Text Structure and Its Ambiguities: Corpus Annotation as a Helpful Guide

Šárka Zikánová

Charles University, Faculty of Mathematics and Physics, Malostranské nám. 25, 118 00 Prague 1, Czech Republic

Abstract

It is typical for natural languages that their texts can be understood differently by individual recipients. A number of scientific disciplines, from cognitive psychology to linguistics, are devoted to this phenomenon. In this study, we focus mainly on linguistic factors, which may lead to different interpretations of coherence relations in the text (simply speaking, what is related to what and how). This work presents a pilot typological survey of disagreements in Czech corpus annotations of coherence relations (discourse relations, coreference, information structure) and their common features. Polysemy (polyfunctionality) and semantic underspecification of coherent expressions (e.g. discourse connectives), generic / abstract meaning of autosemantic words, presence of attribution constructions, word order as a potential marker of information structure and text size appear to be essential factors for disagreement in interpretation. In addition, subjective reception of the relative importance of different text parts plays an important role, too. Based on the observation of the material, we raise questions and propose possible steps for the ongoing research of variability in the perception of text coherence.

Keywords

inter-annoator agreement, human label variation, discourse relations, coreference, information structure

1. Introduction

The availability of digital language resources enables an important step forward in linguistic research, both for its theoretical as well as applicational orientation. The originally collected data serving mostly for the study of the lexical studies and those of the study of syntax proper gave an impulse to enrich them by various more sophisticated annotation systems dealing with most different phenomena, going beyond the sentence boundary and incl. e.g. text coherence and phenomena related to inferencing, and elaborating more levels of granularity in the annotation. The annotated data serve for different tasks in the computational processing of natural languages – as training and testing data for the development of language models.

Human data annotation is a process based on interpretation of observed phenomena and thus may lead to different outcomes. This variation is caused by various factors. Some of them are connected with the shortcomings of the *annotation scenario* (e.g., not providing instructions for the solution of some cases) or with the leaks of the *underlying theory* (e.g., non-intuitive solutions or discerning too fine categories, very close to each other). Other cases of inter-annotator disagreement are connected with the *learning process of annotators*: especially the first annotated batches of data may be influenced by the annotators'

unfamiliarity with the annotation scenario. That is the reason why these data are often re-annotated later. To prevent these kinds of inconsistent analysis of the data, annotators usually attend frequent trainings; simultaneously, their feedback at the beginning of the annotation may improve annotation scenario and point out some problematic points in the underlying theory. Before releasing data, *annotators' mistakes* are searched for and corrected, e.g. a simple overseeing of phenomena that should be marked; nevertheless, some of the mistakes can remain even in the final data. Last, but not least source of the disagreement in the annotation is *language vagueness, polysemy and homonymy*: in some cases, a language itself allows for several understandings of a sentence.

Computational linguistics offers several methodological approaches to this variability of the data annotation. One of the solutions is unification: a gold standard is set, e.g. by majority voting or by a third judge.

Another, more demanding way of data unification is a joint annotation, when annotators mark the data together, discussing each single case and marking the result of their discussion only.

In order to accept and capture the uncertainty annotators can face while marking language phenomena, some annotation scenarios with hierarchical classifications allow the use of more general levels of the classifications, not discerning the finest classification differences in dubious cases. Another way how to mark the annotators' certainty is a separate marking of their confidence as a specific feature (e.g., (a) a discourse relation is marked as a conjunction and (b) the annotator was absolutely sure about his solution). It is necessary to say that annotator's high certainty does not necessarily mean that his solution

is the only possible one; in some cases, another annotator can be equally convinced about a different reading.

Unification is not the only way how to handle the data. Some researchers argue that unification may result in biased data missing important information about variability of language understanding [1]. Consequently, biased language models are developed based on this data. Therefore, annotators are allowed to mark multiple description of the same phenomenon in some approaches, (e.g., in the Penn Discourse Treebank 3.0 [2], a single discourse relation can be marked as an *instantiation* and *cause* at the same time, if the annotator understands it in this way). Other annotation projects publish their data with partial or complete multiple annotations carried out by different annotators; in such data, personal solutions of similar language phenomena can be observed systematically (cf. Czech RST Discourse Treebank, [3]).

2. Aim of the study

In our research, we deal with the annotation variation from a different perspective, from the linguistic and psycholinguistic point of view, with focusing on human language understanding. We use data with variations as a source of phenomena that are regularly understood in different ways and we search for possible common features of different readings. We pay special attention to the cues that are inherent to a language, rather than to the diversity among humans receiving the texts.

Questions of human language understanding have been addressed on a theoretical level, e.g. in psycholinguistics or lexical and syntactic semantics. In our study, we want to take use of our practical long term experience with large amounts of language data and possibly to offer some new insights into the variation of language interpretation or to contribute to theoretical discussions with practical findings.

3. Data: Text Coherence Annotation

Multiple reading may result at many language levels and perspectives, such as lexical semantics (cf. polysemy of the word *bank* as an institution and as a river bank), morphology (homonymous singular and plural form, like *sheep* or *fish*), syntax (*having an old friend for dinner*) etc. Our research is restricted to the area of text coherence in general. Specifically, our data cover multiple annotations of the following phenomena: discourse relations, coreference, and information structure (3.1–3.3).

3.1. Discourse relations

Discourse relations connect so called discourse arguments (clauses, sentences or larger text segments) and express certain semantic relation between the arguments. They are prototypically expressed by discourse connectives (conjunctions, subjunctions, discourse adverbs etc.), but they may be formally unexpressed, either. The former type of relations is called explicit discourse relations, the latter relations are implicit.

