Human evaluation of automatically generated classifiers in Mandarin Chinese

by

Amber de Bruijn

First supervisor: Prof. dr. Kees van Deemter
Second supervisor: Dr. Guanyi Chen

A thesis submitted in fulfilment of the requirements for the degree
Master of Science
in
Artificial Intelligence

Department of Information and Computing Sciences
Faculty of Science
Utrecht University
31 March 2023

Universiteit
Utrecht

Abstract

One of the difficulties of artificially generating Mandarin Chinese text is the question of which classifier - a linguistic unit obligatory in numeral expressions - to choose in a given context. Several algorithms for classifier choice have recently been developed and assessed using a corpus-based evaluation. The best-scoring algorithm was a BERT classification model. However, evaluating classifiers based on a corpus provides a conservative score: it classifies each non-matching classifier as incorrect, while native speakers might acknowledge multiple different classifiers as a correct option. Since the ultimate goal of NLG should be the generation of texts that are useful to humans, we decided to perform a human evaluation in addition to the corpus-based one. We conducted two experiments; the first was a standard NLG evaluation, and the second was a more linguistically motivated experiment focusing on only true classifiers (a specific subset of Mandarin classifiers). We found that, according to human readers, BERT consistently performs better than the other models, agreeing with the corpus-based evaluation. However, we found no difference in the evaluation scores between BERT and the human-produced sentences in the corpus. This is remarkable, because the corpus-based evaluation suggests a large gap between BERT's score and the corpus' score. This result suggests human readers are more accepting of variations in classifier choice than previously thought.

Acknowledgements

First and foremost, I want to thank Kees van Deemter and Guanyi Chen. Their suggestions, guidance, and help have been invaluable. They have taught me that good research entails much more than gathering and analysing numbers; discussions and teamwork are indispensable.

I want to thank Professors Muyun Yang and Tiejun Zhao from Harbin Institute of Technology for their help in recruiting participants and for their feedback on the pilot.

For other contributions, I thank Yupei Du for his feedback on the pilot, and Qixiang Fang and Albert Gatt for their suggestions regarding statistical analyses.

On a personal note, I'd like to thank my partner, sisters, parents, and friends for their support, their belief in me and for being there for me when I needed them.

Lastly, I want to thank the coffee bar ladies at the public library in Amersfoort, who were always open to a conversation during the months I worked on my thesis.

Contents

1 Introduction

Classifier words, or simply *classifiers*, are linguistic units that are obligatory when a noun is modified by a numeral. They are a common linguistic phenomenon in many East-Asian, Australian, African and American languages [Allan, 1977]. In non-classifier languages such as English, uncountable (mass) nouns need some measure word - a kind of classifier - in numeral expressions (e.g. “three drops of oil”, in which the measure word is “drops”). Countable (count) nouns can be modified by a numeral without needing an additional element (e.g. “three pens”). In classifier languages, of which Mandarin Chinese[1] is a prime example, all nouns must be accompanied by a classifier in numeral expressions. Comparing the phenomenon directly with English mass nouns, the construction of the numeral expression is similar; in Mandarin *san di you* “three drops of oil”, the classifier *di* 滴 can be directly translated to English “drops”. On the other hand, a direct comparison like this is not possible with English count nouns; in Mandarin *san zhi bi* “three pens”, the classifier *zhi* 支 is obligatory, while such an element does not exist in English.

Mandarin classifiers can roughly be divided into measure words (modifying nouns corresponding to English mass nouns) and true classifiers (modifying nouns corresponding to English count nouns). While measure words create a measure to quantify the noun, true classifiers only classify the noun. Choosing the right classifier to use is not a straightforward task. For measure words, you need to know the quantity of the entity you are talking about. The problem is different for choosing the right true classifier. True classifiers often provide a semantic indication of the corresponding noun, where it is important to note that a single classifier can be associated with multiple semantically distinct groups of things. Simultaneously, multiple classifiers can often be paired with the same noun. This does not change the noun’s meaning, but instead influences various other semantic qualities (e.g. politeness, number, formality). To complicate true classifier choice even further, in some cases it is not clear why one classifier is chosen over the other, or the classifier may even affect the meaning of the noun.

These are difficulties we encounter when we want to automatically generate Mandarin classifiers. Multiple NLG models have been developed that generate Mandarin classifiers, the most recent ones described in [Järnfors, 2021]. Of these most recent models, one is an LSTM model, two models are based on BERT, and another one is a transparent rule-based model. The performance of these models has been assessed by a corpus-based evaluation in [Järnfors, 2021], and by comparing the models’ results with those of human participants who performed the same task in [Chen, 2022]. Based on these evaluations, the models seem to perform quite well. However, an evaluation of the models’ performance carried out by human judges is preferable, to account for possible difficulties with automatic evaluation (e.g. a corpus saying that a classifier choice is wrong because it did not get a match, while humans would think it is perfectly fine). Additionally, it is not certain how important the choice of true classifiers is to human readers.

Therefore, we will subject these algorithms to a new evaluation based on human judgement. We propose two experiments:

1. a standard human evaluation of the performance of the models by [Järnfors, 2021];
2. and a second one to explore the more linguistic question of the importance of true classifier choice.

The first experiment looks at the choices made by the algorithms `RULE`, `LSTM`, and `BERT` (the rule-based model, LSTM model, and BERT classification model from [Järnfors, 2021], respectively). We decided to use these three because `BERT` performed best in the corpus-based evaluation and `LSTM` performed second best. For `RULE`, we want to see whether native speakers agree with the corpus-based evaluation, or if they are more lenient with `RULE`’s choices.

The second experiment focuses on only true classifiers, with an additional distinction between a randomly chosen dataset and a dataset only containing infrequent classifiers. For this experiment, we look at the choices made by `RULE` and `BERT`, the same algorithms from the previous experiment,

[1] Henceforth referred to as “Mandarin”, unless specified otherwise.

and GE, a model that always chooses the general classifier *ge* 个[2]. We selected these models because RULE performed remarkably well with only true classifiers in the corpus-based evaluation and BERT performed best of all algorithms. For GE, we want to test how native speakers perceive the use of *ge* 个 regardless of context.

For both experiments, participants are presented with sentences containing a classifier, either the one present in the corpus or chosen by one of the algorithms. They judge the sentences based on Clarity ("this sentence is clear") and Fluency ("this sentence was written by a native speaker"), using a seven-point Likert scale.

For the first experiment, we answer two questions:

- How does the corpus-based evaluation of the models RULE, LSTM, and BERT compare to the human evaluation?
- Are BERT's classifier choices rated as more fluent than the choices made by RULE and LSTM?

For the second experiment, we want to know how the classifier choices by the models CORPUS, GE, RULE, and BERT compare to each other, both with frequent and infrequent head nouns. The two specific questions we answer are:

- Given that BERT performs better than the other models, are its chosen classifiers in the first group (random) rated as more fluent compared to the second group (infrequent)?
- How does the choice of the general classifier *ge* 个 compare to the choices by CORPUS, RULE, and BERT?

2 Classifiers in Mandarin Chinese

Classifiers are linguistic units that occur in numeral expressions (i.e. a larger unit that consists of a noun modified by a numeral). In typical classifier languages, such as Mandarin, the classifier is an obligatory word in a numeral expression. The word order is fixed; the classifier follows the numeral and precedes the noun [Zhang, 2013].

(1) a. Yaoyao kanjian-le san di you.
Yaoyao see-PRF three CL oil
'Yaoyao saw three drops of oil.'
b. Yaoyao kanjian-le san zhi bi.
Yaoyao see-PRF three CL pen
'Yaoyao saw three pens.'
Example from [Zhang, 2013, Chapter 1, p. 1][3]

Looking at example 1a, comparing the Mandarin numeral expression with the English one seems clear. Because "oil" is a mass noun in English, it needs a measure word to express how much oil is being talked about. The classifier *di* 滴 can be directly translated to English "drops". On the other hand, in example 1b, there does not seem to be a direct English translation in this context for *zhi* 支. "Pen" is a count noun in non-classifier languages such as English, which means it does not need a measure word in order to be grammatical. Here lies the distinction with classifier languages, where classifiers are always required in numeral expressions.

[2] The true classifier *ge* 个 can be seen as a general classifier, meaning it can be paired with almost any head noun. Section 2.2 provides more details on the distinctive use of *ge* 个.

[3] Abbreviations used in glosses:
PRF = perfect aspect
CL = classifier

2.1 Categorisation and semantic qualities of classifiers

There are roughly two distinct groups of classifiers in Mandarin: **measure words**, which modify nouns that roughly correspond to mass nouns in non-classifier languages, such as in example 1a; and **true classifiers**[4], such as in example 1b. Where the mass-count distinction happens at the noun level in English, it happens at the classifier level in Mandarin; measure words create a measure to quantify the noun, true classifiers only classify the noun [Cheng and Sybesma, 1998].

True classifiers are not assigned randomly to a noun. A classifier often provides a semantic indication for the upcoming head noun, as in example 1b, where *zhi* 支 can be translated as the indication for "stick-like object", or in example 2, where *tiao* 条 refers to rope-like things:

(2) a. yi tiao she
one CL snake
"a snake"

b. yi tiao he
one CL river
"a river"

c. yi tiao jie
one CL street
"a street"
Example from [Zhang, 2007, p. 46]

This semantic indication is not a given one-on-one relationship between a classifier and the nouns it pairs with. In the case of *tiao* 条, it can also be used as a classifier for clothing items that one's legs are put through [Zhang, 2007]. A classifier can be associated with multiple semantically distinct groups of things.

One notable aspect of true classifiers is that different classifiers can be used with a head noun, without changing the noun's meaning. Instead, the classifier choice influences other semantic qualities (e.g. distinctions between formal-informal, educated-uneducated, written-colloquial) [Zhang, 2007]. For example:

(3) a. yi ge yang
one CL sheep

b. yi zhi yang
one CL sheep

c. yi tou yang
one CL sheep
"a sheep"
Example from [Zhang, 2007, p. 53]

All sentences in example 3 have the same referential meaning, but there is a difference in formality. In 3a, the general classifier *ge* 个 indicates more colloquial language use, while in 3b and 3c the classifiers *zhi* 只 (indicating animals) and *tou* 头 (indicating domesticated animals) indicate a more formal use.

Both true classifiers and measure words can add more explicit information than the previously mentioned semantic qualities. Looking first at measure words, this is recognisable for English speakers:

[4] Within the linguistic literature, there are multiple labels for these groups of classifiers. Measure words can for example also be called mass classifiers, mass-noun classifiers, mensural classifiers or massifiers, while true classifiers can also be called count classifiers, count-noun classifiers or sortal classifiers. To keep consistency with [Järnfors, 2021], I will use the terms *measure words* and *true classifiers* to refer to these two groups.

(4) yi bei kafei / yi ting kafei
one cup coffee / one can coffee
“a cup of coffee” / “a can of coffee”
Example from [Järnfors, 2021, p. 5]

The measure words *bei* 杯 “cup” and *ting* 听 “can” in example 4 are two distinct containers that contain different quantities of coffee. When looking at true classifiers, these can also add more explicit information. Take, for example:

(5) a. yi ge laoshi / yi wei laoshi
one CL teacher / one CL.POL teacher
“a teacher”

b. yi ge ren / yi qun ren
one CL person / one CL.PL person
“a person” / “a group of people”
Example from [Järnfors, 2021, p. 5][5]

In example 5a, the difference between *ge* 个 and *wei* 位 is politeness; while *ge* 个 is a neutral classifier, the use of *wei* 位 indicates a polite register. In example 5b, the difference between *ge* 个 and *qun* 群 is number; *ge* 个 indicates an individual, while *qun* 群 is an indicator of plurality, denoting a group [Järnfors, 2021].

2.2 True classifiers: the general classifier *ge* 个

Within the true classifiers, the classifier *ge* 个 is a special case. Unlike other true classifiers, *ge* 个 does not provide a semantic indication of the head noun [Zhang, 2013]. Besides nouns for which *ge* 个 is the default classifier (e.g. *ren* 人 “person”), it can be used in the place of other true classifiers as a more general classifier. This use of *ge* 个 as a general classifier has been observed both in child language acquisition and language loss; it is the first classifier children learn and use in all contexts before using other classifiers, and it is the classifier aphasic patients often fall back to when they cannot access the right true classifier [Cheng and Sybesma, 2015]. Besides situations concerning language loss and acquisition, the use of *ge* 个 as a general classifier has been observed in informally spoken language; even though native speakers are taught to use a specific classifier in a large variety of cases, when observing everyday language use, both children and adults tend to use *ge* 个 instead of the “right” specific classifier [Erbaugh, 1986].

2.3 Difficulties when choosing classifiers

There are many instances in which choosing the right classifier can be difficult for non-native speakers. Besides more obvious differences between classifiers, like quantity, differences can get as subtle as indicating formality. To complicate this further, classifier choice may sometimes be completely arbitrary, where it is difficult to determine what exactly the differences between classifiers are for a given noun are [Zhang, 2013]. The relationship between classifiers and nouns can sometimes even be reversed when a particular isolated noun has multiple meanings; instead of the classifier only conveying information about semantic qualities as discussed in section 2.1, the classifier can play a deciding role in determining the meaning of the noun [Zhang, 2007].

[5] Abbreviations used in glosses:
POL = polite register
PL = plural

3 Models for classifier choice

3.1 What is NLG?

Natural Language Generation (*NLG*) is a subfield of Natural Language Processing (NLP), which is in turn a subfield of Artificial Intelligence (AI). NLG is concerned with the automatic production of understandable texts in natural (i.e. human) languages [Reiter and Dale, 1997]. There are many applications of NLG, using methods like text-to-text generation, in which already existing texts are taken as input to produce a new text (e.g. machine translation, summarisation to make texts more concise, simplification to make complex texts more accessible, and automatic text correction), or data-to-text generation, in which algorithms use non-linguistic data to produce readable texts (e.g. producing news reports, sports reports, weather and financial reports, and summaries of patient information from medical data) [Gatt and Krahmer, 2018]. Perhaps the best known NLG application at the moment is ChatGPT[6].

Many NLG applications can be classified as *practical NLG*, invested in creating practical applications, tools to aid writers and to produce texts that otherwise would not have been written. The main goal of these programs is to produce useful texts. On the other hand is *theoretical NLG*, in which NLG is used to gain understanding and produce theories of linguistic phenomena. Here, the main goal is to mimic and understand human language use as closely as possible [Van Deemter, 2016].

In any case, evaluating how well a model performs for its intended use is important. There are several ways in which NLG models can be evaluated: by automatic metrics requiring no training, machine learning metrics, and human evaluation metrics. Every method has its own advantages and disadvantages, for example, in terms of speed, cost, suitability for the task, and repeatability [Celikyilmaz et al., 2021].

In this thesis, we focus on two different evaluation methods; a non-trained automatic method[7] and a human-centric method[8]. We will compare these methods using models that generate Mandarin classifiers. Additionally, the experiments in this thesis approach the more theoretical side of NLG; instead of only focusing on generating classifiers (e.g. for applications like machine translation), we will look at how computational models can also be used to attempt learning about native speakers' perception of classifier choice in written text.

3.2 Classifier generation

We believe classifier selection to be a non-trivial task: choosing the appropriate classifier depends highly on context (as discussed in section 2). Approaches based on looking up a noun in e.g. dictionaries to find the appropriate classifier are questionable, as multiple classifiers can be associated with a single noun; which classifier fits best is based on context. This is why the use of algorithms, preferably a model that takes context into account, would be more suitable than a look-up-based approach. Additionally, knowledge of native speakers' perception of chosen classifiers can help determine which model would be best to use if we want to achieve generated text that reads as if written by a native speaker.

3.2.1 Earlier generation of classifiers

Over the years, there have been multiple approaches to the automatic generation of classifiers in Mandarin. Following the overview of [Järnfors, 2021], the earliest approach mentioned is the use of Support Vector Machines (SVM) trained to assign classifiers to nouns [Guo and Zhong, 2005]. In 2008, there has been a study exploring a statistical model that generates classifiers for English-to-Chinese statistical machine translation [Zhang et al., 2008]. Another approach has been the implementation of a database of semantic features of classifiers and their associated nouns [Gao, 2011]. Two other

[6] ChatGPT is an NLG model developed by OpenAI. More information: https://openai.com/blog/chatgpt#OpenAI.

[7] Accuracy; further discussed in section 3.3.

[8] Our experiments; further discussed in section 4.

studies made use of an already existing database: a study from 2012 using a bilingual Chinese-English Wordnet[9], to generate classifiers based on semantic hierarchies [Wen et al., 2012]; and a 2016 study using the Chinese Open Wordnet [Da Costa et al., 2016].

In 2017, [Peinelt et al., 2017] proposed a different approach. Arguing that classifier choice is highly contextual, they constructed a large-scale corpus, the ChineseClassifierDataset (*CCD*)[10], to train context-aware machine learning models. They implemented three baseline models, based on some of the previously mentioned studies: an algorithm that always chooses the general classifier *ge* 个; an algorithm that assigns the classifier that most frequently occurred together with a head noun during training (unseen head nouns get assigned *ge* 个); and an algorithm that assigns classifiers using the Chinese Open Wordnet (again, unseen head nouns get assigned *ge* 个). After these baselines, they implemented three machine learning models (SVM, Logistic Regression, and bidirectional LSTM[11]). Their best-scoring model, the LSTM, seems to be the most sophisticated in the sense that it outperforms all baselines and it requires neither manual feature engineering nor extensive pre-processing.

