Analyzing the Effect of Linguistic Instructions on Paraphrase Generation

Teemu Vahtola[1] **Songbo Hu**[2] **Mathias Creutz**[1]
Ivan Vulić[2] **Anna Korhonen**[2] **Jörg Tiedemann**[1]
[1]Unversity of Helsinki, Finland
[2]Language Technology Lab, University of Cambridge, UK
{teemu.vahtola, mathias.creutz, jorg.tiedemann}@helsinki.fi
{sh2091, iv250, alk23}@cam.ac.uk

Abstract

Recent work has demonstrated that large language models can often generate fluent and linguistically correct text, adhering to given instructions. However, to what extent can they execute complex instructions requiring knowledge of fundamental linguistic concepts and elaborate semantic reasoning?

Our study connects an established linguistic theory of paraphrasing with LLM-based practice to analyze which specific types of paraphrases LLMs can accurately produce and where they still struggle. To this end, we investigate a method of analyzing paraphrases generated by LLMs prompted with a comprehensive set of systematic linguistic instructions. We conduct a case study using GPT-4, which has shown strong performance across various language generation tasks, and we believe that other LLMs may face similar challenges in comparable scenarios.

We examine GPT-4 from a linguistic perspective to explore its potential contributions to linguistic research regarding paraphrasing, systematically assessing how accurately the model generates paraphrases that adhere to specified transformation rules. Our results suggest that GPT-4 frequently prioritizes simple lexical or syntactic alternations, often disregarding the transformation guidelines if they overly complicate the primary task.

1 Introduction

Large language models (LLMs) can, without doubt, generate fluent and linguistically correct language with relevance to given prompts (Sottana et al., 2023). However, to what extent can they follow complex linguistic instructions and execute them in a meaningful way? To this end, we propose a systematic approach for analyzing LLMs in performing explicit, theoretically grounded paraphrase transformations in English, using a validated list of 25 linguistic operations (Bhagat and Hovy, 2013).

It is necessary to have knowledge of fundamental linguistic concepts to follow those specialized instructions. This study provides insight into the capabilities and limitations of LLMs when faced with such a demanding task. Extending our understanding on the connections between linguistically grounded theories of paraphrasing and the practical abilities of LLMs, we hope to improve paraphrasing performance with explicit linguistic operations, with potential applications in text simplification (Nisioi et al., 2017), computer-assisted language learning (Mayhew et al., 2020), machine translation (Callison-Burch et al., 2006; Mehdizadeh Seraj et al., 2015) and automatic summarization (Gupta and Gupta, 2019).

We conduct a case study analyzing paraphrases generated by a representative state-of-the-art LLM, GPT-4 (Achiam et al., 2023), focusing on the abilities of the model to create meaning-preserving and diverse paraphrases using systematic instructions related to the 25 paraphrasing categories of Bhagat and Hovy (2013). Our analysis further looks into the complexity of individual transformations and how GPT-4 copes with them with varying degrees of in-context learning (Brown et al., 2020; Dong et al., 2024). Furthermore, we study how humans perceive the produced paraphrases in terms of semantic similarity and linguistic diversity.

The **contributions** of the paper are the following: **(1)** Our study connects a descriptive theory of paraphrasing with generative language models and human perception of sentence-level semantic similarity. **(2)** We conduct a limited case study, in-

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 755–766
March 3-4, 2025 ©2025 University of Tartu Library

Full Name	**Abbreviation**
synonym substitution	synonym
antonym substitution	antonym
converse substitution	converse
change of voice	voice
change of person	person
pronoun/co-referent substitution	pron./co-ref.
repetition/ellipsis	repetition
function word variations	func. word
actor/action substitution	actor/action
verb/’semantic-role noun’ substitution	verb/sem. noun
manipulator/device substitution	manip./device
general/specific substitution	gen./spec.
metaphor substitution	metaphor
part/whole substitution	part/whole
verb/noun conversion	verb/noun
verb/adjective conversion	verb/adj.
verb/adverb conversion	verb/adv.
noun/adjective conversion	noun/adj.
verb-preposition/noun substitution	vp./noun
change of tense	tense
change of aspect	aspect
change of modality	modality
semantic implication	sem. impl.
approximate numerical equivalences	num. eq.
external knowledge	ext. knowl.

Table 1: This table lists all the paraphrase defining transformations from Bhagat and Hovy (2013), along with their abbreviations as used throughout this paper, particularly in Figure 2.

vestigating a systematic approach for analyzing the ability of LLMs to follow complex instructions and how different degrees of complexity influence the result of generated paraphrases. **(3)** To facilitate further research on controlled paraphrase generation and the variability of human language, we publicly release the set of automatically generated sentence pairs exhibiting diverse transformations, accompanied by their corresponding human annotations, at `https://github.com/Helsinki-NLP/paraphrase-instructions`.

2 Background

Paraphrasing denotes variability in expressed meaning. Vague definitions such as this one are typical ways of framing the concept of paraphrasing in NLP research (Vila et al., 2014). However, previous research in (computational) linguistics has presented various, more fine-grained typologies that outline the linguistic transformations defining paraphrasing.

Through the lens of existing paraphrase theories (Mel’čuk, 2012; Honeck, 1971; Harris, 1957), Bhagat and Hovy (2013) empirically validate paraphrase examples from two corpora: the Multiple-translation Corpus (Huang et al., 2002) and the Microsoft Research Paraphrase Corpus (MRPC) (Dolan et al., 2004). They outline 25 concrete operations with systematic linguistic instructions of transformations that produce sentences with near-equivalent meaning. The perspective to these operations is mostly lexical, focusing on the specific lexical changes that can be made at the sentence or phrase level to create paraphrases (Bhagat and Hovy, 2013). However, several of the operations trigger changes that would traditionally fall within the domain of syntactic theory. One such operation would be *ellipsis*. We list all the transformations defined in Bhagat and Hovy (2013) in Table 1.

