Annotation of subtitle paraphrases using a new web tool

Mikko Aulamo[0000-0002-3253-2744], Mathias Creutz[0000-0003-1862-4172], and Eetu Sjöblom[0000-0001-5531-8196]

Department of Digital Humanities
University of Helsinki
{mikko.aulamo, mathias.creutz, eetu.sjoblom}@helsinki.fi

Abstract. This paper analyzes the manual annotation effort carried out to produce Opusparcus, the Open Subtitles Paraphrase Corpus for six European languages. Within the scope of the project, a new web-based annotation tool was created. We discuss the design choices behind the tool as well as the setup of the annotation task. We also evaluate the annotations obtained. Two independent annotators needed to decide to what extent two sentences approximately meant the same thing. The sentences originate from subtitles from movies and TV shows, which constitutes an interesting genre of mostly colloquial language. Inter-annotator agreement was found to be on par with a well-known previous paraphrase resource from the news domain, the Microsoft Research Paraphrase Corpus (MSRPC). Our annotation tool is open source. The tool can be used for closed projects with restricted access and controlled user authentification as well as open crowdsourced projects, in which anyone can participate and user identification takes place based on IP addresses.

Keywords: annotation · paraphrase · web tool · inter-annotator agreement · subtitle.

1 Introduction

This paper introduces an online tool for annotating paraphrases and evaluates annotations gathered with the tool. Paraphrases are pairs of phrases in the same language that express approximately the same meaning, such as *"Have a seat."* versus *"Sit down."*. The annotated paraphrases are part of Opusparcus [3], which is a paraphrase corpus for six European languages: German (de), English (en), Finnish (fi), French (fr), Russian (ru), and Swedish (sv).

The paraphrases in Opusparcus consist of movie and TV subtitles from OpenSubtitles2016 parallel corpora [9], which are part of the larger OPUS corpus.[1] We are interested in movie and TV subtitles because of their conversational nature. This makes subtitle data ideal for exploring dialogue phenomena and properties of everyday, colloquial language [11,17,10]. In addition, the data could prove

[1] http://opus.nlpl.eu/

useful in modeling semantic similarity of short texts, with applications such as extraction of related or paraphrastic content from social media. Our data could also be valuable in computer assisted language learning to teach natural everyday expressions as opposed to the formal language of some well-known data sets, consisting of news texts, parliamentary speeches, or passages from the Bible. Additionally, paraphrase data is useful for evaluating machine translation systems, since it provides multiple correct translations for a single source sentence.

Opusparcus consists of three types of data sets for each language: training, development and test sets. These data sets can be used, for instance, in machine learning. The training sets consist of millions of sentence pairs and their paraphrases are paired automatically using a probabilistic ranking function. The training sets are not discussed further in the current paper, which instead focuses on the manually annotated development and test sets. The development and test sets contain a few thousands of sentence pairs. Each of the pairs has been checked by human annotators in order to ensure as high quality as possible. The annotation effort took place using the annotation tool, which is presented in more detail below.

The source code of the annotation tool is public.[2] A public version of the tool is online for anyone to test.[3] The data gathered with the tool along with the rest of Opusparcus is available for downloading.[4]

The paper is divided into two main parts: First the setup of the annotation task is described together with the design of the annotation tool. Then the annotations produced in the project are evaluated.

2 Setup

In the beginning of the project, we faced many open questions. In the following, we discuss the options we considered when setting up the annotation task. We also describe why we created our own annotation tool and how the tool works.

2.1 Annotation scheme

An essential question when determining the paraphrase status of sentence pairs, is what rating scheme to use. The simplest scheme is to have two categories only, as is the case with the Microsoft Research Paraphrase Corpus (MSRPC) [4]: "Raters were told to use their best judgment in deciding whether 2 sentences, at a high level, 'mean the same thing'."

Another well known resource, the Paraphrase Database (PPDB) [6] contains automatically extracted paraphrases; however, the construction of PPDB also

[2] https://github.com/miau1/simsents-anno
[3] https://vm1217.kaj.pouta.csc.fi
[4] Available through the Language Bank of Finland: http://urn.fi/urn:nbn:fi:lb-2018021221

involved manual annotation to some extent: "To gauge the quality of our paraphrases, the authors judged 1900 randomly sampled predicate paraphrases on a scale of 1 to 5, 5 being the best."

In a later version, PPDB 2.0 [12], there is further discussion: "Although we typically think of paraphrases as equivalent or as bidirectionally entailing, a substantial fraction of the phrase pairs in PPDB exhibit different entailment relations. [...] These relations include forward entailment/hyponym, reverse entailment/hypernym, non-entailing topical relatedness, unrelatedness, and even exclusion/contradiction."