3.2.2 Models in [Järnfors, 2021]

The 2021 Master's thesis by Jani Järnfors, [Järnfors, 2021], builds further upon the approach by [Peinelt et al., 2017]. With Peinelt et al. providing the most promising results thus far, this allowed for making comparisons and gaining more insight into their results.

As in Peinelt et al., the CCD was used as the dataset for generating classifiers. Four models were used to produce classifiers:

- **Rule-based model**, `RULE`, similar to the one in [Peinelt et al., 2017]. This relatively simple algorithm looks only at the head noun in a given sentence. The algorithms works as follows:

  1. Given a head noun, assign the most frequent classifier associated with it in the training data.
  2. If two or more classifiers are equally frequent, one of the classifiers is randomly assigned.
  3. If the head noun does not appear in the training data, then the general classifier *ge* 个 is assigned.

- **LSTM model**, `LSTM`, similar to the one in [Peinelt et al., 2017]. This is a bidirectional LSTM model, framed as a multi-class classification task with 172 classes[12]. The hidden representation of the last time step is used to make predictions.

- Two BERT[13] models. Since BERT's release in 2018, it has produced state-of-the-art results in results on multiple NLP tasks [Devlin et al., 2018]. Various implementations are available, and BERT's performance depends on how the model has been trained. For both BERT implementations, the `bert-base-chinese`[14] version has been used.

  - **BERT masked-language model**, `MLM`. As the original BERT is based on a masked-language model, `MLM` predicts classifiers without fine-tuning on the task of classifier selection.
  - **BERT classification model**, `BERT`. This model demonstrates the classic use of BERT. Here, BERT is fine-tuned on the CCD as a multi-class classification task with 172 classes.

[9] Based on WordNet, a large lexical English database often used in natural language processing [Fellbaum, 2005].

[10] The ChineseClassifierDataset (CCD) is based on three publicly available POS (parts of speech) tagged Chinese corpora: the Lancaster Corpus of Mandarin Chinese, the UCLA Corpus of Written Chinese and the Leiden Weibo Corpus. The CCD is publicly available: https://github.com/wuningxi/ChineseClassifierDataset.

[11] Long Short-Term Memory [Hochreiter and Schmidhuber, 1997].

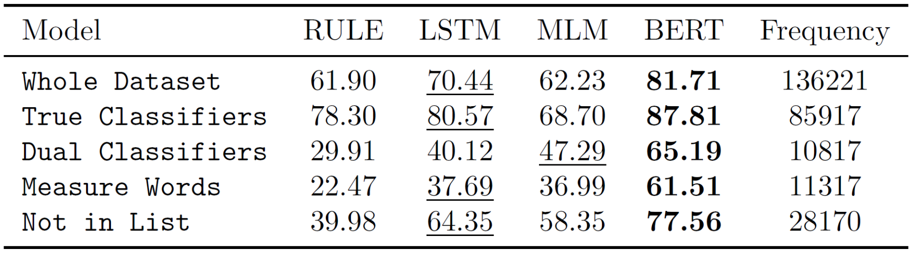
[12] A total of 172 classes, as the final dataset contained 172 distinct classifiers.

[13] Bidirectional Encoder Representations from Transformers [Devlin et al., 2018]

[14] https://huggingface.co/bert-base-chinese

3.3 Evaluation of existing models

After implementing the four models in [Järnfors, 2021], an automatic corpus-based evaluation was carried out. Additional to accuracy, the metrics used to evaluate the models were precision, recall, and F1, both macro-averaged and weighted-averaged. However, since the focus in [Järnfors, 2021] lies on the accuracy scores, we will focus on those results in this section.



Model	RULE	LSTM	MLM	BERT	Frequency
Whole Dataset	61.90	70.44	62.23	**81.71**	136221
True Classifiers	78.30	80.57	68.70	**87.81**	85917
Dual Classifiers	29.91	40.12	47.29	**65.19**	10817
Measure Words	22.47	37.69	36.99	**61.51**	11317
Not in List	39.98	64.35	58.35	**77.56**	28170

Figure 1: *Table 2* from [Järnfors, 2021, p. 28]. Accuracy results for all four models (RULE, LSTM, MLM, BERT). Results are on the whole dataset (top row) and several categories of classifiers. The highest accuracy is boldfaced, while the second highest is underlined. Important: we found different accuracy scores for the whole dataset: 61.73% for RULE and 73.86% for LSTM. We use these updated values in this thesis.

Accuracy results for all models are displayed in Figure 1. After re-implementing RULE and LSTM for our experiments, we found differences in accuracy scores for the whole dataset: 61.73% for RULE and 73.86% for LSTM. In the remainder of this thesis, we use these updated values[15].

Overall, the BERT classification model outperforms all other models, with an accuracy score of 81.71%. The second best model is LSTM, with a score of 73.86%. RULE scores lowest with 61.73%, a result comparable with MLM. All models score considerably better with only true classifiers than the other classifier categories. When looking at only true classifiers, it is noteworthy that the accuracy RULE (78.30%) seems comparable to LSTM (80.57%) and comes closer to BERT's result (87.81%) than in the other classifier categories.

3.3.1 Speaker experiments in [Chen, 2022]

To examine the difficulty of the task of choosing classifiers, [Chen, 2022] has conducted two experiments with human participants who had to perform the same task as the algorithms. The first experiment used randomly sampled sentences from the CCD and had participants choose a classifier for each sampled sentence. Surprisingly, the accuracy of the human participants (70.97%) seemed to be lower than that of BERT, closer to the LSTM's performance. For the second experiment, participants performed the same task. The difference was that in this experiment, instead of randomly sampled CCD sentences, only sentences containing infrequent classifiers were used. In this case, accuracy dropped to 41.82%. These results come with the caveat that the human speakers, as opposed to the algorithms, are possibly not familiar with the kind of language used in the CCD and have not trained for the task with similar sentences. However, from these results, it can still be concluded that the performance of the models seems promising. In terms of accuracy score, human speakers are not coming close to full accuracy when compared to the corpus. They do not seem to perform noticeably better than BERT when performing the same task. Additionally, besides the low accuracy score in the second experiment, agreement between participants was notably lower than in the first experiment. Since BERT had a comparably low accuracy score, it can be concluded that classifier selection is a non-trivial task, both for algorithms and native speakers.

[15] Due to time constraints, we did not compute accuracy scores for RULE and LSTM's individual classifier categories (i.e. true classifiers, dual classifiers, measure words, and not in list). We will not perform inferential statistics on or draw important conclusions from the non-updated accuracy scores.

4 Human evaluation of classifier choice

4.1 Evaluation in NLG

It is important to keep in mind that the evaluation of the algorithms' results in [Järnfors, 2021] and the human participants' results in [Chen, 2022] has only been done with a corpus. In an automatic corpus-based evaluation, the generated text and the target text in a corpus are compared. The resulting score indicates the similarity between these texts. There are benefits to using corpora for evaluating generated text; they are fast, cheap, and widely used [Celikyilmaz et al., 2021], providing a straightforward way to compare models. However, there are drawbacks too: there is much variability to how much human judgements and automatic evaluation metrics correlate [Gatt and Krahmer, 2018]. Ultimately, the goal of NLG should be to generate texts that are useful to humans. If people perceive a generated text differently than the automatic evaluation score reflects, it can be argued that the automatic metric is less suitable for assessing the performance of an algorithm. Additionally, comparing generated text directly to a target text provides a conservative evaluation approach; there is only one target in the corpus, and if the generated text does not exactly match, the metric classifies it as "wrong". However, in many cases it could be possible to use multiple different words to express the same message while still being an acceptable text to native speakers. Returning to the context of Mandarin classifiers, it could be possible to use multiple different classifiers within the same context. The corpus-based evaluation would classify every non-matching classifier as incorrect, while native speakers might disagree with this evaluation and say it is correct.

To get a better understanding of the performance of the models, it would be preferable to conduct a human evaluation to eliminate possible limitations of a corpus-based evaluation (e.g. the corpus suggesting a classifier choice is not correct because it does not occur in the corpus, while native speakers would say it is correct).

4.2 Research questions and hypotheses

Before stating the research questions and hypotheses, it would be helpful to briefly introduce the two experiments. This provides context to better understand the questions we ask.

We will conduct two experiments:

- Experiment 1: We conduct a "traditional" or standard human NLG evaluation of the models in [Järnfors, 2021], with the goal of comparing the automatic and human evaluations to each other. We take a random selection of sentences from the corpus, containing all classifiers (i.e. true classifiers, dual classifiers, and measure words). We compare the choices made by three models in [Järnfors, 2021]:

    1. `RULE`: the classifier chosen by the rule-based model.
    2. `LSTM`: the classifier chosen by the bi-directional LSTM.
    3. `BERT`: the classifier chosen by the BERT classification model.

- Experiment 2: Additionally to Experiment 1, we want to examine classifier choice for true classifiers (so when there is a choice between *ge* 个 and more specific classifiers). We know that sometimes various algorithms make different classifiers choices compared to the corpus and each other. However, we do not know to what extent these differing choices matter to human readers. Classifier choice is a difficult task, especially in "worst-case scenarios", where the models encounter infrequent head nouns that only rarely occur in training data if they occur at all. That is why we additionally want to compare classifier choice with random classifiers in the target text[16] to classifier choice with infrequent classifiers in the target text. Therefore, we create two groups of sentences: one group with random classifiers, in which the sentences are randomly

[16] We decided on using randomly selected sentences to simulate "natural" language use, i.e. mostly frequent classifiers while sometimes encountering more infrequent classifiers.

selected, and a second group with infrequent classifiers, in which the sentences are selected based on the frequency of specific classifiers in the corpus. The classifier choices are made by three models:

1. `GE`: always choose *ge* 个.
2. `RULE`: the classifier chosen by the rule-based model in [Järnfors, 2021].
3. `BERT`: the classifier chosen by the BERT classification model in [Järnfors, 2021].

As a baseline, our participants in both experiments also judge `CORPUS`, the classifier as it appears in the CCD. For both experiments, we will primarily be looking at Fluency, as we expect this to show more variation in answers than Clarity (from the intuition that a given utterance can be clear while it simultaneously is obvious that a native speaker did not write it).

4.2.1 Experiment 1: research questions and hypotheses

The two research questions concerning the first experiment are:

- How does the corpus-based evaluation of the models `RULE`, `LSTM`, and `BERT` compare to the human evaluation?
- Are `BERT`'s classifier choices rated as more fluent than the choices made by `RULE` and `LSTM`?

In [Järnfors, 2021], the metric used to score the algorithms was accuracy. This means that on the sentence level, the generated classifier either matched with the one in the corpus, or it did not match. We take this as a starting point to make expectations about our data: we will divide our data into one group containing all Likert scores for generated classifiers that match with the one in the corpus, and the other group containing all Likert scores for generated classifiers that did not match with the one in the corpus.

This leads to our hypothesis for the first question:

1. The values in the first group (the chosen classifier matches with the corpus) are higher than those in the second group (the chosen classifier does not match with the corpus).

   Additionally, we have two hypotheses concerning `BERT`:

2. `BERT` is more fluent than `RULE`.
3. `BERT` is more fluent than `LSTM`.

4.2.2 Experiment 2: research questions and hypotheses

The broad research question for the second experiment is:

- How do the classifier choices by the models `CORPUS`, `GE`, `RULE`, and `BERT` compare to each other, both with frequent and infrequent head nouns?

More specifically, we ask:

- Given that `BERT` performs better than the other models, are its chosen classifiers in the first group (random) rated as more fluent compared to the second group (infrequent)?
- How does the choice of the general classifier *ge* 个 compare to the choices by `CORPUS`, `RULE`, and `BERT`?

We have five hypotheses (where hypotheses 2 to 5 are concerned only with the random group):

1. BERT random is more fluent than BERT infrequent.
2. GE is less fluent than CORPUS.
3. GE is less fluent than RULE.
4. BERT is more fluent than GE.
5. BERT is more fluent than RULE.

5 Experiment 1: Standard NLG evaluation

5.1 Constructing the dataset

We used the CCD as our starting point for collecting the sentences. Both [Järnfors, 2021] and [Chen, 2022], which this research builds upon, used the same corpus. Because this experiment aims to be a traditional NLG evaluation, we randomly selected 200 sentences. We did not use filters, and all categories of classifiers were included.

For each of these 200 sentences, we list four classifiers; one from the corpus, and three generated by algorithms:

- CORPUS: The classifier as it appears in that specific sentence in the CCD.
- RULE: The classifier generated by the rule-based model in [Järnfors, 2021]: given a head noun, assign the classifier most frequently associated with it in the training data; if at least two classifiers are equally frequent, one of them is randomly assigned; if the head noun is not present in the training data, assign the general classifier *ge* 个.
- LSTM: The classifier generated by the LSTM model in [Järnfors, 2021].
- BERT: The classifier generated by the BERT classification model in [Järnfors, 2021].

After generating four classifiers for every sentence, we have a dataset containing 800 items.

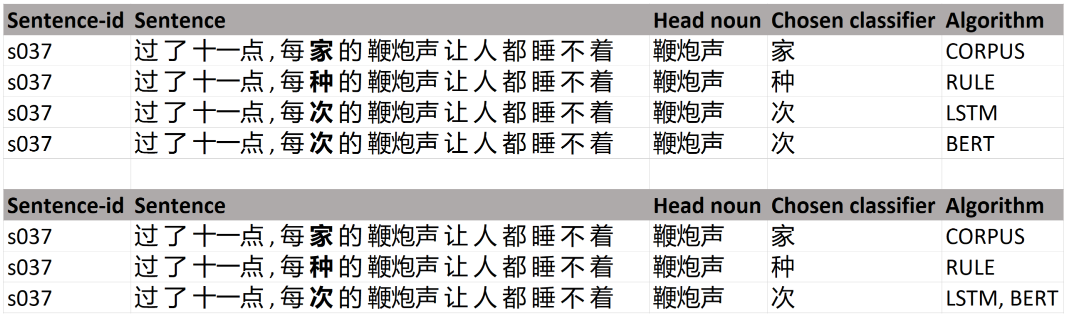
It would not be useful or efficient for the participants to rate the exact same sentence twice. Therefore, we removed duplicates of those items from the dataset for which two or more algorithms provide identical classifier choices (i.e. when given a sentence, if two out of four choices are identical, we remove one of the identical sentences; similarly, if three out of four choices are identical, we remove two of those sentences; and if all four choices are identical, we remove three of those sentences). An example is given in Figure 2.

After removing identical items, we have a final dataset containing 321 items: 112 items for which there was only one choice (112 sentences); 120 items for which there were two choices (60 sentences); 69 items for which there were three choices (23 sentences); and 20 items for which there were four choices (5 sentences). The full dataset can be found in Appendix A.

5.2 Experiment design

The experiment is conducted online, using Qualtrics Survey Software - UU[17]. Participants could partake at a time and place of their choosing. We have set the default language to Chinese (simplified), to ensure participants who do not understand English can read the messages provided by Qualtrics itself (e.g. when the program warns the participant that a question is not answered yet). All other text (i.e. everything we have written) is provided in both Chinese and English. The questionnaire

[17] Qualtrics is an online survey tool for setting up and distributing questionnaires, and collecting and analysing data. It is available in general via https://www.qualtrics.com, and for UU staff and students via survey.uu.nl.



Sentence-id	Sentence	Head noun	Chosen classifier	Algorithm
s037	过了十一点,每**家**的鞭炮声让人都睡不着	鞭炮声	家	CORPUS
s037	过了十一点,每**种**的鞭炮声让人都睡不着	鞭炮声	种	RULE
s037	过了十一点,每**次**的鞭炮声让人都睡不着	鞭炮声	次	LSTM
s037	过了十一点,每**次**的鞭炮声让人都睡不着	鞭炮声	次	BERT

Sentence-id	Sentence	Head noun	Chosen classifier	Algorithm
s037	过了十一点,每**家**的鞭炮声让人都睡不着	鞭炮声	家	CORPUS
s037	过了十一点,每**种**的鞭炮声让人都睡不着	鞭炮声	种	RULE
s037	过了十一点,每**次**的鞭炮声让人都睡不着	鞭炮声	次	LSTM, BERT

Figure 2: An example of how we removed identical items. The upper table shows the four items generated for sentence *s037*. LSTM and BERT chose the same classifier; instead of showing this exact item twice, we only have to show it once to the participants to get their judgement for both models. CORPUS and RULE have a different classifier, so we keep those items as is. After removing the duplicate item, the upper table can be compacted into the lower table.

