Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

The Role of Paragraph in the Corpora of Annotated Texts

Inkova O. Y.
Institute of Informatics Problems,
FRC CSC RAS, Moscow, Russia;
University of Geneva, Geneva,
Switzerland
Olga.Inkova@unige.ch

Nuriev V. A.
Institute of Informatics Problems,
FRC CSC RAS, Moscow, Russia
nurieff.v@gmail.com

Popkova N. A.
Institute of Informatics Problems,
FRC CSC RAS, Moscow, Russia
Natasha__popkova@mail.ru

Abstract

The paper focuses on the function of paragraph both in text organization and in text annotation from the point of view of coherence. Taking as examples three major types of corpora (the RST, ANNODIS, and PDTB corpora), it shows whether and to what extent the existing approaches account for the paragraph when a discourse relation gets annotated. Then it presents the theoretical principles underlying text annotation in two databases: the Supracorpora database of connectives and the Supracorpora database of hierarchical logical-semantic relations (a new linguistic resource). Text coherence is shown to result from the interaction of various discourse phenomena, acting at the level of local and global structures. In this approach, the paragraph is assigned to the meso-level, positioned between local and global levels. The researcher may analyze the internal organization of the paragraph, limiting oneself to the inter-sentential level. Yet, to analyze and describe how paragraphs follow one another in the text, it is necessary to operate at the supra-sentential level, adopting a conceptual apparatus fundamentally different from the one for the description of local text structure.

Keywords: paragraph; text annotation; corpus; discourse relations; database
DOI: 10.28995/2075-7182-2022-21-273-283

Роль абзаца в корпусах аннотированных текстов

Инькова О. Ю.
ИПИ ФИЦ ИУ РАН, Москва, Россия;
Женевский университет,
Женева, Швейцария
Olga.Inkova@unige.ch

Нуриев В. А.
ИПИ ФИЦ ИУ РАН,
Москва, Россия
nurieff.v@gmail.com

Попкова Н. А.
ИПИ ФИЦ ИУ РАН,
Москва, Россия
Natasha__popkova@mail.ru

Аннотация

Статья рассматривает функции абзаца в структуре текста, а также при аннотировании текстов с точки зрения связности. На примере трех наиболее известных корпусов (корпуса, созданные на основе Теории риторической структуры, корпус ANNODIS и PDTB) авторы анализируют существующие подходы и то, в какой степени абзац учитывается при определении дискурсивного отношения или, наоборот, его отсутствия. Авторы формулируют теоретические принципы, лежащие в основе аннотирования в двух базах данных: надкорпусной базы данных коннекторов и надкорпусной базы данных иерархии логико-семантических отношений, нового лингвистического ресурса. Показано, что связность текста осуществляется в результате взаимодействия дискурсивных явлений различной природы, действующих на уровне как локальной, так и глобальной структуры. Абзац при таком подходе является единицей мезоуровня, промежуточного между локальным и глобальным. Если внутренняя организация абзаца может быть описана на межфразовом уровне, то следование абзацев в структуре текста должно быть описано на сверхфразовом уровне и в терминах, принципиально отличных от используемых для описания локальной структуры текста.

Ключевые слова: абзац; аннотирование текстов; дискурсивные отношения; база данных

1 Introductory remarks

It is known that nowadays for the annotation of discourse relations there are several approaches available. All of them are directly related to the theoretical approach that underlies the understanding of discourse relation and, more broadly, text coherence. We will start with a brief overview of resources where texts are annotated in terms of discourse relations (the RST, ANNODIS, and PDTB corpora), showing their specifics and annotation theoretical guidelines, primarily focusing on the function of paragraph in text organization. Then we will demonstrate how some theoretical assumptions have been adopted for the text annotation in the new linguistic resource – the Supracorpora database of hierarchical logical-semantic relations.

2 The RST corpora

In the annotation of discourse (or rhetorical) relations, the most common theoretical approach is known to be the Rhetorical Structure Theory (RST, Mann & Thompson 1988). The four main theoretical principles underlying it say that: (1) no piece of discourse should be left out of the analysis (completeness condition); (2) all text fragments are interconnected (connectedness condition); (3) the same discourse units (DUs), i.e. clauses, can be connected by only one relation (uniqueness condition); (4) DUs directly follow each other and cannot overlap between themselves (adjacency condition).

The text annotation has to satisfy these three conditions. Therefore

- the relations connecting DUs, however diverse they may be, are of the same order, be it syntactic (explanatory and relative clauses) or semantic dependence, anaphoric repetition, thematic progression, or text structure (division into paragraphs, chapters, title, author, etc.);
- in the hierarchical text structure, the elements of either global or local structure are connected by the same rhetorical relations, and the use of these relations is recursive;
- the entire text can be presented as a single graph.

That the view of the RST on text organization and coherence is somewhat simplified has been repeatedly pronounced by representatives of different linguistic schools. They have shown text coherence to be built simultaneously at several various levels: the genre of the text, its thematic organization, the communicative intentions of the speaker, the level of propositional content, and the level of discourse relations (that are understood in a narrower sense than in the RST). We will not dwell on this issue. For more on this see, for example, Adam 2012, Inkova 2019, Webber et al. 2012.

Presenting the text as a single graph limits the size of annotated texts. While the founders of the RST claim that the text length does not matter[1], to make a single graph is possible only for small texts. The Ru-RSTreebank Annotation Manual (https://rstreebank.ru/), adopting the RST principles, even specifies that graphs are built only within paragraphs.

