 1991; Chenevoy, 1993), in Kreich (1993), who scopes with the detection of the logical structure of letterheads, and Fisher *et al.* (1990), who are interested in capturing the geometrical structure and some logical elements like title, footer, page number, and paragraph.

Grammars that extend SGML with recognition rules are used by Klein and Fankhauser (1997) and Klein and Abecker (1999) for the generation of the logical structure. Akindele and Belaïd (1995) describe an approach which infers tree grammars for document structure modelling. The "OSCAR-II"-project makes use of a knowledge representation coded as an attributed, context-free grammar and demonstrated on legal texts and chapters of a scientific book (Hu, 1994; Ingold, 1991). In the "TWIG"-project Conway (1993) detects the logical structure using a page layout grammar and a logical-structure grammar; and Nagy *et al.* (1992) describe a document layout analysis for technical journals based on a publication-specific page grammar. Furthermore, there are grammar-based approaches on the base of Hidden Markov Models for the identification of page classes (Cesarini *et al.*, 1999; Frasconi *et al.*, 2001).

Statistical or probabilistic methods are applied in the "CIDRE" project, i.e. logical structure detection based on generalized n-grams (Brugger *et al.*, 1997). Furthermore, they can be found in the approach of Etemad *et al.* (1997) where the page segmentation is based on a neural network, and by Ittner and Baird (1993), who exploit statistical decision theory for the layout analysis.

Handcrafted models are often considered as expensive, rigid and limited to a few document types. Thus, most approaches focus on learning techniques which allow the automatic model generation from a set of training examples. As almost all research interest comes from the domain of computer science and research on artificial intelligence, the central research interest lies in the development of algorithms that best perform document analysis. To demonstrate the performance of those algorithms, small documents like business letters or only parts of larger documents like chapters or articles are sufficient. Furthermore, it should be noted that a considerable amount of studies limit themselves to one-page documents. Therefore, the generation of extensive models is not the prominent objective of these approaches and they do not offer any solution to scope with generic types of large, multi-page documents like books and journals.

But books and journals are the documents which make up the main part of digital libraries. And to meet the needs of digital libraries, the whole structure of those voluminous documents has to be taken into account. The approach presented in this paper differs from the studies cited above as our interest is not directed towards the development and evaluation of new algorithms, but towards the complete and systematic modelling of the logical structures which can be found in standard documents stored in libraries and archives. Therefore, we believe it to be a prerequisite to provide an extensive model for this structure, a grammar for books and journals.

Owing to fundamental changes in the printing technology at the beginning of the nineteenth century which led to a standardisation of printed objects, we decided to concentrate on books and journals published after 1820 – more information about the corpus used is given below. Documents published before 1820 are very heterogeneous and can hardly be described by a generic model. Our model is conceived as a basis for the further software development within the METAe-project. It is about to be implemented as a core module in the software package metadata engine.

3. Elements and relations of printed objects

The ability of a human reader to intuitively understand their logical functions while perceiving the layout elements of books and journals is based on the same conventions and traditions which determine the production process of printing objects. That is to say, e.g. a chapter heading is recognised as such because of the common cultural framework governing the appearance of layout elements. This is actually a semiotic system (Eco, 1973) – like road signs or language.

If the knowledge of these conventions is implemented in the recognition software, it may improve the results of the automatic capturing of document structures dramatically. So we analysed and described the document structures of books and journals. These descriptions serve to build up a knowledgebase for the software development. The semiotic system which we revealed looks very much like a simple natural language: documents are structured

hierarchically, they have elements belonging to classes, and they are governed by a syntax. Document structures may even have recursive elements. Hence, we based our description on a grammar model used in natural language analysis and capable of dealing with such structures: recursive transition networks and augmented transition networks.

Recursive transition networks were first used to parse the syntax of natural language phrases (Woods, 1970). A recursive transition network is a directed graph made up of nodes indicating the different states and arcs determining which elements have to be found in order to be able to effect the transitions, i.e. find a path through the network. An arc may be a terminal symbol or it may be labelled with the name of another state. When such an arc is encountered, the path leads to the sub-graph representing the state indicated. Thus, recursive transition networks prove to be very practical when modelling complex structures by moving repetitive substructures to small graphs of their own. Augmented transition networks are recursive transition networks enhanced by arbitrary conditions which “rule out meaningless analyses and take advantage of semantic information to guide the parsing” (Woods, 1970). Basically, this is very much the same as enhancing a context-free grammar to a context-sensitive grammar. The conditions are attached to each arc and must be met in order for the arc to be followed. Recursive as well as augmented transition networks are in fact very similar to final state machines, which are well known in the domain of computer science, and have been found to be by far equivalent and even more powerful than final state grammars (Pereira and Warren, 1980). Although it is rather difficult to construct a transition network which covers the whole range of phenomena included in the syntax of a natural language, transition networks have proven efficient for applications which do not deal with linguistic constructs too complex. Therefore, we found augmented transition networks to be the best choice to describe the logical structure of books and journals. They are quite easy to build, rather simple to communicate to computer scientists, and straightforward to convert into a programming language.

