Published in: Tapio Pahikkala, Sampo Pyysalo, Filip Ginter und Tapio Salakoski (eds.): Advance in natural language processing. 5th International Conference on NLP FinTAL 2006 Turku, Finnland, August 23-25, Bd. 2006. Berlin, Heidelberg: Springer-Verlag, pp. 245–256.

Discourse Segmentation of German Written Texts

Harald Lüngen, Csilla Puskás,
Maja Bärenfänger, Mirco Hilbert, and Henning Lobin

Justus-Liebig-Universität GieSSen
FB 05 - Applied and Computational Linguistics Otto-Behaghel-Str. 10 D
{luengen,puskas}@uni-giessen.de

Abstract. Discourse segmentation is the division of a text into minimal discourse segments, which form the leaves in the trees that are used to represent discourse structures. A definition of elementary discourse segments in German is provided by adapting widely used segmentation principles for English minimal units, while considering punctuation, morphology, sytax, and aspects of the logical document structure of a complex text type, namely scientific articles. The algorithm and implementation of a discourse segmenter based on these principles is presented, as well an evaluation of test runs.

1 Introduction

In one subproject of the DFG research group *Text-technological modelling of information*, a discourse parser for a complex text type, i.e. scientific articles, is being developed. Discourse parsing according to Rhetorical Structure Theory (RST, [1]) deals with automatically assigning a text a hierarchical (tree) structure marking *discourse segments* (text spans) and functional-argumentative relations such as BACKGROUND, CONCESSION, and CONTRAST between them. Most discourse relations are binary, and one of the arguments has the status of being a *nucleus* (the more salient piece of information according to the author's intentions) while the other one is the *satellite* (containing supporting information that can potentially be omitted). Discourse segments can be complex or elementary, the latter being the minimal propositional units at the leaves of a discourse tree. The segmentation of an input document into elementary discourse segments is the first step in the discourse parsing process, cf. [2]. In our parsing architecture, this step is performed by a preprocessing component, a discourse segmenter. This paper introduces the discourse segmenter, i.e. is about how the minimal units of discourse should be defined for the relevant language (German), and how they are automatically recognised in text documents.

2 Requirements

What is an elementary discourse segment? In many systems based on RST, the definition is that of an *elementary discourse unit* (EDU) which seems to have been introduced by Marcu, e.g. "*[e]dus* are defined functionally as clauses or clause-like units that are unequivocally the NUCLEUS or SATELLITE of a rhetorical relation that holds between two adjacent spans of text" [3]. This definition includes types of main and subordinate clauses, but also certain phrase types. The phrase *in spite of the bad weather conditions*, for example, is an EDU because the preposition *in spite of* introduces a CONCESSION relation between two propositions just like the subordinating conjunction *although*.

This definition of EDUs has been operationalised in terms of a set of criteria relating to English punctuation and grammar and has been applied to the segmentation and manual RST annotation of a large corpus of newspaper articles by Carlson and Marcu [4]. EDUs have subsequently also been used in the discourse parsers proposed in [5], [6], and [7].

We work with a different application scenario, text type, and language than previous approaches to automated discourse segmentation such as [2], [5], and [8]. For the development of our discourse parser and segmenter we use a corpus of 47 German scientific articles in the discipline of linguistics from the journal *Linguistik Online*[1]. The discourse parser is to be used in a hypertext system that supports students in the explorative and selective reading of scientific articles, based on highlighting text structure and on providing automatically generated link lists to different structural elements that contain rhetorically salient parts of the text. Articles chosen by the students themselves shall be automatically analysed by the discourse parser and annotated with an RST structure. Thus, the definition of a minimal unit of discourse is guided by the question whether it will be part of a discourse relation where the nucleus is semantically independent enough so that the satellite can be realised as a separate hypertext unit.

Although we take the methodology to define EDUs as introduced in [4] as a model, we have chosen not to adopt the term *EDU* itself since in several respects we deviate from the definition of English EDUs. Our criteria for segmenting a German text into elementary discourse *segments* (EDSs) refer to the following levels of information: a) logical document structure, b) punctuation, and c) morphology and syntax, including lexical discourse markers.