<Arg1: *She enjoyed working in the office*> <Arg2: ***because***$_{REASON}$ *she had pretty flowers there.*>

In our data, we work with the data from the following discourse corpora:

(a) **Prague Dependency Treebank 2.0 [12] and 3.0 [13]**. The annotation scenario of the Prague Dependency Treebank was motivated by the approach of the Penn Discourse Treebank ([14], following the Lexical Tree-Adjoining Grammar [15]) and is based on the Functional Generative Description [16] as applied in the family of Prague Dependency Treebanks. It discerns 23 semantic types of discourse relations, such as conjunction, disjunction, concession, generalization etc.; the discourse connectives are marked explicitly. The annotation is carried out on so called tectogrammatic (syntactico-semantic) dependency trees which allows the discourse annotation to be related to syntactico-semantic level of a language. The data in the corpus are in Czech.

(b) **Enriched Discourse Annotation of Prague Discourse Treebank Subset 1.0 (PDiT-EDA 1.0, [17]** The annotation scenario follows the approach of the Prague Dependency Treebank; the annotation is enriched with marking of implicit discourse relations.

(c) **Data comparing underspecification of discourse connectives in five languages (English, French, Czech, Hungarian, Lithuanian) as published in [7].** The annotation scenario is based on the Crible's classification of discourse relations [7] discerning 15 discourse relations (e.g., opening, addition, topic-shift). Unlike the Praguian discourse approach, Crible's classification takes into account broader pragmatic aspects of discourse (so called domains), explicitly discerning ideational, rhetorical, sequential, and interpersonal domains where the discourse relations are used.

(d) **Czech RST Discourse Treebank 1.0 [3]**. The annotation scenario is based on the Rhetorical Text Structure Theory as applied in the Potsdam Commentary Corpus [18]. This theory assumes that text as a whole is built from a smaller segments which are all interconnected by discourse relations, without any part being left aside. It discerns 37 discourse relations (e.g., concession, concession as nucleus, textual preparation). A specific feature of RST is that it puts emphasis on different levels of com-

Phenomenon	Source	Language	Amount of multiple annotations	Reference
Discourse relations	Prague Dependency Treebank 2.0	Czech	44 documents, 2084 sentences; 2 annotators	[4], [5]
	Enriched Discourse Annotation of Prague Discourse Treebank Subset 1.0	Czech	12 documents, 233 sentences; 2 annotators	[6]
	Unpublished parallel multilingual annotation of discourse connectives in TED talks in five languages	English, Czech, French, Hungarian, Lithuanian	3 documents, 234 sentences, 4720 words in the original English; 1-2 annotators for each language	[7]
	Czech RST Discourse Treebank 1.0	Czech	5 documents, 63 sentences, 2 annotators	[8]
Coreference	Prague Dependency Treebank 2.0	Czech	2 annotators, the number of of texts and sentences is not presented	[9]
	Prague Dependency Treebank 3.0	Czech	5 documents, 180 sentences, 2-3 annotators	[10]
Information structure	Prague Dependency Treebank 2.0 Control data annotated independently from the PDT annotation scenario	Czech	879 sentences annotated by 6 annotators, 9825 sentences annotated by 3 annotators	[11]

Table 1
Multiple annotations of text coherence (data overview)

municative importance of discourse arguments, marking more important and less important parts (nucleus and satellite, respectively) in every discourse relation. Relations with balanced importance of both parts are described as multinuclear.

3.2. Coreference

Coreferential relations connect expressions with the same reference, such as **The girl** looked into **her** map, **she** looked like **she** was enjoying the adventure. **Madelein** had a great sense of orientation. The arguments of coreferential relations are prototypically noun phrases (nouns, pronouns) including dropped phrases (While **[she]** walking through the landscape, **she** admired the nature's beauty.). A coreferential relation may also hold between a larger text segment, such as a whole thought or paragraph and a summarizing pronoun it / this etc.

We use coreference data including disagreement in the annotation coming from the Prague Dependency Treebank 2.0 [12] and 3.0 [13] where coreference is a part of multi-level annotation including discourse and syntactic semantics (see above).

3.3. Information Structure

Information structure of a sentence expresses a communicative importance of single parts of a sentence in a given context. In general, it captures a topic (what the sentence is about) and a focus of a sentence (what new information is said about the topic), cf. (context: *There is a cat under the tree.*) It $_{\text{TOPIC}}$ **is ready for a jump** $_{\text{FOCUS}}$.

Our data about information structure come from an experiment carried out on the data of the Prague Dependency Treebank 2.0 [12] where information structure is marked on dependency trees on the tectogrammatic (syntactico-semantic) level.[1]

---

[1]According to the Functional Generative Approach [16], a tectogrammatic tree consists of nodes which prototypically correspond to autosemantic words; the nodes are connected by edges expressing syntactico-semantic relations (e.g., Actor, Patient, Addressee). As for the information structure, each node is ascribed a value of contextual boundness (contextually bound, contextually non-bound, contrastively contextually bound). The nodes are ordered from the left to the right according to their so called communicative dynamism, i.e. measure to which they contribute to the development of information flow in the sentence. The values of topic and focus can be derived from these two features (contextual boundness and communicative dynamism.)