Correctly applying these transformations in automatic paraphrase generation requires the model to process fundamental linguistic concepts and accurately recognize the phrase-level transformations triggered by the defined lexical operations. Furthermore, not every transformation is appropriate for every context. Therefore, the model must thoroughly process the definition and have intricate semantic reasoning abilities to construct sentence pairs that are appropriately suited for the intended transformation. To this end, we analyze the capabilities of LLMs in producing paraphrastic sentence pairs given systematic linguistic instructions. The transformations span from simple local changes, such as synonym substitution (*to build/to construct*) or change of aspect (*studying/studies*), to more complex alterations, such as converse substitution (*buy/sell*).

Along with systematic, descriptive definitions, Bhagat and Hovy (2013) provide 1–3 examples for each paraphrase transformation. Synonym substitution, for example, is defined as follows:[1]

> *Replacing a word/phrase by a synonymous word/phrase, in the appropriate context, results in a paraphrase of the original sentence/phrase. This category covers the special case of genitives, where the clitic ’s is replaced by other genitive indicators like of, of the, and so forth. This category also covers near-synonymy, that is, it allows for changes in evaluation, connotation, and so on, of words or phrases between paraphrases. Example:*
>
> 1. *Google bought YouTube. ↔ Google acquired YouTube.*
> 2. *Chris is slim. ↔ Chris is slender. ↔ Chris is skinny.*

These definitions followed by a small number of examples can be utilized as such in prompts for

[1] For an exhaustive list of the definitions and examples of the paraphrase transformations, we refer the reader to Bhagat and Hovy (2013).

few-shot in-context learning, where an LLM is instructed to generate sentence pairs incorporating the specific transformations. As few-shot learning has been shown to be an effective approach for applying LLMs in various tasks (Brown et al., 2020), we focus on leveraging the framework of Bhagat and Hovy (2013) for evaluating few-shot learning with GPT-4 across a wide range of linguistic operations related to paraphrasing.

In a contemporary work, Meier et al. (2024) analyze various paraphrase types generated by GPT-3.5 by employing more abstract linguistic definitions of paraphrase phenomena as defined by Barrón-Cedeño et al. (2013) and Vila et al. (2014). These phenomena comprise abstract linguistic properties, such as changes based on *morpholexicon*, *structure*, and *semantics*. Each of these classes is further divided into subclasses and types, where one type (e.g., *same-polarity substitution*) can include multiple concrete transformations (e.g., *synonymy*, *general/specific substitution*, or *exact/approximate alternations*) (Barrón-Cedeño et al., 2013). Meier et al. (2024) select 10 of such types for their analysis. Many of the selected types focus on local substitutions, such as *inflectional changes*, *punctuation changes*, and *spelling changes*, while only a few focus on global changes that require intricate contextual understanding. As opposed to this, we use the typology of Bhagat and Hovy (2013), which provides an empirically validated list of concrete linguistic transformations for generating paraphrases, along with their linguistic definitions and examples, covering a wider range of local and contextual transformations. These concrete definitions enable a precise assessment of which specific linguistic features are well-represented by the chosen LLM and which areas the model still lacks sufficient knowledge in.

3 Experimental Details

3.1 Data Generation

We apply GPT-4[2] (Achiam et al., 2023) via the API to generate potential paraphrase pairs following a comprehensive list of paraphrasing operations (Bhagat and Hovy, 2013). We selected GPT-4 as a representative and powerful LLM after initial experiments with various LLMs suggested that GPT-4 produced the most fluent output, which is essential for accurately analyzing our setting. We use the default values provided by the OpenAI package for all hyper-parameters. Additionally, we configured the response output of the model to JSON mode, following the text generation guidelines recommended by OpenAI.[3]

Template 1: System Prompt

You are a helpful assistant designed to output JSON.

Synonym substitution: Replacing a word/phrase by a synonymous word/phrase, in the appropriate context, results in a paraphrase of the original sentence/phrase. This category covers the special case of genitives, where the clitic 's is replaced by other genitive indicators like of, of the, and so forth. This category also covers near-synonymy that is, it allows for changes in evaluation, connotation, and so on, of words or phrases between paraphrases. Example:

(a) Google bought YouTube. $\Longleftrightarrow$ Google acquired YouTube.
(b) Chris is slim. $\Longleftrightarrow$ Chris is slender. $\Longleftrightarrow$ Chris is skinny.

Template 2: User Prompt for Simple Sentences

Could you give me 10 more examples following the given description? Return the examples as a list of json objects.

Template 3: User Prompt for Complex Sentences

Could you give me 15 more examples following the given description? Generate 5 compound sentences, 5 complex sentences, and 5 compound-complex sentences to showcase a variety of syntactic structures. It is enough to perform the operation in only one of the clauses. Return the examples as a list of json objects.

Figure 1: Prompt templates we use for generating the paraphrases.

[2] gpt-4-turbo-2024-04-09 is used.

[3] https://platform.openai.com/docs/guides/text-generation/json-mode

Sentence type	**Example sentence**
Simple	The company employs 100 workers.
Simple	The teacher explained the concept clearly.
Complex	Although it was raining, we played football.
Compound-complex	She loves running in the morning, and when she returns, she makes breakfast.
Compound-complex	She opened a savings account, and she deposited her birthday money, while her parents watched proudly.

Table 2: A randomly sampled set of five generated sentences along with their corresponding sentence types.