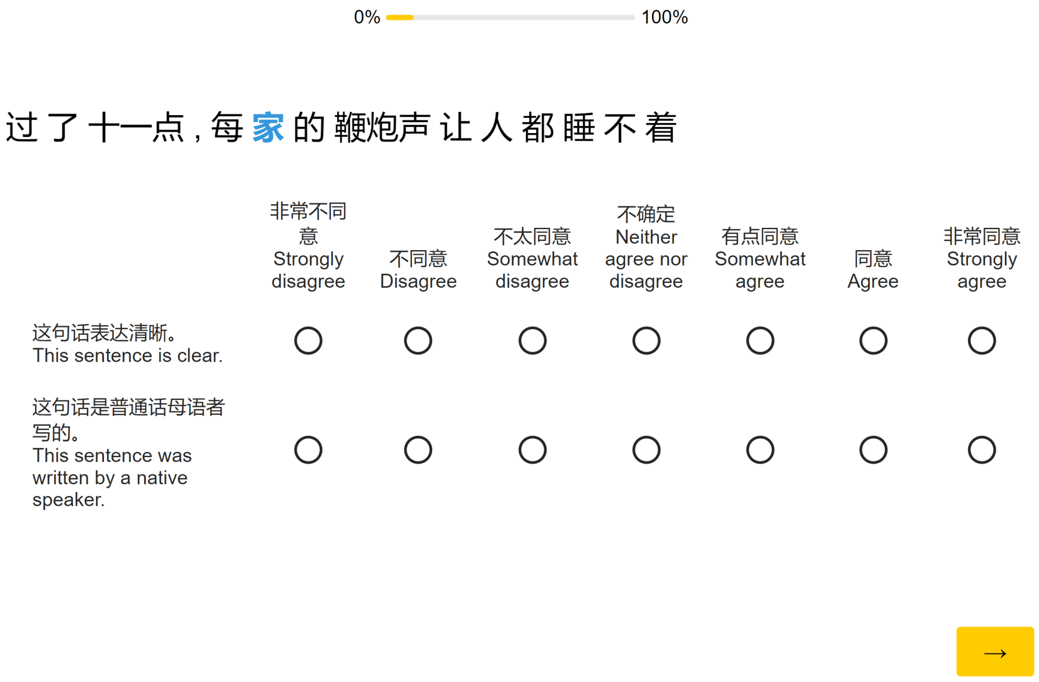
starts with a page containing general information about the experiment, ensuring participants can give their informed consent for participating. The second page contains a series of personal questions for statistics research. Then follows the experiment section.

In the experiment section, participants see one sentence item at a time. Because this concerns intuitive language use, it was important to get answers as intuitively as possible within the constraints of this questionnaire design. Therefore, we did not provide the possibility to return to previous sentences, to ensure the participants could not retroactively change their answers. To make the generated classifier easily identifiable, it has been highlighted (i.e. boldfaced and coloured blue). For every classifier choice, the participants answered two statements:

- CLARITY: This sentence is clear. 这句话表达清晰。
- FLUENCY: This sentence was written by a native speaker. 这句话是普通话母语者写的。

The participants answered the questions on a 7-point Likert scale, with the following options: strongly disagree 非常不同意; disagree 不同意; somewhat disagree 不太同意; neither agree nor disagree 不确定; somewhat agree 有点同意; agree 同意; strongly agree 非常同意.

Our dataset consists of 321 sentence items, which we deemed too much to ask from each of our individual participants. To present each participant with a more reasonable number of items, we divided the dataset into four groups, consisting of 82-83 items each. This was done manually, to ensure every group contained roughly the same amount of variation in classifier choices (as determined by the algorithms' classifier choices) and minimise differences. For each group, the order of the items was randomised. After randomisation, we made sure no sentence item directly followed a variation of the same sentence, to decrease the possibility that participants base their answers on the previous one. Each group of items was incorporated into a separate version of the experiment, producing four versions of the experiment.



过 了 十一点 , 每 **家** 的 鞭炮声 让 人 都 睡 不 着

	非常不同意 Strongly disagree	不同意 Disagree	不太同意 Somewhat disagree	不确定 Neither agree nor disagree	有点同意 Somewhat agree	同意 Agree	非常同意 Strongly agree
这句话表达清晰。 This sentence is clear.	○	○	○	○	○	○	○
这句话是普通话母语者写的。 This sentence was written by a native speaker.	○	○	○	○	○	○	○

Figure 3: Display of how the sentences were presented to the participants. The model-chosen classifier is boldfaced and highlighted in blue.

5.2.1 Pilot

Before finalising the experiments, we conducted a pilot. This pilot was performed with the dataset from Experiment 2[18]. Sentence items from Experiment 2's group 1 were used, minus the classifiers generated by BERT (we first wanted feedback on the survey length, to ensure the experiment with only three models would not already be too long before adding the fourth one). We invited staff from Harbin Institute of Technology and Utrecht University (all native speakers of Mandarin) to participate in the pilot. We have mainly received feedback on being unable to return to the previous question, having no way to know how many questions are left, and some sentences looking strange. We addressed the first two pieces of feedback by adding an explanation in the introduction that it is not possible to return to the previous questions, and adding a progress bar. As for the feedback on some strange-looking sentences; the CCD is partly based on data collected from social media, in which language can be used differently (more informally) from how it is written in e.g. books or news articles. Since these sentences are understandable, represent colloquial language use, and form a relatively large part of the corpus, we decided to keep all sentences.

5.3 Participants

A total of 8 participants completed the experiment, 3 female and 5 male, with ages ranging from 25 to 62 years (M = 36.0, SD = 12.0). All participants had university backgrounds: three participants had backgrounds in computer science, two in management, one in psychology, one in media, and one in statistics. The native language of all participants is Mandarin Chinese. Regarding languages learnt later in life, 7 participants report they know English, and one of them knows some Japanese. All

[18] We used the dataset from Experiment 2 because, due to some limitations, this experiment was finalised before Experiment 1. Because the design for both experiments is almost identical, excluding the datasets, performing a pilot only once would provide us with enough feedback for both.

participants had given their informed consent before participating in the experiment.

The participants were divided over the four versions of the experiment; each version was completed by two participants.

5.4 Overview gathered results

For analysing the data, we use assigned numbers to represent the participants' answers:

1. strongly disagree 非常不同意
2. disagree 不同意
3. somewhat disagree 不太同意
4. neither agree nor disagree 不确定
5. somewhat agree 有点同意
6. agree 同意
7. strongly agree 非常同意

The frequency of answers is given in Tables 1 and 2 and visually represented in Figures 4 and 5.

Experiment 1 - Clarity								
model	1	2	3	4	5	6	7	total
CORPUS	36	34	23	18	60	110	119	400
RULE	58	46	30	20	61	95	90	400
LSTM	48	39	29	19	55	112	98	400
BERT	40	33	26	19	59	110	113	400

Table 1: The frequency of answers concerning Clarity.

Experiment 1 - Fluency								
model	1	2	3	4	5	6	7	total
CORPUS	39	33	28	29	58	102	111	400
RULE	63	49	28	38	51	87	84	400
LSTM	49	44	26	33	56	99	93	400
BERT	41	36	21	34	59	101	108	400

Table 2: The frequency of answers concerning Fluency.

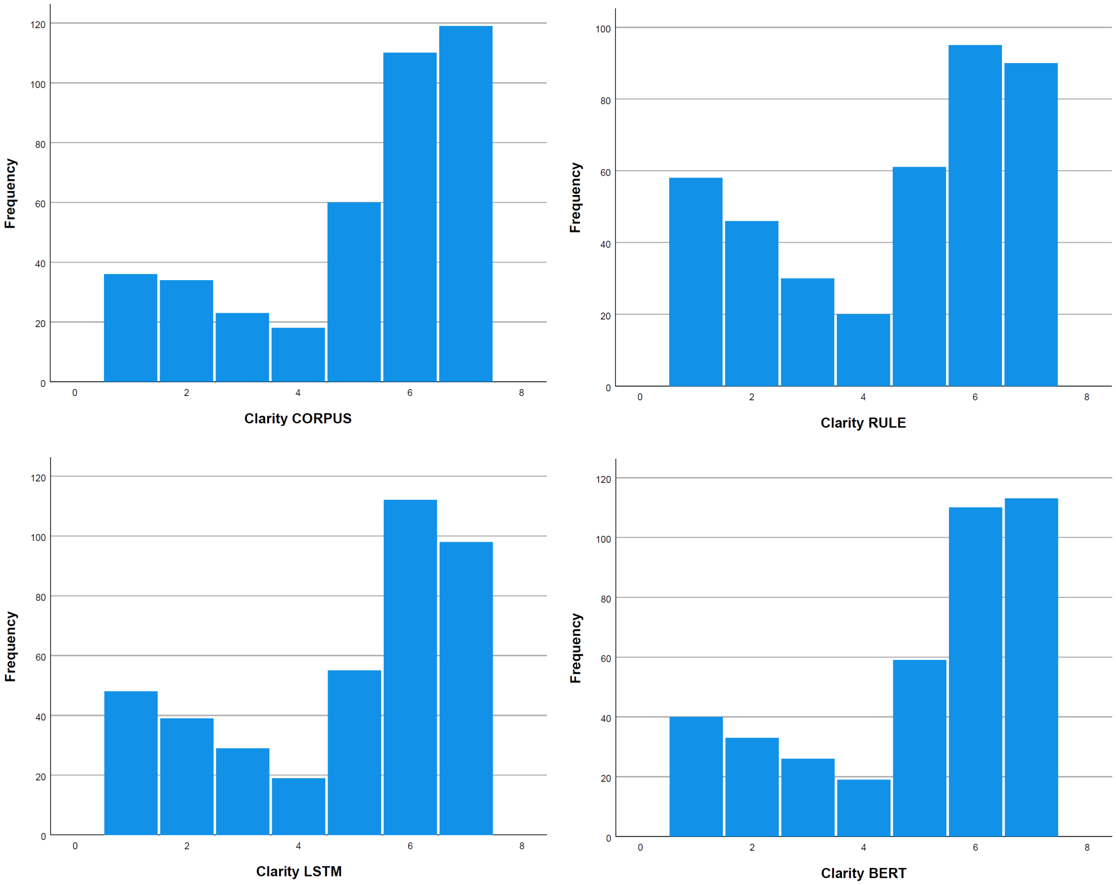


Figure 4: Graphs showing the frequency of Likert scores concerning Clarity for `CORPUS` (top left), `RULE` (top right), `LSTM` (bottom left), and `BERT` (bottom right).

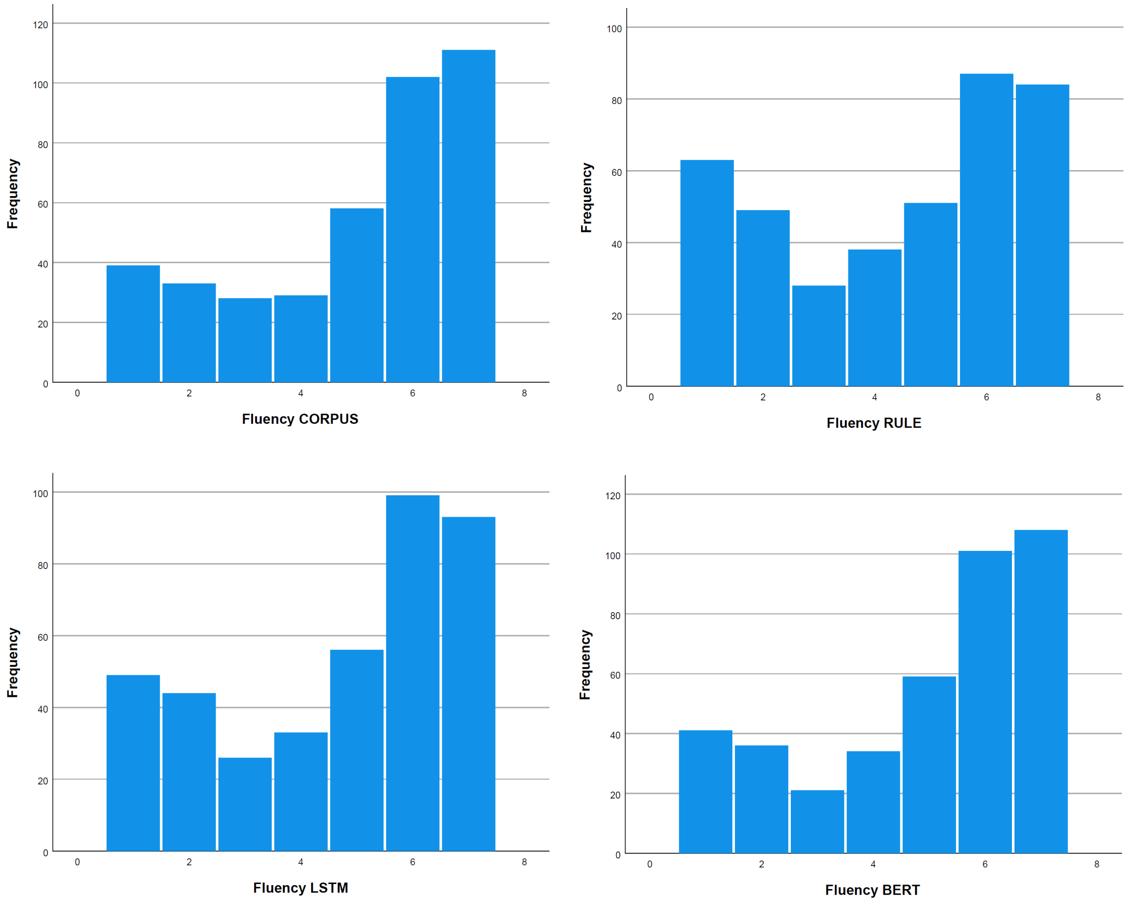


Figure 5: Graphs showing the frequency of Likert scores concerning Fluency for CORPUS (top left), RULE (top right), LSTM (bottom left), and BERT (bottom right).

5.5 Statistical analysis

There have been several studies in which automatic evaluations have been compared to human evaluations. Likert scale data has been compared to automatic evaluations in e.g. [Stent et al., 2005] (using the means of Likert scores (and also normalising them to the range [0,1]) and compare those to five automatic evaluation metrics for two sets of paraphrases, using Spearman's ρ), [Reiter and Belz, 2009] (using the means of Likert scores to compare expert and non-expert judgements using Pearson's r; and comparing the means of the human evaluations to the scores of five automatic evaluation metrics for six algorithms that generate weather forecasts, again using Pearson's r), and [Belz et al., 2010] (using the means of Likert scores (for individual text evaluation) and compare those to six automatic evaluation metrics for eight systems, using Pearson's r). There are many similar analyses we could base ours on. However, in these papers, the data has been analysed using the mean, which is a parametric method. Because data from Likert scales is ordinal[19], we are hesitant to use this for our own experiment. Preferably, we wanted to find a non-parametric way to do our analysis.

It did not prove easy to find other literature comparing Likert scores to accuracy results in a non-parametric way. One paper in which the data is analysed using non-parametric methods is [Elliott and Keller, 2014]. In this paper, the authors examined the correlation between human judgements and automatic evaluation metrics on textual image descriptions. The human judgements were measured using Likert scales, while the automatic evaluations were provided by BLEU (unigram and smoothed), ROUGE, TER, and Meteor. Correlations between the human judgements and automatic evaluation metrics were estimated at the sentence level (i.e. for each image description, the Likert score was compared to the automatic score from one of the metrics) using Spearman's ρ. Because this analysis uses only non-parametric measures, it provides an interesting example of how to look for correlations in Likert data.

However, like the other studies mentioned previously, their automatic evaluation metrics use quantitative data. In contrast, our automatic evaluation is categorical (i.e. there either is a match with the corpus or there is not). This means it would not be insightful to look further into a correlation (especially since, at the sentence level, there either is a match or not, while at the higher level (i.e. the overall accuracy scores per model), we would only have four data points - one accuracy score per model). Since the accuracy scores are categorical, we can better make use of them by splitting our data into two groups and comparing them to each other.

5.5.1 Mood's Median Test

At the sentence-level, the corpus-based evaluation is categorical. This is a good starting point for us to divide the data from `RULE`, `LSTM`, and `BERT` into two groups: one where the generated classifier matched with the corpus, and one where the generated classifier did not match with the corpus. We look at Fluency only, because we have the intuition that a given utterance can be clear while it simultaneously is obvious that a native speaker did not write it. The distribution of answers per group is shown in Figure 6.

The data is not paired, and, as shown in Figure 6, the distributions of the groups are not similar in shape. Instead of randomly choosing a point to split the data and compare distributions, the median seems like a clearer point of interest. Therefore, we chose to use Mood's Median Test, to look at the difference in the distribution of answers above and below the median.

[19] We treat Likert scales as ordinal, because you cannot, for example, compare the distance between "somewhat disagree"-"neither agree nor disagree" to the distance between "agree"-"strongly agree".