4. Methodology

In the present study, we search for general language features of sentences (words, contexts) allowing for variable readings of text structure. For this purpose, we collect occurrences of inter-annotators' disagreement in the language corpora (see Table 1) and classify them manually, putting aside occurrences of disagreement resulting obviously from other types of reasons (annotator's mistake, technical solutions of the applied theory). We concentrate on the semantic and grammatical features of the examined sentences and expressions.[2]

The results are compared and supplemented by a meta-analysis of reports on annotations of single corpora; unfortunately, due to space limitations, the annotation reports often describe reasons of inter-annotators' disagreement very shortly.

4.1. Measuring inter-annotator disagreement on a text structure

On the most general level, measuring inter-annotator agreement of textual phenomena concerns with two criteria:

(a) How often all the annotators found a certain phenomenon (e.g., a discourse relation). E.g. one annotator may ignore a case which should be marked whereas the other one does not. This would be a case of **a disagreement on the existence** of the phenomenon. (Dis)agreement on the existence is usually measured with the F1 measure (a harmonic average of precision and recall).

(b) Within the cases where all the annotators agree on the existence of a certain phenomenon, it is measured how often annotators agree on the classification of the found phenomenon. If one annotator assigns a discourse relation the semantic type *conjunction*, whereas the other one sees it as *gradation*, it is a case of **a disagreement on the type** of the phenomenon. (Dis)agreement on the type is prototypically evaluated as a simple percentage match or with the Cohen's kappa measure.

Both types of disagreement are relevant to our research: we are looking for linguistic features that can cause one annotator not to recognize a certain type of contiguity while another does. We are equally interested

in the linguistic reasons why annotators ascribe different meanings to one coherence relation.[3]

5. Analysis

In our data, which includes the annotation of discourse relations, coreference, and information structure, we have identified seven areas (factors) that repeatedly influence different readings of textual coherence by annotators.

5.1. Synsemantic signals of coherence relations: polysemy

Some words function primarily in the text as explicit markers of coherence relations (discourse connectives for discourse relations, some pronouns for anaphoric relations). However, these words are often polysemous (polyfunctional) as lexical units: they can also be used in other, coherence-unrelated roles in the text. For example, conjunctions can have a connecting function in discourse relations, but they can also become particles and function as communication expressions without connecting function (cf. Czech *Já peníze nemám,* ***ale***_{CONJUNCTION} *můj bratr je má. I have no money,* ***but***_{CONJUNCTION} *my brother has.* vs ***Ale***_{PARTICLE} *prosím vás! Co to říkáte?* ***But***_{PARTICLE} *please! What are you saying?*).

Similarly, in coreferential relations, e.g. the word *it* can perform a pronominal function and be part of a coreferential chain (**She played great. I really liked it.**), but it can also function as a grammatical word without any reference (*The weather is fine.* **It is not raining anymore.**). The presence of such synsemantic expressions in the text does not signal the presence of a coherence relation clearly; thus, recipients may disagree about the existence of a relation depending on their readings of the function of the polysemous word, as in the discourse annotation example 1:

(1)   Annotation 1: explicit discourse relation expressed by a discourse connective *přece* (*because*)
<Arg1: *Neptejte se mě, proč jsem přijel do Prahy.*>
<Arg2: *Je to* ***přece*** _{EXPLICATION} *normální sem přijet.*>

---

[2]This method has its restrictions: it may be questionable how far we interpret the real reasons of inter-annotators' disagreement correctly: what we see as a variation based on a language feature, could have be seen by an annotator just as his clear oversight. We do not have annotators' explanations for their solutions. These questions are being solved by the present-day research by Anna Nedoluzhko; for the time being, we find this method appropriate for the present analysis as a pilot study.

[3]General information on measuring inter-annotator agreement can be found in [19].

Many annotation projects adapt their measurement methods to more precisely suit the phenomena under investigation. E.g. in the case of discourse relations, the agreement on existence can be considered strictly as the case where both annotators agree on the exact scope of both discourse arguments and assign it to a certain discourse connective as an agreement on existence. For a looser approach, which respects that the exact localization of arguments can be difficult in some cases, the mere matching of a discourse connective can be considered an agreement on existence. In this case, it does not matter which words exactly the annotators mark as parts of single discourse arguments [9].

Don't ask me why I came. **Because** EXPLICATION *it's normal to come here.*

Annotation 2: no explicit discourse relation, the word *přece* (*after all*) expresses the stance of the speaker
Neptejte se mě, proč jsem přijel do Prahy. Je to **přece** *normální sem přijet.*
Don't ask me why I came. **After all**, *it's normal to come here.*

(according to [6, p. 63]; multiple annotation of the PDiT-EDA 1.0 [17])

5.2. Synsemantic signals of coherence relations: underspecification

Other cases of disagreement are based on the semantic underspecification of words signaling coherence relations: in these cases, the annotators agree on the existence of a certain relation, but they disagree on the assessment of its meaning (disagreement on type). This disagreement is typical for discourse relations, signaled by discourse connectors with a vague meaning, cf. (2):

(2)   <Arg1: *Za nabídku by se nemusel stydět ani Don Carleone – nebylo možné jí odolat.*>
<Arg2: **A tak** CONJUNCTION / RESULT *do roka a do dne dostalo práci 440 shanonských občanů a do pěti let jich bylo už desetkrát tolik.*>

<Arg1: *Not even Don Carleone would have to be ashamed of that offer – it was impossible to resist.*>
<Arg2: **And so** CONJUNCTION / RESULT *440 people of Shannon got a job within a year and a day, and within five years, they were already ten times as many.*>

([4]; multiple annotation of the PDT 2.0, [12])

Different understandings of underspecified discourse conjunctions are also evident in the dataset reported in [7], which contains the original English subtitles of TED talks and their equivalents in four languages. In the following document, the original English conjunction *but* (underspecified discourse connective with contrastive meaning) is translated using the Czech *a* (*and*, underspecified discourse connective with a simple conjunctive meaning).