2.1 EDSs Induced by Logical Document Structure

The logical structure of the documents in our corpus, i.e. their hierarchical division in sections, titles, paragraphs etc. is annotated according to the so-called DOC annotation scheme which was developed in co-operation with a partner project. It comprises about 60 elements from the DocBook DTD [9] plus 14 additional elements for scientific articles such as `<caption>` as well as XHTML elements, integrated in one XML schema using namespace technology. Our segmenter expects a text plus its DOC annotation as input.[2]

The textual content of certain DOC elements shall directly correspond to an EDS (Table 1 shows some). Some of them, e.g. `<blockquote>`, `<blockemphasis>`,

[1] `http://www.linguistik-online.de/`

[2] In later stages of the project, a tool to convert other document formats to DOC will be developed.

Table 1. Elementary discourse segments according to the DOC annotation layer

DOC Element	Semantics
`<title>`	The title of the whole article, or of sections
`<programlisting>`	Code
`<bibliomixed>`	An entry in the bibliography
`<glossterm>`	A term in a definition list
`<ackno>`	Acknowledgments
`<blockquote>`	A quotation that is set apart from the running text
`<blockemphasis>`	Text that is set apart from the running text
`<footnote>`	Text in a footnote
`<log:mediaobject>`	(Empty elements containing) figures, i.e. images or diagrams
`<log:caption>`	The caption of a table, or a figure
`<log:tgroup>`	The body of a table

and `<footnote>`, may contain text that could potentially be further segmented by the punctuational and grammatical criteria. In view of the explorative reading scenario sketched above, however, we want them to always correspond to EDSs. We think that this specification makes sense for other languages and application scenarios, too.

2.2 EDSs Induced by Grammar and Punctuation

In the following, the main types of EDSs according to grammatical and punctuational criteria are formulated independently of the representation of grammatical analysis produced by the syntactic parser that we employ in the segmenter.

1. *Main clauses:* all simplex main clauses form an EDS. Main clauses are separated from other segments by punctuation, and/or coordinating conjunctions. Example: *[Die schwedische Kolonisation dauerte über sog. 600 Jahre.] [und zur selben Zeit sind Handwerker und Kaufleute aus dem ganzen Ostseeraum nach Finnland gezogen.]*[3]
2. *Modal subclauses:* modal subordinate clauses (marked by a modal subordinating conjunction), including modal infinitival constructions (marked by *ohne zu* or *um zu*). Example: *[Es ist auch üblich, dass man zu Hause sowohl Finnisch als auch Schwedisch redet.] [da Ehen oft über die Sprachgrenze hinweg geschlossen werden.]*
3. *Coordinated clauses:* Only in coordinations of the categories S, $\bar{S}$, and VP are the coordinated parts EDSs. Thus the subject or a subject plus a grammatical auxiliary may be elliptified in an EDS; this is parallel to the definitions for English in [4]. Example: *[Das Land gehörte 600 Jahre lang zu dem schwedischen Reich] [und wurde im Jahre 1809 ein autonomes Grossherzogtum unter dem russischen Zaren.]* We additionally include cases where units consisting of Subject + Complement are coordinated. i.e. in these cases a verb may

[3] Unless otherwise stated, the examples given are taken from [10].

be elliptified in a resulting EDS. Example: *[Ein Drittel von ihnen wohnt in Ostrobothnia (...) an der Westküste des Landes,] [die anderen in Südfinnland und auf den Åland-Inseln.]*