This comes, in turn, from the consideration that the markers of global and local structures have the same functions, and the minimal unit of global structure, especially in a newspaper article, equals a paragraph. Hence, its function – keeping the text coherent – is comparable to the rhetorical relation that connects the elements of local structure, for example, in the fragment: *Он заболел, поэтому не пришел.* Such role of paragraph in text organization traces back to the works of Kenneth Lee Pike and Robert E. Longacre. Since the languages of the Philippines and Papua and New Guinea are known to have some specific identifiers of the beginning and end of a paragraph, linguistic scholars who follow the ideas of Pike (1982) decided to assign paragraph to the fourth level of grammatical units of surface structure. There are words and syntagmas at the first level, clauses – at the second level, and sentences – at the

[1] "It is insensitive to text size, and has been applied to a wide variety of sizes of text" (Mann, Thompson 1988: 243).

third level. The paragraph thus belongs to the inter-sentential level of linguistic analysis. As for Longacre, he considers "the paragraph as a grammatical unit" (Longacre 1979) and proposes in his latest classification of paragraphs (Longacre 1996) rubrics resembling the rhetorical relations that one might find in the RST.

Later we will return to the functions of paragraph and its role in making the text coherent as we understand it. However, now it is to be noticed that while the connective *поэтому* in our example serves to convey the connection between two minimal DUs, the function of paragraph is to signal, on the contrary, the weakening of this connectedness between two larger DUs.

3 The ANNODIS corpus

The notion of text as a complex multi-level and multi-parameter system reflects in the ANNODIS corpus (http://redac.univ-tlse2.fr/corpus/annodis/). This resource builds on the Segmented Discourse Representation Theory (SDRT) (Asher 1993, Ascher, Lascarides 2003) and consists of several independent subcorpora. It aims to annotate various discourse phenomena, each of which contributes to text organization: rhetorical relations and two types of hierarchical structures, namely thematic chains (sequences of semantic blocks with a common topic) and enumerative structures. Through its annotations, the ANNODIS corpus shows how these three phenomena interact between them, making the text coherent. For example, for enumerative structures, one might see

- what rhetorical relation may hold between the initial sentence that specifies the enumeration and the enumeration itself (as a rule, it is Elaboration or Motivation),
- what relations, besides additive ones, can connect the members of the enumerative series.

The texts differ in length and genre, and they are annotated entirely. The annotation does not take into account the division into paragraphs, however, theoretically important is that the minimal DU can be a language unit less than a clause, and several minimal DUs can fall within its scope. We are talking about the so-called frame expressions, the function of which is to create semantic blocks of sentences that should be interpreted in relation to a single criterion (spatial, temporal, communicative). See (Charolles 1997, Inkova 2021) for details.

(1) *Согласно креационистской гипотезе*, которая имеет самую длинную историю, создание жизни есть акт божественного творения. Свидетельством этому является наличие в живых организмах особой силы, «души», управляющей всеми жизненными процессами. Гипотеза креационизма навеяна религиозными воззрениями и к науке отношения не имеет. (Л.А. Михайлов, Концепции современного естествознания. Учебник для вузов; books.google.it; accessed 12 January 2022)

The frame expression *согласно креационистской гипотезе* helps refine the interpretation of the first sentence, in which it occurs, opening the paragraph. This expression also refines the interpretation of the second sentence. Otherwise, it would give the impression that the author argues in favor of the correctness of this hypothesis. And the author, on the contrary, refutes its accuracy, explaining it in the third and last sentence of the paragraph.

The function of frame expressions in the text is thus twofold: on the one hand, they serve to integrate, combining minimal DUs into larger ones, and on the other hand, they divide into segments, signaling a weaker connectedness between semantic blocks that should be interpreted "separately" (in this case due to different speakers). As we will see, the paragraph can assume the same functions in the text, often signaling that the scope of the frame expression comes to its end. Both the SDRT and the RST consider "relations" conveyed by frame expressions as "rhetorical" ones (Prévot et al. 2009, Vieu et al. 2005).

4 The Penn Discourse Treebank

In the Penn Discourse Treebank (PDTB), they annotate, first of all, discourse relations that can potentially be expressed by a connective, i.e. the understanding of "rhetorical relation" is narrower than in the RST. Since the texts in the corpus (articles from the *World Street Journal*) are annotated entirely, the observations have resulted in three theoretical conclusions that distinguish this approach from the RST.

First, if it is impossible to place a connective between adjacent DUs, the relation cannot be qualified as discourse one. We are talking about such cases, “where the second sentence only serves to provide some further description of an entity in the first sentence” (PDTB Research Group 2008: 1). In (2) **EntRel** indicates this state of affairs:

(2) Hale Milgrim, 41 years old, senior vice president, marketing at Elecktra Entertainment Inc., was named president of Capitol Records Inc., a unit of this entertainment concern. **EntRel** Mr. Milgrim succeeds David Berman, who resigned last month. [Example borrowed from (PDTB Research Group 2008: 23)]

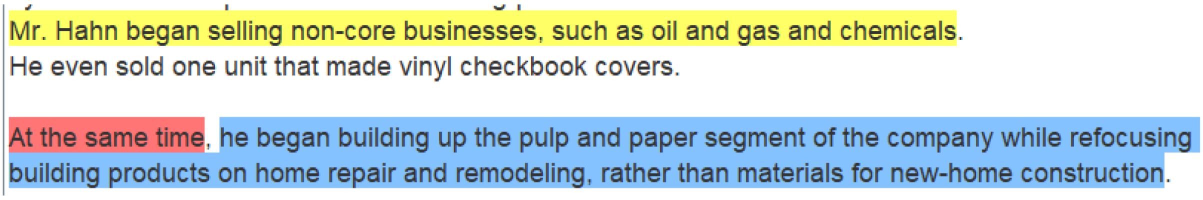
Secondly, even this kind of relation may not hold between adjacent sentences; see **NoRel** in (3).

