Complex sentences as leaky units in discourse parsing

Nynke van der Vliet and Gisela Redeker

University of Groningen
n.h.van.der.vliet@rug.nl, g.redeker@rug.nl

Abstract. It is usually assumed that complex sentences with multiple clauses function as rhetorical units in discourse. We show that there are rare but systematic exceptions to this general assumption: structures where a sentence-external unit attaches to one of the clauses in a complex sentence before the combined span joins the rest of the complex sentence. In our RST-annotated Dutch text corpus, 13% of the complex sentences have such 'leaky' boundaries. We distinguish four structural types and argue that only two rather infrequent types pose a serious problem for sentence-first discourse parsing.

Keywords: RST, discourse structure, Dutch

1 Introduction

The assumption that the hierarchical structure of a text correlates with the orthographic layout of a text is often implemented in discourse parsers [1-3]. In these systems the sentences, paragraphs and sections in a text correspond to the hierarchical spans in the rhetorical representation of the text. In particular, the clauses in complex sentences are usually combined before considering combinations with other units.

Our analysis of a manually annotated corpus shows that there are cases where the segments of one complex sentence do not constitute a hierarchical text span. Instead, at least one of the sentence segments is attached to (a segment of) another sentence. In our corpus, such 'leaky' sentence boundaries occur in 13% of the complex sentences. In this paper we will explore these structures and their consequences for automatic discourse parsing.

2 Corpus

Our corpus contains 80 Dutch texts and covers a range of text genres, including, in particular, expository texts, whose main purpose is to present information to the reader, and persuasive texts that aim to affect the readers intentions or actions. For the expository subcorpus, 20 texts have been selected from online encyclopedias on astronomy[1] and 20 from a popular scientific news website.[2] The persuasive texts are

[1] http://www.astronomie.nl; http://www.sterrenwacht-mercurius.nl/encyclopedie.php5
[2] http://www.scientias.nl/categorie/astronomie

20 fundraising letters from humanitarian organizations and 20 commercial advertisements from lifestyle and news magazines. The texts vary in length between a minimum of approximately 190 words and a maximum of approximately 400 words.

For the analysis of discourse structures, we chose the widely used Rhetorical Structure Theory (RST) [4,5]. RST describes the hierarchical structure of text by means of the relations between text parts. The analysis yields non-binary labeled tree structures, in which every part of the text has a role or function to play with respect to other parts of the text[3]. All annotations were done separately by at least two expert annotators using O' Donnell's RSTTool,[4] and then discussed and reconciled.

We computed inter-annotator agreement for the initial versions of the expert annotators of the RST analysis for two fundraising letters and two encyclopedia texts, using the methods proposed in [6]. On average, the kappa for agreement on the spans was 0.88, on nuclearity 0.82, and on the RST relation labels 0.57. For more details about the corpus and the annotations, see [7].

3 Analysis

Table 1 shows the number of sentences and complex sentences in the different genres in the corpus. The third data column shows the number of 'leaky' sentences: complex sentences that are not represented by one hierarchical span of that sentence in the RST analysis, but by a larger structure in which (at least) one of the sentence segments is first attached to another sentence.

Table 1. Leaky complex sentences in Encyclopdia Entries (EE), Popular Scientific News (PSN), Fundraising Letters (FL) and Advertisements (AD)

Genre	Sentences (Se)	Complex sentences (C)	Leaky (L)	L/C	L/Se
EE	396	167	14	8%	4%
PSN	435	121	15	12%	3%
FL	467	106	17	16%	4%
AD	399	91	16	18%	4%
Total	*1697*	*485*	*62*	*13%*	*4%*

3.1 Leaky structures

There are 73 leaky structures, i.e., instances where a sentence-external segment attaches to a segment inside a complex sentence, in our corpus. Sometimes two or (in one case) three such instances occur with one complex sentence. We identified four types of leaky structures by classifying them according to the internal structure of the complex sentence and the type of inter-sentential link in the RST structure. The four

[3] Full relation definitions are available on the RST website http://www.sfu.ca/rst.
[4] Available from http://www.wagsoft.com/RSTTool

types differ in complexity and, we will argue, in the consequences for a strictly bottom-up sentence-first parsing strategy.

We will describe these structures with the notation N(N:NS)(1a,sp(1b,2)), where N = Nucleus, S = Satellite, sp = span, and 1a, 1b, and 2 are the text segments.[5] The segments marked with a and b are the segments of the complex sentence. Thus, NNN(1a,1b,2) corresponds to the RST structure illustrated in figure 1 below (left diagram), and N(N:NS)(1a,sp(1b,2) to the righthand structure in figure 1.