The production of printed material is a process based on a set of principles and conventions established by editors, publishing houses and the printing and binding industry over the years. Because of these “rules”, humans are able to recognize a collection of leaves as a book, to distinguish it from a journal and to identify chapters and subchapters without even reading and understanding the content.

Printed documents consist of a set of elements and the relationship between them. Basically, there are two types of rules defining the relationship between the elements of printed objects:

- (1) Rules which define the geometric and typographical relationship between elements, e.g. the font size of the chapter title has to be larger than the “default” font size used for the running text; it has also to be larger than the font size used for the titles of subchapters, if there are any subchapters.

Most of the handcrafted models described in the literature encode thresholds or absolute values of presentational and geometric characteristics in order to allow the association of logical labels to physical blocks (Palowitch and Stewart, 1995). Such a model does not embrace a document class in the sense of “books”, “journals” or “letters”, etc., but describes a specific family of publications like the volumes of one series or the issues of a concrete journal. Or it can be used only for the labelling of a few elements such as headers or footnotes. If a model should be applicable to a broad class of documents like “prose books”, “scientific books”, various journals, etc., abstract relations instead of document specific formatting rules have to be specified in the model.

- (2) Syntactical rules, i.e. the sequence in which elements of printed documents can appear, e.g. a preface never follows the chapters of a book, a table of contents does never appear in front of a title page or as part of the running text and so on.

A good model which encodes all this information can perform two tasks within a document analysis system: first, it can help to identify logical components. Owing to the fact that normally there is no one-to-one-mapping between physical and logical structure, reverse encoding processes have to resolve ambiguities in the logical interpretation of physical components. To clear these ambiguities, taking into account the whole sequence of elements may help. Second, it happens that errors occur during the printing as well as during the scanning process. A document understanding system based on an extensive model can detect such errors as an inadmissible deviation from allowed, well-formed structures.

4. The application of augmented transition networks for logical structure modelling

This chapter describes how the structure found in monographs can be represented and modelled using recursive transition networks. The examples are taken from the grammar for monographs we have developed so far. It should be pointed out that this is not the whole grammar and the examples given are slightly simplified to ease comprehension.

For the generation of the grammar of books and journals a top-down approach was applied. Starting at the top level of a volume of a typical prose book, a structure can be found which consists of three components (see Figure 1).

The first component contains pages with a heterogeneous layout structure; the second component consists of a set of pages which show a balanced, repetitive structure; the last set of pages makes up the third component which can be identified because the layout structure obviously differs from the one of the second component. Concerning the logical structure, the first part, “front”, contains different prefatory matters like title page, preface, and table of contents; the middle part, “main”, contains the main text body of the work, i.e.

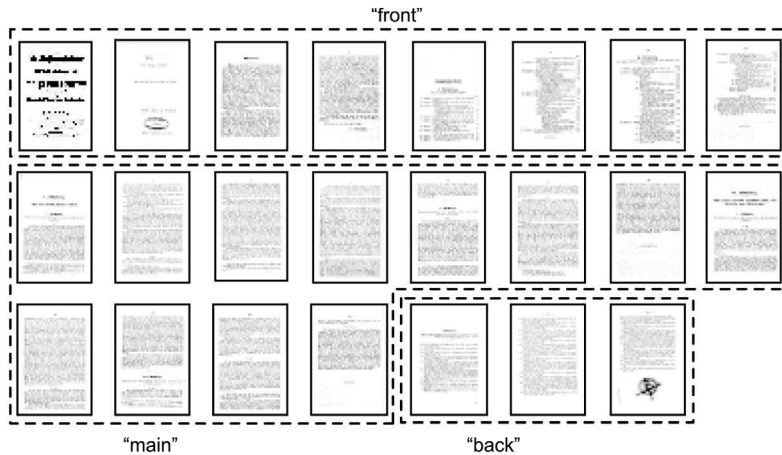


Figure 1.
The three components
of a typical prose book (the
pages shown are taken
from different books and
arranged for
demonstration purposes)

the chapters and subchapters of the book, whereas the third part, “back”, contains elements like bibliography lists, appendices, and registers. A rather similar grading is used by the “Text encoding initiative”: TEI distinguishes between “front matter”, “body”, and “back matter” as a default high level structure for all documents (Sperberg-McQueen and Burnard, 1999).