(3) Jacobs Engineering Group Inc.’s Jacobs International unit was selected to design and build a microcomputer- systems manufacturing plant in County Kildare, Ireland, for Intel Corp. *Jacobs is an international engineering and construction concern*. **NoRel** *Total capital investment at the site could be as much as $400 million, according to Intel*. [Example borrowed from (PDTB Research Group 2008: 25)]

In the description of **NoRel** cases, importantly, there is a mention of paragraph: “NoRel indicates (of adjacent sentences within a paragraph) that no relation holds between them” (PDTB Research Group 2019: 5). This means by default that the paragraph has two aforesaid functions. It can: (1) create a single semantic block, within which sentences should be interpreted together; (2) create a boundary between semantic blocks, signaling no immediate connection between them.

Thirdly, it is assumed that DUs connected by a discourse relation do not have to immediately follow each other (as opposed to the adjacency condition in the RST).

To visualize, in particular, such cases, the PDTB uses colors, which allows you to see the boundaries of DUs. However, the PDTB corpus has a significant flaw: its annotation does not account for the hierarchy of discourse relations. See Figure 1.

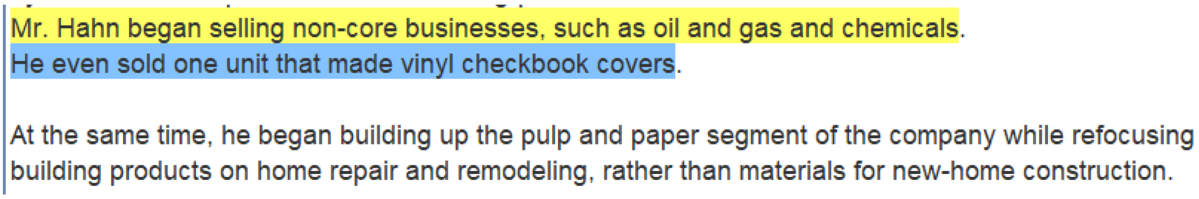


Mr. Hahn began selling non-core businesses, such as oil and gas and chemicals.
He even sold one unit that made vinyl checkbook covers.

At the same time, he began building up the pulp and paper segment of the company while refocusing building products on home repair and remodeling, rather than materials for new-home construction.

Figure 1: The annotation of the explicit “temporal synchronous”[2] relation in the PDTB

In Figure 1, we see that the temporal synchronous relation expressed by *at the same time* (highlighted in red) connects arguments 1 and 2, highlighted in yellow (argument 1) and blue (argument 2) and separated by another sentence (without highlighting). This latter, in turn, is argument 2 in the implicit conjunction relation (see Figure 2) for the same argument 1, thus included in two relations. In the annotation, *in fact* is the connective conveying the conjunction relation.



Mr. Hahn began selling non-core businesses, such as oil and gas and chemicals.
He even sold one unit that made vinyl checkbook covers.

At the same time, he began building up the pulp and paper segment of the company while refocusing building products on home repair and remodeling, rather than materials for new-home construction.

Figure 2: The annotation of the implicit conjunction relation in the PDTB

[2] The tag is used in the PDTB.

5 The Supracorpora database of hierarchical logical-semantic relations

A new linguistic resource, the Supracorpora database of hierarchical logical-semantic relations (hereupon referred to as the SDB of hierarchical LSRs), aims to enlarge the annotation capabilities provided by the Supracorpora database of connectives (the SDB of connectives[3]). Firstly, it shows the boundaries of text fragments connected by explicit and implicit relations. Secondly, it visualizes the relation hierarchy. Thus, the "вопреки ожидаемому" ("contrary to the expected state of affairs") relation expressed by the Russian connective *но* is annotated in the SDB of connectives as shown in Figure 3.

– Слушай, – сказал он Зосимову, – ты малый славный, **но** ты, *кроме* всех твоих скверных качеств, *еще и* потаскун, это я знаю, *да еще* из грязных.	**но** <"вопреки ожидаемому"> <сложное предложение> <начальная> <p CNT q> <CNT> <SuperCNT>

Figure 3: The annotated occurrence of the Russian connective *но* in the SDB of connectives

If the focus is on this relation, the left context of the fragment where the connective (CNT) *но* occurs is wider than needed (the part between the dashes is unnecessary). And in its right context, there are two more LSRs: additive propositional relation (*кроме... еще и*) and additive illocutionary relation (*да еще*). This relation hierarchy is not visible in the annotation. Its only mark is the SuperCNT tag, i.e. *но* is the "embedding connective" (in terms of the SDB), and the italicization of "embedded" connectives that fall within the scope of *но*.

