# The Database of Constructions with Lexical Repetitions "RepLeCon" and Inter-Annotator Agreement


**Olga Blinova and Elena Vilinbakhova**

## 1 Introduction

This chapter discusses some issues emerged in the development of the "RepLeCon" database. The "RepLeCon" covers a variety of Russian constructions with lexical repetitions and its equivalents in English, German, French, Italian, and Spanish. Being the product of the project on lexical repetitions carried out at St. Petersburg State university since 2019, the "RepLeCon" aims to provide a resource for theoretical linguistic studies and its practical applications. Possible directions of research, which can benefit from the use of the "RepLeCon," include the analysis of structural and interpretive properties of Russian constructions with repetitions, as well as their use in translations to and from European languages.

More specific research questions are concerned with the types of passage, that is, argumentative, narrative, or instructive, in which the particular construction is employed, evaluation of the referent conveyed by it use or its rhetorical relations with the preceding and following discourse fragments, a. m. o.

The "RepLeCon" is composed of parallel units, that is, aligned sentences in Russian and one of the European languages listed above, extracted from parallel corpora, such as Russian National Corpus (henceforth – RNC) and Open Subtitiles on the platform Sketch Engine. The data from RNC represents mostly literary texts and their translations, while Open Subtitiles provides subtitle pairs sorted by movie and release year and includes dialectal expressions and slang.

O. Blinova (✉)
Saint Petersburg State University, St. Petersburg, Russia

National Research University Higher School of Economics, St. Petersburg, Russia
e-mail: o.blinova@spbu.ru

E. Vilinbakhova
Saint Petersburg State University, St. Petersburg, Russia
e-mail: e.vilinbakhova@spbu.ru



R. Bolgov et al. (eds.), *Digital Geography*, Springer Geography,
https://doi.org/10.1007/978-3-031-50609-3_8

The initial data structure has been developed in *.xls spreadsheets format along with instructions for their compilation. For corpus data extraction, we used a series of queries designed for each construction. Since the searches are accomplished on bilingual data, all qualified Russian repetitions found in the original Russian texts or in translations from European languages along with their foreign equivalents have been taken into account. As a result, we obtained from parallel corpora pairs like (1), in which Russian comparative tautology employed by the Russian author Maxim Gorky is translated as "people are like people" into English, and (2), where the original fragment by John Galsworthy is in English and the use of the Russian conditional tautology is a decision of the translator.

(1) *Poroju u materi javljalosʹ nedovolʹstvo synom, ona dumala: "Vse* **ljudi – kak ljudi***, a on – kak monax."* [Maxim Gorky. Mother (1906)].
"Every now and then she felt a certain dissatisfaction with him, and she thought: 'All **people are like people**, and he is like a monk.'" (translated by D.J. Hogarth, 1921)

(2) *But, as Dartie said:* **There was nothing like pluck!** [John Galsworthy. The Man of Property (1906)].
"No, kak govoril Darti, už **esli povezet, to povezet**!" (translated by N. Volžina, 1946).

Next, the *.xls spreadsheets are imported into an OpenOffice database. The user can operate queries in several modes including SQL mode. Data retrieving from 19 base tables is now available. Each table describes specific type of construction with lexical repetition (see below for details). The database allows one to make queries using main fields such as "source context," "target context," "corpus," "subcorpus," "L1 (source language)," "L2 (target language)," "N of repetitions," and many others.

At present, the project includes 19 Russian constructions, including various types of tautologies, such as equative (NP-Nom Cop NP-Nom), disjunctive (VP ili VP), conditional (esli VP, to VP), or comparative (NP-Nom tak NP-Nom), reduplicated VPs with or without negation (VP-Inf (ne) VP), constructions with temporal nouns conjoined by a preposition (NP-Nom za NP-Instr, NP-Nom v NP-Acc, etc.).

The following five general categories were introduced in the "RepLeCon" database:

- Context description
- Structural features
- Semantic features
- Pragmatic features
- The degree of equivalence of translations of constructions

Context description consists of following fields: (i) full and (ii) short context in both source and target languages, (iii) corpus metadata (e.g., the title and author) for both languages, (iv) the subcorpus used, (v) its size in tokens, and (vi) language of the original text and its translation.

The description of structure of constructions includes (i) repeated material (lexemes), (ii) number of repetitions in a fragment, (iii) part of speech (PoS henceforth)

of the head in a repeated phrase, (iv) grammatical tags (including PoS) for all elements, as given in the corpus for both original and translation fragments, (v) extension of repeated elements, and (vi) their codependent elements.

Semantic description covers (i) the degree of semantic similarity between repeated elements, (ii) type of information to which the construction makes reference, (iii) referential status indicators (if any available), and (iv) idiomatic status of the construction as defined in [1].

The description of the pragmatic features of constructions focuses on (i) speech event structure, modus (oral, written, etc.), (ii) type of passage (narrative, descriptive, etc.), (iii) rhetorical relations of construction with preceding and following discourse, (iv) markers of such relations, if available, and (v) evaluation of the referent, conveyed by a construction.