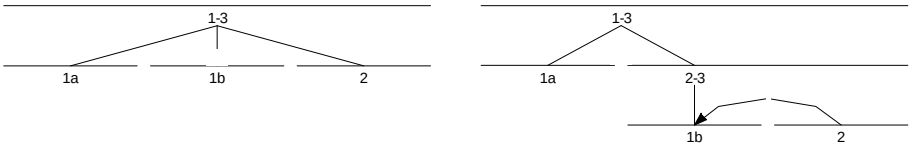

Fig. 1. Examples of RST structures

Type 1: Same-level coordination structures. In these cases, the segments of a complex sentence are connected by a multi-nuclear relation that also includes one or more segments of another sentence: NNN(1a,1b,2) or NNN(1,2a,2b). We found eight such cases in our corpus. In the example in figure 2 below, the three legs of the conjunction describe what happens to the elementary particles that fly through the atmosphere. A sentence-first analysis, where segments 1 and 2 are joined first to form a separate hierarchical span, would suggest that the first two legs of the conjunction are more closely related to each other than to the third leg, which is in our view less plausible.

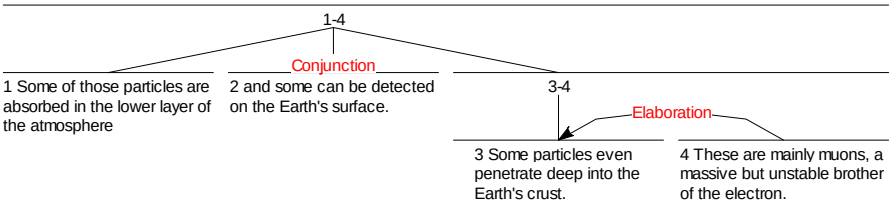

Fig. 2. Same-level coordination structure [EE13:26-29]

The coordination structures in our corpus concern Conjunction, List and Sequence relations. In all cases, a representation in which the intra-sentential relations are combined first would involve an often not well-motivated stacked structure of two multinuclear relations of the same kind.

[5] RST distinguishes two kinds of relations: The asymmetric *mononuclear* relations like *Elaboration* or *Justify* relate a nucleus N (centrally important) and a *satellite* S (additional information, which could in many cases be left out without rendering the text incoherent). The symmetric *multinuclear* relations like *List* or *Joint* relate discourse entities of equal status.

Type 2: Same-level subordination structures. Here the segments of a complex sentence are connected by a subordinating relation, but at the same level there is a relation between the nucleus and another sentence. In figure 3 below, the nucleus in segment 3 has two Motivation satellites, one of which consists of two other sentences. Combining the intra-sentential segments 3 and 4 before attaching the sentences in (1-2) would result in a Motivation relation between (1-2) and (3-4), which is not plausible, as 4 is giving a quite different motivation than (1-2).

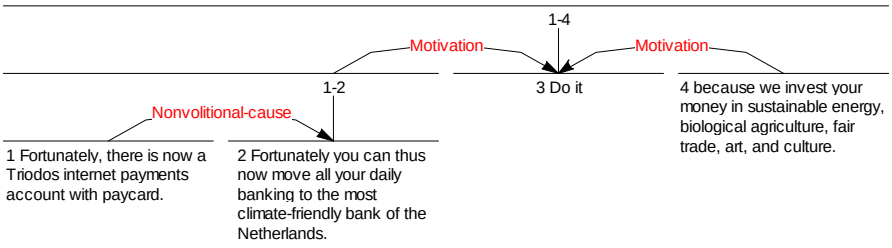

Fig. 3. Same-level subordination structure [AD06:11-14]

The 41 cases of this type in our corpus instantiate three of the four possible configurations: There are 18 cases with SNS(1a,1b,2), 16 with NSS(1a,1b,2), seven with SNS(1,2a,2b) and none with SSN(1,2a,2b).

For all these structures, a sentence-first analysis would include the nucleus and satellite of the intra-sentential relation in the nucleus of the inter-sentential relation. Depending on the inter-sentential relation, this is more or less problematic. As shown above, inter-sentential Motivation relations (11 cases) can lead to implausible sentence-first representations. Elaboration relations (22 cases) are less problematic. This is due to the special nature of this relation (discussed, e.g., in [8]): The Elaboration satellite "presents additional detail about the situation or *some* element of subject matter which is presented in the nucleus" (RST relation definition). It is thus fine for a satellite to elaborate only a part of the nucleus.