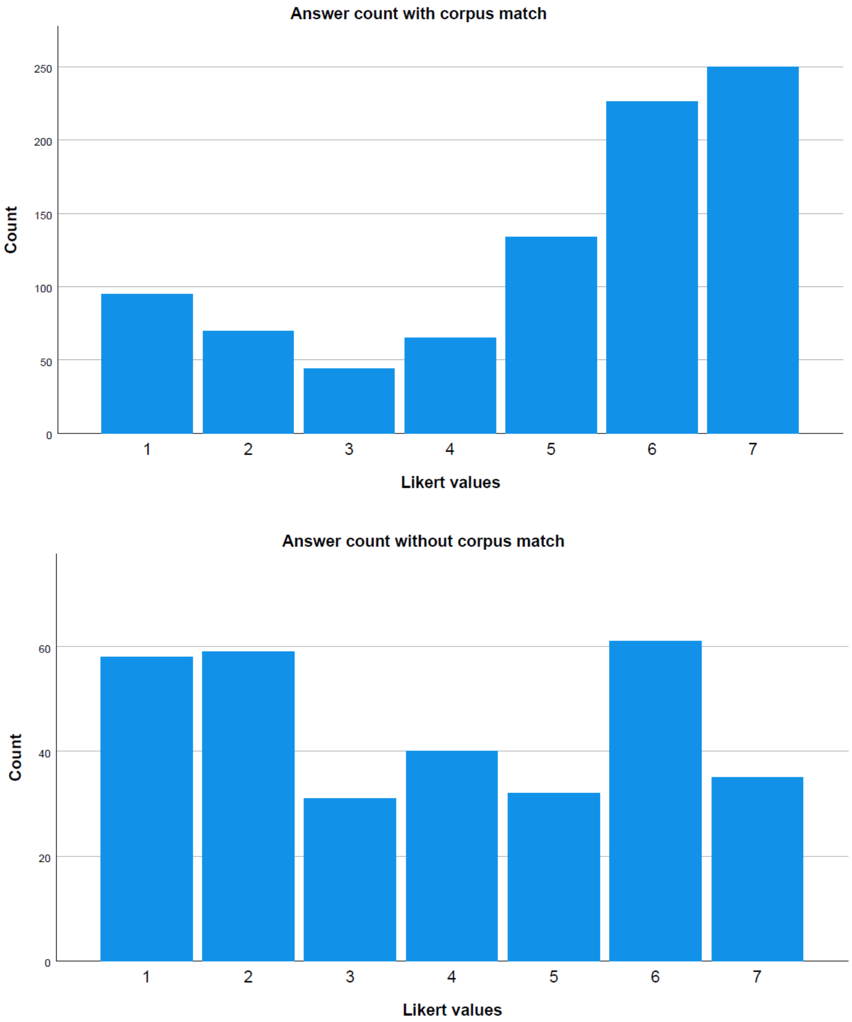


Figure 6: Answer count of classifiers that match with the corpus (top graph, N = 884) and classifiers that do not match with the corpus (bottom graph, N = 316).

	Less than or equal to median	Greater than median	Total
Corpus match	408	476	884
No corpus match	220	96	316
Total	628	572	1200

Table 3: Frequencies of classifiers that match with the corpus and classifiers that do not match with the corpus, compared to the overall median 5.

In SPSS, we performed Mood's Median Test. The Grand Median (i.e. the overall median of all data combined) is 5. Table 3 shows the frequency of answers below or equal to the median and answers above the median for both groups. The difference in distribution between the two groups is visualised in Figure 7. The last part of Mood's Median Test consists of a Chi-Squared Test, which gives us the following result:

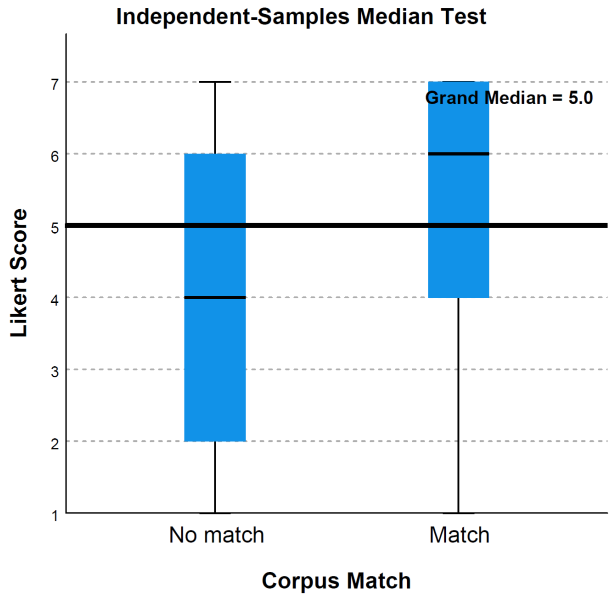


Figure 7: The difference in distribution between classifiers that match with the corpus and classifiers that do not match with the corpus, compared to the overall median 5.

- $\chi^2(1, N = 1200) = 53.33, p < .001$, with effect size[20] $\phi = .21$

We can conclude that there is a difference between generated classifiers that match the corpus and generated classifiers that do not match the corpus.

5.5.2 BERT compared to RULE and LSTM

We had two hypotheses concerning BERT and how it compares to the other two algorithms. We look at the Fluency scores. For each sentence that was selected from the CCD, we tested the chosen classifier for each model. Because we did this, we can for each sentence directly compare the classifiers (because for each sentence, a single participant judged all four chosen classifiers). To test these two hypotheses, we use the Wilcoxon Signed-Rank Test, because this is a non-parametric test for paired data:

- BERT is more fluent than RULE: $p < .001$, with effect size[21] $r = .31$.
- BERT is more fluent than LSTM: $p < .001$, with effect size $r = .19$.

BERT is indeed judged as more fluent than the other two algorithms. We interpret these results further in section 7.1.1.

[20] To calculate the effect size for Mood's Median Test, and subsequently in section 5.5.3 for the other Chi-Squared Tests, we use ϕ. This value is calculated using the χ^2 value with the formula $\phi = \sqrt{\frac{\chi^2}{N}}$, where N is the total sample size [Fritz et al., 2012]. The effect size ϕ can be interpreted as: .10 is a small effect, .30 is a medium effect, and .50 is a large effect [Cohen, 1988, pp. 224-225].

[21] To calculate the effect size for Wilcoxon's Signed-Rank Test, and subsequently for Experiment 2's Wilcoxon's Signed-Rank and Mann-Whitney U Tests (section 6.5), we use the r as proposed in [Cohen, 1988]. This value is calculated using the formula $r = \frac{z}{\sqrt{N}}$, where z is the z-value and N is the number of pairs (Wilcoxon's Signed-Rank) or the total sample size (Mann-Whitney U). The effect size r can be interpreted as: .10 is a small effect, .30 is a medium effect, and .50 is a large effect [Cohen, 1988, pp. 79-80].

5.5.3 Post hoc analysis: Chi-Squared Tests at different cut-off points

After doing Mood's Median test, we thought it interesting to also look at the frequency distribution at different cut-off points on the Likert scale. Would we still find significant differences if we split the data at points that are more conservative or more liberal than the median? Again, we only look at Fluency scores. The relative distributions are shown in Figure 8. We performed the χ^2 test to compare the distributions.

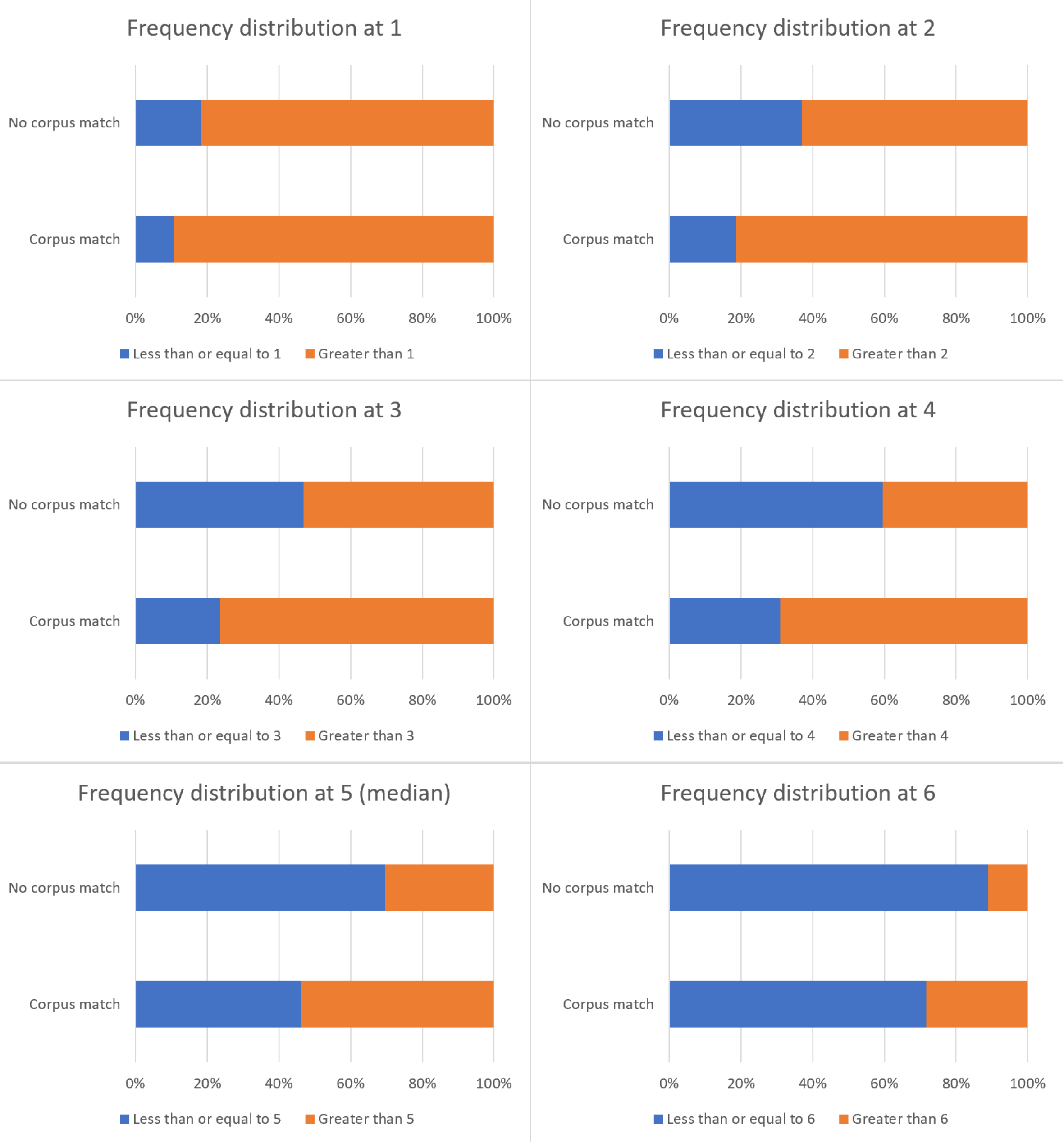


Figure 8: Differences in distribution between classifiers that match with the corpus and classifiers that do not match with the corpus, at all cut-off points. The distribution at 5, the median, is visualised for completeness; we performed no post hoc Chi-Squared Test for this point because we already performed Mood's Median Test.

- When we put the cut-off point at lower or equal to 1, the Chi-Square Test shows there is a significant difference, $\chi^2(1, N = 1200) = 12.11, p < .001$, with effect size $\phi = .10$.
- When we put the cut-off point at lower or equal to 2, the Chi-Square Test shows there is a significant difference, $\chi^2(1, N = 1200) = 43.65, p < .001$, with effect size $\phi = .19$.
- When we put the cut-off point at lower or equal to 3, the Chi-Square Test shows there is a significant difference, $\chi^2(1, N = 1200) = 59.92, p < .001$, with effect size $\phi = .22$.
- When we put the cut-off point at lower or equal to 4, the Chi-Square Test shows there is a significant difference, $\chi^2(1, N = 1200) = 79.85, p < .001$, with effect size $\phi = .26$.
- When we put the cut-off point at lower or equal to 6, the Chi-Square Test shows there is a significant difference, $\chi^2(1, N = 1200) = 38.05, p < .001$, with effect size $\phi = .18$.

Considering a Bonferroni correction, to account for the use of five separate Chi-Sqaured Tests, the resulting p-values are still so small that distributions between the groups differ significantly at all cut-off points, with non-negligible effect sizes. Since a difference is shown for all Likert values instead of only the median, this gives us stronger indications of a difference between generated classifiers that match the corpus and generated classifiers that do not match the corpus.

Using the results from Mood's Median Tests and these post hoc tests, we answer the research question in the Discussion, section 7.1.1.

5.5.4 Post hoc analysis: Correlation between Clarity and Fluency

To get a better feel for the data, we took the means of the Likert scores for all models, as shown in Table 4.

model	Clarity	Fluency	Corpus-based
`CORPUS`	5.095	4.960	100.00
`RULE`	4.563	4.405	61.73
`LSTM`	4.805	4.680	73.86
`BERT`	5.015	4.923	81.71

Table 4: In the columns for Clarity and Fluency, the means of Likert scores are shown. In the column for the corpus-based evaluation, the accuracy scores are shown (i.e. the percentage of generated classifiers that matches the one present in the corpus).

The first thing that we noticed were the similarities between Clarity and Fluency. As these values already seemed quite similar when looking at the raw data in section 5.4, we decided to look closer into this. We used Spearman's Rank Correlation to compare Clarity and Fluency on the sentence level:

- We found a very strong positive correlation, $\rho(1598) = .90$, $p < .001$.

We interpret this result further in section 7.2.1 in the Discussion.

5.5.5 Post hoc analysis: Differences between individual models

Looking again at Table 4, the other thing we noticed were the near identical values between `CORPUS` and `BERT` in the human experiment for both Clarity and Fluency. This was not something we had expected, especially since the gap between `CORPUS` and `BERT` is relatively large in the corpus-based evaluation. We decided to compare `BERT` to the corpus and the other models to explore the (lack of) differences we notice in the table. We use Wilcoxon's Signed-Rank Test, as this is a non-parametric test that works with paired data:

- Clarity - CORPUS and BERT: $p = .065$, with effect size $r = .09$.
- Fluency - CORPUS and BERT: $p = .480$, with effect size $r = .04$.
- Clarity - RULE and BERT: $p < .001$, with effect size $r = .30$.
- Clarity - LSTM and BERT: $p < .001$, with effect size $r = .19$.

We interpret these results further in section 7.2.2 in the Discussion.

6 Experiment 2: True classifiers

Given the outputs by various algorithms, is the choice between classifiers important to human readers of Mandarin? For this experiment, we want to examine those sentences in which there is a genuine choice between *ge* 个 and more specific classifiers. To that extent, it is important to only look at true classifiers (as *ge* 个 can only be used in these cases).

We know from previous evaluations that various algorithms make different choices compared to both the corpus and each other. However, we do not know to what extent these differences matter to human readers, or if they matter to them at all. We will be looking at the classifier choices provided by the corpus and the three algorithms GE, RULE, and BERT, because:

- CORPUS: Since the sentences in the corpus are produced by humans, they can function as a baseline for the choices made by the algorithms.
- GE: In this first experiment, we also specifically look at infrequent classifiers. We want to know people's opinions on the use of *ge* 个 in these contexts.
- BERT: Of the four algorithms provided and tested by [Järnfors, 2021], BERT seems to be performing best. Since it also in some cases seems to perform the task better then humans, we want to know how people evaluate BERT's choices compared to the classifiers in the corpus.
- RULE: This rule-based algorithm had a relatively high accuracy in predicting true classifiers, given that it is such a computationally simple algorithm compared to the other ones in [Järnfors, 2021]. Since it is so simple, we mainly want to compare people's opinions on RULE's choices to those of BERT, which is a far more complex algorithm.

We want our participants to see the context around the classifiers, which means we show them the sentences as they appear in the corpus (for each algorithm, we only change the classifier if it is different from the one in the corpus). Our dataset consists of 200 sentences: 100 sentences that were randomly selected, to approximate "normal" language use, and 100 sentences including some of the most infrequent classifiers. The inclusion of infrequent classifiers is important because in these cases it will be harder for the algorithms to make a choice (since the algorithm has rarely or never seen the classifier before).

6.1 Constructing the dataset

Similar to Experiment 1, we used the CCD as our starting point for collecting the sentences. For constructing our own dataset, we started by following the same methodology as the human experiment in [Chen, 2022]. Our dataset consists of two groups of sentences:

1. For the first group, sentences were randomly selected from the CCD. Each sentence was manually filtered to make sure we only include true classifiers and exclude noise[22]. Using this method, we collected 100 sentences in the first group.

[22] "Noise" in the context of our experiments means "sentences that are not intelligible and sentences in which a word was, given its context, wrongly tagged as classifier".

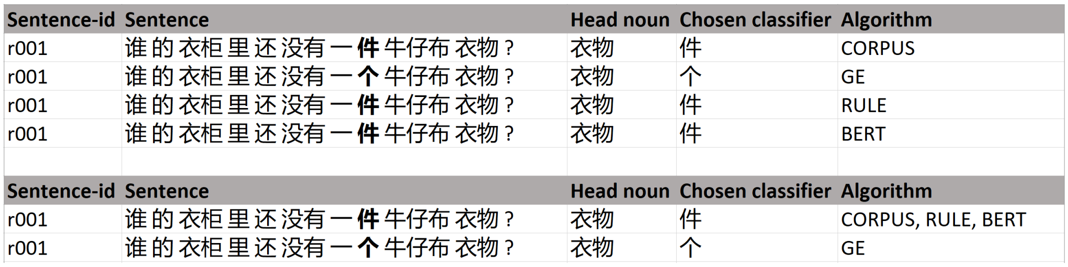
2. For the second group, we wanted to cover infrequent classifiers, to include how human participants would judge cases in which the algorithms would have more difficulty predicting the classifier. We selected classifiers which appeared less than 500 times in the CCD. For each classifier that met this criterion, we randomly selected three sentences. Just like for the first group, these were manually filtered to only include true classifiers and exclude noise. Using this method, we collected 100 sentences in the second group.