4. *Embedded segments:* embedded segments are segments marked by punctuation (brackets, dashes, or commas) which disrupt other EDSs, and which themselves are EDSs. Exception: Brackets that contain only figures (e.g. *(1999)*) are not segmented. Example: *[Problematisch ist jedoch, dass in Finnland mehrere samische Sprachen [(Nordsamisch, Skoltsamisch und Enaresamisch)] gesprochen werden.]* Note: Embedded segments are not internally segmented.
5. *Quotations* that are delimited by quotation marks and are introduced by reporting verbs. Note: Quotations shall *not* be internally segmented. Example: *["Ein Kind hatte im Spielzeugladen eine Wunschliste hinterlegt. Ich war froh, dass noch ein Aufziehauto für 3,50 Euro zu vergeben war".] [berichtet Rolfs.]*[4] Exception: Quotations that are built into running text without attributional constructions are not separated at all, i.e. in these cases the quotation marks are simply ignored, and segment boundaries are assigned as usual.
6. *Clausal complements of reporting verbs* such as (*meinen, sagen, feststellen*) in connection with a citation or quotation (inducing the rhetorical relation of ATTRIBUTION, cf. [4]). Example: *[Allardt (2000:8) meint], [dass die Einstellung während der letzten Jahrzehnten sich positiv entwickelt hat.]*
7. *Clausal complements and relative clauses preceded by adverbials:* Clausal complements of verbs or nouns, or relative clauses that are preceded by a *discourse marking adverbial* such as *nämlich, namentlich, besonders, insbesondere, d.h., vozugsweise.* Example: *[Das Ergebnis stimmt mit einer ziemlich allgemein verbreiteten Auffassung überein,] [nämlich dass das Sprachprogramm der finnischen Schulen allzu schmal ist.]*
8. *Prepositional phrases of attribution*, i.e. one of the prepositions *nach, laut, gemäSS* + a named entity, or a pronoun referring to a named entity, in connection with a citation or quotation. Example: *[Nach Allardt] (...)][hängt dieses damit zusammen, dass die Finnischsprachigen daran gewöhnt sind, mit den Finnlandschweden Finnisch zu sprechen.]*
9. *Appositives.* Appositives are NPs that can be used postnominally as supplements to NPs, with which they mostly agree in number and case. Appositions sometimes start with a discourse-marking adverbial, too.[5] Example: *[Dazu hat das Land seit 1995 drei offizielle Minderheitssprachen,] [Samisch, Romani und Gebärdensprache.]*
   Appositives frequently occur as embedded segments.
10. *PPs that are separated by a comma.* These are similar to the "discourse-salient phrases" in [4], only we do not inventorise a list of strong discourse cues but define every adverbial PP that is separated from the rest of the clause by a comma to be an EDS. Since PPs are not usually separated by

[4] This example is taken from the newspaper article [11].

[5] Cf. *grammis - das grammatische Informationssystem des Instituts für deutsche Sprache*, `http://hypermedia.ids-mannheim.de/index.html`.

commas, the use of a comma in such cases can be considered a strong discourse cue employed by the author. Example: *[Gleichzeitig entstand aber eine Gegenbewegung.] [für das Bewahren der schwedischen Sprache in Finnland.]*

With EDSs defined in the above fashion, note that the following clause types are *not* EDSs:

- Clausal subjects and clausal complements of verbs and nouns, (with the exception of the attributional complements described under 6. and 7.)..
- Restricting relative clauses. Following [1], we do not regard restricting relative clauses as EDSs because unlike other satellites, they contribute to the semantic interpretation of their head noun, and in that way can never be omitted. This treatment is in contrast to [4].
- Conditional clauses: *wenn..., dann..., je..., desto* etc. Unlike in other so-called mononuclear constructions, the nucleus in constructions related by the CONDITION relation seems not comprehensible without the satellite, thus we regard them as together forming one EDS. This treatment is also in contrast to [4].
- Proportional clauses, i.e. clauses combined by comparative connectives such as *mehr... als, weniger... als, so (ADJ)... wie.* Unlike in [4], such a construction is not split into separate EDSs, because neither of its parts seems more salient than the other in terms of nuclearity.

A consequence of denying certain clause types the status of EDS (most notably complemental clauses and restricting relative clauses) is that potential EDSs that are subordinate to such non-segmentable clauses cannot be EDSs, either. Consider a sentence from [10], the clause structure of which is indicated by labelled bracketing:

> $_S$*[Es ist aber symptomatisch,* $_S$*[dass alle Streitigkeiten sofort vergessen wurden,* $_S$*[als eine gemeinsame Gefahr von AuSSen drohte,* $_{NP}$*[d.h. Russifizierung in der Periode 1890-1917 und zwei Kriege in den Jahren 1939-1945]]]].*

According to criterion 2 above, an EDS boundary could potentially be introduced between *wurden,* and *als* because it is the beginning of a modal subclause. At the same time, no boundary is to be inserted at the previous subordination, i.e. between *symptomatisch,* and *dass*, because an ordinary sentential subject is starting. This means that the correct segment to attach the second subclause to will not be available in the discourse structure, and attaching it to the remaining matrix clause + first subclause EDS would yield a descriptively inadequate structure as in Fig. 1.[6] Thus, we introduce a general exception pertaining to all segmentable clause types as listed above:[7]

- Any potential EDS shall *not* be segmented if it is coordinate or subordinate to a clause that is *not* segmented according to the criteria above, either.

[6] For drawing RST trees we employ the tool sketched in [12].