Theoretically, the annotation in the SDB of hierarchical LSRs builds on the principles rather close to those of the PDTB. However, the annotation does not cover the whole text, since it is, first of all, the occurrence of a connective that gets annotated (although the SDB allows annotating implicit relations as well[4]). Therefore, there are no limitations on the length of annotated texts, which is important, as the SDB processes texts of significant length, primarily fictional, scientific, and newspaper ones. Hence, there is no need to resort to the criterion of (typo)graphic paragraph[5]. To our mind, the relations between the elements of local and global text structures are fundamentally different, and the paragraph itself occupies an in-between level, or the "meso-level" (the term is from Adam 2018). The following arguments can prove this position.

1) Regarding its internal organization, the paragraph can, in most cases, be defined through the connection between sentences at the inter-sentential level (morphology, semantics, and syntax). Yet, how paragraphs follow each other in the text is subject to discourse laws and needs to be described at the supra-sentential level in terms other than "rhetorical relations" at the local level. Such relations cannot explain how sentences merge into larger – semantically homogeneous and macrostructural – discourse units. Cf. the terms Longacre (1968) uses to classify paragraphs, resting upon major types of text passages (narrative, explanatory, expository, hortatory, procedural, and dialogue paragraphs).

2) The sentences in a paragraph do not make up a simple chain, since they are discursively heterogeneous. Of the greatest importance are the opening and final sentences, and the graphic paragraph is to emphasize this importance. The opening sentence introduces the topic that will evolve throughout the sentences grouped in the paragraph. And the incomplete line ending the paragraph signals that the previous information is detached from the subsequent information. The psycholinguistic experiments

[3] For more details about the architecture of the SDB of connectives, its interface and functionality, see Inkova 2018, Inkova & Popkova 2017. For the architecture of the SDB of hierarchical LSRs and its functional content, see Durnovo et al. 2022.

[4] An implicit relation gets annotated only if it becomes explicit in the target text, or vice versa, if a relation is explicit in the source text and becomes implicit in translation.

[5] This criterion is likely to be artificial, which is clear from the Ru-RSTreebank Annotation Manual, already quoted earlier: "If there are less than three clauses in a paragraph, we attach it, depending on the meaning, to the 'tree' of the previous or next paragraph. If the text does not show a distinct division into paragraphs, and, for example, there are many quotes from various sources (see news texts) – follow the meaning" (https://docs.google.com/document/d/1wd-sgGyIo5AQq2IPj6jWa_QmU0fUohXj48qsfVDgcBs/edit, p. 1).

showing a slowdown in reading speed in these zones prove the integrating and demarcating functions to pertain to the opening and final sentences of the paragraph (Coirier, Gaonac'h & Passerault 1996, ch. 14). The same is true for language data. If a connective marks the border between paragraphs or even chapters, it certainly connects not paragraphs or chapters, but the last and first sentences of consecutive paragraphs. Cf. (4), where *однако ж* begins the third chapter of the second part of the novel. The connective expresses the "contrary to the expected state of affairs" relation that holds between the sentence closing the last paragraph of the previous chapter ("Затем наступило беспамятство") and the sentence opening the paragraph of the third chapter ("Он не то что уж был совсем в беспамятстве"). Then there follows the description of what Raskolnikov remembered and what he forgot. The paragraph ends with his recovery.

(4) Она сошла вниз и минуты через две воротилась с водой в белой глиняной кружке; но он уже не помнил, что было дальше. Помнил только, как отхлебнул один глоток холодной воды и пролил из кружки на грудь. Затем наступило беспамятство.
III
Он, *однако ж*, не то чтоб уж был совсем в беспамятстве во всё время болезни: это было лихорадочное состояние, с бредом и полусознанием. Многое он потом припомнил. <…> Наконец он совсем пришел в себя. [Ф. М. Достоевский. Преступление и наказание (1866)]

In general, the internal structure may differ from one paragraph to another; it cannot be described in terms of "rhetorical relations" at the local level of text coherence. Cf., for example, the structure of the argumentative paragraph below.

3) While, as shown above, paragraphs help readers interpret the text, the (typo)graphical paragraph is often unnecessary. Firstly, language signals other than paragraphs also help see the topical unity (anaphoric repetitions, connectives, temporal markers, indicators of topic change, headings, subheadings, etc.). Secondly, many corpora (the Russian National Corpus, Frantext) overlook the division into paragraphs. Thirdly, different editions of the same text may have different divisions into typographical paragraphs (cf. Adam 2018, ch. 4). Fourthly, from the translation perspective, translators can change the paragraphing of the source text (Adam 2018: 66-67, Nuriev 2021: 371-384). Moreover, the concept of the paragraph itself is not universal. While most European languages distinguish between the sentence and the paragraph, languages such as, for example, Japanese, Soddo (Ethiopia), Newar (Nepal) or Godié (Ivory Coast) do not.

4) On the other hand, sometimes paragraph boundaries happen to be somewhat misleading: (typo)graphic paragraphs in the text may not coincide with semantic paragraphs mentally reconstructed by the reader. See the Anglo-Saxon opposition between *orthographic paragraph (o-paragraph)* and *semantic paragraph (s-paragraph)*. In this regard, we recall the well-known experiment of Teun van Dijk (1981: 183-190), who splits eleven graphic paragraphs of a Newsweek article (the news story type) into thirteen semantic paragraphs. The semantic paragraph is semantically coherent, which is usually described in terms of the topical or thematic unity (cf. "thematic paragraph" in Givón 1983: 8, and also, among others, Bain 1867, Albadalejo Mayordomo & Garcìa Berrio 1983, Adam 2018: 65-82, Hoey 2005, Hoffmann 1989) and is quite obvious to the readers when they move from one paragraph to another. The (typo)graphical paragraph, on the contrary, can be a mere convention determined by other factors, including those of extralinguistic nature (for example, the text layout strategies or the editorial traditions, etc.).