Combining these two groups results in a dataset containing 200 sentences. For each of these sentences, we list four classifiers; one from the corpus, and three generated by algorithms:

- `CORPUS`: The classifier as it appears in that specific sentence in the CCD.
- `GE`: Always assign the general classifier *ge* 个.
- `RULE`: The classifier generated by the rule-based model in [Järnfors, 2021]: given a head noun, assign the classifier most frequently associated with it in the training data; if at least two classifiers are equally frequent, one of them is randomly assigned; if the head noun is not present in the training data, assign the general classifier *ge* 个.
- `BERT`: The classifier generated by the BERT classification model in [Järnfors, 2021].

After generating four classifiers for every sentence, we have a dataset containing 800 items.

Because it would not be efficient for the participants to rate the exact same sentence twice, we removed identical items using the same method as in Experiment 1 (i.e. when given a sentence, if two out of four choices are identical, we remove one of the identical sentences; similarly, if three out of four choices are identical, we remove two of those sentences; and if all four choices are identical, we remove three of those sentences. An example is given in Figure 9).



Sentence-id	Sentence	Head noun	Chosen classifier	Algorithm
r001	谁的衣柜里还没有一**件**牛仔布 衣物？	衣物	件	CORPUS
r001	谁的衣柜里还没有一**个**牛仔布 衣物？	衣物	个	GE
r001	谁的衣柜里还没有一**件**牛仔布 衣物？	衣物	件	RULE
r001	谁的衣柜里还没有一**件**牛仔布 衣物？	衣物	件	BERT

Sentence-id	Sentence	Head noun	Chosen classifier	Algorithm
r001	谁的衣柜里还没有一**件**牛仔布 衣物？	衣物	件	CORPUS, RULE, BERT
r001	谁的衣柜里还没有一**个**牛仔布 衣物？	衣物	个	GE

Figure 9: An example of how we removed identical items. The upper table shows the four items generated for sentence *r001*. Because `RULE` and `BERT` chose the same classifier as `CORPUS`, we only have to show this item once to the participants to get their judgement for these models. `GE` chose a different classifier, so we keep that item as is. After removing these duplicate items, the upper table can be compacted into the lower table.

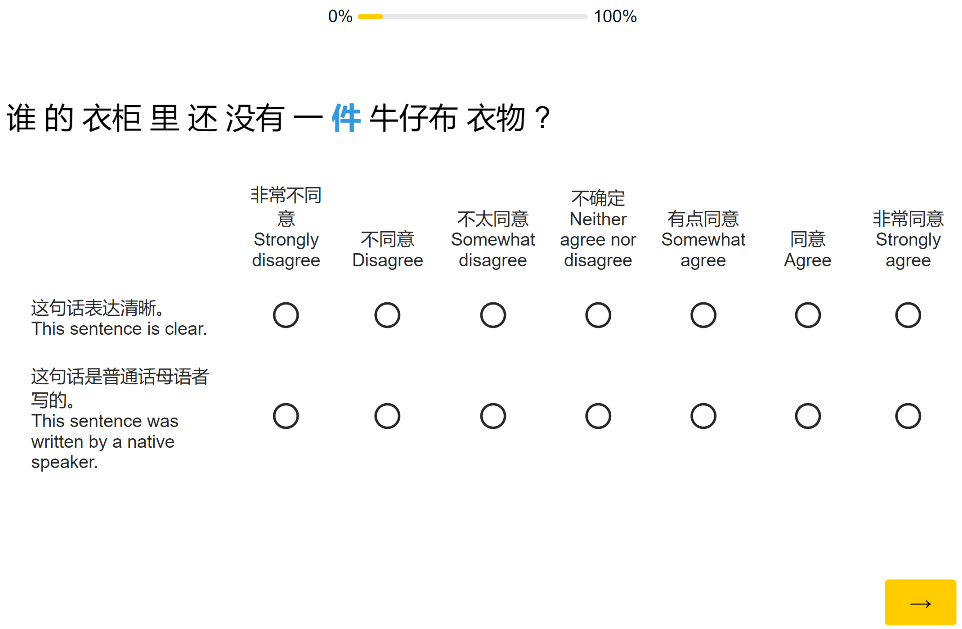
After removing identical items, we have a final dataset containing 403 items: 63 items for which there was only one choice (63 sentences); 160 items for which there were two choices (80 sentences); 144 items for which there were three choices (48 sentences); and 36 items for which there were four choices (9 sentences). The full dataset can be found in Appendix B.

6.2 Experiment design

The experiment is conducted online, using Qualtrics Survey Software - UU[23]. Participants could partake at a time and place of their choosing. We have set the default language to Chinese (simplified),

[23] Qualtrics is an online survey tool for setting up and distributing questionnaires, and collecting and analysing data. It is available in general via https://www.qualtrics.com, and for UU staff and students via survey.uu.nl

to ensure participants who do not understand English can read the messages provided by Qualtrics itself (e.g. when the program warns the participant that a question is not answered yet). All other text (i.e. everything we have written) is provided in both Chinese and English. The questionnaire starts with a page containing general information about the experiment, ensuring participants can give their informed consent for participating. The second page contains a series of personal questions for statistics research. Then follows the experiment section.



谁 的 衣柜 里 还 没有 一 **件** 牛仔布 衣物？

	非常不同意 Strongly disagree	不同意 Disagree	不太同意 Somewhat disagree	不确定 Neither agree nor disagree	有点同意 Somewhat agree	同意 Agree	非常同意 Strongly agree
这句话表达清晰。 This sentence is clear.	○	○	○	○	○	○	○
这句话是普通话母语者写的。 This sentence was written by a native speaker.	○	○	○	○	○	○	○

Figure 10: Display of how the sentences were presented to the participants. The model-chosen classifier is boldfaced and highlighted in blue.

In the experiment section, participants see one sentence item at a time. Because this concerns intuitive language use, it was important to get answers as intuitively as possible within the constraints of this questionnaire design. Therefore, we did not provide the possibility to return to previous sentences, to ensure the participants could not retroactively change their answers. To make the generated classifier easily identifiable, it has been highlighted (i.e. boldfaced and coloured blue). For every classifier choice, the participants answered two statements:

- CLARITY: This sentence is clear. 这句话表达清晰。
- FLUENCY: This sentence was written by a native speaker. 这句话是普通话母语者写的。

The participants answered the questions on a 7-point Likert scale, with the following options: strongly disagree 非常不同意; disagree 不同意; somewhat disagree 不太同意; neither agree nor disagree 不确定; somewhat agree 有点同意; agree 同意; strongly agree 非常同意.

Our dataset consists of 403 sentence items, which we deemed too much to ask from each of our individual participants. To present each participant with a more reasonable number of items, we divided the dataset into five groups, consisting of 80-81 items each. This was done manually, to ensure every group contained roughly the same amount of variation in classifier choices (as determined by the algorithms' classifier choices) and minimise differences. For each group, the order of the items was randomised. After randomisation, we made sure no sentence item directly followed a variation of the same sentence, to decrease the possibility that participants base their answers on the previous one. Each group of items was incorporated into a separate version of the experiment, producing five versions of the experiment.

6.3 Participants

A total of 29 participants completed the experiment, 6 female and 23 male, with ages ranging from 18 to 31 years (M = 21.9, SD = 2.1). All participants were students from Harbin Institute of Technology; 11 Bachelor's students, 15 Master's students, and 3 PhD candidates. Of the 29 participants, 27 had a study background in computer science and 2 had other technical backgrounds. The native language of all participants is Mandarin Chinese, with two participants having an additional native language (one Cantonese, the other Sichuanese). Regarding languages learnt later in life, 26 participants know at least some English, one knows Sichuanese, and one knows some Japanese. All participants had given their informed consent before participating in the experiment.

The participants were divided over the five versions of the experiment. Version 1 was completed by six participants; version 2 was completed by six participants; version 3 was completed by five participants; version 4 was completed by five participants; and version 5 was completed by seven participants.

6.4 Overview gathered results

For analysing the data, we use assigned numbers to represent the participants' answers:

1. strongly disagree 非常不同意
2. disagree 不同意
3. somewhat disagree 不太同意
4. neither agree nor disagree 不确定
5. somewhat agree 有点同意
6. agree 同意
7. strongly agree 非常同意

The frequency of answers is given in Tables 5 to 8 and visually represented in Figures 11 to 14.

Experiment 2 - Clarity - random								
model	1	2	3	4	5	6	7	total
CORPUS	27	24	42	23	76	175	221	588
GE	43	37	46	23	78	173	188	588
RULE	38	30	47	22	69	180	202	588
BERT	30	26	42	25	75	173	217	588

Table 5: The frequency of answers concerning Clarity for the random group.

Experiment 2 - Fluency - random								
model	1	2	3	4	5	6	7	total
CORPUS	40	35	42	25	65	159	222	588
GE	65	51	53	29	51	152	187	588
RULE	61	35	46	21	60	162	203	588
BERT	45	36	45	27	61	162	212	588

Table 6: The frequency of answers concerning Fluency for the random group.

Experiment 2 - Clarity - infrequent								
model	1	2	3	4	5	6	7	total
CORPUS	26	28	42	27	69	185	196	573
GE	42	62	61	42	81	182	103	573
RULE	44	45	62	35	74	169	144	573
BERT	26	25	42	39	73	182	186	573

Table 7: The frequency of answers concerning Clarity for the infrequent group.

Experiment 2 - Fluency - infrequent								
model	1	2	3	4	5	6	7	total
CORPUS	28	43	53	36	66	147	200	573
GE	77	97	83	53	70	109	84	573
RULE	69	75	68	50	54	121	136	573
BERT	33	49	51	51	61	151	177	573

Table 8: The frequency of answers concerning Fluency for the infrequent group.

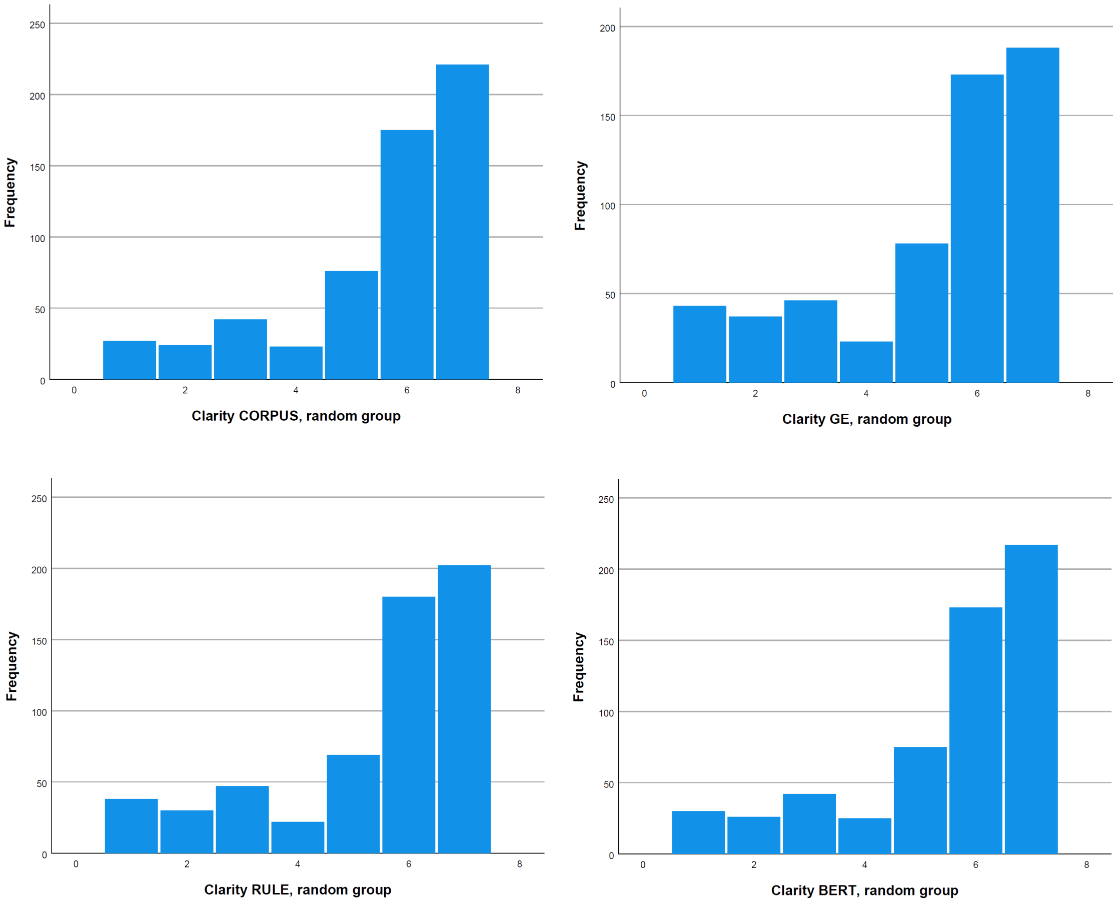


Figure 11: Graphs showing the frequency of Likert scores concerning Clarity in the random group for CORPUS (top left), GE (top right), RULE (bottom left), and BERT (bottom right).

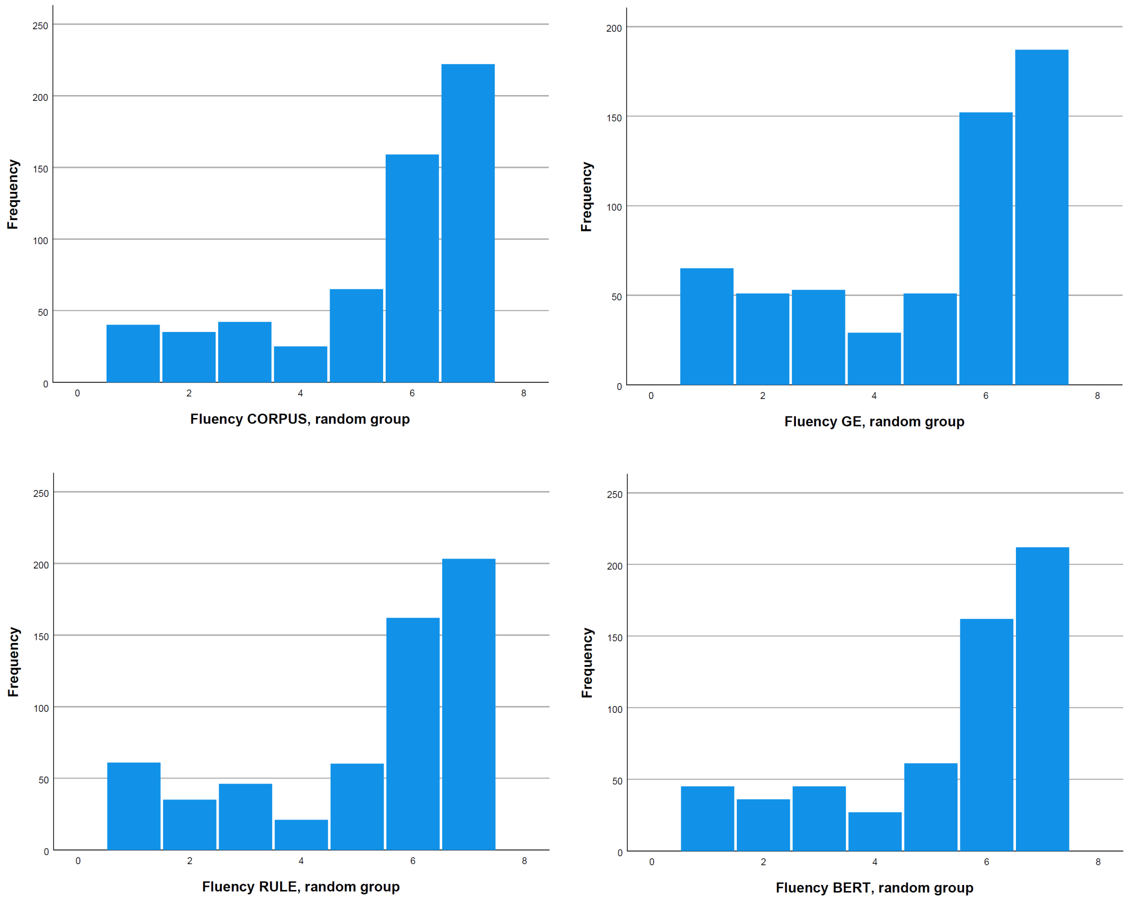


Figure 12: Graphs showing the frequency of Likert scores concerning Fluency in the random group for `CORPUS` (top left), `GE` (top right), `RULE` (bottom left), and `BERT` (bottom right).