Type 3: Multi-level coordination structures. In these structures, the segments of the complex sentence are connected to another sentence in a multi-level structure that contains a coordination relation. The most common structure is N(N:NS)(1a,sp(1b,2)) (6 cases in the corpus), where the right leg of the coordination relation between 1a and 1b is elaborated by the next sentence (or a larger text span). In the example in figure 4 below, segment 3 elaborates on how the people learn how to protect themselves, but not on the emergency help that GUK provides. We could also attach segment 2 on top of a coordination relation between segment 1a and 1b ((N:NN)S(sp(1a,1b),2)), but that structure does not represent the fact that segment 3 only elaborates on segment 2.

In general, creating an alternative structure in which sentences are combined first in these cases results in an implausible representation with both nuclei of the multinuclear relation inside the scope of the asymmetric inter-sentential relation. As

with the same level coordination structures, it seems that inter-sentential Elaboration structures are less problematic than other discourse relations, due to the fact that elaboration satellites can easily elaborate only a part of their nucleus.

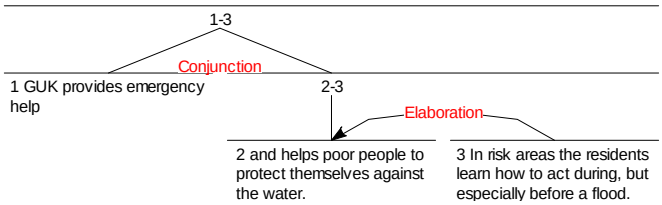

Fig. 4. Multi-level coordination structure N(N:NS)(1a,sp(1b,2)) [FL05:20-22]

Figure 5 shows one of our two cases with the structure S(N:NN)(1a,sp(1b,2)). In this example, the Concession relation would not make much sense if segment 2 wasn't enhanced by span (3-4) explaining why crème fraîche is not a household staple (and why the alternative product advertised here should be preferred). In cases like this, any representation in which the segments of the first sentence together form one hierarchical span would be odd. The same holds for the structures (S:NN)N(sp(1,2a),2b) and N(S:NN)(1a,sp(1b,2)), each occurring only once in the corpus.

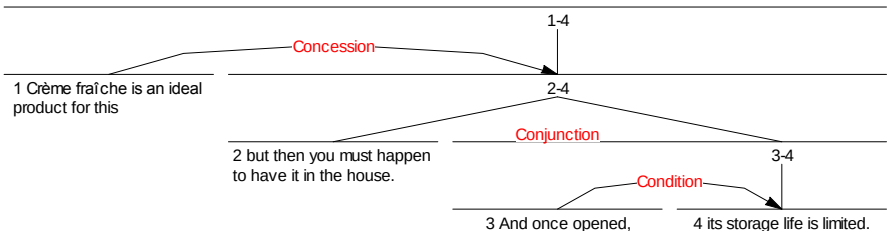

Fig. 5. Multi-level coordination structure S(N:NN)(1a,sp(1b,2)) [AD03:5-8]

Type 4: multi-level subordination structure. In multi-level subordination structures, the segments of the complex sentence are connected to another sentence in a multi-level structure that contains only subordination relations. A problematic structure is N(S:NS)(1a,sp(1b,2) (8 cases in the corpus), as shown in figure 6 below. In this example, segment 3 elaborates on the half year that the space telescope has been active, by describing what the telescope has found in this short time. Segment 2 explains why it is remarkable that WISE has already discovered 25.000 asteroids (segment 1). In a sentence-first representation, segment 3 would have to be attached to the span (1-2), of which segment 1 would be the nucleus. This is clearly not a good solution, because segment 3 only elaborates segment 2, not segment 1. A representation in which intra-sentential segments are combined first violates the RST constraint that whenever two text spans are connected through a rhetorical relation, that relation also holds between the most salient parts (the nuclei) of the constituents

[9]. The same applies to structures of the type (S:SN)N(sp(1,2a),2b) (1 case in the corpus).

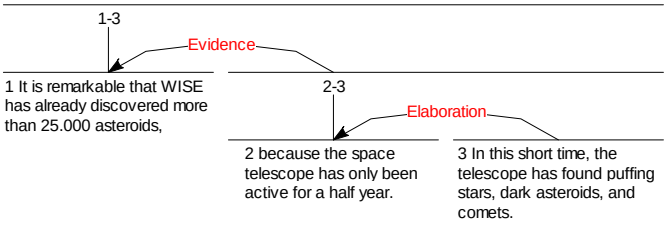

Fig. 6. Multi-level subordination structure N(S:NS)(1a,sp(1b,2) [PSN12:14-16]

From the example in figure 6 it is clear that, unlike the cases described in the previous two categories, inter-sentential Elaborations are in this structure *not* less problematic than other discourse relations.