In paraphrase generation, a set of source sentences is typically given, and the task is to generate target sentences with the same meaning. In our experiment, however, we let the model generate both the source and the target sentences given the definition and 1–3 examples. Since not all transformations are possible on just any source sentence, this allows for the model to come up with suitable source/target pairs for each transformation. Moreover, we believe that our approach more effectively encourages the model to engage in deeper semantic reasoning. When provided with a source sentence, the model is already primed towards a certain transformation, potentially making the task simpler. In contrast, when given only a description of a paraphrase operation along with a few examples, the model must first fully identify the relationship between the description and the examples to generate an appropriate source sentence.

We leverage the definitions and examples given in Bhagat and Hovy (2013) as prompts for the LLM, and request it to produce 25 sentence pairs following the definitions of each of the 25 transformations. Our initial experiments suggest that when we only use the definition and the examples as the prompt, the model predominantly generates rather short sentences with simple syntactic structures, which may constrain its ability to execute more complex paraphrasing transformations. Therefore, we explicitly prompt the model to generate *compound*, *complex* and *compound-complex* sentences. Table 2 presents randomly sampled examples of various sentence types.

The prompts are composed of two parts: system prompts and user prompts, as illustrated in Figure 1. For each paraphrase operation described in Bhagat and Hovy (2013), we construct a system prompt following Template 1, adapting the transformation definition and examples as needed. To generate simple sentence pairs, we use Template 2 as the user prompt. For syntactically complex sentence pairs, we employ Template 3. These templates are specifically crafted to guide the model in producing sentence pairs with varying levels of syntactic complexity.

Eventually, we generate 10 simple sentences and 5 each of compound, complex, and compound-complex sentences for every paraphrase transformation.

3.2 Collecting Annotations

We collect manual annotations by four independent annotators to the generated sentences to answer three key questions: **(1)** Does the generated sentence pair follow the given definition of a paraphrase transformation? **(2)** Are the generated sentences paraphrases of each others? **(3)** To what extent are the generated sentences semantically equivalent? Each sentence pair is annotated by all annotators. For evaluating the third question concerning semantic equivalency, we follow previous work involving manually annotating paraphrases (Creutz, 2018; Kanerva et al., 2021), and use the four-point Likert scale with the following scores and associated descriptions: *4: Full paraphrases, 3: Paraphrases in some contexts, 2: Semantically similar sentences but not paraphrases, 1: Unrelated sentences.*

The annotators are fluent speakers of English, and knowledgeable of fundamental linguistic concepts.[4] They are provided with the definitions and examples of each paraphrase operation, as well as

[4] In addition to some of the authors, we involve colleagues as annotators, bringing the total number of annotators to four. Each example is annotated by all four annotators to better capture the range of human variability and subjectivity in evaluating paraphrases.

Annotator	Para. Acc.	Trans. Acc.
1	0.824	0.688
2	0.869	0.677
3	0.821	0.677
4	0.872	0.744
Average	0.847	0.696

Table 3: Model performance on paraphrase accuracy (Para. Acc.) and transformation accuracy (Trans. Acc.), evaluated by four annotators. Paraphrase Accuracy measures whether the generated sentence pairs qualify as paraphrases. Transformation Accuracy measures whether the sentence pairs adhere to the predefined transformation operation.

the generated sentence pairs. Appendix A shows a screenshot of the customized annotation tool.

4 Results and Discussion

4.1 Paraphrase and Transformation Accuracy

We first focus on evaluating the model's performance with respect to the aforementioned questions **(1)** and **(2)**. By *transformation accuracy* we understand the proportion of generated sentence pairs that successfully follow the desired transformation operation (Question 1). By *paraphrase accuracy* we understand the proportion of generated sentence pairs that are true paraphrases (Question 2).

Table 3 presents the obtained paraphrase and transformation accuracies for all the generated sentence pairs, as assessed by our four expert annotators. It can be seen that GPT-4 generally performs well at providing alternative expressions that convey the same meaning (average paraphrase accuracy is 84.7 %). However, it shows clear limitation in accurately following the specified transformations (average transformation accuracy is 69.6 %). Furthermore, the evaluation results indicate that the scores provided by the annotators are consistent and similar. To demonstrate the reliability of our measurement approach, we compute Fleiss' Kappa for the two binary variables in our dataset: paraphrase accuracy and transformation accuracy. The Fleiss' Kappa scores were 0.53 for paraphrase accuracy and 0.71 for transformation accuracy. These scores indicate moderate and substantial agreement among annotators, respectively, demonstrating the robustness of our evaluation methodology and the inherent subjectivity in evaluating paraphrases.

Figure 2 presents the paraphrase and transformation accuracies for each individual paraphrase transformation operation, averaged over the different annotators. The figure clearly illustrates that the model achieves high results in paraphrase and transformation accuracies for specific, local changes, such as synonym substitution, antonym substitution, change of voice, and change of aspect. In contrast, the model appears to struggle with transformations that require a more nuanced understanding of context, such as converse substitution, actor/action substitution, or verb/adverb conversion.

Next, we provide an analysis across the various types of paraphrase transformations to better understand where the model succeeds and the kinds of mistakes it makes when it struggles.

4.2 Qualitative Analysis

Figure 3 illustrates the correlation between paraphrase and transformation accuracy. All transformations except one are located either in the top row or the right-most column of Figure 3, meaning that either the transformation or the paraphrasing was performed successfully (accuracy > 75 %). This is an excellent result.

The top right corner represents the most successful transformations, with a high transformation accuracy combined with a high paraphrase accuracy. There are ten such transformations corresponding to 40 % of all 25 types. These are fairly straightforward or local transformations, such as replacing synonyms within sentences (*started* vs. *began*) or substituting a word with its negated antonym (*happy* vs. *not sad*). Approximate numerical equivalence (mapping between units) and external knowledge (the *Louvre* is a museum) are also found here. This outcome is not too surprising given that a number of well-known paraphrase corpora, such as PPDB (Ganitkevitch et al., 2013) and MRPC (Dolan et al., 2004), contain similar examples (cf. Rajana et al., 2017; Bhagat and Hovy, 2013) and the model has most likely been trained on such data. Moreover, knowing that 125 miles corresponds to about 200 kilometers can be memorized from the training data rather than actually being calculated by the model.