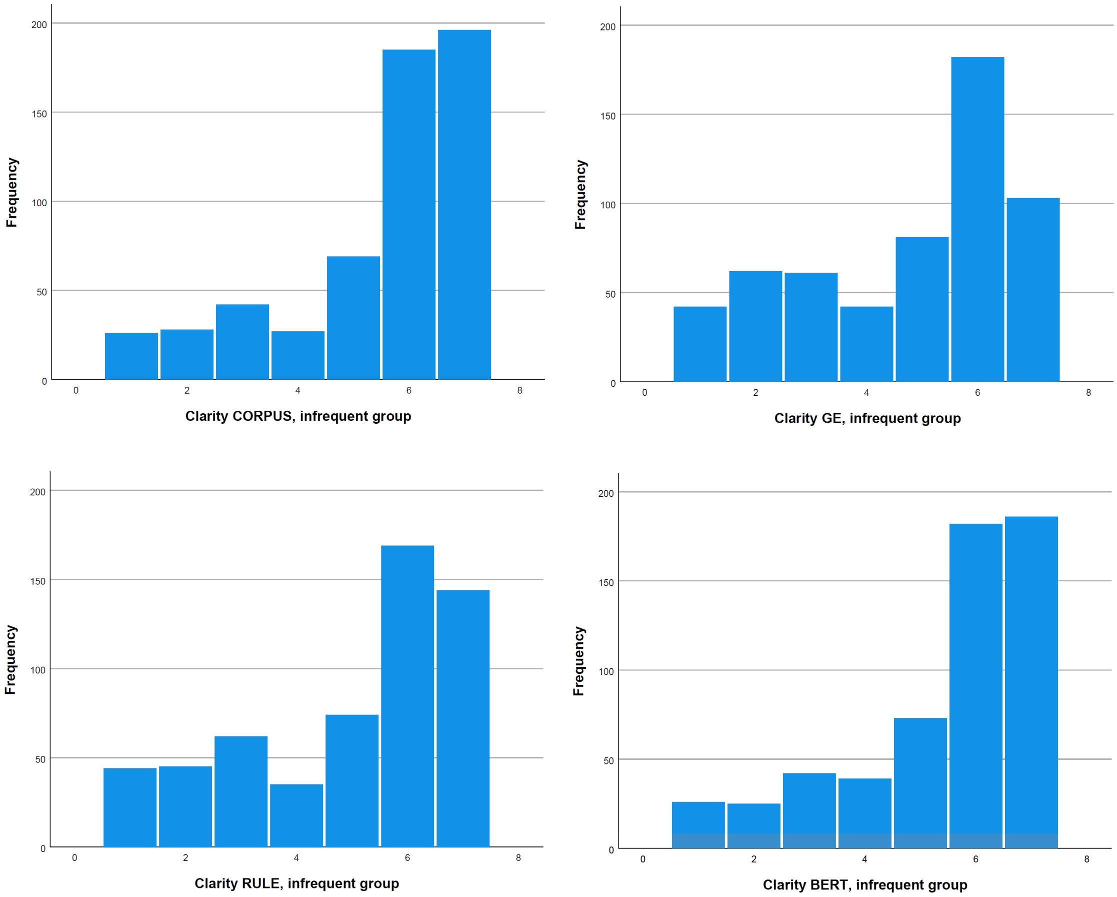


Figure 13: Graphs showing the frequency of Likert scores concerning Clarity in the infrequent group for `CORPUS` (top left), `GE` (top right), `RULE` (bottom left), and `BERT` (bottom right).

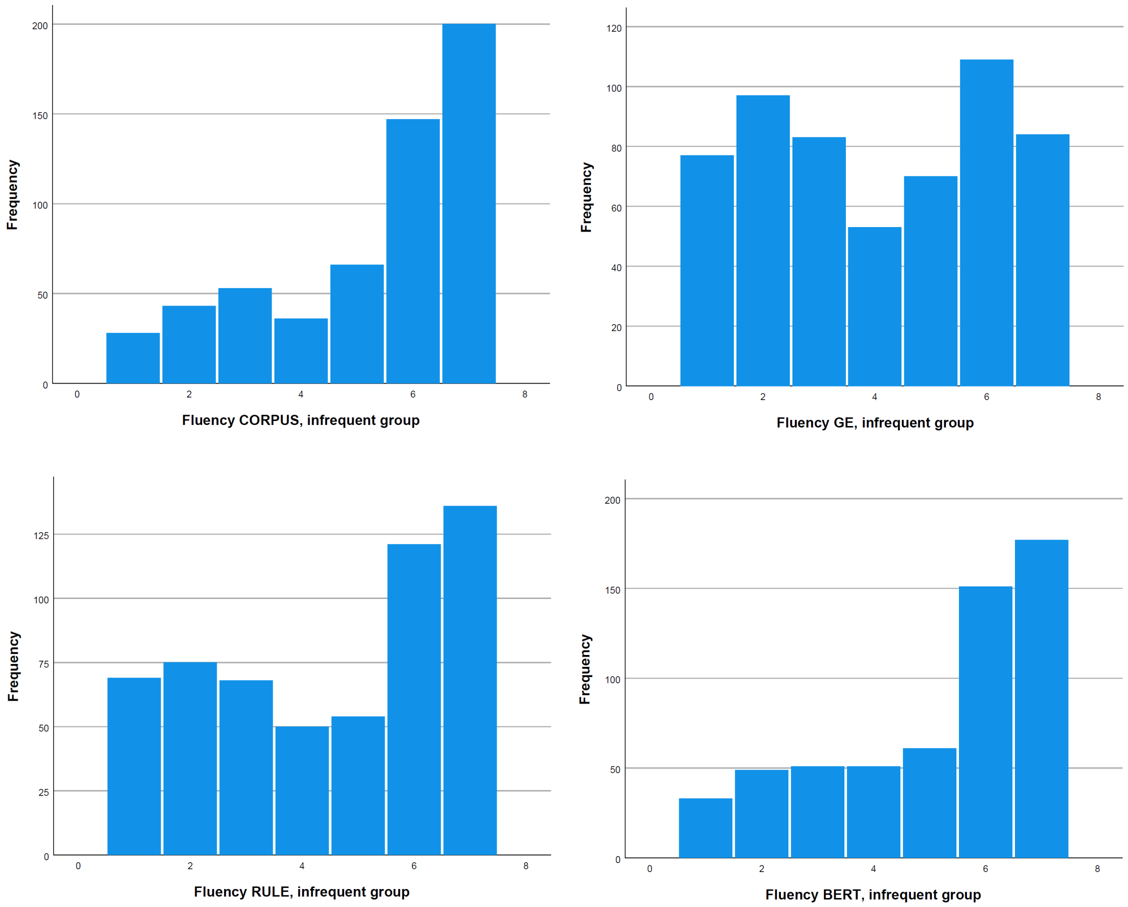


Figure 14: Graphs showing the frequency of Likert scores concerning Fluency in the infrequent group for `CORPUS` (top left), `GE` (top right), `RULE` (bottom left), and `BERT` (bottom right).

6.5 Statistical analysis

After obtaining the data, we tested the five hypotheses. Because all hypotheses are about the Fluency dataset, we use a Bonferroni correction to adjust the significance threshold to .010. We do this to reduce the probability of false positives (i.e. a hypothesis possibly being confirmed by accident). For the first hypothesis (`BERT` random compared to `BERT` infrequent), we use the Mann-Whitney U Test, as this is a non-parametric test for comparing independent data. For the other four hypotheses, we can directly compare the classifiers scores for each sentence. Therefore, we use the Wilcoxon Signed-Rank Test, a non-parametric test for paired data. These are the results:

1. `BERT` random is more fluent than `BERT` infrequent: $p = .043$, with effect size $r = .06$.
2. `GE` is less fluent than `CORPUS`: $p < .001$, with effect size $r = .30$.
3. `GE` is less fluent than `RULE`: $p < .001$, with effect size $r = .14$.
4. `BERT` is more fluent than `GE`: $p < .001$, with effect size $r = .25$.
5. `BERT` is more fluent than `RULE`: $p = .012$, with effect size $r = .10$.

We could not find a significant difference between `BERT` random and `BERT` infrequent. The result of .043 does not get near the threshold of .010. The performance of `BERT` with the infrequent group is better than we expected. Using *ge* 个 in every context clearly seems to be regarded as less fluent than using a classifier chosen by human speakers (i.e. `CORPUS`) and the other models. The fluency of `RULE` and `BERT` differs much less; the result of .012 was not significant and we cannot assume a difference. As `RULE` is a rule-based algorithm working with just three rules, it performs relatively well compared to the machine learning results of `BERT`. We interpret these results further in section 7.1.2.

6.5.1 Post-hoc analysis: Correlation between Clarity and Fluency

In the previous experiment, we found a very strong correlation between Clarity and Fluency (see section 5.5.4). Because the raw data for this second experiment shows comparable similarities, we are curious to see if we can also find such a correlation here. We used Spearman's Rank Correlation to compare Clarity and Fluency on the sentence level for both the random group and the infrequent group:

- For the random group, we found a strong positive correlation, $\rho(2350) = .85$, $p < .001$.
- For the infrequent group, we found a strong positive correlation, $\rho(2290) = .81$, $p < .001$.

We interpret these results further in section 7.2.1 in the Discussion.

6.5.2 Post-hoc analysis: Differences between `CORPUS` and `BERT`

In the previous experiment, we could not find a difference between the noticeably similar results by `CORPUS` and `BERT`. We are curious whether `CORPUS` and `BERT` show these similarities in this experiment too. Therefore, we use Wilcoxon's Signed-Rank Test to compare these two models:

- Clarity - random group - `CORPUS` and `BERT`: $p = .063$, with effect size $r = .08$.
- Fluency - random group - `CORPUS` and `BERT`: $p = .011$, with effect size $r = .10$.
- Clarity - infrequent group - `CORPUS` and `BERT`: $p = .366$, with effect size $r = .04$.
- Fluency - infrequent group - `CORPUS` and `BERT`: $p = .017$, with effect size $r = .10$.

We interpret these results further in section 7.2.2 in the Discussion.

7 Discussion

In this section, we summarise the experiments, interpret the results, and discuss other findings and ideas for future research.

7.1 Answering the research questions

We return to the research questions presented in section 4.2 and answer them using the analyses presented in section 5.5 for Experiment 1, and section 6.5 for Experiment 2.

7.1.1 Experiment 1: Standard NLG evaluation

First, we presented our standard NLG evaluation, in which all classifiers were included (i.e. true classifiers, dual classifiers, and measure words). We were building upon [Järnfors, 2021] by adding a human evaluation to the already existing corpus-based evaluation. Our research questions are:

- How does the corpus-based evaluation of the models `RULE`, `LSTM`, and `BERT` compare to the human evaluation?
- Are `BERT`'s classifier choices rated as more fluent than the choices made by `RULE` and `LSTM`?

Because the corpus-based accuracy results are binary on the sentence level (i.e. a chosen classifier either matches with the corpus or it does not), we divided the data into two groups, and followed with the hypothesis:

1. The values in the first group (the chosen classifier matches with the corpus) are higher than those in the second group (the chosen classifier does not match with the corpus).

   Additionally, we had two hypotheses concerning `BERT`:

2. `BERT` is more fluent than `RULE`.
3. `BERT` is more fluent than `LSTM`.

To test the first hypothesis, comparing the first and second groups, we used Mood's Median Test. The difference was highly significant ($p < .001$), with a small-medium effect size ($\phi = .21$). From this, we conclude our hypothesis can be confirmed. After performing Mood's Median Test, we got curious about whether this difference would also show at other cutoff points than the median. So, post hoc, we conducted five consecutive Chi-Squared Tests at the remaining Likert values (i.e. at 1, 2, 3, 4, and 6, since 5 was the median value). All five tests showed high significance ($p < .001$), with effect sizes of $\phi = .10$, $\phi = .19$, $\phi = .22$, $\phi = .26$, and $\phi = .18$ for cutoff points 1, 2, 3, 4, and 6, respectively. These post hoc results provide additional support for the difference between the first and second groups and the confirmation of the hypothesis.

In the corpus-based evaluation, the number of matches with the corpus determines the accuracy score; more matches with the corpus means a higher accuracy score. Returning to the research question, we can conclude that it is also the case for human evaluation that chosen classifiers that match with the classifier in the corpus are valued higher than classifiers that do not match with the corpus. The corpus-based and human evaluations seem to be at least somewhat comparable in this regard.

However, after some additional post hoc testing, we could not find significant differences between `CORPUS` and `BERT`. Human evaluators could be more lenient than corpus-based scores when classifiers do not match the corpus. This difference between the automatic and human evaluations can indicate the use of accuracy as an evaluation metric could be flawed, at least if we want to know a generated text's acceptability to human readers. The post hoc tests and their results are discussed further in section 7.2.2.

To test the second and third hypotheses, we used Wilcoxon's Signed-Rank Test. Comparing BERT and RULE, we found a highly significant difference ($p < .001$) with a medium effect size ($r = .31$). Likewise, when comparing BERT and LSTM we found a highly significant difference ($p < .001$) with a small-medium effect size ($r = .19$). We can conclude that BERT's classifier choices are regarded as more fluent than those made by the other algorithms.

7.1.2 Experiment 2: True classifiers

After the first experiment, we present a more linguistically motivated experiment in which we focus on true classifiers only. From the previous experiment, we keep the algorithms RULE and BERT, since in the automatic evaluation, RULE performed notably better with only true classifiers, and BERT outperformed all other models. Additionally, we look at a new model, GE, which only assigns the general classifier *ge* 个. We did this because this more general classifier would be acceptable as a substitution for all other true classifiers (as observed in child language acquisition, aphasic language loss, and informal everyday language use, discussed in section 2). We were interested in how readers would rate the use of this classifier. Furthermore, we were interested in the difference between two groups of true classifiers: randomly sampled classifiers (to imitate a more "natural" use of language, in which a majority of sentences uses frequent classifiers like *ge* 个) and infrequent classifiers (because those reduce the possibility that the classifier has been seen before in training data, which could provide difficulties for some algorithms). We stated our broad research question as:

- How do the classifier choices by the models CORPUS, GE, RULE, and BERT compare to each other, both with frequent and infrequent head nouns?

We divide this question into two more specific sub-questions:

- Given that BERT performs better than the other models, are its chosen classifiers in the first group (random) rated as more fluent compared to the second group (infrequent)?
- How does the choice of the general classifier *ge* 个 compare to the choices by CORPUS, RULE, and BERT?

We had five hypotheses (where hypotheses 2 to 5 are concerned only with the random group):

1. BERT random is more fluent than BERT infrequent.
2. GE is less fluent than CORPUS.
3. GE is less fluent than RULE.
4. BERT is more fluent than GE.
5. BERT is more fluent than RULE.

For testing the first hypothesis, we used the Mann-Whitney U Test. We did not find a significant difference ($p = .043$), with a very small effect size ($r = .06$). We expected BERT to perform better with the higher amount of training data in the random group. However, we did not find a difference in fluency between the two groups. Therefore, to answer the first sub-question, we can say BERT does not seem to perform better with the first group compared to the second group. It seems that BERT performs better than we expected with only little training data.

For testing hypotheses 2 to 5, we used Wilcoxon's Signed-Rank Test. Looking at the hypotheses that concern *ge* 个, we find a highly significant difference ($p < .001$) with medium effect size ($r = .30$) when comparing GE with CORPUS (hypothesis 2), a highly significant difference ($p < .001$) with small effect size ($r = .14$) when comparing GE with RULE (hypothesis 3), and a highly significant difference ($p < .001$) with small-medium effect size ($r = .25$) when comparing GE with BERT (hypothesis 4). For the second sub-question, we can say the use of *ge* 个 in all contexts is perceived as less fluent compared to the choices made by both native speakers and the algorithms RULE and BERT. Looking at only effect size, the differences between GE on the one hand and CORPUS and BERT on the other are more prominent than the difference between GE and RULE. It could be that RULE chooses the general classifier slightly more compared to native speakers and BERT, which would be a possible explanation for this difference in effect size.

For our last hypothesis concerning RULE and BERT, we do not find a significant difference ($p = .012$) and a small effect size ($r = .10$). Considering RULE is a simple rule-based algorithm, it performs better than we expected compared to the state-of-the-art algorithm BERT. It seems that, compared to CORPUS and BERT, RULE is more suitable for choosing true classifiers than we initially thought.

7.2 Other findings: Post hoc analyses

During and after data collection, multiple questions arose about our findings. Most notably, we were curious about the, to us, unexpected similarities between Clarity and Fluency and between CORPUS and BERT. Therefore, we performed multiple additional post hoc tests, as presented in section 5.5 for Experiment 1, and section 6.5 for Experiment 2. In the following section, we interpret the results.

7.2.1 Clarity and Fluency

The first thing we noticed was that, in both experiments, the results for Clarity and Fluency seemed similar. Before analysing the data, we expected Clarity to be judged higher than Fluency. Our reasoning was that the use of a given classifier could be clear, even though it would not be the classifier a native speaker would use (compare this to a Mandarin L2 learner who does not use the most appropriate classifier in a given context; a native speaker would in many cases still understand what the Mandarin learner was trying to convey).

Because the Clarity and Fluency results seemed so similar, we conducted some post hoc tests to look for a correlation in both experiments. We used Spearman's Rank Correlation to compare the Likert values on the sentence level. The comparison for Experiment 1 showed a significant ($p < .001$) strongly positive correlation ($\rho = .90$) between Clarity and Fluency. Similarly, for Experiment 2, both the random group and the infrequent group showed a significant ($p < .001$) strongly positive correlation ($\rho = .85$ for the random group, $\rho = .81$ for the infrequent group).

There are two main reasons we can imagine that explain this Clarity-Fluency correlation:

- One or both concepts were not interpreted as we intended. It is a known problem within NLG evaluation when employing questionnaires to gauge participants' opinions; many different approaches and terms are used to describe the qualities of a text (which could be difficult to interpret for the participants, and are sometimes even vague to researchers themselves), and standardised methodology and terminology are still absent within the field [Howcroft et al., 2020]. There is the additional complexity of conducting experiments with laypeople; we aimed to explain these concepts with a short and understandable statement (*This sentence is clear.* 这句话表达清晰。for Clarity, and *This sentence was written by a native speaker.* 这句话是普通话母语者写的。for Fluency). Still, we cannot be certain that a layperson interprets these statements in the same manner as someone familiar with NLG text evaluation. Participants could have been overthinking what it means for (part of) a sentence to be clear[24].