Without going into details on the relationship between semantic and graphic paragraphs, one can say that the division into paragraphs is rather free from strict formal or grammatical laws. So it would be a clear exaggeration to argue that each new graphic paragraph introduces a new topic breaking the referential unity and that any paragraph has only one topic. "In the paragraph, we have uncovered the specific "play" of mild-level structure, which both builds upon smaller components, and acts as a building-block of much larger object. In this looking both "below" and "above" itself, paragraph enjoys a uniquely central position in the economy of texts" (Algee-Hewitt et al. 2015: 22). To recognize the paragraph as an in-between – meso-level – text unit means to connect it with both local (inter- and super-sentential) text structure and the global one. The latter, notably, establishes a hierarchy of text passages that, in turn, largely depends on the genre and stylistic conventions and the publishing traditions.

Building on these theoretical considerations, the SDB of hierarchical LSRs approaches the question of hierarchical text structure regardless of whether the fragments connected by LSRs are in the same paragraph or different paragraphs or even chapters. For example, let us take an argumentative paragraph and compare its graphic, thematic, and discourse design. Traditionally, it should have a tripartite structure, corresponding to an argumentative passage: the first part introduces a thesis, the second part is an argument, and the third part gives a conclusion. Cf. (5), describing how the "dead civil servant" takes the overcoat from the significant personage:

(5) 1. «А! так вот ты наконец! 2. наконец я тебя того, поймал за воротник! 3. твоей-то шинели мне и нужно! 4. не похлопотал об моей, 5. да еще и распек, – 6. отдавай же теперь свою!» [Н. В. Гоголь. Шинель (1842)]

DUs 1-3 introduce a thesis, DUs 4-5 explain why the overcoat needs to be taken, and DU 6 makes a conclusion. However, it is not the graphic paragraph that reinforces the unity of the argumentative passage here (there are very few paragraphs in the story). The quotation marks enclose the character's dialogue cue and draw the boundaries of the passage.

We see a different situation in the following excerpt from Ivan Goncharov's *Oblomov*. The argumentation that we have to abridge is contained in 15 graphic paragraphs and 18 graphic sentences (labeled by **P**), i.e. the number of paragraphs and the number of sentences (differing significantly in length and complexity) are almost identical.

(6) §1. **P1** Утешься, добрая мать: твой сын вырос на русской почве – не в будничной толпе, с бюргерскими коровьими рогами, с руками, ворочающими жернова. **P2** Вблизи была Обломовка: *там* <...>! **P3** *Там* <...>; *там* <...>.
§2. **P4** *Да и* в самом Верхлёве стоит, хотя большую часть года пустой, запертой дом, но туда частенько забирается шаловливый мальчик, и *там* видит он <...>, – видит <...>; видит <...>.
§3. **P5** Он в лицах проходит <...>; читает <...> …
§4. **P6** Года в три раз этот замок вдруг наполнялся народом, <...>.
§5. **P7** Приезжали князь и княгиня с семейством: князь, <...>; княгиня <...>.
§6. **P8** Она казалась <...>.
§7. **P9** *Зато* в доме, кроме князя и княгини, был целый, такой веселый и живой мир, что Андрюша <...>.
§8. **P10** Тут были князья Пьер и Мишель, из которых первый <...>.
§9. **P11** Другой, Мишель, только лишь познакомился с Андрюшей, как <...>.
§10. **P12** Дня через три Андрей, <...>, разбил ему нос.
§11. **P13** Были еще две княжны, <...>.
§12. **P14** Была их гувернантка, <...>. **P15** Она <...>!
§13. **P16** *Потом* был немец, <...>, *потом* учитель музыки, <...>, *потом* целая шайка горничных, *наконец* стая собак и собачонок.
§14. **P17** Все это наполняло дом и деревню шумом, гамом, стуком, кликами и музыкой.
§15. **P18** С одной стороны Обломовка, с другой – княжеский замок, с широким раздольем барской жизни, встретились с немецким элементом, и не вышло из Андрея ни доброго бурша, ни даже филистера. [И. А. Гончаров. Обломов (1848-1859)]

§1. **P1** introduces the following thesis: Stolz's mother should take comfort in that he will not become one hundred percent German. **P2-P3** give the first argument: in the neighborhood, there is Oblomovka with its Russian way of life, the description of which is an unmarked enumeration, emphasized only by the parallelism of three *там*. The second argument – the princely house at Verhlyovo with its residents and customs – is introduced by the connective *да и* and separated from the first argument by the graphic paragraph. But due to its complexity, this argument is also divided into paragraphs. Being a kind of macro-argument, it describes both what Andrey sees in the house (§§2-3) and the inhabitants of the house who come once in three years (§§4-14). This description could make up a single graphic paragraph since each graphic paragraph here equals in size a graphic sentence of insignificant length. §15 concludes in support of the thesis (**P1**). Due to limitations in volume, we will not dwell much on the internal organization of the argument and will show its scheme (see Figure 4).