Recursive transition networks are similar to finite state transition graphs as they consist of a set of nodes which are connected by labelled, directed graphs. Transitions from one node to the next are effected if the input corresponds to the label of the concerning arc. The nodes correspond to states in a finite state machine, but in contrast to finite state transition graphs, recursive transition networks allow not only terminal symbols but also non-terminal symbols as labels of the arcs. Non-terminal symbols are state names which represent complex structures that are described by another (or the same) transition graph (Woods, 1970). Thus, applying recursive transition graphs, complex hierarchies can be described by packing substructures into non-terminal symbols and therefore allowing for abstractions of the different levels of a hierarchical structure. Following this model, the basic document structure described above can be represented as follows (see Figure 2).

While the node “vol”, “volume”, is the start state, the nodes “Q1” and “Q2” represent intermediate states and the node “Q3/i” indicates the final state, i.e. the possible end of the graph. The arcs “front”, “main”, and “back” represent the three different components of a volume. The complex substructures of these three components are described in a next step by an own transition graph for each of them. In order to represent the book’s component called “front” properly, it is important to focus on the logical structure of the pages and their content (see Figure 3).

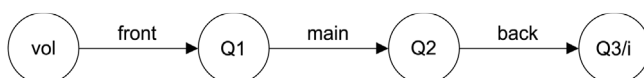


Figure 2.
A graph for the basic
document structure

First in line is a title page, followed by a copyright statement and a print statement on the next page. The rest of the “front” component consists of a two page preface and a table of contents filling four pages. This structure can be represented by the following graph (see Figure 4).

While the graph for a volume shown in Figure 2 may be almost called generic, obviously the graph in Figure 3 will just cover a subset of volumes on a bookshelf, some will have no preface, some lack a table of contents, others will show a blank page after the title page and some will have a table of contents followed by a preface, just to mention some possible variations.

Actually, the analysis of prose books, which were published between 1820 and 1950, revealed that only the title page is a mandatory element, whereas all the other elements are not. On the other hand, the elements in the “front” component tend to be arranged in a wide variety of orders. Moreover, there may be additional elements like advertisements or photographs before the title page. Therefore, the generic logical structure of the “front” component of books can be described as follows (see Figure 5).

The arc “tip”, “title page”, has been placed in the middle and two new non-terminal symbols have been introduced in this graph: “beft”, which stands for “before title page” and indicates those parts of the whole title matter which precede the title page, and “aft”, which stands for “after title page”, i.e. the parts of the title matter which follow the title page. It should be noted that the graph shows an unlabelled arc leading from “front” to “Q4”. That means that the arc “beft” may be bypassed altogether: a book may simply start with its title page. And that “Q5/i” is a final state, so that there may be no “aft” at all. The possibility to introduce abstract non-terminal symbols like “beft” and “aft” and to pack whole substructures into them is one of the powerful features of recursive transition network. If one had to use terminal symbols only at this point of the analysis as in a finite state grammar

Figure 3.
The elements of the “front” component (pages arranged for demonstration purposes)

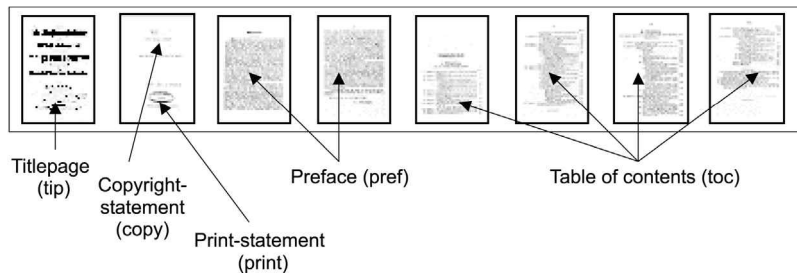


Figure 4.
A graph of the elements from Figure 3

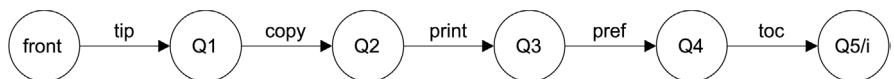
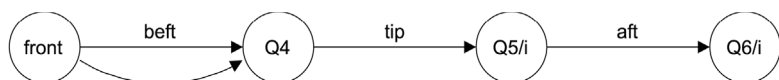


Figure 5.
A generic graph of the “front” component



the graph would already look very complicated and confusing because of the high degree of variation in the front matter of prose books.