- Alternatively, these concepts may influence each other more than we initially realised. Maybe it is the case that the Fluency of a given phrase contributes to its Clarity. Considering participants saw the classifiers not only within the noun phrase, but were given the context of a full sentence, a wrongly selected classifier could result in the meaning of the sentence becoming less clear.

[24] i.e. It could have been the case that most classifier choices were perfectly clear. However, because we ask the Clarity rating for each sentence, participants could have interpreted "clear" in an unusual manner to ensure they had variation in their answers.

It is worth noting that these correlations apply only within the context of the two experiments we conducted; this could be either an issue with how the concepts of Clarity and Fluency were defined and interpreted, or it could be a correlation possibly within the context of classifier selection in Mandarin[25]. We do not know if this correlation is applicable within a broader context of NLG.

7.2.2 Corpus classifiers and BERT's results

BERT has yielded interesting results in both experiments, especially compared to CORPUS. In the first experiment, while BERT was evaluated significantly better compared to RULE and LSTM ($p < .001$ in all cases; medium effect sizes compared to RULE Clarity ($r = .30$) and Fluency ($r = .31$), and small-medium effect sizes compared to LSTM Clarity ($r = .19$) and Fluency ($r = .19$)), when we compare BERT to CORPUS we do not find a significant difference ($p = .065$, $r = .09$ for Clarity and $p = .480$, $r = .04$ for Fluency). In the second experiment, we similarly find no significant differences and (very) small effect sizes between BERT and CORPUS ($p = .063$, $r = .08$ for Clarity in the random group; $p = .011$, $r = .10$ for Fluency in the random group; $p = .366$, $r = .04$ for Clarity in the infrequent group; and $p = .017$, $r = .10$ for Fluency in the infrequent group).

Looking back at the corpus-based accuracy results presented in Table 4, BERT had an accuracy score of 81.71%, while CORPUS has an accuracy score of 100% (as should be expected). Since we could not find meaningful differences between the classifiers used by native speakers in CORPUS and the classifiers chosen by BERT in our human evaluations, this seems to support the classic argument against using automatic metrics alone: if a classifier does not match with the corpus, it does not necessarily mean native speakers interpret it as "wrong". The classifier choice does not seem to matter as much to readers as the corpus-based accuracy scores suggest.

7.3 Limitations & future research

In this section, we share the ideas that arose while working on the experiments. Some came about as a result of curiosity after analysing the data, while others result from limitations we encountered. In all cases, we could unfortunately not look into them further because of time constraints.

To follow up on our human evaluation, a task-based experiment would provide some interesting additional information. Participants would be shown phrases containing a classifier chosen by an algorithm, and afterwards, reaction times are compared to see whether there are any differences between algorithms. This could be done both for the standard NLG evaluation and the evaluation focussing on true classifiers. For example, in Experiment 2, we found GE to be rated as less fluent than CORPUS, RULE, and BERT. In a task-based experiment, participants would be presented with a (part of a) sentence containing a true classifier - either *ge* 个 or a specific one. The participants are asked a question about the text they are presented with; their task is to answer as quickly as possible. If it is indeed the case that *ge* 个 is rated lower than more specific classifiers, we would expect the reaction times for GE to be longer than those for CORPUS, RULE, and BERT.

Looking at the strong correlation between Clarity and Fluency in both experiments, it is imaginable that these concepts influence each other. Possibly, they are more intertwined in the context of Mandarin classifiers. It would be interesting to examine whether readers only distinguish between Clarity and Fluency in specific situations. To do this, we could return to our gathered data and focus on the sentences in which the scores for Clarity and Fluency diverged the most. Examining the chosen classifiers and their context, maybe a pattern could be found.

For the overall evaluation, we found significant differences between some models, and no significant differences between others. In both experiments, we did not find a significant difference between CORPUS and BERT. We are curious about what can be found if we focus on the level of individual classifiers.

[25] With the knowledge we have after conducting the experiments, we are unsure if it was a good idea to include Clarity. We can imagine it is possible to be confused about the meaning of a sentence when a wrong measure word is selected, but especially in Experiment 2, using a distinct true classifier should not have caused noticeable issues with understanding the meaning of a phrase.

We would start by looking at one specific classifier: where does this classifier appear in the corpus? If we have found those places, we look at the classifier choices made by BERT: if BERT's classifier does not match with CORPUS, what classifier did BERT choose? We move from one classifier to the next; perhaps there is a pattern to certain classifiers that are confused often, or a pattern to contexts which tend to have more "flexibility" in classifier choice.

Lastly, we address some feedback about our analyses. We had assumed independence in our observations, which is reflected in the statistical tests we used. However, it could be argued that this assumption was not entirely justified; individual participants only judged part of the sentences, while simultaneously, individual sentences were only judged by part of the participants. This makes the data dependent; results could be affected by unaccounted-for differences between groups of participants, between groups of sentences, or both. If we were to account for this dependency in our data, we would have to perform some multilevel analysis. We could not do this due to time constraints, but if, at some point, we have the time, we could look further into analysing the data in this way.

8 Conclusions

In this thesis, we have subjected a number of algorithms that generate Mandarin classifiers to a new evaluation based on human judgement. This human-based evaluation is meant as an addition to the previous corpus-based approach. To achieve this new evaluation, we conducted two experiments in which we compared participants' judgement of classifiers chosen by various models.

For the first experiment, we created a dataset from 200 randomly chosen sentences from the CCD corpus (Appendix A). This was a standard human NLG evaluation, where participants evaluated classifiers generated by the models RULE and LSTM as described in [Peinelt et al., 2017], and BERT from [Järnfors, 2021].

For the second experiment, we created a dataset from 100 random classifiers and 100 infrequent classifiers from the CCD corpus (Appendix B). Participants evaluated only true classifiers generated by the models GE, RULE, and BERT.

Firstly, concerning the algorithms, we have learned that human and corpus-based evaluations showed similar results between models. When all classifiers are evaluated, BERT consistently scores higher than the other models. When looking at true classifiers only, BERT performs just as well with infrequent classifiers as with random classifiers, suggesting it can achieve good results with only little training data. Remaining within the context of true classifiers, RULE performs very well for such a simple algorithm; we did not find a difference comparing its results to those of BERT.

Notably, classifiers from the corpus (i.e. sentences produced by humans) were not evaluated higher than those generated by BERT under all circumstances. Within the context of this corpus, BERT showed a human-like performance.

Additionally, we can conclude something about evaluation: the similar results between CORPUS and BERT showed that an automatic evaluation based on accuracy does not exactly resemble how readers perceived the fluency of the sentences. If we want to explore how well a text will be perceived by human readers, it is a good idea not to rely on automatic evaluation metrics alone.

Lastly, given the corpus-based evaluation, we did not expect RULE to perform on par with BERT with true classifiers. Readers seemed to be more accepting of variations in classifier choice than the corpus-based evaluation made us believe.

References

[Allan, 1977] Allan, K. (1977). Classifiers. *Language*, 53(2):285–311.

[Belz et al., 2010] Belz, A., Kow, E., Viethen, J., and Gatt, A. (2010). Generating referring expressions in context: The grec task evaluation challenges. *Empirical methods in natural language generation: Data-oriented methods and empirical evaluation*, pages 294–327.

[Celikyilmaz et al., 2021] Celikyilmaz, A., Clark, E., and Gao, J. (2021). Evaluation of text generation: A survey. *arXiv eprint arXiv:2006.14799v2.*

[Chen, 2022] Chen, G. (2022). *Computational generation of Chinese noun phrases.* PhD thesis, Utrecht University.

[Cheng and Sybesma, 1998] Cheng, L. L.-S. and Sybesma, R. (1998). yi-wan tang, yi-ge Tang: Classifiers and massifiers. *Tsing-Hua Journal of Chinese Studies*, 28(3):385–412.

[Cheng and Sybesma, 2015] Cheng, L. L.-S. and Sybesma, R. (2015). Mandarin. In Kiss, T. and Alexiadou, A., editors, *Syntax - Theory and Analysis. volume 3*, pages 1518–1559. De Gruyter.

[Cohen, 1988] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Routledge.

[Da Costa et al., 2016] Da Costa, L. M., Bond, F., and Gao, H. H. (2016). Mapping and generating classifiers using an open Chinese ontology. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 249–256.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[Elliott and Keller, 2014] Elliott, D. and Keller, F. (2014). Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457.

[Erbaugh, 1986] Erbaugh, M. S. (1986). Taking stock: The development of Chinese noun classifiers historically and in young children. *Noun Classes and Categorization*, pages 399–436.

[Fellbaum, 2005] Fellbaum, C. (2005). WordNet and wordnets. In Brown, K., Anderson, A., Bauer, L., Berns, M., Miller, J., and Hirst, G., editors, *Encyclopedia of Language and Linguistics, Second Edition*, pages 665–670. Elsevier.

[Fritz et al., 2012] Fritz, C. O., Morris, P. E., and Richler, J. J. (2012). Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General*, 141(1):2.

[Gao, 2011] Gao, H. H. (2011). E-learning design for Chinese classifiers: Reclassification of nouns for a novel approach. In *International Conference on ICT in Teaching and Learning*, pages 186–199. Springer.

[Gatt and Krahmer, 2018] Gatt, A. and Krahmer, E. (2018). Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

[Guo and Zhong, 2005] Guo, H. and Zhong, H. (2005). Chinese classifier assignment using SVMs. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.*

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Howcroft et al., 2020] Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., Van Miltenburg, E., Santhanam, S., and Rieser, V. (2020). Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182.

[Järnfors, 2021] Järnfors, J. J. (2021). Computational models of classifier choice in Mandarin: from rule-based to BERT. Master's thesis, Utrecht University.

[Peinelt et al., 2017] Peinelt, N., Liakata, M., and Hsieh, S.-K. (2017). ClassifierGuesser: A context-based classifier prediction system for Chinese language learners. In *Proceedings of the IJCNLP 2017 (System Demonstrations)*, pages 41–44.

[Reiter and Belz, 2009] Reiter, E. and Belz, A. (2009). An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

[Reiter and Dale, 1997] Reiter, E. and Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

[Stent et al., 2005] Stent, A., Marge, M., and Singhai, M. (2005). Evaluating evaluation methods for generation in the presence of variation. In *CICLing*, volume 2005, pages 341–351. Springer.

[Van Deemter, 2016] Van Deemter, K. (2016). *Computational models of referring: a study in cognitive science*. MIT Press.

[Wen et al., 2012] Wen, H. M. S., Eshley, G. H., and Bond, F. (2012). Using WordNet to predict numeral classifiers in Chinese and Japanese. In *Proceedings of the 6th Global Wordnet Conference (GWC)*, pages 211–218.

[Zhang et al., 2008] Zhang, D., Li, M., Duan, N., Li, C.-H., and Zhou, M. (2008). Measure word generation for English-Chinese SMT systems. In *Proceedings of ACL-08: HLT*, pages 89–96.

[Zhang, 2007] Zhang, H. (2007). Numeral classifiers in Mandarin Chinese. *Journal of East Asian Linguistics*, 16(1):43–59.

[Zhang, 2013] Zhang, N. N. (2013). *Classifier structures in Mandarin Chinese*. De Gruyter Mouton.

Appendices

A Dataset Experiment 1

We used these sentences to create our dataset for Experiment 1. The column **Sentence** contains the sentence as it appears in the CCD. The location of the classifiers is indicated with <CL>. The column **Head noun** contains the head noun of the noun phrase the classifier belongs to. The column CORPUS contains the classifier as it appears in the CCD. The columns RULE, LSTM, and BERT contain the classifiers chosen by the respective models. This data can also be found here: https://github.com/AmberdeBruijn/Thesis-datasets.

Sentence	**Head noun**	CORPUS	RULE	LSTM	BERT
即使是错误那就让它错一 <CL> 时间先吧有点享受现在这样 HHHHHHHHHHHHHHHHHHHHHHHHH-HHHHHHH	时间	段	段	段	段
way to 绍兴三桥旁的商品房挺有特点的什么时候你设计一 <CL> 给哥们儿瞧瞧	瞧瞧	个	个	个	套
可怜的强哥, 不就是查 <CL> 房, 被学生喊说狼来啦!	房	个	间	个	个
第一 <CL> 发现, 原来脱离现实是很可怕, 很可悲的事。	发现	次	次	次	次
男士时尚假两 <CL> 毛衣欢迎大家光临本店, 本店物美价廉, 质量保证!	毛衣	件	件	件	件
第二 <CL> 生日礼物, 谢谢骨头的骨头	礼物	份	份	份	份
外婆总🄵叨不能吃 junk food 哈可是我們四 <CL> 总帶了滿滿袋回來, 从二樓躲到三樓偷偷干掉或许我真的🄵長大	总帶	个	个	个	个
我今天在微盘签到获得了 156 M 免费空间, 好运指数 3 <CL> 星, 你也来试试手气吧新浪网旗下云存储网盘!	星	颗	颗	颗	颗
HC 年会个人很喜欢来 G 的几 <CL> 照片。	照片	组	张	张	张
今天要和一 <CL> 酒神喝酒好恐怖啊	酒神	群	个	群	群
上传了 1 <CL> 照片到相册好友	照片	张	张	张	张
这几天买早饭是 <CL> 大问题啊	问题	个	个	个	个
咦, 洋子也算蹭得累嗎动漫新 <CL> MAD 傲娇的公主们来自	傲娇	番	种	番	番
是不是太委屈求全太卑微的爱一 <CL> 人他是不懂的珍惜你的呢	人	个	个	个	个
又睡醒了真心觉得这个年过的一 <CL> 意思也没有	意思	点	个	点	点
于是只好禁止自己再做分享这 <CL> 虚荣的事	事	种	件	种	种
喝了一 <CL> 啤酒, 还喝了一中白酒!	啤酒	瓶	瓶	瓶	瓶
昨天对着财神祷告过了, 也就 3 <CL> 愿望。	愿望	个	个	个	个
再见长春带着两 <CL> 冠军回家了福地	冠军	个	个	个	个
有 <CL> 家具的线条很有特色, 对于这样的家具, 要选用能与之颜色相配套的其他家具。	家具	些	款	件	种
死北佬老娘我看到一 <CL> 鞋好喜欢最大 39 , 我可悲的竟然穿小了。	鞋	双	双	双	双
其实, 我曾经是 <CL> 问题女孩, 表象上。	女孩	个	个	个	个
总有一 <CL> 人会对你说完这些话	人	个	个	个	个
一生要爱几 <CL> 人, 才能喊停。	人	个	个	个	个

去年的五月份, 高三学长学姐们最紧张的时候, 整个校园笼罩在一 <CL> 初夏, 忙碌的氛围中, 拍的不好。	氛围	种	种	片	片
但是要承认这 <CL> 比 LUCKY7 好看。	好看	部	个	张	张
你根本无法真正了解另外一 <CL> 人!	人	个	个	个	个
明天金谷仓, 卡门或者附近的某 <CL> 店, 安安静静的聊聊天如何?	店	个	家	家	家
haypi new 了 <CL> 爷啦发红包放鞭炮龙啸大红灯笼福到啦	爷啦	个	个	个	个
不懂得爱的人说爱情是 <CL> 玩物, 我希望它只给我带来快乐, 而不是痛苦。	玩物	件	个	个	个
15 分钟后飞奔回宿舍带上行李再飞奔车站, 回家了, 归心似箭, 各 <CL> 鸡冻啊!	鸡冻	种	个	种	种
EMS 的速度真的不敢恭维, 比几 <CL> 钱运费的 X 通 X 达的速度还要蜗牛很多很多。	钱	块	块	次	块
例如五年前, 我去云南宁蒗县永宁区托甸普米族村寨调查, 一老叟与我在火塘边交谈, 只见他一边谈一边拿一 <CL> 荆棘往他眼前挥舞。	荆棘	束	片	片	根
各 <CL> 憋屈。	憋屈	种	个	种	种
中国首 <CL> 儿童赛车主题社区, 带上朋友一起来玩库库马力吧!	儿童	款	名	次	家
真后悔晚上去超市没有拿那 <CL> 康师傅红烧牛肉面回来	牛肉面	桶	碗	条	碗
过了十一点, 每 <CL> 的鞭炮声让人都睡不着	鞭炮声	家	种	次	次
新年的爆竹很给力, 绽开对你的 <CL> 思绪, 缤纷的祝福洒进你的世界里春节的问候满满地, 包裹我沉甸甸的心意, 委托手机里的短信传递给你春节如意!	思绪	点	种	点	丝
老妹家翻出来一 <CL> 老照片, 啊年轻真好给老姐看一看你的成长蜕变记录吧	照片	张	张	张	张
凌晨 4 点 37 分, 雪停了, 四周一 <CL> 寂静。	寂静	片	片	片	片
硬要给我找 <CL> 模板, 奥斯尔吧!	模板	个	个	个	个
今天是大年二十九啦可是今年可能连一 <CL> 红包都没了	红包	个	个	个	个
早起去吃 XXX , 然后泥马连六 <CL> 小时的 lab 就这一学期拼了最后一得色了美图秀秀 iPhone 版	小时	个	个	个	个
人生中第一 <CL> 在電視劇🄵面遇到和我名字一模一樣的人耶!	遇到	次	次	次	次
一 <CL> 月, 可以干多少事啊!	月	个	个	个	个
起身冲 <CL> 热水凉好舒服	热水	个	杯	个	个
不过我承认爱情真他妈值不了几 <CL> 钱, 尽他妈花钱。	钱	个	块	个	个
全身上下各 <CL> 痛我们真的老了	痛	种	种	种	种
马上又到年三十了, 确找不到一 <CL> 快乐。	快乐	点	个	种	丝
上 <CL> 吃妈妈做的饭是啥时候不好意思, 我老妈不做饭, 都是老爸做的来自	饭是啥	次	次	次	次
出去逛 <CL> 街, 走起	街	个	条	个	个
这就是我们家四 <CL> 人的团圆饭。	人	个	个	个	口
曾经住在里面的, 都是传说中的神奇人物高官或者富贾, 以及他们的佳人, 和一 <CL> 段妙秘往事。	段妙秘	段	个	起	段