[7] Note that from the segmentation criteria suggested in [4], the same problem arises, albeit in fewer cases.

Listing 1.1. XML format SEG for segmented text

```
<cds type="para" docIdref="i1119">
  <sds id="s87">
    <eds id="e149">Die Frage der beiden Nationalsprachen ist für die finnische Bevölkerung
       so gut wie eine Selbstverständlichkeit,
    </eds>
    <eds id="e150"> aber vor 150 Jahren war die Sprachfrage ein heikles Thema.
    </eds>
  </sds>
  <sds id="s88">
    <eds id="e151">Es hing mit dem Nationalitätsgedanken zusammen,
    </eds>
    <eds id="e152"> obwohl Finnland damals zu Russland gehörte.
   </eds>
 </sds> [...]
</cds>
```

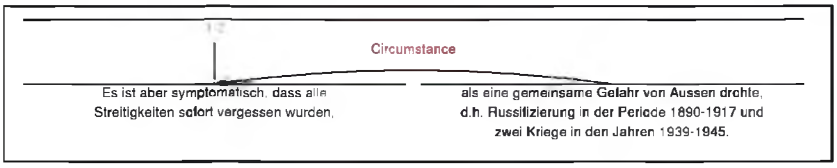
Circumstance
Es ist aber symptomatisch, dass alle Streitigkeiten sofort vergessen wurden,
als eine gemeinsame Gefahr von Aussen drohte, d.h. Russifizierung in der Periode 1890-1917 und zwei Kriege in den Jahren 1939-1945.

Fig. 1. Example of a potential segment boundary in a subordinate clause leading to a descriptively inadequate discourse structure

2.3 XML Format for Segmented Text

We store a discourse-segmented text in an XML annotation layer called SEG, where EDSs are contained in an element called <eds>. But not only EDSs are marked, additionally there are <sds> elements for SDSs (*sentential discourse segments*), i.e. text segments that correspond to sentences, as well as <cds> elements for CDSs (*complex discourse segments*), i.e. text segments that correspond to elements on the DOC layer. After having been identified by the segmenter, their purpose is to serve as input to and to guide the parsing cycles in the discourse parser, cf. [13]. Listing 1.1 shows an example of text annotated according to the SEG format.[8]

3 Algorithm

Segmentation is performed in three major phases corresponding to the identification of CDS, SDS, and EDS boundaries, cf. Figure 2. The result annotation layer SEG is constructed in a top-down fashion in the three phases. The basic idea for EDS recognition is to first determine all *potential* EDS boundaries by looking at punctuation and coordination, and then to successively remove those that can be established as non-EDS-marking by looking at their syntactic features.

[8] An extract from [10].

The segmentation component is implemented in Perl, using the LibXML and LibXSLT libraries to process the XML input document. Each phase is realised in one perl module. During the segmentation process, the syntactic parser *Machinese Syntax* from Connexor Oy. is repeatedly called. It provides output in XML ("CNX" in Fig. 2), containing morphological and syntactic tags for each token, as well as dependency relations between words based on Functional Dependency Grammar [14].

Phase 1: CDS recognition. The elements of the DOC annotation layer are all straightforwardly transformed into `<cds>` elements, still distinguished by a `@type` attribute specified e.g. for the value `para`, `sect`, `table`, or `title`. The specification `@type="eds"` marks those `<cds>` that at the same time correspond to EDSs according to Table 1.

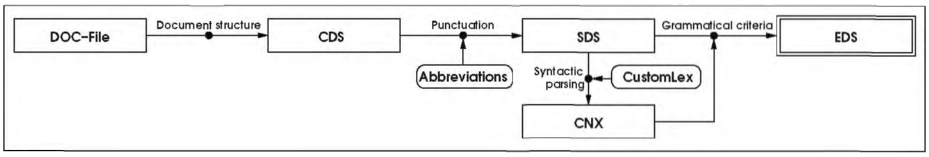

Fig. 2. Three phases of segment identification

Phase 2: SDS recognition and syntactic parsing. In the second phase, the textual content of those `<cds>` elements with `@type="para"` is further segmented. By using punctuation and a list of stop words abbreviations, the sentence boundaries are determined and `<sds>` tags are added to the SEG annotation layer. Sentence boundaries inside quotations and parentheses as described in criteria 4 and 5 in Sect. 2.2 as well in certain DOC elements (Sect. 2.1) are ignored. Then for each SDSs obtained, the syntactic parser *Machinese Syntax* is called. The reason for not doing this in a preprocessing step over the whole document is that the parser has its own internal rules to detect sentence and paragraph boundaries which may contradict the boundaries determined here via the DOC annotation layer.