We will now go further into detail focusing on the “aft” element. The graph representing “after title page” has to cover the existing variation by offering many different paths (see Figure 6).

The graph starts with three possibilities: there may be a blank page, a copyright statement or an empty arc. The copyright statement may be followed by a print statement or another empty arc and the first empty arc may be followed by a print statement or an empty arc. After the state “Q8” there is either a table of contents followed by a preface or a preface followed by a table of contents. But there may be also a preface only or a table of contents only.

It should be noticed that there is a path through the unlabelled arcs of the graph, which allows the “aft” component to be entirely empty. But this possibility is already covered by the final state “Q5/i” in Figure 5. So the state “aft” is only entered if something follows the title page and therefore the empty path mentioned will actually never be used. Despite the many variations the “aft” component covers, the succession of the different elements is still subject to a certain order. Thus, if there is for example a copyright statement, it always appears before the preface or the table of contents.

While the “front” component of a prose book can be characterised as a succession of more or less unique elements in varying orders, the “main” component is typically made of balanced structures and recurring elements. The “main” component of prose books consists either of a text body (“body”) or of a set of chapters (“caps”) (see Figure 7a). In other words, a prose book is either structured into chapters or it is not. A set of chapters (“caps”) may have any number of chapters – this is indicated by the looping arc at state “Q13/i” – but is made up of at least two chapters (see Figure 7b). A chapter (“cap”) has a title (“tit”), possibly a subtitle (“tit.sub”) and a text body – or again a set of chapters, i.e. subchapters (see Figure 7c). Finally, a text body (“body”) is a series of paragraphs (“p”) that can as well include graphics, images, tables and formulas (“non-p” elements) (see Figure 7d and 7e).

Although the model for the “main” component shown in Figure 7 is a bit simplified, because some elements like introductions, which may appear at the start of a set of chapters, mottos, which can be encountered between the title of chapter and its text body, or blank pages, which may appear at the end of a chapter to make the following chapter start on a right hand page, have been stripped from the model for the sake of comprehensibility, it can be seen that the model is able to cover the variety of different structures which can be found in prose books.