In addition to assessing the degree of paraphrasticity, the annotation schemes can include information about the types of paraphrase relations a phrase pair contains. Vila et al. [16] propose a complex scheme based on extensive linguistic paraphrase typology. It consists of 24 different type tags and the annotations also include the scopes for different paraphrase relations, such as lexical, morphological or syntactic changes. Other complex schemes have also been developed. Kovatchev et al. [7] extend the typology and annotation scheme of Vila et al., whereas Barrón-Cedeño et al. [1] present a scheme based on an alternative typology.

When designing the Opusparcus corpus we wanted to annotate *symmetric* relations and find out whether two sentences essentially meant the same thing. This excluded the different (asymmetric) entailment options from our emerging annotation scheme. Furthermore, having only two classes (paraphrases versus non-paraphrases) seemed too limited, because of some challenges we faced with the data. In our system, the sentence pairs proposed as paraphrases are produced by translating from one language to another language and then back; for instance, English: *"Have a seat."* → French: *"Asseyez-vous."* → English: *"Sit down."* Here "translation" actually means finding subtitles in different languages that are shown at the same time in a movie or TV show. We have found that translational paraphrases exhibit (at least) two types of near-paraphrase cases:

1. Scope mismatch: The two phrases mean almost the same thing, but one of the phrases is more specific than the other; for instance: *"You?"* ↔ *"**How about** you?"*, *"Hi!'* ↔ *"Hi, **Bob!**"*, *"What are you doing?"* ↔ *"What **the hell** are you doing?"*
2. Grammatical mismatch: The two phrases do not mean the same thing, but the difference is small and pertains to grammatical distinctions that are not made in all languages. Such paraphrase candidates are typically by-products of translation between languages; for instance: *"I am **a** doctor."* ↔ *"I am **the** doctor."*, or French *"**Il** est là."* ↔ *"**Elle** est là."*. The French example could mean either *"He is here."* ↔ *"She is here."* when referring to animate objects, or just *"It is here."* when talking about inanimate things. It does not appear crucial to distinguish between grammatical gender in the latter case.

Another aspect that caught our attention initially was whether it would be necessary to distinguish between *interesting* and *uninteresting* paraphrases.

There are fairly trivial transformations that can be applied to produce paraphrases, such as: *"I **am** sorry."* ↔ *"I'**m** sorry."*, *"**Al**right."* ↔ *"**All** right."*, or change of word order, which is common in some languages; an English example could be: *"I don't know **him**."* ↔ *"**Him** I don't know."* If a computer were to determine whether such phrase pairs were paraphrases, a very simple algorithm would suffice, and the data would not be too interesting from a machine learning point of view.

Taking these considerations into account, an initial six-level scale was planned for assessing to what extent sentences meant the same thing: 5 – Excellent, 4 – Too similar, and as such uninteresting, 3 – Scope mismatch, 2 – Grammatical mismatch, 1 – Farfetched, 0 – Wrong. However, this scheme immediately turned out to be impractical. The scale does not produce a simple range from good to bad. For instance, in case of 5 (excellent) or 4 (too similar), the annotator first has to decide whether the sentences are paraphrases or not, and in case of paraphrases, whether they are interesting or not.

A four-grade scale was adopted instead: 4 – Good example of paraphrases, 3 – Mostly good example of paraphrases, 2 – Mostly bad example of paraphrases, and 1 – Bad example of paraphrases. Note that the scale has an even number of entries, so that the annotator needs to take sides, and indicate a preference towards either good or bad. There is no option for "cannot tell" in the middle, in contrast to the five-grade scale of PPDB [6]. Nonetheless, a fifth so-called "trash" category was created, to make it possible for the annotators to discard invalid data.

The number of too similar sentence pairs have been reduced in a prefiltering step, where edit distance is used to measure sentence similarity. In this way, we avoid wasting annotation effort on trivial cases. When it comes to scope mismatch and grammatical mismatch, the annotators must make decisions to their best judgment and the characteristics of the language they are annotating; these cases need to be annotated as either "mostly good" (3) or "mostly bad" (2) examples of paraphrases. The instructions shown to the annotators are displayed in Table 1.

2.2 Why did we build our own tool?

Before tackling the annotation task, we evaluated whether to use an existing annotation tool or build one ourselves. Using an existing tool is potentially less expensive, and existing services usually offer ways of storing and backing up data and securely handling user authentications.

We tried using WebAnno [18], which is a web-based annotation tool designed for linguistic annotation tasks. With WebAnno, one can design one's own annotation projects, assign users and monitor the projects. WebAnno turned out to be too slow to use for our purposes: the user has to highlight the part they want to annotate and then type in the annotation category. Working with WebAnno is useful for annotating linguistic relations but unnecessarily complicated for simply choosing one of our five annotation categories.