(3)   English original:
Today I want to talk to you about the mathematics of love. Now, I think that we can all agree that mathematicians are famously excellent at finding love. **But** *it's not just because of our dashing personalities, superior conversational skills and excellent pencil cases.*

Czech translation:
Dnes vám chci povědět něco o matematice lásky. Myslím, že se shodneme na tom, že matematici jsou v oblasti lásky proslulí svými schopnostmi. **A** *nestojí*

za tím jen okouzlující charakter, neobyčejný konverzační um či ostře nabroušené tužky.

(Dataset of the research reported in [7])

The interchangeability of these words in the given contexts raises certain theoretical questions: for example, what level of text coherence is necessary for the recipient? In the examples given, it seems sufficient to signal that the two arguments are connected by a discourse relation. Which meaning type is specifically involved seems to be irrelevant.

Both examples, (2) and (3) lead at the same time to another question, namely the nature of the semantic types of discourse relations. In the annotations, we differentiate the individual types very precisely; but in fact, contrastivity, like causality, can be scalar, gradual, can be located on the same axis with conjunction, and different recipients can only perceive different degrees of contrastivity or causality. This property of discourse semantic types can be verified using psycholinguistic experiments.

5.3. Autosemantic words in coherence relations: genericity and abstractness

Based on the analysis of the data, we make the assumption that autosemantic words with a concrete, non-abstract meaning (cf. concrete *to bake* versus abstract *to do*) and expressions with a specific, not generic reference (*the boy* vs. *the youth as such*) are generally more accessible and representable for the recipients. In this context, we observe that words with an abstract meaning or with a generic reference can complicate the understanding of the text coherence structure: in sentences with these expressions, inter-annotator disagreement occurs more often.

Regarding coreferential relations, Nedoluzhko [10, p. 221] states that "The more nouns with abstract meaning and expressions with generic reference in the text, the smaller the agreement." It is often difficult to estimate, for example, whether concepts of two abstract expressions fully overlap (and are therefore fully coreferential), or one is a part of the other, or they are independent, cf. (4).

(4)   (context: interview with child psychiatrists who published the Czech book Children, Family and Stress)
- Materiálům, které dnes máte k dispozici, předcházel **dlouholetý výzkum**.
- Zdeněk Dytrych: Od roku 1969, kdy jsme založili v bývalém Výzkumném ústavu psychiatrickém Oddělení pro výzkum rodiny, se hlavně zabýváme touto problematikou.
Měli jsme samozřejmě řadu spolupracovníků a za pětadvacet let jsme v týmu udělali téměř nekonečnou **řadu prací**.
Tak například rozsáhlý **výzkum rozvodovosti**.

*- The materials you have at your disposal today were preceded by **a long-term research**.*

- Zdeněk Dytrych: Since 1969, when we founded the Department for Family Research in the former Research Institute of Psychiatry, we have mainly been dealing with this issue.

*Of course, we had a number of collaborators, and in twenty-five years we have done an almost endless **amount of work** as a team.* [lit.: *endless amount of works* (plural) which can mean *publications* as well, ŠZ]

*For example, extensive **research on the divorce rate**.*

([10, p. 223–226]; multiple annotation of the PDT 3.0[13])

In example (4), the question is how the last sentence is related to the previous text – what is *the research on the divorce rate* supposed to serve as an example of? One annotator sees the phrase *research on the divorce rate* as an example of *a series (amount) of works* in the previous sentence, while the other one sees it as an example of the *long-term research* in the first sentence. Is *a series (amount) of works (publications?)* the same as *research*? Or are *the works (publications)* only the result of *research*, i.e. one part of it? Similar contradictions are quite common in the understanding of the coreference of generic and abstract terms.

Also in the annotation of discourse relations, words with an abstract, non-specific meaning result in the inter-annotators' disagreement [5]. This is the case of sentences including verbs with an abstract, general meaning. As the authors say, "The disagreement occurs when it is not clear whether the potential discourse connective refers to the whole sentence as an independent abstract object (discourse argument), or just to its complement, typically a nominal phrase." [5, p. 2003]. Thus, in example (5), the disagreement between annotators shows that it is questionable whether the second part of the sentence (*while chimneys…*) is related to the whole previous clause including the verbs with abstract meaning (*it is possible to note a small, but distinctive difference between…*), or just to the nominal phrase (*a small, but distinctive difference between…*).[4]

(5) <Arg1: *Při prohlídce střech Šternberského paláce si lze všimnout **drobného, avšak charakteristického rozdílu mezi přístupem památkářů koncem 80. let a nyní:*** COLON > SPECIFICATION <Arg2: *zatímco komíny staré sněmovny byly zbourány jako zbytečné a zůstala jen holá střecha, dělníci KDM mají přikázáno komíny všech čtyř objektů nejen ponechat, ale dokonce mírně přizdobit, aby tradiční kolorit malostranských střech časem nezmizel.>*

<Arg1: *When observing the roofs of the Sternberg Palace it is possible to note **a small, but distinctive difference between the approaches of preservationists of late 80's and now:*** COLON > SPECIFICATION <Arg2: *while chimneys of the old Parliament were demolished as functionless and only a clear roof was retained, the KDM workers are ordered not only to maintain chimneys of all the four objects, but even to decorate them slightly, so that the traditional local atmosphere of Lesser Town roofs does not eventually disappear.>*

([5, p. 2004]; multiple annotation of the PDT 2.0[12])

In fact, this is a disagreement on which level the given phenomenon should be captured (in this case, coreference or discourse). It is rather an academic question how to annotate these cases consistently. As for the recipients themselves, the difference in the annotation does not mean a difference in the understanding of the text, as the language levels and perspectives are inter-related and the annotators can ascribe single phenomena to different levels without understanding the text coherence in a different way.