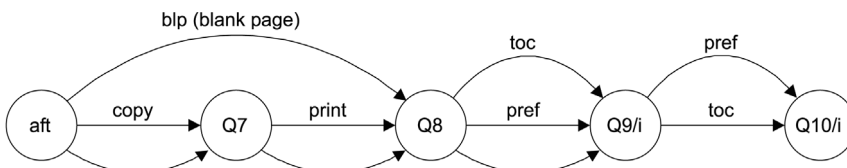


Figure 6.
A generic graph of the
“after title page”
component

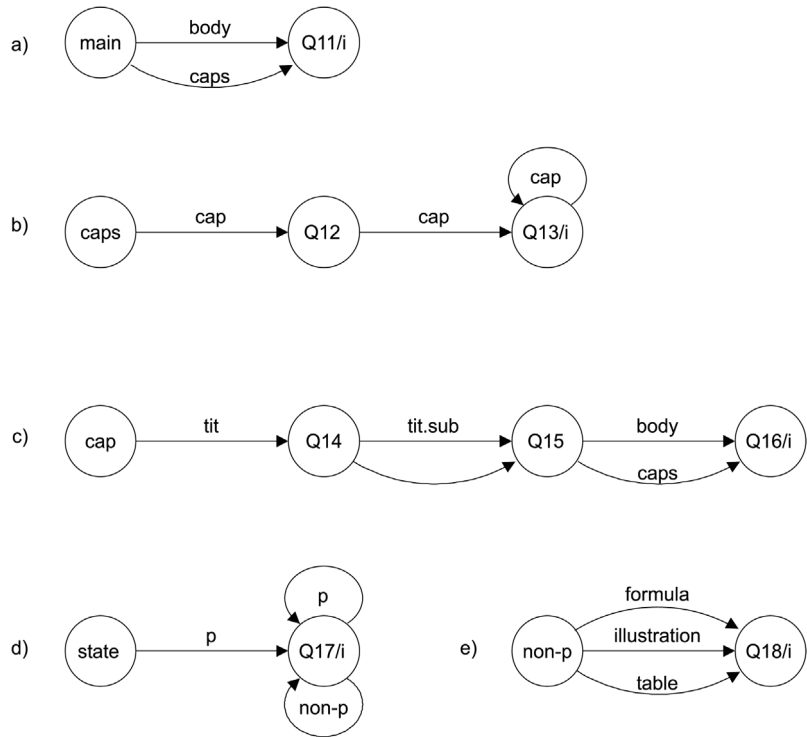


Figure 7.
A set of graphs of the
“main” component of a
prose book

In our grammar for books and journals, one group of elements such as titles, subtitles, paragraphs, formulae, etc. are terminal symbols or end elements which correspond to different kinds of text blocks. Another group of end elements such as title pages or dedications do not correspond to text blocks but to whole pages. Most of these end elements can be further decomposed into lines, words, and characters, and can be described according to their characteristics, e.g. a text block type corresponding to a chapter title: has a font size which is larger than the default font size of the analysed book; may be aligned in the centre; and consists of less text lines than the following text block, unless it is a subtitle or a chapter title of a subchapter.

The last condition may need some further explanation. As already mentioned the “main” component of a prose book reveals a balanced structure. Therefore, a chapter as a logical substructure never appears alone: there are at least two. The same is true for subchapters as well as for subchapters of subchapters. The problem implied in the last condition characterising a chapter title is, how to figure out whether a text block following a chapter title is simply a sub-title of that chapter or the title of a subchapter (see Figure 8).

To resolve the ambiguity of subtitle versus title of a subchapter the recursive transition network must be augmented by a condition defining that titles of chapters belonging to the same level always share the same characteristics like font size, type, and face. Or to put it the other way round:

chapter titles of different levels show differences in font size or style or face. Therefore the titles of the first, second or third chapter never differ in font size or style or face. This is also true for any chapter title of a subchapter belonging to the same sub-level. Thus, the structure of the example shown in Figure 8 is recognised either as consisting of a subtitle and a following chapter of the same level or as a succession of subchapters, depending on the characteristics of the next chapter or subchapter title encountered (see Figure 9).

We will stop at this point to describe and explain our approach how to model the logical structure of books and journals using augmented transition networks. It must be admitted that the grammar we have developed so far is more complicated than the examples given, merely because it covers a lot more layout elements which can be found in books and journals, especially in the front and back matter. It should be mentioned as well that journals may raise some special problems due to structural elements which can only be successfully dealt with when taking into account multiple issues and volumes.

Our research is based on a corpus of books and journals in English and German (see <http://meta-e.uibk.ac.at/docs/grammar-corpus.html>). For the generation of the model for monographs, we have analysed about 70 volumes published between 1820 and 1940. Approximately 30 books were taken from the microfiche edition of the library of Corvey Castle, one of the largest private book collections of literary works from the nineteenth century (Edition Corvey, 1989-1996). A total of 20 volumes of specialist books from the domain of law were provided by the digital library of the Max-Planck-Institut für europäische

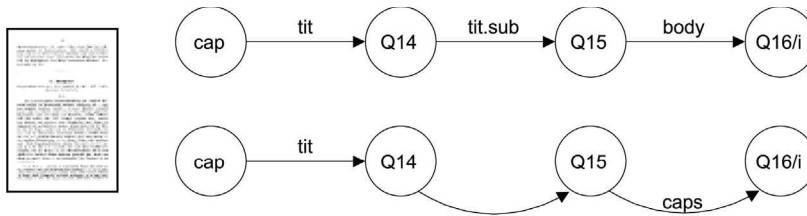


Figure 8.
A single page with an
ambiguous structure
and the two graphs
representing the two
possible structures

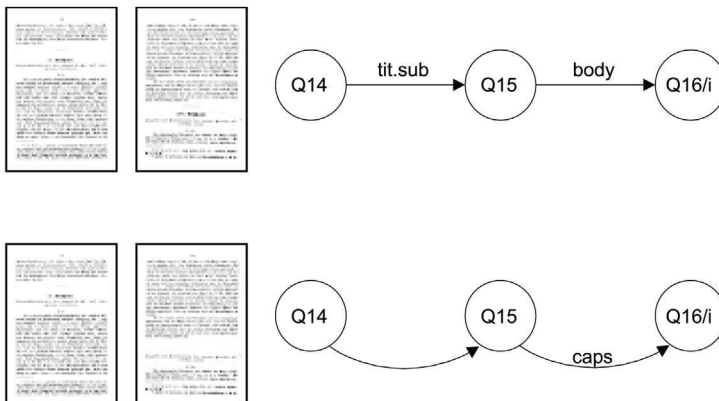


Figure 9.
The ambiguous
structures shown in
Figure 8 are resolved by
taking into account the
follow-up pages