Table 1. The five annotation categories used, with instructions and examples for the annotators. The colors mentioned correspond to the color of a button in the user interface of the tool.

Category	Description	Examples
Good, *"Dark green"*, 4	The two sentences can be used in the same situation and essentially "mean the same thing".	*It was a last minute thing.* ↔ *This wasn't planned.* *Honey, look.* ↔ *Um, honey, listen.* *I have goose flesh.* ↔ *The hair's standing up on my arms.*
Mostly good, *"Light green"*, 3	It is acceptable to think that the two sentences refer to the same thing, although one sentence might be more specific than the other one, or there are differences in style, such as polite form versus familiar form. There may also be differences in gender, number or tense, etc if these differences are of minor importance for the phrases as a whole, such as masculine or feminine agreement of French adjectives.	*Hang that up.* ↔ *Hang up the phone.* *Go to your bedroom.* ↔ *Just go to sleep.* *Next man, move it.* ↔ *Next, please.* *Calvin, now what?* ↔ *What are we doing?* *Good job.* ↔ *Right, good game, good game.* *Tu es fatigué?* ↔ *Vous êtes fatiguée?* *Den är fånig.* ↔ *Det är dumt.* *Olet myöhässä.* ↔ *Te tulitte liian myöhään.*
Mostly bad, *"Yellow"*, 2	There is some connection between the sentences that explains why they occur together, but one would not really consider them to mean the same thing. There may also be differences in gender, number, tense etc that are important for the meaning of the phrases as a whole.	*Another one?* ↔ *Partner again?* *Did you ask him?* ↔ *Have you asked her?* *Hello, operator?* ↔ *Yes, operator, I'm trying to get to the police.* *Isn't that right?* ↔ *Well, hasn't it?* *Get them up there.* ↔ *Put your hands in the air.* *I thought you might.* ↔ *Yeah, didn't think so.* *I am on my way.* ↔ *We are coming.*
Bad, *"Red"*, 1	There is no obvious connection. The sentences mean different things.	*She's over there.* ↔ *Take me to him.* *All the cons.* ↔ *Nice and comfy.*
Trash	At least one of the sentences is invalid in some of the following ways: – The language of the sentence is wrong, such as an English phrase in the French annotation data. – There are spelling mistakes or the sentence is syntactically misformed. However, sloppy punctuation or capitalization can be ignored and the sentence can be accepted.	*Estoy buscando a mi hermana.* ↔ *I'm looking for my sister.* *Now, watch what yöu're saying.* ↔ *Watch your mouth.* *Adolfo Where can I find?* ↔ *Where I can find Adolfo?*

Amazon Mechanical Turk[5] (AMT) is similar to WebAnno in the sense that users can design their own annotation task, but the main selling point of AMT is that the annotations are made using crowdsourcing. AMT utilizes a global marketplace of workers who are paid for their work effort. According to Snow et. al [15], linguistic annotation tasks can be carried out quickly and inexpensively by non-expert users. However, it is important that the annotators are proficient in the language they are annotating in order to obtain reliable annotations.

In the end, we decided to implement our own tool, because it needs to perform a specific task in a controllable setting.

2.3 Design choices

Before implementing the annotation platform, the design has to be thought out thoroughly to serve the annotation task. It is important that the annotation process is simple and convenient. This makes the task pleasant for the annotators, while simultaneously benefiting the ones conducting the project by allowing annotations to be gathered faster.

Web-based tool. In order to allow the annotators an easy access to the tool, we decided to make it accessible with a web browser. In this way the annotators can evaluate sentence pairs anywhere and anytime they like. This also allows for easy recruitment of new annotators by creating new user accounts and sharing the link to the interface.

The main annotation view is meant to be simple and informative (Figure 1). The person annotating sees two sentences in a given language and evaluates the similarity on a scale from 1 to 4 by pressing the corresponding number key or by clicking the button. In addition to the four similarity category buttons, there is a button to discard the sentence pair. The discard button has no shortcut key on the keyboard in order to avoid the category being chosen accidentally. The criteria for each category are visible below the sentence pair. The annotator can also see their progress for each language at the top of the page. By clicking their username at the top of the page, the user can enter their user page. Here the user can switch between the languages they were assigned to annotate, change their password and see their 100 most recent annotations and edit them.

In addition to being able to make annotations, admin users have access to special features. They can add new users, view annotation statistics per language or per user and search for and read specific annotations.