Finally, the characteristics of the translations of constructions include one basic field: source-target correspondence, that is, whether the translation is in formal and/or functional correspondence with the original construction. For instance, in (1) above the aligned fragments in the original – *ljudi kak ljudi* – and in translation – "people are like people" – exhibit both formal similarity and functional equivalence. On the contrary, in (2) the Russian conditional tautology *esli povezet, to povezet* chosen by the Russian translator, is different from the original fragment "There was nothing like pluck!".

All our data is manually annotated in order to identify values of the parameters listed above. While the identification of values of the parameters describing context, structure, and semantics does not require metalinguistic judgement and does not trigger subjective interpretations of the annotators, the parameters describing pragmatics and characteristics of translation may trigger subjective judgements. Hence, for the latter type of parameters, we established the procedure in which two independent annotators rated the data based on pre-established annotation guidelines.

In this study, we discuss the results of the annotation of rhetorical relations for equative tautologies, such as *Friends are friends*, analyze the factors that influence the annotators' inter-annotator agreement, and provide explanation for the observed phenomena.

## 2 Theoretical Background

As mentioned above, in the present work, we deal with the subset of pragmatic features, namely, rhetorical relations, applied to equative tautologies. In this section, we provide an overview of the necessary notions and justification for our choice of data.

The analysis of rhetorical relations was first presented in Mann and Thompson's Rhetorical Structure Theory (henceforth RST) [2]. Their account characterizes different patterns found in discourse and the ways they are connected to each other. In the subsequent works, RST was modified in a number of aspects, see [3–8], and for

our study, we take the approaches of Jasinskaja and Karagjosova [8] for monologues and Asher and Lascarides [6] for dialogues.

In [8], the authors, on the one hand, keep intact the main tenets of RST and, on the other hand, restrict the list of RR to *Contrast*, *Elaboration*, *Explanation*, *Narration*, *Parallel*, and *Result*, based on general principles of cause and effect, contiguity, and resemblance. Their classification is checked against naturally occurring data, as attested in [9, 10]. Later in the course of work, we decided to add the rhetorical relation of *Condition* from the original version of RST.

For dialogues, the classification of Asher and Lascarides in [6] includes the relations of *Question-Answer Pair*, *Indirect Question-Answer Pair*, *Partial Question-Answer Pair*, *Denial*, and *Acknowledgement*. Let us look at the RR in more detail.

First, we consider RR in monologues, in which both discourse fragments involved in a relation are uttered by the same speaker.

1. **Condition**. *Condition* recognises how the realization of one situation depends on the realization of another situation [2, p. 275]. For instance, in (3) acknowledgment of the general claim that every woman is special is a pre-condition for the acceptance of feelings of particular woman Dee Dee.

(3) *You said that every woman is special in her own way. That there's no two alike. So if that's true … then* ***Dee Dee is Dee Dee****. Let her be who she is*. [OpenSubtitles2011, Dr. T & the Women, 2000].
"Esli èto pravda, to … **Didi est′ Didi**."[1]

2. **Elaboration**. In this RR, both discourse fragments refer to the same situation, but one of them is more general, and another is more specific. The specific unit usually provides additional information to ensure the better understanding of the general unit. In (4), the tautological utterance is clarified by the follow-up statement, specifying the particular aspect of Tom's character.

(4) ***Tom's always gonna be Tom****. He's like a guided missile, locks on … That's it*. [OpenSubtitles2011, Street Kings, 2008].
"**Tom – èto Tom**. Nacelitsja, kak raketa, i vse."

3. **Explanation**. In contrast to the previous RR of *Elaboration*, in *Explanation* two discourse fragments refer to different situations framed as cause and effect. They are given in the reverse order, since the cause explains the effect. This is the case of (5), where the tautology *Policy is policy* serves as an explanation for the speaker's decision.

(5) *I'm afraid I have no choice.* ***Policy is policy***. [OpenSubtitles2011, Lexx. Season 2, Episode 2, 1997–2002].
"Bojus′, u menja net vybora. **Pravila est′ pravila**."

4. **Contrast**. *Contrast* requires both similarity and dissimilarity of its alternatives [11], and the last contrasted element is presented as more important and relevant than others, see [6].

[1] For the sake of brevity, for Russian translations, we provide a minimal context, which includes an equative tautology and a discourse fragment involved in the relevant rhetorical relation.

(6) *Corporal Henderson … I don't mean to leave you short-handed, but* ***orders are orders***. [OpenSubtitles2011, Saving Private Ryan, 1998].

"Kapral Xenderson … ne xotelosʹ by vas ostavljatʹ bez bojca, no **prikaz estʹ prikaz**."

5. **Narration**. In *Narration*, two discourse fragments refer to different situations, which are adjacent in space and time. For instance, in (7) the speaker tries to persuade the interlocutor that she is not the murderer and tells him that after meeting another person, she forgave and forgot the previous partner, the latter message expressed by a conventional tautology.