5.4. Attribution: verbs of thinking and saying

Attribution is the relation between the (named) author of a section of text and his speech. A typical component in the attribution construction is the author's name, the verb of thinking or speaking or another form expressing speech (colon, phrases such as *according to*) and the direct / indirect speech itself (dictum). A language has means how to distinguish the author's speech from the reported speech. Nevertheless, with attributive constructions it is often difficult to distinguish how far discourse relations extend and what is the scope of their arguments, especially when it comes to verbs of thinking and saying. In these cases, annotators often disagree in their interpretations, cf. examples (6) and (7).

(6) Annotation 1: the discourse connective *ale* (*but*) relates the second sentence to the whole previous sentence including the verb of thinking phrase *vím, že* (*I know that*).

"<Arg1: *Vím, že se nás Rusů bojíte, že nás nemáte rádi, že námi trochu pohrdáte.>* <Arg2: ***Ale*** OPPOSITION *Rusko není jenom Žirinovskij, Rusko není jenom vraždění v Čečensku.>*"

"<Arg1: *I know that you are afraid of us Russians, that you dislike us, that you despise us a little.>* <Arg2: ***But*** OPPOSITION *Russia is not only Zhirinovsky, Russia is not only murdering in Chechnya.>*"

Annotation 2: the discourse connective *ale* (*but*)

---

[4]According to the approach of the Prague Dependency Treebank 2.0, a colon is understood as an explicit discourse connective ([20]).

relates the second sentence to the content of the thought only, without the governing verb of thinking.

"*Vím, že* <Arg1: *se nás Rusů bojíte, že nás nemáte rádi, že námi trochu pohrdáte.*> <Arg2: **Ale** OPPOSITION *Rusko není jenom Žirinovskij, Rusko není jenom vraždění v Čečensku.*>"

"*I know that* <Arg1: *you are afraid of us Russians, that you dislike us, that you despise us a little.*> <Arg2: **But** OPPOSITION *Russia is not only Zhirinovsky, Russia is not only murdering in Chechnya.*>"

([9, p. 777]; multiple annotation of the PDT 2.0 [12])

(7) Annotation 1: the discourse connective *tudíž* (*therefore*) relates the second sentence to the whole previous sentence including the governing verb of saying phrase *trvají památkáři* (*preservationists insist*); the relation of reason is broader.

<Arg1: *Na tom, aby ve Šternberku ani v paláci Smiřických nevznikaly žádné příčky, trvají památkáři.*> <Arg2: *Poslancům* **tudíž** REASON *nebude dopřáno žádné velké soukromí.*>

<Arg1: *Preservationists insist that no partition walls will be built up neither in the Sternberg Palace nor in the Smiřický Palace.*> <Arg2: **Therefore,** REASON *MP's will not enjoy great privacy.*>

Annotation 2: the discourse connective *tudíž* (*therefore*) relates the second sentence to the content of the saying only (dictum), the Arg1 is smaller; the meaning of the whole causal relation is different.

<Arg1: *Na tom, aby ve Šternberku ani v paláci Smiřických nevznikaly žádné příčky,*> *trvají památkáři.* <Arg2: *Poslancům* **tudíž** REASON *nebude dopřáno žádné velké soukromí.*>

([5, p. 2005]; multiple annotation of the PDT 2.0[12])

In general, attribution is one of the ways of text arrangement, in addition to e.g. parentheses, meta-comments on the communication etc. All of these ways represent a digression from the baseline of a simple main narrative with a single narrator. As such, they can be a source of different interpretations of the text: people can differ in what they regard as author's speech and what as reported speech, what as part of the main line and what as a parenthesis, etc. (see subsection 5.6 below).

5.5. Word order

So far, we have observed cases of disagreement between annotators, which result from the lexical properties of expressions ensuring coherence (underspecification vs. specificity, abstractness vs. concreteness) and from the syntactic structure (governing verb of saying/thinking vs. dictum itself). Word order is another area that plays an important role in ensuring the coherence of the text and can also become subject to different interpretations.

In Czech, similarly as in other Slavic languages, the word order is relatively free, with few grammatical restrictions. It is used to express information structure of a sentence: the information belonging to the topic is prototypically placed in the sentence to the left, the focus is usually located to the right. However, it is also possible to use a marked word order, when the topic and focus occupy various places in the sentence and are distinguished by intonation, the use of focalizers, or deduced from the context. This freedom in the formal expression of information structure results in some cases in inter-annotator disagreement. Often, annotators interpret differently information structure of the left part of a sentence: some tend to consider it less important, disregarding the used expressions, because it is prototypically a topic position; others are more driven by context and other indicators of possible focus.