Phase 3: EDS recognition. In phase 3, elementary discourse segments (EDS) are determined according to the grammatical criteria presented in Sect. 2.2. To this end, the segmenter accesses morphosyntactic information from the syntactic parser, i.e. POS-tags and information about the finiteness of verbs.

To identify EDS boundaries within an SDS, firstly, all *potential EDS boundaries* are marked and numbered. Potential boundaries are commas (except commas in numbers such as in *27,8%*), the lexical discourse markers *und* and *oder* as well as parentheses. Subsequently, boundary markers within parentheses as well as within quotation marks are deleted according to criteria 4 and 5.

At the beginning of the grammatical analysis, the POS tags of a sentence are used to build up phrasal information (NP and PP) and to store it in a string variable called $ic associated with the current SDS.[9] During the analysis, the potential boundary markers are one after the other tested for being a non-boundary, that is, whether they mark enumerations, relative clauses, clausal subjects and complements, proportional clauses, and infinitival complements. Only if all tests are negative, a boundary will be preserved.

An enumeration is a coordination of PPs, NPs, or APs. The recognition of such coordinations is achieved by looking at the variable $ic as explained above. From $ic = "P ART N und ř1ř ART N", simple phrases ($ic = "P NP und ř1ř NP"), then complex phrases ($ic = "PP und ř1ř NP") before and after the conjunction are generated and each time compared with each other. If the POS or phrasal categories match, the potential boundary is identified as enumerative and the actual boundary flag for the respective marker is set to 0.

Clausal subjects and complements start with the subordinating conjunctions *dass*, *ob*, or a wh-pronoun, or are infinitival constructions starting with *zu*. Their identification is combined with a check for attributional constructions which form the matrix clauses of clausal complements but are still EDSs according to criterion 5.

The results of the tests (value 0 or 1 for ***boundary*** or ***non-boundary***) are stored in a complex data structure associated with the current SDS. After the results of all tests for one SDS are available, these are evaluated and actual non-boundary markers are removed in a function called `ignore()`. The order in which the results are evaluated is crucial for the determination of EDSs. The whole process of evaluation (the function `remove-marker()`) of the test results is shown in Figure 3. First, those markers that are associated with the conjunctions *und* and *oder* are removed if the conjunctors were only enumerative. Then, those markers that are associated with a comma are evaluated in order (see Figure 3). If the current comma marker is associated with a sentential subject or complement, or a proportional clause, or an infinitival complement, `ignore()` is called, removing the marker itself and all other markers up to the following comma marker. If the current comma marker is associated with an enumeration, only the current marker is removed. If it is associated with a relative clause, then not only all markers up to the next comma marker ($next) are removed, but also the marker after that one. If $next marked an enumeration or a relative clause, $next is re-calculated, and `ignore()` is called again. This procedure represents a default solution for the general exception sketched in 2.2, i.e. currently all clauses following a main clause and a sub-clause EDS are treated as being sub-subordinated. After this evaluation of tests for potential EDS boundaries, each EDS established so far is checked for further internal boundaries brought about by EDSs below the clause level, i.e. phrases. Currently only attributional PPs are checked for (e.g. *nach Allardt*).

[9] Alternatively, phrasal information could be derived from the dependency structure information in the parser output. But since the POS information seems more reliable and is easier to use, for the time being we use only POS information.