(7) *– Were you upset when he broke it off with you?*

*– What?*

*– Well, like you said, you… you didn't have any legal options.*

*– Not upset enough to kill. I moved on. I met someone after a couple months, and* ***bygones are bygones****, right?* [OpenSubtitles2011, unidentified source].

"Ja vstretil koe- kogo čerez neskolʹko mesjacev, i **prošloe – èto prošloe**, pravilʹno?"

6. **Parallel**. This RR emphasizes similarity and parallelism between its parts. For instance, in (8) both tautologies *Family is family* and *Blood is blood* are used as arguments for the position of the speaker that one should stand for one's family unconditionally, in an automatic way.

(8) *After what I've done for you, it should be automatic.* ***Family is family***. ***Blood is blood***. *You don't ask questions. You protect your own*. [OpenSubtitles2011, Cassandra's dream, 2007].

"**Semʹja estʹ semʹja**! **Krovʹ estʹ krovʹ**!"

7. **Result**. *Result* is based on the cause-and-effect principle: the event described in first discourse fragment causes the event presented in the second part. In (9) the tautology *Facts are facts* evoking some mutually known events is used as a reasoning for the interlocutor's right to speak to her father in a disrespectful way.

(9) ***Facts are facts****, so listen up. You may be my father, but I am never going to be your daughter. You got that?* [OpenSubtitles2016, The Vampire Diaries, 2009–2017].

"**Fakty estʹ fakty**, tak čto slušaj."

Now let us turn to RR in dialogues, in which discourse fragments involved in a relation are uttered by two distinct speakers.

1. **Question-Answer Pair**. This RR suggests that the contribution is a direct response to the question [6, p. 316], as exemplified by (10).

(10) *– What's your thesis?*

*– All I'm saying is* ***people are people****. We do what we do…* [OpenSubtitles2011, United States of Tara, 2010].

"– I kakov vaš tezis?

– Prosto **ljudi estʹ ljudi**."

2. **Indirect Question-Answer Pair**. This RR holds when the contribution is not a direct response to the question, but the questioner can infer the necessary information from it [6, p. 316]. In (11) the interlocutor utters as a response a

so-called tautology of value [12], which implies that an entity should be appreciated by virtue of belonging to a particular category, that is, "a ship is valuable because it is a ship."

(11) – *With all respect, doctor, I'm counting on Excelsior.*
– *Excelsior? Why would you want that bucket of bolts?*
– ***A ship is a ship***. [OpenSubtitles2011, Star Trek IV: The Voyage Home, 1986].
"– Èksel′sior? Začem vam èta, prosti gospodi, gruda železa?
– **Korabl′ est′ korabl′**."

3. **Partial Question-Answer Pair**. *PQAP* implies that the contribution may rule out some true answers, but is not sufficiently informative that the questioner can compute a direct answer from it [6, p. 319]. For instance, in (12) the speaker's response, while relevant to her daughter's question, is too elusive and does not admit to infer a clear message about her understanding of the notion of spirit.

(12) – *Mama, do you know what a spirit is? You don't know and I do.*
– ***A spirit is a spirit***. [OpenSubtitles2011, Mama, 2013].
"– Mama, ty znaeš′, čto takoe dux? <…>
– **Dux – èto dux.**"

4. **Acknowledgment**. *Acknowledgment* is used to express the acceptance of the previous utterance. This is the case of (13), when the speaker uses a tautology to indicate that the interlocutor's statement describing the behaviour of Hank is fully consistent with her expectations.

(13) – *EMT's pump him full of morphine?*
– *He refused it. Said it would ruin his sobriety.*
– *Yeah, that's just* ***Hank being Hank***. [OpenSubtitles2011, Terriers, 2010].
"– On otkazalsja. <…>
– Nu da, **Xènk kak vsegda Xènk**."

5. **Correction**. This RR holds when the contribution is aimed at correcting the previous utterance, as in (14). Here the speaker cannot deny the fact that she indeed stopped at a red light, but she corrects the previous interlocutor's scornful comparison to a more favourable alternative.

(14) – *Chasing kidnappers and you stop at a red light. Like an old lady!*
– *Well,* ***the law's the law***. [OpenSubtitles2011, Remote Control, 1988]
"– Presleduja poxititelej, ty ostanavlivaeš′sja na krasnyj. Kak staruška kakaja.
– Nu, **zakon est′ zakon**."

6. **Denial**. *Denial* is employed to rebut the previous contribution. For instance, in (15) the speaker objects to the description of theft as liberating funds with a so-called deep tautology, which indicates that "being an A does not admit of degrees" [13, p. 287].

(15) – *They stole?*
– *They liberated funds.*
– *Theft is theft. There is no grey area.* [OpenSubtitles2011, Black Rain, 1989]
"– Oni vydelili sebe fondy.
– **Kraža est′ kraža**."