This variability applies especially to adverbials located before the verb in the surface word order, focalized phrases and predicate verbs in the left part of the sentence [11]. The example (8) presents an ambiguous interpretation of the conditional phrase at the beginning of the sentence; one of the annotators considers it to be a part of the very message of the sentence, the other as a mere unimportant circumstance. Thus, both perceive the given sentence as a response to a different (unspoken) context, as shown by the contextual questions at the end of each interpretation. (The expressions in topic are underlined; the focus is marked with bold characters.)

(8) (Context: *Po ekonomech, kteří nyní už opouštějí školu se znalostí pravidel hry v tržním prostředí, je hlad. Co hodláte udělat, aby jich bylo dost?*
The economists are now requested who leave the school with a knowledge of the life in the market environment. How do you intend to provide a sufficient number of them?)

Annotation 1:
[*Při využití všech výukových prostor od rána až do večera*] 0-subject *jsme schopni ročně přijmout ke studiu okolo 2500 studentů*.
Lit.: **[*When using all classrooms from morning till evening*]** <u>we</u> **are able a_year to_accept to_studies about 2500 students.**
[When using all our classrooms during the whole day], we are able to accept about 2500 new students a year.
(How is your present-day situation?)

Annotation 2:
[*Při využití všech výukových prostor od rána až do* <u>*večera*</u>] *jsme schopni ročně přijmout ke studiu okolo 2500 studentů*.

(How will your situation be if you take full advantage of your present-day capacities?)

([11]; control multiple annotation of the PDT 2.0, [12])

In example (9), there is a collision between two indicators of importance (belonging to the topic / focus): the observed phrase is located at the beginning of the sentence, a place typical for the topic; but at the same time it is emphasized by the focalizer. Annotators perceive its role in the information structure of the sentence differently.

(9) (Context: *Oskar… Firmě Ilja Běhal a spol., zajišťující umělecko-kovářské a restaurátorské práce hlavně na střední Moravě.*
The Oscar prize… for the firm Ilja Běhal & Co. which deals with smith craft and conservatory works mainly in central Moravia.)

Annotation 1:
[Zejména FOCALIZER ***v Olomouci]*** <u>*firma*</u> **svými výrobky přispívá ke zvýraznění koloritu historického jádra města**.
Lit.: *[Especially* FOCALIZER ***in Olomouc]*** <u>*firm*</u> **with_its products helps accentuation of_colouring of_historical centre of_city**.
[Especially in Olomouc], the firm helps to accentuate the colouring of the historical centre of the city with its products.
(What does the firm do? What can we say about the firm?)

Annotation 2:
[Zejména v Olomouci] <u>*firma*</u> **svými výrobky přispívá ke zvýraznění koloritu historického jádra města**.
(What does the firm do especially in Olomouc?)

([11]; control multiple annotation of the PDT 2.0 [12])

In example (10), a striking feature of verbs can be seen: expressions dependent on the verbs often tend to be communicatively more important than the verbs themselves. This can make the role of predicate verbs in the information structure unclear: annotators do not agree whether to classify them as focus or as topic. We have already observed the unclear importance of verbs with respect to dependent parts in examples (5, unclear role of a verb with general meaning in a discourse structure) and (6-7, unclear role of a verb of thinking/saying in a discourse structure, compared to the clear role of dictum).

(10) (Context:
- *Nářky lidí známe ze svého nejbližšího okolí. Jejich frekvence spíš vzrůstá, než aby se tenčila. Proč?*
- *We know these complaints from our nearest vicinity. Their frequency is getting rather higher than lower. Why?*)

Annotation 1:

Nejvíc [kritizují a rozčilují se] **neschopní**.
Lit.: **Most [criticize and get_angry] incompetent**.
Incompetent employees criticize and get angry most of all.
(What happens?)

Annotation 2:
Nejvíc [kritizují a rozčilují se] **neschopní**.
(Who criticizes and gets angry most of all?)

([11]; control multiple annotation of the PDT 2.0 [12])

5.6. Core of the message: subjective perception of the relative importance

At this point, we allow ourselves a small digression inspired by the information structure. In many kinds of coherence annotations, we see that annotators differ in what they consider to be important, central, at a given place in the text.

As the previous subsection showed, the variety of understanding of coherence relations often comes from certain linguistic forms (specific word order pattern, etc.). However, the language itself often does not provide a clue: we cannot tell which phrase or syntactic construction was vague enough to allow for multiple readings. The diversity here comes from the different experience of the recipients, from their expectations and knowledge of the world. This type of inter-annotator disagreement is difficult for linguistics to grasp. Nevertheless, since we can document it well in our data, we take the liberty of presenting a few of these phenomena here, which can serve as inspiration for e.g. psycholinguistic research.

At the local level, subjectivity can be seen in the perception of importance in the information structure (cf. [21]), i.e. what people see as a topic / focus of a sentence. Furthermore, this variation is found in discourse relations in Rhetorical Structure Theory, which differentiates between a more substantial and a less substantial arguments of a discourse relations (nucleus and satellite, respectively; cf. [8]). See the following example (11) where adjacent sentences have the same syntactic structure connected by the phrase *not only – but also*. One of the annotators considers both parts of these sentences to have the same level of importance and marks a multinuclear relation of *contrast* between them. The other one understands the second parts (starting with *but also*) as emphasized, more important, marking thus the relation as *antithesis* with the nucleus in the second part.