There are more transformations in the right-most column (21) than in the top row (13), indicating that the system more accurately generates paraphrases than the desired transformation types.

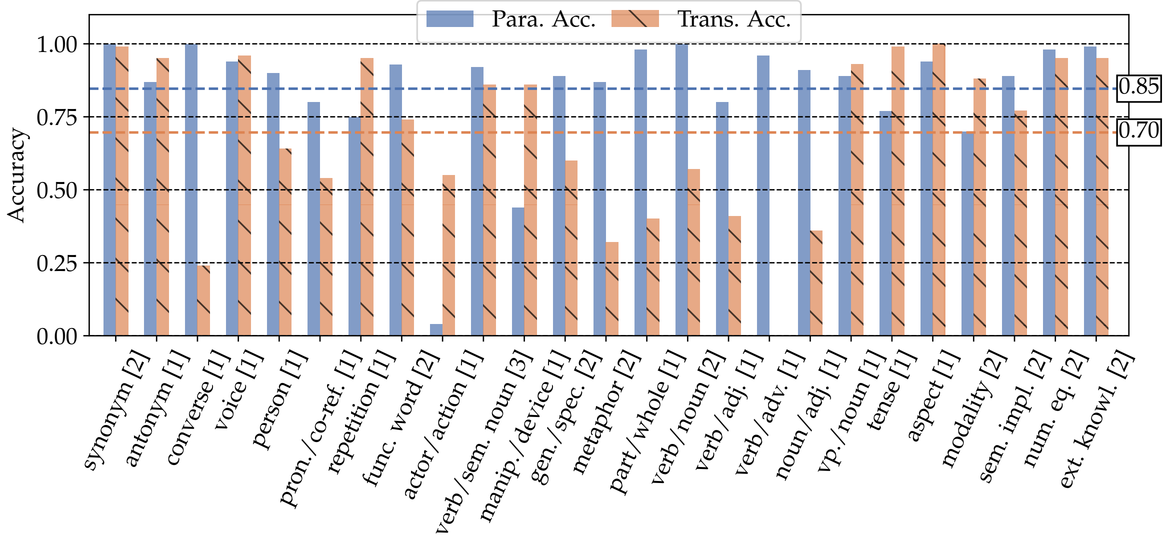


Figure 2: Model performance on Paraphrase Accuracy (Para. Acc.) and Transformation Accuracy (Trans. Acc.). This figure highlights the aggregated mean values for each metric across the 25 transformation operations, indicated by dashed horizontal lines. Abbreviations representing each operation are used, with full names provided in Table 1. All the results are based on annotations by four expert annotators. The number of examples provided by Bhagat and Hovy (2013) for each operation is noted in square brackets. For example, *synonym [2]* indicates two examples for the synonym substitution operation.

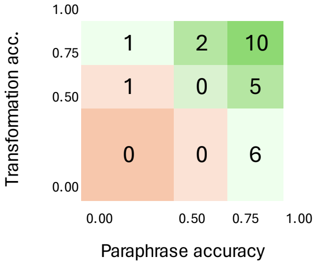


Figure 3: Distribution of the 25 transformations into 3×3 distinct bins depending on paraphrase and transformation accuracy. There are three intervals on the axes, corresponding to accuracies between 0.0 and 0.5, above 0.5 up to 0.75, and between 0.75 and 1.0, respectively.

Failures to capture the desired transformation, while still producing a valid paraphrase, include the following mistakes: (1) using change of voice (*buy/be bought*) instead of converse substitution (*buy/sell*), verb/noun conversion (*to try/make an attempt*) or verb/adjective conversion (*to clean/make clean*), (2) confusion between the categories part/whole (*room/house*) vs. general/specific (*astronomical body/sun*), (3) poor metaphor generation capacity (*"a sea of people"* vs. *"an ocean of people."*). Apart from the very demanding task of creating metaphors, the failures here are artefacts of somewhat artificial, grammatical distinctions, such that participle forms of verbs (*interested*) do not qualify as adjectives (*curious*).

Failures to reliably produce paraphrases while still being faithful to the desired transformation (top row, left and center) comprise manipulator/device substitution (*"The photographer (vs. camera) took stunning photos"*) and change of modality (*finds/can find*), which in fact can alter the meaning. Nevertheless these types have been included in the paraphrase taxonomy of Bhagat and Hovy (2013), which may seem odd. While it is possible to produce paraphrases within the limits of the above transformations, it requires strong semantic reasoning abilities from the model. It must first generate a source sentence that is comprised of (potentially limited) concepts that are suitable for such transformations and then create an effective paraphrase as a target sentence.