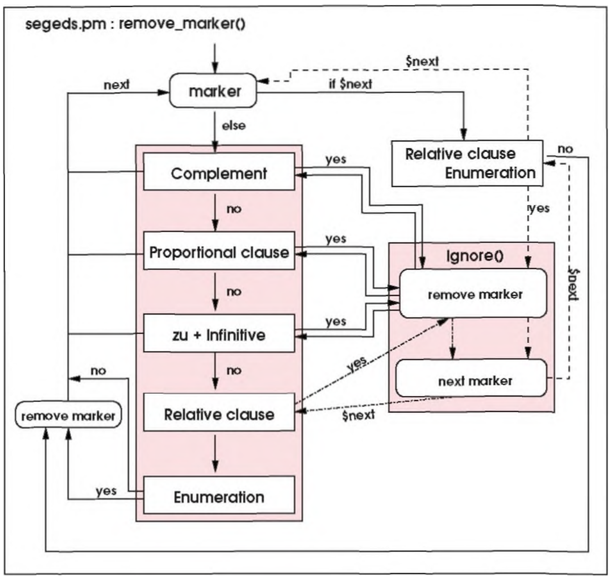

Fig. 3. The function `remove-marker()`

4 Results and Discussion

We performed test runs of the segmenter on six different texts. Four of them were scientific articles from our corpus (A-003, A-010, A-040, and A-023). To evaluate the performance on other text types as well we additionally segmented a web-published article on hypertext (L) and one newspaper article (Z). Manual segmentations of all six texts were provided by experts and served as "master" annotations containing the correct segmentations, against which precision and recall were then calculated. Table 2 gives statistics of the test texts as well as three groups of results of test runs.

The first group shows the performance of a baseline segmenter on all six texts. It executes CDS and SDS segmentation as described above and on top of that simply converts each comma into an EDS boundary. The second group shows the results of the segmenter applying the complete segmentation procedure as described in Sect. 3 to the texts. The final group shows the performance of pure sentence segmentation based on the CDS and SDS segmentation as described above, evaluated against the manually produced sentence segmentations of the six texts.

For all six texts, the performance of the fully informed EDS segmenter was significantly better than the baseline version. The SDS segmenter performed well in general, however text L additionally contained XHTML elements on the DOC layer, which the segmenter simply ignores, but which were considered boundary markers in the master segmentation (e.g. <`xhtml:br`>). This was the cause of many segmentation errors in text L.

Table 2. Results

	Texts	A-003	A-010	A-040	A-023	L	Z
Statistics	# wordforms	12323	2239	6560	5450	3138	1448
	# master eds	758	154	497	338	292	136
	# master sds	470	103	300	231	148	90
Baseline-EDS	% Precision	0.34	0.41	0.39	0.25	0.45	0.43
	% Recall	0.59	0.66	0.62	0.48	0.62	0.59
Segmenter-EDS	% Precision	0.80	0.82	0.78	0.60	0.74	0.76
	% Recall	0.80	0.88	0.77	0.67	0.66	0.80
Segmenter-SDS	% Precision	0.89	0.98	0.92	0.84	0.85	0.98
	% Recall	0.93	0.99	0.92	0.90	0.74	0.99

Text A-003 is the longest text and the one that we inspected most closely when implementing and debugging the segmenter. The texts A-010 and Z were not previously inspected in that way but recall is equally good or even better for them.

The EDS segmenter still has some shortcomings regarding the implementation of some of the segmentation criteria, which were considered not too significant for text A-003. In some texts, however, they produce a higher fraction of errors. Sometimes, for example, the EDS segmenter does not recognise attributional constructions, firstly because not all possible verbs of attribution are inventorised yet, and secondly because even for humans it is sometimes difficult to distinguish attributional constructions from non-attributional ones according to criterion 6. Likewise, the segmenter sometimes does not recognise appositives (criterion 9) well because they can be confused with NP enumerations. In the sentence *Dazu hat das Land seit 1995 drei offizielle Minderheitssprachen, Samisch, Romani und Gebärdensprache*, for example, the first comma is a segment boundary separating the trailing appositive from the main clause. The second comma, however, marks only an enumeration of NPs. The problem is that NP-enumerations are identified by checking for consecutive NPs that agree in their case value, which also holds for appositives, i.e. such constructions are functionally ambiguous.

A second type of segmentation errors is caused by faulty analyses of the syntactic parser which tend to occur with very long and complex sentences. Some such errors could be avoided by tuning our segmenter accordingly. Relative pronouns, for example, are sometimes POS-tagged as determiners, so our implementation checks whether forms such as *der, die, den* are rather relative pronouns, by additionally looking at the POS tags in the context.

The principle of not segmenting certain sub-subordinated or sub-coordinated clauses as described in Sect. 2.2 has proven difficult to implement, because with multiple subordinations it is not uncommon that the output from the syntactic parser is already faulty. At the moment we have implemented a default strategy that regards every subclause following a subclause that was preceded by a main clause as sub-subordinated. Though this seems to cover the majority of cases, it is also the cause of several unidentified boundaries that affect recall figures.

A third type of error turned out to be an author's omission of commas or using too many commas. Several segmentation errors in text A-040 proved to be due to such mispunctuations. A solution could be to rely more systematically on lexical discourse markers as boundary signals as in [2], however for most texts we do not expect this to improve recall figures significantly.