Sharing the task. Each sentence pair has to be annotated by two different annotators. We do not hand out complete batches of sentence pairs for annotation, in order to avoid dealing with unfinished batches. Instead, our tool finds the next sentence pair dynamically. Within a given language, all annotators annotate sentence pairs from the same sentence pair pool. The algorithm looks for

[5] http://mturk.com

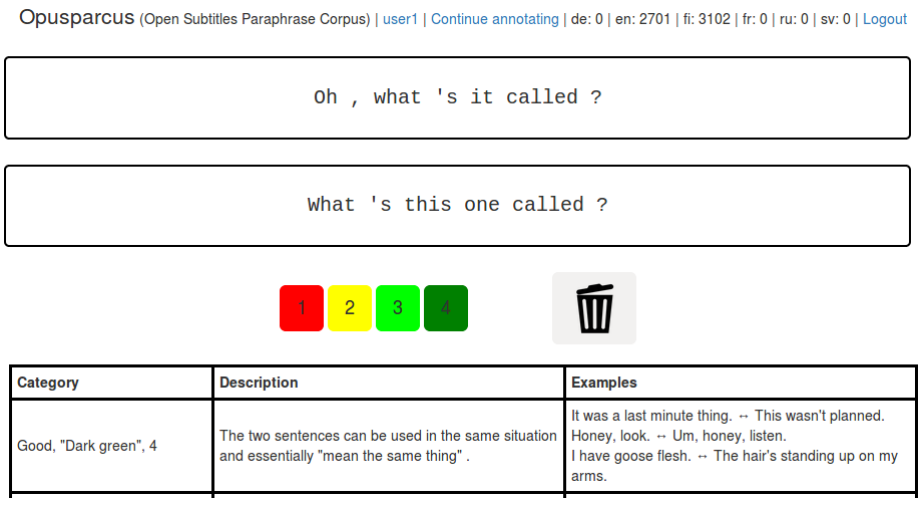

Fig. 1. Main annotation view. The lower portion of the page, containing criteria for each category, is not fully visible in this figure (See Table 1 for full criteria).

the first pair that has been annotated by another annotator, but lacks a second annotation. If such a pair is not found, the algorithm finds the first pair that has no annotations. The users can stop annotating anytime they like without feeling the pressure of having unfinished work and continue again when it is convenient for them.

2.4 Structure of the tool

The annotation tool is written in Python and it uses the Django web framework[6]. The database used is PostgreSQL[7]. The application runs in a cPouta virtual machine by CSC[8], a Finnish information and communication technology provider, but it can be run on any server, for example on Heroku[9], a cloud computing service.

We have chosen to use Django, one of the most popular web frameworks for Python. Django has a prebuilt admin page, which allows multiple admins to easily manage users without each of them having access to the backend of the tool. Django also has a database API, which allows the developer to use Django's methods instead of raw SQL commands. This makes database interactions more intuitive and concise. Additionally, Django has built-in methods for handling security risks, which is important to us, since we are dealing with users with passwords.

[6] https://www.djangoproject.com/
[7] https://www.postgresql.org/
[8] https://www.csc.fi/
[9] https://www.heroku.com/

There are two versions of our tool: one that requires registration and logging in, and one that is open for anyone to use. Each annotator for the private tool was approved by admins. This makes it time consuming to have a large group of annotators. The public tool is open for anyone, but there still has to be two annotations from two different annotators for each sentence pair. The users are tracked by their IP addresses, which is not by any means a perfect way of identifying individual users. An open tool is a good way of gathering large amounts of annotated data, but the tool has to have mechanisms for detecting and filtering out random and noisy annotations. In the end, we decided to use annotations only from the private tool.

3 Evaluation

Eighteen persons participated in the annotation effort. The annotators were recruited among researchers and students at the university, as well as family members and friends. The German data was annotated by native German speakers and a skilled speaker of German as a second language. The English data was annotated by non-native but highly skilled English speakers. The Finnish data was annotated by native Finnish speakers. The French data was annotated by a native French speaker and skilled non-native French speakers. The Russian data was annotated by native Russian speakers, and the Swedish data was annotated by native Swedish speakers. Table 2 shows the total number of paraphrases annotated as well as the number of annotators who contributed the most for each language.

Table 2. Number of annotated paraphrase pairs, discarded paraphrase pairs and primary contributors for each language. A paraphrase pair is discarded when at least one of the two annotators marks the pair as "trash" or when the annotators disagree significantly (by choosing categories that are two or three steps apart). Discarded paraphrase pairs are not included in the final data set. Primary contributors are the annotators who annotated at least 10% of the sentence pairs for a given language. Four annotators were primary contributors for two languages simultaneously and one user was a primary contributor for three languages, which explains the total number of primary contributors.