The cases when there is no relation with the preceding or the following discourse fragment, that is, the construction is used in the absolute initial or final position, are marked as N/A (not applicable).

Further, to our data we apply the distinction between multinuclear vs. mononuclear (nucleus-satellite) relations based on the status of two discourse items with respect to each other. If both items are equal, they are regarded as two nuclei; this is the case for *Contrast*, *Parallel*, *Narration*, and *Result*. If one item is dependent on the other, they are regarded as satellite and nucleus; this is the case for *Condition*, *Elaboration,* and *Explanation*. For the RR in dialogues, the reactive utterances are marked as satellites.

Besides, we record linguistic expressions that serve as markers for the identified RR.

In our study, we look at the rhetorical relations of constructions *X cop X*, labeled in the literature as equative tautologies and discussed with regard to their argumentative force due to the literal truthfulness [14–16].

## 3 Improving Annotation Validity Through Agreement Measurement

### *3.1 Inter-Annotator Agreement (IAA) Methods*

Computational linguistics traditionally uses the practice of annotation of single item by several people and then comparing the annotations. The practice is applied to improve annotation guidelines, to identify phenomena that are hard to annotate, to assess the range of possible interpretations, and to improve annotation validity, see for example, [17].

Comparison of annotator judgments can be performed using calculation of agreement indices. The general scheme can be described as follows. Multiple annotators independently perform linguistic markup by following the guideline. Then, using statistical measures, the consistency of the annotation is assessed, that is, some coefficient of agreement is calculated. If the coefficient is below a certain threshold value, the guideline is subject to revision. After the improvement of the annotation scheme, the markup is performed again. After a certain number of iterations, implying guideline revision and re-annotation procedure, the guideline has to become clearer, and the annotation process should give reproducible results.

The calculation and assessment of inter-annotator agreement is used in particular in corpus pragmatics and in the building of discourse-annotated corpora, for instance, for creation of speech-act annotated corpora, prosodically annotated corpora, coreference-annotated corpora, discourse-tagged corpora in the framework of Rhetorical Structure Theory, see for example, [18–20], and many others.

As stated in [21, p. 700], "Inter-annotator (or inter-coder) agreement has become the quasi-standard procedure for testing the accuracy of manual annotations. This

process is based on the assumption that if multiple coders agree in their coding decisions of the same material, we can be certain that – at least for this set of data and this set of coders – annotations are free of unsystematic and distorting variations."

According to [22, 23], the main sources of disagreement between annotators are:

- gaps in the annotation guide, causing differences in comprehending instructions,
- difficulty in interpreting a markable item, existence of debatable cases, or several possible interpretations,
- annotators' carelessness and their openness to distractions.

On the whole, agreement values are influenced by research domain, number of categories in a coding scheme, number of annotators, the presence or absence of annotators' training and its intensity, the annotation purpose, and the method used for the calculation of agreement index, see [21, 24].

### *3.2 Agreement Measures*

The commonly used and simplest measure of agreement is **percentage agreement**. To calculate the measure value, it is enough to divide the number of items on which two annotators agree by the total number of items. The measure is criticized as biased, since it lacks the correction for chance agreement [25, 26].

When the number of annotators is more than two, the following methods for calculating percentage agreement values are used [21, p. 705]:

I. **pairwise method**, involving calculation of the average agreement across all pairs of annotators,
II. **majority method**, when agreement is assigned if a certain proportion of annotators (two out of three, three out of five, etc.) label certain item with same category,
III. **consensus method**, when agreement is assigned if all coders make consentient decisions on an item.

As suggested in (Ibid.), under otherwise equal conditions, agreement should be calculated as consensus rather than as a (average) pairwise or majority agreement.

The percentage agreement measures are traditionally opposed to **more robust chance-corrected agreement measures**. The best-known chance-corrected coefficients for measuring agreement between two annotators (Bennett, Alpert and Goldstein's S, Scott's pi, and Cohen's kappa) use the Formula (1), where $A_o$ is observed agreement, $A_e$ is expected agreement, see [26], and many others.

$$\text{Inter} - \text{annotator agreement} = A_o - A_e \,/\, 1 - A_e \tag{1}$$

In case the number of annotators is more than two, generalized versions of the coefficients (multi-pi, generalized Scott's pi, and multi-kappa, generalized Cohen's kappa) are used.

"K coefficient of agreement" (K) can be considered as a variant of Cohen's kappa (Cohen 1960), or Fleiss' kappa, see the discussion in [27], that is a generalization of pi rather than kappa [28]. As stated in [26], "K <a variant of Cohen's kappa> quickly became the de facto standard for measuring agreement in computational linguistics."

Meanwhile, there are limitations in the use of both pi and kappa coefficients due to the fact that "all disagreements are treated equally"; however, **when marking up semantic and pragmatic phenomena, taking into account the degree of disagreement between annotators becomes essential** (Ibid). Accordingly, for some research tasks, it is desirable to use coefficients, capable to differentiate between types of disagreements.