(11) <Arg1: *Jan Kotík* **nemaluje** *jen* *očima a rukou,*> CONTRAST / ANTITHESIS <Arg2: *ale také mozkem.*>
<Arg1: *Jeho obrazy tedy vyžadují* **nejen** *citlivost a vnímavost,*> CONTRAST / ANTITHESIS <Arg2: *ale také přemýšlení.*>

<Arg1: *Jan Kotík paints* **not only** *with his eyes and*

hands,> CONTRAST / ANTITHESIS *<Arg2: **but also** with his brain.>*

*<Arg1: Therefore his paintings require **not only** sensitivity and receptivity,>* CONTRAST / ANTITHESIS *<Arg2: **but also** thinking.>*

At the global level, in the annotations according to Rhetorical Structure Theory, the perceptual importance of individual parts of news reports differs, too. Typically, while one annotator understands the introductory part as a central message to which details are added in the following text, the other perceives the same part as a preparation to which the own message is associated afterwards. ([8]).

5.7. Text dimensions

Inter-annotator agreement can also be affected by text dimensions. As coreference research shows, the larger the network of possible antecedents for a given word in a text, the greater the disagreement between annotators ([10, p. 221]; cf. the opportunities for disagreement in example 4). The author further states that divergent interpretations of coreference can also be chained: if annotators differ in the interpretation of expressions at the beginnings of the coreference chain, their different interpretations can be reflected in other expressions with a similar meaning in the text.

It is a question of how the size of the text affects the variability of understanding in other coherence relations, such as discourse relations and information structure. We have not yet conducted research in this direction. For discourse relations, there can theoretically be more potential arguments in a large text that are connected by a discourse connective. If the text is longer, it will probably also be more layered in terms of author's and reported speech, metacommunication, insertions, etc., which again offers more possibilities for different understandings of discourse and other relation. On the other hand, a longer text can more accurately describe the context in which the discourse relations are interpreted, and thus contribute to the clarity of understanding. In this regard, another question arises: whether there is a difference in the variability in the understanding of coherence relations at the beginning of the text (where the text is still short, there are few potential members of different relations available, but also little context) and in its later parts.

6. Conclusion

In this study, we observed what common features the occurrences of inter-annotator disagreement have in coherence relations, specifically in discourse relations, coreference and information structure. We were mainly con-

cerned with the features given by the language itself; we only marginally stopped at cases of disagreement that result from the difference of speakers. We have also formulated some questions that can be the subject of further research.

Coherence relations can be divided into formally expressed (e.g. in the discourse structure relations expressed by an explicit discourse connective or an information structure expressed by word order) and unexpressed relations that are understood from the context (e.g. coreference relation between the words *text* and *chapter* in a specific text).

In formally unexpressed relations, disagreement occurs naturally: it depends on the recipients what they infer from the context. Formally expressed relations can be also interpreted differently. There may be disagreement on the very existence of a coherence relation; this disagreement is usually based on the polysemy (polyfunctionality) of the linguistic form (expression), which in some contexts functions as a signal of coherence, but not in others. In addition, coherence signals can also lead to a different perception of the semantic type of a discourse relation (in cases where speakers agree on its existence): this is caused by the semantic underspecification of language forms that express coherence (discourse connectives). The general question arises whether, as recipients, we need to understand textual coherence in detail in all contexts, i.e. distinguish not only the simple existence of coherence relations, but also their semantic coloring. What level actually represents a functional and sufficient understanding of the text?

Lexical specificity plays an important role in the understanding of autosemantic words, too; these expressions do not function primarily as signals of coherence. Coreference research shows that for abstract and generic nominal phrases in a text, recipients determine with difficulty whether the words have the same content; in contrast, for words with a concrete, specific meaning, coreference is easier to determine. The same applies to the semantic concreteness of verbs: for verbs with more vague, general meanings, it is difficult for annotators to determine whether or not they are part of discourse arguments. Their meaning seems to be too insignificant, whereas the content of their dependent words is more important.

This observation also applies to the verbs of thinking and saying in the relation of attribution, where the content of reported speech seems to be communicatively more essential than the act of communication itself. In the case of attribution, there is another reason for the diverse interpretation of the text: it represents one of the forms of text arrangement (alongside parentheses, meta-comments on the communication, etc.), i.e. a complication in the simple basic line of the narrative. It thus provides the possibility for different recipients to interpret the overall structure of the text differently.

In addition to individual words, such as various coherence operators or autosemantic expressions, word order can also cause a disagreement in text understanding. Specifically, in Czech and other Slavic languages, word order affects the understanding of the information structure. If expressions with higher communicative dynamism (informativeness) appear in the left, topical part of the sentence, which has a prototypically low communicative dynamism, typical contradictions in their evaluation occur.

In many types of annotation, it turns out that annotators perceive the importance of individual parts of the text and their (hierarchical) connections differently. These disagreements are often not so much caused by the special properties of the text as by differences between the annotators (specifically, it may be knowledge of the language, knowledge of the world, expectations, experience with different text genres, etc.). This area seems particularly suitable for future psycholinguistic research focusing on specific domains of coherence. Here, for example, it is possible to examine the influence of respondents' literacy on the understanding of coreference in abstract words or the process how children learn the text arrangement.

The last factor we dealt with is text dimensions. Its effect on different readings was described in coreference (the longer the text, the greater the disagreement in interpretation). For other coherence relations, this factor is still unexplored. We hypothesized that for discourse relations and information structure, text dimensions could influence the degree of disagreement in both directions; the degree of disagreement may also vary by place in the text and amount of preceding context (early vs. later in the text). These ideas suggest possible directions for further research on different text comprehension coherence.