Language	Annotated paraphrase pairs	Discarded paraphrase pairs	Primary contributors
German (de)	3483	315	3
English (en)	3088	188	2
Finnish (fi)	3703	194	4
French (fr)	3847	543	4
Russian (ru)	4381	672	2
Swedish (sv)	4178	390	4
Total	22,680	2302	13

In the following, we evaluate the annotations in terms of inter-annotator agreement as well as annotation times and session lengths. We want to make sure that the annotations are good quality and that fatigue or carelessness was not a detrimental factor in the process.

3.1 Inter-annotator agreement

The results of the annotation of the Opusparcus development and test sets have been published earlier in connection with the release of the corpus [3]; a detailed breakdown is presented, showing the number of sentence pairs that end up in different categories.

The current paper extends the analysis by taking a closer look at inter-annotator agreement. It would also be interesting to study *intra*-annotator agreement (intra-rater reliability) to find out how consistently our annotators performed on data that they had already annotated before. However, we never displayed the same sentence pairs twice to the same annotator, so we cannot assess the reliability of individual annotators, only to what extent they agreed or disagreed with other annotators.

Distributions over annotation categories. The annotators were shown sentence pairs and needed to decide between five options. For every sentence pair, two annotations were obtained, because two annotators made two independent choices. Figure 2 shows the distributions of all annotation choices made, separately for each language. It is obvious that not all annotation categories occur as frequently, and there are differences across languages. The language-specific differences are explained, at least partly, by the amount of available data from which to produce sentence pairs for annotation. In a preprocessing step, the sentence pairs were ranked automatically, most "promising" sentences first. The English data set was the largest one, and 70 % of the annotated pairs turned out to be "good" or "mostly good" paraphrases. By contrast, the Swedish material was the smallest one and only about half of the pairs were tagged as paraphrases.

Discounting for chance agreement. To assess the level of agreement between annotators, Cohen's kappa score [2] is frequently used in the literature. In Cohen's own words, kappa (or κ) is "[a] coefficient of interjudge agreement for nominal scales. [...] It is directly interpretable as the proportion of joint judgments in which there is agreement, after chance agreement is excluded."

There are two main ways of computing the probability that agreement occurs by pure chance: either the distribution of proportions over the categories is taken to be equal for all annotators or the annotators have their own individual distributions, as originally suggested by Cohen [5]. To use individual distributions is complicated in our case, since we assign each sentence pair dynamically to two annotators in our annotator pool. Hence, we have a large number of batches, each annotated by different pairs of annotators. However, in practice the two approaches tend to produce very similar outcomes [5], and consequently we base

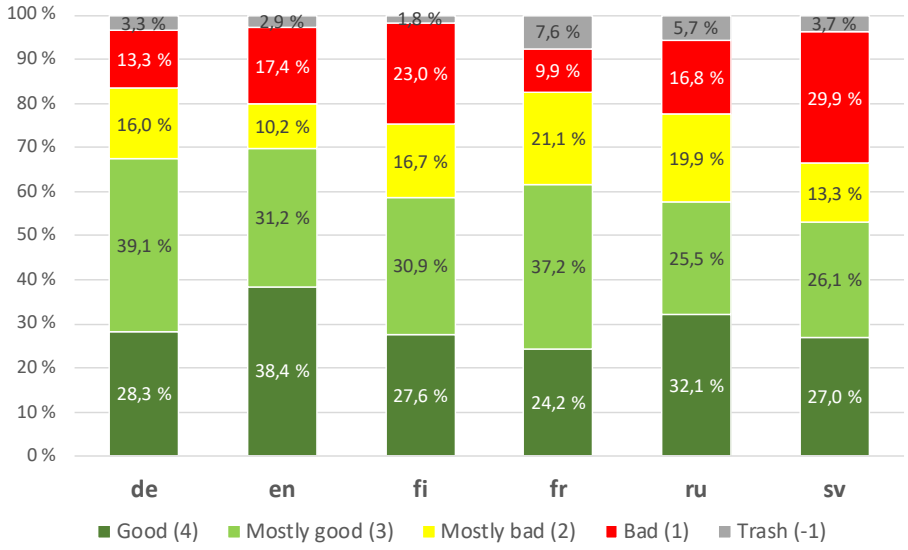

Fig. 2. Proportions of annotation events falling into each of the five annotation categories. The proportions are different in each of the six annotation languages.

our kappa calculations on one common distribution per language (shown in Figure 2). In fact, we did verify the hypothesis that both calculations produce very similar results, by examining the languages where one pair of annotators had co-annotated more than half of the sentence pairs. When we used annotator-specific distributions in the calculations, the resulting chance agreement probabilities differed by at most one percentage point from the probabilities based on one common distribution.