Among such coefficients are:

- Krippendorff's alpha,
- Weighted kappa κw (another member of the kappa family measures).

The calculation of these coefficients (called weighted coefficients) implies the determination of the disagreement and can be illustrated using Formula (2), where $D_o$ is observed disagreement and $D_e$ is expected disagreement.

$$\text{Inter} - \text{annotator agreement} = 1 - D_o / D_e \tag{2}$$

In [29] is also proposed weighted coefficient **beta**, which can be considered as a generalization of weighted kappa ($\boldsymbol{\kappa_w}$) to multiple annotators.

So, we briefly reviewed some basic agreement measures. Several parameters directly affect the choice of a particular measure, including the number of annotators and the type of data.

### *3.3 The Proposed IAA Calculation Scheme*

To make the annotations in "RepLeCon" maximally reliable, inter-annotator agreement will be used in biased cases, which are some structural, semantic, and pragmatic annotation. Our goal is to improve the validity of annotated data.

We use the following data evaluation scheme. At the preliminary stage, the fields of the "RepLeCon" database were divided into "technical" fields (not involving judgment and analysis, but requiring technical work to copy metadata and other "prepared" information from the source corpora) and "interpretable" ones.

Then all "interpretable" fields were annotated by two coders with year-long annotation experience; then the data was analyzed using agreement measures.

### 3.4 The Obtained IAA Calculation Results

In this chapter, we present an analysis of the annotation results of 130 constructions with lexical repetitions. We examined annotation consistency in the fields where information about the discursive and pragmatic features of constructions with lexical repetitions is encoded, more precisely, about the rhetorical relations of constructions with the preceding and following discourse.

The fields in which cases of inconsistent markup are observed and the number of cases where consistent markup is observed are presented in Table 1.

We used the R "irrCAC" package [30] to compute agreement coefficients. The coefficients were calculated for each of the four fields ("rr-prec," "role-prec," "rr-foll," and "role-foll") separately. The computing is based on a raw dataset with the answers of two annotators (A1 and A2), see Table 2 as an example, which shows the results of the markup of ten constructions with lexical repetitions.

Table 3 shows values of the coefficients (see also Fig. 1 below). We calculated six coefficients of agreement: Percent Agreement, Gwet's AC (AC1), Fleiss' Kappa, Krippendorff's alpha, Conger's Kappa, and Brennan and Prediger's agreement coefficient.

In addition, the rows of the table for each coefficient present: pa (the percent agreement), pe (the percent chance agreement), coeff.val (the estimated value of an agreement coefficient), coeff.se (the agreement coefficient standard error), and conf.int (the confidence interval), see [31] for details.

Percent agreement exceeds 90% in all cases, except for the examples of marking the role of the construction in rhetorical relation with the following discourse fragment (rr-foll). Gwet's AC (AC1) values are in the range [0.78; 0.98]; Fleiss' Kappa, Krippendorff's Alpha, Conger's Kappa values are in the range [0.73; 0.97]; Brennan and Prediger's agreement coefficient (AC) values are in the range [0.78; 0.98].

If we interpret the observed data by applying coefficient thresholds (as suggested, for example, by Landis and Koch in [32]), then we can assume that (according to the benchmark scale) values falling in the range [0.81; 1.00] could considered as "almost perfect"; values falling in the range [0.61; 0.80] could considered as "substantial."

Thus, it is possible to formulate the first conclusion, according to which annotators perform markup in a consistent way, so the instruction is compiled quite successfully.

Meanwhile, we remember that annotators could give a consistent score by chance. The probability of a random response depends among other things on the

**Table 1** Number of agreed responses

| Field | *N* of consistent annotation |
|---|---|
| role-prec | 128 |
| role-foll | 127 |
| rr-prec | 126 |
| rr-foll | 104 |

**Table 2** Dataset example

| rr-prec | |
|---|---|
| A1 | A2 |
| PQAP | PQAP |
| PQAP | PQAP |
| E | E |
| P | P |
| C | C |
| DEN | DEN |
| PQAP | PQAP |
| E | E |
| X | X |
| DEN | DEN |