Figure 7 shows one of the five cases with the structure S(N:NS)(1a,sp(1b,2). Segment 4 gives extra information about which dwarf planets are affected by the expansion of the group described in segment 3. Segment 1 and 2 describe the condition under which the group of dwarf planets will become larger. In a sentence-first representation, segment 4 would be outside the scope of the Condition relation, which is not plausible. In three of the five cases of this subtype, the intra-sentential relation is a Condition relation. These cases would all lead to implausible sentence-first representations. The two remaining cases (one with an Evaluation and one with a Concession relation) are less problematic.

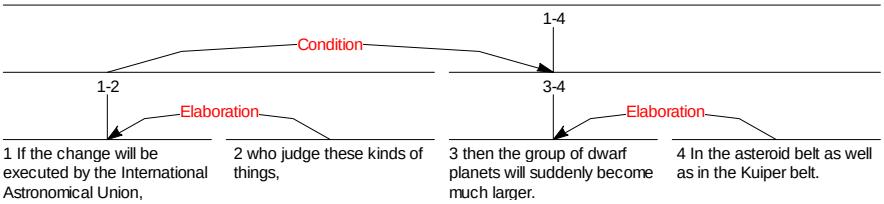

Fig. 7. Multi-level subordination structure S(N:NS)(1a, sp(1b,2)) [PSN16:19-22]

4 Discussion

Our discussion of the four types of 'leaky' complex sentence structures has shown that structures of types 1 and 2 tend to be rather unproblematic for sentence-first parsing. In type 1 (same-level coordination structures), the stacked coordinations in a sentence-first parse – with the inter-sentential multinuclear relation applying after the intra-sentential one – would be only slightly less plausible than our analysis. In type 2 (same-level subordination structures), sentence-first alternatives are unproblematic if the breaching inter-sentential relation is an Elaboration (22 of our 41 cases); for other relations, however, sentence-first parses are often clearly suboptimal (e.g. including

one Motivation satellite inside the nucleus of another). Structures of types 3 and 4, by contrast, are always problematic in the sense that a sentence-first parse would yield implausible or even unacceptable readings.

Distribution. Leaky complex sentence structures occur in all four genres in our corpus, with an average occurrence of 15 cases per 100 complex sentences. The frequencies of the four types of structures are shown in table 2. Same-level coordination and subordination structures (types 1 and 2) together account for 49 (67%) of our cases, and same-level subordination structures (type 2) are by far the most frequent in all genres. The problematic types 3 and 4 are very rare in the encyclopedia texts and most frequent in the popular science news and in the fundraising letters. Overall, their average occurrence is five cases in 100 complex sentences.

Table 2. Leaky structures per type (counts and occurrences per 100 complex sentences)

Genre	Same-level coordination	Same-level subordination	Multi-level coordination	Multi-level subordination
EE	3 (1.8)	12 (7.2)	0 (0.0)	3 (1.8)
PSN	1 (0.8)	9 (7.4)	1 (0.8)	6 (5.0)
FL	1 (0.9)	10 (9.4)	5 (4.7)	4 (3.8)
AD	3 (3.3)	10 (11.0)	4 (4.4)	1 (1.1)
Total	*8 (1.6)*	*41 (8.5)*	*10 (2.1)*	*14 (2.9)*

Why do these sentence boundaries leak? The idea that discourse structure is not always built up hierarchically from intra-sentential to inter-sentential structures is rather counter-intuitive. One might suspect that these are somehow production errors, e.g. afterthoughts or simply faulty or sloppy punctuation. The example in figure 7 above could be read as such a case. Note however that the separate presentation of the sentence fragment in segment 4 serves the rhetorical purpose of highlighting that information – in fact it would be very inelegant to include this PP in the already long sentence presented in segments (1-3). More generally, note that the genres in our corpus tend to be rather well edited (especially the advertisements and fundraising letters), making it unlikely that 13% of the complex sentences should be badly written or carelessly punctuated.

Could this be an artifact of our analytical model? RST assumes that discourse structure can be represented with non-binary trees with symmetric and asymmetric relations. We will discuss each of these components in turn.