Rechtsgeschichte (<http://www.mpier.uni-frankfurt.de/>). These books are especially interesting because of the fine-grained chapter structure they reveal. Another 20 volumes originate from the “Making of America” project of Cornell University and the University of Michigan (<http://cdl.library.cornell.edu/moa/> and <http://moa.umdl.umich.edu/>). We took care of covering the whole period by selecting volumes according to their publishing date. This set of books was used to construct and refine the model. The model was then verified against a set of 60 books from the same sources.

From about 40 journals more than 70 volumes and 100 issues were used to generate the model for journals and serials. Scientific as well as popular journals were among them. The journals were taken from the University Library of Innsbruck and from the “Making of America” project. For verification purposes ten journals were used, once again from the University Library of Innsbruck and from the digital library “Gallica” of the Bibliothèque nationale de France (<http://gallica.bnf.fr/>).

During the generation and verification of the model statistical data have been collected to determine the probability of a certain arc in the graphs to be passed successfully. Figure 10 shows the model for the element “vol”, “volume”.

Figure 10 is taken from the original grammar. It only differs from the example given in Figure 2 as the state following the arc “main” is a final state. All the books we have analysed had a “front” and a “main” element, but some lacked a “back”. Figure 11 shows that most of the books were structured in chapters and only a few had only a “body”. Those books were fiction – specialist books and scientific literature from our corpus were always made up of chapters. Finally, in Figure 12 the model for “cap”, “chapter”, is shown. It is

Figure 10.
Probabilities in the graph describing the element “volume” of a book

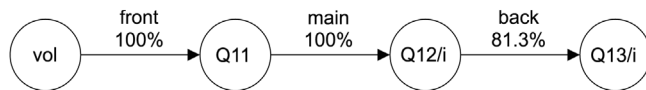


Figure 11.
Probabilities in the graph describing the element “main”

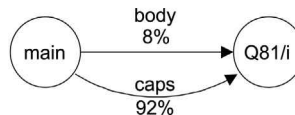
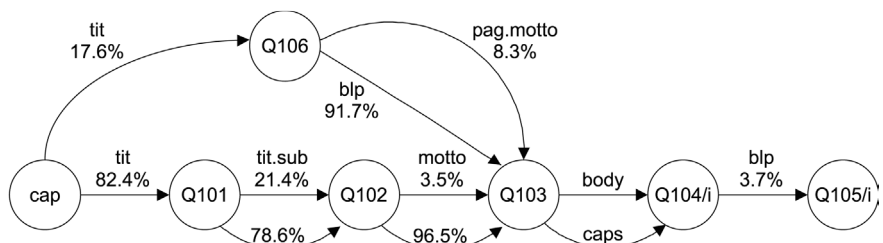


Figure 12.
Probabilities in the graph describing the element “chapter”



more complicated than the simplified version in Figure 7. Additional elements are “motto”, i.e. a quotation which follows a title or a subtitle, and “pag.motto”, “page motto”, where a motto appears alone on one page.

No probabilities are given for the arcs “body” and “main” leading from “Q103” to “Q104”, because “caps”, “chapters”, leads to a state, which in turn has “cap”, “chapter”, as an arc to describe the recursive structure of chapters with subchapters. Thus, figures on probability would not make much sense unless the current level of the subchapter is known.

5. Conclusion

Starting with the idea that the logical structure of books and journals is part of a semiotic system, we modelled a grammar of books and journals using augmented transition networks. This grammar describes well-formed logical structures of books and journals published in German- and English-speaking countries between 1820 and 1950. Our attempt to describe the rules governing the layout of books and journals in terms of syntactical and abstract layout relations can help automated document understanding. The correct identification and detailed representation of the logical structure of digitised books and journals is a primary need in the world of digital libraries.

The grammar has proven to be fit to deal with most books and journals from the period mentioned, but it has to be admitted that structural exceptions may cause the model to fail. The grammar of books and journals is currently being implemented in the automatic document analysis of the metadata engine.

Other types of documents do not lie within the scope of the METAe project, but our approach could be easily applied to such types as anthologies, dramatic poetry, or dictionaries.