We evaluate inter-annotator agreement in three different ways. In the first evaluation, we retain all distinctions between the five annotation categories. This means, for instance, that we consider the annotators to disagree if one annotator opts for "Good" and the other one "Mostly good" in a particular case. The results are shown in Table 3. To verbally assess what the kappa values actually tell us about inter-annotator agreement, we have adopted a guideline proposed by Landis and Koch [8], which is commonly used for benchmarking in spite of being fairly arbitrary, as already stated in the original paper.

Table 3 demonstrates that the level of agreement between the five categories "Good", "Mostly good", "Mostly bad", "Bad", and "Trash" ranges between fair and moderate. The average level of agreement is 59.9 % with a kappa value of 0.46. Thus, in general there are differing views among the annotators on how to judge paraphrase status on this four-level scale (plus trash).

Next, we relax the conditions of agreement and merge the two categories "Good" and "Mostly good" paraphrases into one single class "Is paraphrase", and similarly merge the categories "Bad" and "Mostly bad" into one class "Is not

Table 3. Inter-annotator agreement across all five annotation categories. Results are shown for each language separately, and the arithmetic mean of all languages combined is also reported. The columns from left to right contain the language of the data, the measured level of inter-annotator agreement, the expected level of chance agreement, the kappa value, and a verbal assessment of how to interpret the kappa value, according to Landis and Koch [8].

Language	Agreement	Chance	Kappa	Assessment
de	58.1%	27.7%	0.42	moderate
en	66.4%	28.6%	0.53	moderate
fi	65.3%	25.3%	0.54	moderate
fr	55.9%	25.7%	0.41	moderate
ru	50.7%	23.9%	0.35	fair
sv	62.8%	24.9%	0.50	moderate
Average	59.9%	26.0%	0.46	moderate

paraphrase". The trash category is maintained as a third class. The results for this division are shown in Table 4. The average level of agreement is now 83.1 % with a kappa value of 0.66, which can be characterized as substantial agreement. Interestingly, very similar values are reported for the Microsoft Research Paraphrase Corpus (MSRPC) [4], where annotators were supposed to decide whether sentences from the news domain were paraphrases or not. The inter-annotator agreement for MSRPC was 84 % and kappa was 0.62. Thus, these two tasks are very similar and so is the observed level of agreement.

Table 4. Inter-annotator agreement across two main categories (paraphrase or not) plus the trash category. The columns contain the same types of information as Table 3.

Language	Agreement	Chance	Kappa	Assessment
de	82.6%	54.1%	0.62	substantial
en	89.2%	56.1%	0.75	substantial
fi	86.0%	50.0%	0.72	substantial
fr	79.1%	47.9%	0.60	moderate
ru	76.6%	47.0%	0.56	moderate
sv	84.9%	47.0%	0.72	substantial
Average	83.1%	50.3%	0.66	substantial

Since our paraphrase annotation is based on a four-grade scale ranging from "good" to "bad", we decided to evaluate agreement in a third way, where adjacent choices are considered to be in agreement. In this scheme "good" and "mostly good" match, and so do "mostly good" and "mostly bad" as well as "mostly bad" and "bad". Table 5 presents the results of this calculation. Not surprisingly, inter-annotator agreement increases (to 92.5 % on average), but so does the expected level of agreement by chance (60.7 %). The kappa score is

0.81. It is interesting to note that although the likelihood of agreement by pure chance increases, inter-annotator agreement increases to such an extent that the overall kappa score suggests "almost perfect" agreement.

Table 5. Inter-annotator agreement when adjacent annotation categories are considered to be in agreement. The columns contain the same types of information as Tables 3 and 4.

Language	Agreement	Chance	Kappa	Assessment
de	93.1%	66.6%	0.79	substantial
en	95.8%	62.5%	0.89	almost perfect
fi	95.8%	60.3%	0.90	almost perfect
fr	90.3%	63.6%	0.73	substantial
ru	87.5%	57.1%	0.71	substantial
sv	92.7%	53.9%	0.84	almost perfect
Average	92.5%	60.7%	0.81	almost perfect

Discussion. The authors behind the MSRPC corpus consider their annotation task to be "ill-defined", but they were surprised at how high inter-rater agreement was (84 %) [4]. Our setup was similar in the sense that our annotators did not typically receive any further instructions than the descriptions and examples shown in the annotation tool (see Table 1). Highest agreement is observed for English, Finnish and Swedish, languages where the people most involved in the paraphrase project performed a substantial part of the annotation effort. This indicates that deeper involvement in the project contributes to more convergent views on how to categorize the paraphrase data. Why Russian and French have the lowest degrees of agreement is unclear. These languages seemed to have the noisiest data, French because of complicated orthography, and Russian possibly because of OCR errors, which produce Latin letters into Cyrillic text.