Treeness. The assumption that discourse structure can be represented by a strictly hierarchical structure (a tree) (argued for, e.g., in [10]) implies that each segment can only have one immediately dominating (parent) node. This constrains the accessibility of subordinate segments for further attachments (strong empirical support for such a – in their case weaker – constraint is reported in [11]).

Non-binarity. Our tree structures are allowed to be non-binary. This includes not only multi-segment coordinate structures (in multinuclear RST relations), but also multi-satellite structures (as in figure 3 above). Our type-1 leaky structures cannot occur in binary trees. Note, however, as we have argued elsewhere [7], that the stacking of binary coordinate structures to represent a coordination of more than two segments tends to invite unintended hierarchical interpretations. For structures involving multiple satellites attaching to one nucleus in our analysis, the situation would in fact be worse with binary trees, as they exclude the same-level option and force a decision to apply the intra-sentential relations either before or after the inter-sentential one.

Symmetric and asymmetric relations. This distinction affects our judgments of the acceptability of alternative structures, as only nuclei and not satellites in a span are directly involved in relations of that span to other spans or segments. The notions of coordination and subordination in discourse are, however, generally accepted (see e.g. [12]) and can be traced back at least as far as [13]. Their role in the leaky structures should therefore not be dismissed as nuisance or an artifact.

5 Future Work

Complex sentences are an important source of local structure information in discourse. They provide grammatically derivable clues to coordinating and subordinating intra-sentential relations, e.g. through conjunctions and adverbials (in our corpus, 69% of the intra-sentential RST relations are explicitly signaled, compared to only 16% of the inter-sentential relations). We will therefore continue to explore sentence-combining-first strategies in discourse parsing. In particular, we will look for clues (e.g., connectives or lexical cohesion) to detect possible leaks at the boundaries of complex sentences.

Acknowledgement. This work has been supported by grant 360-70-280/282 of the Netherlands Organization for Scientific Research (NWO) as part of the program Modelling Textual Organization (http://www.let.rug.nl/mto/).

References

1. Hernault, H., Prendinger, H., du Verle, D.A., Ishizuka, M.: HILDA: A Discourse Parser Using Support Vector Machine Classification. Dialogue and Discourse 1(3), 1−33 (2010)
2. Le Thanh, H.: An Approach in Automatically Generating Discourse Structure of Text. Journal of Computer Science and Cybernetics 23(3), 212−230 (2007)
3. Marcu, D.: The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. Computational Linguistics 26(3), 395−448 (2000)
4. Mann, W.C., Thompson, S.A..: Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. Text 8(3), 243−281 (1988)
5. Taboada, M., Mann, W.C.: Applications of Rhetorical Structure Theory. Discourse Studies 8(4), 567−588 (2006)

6. Marcu, D., Amorrortu, E., Romera, M.: Experiments in Constructing a Corpus of Discourse Trees. In: Proceedings of the ACL99 Workshop on Standards and Tools for Discourse Tagging, pp. 48−57. Association for Computational Linguistics, Stroudsburg, PA (1999)
7. van der Vliet, N., Berzlánovich, I., Bouma, G., Egg, M., Redeker, G.: Building a Discourse-annotated Dutch Text Corpus, In: Dipper, S., Zinsmeister, H. (eds.) Bochumer Linguistische Arbeitsberichte, vol. 3, pp. 157−171. Bochum Linguistics Department, Bochum, Germany (2011)
8. Knott, A., Oberlander, J., O'Donnell, M., Mellish, C.: Beyond elaboration: The interaction of relations and focus in coherent text. In Sanders, T., Schilperoord, J., Spooren, W. (eds.) Text representation: linguistic and psycholinguistic aspects, pp. 181–196. Benjamins, Amsterdam (2001)
9. Marcu, D.: Building up Rhetorical Structure Trees. In: Proceedings of the Thirteenth National Conference on Artificial Intelligence, pp. 1069−1074. AAAI Press, Menlo Park, California (1996)
10. Egg, M., Redeker, G.: How complex is discourse structure? In: Proceedings of LREC'10, Malta, pp. 1619–1623. ELRA, Paris (2010)
11. Afantenos, S.D., Asher, N.: Testing SDRT's Right Frontier. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pp. 1-9. Beijing (2010)
12. Asher, N., Vieu, L.: Subordinating and Coordinating Discourse Relations. Lingua, 115(4), 591−610 (2005)
13. Grosz, B., Sidner, C.: Attention, intention and the structure of discourse. Computational Linguistics 12, 175–204 (1986)