References

- Akindele, O.T. and Belaïd, A. (1995), “Construction of generic models of document structures using inference of tree grammars”, in *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR'95), Montréal, 14-16 August*, IEEE Computer Society Press, Los Alamitos, CA, pp. 206-9.
- Altamura, O., Esposito, F. and Malerba, D. (1999), “WISDOM++: an interactive and adaptive document analysis system”, in *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR'99), Bangalore, 20-22 September*, IEEE Computer Society Press, Los Alamitos, CA, pp. 333-69.
- Bayer, T. (1993), “Understanding structured documents by a model based document analysis system”, in *Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR'93), Tsukuba Science City, Japan, 20-22 October*, IEEE Computer Society Press, Los Alamitos, CA, pp. 565-8.
- Bayer, T., Bohnacher, U. and Mogg-Schneider, H. (1994), “InfoPortLab – an experimental document analysis system”, in *Proceedings of the IAPR-Workshop on Document Analysis Systems (DAS 94), Kaiserslautern, 18-20 October*, DKFI, Kaiserslautern, pp. 297-312.
- Brugger, R. (1998), “Eine statistische Methode zur Erkennung von Dokumentstrukturen”, PhD thesis, University of Fribourg, available at: www-iiuf.unifr.ch/~brugger/papers/da/da.html.html (accessed 28 June 2001).

- Brugger, R., Zramdini, A. and Ingold, R. (1997), "Modeling documents for structure recognition using generalized N-Grams", in *Proceedings of the Fourth International Conference on Document Analysis and Recognition (ICDAR'97), Ulm, Germany, 18-20 August*, IEEE Computer Society Press, Los Alamitos, CA, pp. 56-60, available at: ftp-iiuf.unifr.ch/pub/giraf/papers/icdar97brugger.pdf (accessed 21 June 2001).
- Cesarini, F., Franceconi, E., Gori, M. and Soda, G.A. (1999), "Two level knowledge approach for understanding documents of a multi-class domain", in *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR'99), Bangalore, 20-22 September*, IEEE Computer Society Press, Los Alamitos, CA, pp. 135-8.
- Chenevoy, Y. (1993), "Reconnaissance structurelle de documents imprimées: études et réalisations", PhD thesis, CRIN-Nancy.
- Chenevoy, Y. and Belaïd, A. (1991), "Hypothesis management for structured document recognition", *ICDAR 91. First International Conference on Document Analysis and Recognition, Saint-Malo, September*, AFCET, Paris, pp. 121-9.
- Conway, A. (1993), "Page grammars and page parsing – a syntactic approach to document layout recognition", in *Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR'93), Tsukuba Science City, Japan, 20-22 October*, IEEE Computer Society Press, Los Alamitos, CA, pp. 761-4.
- Dengel, A., Bleisinger, R., Hoch, R., Hönes, F. and Malberg, M. (1994), "OfficeMAID – a system for office mail analysis, interpretation and delivery", in *Proceedings of the IAPR-Workshop on Document Analysis Systems (DAS 94), Kaiserslautern, 18-20 October*, DKFI, Kaiserslautern, pp. 253-75.
- Dori, D., Doermann, D., Shin, C., Haralick, R., Phillips, I., Buchman, M. and David, R. (2000), "The representation of document structure: a generic object-process analysis", in Bunke, H. and Wang, P.S.P. (Eds), *Handbook of Character Recognition and Document Image Analysis*, World Scientific Publishing Company, Singapore, pp. 421-56.
- Eco, U. (1973), *Il segno*, ISEDI, Istituto Editoriale Internazionale, Milano.
- Edition Corvey. Literatur des 18. und 19. Jahrhunderts aus der Fürstlichen Bibliothek Corvey (1989-1996)*, Belsler, Stuttgart, Zürich.
- Etemad, K., Doermann, D. and Chellappa, R. (1997), "Multiscale segmentation of unstructured document pages using soft decision integration", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19 No. 1, pp. 92-6.
- Fisher, J.L., Hinds, S.C. and D'Amato, D.P. (1990), "A rule-based system for document image-segmentation", in *Proceedings of the 10th International Conference Pattern Recognition (ICPR), Vol. 1, Atlantic City, NJ, 16-21 June*, IEEE Computer Society Press, Los Alamitos, CA, pp. 567-72.
- Frasconi, P., Soda, G. and Vullo, A. (2001), "Text categorization for multi-page documents: a hybrid naive Bayes HMM approach", in *First ACM-IEEE Joint Conference on Digital Libraries, JDCL 2001, Roanoke, VA, 24-28 June, Proceedings*, ACM, pp. 11-20.
- Hu, T. (1994), "New methods for robust and efficient recognition of the logical structures in documents", PhD thesis, University of Fribourg, Fribourg.
- Ingold, R. (1991), "A document description language to drive document analysis", in *ICDAR 91. First International Conference on Document Analysis and Recognition, Saint-Malo, September*, AFCET, Paris, pp. 294-301.
- Ittner, D.J. and Baird, H.S. (1993), "Language-free layout analysis", in *Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR'93), Tsukuba Science City, Japan, 20-22 October*, IEEE Computer Society Press, Los Alamitos, CA, pp. 336-40.
- Klein, B. and Abecker, A. (1999), "Distributed knowledge-based parsing for document analysis and understanding", in *Proceedings of the IEEE Forum on Research and Technology*