3.2 Annotation times

Measuring annotation times reveals information on annotator behavior. Especially interesting behavior is such that would affect the reliability of the annotation effort, e.g. signs of fatigue or maliciousness. With annotation times, we mean the time elapsed between two annotation events for a user.

Many annotators started the annotation task with slow annotations. In Figure 3 we see this effect for user2 and user4. The slow start is more clearly visible for user4. The fastest times before the 200 annotation mark are slower than after that. Additionally, the times are slightly faster after about 1000 annotations. This indicates that the user first took his time annotating to get familiar with the task. Once the user figured out the nature of the work, he increased his annotating speed and maintained it or slightly increased it for the rest of

the task. The same effect is observable for user2 at the beginning of both of the annotated languages but to a lesser extent. Additionally, the annotation speed for native Russian speaker user2 decreases when he switches from annotating Russian to French. We did not observe signs of slowing down because of fatigue for any annotator. Neither did we experience any maliciousness from the users' side, e.g. very fast consecutive annotations.

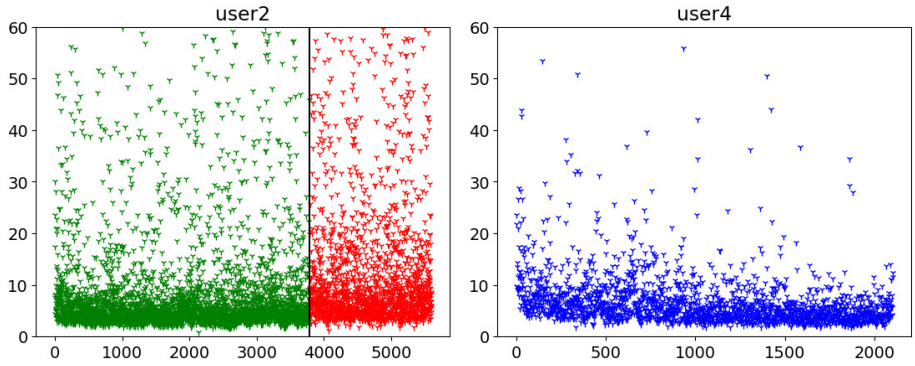

Fig. 3. Annotation times for user2 and user4. The time difference between two annotations in seconds is shown on the y-axis, and the number of annotations on the x-axis. Time differences greater than 60 seconds are excluded. Different colors represent different languages. User2 annotated Russian (green markers on the left side of the horizontal line) and French (red markers on the right side of the horizontal line), User4 annotated Swedish (blue markers).

Annotation behavior and strategies are also reflected in the amount of time people spend annotating in a single session. We define an annotation session to consist of annotation events where the time between two consecutive events is less than five minutes. Figure 4 shows the number of sessions of different lengths, as well as the cumulative proportion of annotation events for all users.

Most of the annotation sessions are relatively short, and consequently a large proportion of the annotations come from short sessions. As we mentioned above, we cannot assess the reliability of individual annotators using intra-annotator agreement measures, but a look at the session lengths and annotation results suggests no difference in quality of the annotators who worked in short sessions in comparison to those who preferred longer sessions. Based on this we assume that annotator fatigue does not affect the quality of the resulting data set to a large degree.

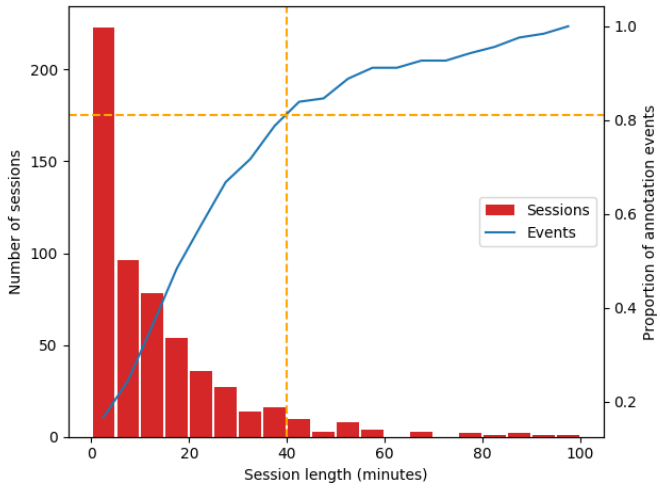

Fig. 4. Session lengths and the cumulative proportion of annotation events for all users. The x-axis shows the session length in minutes, divided into five-minute bins. The y-axis on the left shows the number of sessions and the y-axis on the right shows the proportion of annotation events. The blue line shows the cumulative proportion of annotation events for each five-minute bin. For example, a little over 80% of all the annotations come from sessions that lasted for less than 40 minutes (dashed orange lines).