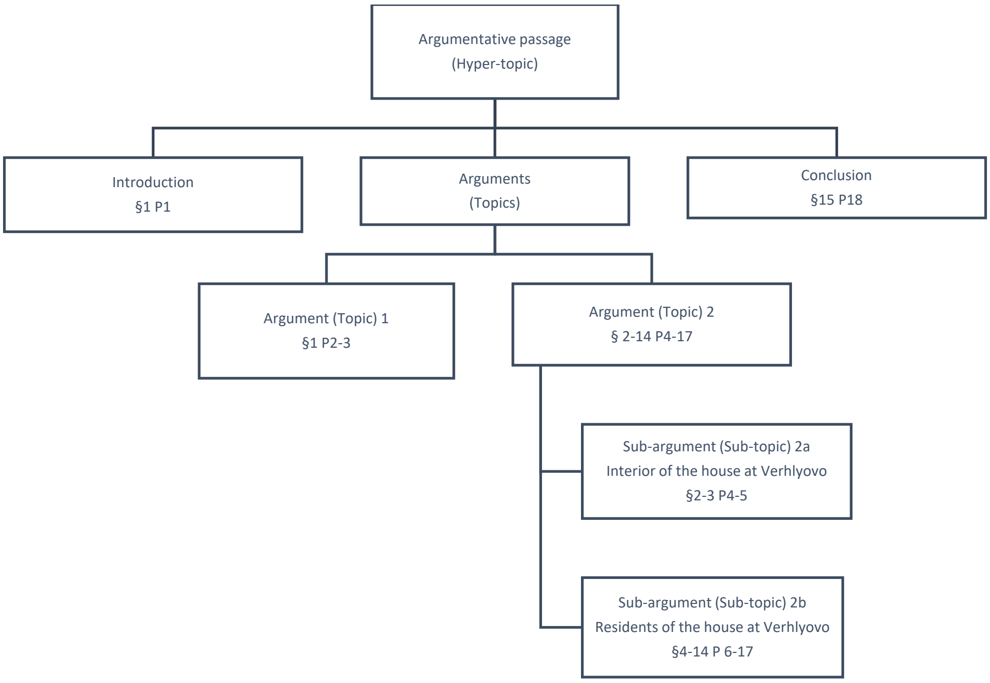


Figure 4: The topical organization of the argumentative passage (6)

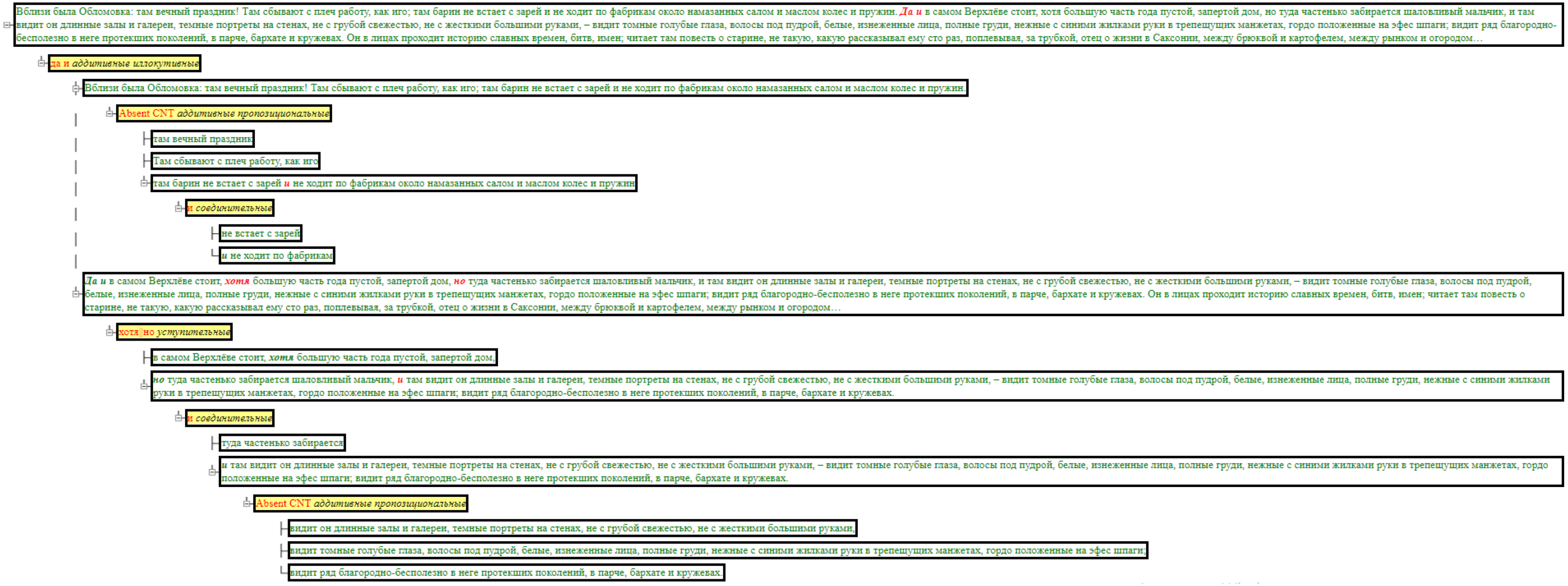
Regarding the discourse relations, the SDB of hierarchical LSRs allows building a graph for the connective *да и* (see Figure 5). It pictures the hierarchy of text fragments **P2-P3** (the left context) and **P4-P5** (the right context) falling within its scope and separated from sub-topic 2 by a strong punctuation mark – ellipsis points. The principal difference from the PDTB annotation is the possibility to visualize the relation hierarchy.