**Table 3** Agreement coefficients

| coeff.name | pa | pe | coeff.val | coeff.se | conf.int |
|---|---|---|---|---|---|
| *rr-prec* | | | | | |
| Percent Agreement | 0.96154 | 0.00000 | 0.96154 | 0.01693 | (0.928,0.995) |
| AC1 | 0.96154 | 0.07169 | 0.95857 | 0.01824 | (0.922,0.995) |
| Fleiss' Kappa | 0.96154 | 0.13970 | 0.95529 | 0.01964 | (0.916,0.994) |
| Krippendorff's Alpha | 0.96169 | 0.13970 | 0.95546 | 0.01964 | (0.917,0.994) |
| Conger's Kappa | 0.96154 | 0.13935 | 0.95531 | 0.01962 | (0.916,0.994) |
| Brennan-Prediger's AC | 0.96154 | 0.07692 | 0.95833 | 0.01834 | (0.922,0.995) |
| *rr-foll* | | | | | |
| Percent Agreement | 0.80000 | 0.00000 | 0.80000 | 0.03522 | (0.73,0.87) |
| AC1 | 0.80000 | 0.09101 | 0.77998 | 0.03876 | (0.703,0.857) |
| Fleiss' Kappa | 0.80000 | 0.18095 | 0.75582 | 0.04315 | (0.67,0.841) |
| Krippendorff's Alpha | 0.80077 | 0.18095 | 0.75675 | 0.04315 | (0.671,0.842) |
| Conger's Kappa | 0.80000 | 0.17793 | 0.75671 | 0.04268 | (0.672,0.841) |
| Brennan-Prediger's AC | 0.80000 | 0.10000 | 0.77778 | 0.03913 | (0.7,0.855) |
| *role-prec* | | | | | |
| Percent Agreement | 0.98462 | 0.00000 | 0.98462 | 0.01084 | (0.963,1) |
| AC1 | 0.98462 | 0.22643 | 0.98011 | 0.01407 | (0.952,1) |
| Fleiss' Kappa | 0.98462 | 0.54713 | 0.96603 | 0.02366 | (0.919,1) |
| Krippendorff's Alpha | 0.98467 | 0.54713 | 0.96616 | 0.02366 | (0.919,1) |
| Conger's Kappa | 0.98462 | 0.54710 | 0.96603 | 0.02366 | (0.919,1) |
| Brennan-Prediger's AC | 0.98462 | 0.33333 | 0.97692 | 0.01625 | (0.945,1) |
| *role-foll* | | | | | |
| Percent Agreement | 0.89231 | 0.00000 | 0.89231 | 0.02729 | (0.838,0.946) |
| AC1 | 0.89231 | 0.13405 | 0.87564 | 0.03220 | (0.812,0.939) |
| Fleiss' Kappa | 0.89231 | 0.59784 | 0.73222 | 0.06581 | (0.602,0.862) |
| Krippendorff's Alpha | 0.89272 | 0.59784 | 0.73325 | 0.06581 | (0.603,0.863) |
| Conger's Kappa | 0.89231 | 0.59391 | 0.73481 | 0.06394 | (0.608,0.861) |
| Brennan-Prediger's AC | 0.89231 | 0.25000 | 0.85641 | 0.03639 | (0.784,0.928) |

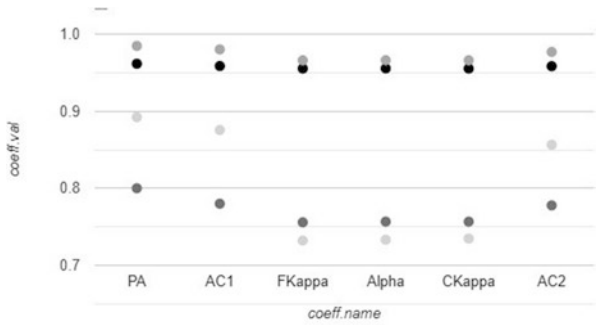


**Fig. 1** Agreement coefficients

**Table 4** Number of categories in the coding scheme

| Fields | Categories |
|---|---|
| role-prec & role-foll | 2 |
| rr-prec & rr-foll (dialogues) | 6 |
| rr-prec & rr-foll (monologues) | 7 |

number of categories in a coding scheme, that is, on the number of possible values of a particular coded variable. If the number of categories in a coding scheme grows, the agreement between annotators decreases, see for example [21]. Note also that the relationship of inverse proportionality between the number of categories and consistency is partly explained by the statistical characteristics of agreement metrics [26].

The number of categories in the relevant fields of the "RepLeCon" database is shown in Table 4 (see also Sect. 2 above for details on categories).

The average percentage agreement for core vs satellite role markup (with respect to both the preceding and the following discourse) is 94%. That is, annotators agree on what is core and what is satellite. The average percentage agreement for marking rhetorical relations (rr) is 88%.

Because of the different number of categories possible in coding the fields with constructions in monological and dialogical fragments, we need to take into account the differences in the type of the speech event. The total number of constructions belonging to the monological fragments is 13. The total number of constructions belonging to the dialogue fragments is 117. The values of the percent agreement for the possible combinations of the type of speech event and the type of rhetorical relation are shown in Table 5.

So, we can expect that the number of consistent responses for the "rhetorical relation type" (rr) and for the "role of the construction in the rhetorical relation" (role) will be statistically significantly different. In addition, we can expect that the number of consistent responses of dialogical fragments is statistically significantly lower than the consistency of the markup of monological ones (since more

**Table 5** Number of consistent annotations (taking into account the type of speech event)

| Fields | *N* of CA | Total | Percent agreement |
|---|---|---|---|
| rr-prec (monologues) | 12 | 13 | 92.31 |
| rr-foll (monologues) | 10 | 13 | 76.92 |
| rr-prec (dialogues) | 114 | 117 | 97.44 |
| rr-foll (dialogues) | 94 | 117 | 80.34 |

categories are provided for dialogical fragments in the markup scheme). Finally, the data allow us to suggest that annotators are worse at coding the type of rhetorical relationship of the analyzed fragment to the following discourse (when compared to the preceding one).