4 Discussion and conclusion

Could the inter-annotator agreement be higher? The creators of MRSPC [4] believe that in their task agreements could be improved through practice and discussion among the annotators. However, they also observed that attempts to make the task more concrete resulted in degraded intra-annotator agreement.

Others have called for more linguistically informed data sets with more fine-grained annotation categories. [13] There is a trade-off, however, between annotation speed and complexity of the annotation task. We have favored a fairly simple intuitive annotation scheme.

The Opusparcus data sets have been used successfully in machine learning for training and evaluating automatic paraphrase detection. [14]

In future work, if we wish to recruit a larger pool of annotators through crowdsourcing, attention needs to be paid to better tracking of the reliability and consistency of individual annotator performance. Additionally, although the colloquial style of the data makes it interesting to work with, the task could be made even more enjoyable, for instance through gamification.

Acknowledgments

We are grateful to the following people for helping us in the annotation effort: Thomas de Bluts, Aleksandr Semenov, Olivia Engström, Janine Siewert, Carola Carpentier, Svante Creutz, Yves Scherrer, Anders Ahlbäck, Sami Itkonen, Riikka Raatikainen, Kaisla Kajava, Tiina Koho, Oksana Lehtonen, Sharid Loáiciga Sánchez, and Tatiana Batanina.

We would also like to thank Hanna Westerlund, Martin Matthiesen, and Mietta Lennes for making Opusparcus available at the Language Bank of Finland (http://www.kielipankki.fi).

The project was supported in part by the Academy of Finland through Project 314062 in the ICT 2023 call on Computation, Machine Learning and Artificial Intelligence.

References

1. Barrón-Cedeño, A., Vila, M., Martí, M.A., Rosso, P.: Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. Computational Linguistics **39**, 917–947 (2013)
2. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960), https://doi.org/10.1177/001316446002000104
3. Creutz, M.: Open Subtitles Paraphrase Corpus for Six Languages. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018)
4. Dolan, B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the Third International Workshop on Paraphrasing (IWP2005) at the Seond International Joint Conference of Natural Language Processing (IJCNLP-05). Asia Federation of Natural Language Processing (January 2005), https://www.microsoft.com/en-us/research/publication/automatically-constructing-a-corpus-of-sentential-paraphrases/
5. Eugenio, B.D., Glass, M.: Squibs and discussions: The kappa statistic: A second look. Computational Linguistics **30**(1) (2004), http://www.aclweb.org/anthology/J04-1005
6. Ganitkevitch, J., Van Durme, B., Callison-Burch, C.: PPDB: The paraphrase database. In: Proceedings of NAACL-HLT. pp. 758–764. Association for Computational Linguistics, Atlanta, Georgia (June 2013), http://cs.jhu.edu/~ccb/publications/ppdb.pdf
7. Kovatchev, V., Martí, T., Salamó, M.: ETPC – a paraphrase identification corpus annotated with extended paraphrase typology and negation. In: LREC (2018)
8. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**(1), 159–174 (1977), http://www.jstor.org/stable/2529310
9. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). Portorož, Slovenia (May 2016)

10. Lison, P., Tiedemann, J., Kouylekov, M.: OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora. In: Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (May 2018)
11. Paetzold, G.H., Specia, L.: Collecting and exploring everyday language for predicting psycholinguistic properties of words. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 669–1679. Osaka, Japan (December 2016)
12. Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., Callison-Burch, C.: PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers). pp. 425–430. Association for Computational Linguistics, Beijing, China (July 2015)
13. Rus, V., Banjade, R., Lintean, M.C.: On paraphrase identification corpora. In: LREC (2014)
14. Sjöblom, E., Creutz, M., Aulamo, M.: Paraphrase detection on noisy subtitles in six languages. In: Proceedings of W-NUT at EMNLP. Brussels, Belgium (2018)
15. Snow, R., O'Connor, B.T., Jurafsky, D., Ng, A.Y.: Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of EMNLP (2008)
16. Vila, M., Bertrán, M., Martí, M.A., Rodríguez, H.: Corpus annotation with paraphrase types: new annotation scheme and inter-annotator agreement measures. Language Resources and Evaluation **49**, 77–105 (2015)
17. van der Wees, M., Bisazza, A., Monz, C.: Measuring the effect of conversational aspects on machine translation quality. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. pp. 2571–2581. Osaka, Japan (December 2016)
18. Yimam, S.M., Gurevych, I., Eckart de Castilho, R., Biemann, C.: Webanno: A flexible, web-based and visually supported system for distributed annotations. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 1–6. Association for Computational Linguistics, Sofia, Bulgaria (August 2013), http://www.aclweb.org/anthology/P13-4001