Thus, we see that various levels of text organization are not identical; they are somewhat superimposed on each other. If the scopes of diverse means contributing to text coherence coincide or overlap, it leads to greater coherence. As for global structure, its analysis should adopt a different conceptual apparatus that makes it possible: (1) to explain how the units of local structure follow one another, building global text structure, and (2) to account for stylistic and genre criteria.

6 Conclusion

To sum up the whole matter, the corpus annotation, accounting for the complex nature of various discourse elements in text organization, appears to be more thorough and theoretically justified than one that uses the same rhetorical relations to annotate units of all hierarchical levels. The former – the multi-level annotation – does not lead to oversimplification and shows more clearly how different discourse phenomena involved in the creation and interpretation of a coherent text interact between them. Cf. the corpus annotation in the ANNODIS project and the RST annotation.

As for the paragraph, since it is a unit of the level between local and global text structures, it is of little relevance when analyzing discourse phenomena at the level of local text structure. The paragraph is even less relevant, as we have seen, for delineating the boundaries of units at the level of global text structure.



Вблизи была Обломовка: там вечный праздник! Там сбывают с плеч работу, как иго; там барин не встает с зарей и не ходит по фабрикам около намазанных салом и маслом колес и пружин. *Да и* в самом Верхлёве стоит, хотя большую часть года пустой, запертой дом, но туда частенько забирается шаловливый мальчик, и там видит он длинные залы и галереи, темные портреты на стенах, не с грубой свежестью, не с жесткими большими руками, – видит томные голубые глаза, волосы под пудрой, белые, изнеженные лица, полные груди, нежные с синими жилками руки в трепещущих манжетах, гордо положенные на эфес шпаги; видит ряд благородно-бесполезно в неге протекших поколений, в парче, бархате и кружевах. Он в лицах проходит историю славных времен, битв, имен; читает там повесть о старине, не такую, какую рассказывал ему сто раз, поплевывая, за трубкой, отец о жизни в Саксонии, между брюквой и картофелем, между рынком и огородом…

да и *аддитивные иллокутивные*

Вблизи была Обломовка: там вечный праздник! Там сбывают с плеч работу, как иго; там барин не встает с зарей и не ходит по фабрикам около намазанных салом и маслом колес и пружин.

Absent CNT *аддитивные пропозициональные*

там вечный праздник

Там сбывают с плеч работу, как иго

там барин не встает с зарей *и* не ходит по фабрикам около намазанных салом и маслом колес и пружин

и *соединительные*

не встает с зарей

и не ходит по фабрикам

Да и в самом Верхлёве стоит, ***хотя*** большую часть года пустой, запертой дом, ***но*** туда частенько забирается шаловливый мальчик, и там видит он длинные залы и галереи, темные портреты на стенах, не с грубой свежестью, не с жесткими большими руками, – видит томные голубые глаза, волосы под пудрой, белые, изнеженные лица, полные груди, нежные с синими жилками руки в трепещущих манжетах, гордо положенные на эфес шпаги; видит ряд благородно-бесполезно в неге протекших поколений, в парче, бархате и кружевах. Он в лицах проходит историю славных времен, битв, имен; читает там повесть о старине, не такую, какую рассказывал ему сто раз, поплевывая, за трубкой, отец о жизни в Саксонии, между брюквой и картофелем, между рынком и огородом…

хотя||но *уступительные*

в самом Верхлёве стоит, ***хотя*** большую часть года пустой, запертой дом,

но туда частенько забирается шаловливый мальчик, *и* там видит он длинные залы и галереи, темные портреты на стенах, не с грубой свежестью, не с жесткими большими руками, – видит томные голубые глаза, волосы под пудрой, белые, изнеженные лица, полные груди, нежные с синими жилками руки в трепещущих манжетах, гордо положенные на эфес шпаги; видит ряд благородно-бесполезно в неге протекших поколений, в парче, бархате и кружевах.

и *соединительные*

туда частенько забирается

и там видит он длинные залы и галереи, темные портреты на стенах, не с грубой свежестью, не с жесткими большими руками, – видит томные голубые глаза, волосы под пудрой, белые, изнеженные лица, полные груди, нежные с синими жилками руки в трепещущих манжетах, гордо положенные на эфес шпаги; видит ряд благородно-бесполезно в неге протекших поколений, в парче, бархате и кружевах.

Absent CNT *аддитивные пропозициональные*

видит он длинные залы и галереи, темные портреты на стенах, не с грубой свежестью, не с жесткими большими руками,

видит томные голубые глаза, волосы под пудрой, белые, изнеженные лица, полные груди, нежные с синими жилками руки в трепещущих манжетах, гордо положенные на эфес шпаги;

видит ряд благородно-бесполезно в неге протекших поколений, в парче, бархате и кружевах.

Figure 5: The graph for the connective *да и* in the SDB of hierarchical LSRs

Acknowledgements

Joint Switzerland-Russia Research Preparation Grant “Annotation methodology in a supracorpora database of connectives” of the State Secretariat for Education, Research, and Innovation.

The research was carried out using the infrastructure of the Shared Research Facilities «High Performance Computing and Big Data» (CKP «Informatics») of FRC CSC RAS (Moscow).

References

[1] Adam J.-M. (2012) The emergentist model in text linguistics [Le modèle émergentiste en linguistique textuelle], Information grammaticale 134. Pp. 30–37.

[2] Adam J.-M. (2018) The paragraph: between the sentences and the text [Le paragraphe : entre phrases et texte]. Paris : Arman Colin.