To test the expectations we formulated, we used Fisher's exact criterion to compare the number of agreed and unagreed responses. The results are as follows. First, the differences in the number of agreed and unagreed responses in the "role" and "rr" fields are statistically significant ($p < 0.001$). Second, the differences in the total number of agreed and unagreed responses when comparing dialogical and monological fragments are statistically insignificant ($p = 0.383$). Differences in "rr" field annotation results ("rhetorical relationship type," see Table 3) when comparing monological and dialogical contexts are also insignificant ($p = 0.517$). Finally, a comparison of the number of agreed and unagreed responses, categorized by "foll" and "prec," showed statistically significant differences ($p < 0.001$).

Let us formulate the following conclusions:

- Annotators consistently determine what is the nucleus and what is the satellite.
- The influence of the number of categories in the annotation scheme on consistency is confirmed by our data (annotators handle the markup in the "role" field significantly better than in the "rr" field).
- Annotators assign tags of rhetorical relations with the preceding discourse (compared to the following one) in a more consistent way.

For a more detailed analysis, it makes sense to use information about the presence of lexical markers of rhetorical relations. To do this, let us turn to the database fields "marked-prec" (marker of rhetorical relation in which a construction consists with a preceding discourse fragment; linguistic expression or, in the absence of the marker, NO value is indicated) and "marked-foll" (marker of rhetorical relation in which a construction consists with a following discourse fragment; linguistic expression or, in the absence of the marker, NO value is indicated). Examples of observable rhetorical relation markers are: *potomu čto* "because," *no* "but," *nu* "well," *prosto* "just," *vsë-taki* "still," as well as *i kogda* "and when," *xotja* "though," *tak čto* "so what," *krome togo* "besides," etc.

Our hypothesis was that the presence of explicit markers in the analyzed context simplifies the rhetorical relations annotation, respectively, and there will be significantly fewer cases of inconsistent annotation in contexts with such markers.

However, among the annotated contexts, we see 46 in which there are markers of rhetorical relations with the preceding discourse and 46 in which there are markers

**Table 6** Number of consistent annotations (taking into account lexical markers of rhetorical relations)

| Field | *N* of CA | *N* of markers |
|---|---|---|
| rr-prec | 126 | 46 |
| rr-foll | 104 | 46 |

of rhetorical relations with the following discourse. The corresponding data for the "rr" fields are presented in Table 6. We see that the hypothesis for the observed data should be rejected, since the number of markers matches, and it is meaningless to calculate, for example, correlation coefficients between the number of cases of consistent coding and the presence of markers.

## 4 Conclusion

In this chapter, we introduced the "RepLeCon" – a database of Russian constructions with lexical repetitions and their equivalents in English, German, French, Italian, and Spanish which deals with their structural, semantic, and pragmatic features. Since the "RepLeCon" is an annotated resource, here we discussed the annotation validity through agreement measurement. We took a sample of equative tautologies, such as *Family is family*, annotated with respect to their rhetorical relations and their role (i.e., the nucleus or the satellite) in these relations.

Based on the data received, our study revealed, first, that raters show either an almost perfect or substantial agreement, in terms of [32], and hence, the annotation instruction is well compiled and does not need significant revision.

Next, it turned out that raters assign the particular rhetorical relations significantly less consistently than the role of tautologies in the rhetorical relation. This suggests that the number of categories in the annotation scheme has an impact on consistency: since there are only two possibilities of the role of constructions compared to a longer list of the rhetorical relations, annotators' decision for the former scheme requires less effort.

Finally, the rhetorical relations with the following discourse fragment are annotated in a less consistent way than the rhetorical relations with the preceding fragment. To provide an explanation to this fact, we examined the hypothesis that the presence of explicit markers in the analyzed context simplifies the rhetorical relations annotation. While this hypothesis is not borne out on the present material, we expect to explore it further on a broader range of data.

More directions for further research include the analysis of other semantic and pragmatic features represented in the "RepLeCon." In particular, the types of passage in which the constructions are employed involve the subjectivity of the judgments and, therefore, are worth being examined in the future.

**Acknowledgments** The first and second sections of this chapter were prepared with the support of RSF grant #19-78-10048 "Structures with Lexical Repetitions from the Viewpoint of Contemporary Linguistic Theories." The third section is supported by St. Petersburg State University, the project #92562973 "Modeling of Russian Megalopolis Citizens' Communicative Behavior in Social, Speech and Pragmatic Aspects Using Artificial Intelligence Methods."