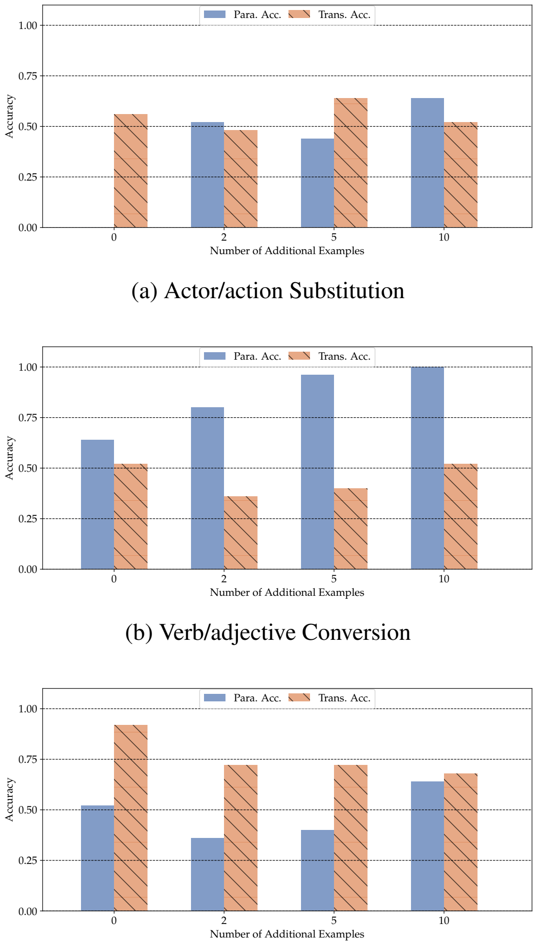
Additionally, the removal of repetition (ellipsis) is sometimes performed too aggressively and the meaning is not preserved (*"The cat chased the mouse and the dog chased the squirrel."* vs. *"The cat chased the mouse and the dog did, too."*). The model may overly prioritize elliptical constructions similar to the example prompt, failing to generalize to different kinds of sentence structures. Specifically in the above example, it fails to recognize that omitting the object in the second clause changes the meaning as the repeated part is the predicate rather than the object.

The poorest result is obtained for actor/action substitution (center left), which mostly generates semantically or grammatically incorrect sentence pairs: *"I love teaching."* vs. *"I love teacher."* This operation is particularly challenging, as it demands deep contextual understanding. Merely replacing an actor, such as *teacher* with a corresponding action, such as *teaching*, is not sufficient for preserving the original meaning if the context does not allow it. The example Bhagat and Hovy (2013) provide for actor/action substitution is: *"I dislike rash drivers (vs. driving)."* It is possible that the training data has limited examples of correctly applying this operation, which can result in poor accuracy in recognizing appropriate concepts and contexts.

4.3 Semantic Equivalence

Our annotators assessed three criteria (Section 3.2), two of which have been analyzed thoroughly above: transformation accuracy (Question 1) and paraphrase accuracy (Question 2). Question 3 on semantic equivalency remains to be studied. Next, we compare the binary annotations of paraphrase accuracy (Question 2) to the 4-level Likert scale annotations (Question 3). The four-level scale offers a more nuanced view on semantic equivalency than the binary paraphrase classification.

Two out of four annotators had virtually perfect correlation between the binary paraphrase category and Likert scale values 4 and 3 ("full paraphrases" and "paraphrases in some contexts"). The other two annotators did very similarly, but in addition, there was a small number of data points (around 3 % and 6 %) in which Likert scale 3 ("paraphrases in some context") rendered the "not paraphrases" binary classification. An example where both annotators classified the example as a non-paraphrase but still assigned it a Likert scale score of 3 is: *"The driver (vs. car) accelerated quickly, but the passenger felt nervous."* Overall, the binary annotations closely align with the detailed results from the 4-level Likert scale. Consequently, we do not conduct further analysis on the relationship between the different annotation granularities but reserve it for future work.



(a) Actor/action Substitution

(b) Verb/adjective Conversion

(c) Manipulator/device Substitution

Figure 4: Model performance for **(a)** Actor/action Substitution, **(b)** Verb/adjective Conversion, and **(c)** Manipulator/device Substitution with increasing number of in-context learning examples. For each example, we append it to the system prompt as shown in Template 1 in Figure 1. Results are based on annotations by one expert annotator.

4.4 Additional In-context Learning

Bhagat and Hovy (2013) do not provide the same number of examples for all of the 25 transformations. In fact, 15 transformations have only 1 example, 9 have 2, and 1 has 3 examples.[5] As LLMs have been shown to generalize well from few-shot learning (Brown et al., 2020), and as we observe a slight correlation between the number of examples and paraphrase and transformation accuracy[6], we experiment whether providing additional examples

[5] The example numbers corresponding to each transformation are shown in Figure 2.

[6] We report a mean paraphrase accuracies of 0.81, 0.91, and 0.90, and mean transformation accuracies of 0.66, 0.77, and 0.80 for operations that have 1, 2, and 3 examples, respectively.

improves GPT-4's performance in the more difficult paraphrase operations (left-most column, and bottom right of Figure 3). The operations we focus on are actor/action substitution, verb/adjective conversion, and manipulator/device substitution, each having 1 provided example in the original prompt.

Figure 4 presents the model accuracies for paraphrasing and the specified transformations for three operations that GPT-4 struggles with. When we add 2, 5, and 10 additional hand-crafted examples to the prompt, we do not see consistent improvement. Additional examples may improve the paraphrasing results, but transformation accuracy does not increase. In fact, higher paraphrase accuracy might even be detrimental to transformation accuracy, because the model prioritizes paraphrasing, if the two criteria seem conflicting. The inconsistency in improving with additional ICL examples suggests that these specific transformations may be challenging to process, possibly due to a lack of training data involving such transformations. Further research is necessary for a deeper understanding of this phenomenon.

5 Related Work

Previous work related to diverse paraphrasing has studied the generation of specific linguistic features, for instance on lexical (e.g., Thompson and Post, 2020) or syntactic level (Iyyer et al., 2018; Chen et al., 2019; Sun et al., 2021, *i.a.*), or controlling for various granularities (Vahtola et al., 2023).

Additionally, previous research has presented various taxonomies of paraphrase types for better understanding of the diverse paraphrase phenomena. Vila et al. (2014) propose a typology of 24 paraphrase types spanning three levels of granularity, while Dutrey et al. (2010) define rephrasing modifications extracted from the revision history of Wikipedia. Less fine-grained categorizations can include for instance differences in specificity or tone (Kanerva et al., 2021). Bhagat and Hovy (2013) propose a list of 25 empirically validated paraphrase transformations with a systematic definition and examples of each transformation.