[3] Albadalejo Mayordomo T., Garcìa Berrio A. (1983) Compositional structure. Macrostructures [Estructura composicional. Macroestructuras], Estudios de Lingüística. Universidad de Alicante 1. Pp. 127–180.

[4] Algee-Hewitt M., Heuser R., Moretti F. (2015) On paragraph. Scale, themes, and narrative form, Pamphlets of the Stanford literary lab, Pamphlet 10. Pp. 1–22.

[5] Asher N. (1993) Reference to Abstract Objects in Discourse. Kluwer Academic Publishers: Dordrecht.

[6] Asher N., Lascarides A. (2003) Logics of Conversation. Cambridge: Cambridge University Press.

[7] Bain A. (1867) English Composition and Rhetoric. A manual. New York: Appleton.

[8] Charolles M. (1997) The framing of discourse [L’encadrement du discours], Cahier de Recherche Linguistique 6. Pp. 1–73. URL: https://hal.archivesouvertes.fr/hal-00665849.

[9] Coirier P., Gaonac’h D., Passerault J.-M. (1996) Text psycholinguistics [Psycholinguistique textuelle]. Paris: Armand Colin.

[10] Durnovo A.A., Inkova O. Yu., Popkova N.A. (2022) Database of hierarchical logical-semantic relations: architecture [Arhitektura bazy dannyh ierarhii logiko-semanticheskih otnoshenij], Systems and Means of Informatics [Sistemy i Sredstva Informatiki] 1. Pp 114–125.

[11] Givón T. ed. (1983) Topic continuity in Discourse: A quantitative cross-language study. Amsterdam/Philadelphia: John Benjamin.

[12] Hoey M. (2005) Lexical priming: a new theory of words and language. London/New York: Routledge.

[13] Hoffmann T. (1989) Paragraphs & anaphora, Journal of Pragmatics 13. Pp. 239–250.

[14] Inkova O.Yu. (2018) The language-specificity of connectives: methods and parameters of description [Lingvospetsifitchnost’ konnektorov: metody i parametry opisaniya], Semantics of connectives: a contrastive study [Semantika konnektorov: kontrastivnoe issledovanie], O. Inkova (ed.), Moscow: TORUS PRESS. Pp. 5–23.

[15] Inkova O. (2019) Logical-semantic relations: classification problems [Logiko-semanticheskie otnosheniya: problemy klassifikatsii], O. Inkova, E. Manzotti, Text coherence: mereological logical-semantic relations [Svyaznost' teksta: mereologicheskie logiko-semanticheskie otnosheniya]. Moscow: Izdatel'skii Dom YaSK. Pp. 11–98.

[16] Inkova O.Yu. (2021) Text incoherence, or some pitfalls of automatic text processing [Nesvyaznost' teksta, ili o nekotorykh podvodnykh kamnyakh na puti avtomaticheskoj obrabotki teksta], Tomsk State University Journal of Philology [Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya] 74. In press.

[17] Inkova O., Popkova N. (2017). Statistical data as information source for linguistic analysis of Russian connectors, Informatics and applications [Informatika i ee primeneniya] 11(3). Pp. 123–131.

[18] Longacre R. E. (1968) Discourse, paragraph and sentence structure in selected Philippine languages, Vol 2. Sentence structure. Santa Ana: Summer Institute of Linguistics.

[19] Longacre R. E. (1979) The Paragraph as Grammatical Unit, T. Givón (ed.) Syntax and semantics. Discourse and Syntax, vol. 12. New York: Academic Press. Pp. 115–134.

[20] Longacre R. E. (1996) The Grammar of Discourse. 2nd ed. (1st ed. 1983). New York/London: Plenum Press.

[21] Mann W., Thompson S. (1988) Rhetorical structure theory: Towards a functional theory of text organization, Text 8. Pp. 243–281.

[22] Nuriev V. (2021) Literary translation through the lens of language experiment (syntactic aspect) [Khudozhestvennyj perevod skvoz' prizmu jazykovogo jeksperimenta (sintaksicheskij aspect)]. D.Sc. thesis. Moscow: The Military University of the Defense Ministry of the Russian Federation. Pp. 371–384.

[23] PDTB Research Group (2008) The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report IRCS-08-01. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania; URL: https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb- annotation- manual.pdf.

[24] PDTB Research Group (2019) The Penn Discourse Treebank 3.0 Annotation Manual; URL: https://doi.org/10.35111/qebf-gk47.

[25] Pike K. L. (1982) Linguistic Concepts: An Introduction to Tagmemics. Lincoln and London: University of Nebraska Press.

[26] Prévot L., Vieu L., Asher N. (2009) A more precise formalization for a less confused annotation: the Elaboration relation [Une formalisation plus précise pour une annotation moins confuse : la relation d'élaboration d'entité], Journal of French Language Studies 19. Pp. 207–228.

[27] van Dijk T. A. (1981) Episodes as units of discourse analysis, D. Tannen (ed.) Analysing Discourse: Text and Talk. Georgetown: Georgetown University Press. Pp. 177–195.

[28] Vieu L., Bras M., Asher N., Aurnague M. (2005) Locating adverbials in discourse, Journal of French Language Studies 15. Pp. 173–193.

[29] Webber B. L., Egg M., Kordoni V. (2012) Discourse Structure and Language Technology, Natural Language Engineering 18 (4). Pp. 437–490.