Acknowledgments

References

[1] B. Plank, The "problem" of human label variation: On ground truth in data, modeling and evaluation, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10671–10682. URL: https://aclanthology.org/2022.emnlp-main.731. doi:10.18653/v1/2022.emnlp-main.731.

[2] R. Prasad, B. Webber, A. Lee, A. Joshi, Penn Discourse Treebank Version 3.0, 2019. URL: https://hdl.handle.net/11272.1/AB2/SUU9CB. doi:11272.1/AB2/SUU9CB.

[3] L. Poláková, Š. Zikánová, J. Mírovský, E. Hajičová, Czech RST Discourse Treebank 1.0, 2023.

[4] P. Jínová, J. Mírovský, L. Poláková, Analyzing the most common errors in the discourse annotation of the Prague Dependency Treebank, in: I. Hendrickx, S. Kübler, K. Simov (Eds.), Proceedings of the 11th International Workshop on Treebanks and Linguistic Theories, Universidade de Lisboa, Edicoes Colibri, Lisboa, Lisboa, Portugal, 2012, pp. 127–132.

[5] Š. Zikánová, L. Mladová, J. Mírovský, P. Jínová, Typical cases of annotators' disagreement in discourse annotations in Prague Dependency Treebank, in: Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010), European Language Resources Association, Valletta, Malta, 2010, pp. 2002–2006.

[6] Š. Zikánová, Implicitní diskurzní vztahy v češtině [Implicit Discourse Relations in Czech], Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic, 2021.

[7] L. Crible, Á. Abuczki, N. Burkšaitienė, P. Furkó, A. Nedoluzhko, G. Oleskeviciene, S. Rackevičienė, Š. Zikánová, Functions and translations of underspecified discourse markers in TED talks: a parallel corpus study on five languages, Journal of Pragmatics (2019) 139–155.

[8] L. Poláková, J. Mírovský, Š. Zikánová, E. Hajičová, Developing a Rhetorical Structure Theory Treebank for Czech, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), European Language Resources Association, Torino, Italy, 2024, pp. 4802–4810.

[9] J. Mírovský, L. Mladová, Š. Zikánová, Connective-based measuring of the inter-annotator agreement in the annotation of discourse in PDT, in: C.-R. Huang, D. Jurafsky (Eds.), Proceedings of the 23rd International Conference on Computational Lin-

guistics (Coling 2010), volume 1, Chinese Information Processing Society of China, Tsinghua University Press, Beijing, China, 2010, pp. 775–781.

[10] A. Nedoluzhko, Rozšířená textová koreference a asociační anafora (Koncepce anotace českých dat v Pražském závislostním korpusu) [Extended nominal coreference and bridging anaphora (An approach to annotation of Czech data in the Prague Dependency Treebank)], Studies in Computational and Theoretical Linguistics, Ústav formální a aplikované lingvistiky, Praha, Česká republika, 2011.

[11] Š. Zikánová, M. Týnovský, Identification of Topic and Focus in Czech: Comparative Evaluation on Prague Dependency Treebank, in: G. Zybatow, U. Junghanns, D. Lenertová, P. Biskup (Eds.), Studies in Formal Slavic Phonology, Morphology, Syntax, Semantics and Information Structure. Formal Description of Slavic Languages 7, Universität Leipzig, Peter Lang, Frankfurt am Main, Germany, 2009, pp. 343–353.

[12] J. Hajič, J. Panevová, E. Hajičová, P. Sgall, P. Pajas, J. Štěpánek, J. Havelka, M. Mikulová, Z. Žabokrtský, M. Ševčíková-Razímová, Z. Urešová, Prague Dependency Treebank 2.0, 2006.

[13] E. Bejček, E. Hajičová, J. Hajič, P. Jínová, V. Kettnerová, V. Kolářová, M. Mikulová, J. Mírovský, A. Nedoluzhko, J. Panevová, L. Poláková, M. Ševčíková, J. Štěpánek, Š. Zikánová, Prague Dependency Treebank 3.0, 2013.

[14] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, B. Webber, The Penn Discourse TreeBank 2.0, in: Proceedings, 6th International Conference on Language Resources and Evaluation, Marrakech, Morocco, 2008, pp. 2961–2968.

[15] B. L. Webber, A. K. Joshi, Anchoring a Lexicalized Tree-Adjoining Grammar for discourse, in: Discourse Relations and Discourse Markers, 1998. URL: https://aclanthology.org/W98-0315.

[16] P. Sgall, E. Hajicová, J. Panevová, The Meaning of the Sentence in its Semantic and Pragmatic Aspects, Springer Science & Business Media, 1986.

[17] Š. Zikánová, P. Synková, J. Mírovský, Enriched discourse annotation of PDiT subset 1.0 (PDiT-EDA 1.0), 2018.

[18] M. Stede, M. Taboada, D. Das, Annotation Guidelines for Rhetorical Structure. (Manuscript)., 2017.

[19] R. Artstein, Inter-annotator agreement, Handbook of linguistic annotation (2017) 297–313.

[20] L. Poláková, Discourse Relations in Czech, Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague, Czech Republic, 2015.

[21] Š. Zikánová, M. Týnovský, J. Havelka, Identification of Topic and Focus in Czech: Evaluation of Manual Parallel Annotations, The Prague Bulletin of Mathematical Linguistics (2007) 61–70.