Detection and generation of diverse paraphrases leveraging a corpus of various paraphrase types (Kovatchev et al., 2018) has been proposed (Wahle et al., 2023). In a concurrent work, Meier et al. (2024) leverage the linguistic phenomena defined in Barrón-Cedeño et al. (2013) to generate specific types of paraphrases. Meier et al. (2024) also gather human annotations to analyze the accuracy of GPT-3.5 across the different paraphrase types and to evaluate how human annotators rank the generated paraphrases. Their findings are in line with ours, suggesting that LLMs struggle with performing more complex paraphrase transformations. Conversely to the framework of paraphrase operations that we use, the phenomena outlined in Barrón-Cedeño et al. (2013) can often manifest themselves in various surface-form alternations (i.e., one *phenomenon* can include multiple *operations*) as they attempt to capture the general phenomena rather than providing specific mechanisms for paraphrasing. Furthermore, we focus on analyzing the performance of LLMs on various specific paraphrase transformations given their detailed linguistic definitions, and connect the theoretical perspectives of paraphrasing with generative language models and human understanding of semantic similarity.

Another line of related work has focused on benchmarking various pretrained language models, such as BERT (Devlin et al., 2019), across a diverse range of downstream tasks, e.g., GLUE (Wang et al., 2018), SentEval (Conneau and Kiela, 2018), and SICK (Marelli et al., 2014), or a limited range of linguistic phenomena (Marvin and Linzen, 2018; Jumelet and Hupkes, 2018; Ettinger, 2020; Vahtola et al., 2022). Diverging from this line of work, we focus on the capabilities of one state-of-the-art LLM and connect human perception of semantic equivalence to the theory and practice of diverse paraphrasing. In particular, we propose a method and conduct a pilot study to analyze how LLMs manage semantic abstractions in the context of systematically defined paraphrase transformations.

6 Conclusions

In this paper, we design a methodology for testing LLMs to analyze whether they can follow theoretically motivated instructions in the case of paraphrase generation. We utilize explicit linguistic prompts to guide complex transformations and evaluate the results based on human assessment.

Using this framework, we conduct a focused case study on the capabilities of GPT-4 in accurately generating paraphrases. This study is based on 25 paraphrase transformations provided in Bhagat and Hovy (2013), whose definitions of the transformations serve as prompts for few-shot learning. We have customized a web-interface for collecting manual annotations for the generated sentences in order to assess how accurately the model produces paraphrases that follow the specified transformations.

Our findings indicate that GPT-4 can effectively follow detailed linguistic instructions to generate paraphrastic sentence pairs through simple, local transformations. However, it often prioritizes simple lexical or syntactic substitutions for paraphrasing instead of following specified transformation guidelines. This is especially true when the transformations trigger more complex alternations, indicating limitations in controllability and its ability to process complex linguistic instructions. Furthermore, increasing the number of examples for few-shot in-context learning does not seem to improve the model's ability to accurately produce paraphrase pairs involving complex operations. This suggests that the model may still lack sufficient proficiency in these linguistic structures. Future work could include a more comprehensive evaluation of how additional few-shot examples, encompassing a broader range of operations, influence performance.

The presented methodology opens many alternative directions for further research. The use of systematic linguistic instructions in text generation tasks is still very much under-explored. Theoretically controlled prompts may help to further understand the abilities of LLMs to generalize and follow explicit rules and guidelines. Such prompts can also be used to compare and benchmark different models about their abstraction capabilities, and the analysis of the results can also be combined with interpretability studies of the network itself in case model weights are openly available.

Limitations

We cover a comprehensive list of transformations, which requires substantial annotations to properly analyze the effect of the instructions. The number of examples for each prompt is still limited in our study but provides a systematic view on linguistically motivated paraphrase generation. Another limitation is the focus on one particular model, GPT-4. Future work could compare the results to other models to deepen our understanding of what and how LLMs learn about human language, even though this is a moving target that is impossible to handle exhaustively. Preliminary studies indicated that GPT-4 is better in handling the complex instructions we used than other available models. This motivated our choice to look at the limitations of state-of-the-art generative models as GPT-4 abilities in this space currently serve as an upper bound for all the other LLMs. Additional prompt engineering may also be possible to further push the results, and chain-of-thought experiments would also be interesting to study in connection with the task. Finally, we would also like to extend the experiments and annotations in order to expand the dataset and the analyses that can be made on top of the collection.

Acknowledgements

At the time of this research, Teemu Vahtola was supported by the Behind the Words project, funded by the Research Council of Finland. Songbo Hu is supported by the Cambridge International Scholarship. We would like to acknowledge the annotators for their valuable contributions to this work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alberto Barrón-Cedeño, Marta Vila, M. Antònia Martí, and Paolo Rosso. 2013. Plagiarism meets paraphrasing: Insights for the next generation in automatic plagiarism detection. *Computational Linguistics*, 39(4):917–947.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the Human Language Technology Conference of the NAACL,* *Main Conference*, pages 17–24, New York City, USA. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning.

Camille Dutrey, Houda Bouamor, Delphine Bernhard, and Aurélien Max. 2010. Local modifications and paraphrases in wikipedia's revision history. *Procesamiento del lenguaje natural*, 46:51–58.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764, Atlanta, Georgia. Association for Computational Linguistics.

Som Gupta and Sanjai Kumar Gupta. 2019. Abstractive summarization: An overview of the state of the art. *Expert Systems with Applications*, 121:49–65.

Zellig S. Harris. 1957. Co-occurrence and transformation in linguistic structure. *Language*, 33(3):283–340.

Richard P Honeck. 1971. A study of paraphrases. *Journal of Verbal Learning and Verbal Behavior*, 10(4):367–381.