5 Summary and Outlook

We presented an automatic discourse segmenter for German written text to be used in the framework of RST-based discourse parsing in a text-technological environment. We defined the notion of an elementary discourse segment (EDS) by adapting the widely used segmentation principles for English EDUs presented in [4] to German while also considering aspects of the document structure of a complex text type, namely scientific articles. Thus the criteria in defining our EDSs are based on logical document structure, syntax, and punctuation.

Unlike the discourse segmenters presented in [3] and [5], we employ a knowledge-based procedure that does not require a large amount of training data (which is not available for German). And unlike the segmenters presented in [3] and [6], we do not presuppose that an input text comes together with its correct syntactic analysis; instead we have integrated a syntactic parser that is used online in the segmentation process.

Our segmenter first performs a segmentation of CDS induced by elements of the logical document structure, then a segmentation of SDS based on logical document structure and punctuation. Subsequently, the syntactic parser is called for each SDS. EDS segmentation is then performed by marking the potential segment boundaries of an SDS and successively checking whether they are not actual segment boundaries, using the syntactic analyses. By first setting potential boundary markers and then eliminating the actual non-boundaries, we have considerably reduced the amount of analysis (i.e. the number of tests required in Phase 3) in comparison with the opposite strategy of directly establishing the actual EDS boundaries. A strength of our approach lies also in the separation of the syntax checks from the evaluation of their results in phase 3. It enables us to modify or extend segmentation criteria easily if desired.

The performance of our system (EDS and SDS segmentation) on three of the six texts used in the evaluation was slightly worse than that of the knowledge-based segmenter for English reported in [6] (79% overall recall), and our recall figures also remain lower than those reported for the statistical approach to discourse segmentation in [5] (85,4% recall of sentence-internal EDU boundaries). However, on account of our evaluation we reported several types of segmentation errors that can be remedied by further use of the syntactic analysis and that we will tackle in the near future to further improve the performance.

References

1. Mann, W.C., Thompson, S.A.: Rhetorical Structure Theory: Toward a functional theory of text organisation. Text 8(3) (1988) 243–281
2. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge, MA (2000)

3. Marcu, D.: A decision-based approach to rhetorical parsing. In: Proceedings of the 37th annual meeting of the ACL, Maryland, Association for Computational Linguistics (1999) 365–372
4. Carlson, L., Marcu, D.: Discourse tagging reference manual. Technical report, Information Science Institute, Marina del Rey, CA (2001) ISI-TR-545.
5. Soricut, R., Marcu, D.: Sentence level discourse parsing using syntactic and lexical information. In: Proceedings of the Human Laanguage Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), Edmonton, Canada (2003)
6. Le Thanh, H., Abeysinghe, G., Huyck, C.: Automated discourse segmentation by syntactic information and cue phrases. In: Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2004), Innsbruck, Austria (2004)
7. Sporleder, C., Lapata, M.: Discourse chunking and its application to sentence compression. In: Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP-05), Vancouver, Canada (2005)
8. Le Thanh, H., Abeysinghe, G., Huyck, C.: Generating discourse structures for written texts. In: Proceedings of COLING'04, Geneva, Switzerland (2004)
9. Walsh, N., Muellner, L.: DocBook: The Definitive Guide. O'Reilly (1999)
10. Saari, M.: Schwedisch als die zweite Nationalsprache Finnlands: Soziolinguistische Aspekte. Linguistik Online 7 (2000) http://www.linguistik-online.de.
11. Krohn, P.: Arm, ärmer, kind. Die Zeit 15 (2005) 27
12. O'Donnell, M.: RSTTool 2.4 – A markup tool for Rhetorical Structure Theory. In: Proceedings of the International Natural Language Generation Conference (INLG'2000), Mitzpe Ramon, Israel (2000) 253 – 256
13. Lobin, H., Bärenfänger, M., Hilbert, M., Lüngen, H., Puskàs, C.: Text parsing of a complex genre. In: Proceedings of the Conference on Electronic Publishing (ELPUB), Bansko, Bulgaria (2006) to appear
14. Tapanainen, P., Järvinen, T.: A non-projective dependency parser. In: Proceedings of the 5th Conference on Applied Natural Language Processing, Washington D.C., Association for Computational Linguistics (1997) 64–71