- Advances in Digital Libraries (ADL'99)*, Baltimore, MD, 9-12 May, IEEE Computer Society Press, Los Alamitos, CA, pp. 6-15, available at: www.gmd.de/publications/report/0048/Text.pdf (accessed 23 June 2001).
- Klein, B. and Fankhauser, P. (1997), "Error tolerant document structure analysis", *International Journal on Digital Libraries*, Vol. 1 No. 4, pp. 344-57.
- Kreich, J. (1993), "Robust recognition of documents", in *Proceedings of the Second International Conference on Document Analysis and Recognition (ICDAR'93)*, Tsukuba Science City, Japan, 20-22 October, IEEE Computer Society Press, Los Alamitos, CA, pp. 444-7.
- Lee, K.-H., Choy, Y.-C. and Cho, S.-B. (2000), "Geometric structure analysis of document images: a knowledge-based approach", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22 No. 11, pp. 1224-40.
- Library of Congress, Network Development and MARC Standards Office (2001), *MARC21 Concise Format for Bibliographic Data*, available at: www.loc.gov/marc/bibliographic/ecbdhome.html (accessed 24 September 2001).
- Nagy, G., Seth, S. and Viswanathan, M. (1992), "A prototype document image analysis system for technical journals", *Computer*, Vol. 25 No. 7, pp. 10-22.
- Niyogi, D. and Srihari, S.N. (1995), "Knowledge-based derivation of document logical structure", in *Proceedings of the Third International Conference on Document Analysis and Recognition (ICDAR'95)*, Montréal, 14-16 August, IEEE Computer Society Press, Los Alamitos, CA, pp. 472-5.
- Palowitch, C. and Stewart, D. (1995), "Automating the structural markup process in the conversion of print documents to electronic texts", in *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries. Digital Libraries '95*, Austin, TX, 11-13 June, available at: www.csdl.tamu.edu/DL95/papers/palowitc/palowitc.html (accessed 21 June 2001).
- Pereira, F. and Warren D. (1980), "Definite clause grammar for language analysis – a survey of the formalism and a comparison with augmented transition networks", *Artificial Intelligence*, Vol. 13 No. 3, pp. 231-78.
- Sainz Palmero, G.I. and Dimitriadis, Y.A. (1999), "Structured document labeling and rule extraction using a new recurrent fuzzy-neural system", in *Proceedings of the Fifth International Conference on Document Analysis and Recognition (ICDAR'99)*, Bangalore, 20-22 September, IEEE Computer Society Press, Los Alamitos, CA, pp. 181-4.
- Sperberg-McQueen, C.M. and Burnard, L. (Eds) (1999), *The Association for Literary and Linguistic Computing (ALLC) Guidelines for Electronic Text Encoding and Interchange. TEI P3 Text Encoding Initiative*, Chicago, Oxford, revised reprint, Oxford, May, available at: www.tei-c.org/Guidelines/index.htm (accessed 9 September 2001).
- Todoran, L., Aiello, M., Monz, C. and Worring, M. (2001), "Logical structure detection for heterogeneous document classes", in *Proceedings of the Document Recognition and Retrieval VIII, SPIE, San Jose, CA, 24-25 January, Proceedings of SPIE*, Vol. 4307, SPIE Press, Bellingham, Washington, DC, pp. 99-110, available at: <http://carol.wins.uva.nl/~todoran/spie01.pdf> (accessed 20 June 2001).
- Woods, W.A. (1970), "Transition network grammars for natural language analysis", *Communications of the ACM*, Vol. 13 No. 10, pp. 591-606.