Shudong Huang, David Graff, and George Doddington. 2002. Multiple-translation chinese corpus. Linguistic Data Consortium.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Jaap Jumelet and Dieuwke Hupkes. 2018. Do language models understand anything? on the ability of LSTMs to understand negative polarity items. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 222–231, Brussels, Belgium. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Jenna Saarni, Maija Sevón, and Otto Tarkka. 2021. Finnish paraphrase corpus. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 288–298, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018. ETPC - a paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Stephen Mayhew, Klinton Bicknell, Chris Brust, Bill McDowell, Will Monroe, and Burr Settles. 2020. Simultaneous translation and paraphrase for language education. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 232–243, Online. Association for Computational Linguistics.

Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. 2015. Improving statistical machine translation with a multilingual paraphrase database. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1379–1390, Lisbon, Portugal. Association for Computational Linguistics.

Dominik Meier, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2024. Towards human understanding of paraphrase types in chatgpt.

Igor Mel'čuk. 2012. *Semantics: From meaning to text. Volume 1*. John Benjamins.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Sneha Rajana, Chris Callison-Burch, Marianna Apidianaki, and Vered Shwartz. 2017. Learning antonyms with paraphrases and a morphology-aware neural network. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 12–21, Vancouver, Canada. Association for Computational Linguistics.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. 2023. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8776–8788, Singapore. Association for Computational Linguistics.

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. AESOP: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Teemu Vahtola, Mathias Creutz, and Jörg Tiedemann. 2022. It is not easy to detect paraphrases: Analysing semantic similarity with antonyms and negation using the new SemAntoNeg benchmark. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 249–262, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Teemu Vahtola, Mathias Creutz, and Jrg Tiedemann. 2023. Guiding zero-shot paraphrase generation with fine-grained control tokens. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 323–337, Toronto, Canada. Association for Computational Linguistics.

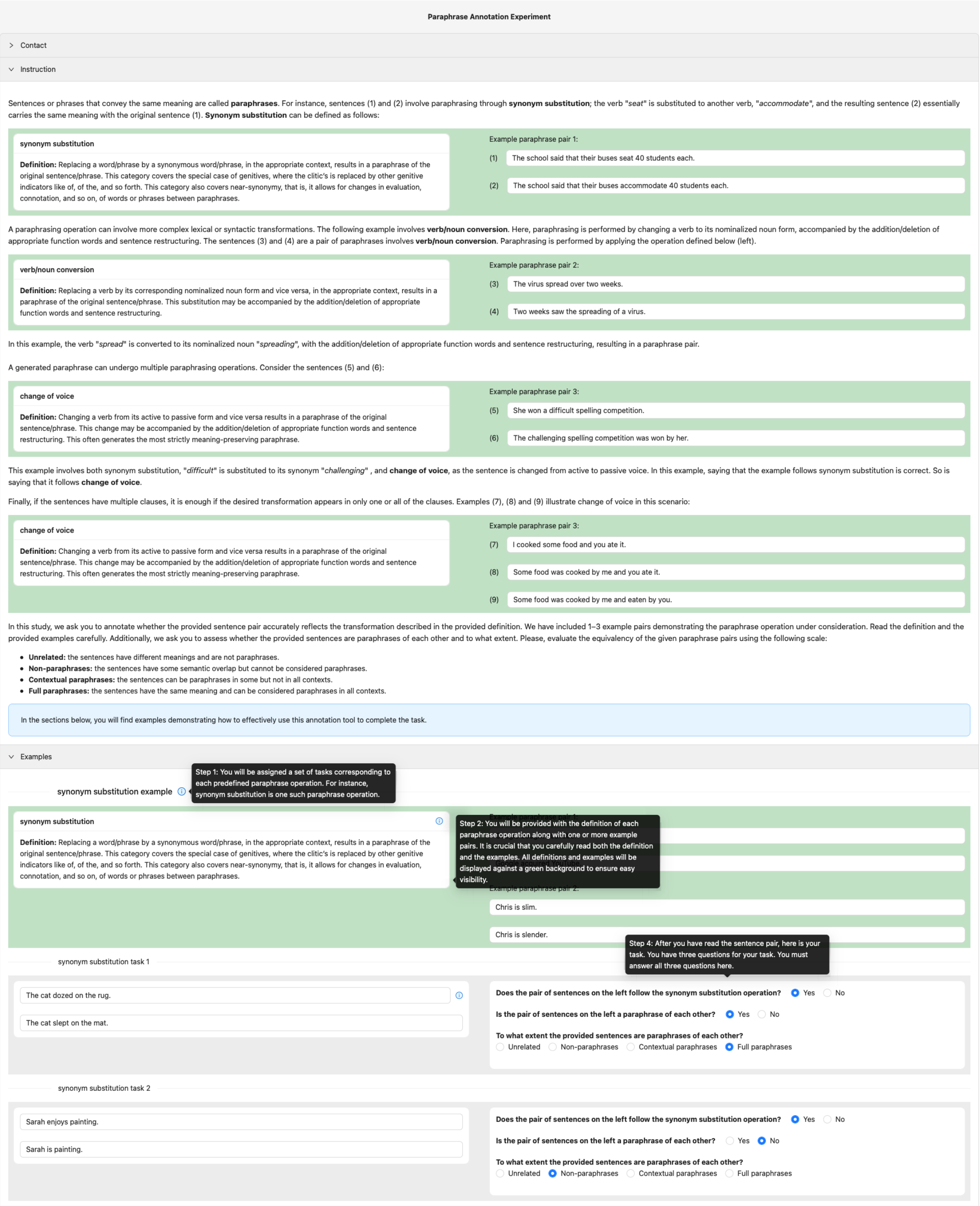
Marta Vila, M Antònia Martí, Horacio Rodríguez, et al. 2014. Is this a paraphrase? what kind? paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(01):205.

Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2023. Paraphrase types for generation and detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12148–12164, Singapore. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

A Annotation Setup

Figure 5 presents an example of the web-based annotation tool we used for collecting the manual annotations.



Paraphrase Annotation Experiment

Sentences or phrases that convey the same meaning are called **paraphrases**. For instance, sentences (1) and (2) involve paraphrasing through **synonym substitution**; the verb "*seat*" is substituted to another verb, "*accommodate*", and the resulting sentence (2) essentially carries the same meaning with the original sentence (1). **Synonym substitution** can be defined as follows:

synonym substitution

Definition: Replacing a word/phrase by a synonymous word/phrase, in the appropriate context, results in a paraphrase of the original sentence/phrase. This category covers the special case of genitives, where the clitic's is replaced by other genitive indicators like of, of the, and so forth. This category also covers near-synonymy, that is, it allows for changes in evaluation, connotation, and so on, of words or phrases between paraphrases.

Example paraphrase pair 1:

(1) The school said that their buses seat 40 students each.

(2) The school said that their buses accommodate 40 students each.

A paraphrasing operation can involve more complex lexical or syntactic transformations. The following example involves **verb/noun conversion**. Here, paraphrasing is performed by changing a verb to its nominalized noun form, accompanied by the addition/deletion of appropriate function words and sentence restructuring. The sentences (3) and (4) are a pair of paraphrases involves **verb/noun conversion**. Paraphrasing is performed by applying the operation defined below (left).

verb/noun conversion

Definition: Replacing a verb by its corresponding nominalized noun form and vice versa, in the appropriate context, results in a paraphrase of the original sentence/phrase. This substitution may be accompanied by the addition/deletion of appropriate function words and sentence restructuring.

Example paraphrase pair 2:

(3) The virus spread over two weeks.

(4) Two weeks saw the spreading of a virus.

In this example, the verb "*spread*" is converted to its nominalized noun "*spreading*", with the addition/deletion of appropriate function words and sentence restructuring, resulting in a paraphrase pair.

A generated paraphrase can undergo multiple paraphrasing operations. Consider the sentences (5) and (6):

change of voice

Definition: Changing a verb from its active to passive form and vice versa results in a paraphrase of the original sentence/phrase. This change may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. This often generates the most strictly meaning-preserving paraphrase.

Example paraphrase pair 3:

(5) She won a difficult spelling competition.

(6) The challenging spelling competition was won by her.

This example involves both synonym substitution, "*difficult*" is substituted to its synonym "*challenging*" , and **change of voice**, as the sentence is changed from active to passive voice. In this example, saying that the example follows synonym substitution is correct. So is saying that it follows **change of voice**.

Finally, if the sentences have multiple clauses, it is enough if the desired transformation appears in only one or all of the clauses. Examples (7), (8) and (9) illustrate change of voice in this scenario:

change of voice

Definition: Changing a verb from its active to passive form and vice versa results in a paraphrase of the original sentence/phrase. This change may be accompanied by the addition/deletion of appropriate function words and sentence restructuring. This often generates the most strictly meaning-preserving paraphrase.

Example paraphrase pair 3:

(7) I cooked some food and you ate it.

(8) Some food was cooked by me and you ate it.

(9) Some food was cooked by me and eaten by you.

In this study, we ask you to annotate whether the provided sentence pair accurately reflects the transformation described in the provided definition. We have included 1–3 example pairs demonstrating the paraphrase operation under consideration. Read the definition and the provided examples carefully. Additionally, we ask you to assess whether the provided sentences are paraphrases of each other and to what extent. Please, evaluate the equivalency of the given paraphrase pairs using the following scale:

- **Unrelated:** the sentences have different meanings and are not paraphrases.
- **Non-paraphrases:** the sentences have some semantic overlap but cannot be considered paraphrases.
- **Contextual paraphrases:** the sentences can be paraphrases in some but not in all contexts.
- **Full paraphrases:** the sentences have the same meaning and can be considered paraphrases in all contexts.

In the sections below, you will find examples demonstrating how to effectively use this annotation tool to complete the task.

synonym substitution example

Step 1: You will be assigned a set of tasks corresponding to each predefined paraphrase operation. For instance, synonym substitution is one such paraphrase operation.

synonym substitution

Definition: Replacing a word/phrase by a synonymous word/phrase, in the appropriate context, results in a paraphrase of the original sentence/phrase. This category covers the special case of genitives, where the clitic's is replaced by other genitive indicators like of, of the, and so forth. This category also covers near-synonymy, that is, it allows for changes in evaluation, connotation, and so on, of words or phrases between paraphrases.

Step 2: You will be provided with the definition of each paraphrase operation along with one or more example pairs. It is crucial that you carefully read both the definition and the examples. All definitions and examples will be displayed against a green background to ensure easy visibility.

Example paraphrase pair 2:

Chris is slim.

Chris is slender.

Step 4: After you have read the sentence pair, here is your task. You have three questions for your task. You must answer all three questions here.

synonym substitution task 1

The cat dozed on the rug.

The cat slept on the mat.

Does the pair of sentences on the left follow the synonym substitution operation? ● Yes ○ No

Is the pair of sentences on the left a paraphrase of each other? ● Yes ○ No

To what extent the provided sentences are paraphrases of each other?
○ Unrelated ○ Non-paraphrases ○ Contextual paraphrases ● Full paraphrases

synonym substitution task 2

Sarah enjoys painting.

Sarah is painting.

Does the pair of sentences on the left follow the synonym substitution operation? ● Yes ○ No

Is the pair of sentences on the left a paraphrase of each other? ○ Yes ● No

To what extent the provided sentences are paraphrases of each other?
○ Unrelated ● Non-paraphrases ○ Contextual paraphrases ○ Full paraphrases

Figure 5: A screenshot of our web-based annotation tool.