## References

1. Lubensky, S.: Russian-English Dictionary of Idioms. Random House, New York (1995)
2. Mann, W.C., Thompson, S.: Rhetorical structure theory: toward a functional theory of text organization. Text. **8**, 243–281 (1988)
3. Polanyi, L.: A formal model of the structure of discourse. J. Pragmat. **12**, 601–638 (1988)
4. Zeevat, H.: Rhetorical relations. In: Maienborn, C., Von Heusinger, K., Portner, P. (eds.) Semantics: An International Handbook of Natural Language and Meaning, pp. 946–970. Walter de Gruyter, Berlin (2011)
5. Lascarides, A., Asher, N.: Temporal interpretation, discourse relations and commonsense entailment. Linguist. Philos. **16**, 437–493 (1993)
6. Asher, N., Lascarides, A.: Logics of Conversation. Cambridge University Press, Cambridge (2003)
7. Asher, N., Lascarides, A.: Strategic conversation. Semant. Pragmat. **6**(2), 1–62 (2013)
8. Jasinskaja, K., Karagjosova, E.: Rhetorical relations. In: Gutzmann, D., Matthewson, L., Meier, C., Rullmann, H., Zimmermann, T. (eds.) The Blackwell Companion to Semantics. Wiley-Blackwell, Hoboken (2021)
9. Jasinskaja, K., Zeevat, H.: Explaining conjunction systems: Russian, English, German. In: Riester, A., Solstad, T. (eds.) Proceedings of Sinn und Bedeutung, vol. 13, pp. 231–246. University of Stuttgart, Stuttgart (2008)
10. Jasinskaja, K.: Corrective contrast in Russian, in contrast. Oslo Stud. Lang. **2**(2), 433–466 (2010)
11. Umbach, C.: On the notion of contrast in information structure and discourse structure. J. Semant. **21**(2), 155–175 (2004)
12. Wierzbicka, A.: Cross-Cultural Pragmatics: The Semantics of Human Interaction. Mouton de Gruyter, Berlin; New York (1991)
13. Bulhof, J., Gimbel, S.: Deep tautologies. Pragmat. Cogn. **9**(2), 279–291 (2001)
14. Levinson, S.: Pragmatics. Cambridge University Press, Cambridge (1983)
15. Miki, E.: Evocation and tautologies. J. Pragmat. **25**(5), 635–648 (1996)
16. Snider, T.: Using tautologies and contradictions. In: Csipak, E., Zeijlstra, H. (eds.) Proceedings of Sinn und Bedeutung, vol. 19, pp. 610–627. LinG, Gottingen (2015)
17. Artstein, R.: Inter-annotator agreement. In: Ide, N., Pustejovsky, J. (eds.) Handbook of Linguistic Annotation. Springer, Dordrecht (2017)
18. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., et al.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput. Linguist. **26**(3), 339–374 (2000)
19. Ostendorf, M., Price, P.J., Shattuck-Hufnagel, S.: The Boston University radio news corpus. Technical Report No. ECS-95-001. Boston University, Boston (1995)
20. Ghaddar, A., Langlais, P.: WikiCoref: an English coreference-annotated corpus of Wikipedia articles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 136–142. European Language Resources Association, Portorož (2016)
21. Bayerl, P.S., Paul, K.I.: What determines inter-coder agreement in manual annotations? A meta-analytic investigation. Comput. Linguist. **37**(4), 699–725 (2011)
22. Carlson, L., Marcu, D., Okurowski, M.E.: Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: van Kuppevelt, J., Smith, R.W. (eds.) Current and

New Directions in Discourse and Dialogue. Text, Speech and Language Technology, vol. 22. Springer, Dordrecht (2003)
23. Meyer, C.M.: A brief tutorial on inter-rater agreement. https://dkpro.github.io/dkpro-statistics/inter-rater-agreement-tutorial.pdf. Last accessed 2022/04/06
24. Gut, U., Bayerl, P.S.: Measuring the reliability of manual annotations of speech corpora. In: Proceedings of the Speech Prosody, Nara, Japan, pp. 565–568 (2004)
25. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Comput. Linguist. **22**(2), 249–254 (1996)
26. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Comput. Linguist. **34**(4), 555–596 (2008)
27. Siegel, S., Castellan, N.J.: Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York (1988)
28. Di Eugenio, B., Glass, M.: The kappa statistic: a second look. Comput. Linguist. **30**(1), 95–101 (2004)
29. Artstein, R., Poesio, M.: Bias decreases in proportion to the number of annotators. In: Proceedings of the 10th Conference on Formal Grammar and the 9th Meeting on Mathematics of Language, pp. 141–150 (2005)
30. Kilem, L.G.: irrCAC: computing chance-corrected agreement coefficients (CAC). R package version 1.0. https://CRAN.R-project.org/package=irrCAC. Last accessed 2022/04/06
31. Klein, D.: Implementing a general framework for assessing interrater agreement in Stata. Stata J. **18**(4), 871–901 (2018)
32. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics. **33**, 159–174 (1977)