发表了一 <CL> 转载博文转载中国教育反思形式主义行大道	博文	篇	篇	篇	篇
奎贤啊, 听说和你的是 <CL> 中国女生, 会是谁呢?	女生	个	个	个	个
女人天生专情, 但她会对某一 <CL> 人绝情。	人	个	个	个	个
第一 <CL> 压岁钱成功厚脸皮	钱	笔	块	次	次
分享一 <CL> 来自虾米网的专辑不能说的秘密电影原声带分享自	专辑	张	张	张	张
漂亮女老师惨遭凌辱第一 <CL> 血 4 欲望号列车黑社会老爸我爱 HK 开心万岁男儿本色画壁断喉弩迷幻公园血色浪漫瑶山大剿匪瑶山大剿匪甜蜜间谍万有引力毒刺 Hello 小姐红磨坊金陵秘事恶夜惊魂甜蜜的冲动男儿本色黑暗终结者我人生的黄金期恶夜惊魂	血	滴	滴	滴	滴
手翘的就敲不到那 <CL> 键上。	键上	个	个	个	个
可以美美的泡 <CL> 澡好开心。	澡好	个	个	个	个
这两 <CL> 表情!	表情	个	个	个	个
向提问最后一 <CL> 了么?	了么	条	个	波	次
开车出门转转给自己买 <CL> 礼物	礼物	份	份	点	点
什么生活也是百感交集莫哀一是, 为什么反映在小说中却成了那么一 <CL> 简单的面孔, 譬如说: 喜剧式的。	面孔	副	张	张	张
对这 <CL> 年轻些的隔壁还不是那么讨厌蕾丝这机会没把握住确实坑爹。	隔壁	支	句	位	种
只有邹静帮我一 <CL> 对单子。	单子	张	张	张	张
期待大厅的新样貌, 同时感叹 SMN 竟然已经走过了 <CL> 春夏秋冬	秋冬	个	个	个	个
生命就像是一 <CL> 回声送出什么就收回什么你播种什么就收获什么给予什么就得到什么。	回声	种	种	种	种
呼深呼吸下明天很快就要来了还有那么十几 <CL> 小时啊祝我 lucky 吧	小时	个	个	个	个
联想乐 pad 就是 <CL> 垃圾货	货	个	个	个	个
五 <CL> 女人的豪尚豪, 吃完去风尚, 寂寞啊。	女人	个	个	个	个
这 <CL> 重逢的场面只会令大家不快乐。	场面	种	种	次	种
好感慨啊, 原来我也曾经这么疯狂而牛逼得玩过一 <CL> 游戏, 帝国琪儿仙亦恋凡尘葡萄剑影部落 209 狂血饮空圣少爷。	游戏	次	个	把	个
输了一 <CL> 该赢的球!	球	场	个	个	场
现在出场的是背号 56 的吾郎选手, 看他神采飞扬的进场看来这 <CL> 势在必得阿!	势	次	个	场	场
我刚刚填写了问卷星上的一 <CL> 问卷慈善该何去何从?	慈善	个	个	个	个
最近以及很长一 <CL> 时间都会用厂花和坤哥刷屏抱歉如果讨厌被刷请取消关注吧。	时间	段	段	段	段
网友调侃赵本山退出春晚大叔买到船很显然, 身体欠佳是赵本山春晚隐退的体面借口, 他还参加了湖南的, 辽宁的多 <CL> 春晚, 体力旺盛着呢。	春晚	个	届	个	场
向日葵昨晚做了 <CL> 乱七八糟的梦, 记不起情节, 但是早上起来眼睛肿了, 难道哭过?	梦	些	个	个	个

发表了博文摘如果有可能我带你去远行如果有可能我带你去远行躺在德德玛的草原数最亮的星如果有可能我带你去远行坐在外婆的沙滩看最白的帆影如果有可能我带你去远行爬上那 <CL> 山	山	座	座	座	座
如果哪天你喝醉了一 <CL> 人走在街头看着城市繁华灯红酒绿会歇斯底里的喊出谁的名字?	人	个	个	个	个
两 <CL> 醉汉驾着汽车狂奔。	醉汉	个	个	个	个
今晚三 <CL> 小祖宗光临我家, 天阿!	祖宗	个	辈	个	代
送小孩子一 <CL> 轩轩可爱了	轩轩	个	个	张	个
终于发现一 <CL> 昨天刚刚买的草莓照片。	照片	张	张	张	张
春晚二 <CL> 联排中, 吴秀波带妆彩排。	联排	次	次	次	次
, 我投给了李小冉这 2 <CL> 选项。	选项	个	个	个	个
上帝创造一 <CL> 男人让他去地球。	男人	个	个	个	个
这是一 <CL> 被赋予了传奇和灵性的宫殿, 里面封存着太多寂寞的亡灵。	宫殿	座	座	个	座
巧合的是这两 <CL> 比赛他都去了现场观看, 曼联现在脆弱的中场需要小贝, 鲁尼太孤独了。	比赛	场	场	场	场
都喜欢做 <CL> 力所不能及的事情, 让那么一些人那么尴尬	力	些	份	件	些
如果某一天醒来阳光里看到你那会是多么幸福的一 <CL> 事	事	件	件	件	件
有好长时间没有回家了当一 <CL> 人坐在一起吃饭的时候就算是神仙也会羡慕不已祝愿所有天下的父母老人孩子身体健康万事如意	人	家	个	个	家
1 月 11 日 0830 第五 <CL> 考试政治。	考试	场	次	场	次
尼玛光陪聊陪笑陪喝拿的小费就 TM 是我一 <CL> 月的工资!	月	个	个	个	个
可惜了, 美女老师被 OOXX 千 <CL> 激情片红楼梦请你原谅我十二生肖传奇说好不分手水浒茶馆断喉弩窈跳淑女李春天的春天断刺一起又看流星雨男人帮烈火红岩潜伏水浒十二生肖传奇断喉弩美人心计牵挂师傅蜗居菊子断喉弩	红楼梦	部	级	部	部
想说的话很多, 但总归还是一 <CL> 诞生日!	诞生日	句	个	点	句
未燃完的半 <CL> 烟!	烟	支	支	支	支
无锡天气 20120127 早今天星期五, 无锡白天阴, 夜间小雨, 气温 83 , 东风 34 <CL> 明天小雨, 气温 62 。	雨	级	场	级	级
18 有人在飞机上拿出方便面来干吃这时一 <CL> 美丽的空姐看到了过来对他说先生我给你泡	空姐	位	个	个	位
烤得金黄的蛋味酥皮红豆沙鸡蛋黄心, 有 <CL> 说不清的味道, 很好吃。	味道	种	种	种	种
我为啥觉得这是一 <CL> 很伤感的故事	故事	个	个	个	个
旁边一 <CL> 楼死人特么一大早吹到现在来威不威啊!	楼	栋	层	栋	栋
要再一 <CL> 认识到过程的温柔和宽容急不得。	宽容	次	颗	次	次
进入广西境内一 <CL> 迷蒙的雨雾旁边的山上都是云雾缭绕	迷蒙	片	片	个	片
后日早上单位游园会啊, 边 <CL> 想来讲声啦, 名额有限, 先到先得	想来	个	种	个	个
身体不舒服脚又还没好心情也不怎么样真想给自己一 <CL> 安眠药就这样去了	安眠药	瓶	颗	颗	颗

在家的时候觉得年年吃一样的团年饭没意思, 现在却很想念那 <CL> 熟悉的味道!	味道	种	种	种	种
第一 <CL> 极速光疗!	极速	次	个	次	次
4 腊月二十八 <CL> 面发, 腊月二十九蒸馒头。	面	把	个	把	把
他说, 马车走了整整一天才走进了那 <CL> 沟。	沟	条	个	条	条
第四 <CL> 双子座人财两失, 欲哭无泪!	双子座	名	名	名	名
三 <CL> 吃货在床上吃张君雅小妹妹!	吃货	个	个	个	个
分享一 <CL> 活动给大家 3000 元的红酒转到就送!	活动	个	个	个	个
为啥子刚刚进了你空间有一 <CL> 怅然若失的感觉	怅然若失	种	种	种	种
两 <CL> 妈妈做的丰盛的晚餐	妈妈	位	个	桌	位
每天七点起是几 <CL> 意思黑眼圈下不去	黑眼圈	个	个	个	个
现在对过年一 <CL> 感觉都没有	感觉	点	种	点	点
老样子, 挂好号后在对面良鹰吃一 <CL> 辣酱面。	辣酱面	碗	个	个	碗
上传了 16 <CL> 照片到相册哈哈	照片	张	张	张	张
现在就开始各 <CL> 难受想睡觉了啊。	难受	种	种	种	种
大清早起床给哥们当伴郎那 <CL> 累呀!	累	个	种	个	个
过去的一年, 得到了什么, 失去了什么, 都不再重要, 关键要有一 <CL> 新的开始。	开始	个	个	个	个
求助我遇到了问题, 谁来帮帮我一 <CL> 国际象棋的问题来自百度知道	象棋	个	款	个	个
有一 <CL> 宽容是因为爱!	宽容	种	颗	种	种
试问, 我什么时候不是 <CL> 女人?	女人	个	个	个	个
两 <CL> 人在酒店的最新娱乐方式。	人	个	个	个	个
开业一周内即 16 日 22 日, 16 <CL> 经典影片带您体验!	影片	部	部	部	部
FM 950 广西音乐广播松总监别老是提我们家蛋叔啊你唱首 <CL> 截棍吧	截棍吧	双	个	次	个
这个航班延误的够水平整整 12 <CL> 小时看凌晨 2 点 30 分美兰机场的客人困	小时	个	个	个	个
奇石岛又完成了一 <CL> 市民的心愿。	心愿	个	个	个	个
给 <CL> 机会行不!	机会	个	个	个	个
电脑族易得 6 <CL> 疾病头痛。	疾病	种	种	种	种
新的一天新的开始, 有些人有些事有 <CL> 话, 永远会记住你不要后悔我也忍够了。	话	些	句	些	些
解说员说那 <CL> 握手是拉莫斯自认为得牌没有问题, 搞笑。	握手	个	个	个	个
我买条短裤了啊, 姐这 <CL> 带的全是西装。	带	次	次	套	次
收到快递一定要当快递员面验货再签收, 垃圾的是 <CL> 专业盗窃集团啊, 有木有请看 http t cn Sapppc 韵达快运的盗窃业绩。	集团	个	个	个	个
一 <CL> 人开开后备箱, 一滩血出来。	人	个	个	个	个
剑侠世界我在电信朱雀区采石矶江津村, 这里是五千万人的江湖, 萝莉御姐任你挑选, 抱 <CL> 妞儿回家过冬咯!	妞儿	个	个	个	个
当时那 <CL> iPhone 同样处于设备验证测试阶段。	iPhone	款	个	个	台
大肠杆菌在消化这些未分解物后就会排气, 这些气体在体内累积, 造成一 <CL> 气压。	气压	股	级	种	种
图中的差别仅仅是 <CL> 开始, 你想象中的 XXOO 是怎样的?	开始	个	个	个	个

能开 <CL> 机么陈雀喜	陈雀喜	个	个	个	个
对生活的一 <CL> 小小建议美国畅销书	建议	点	个	个	点
哎, 这三 <CL> 必须缺一不可, 真的有点让我崩溃!	必须	个	次	个	个
我参与了发起的投票芒果台捧红的超人气明星, 你最喜欢谁李宇春张杰刘忻, 我投给了李斯丹妮 11 快乐女声这 1 <CL> 选项。	选项	个	个	个	个
我是 <CL> 宅男, 现住在南京郊区, 除了写作我喜欢走走路运动。	宅男	个	个	个	个
剑侠世界我在电信金麟区雪芳草巴陵县, 这里是五千万人的江湖, 萝莉御姐任你挑选, 抱 <CL> 妞儿回家过冬咯!	妞儿	个	个	个	个
一 <CL> 小事也要闹, 闹大了就好了?	小事	点	件	点	点
like 这 <CL> 感觉!	感觉	种	种	种	种
上新, 这 <CL> 新品很不错, 分享一下美国红牌澳洲美利奴羊毛修身妩媚羊毛連衣裙, 价格 36000 元, 购买链接, 更多精彩宝贝请看	新品	款	款	款	款
对于一 <CL> 不懂也不想结婚的人来说!	人	个	个	个	个
这 <CL> 葡萄酒干掉一瓶是没问题的很好喝可是后劲很大所以我开始头晕了?	葡萄酒	种	瓶	瓶	瓶
今晚人气吴差除了娥是一 <CL> 人以外	人	个	个	个	个
你可有发现 beyond 的所有专辑中没有一 <CL> 关于爱情的歌曲。	歌曲	首	首	首	首
博山的颜文姜雕塑, 有 <CL> 美丽的传说!	传说	个	个	个	个
昨晚躺在床上饼向我表白了一 <CL> 匪夷所思的事, 那时还没认识他时, 我去他们高数一插班听课, 我一直认为是我先和他说话借他的课件来着原来不是!	事	件	件	件	件
逆战不错喔, 有点好莱坞的感觉活脱脱一 <CL> 使命召唤真人版, 推荐	使命	个	个	个	部
黏嗒嗒 <CL> 小男人	男人	滴	个	滴	滴
请别质疑我尽管我有多坏也不会至于到这 <CL> 地步, 你懂的。	地步	种	种	种	种
本 <CL> 工人阶级假名社会主义中国的领导阶级经济学定义低收入阶层洋名蓝领别名体力劳动者昵称弱势群体外号蚁族社会学定义生存性生活者政治学定义社会不稳定因素经常性称呼失业者政府给的名字下岗工人民政定义低保户真名穷人	工人	名	个	位	名
Love u 下 <CL> 还要跟你出来渲泄情绪, 现在又充满爱的力量了!	要	次	次	次	次
如果你没有走过那一 <CL> 路程, 怎么能抵达到现在?	路程	段	段	段	段
給力啊, 一個帥哥接一個帥哥輪 <CL> 🄵服我, 然後我告訴你他們失敗了你居然要指派最帥的哥們來我家樓下等我, 知道我不忍心麼。	🄵服	番	个	次	次
还是说互联网就是一 <CL> 政治家的工具?	工具	群	个	个	个
婚礼跟拍的写实风格和写意风格, 我无意说哪 <CL> 风格更好。	风格	种	种	种	种
结尾的一 <CL> 我知道你的心脏异于常人, 是长在右边的让我欢快地睡觉去了。	心脏	句	颗	句	句
不过第八 <CL> 最后的扑到真是!	扑到	集	个	个	次

, 这一带有很多疗养院, 另一 <CL> 部队医院是浙江医院, 也在这里。	部队	个	支	家	家
我参与了发起的投票 2012 新浪微博全球最具人气娱乐明星终极票选, 我投给了炎亚纶这 1 <CL> 选项。	选项	个	个	个	个
春晚三十年了, 我也凑 <CL> 热闹, 今天 1 月 23 日我生日, 三十岁生日, 大家祝我生日快乐吧!	热闹	个	个	个	个
开 <CL> 包, 来来来, 大家一起吃, 一起早生贵子!	包	个	个	个	个
这 <CL> 寒假感觉自己又长大了点, 懂得主动和妈妈一起做家务, 一般做完事才会休息, 特别是切菜也感觉更熟练了, 哈哈	寒假	次	个	次	次
4 <CL> 美臀法则久坐也拥有翘臀 1 面向下俯卧, 头部轻松地放在交叉的双臂上 2 缓缓吸气, 同时抬起右腿, 在最高处暂停数秒, 然后边吐气边缓缓放下 3 在胎腿时需注意足尖下压, 并且臀部不能离地。	久坐	个	种	条	条
祝自己做 <CL> 好梦吧!	梦	个	个	个	个
京东商城网购晒单我买了 <CL> Sharon 雪伦豹纹薄纱带帽中长款女士羽绒服 6152842 A 咖啡 XL 175 感觉不错哦优点质量不错, 手感不错。	薄纱	个	层	个	个
突然间发现自己好象不配拥有这么好的一 <CL> 你	你	个	个	个	个
对说完美神起翻唱, 一 <CL> 人搞定所有和声很强大的男生	人	个	个	个	个
不知道恩慈能不能撑过七天生 <CL> 水瓶座宝宝	宝宝	个	个	对	个
, 在一 <CL> 超酷的盒子里!	超酷	个	款	个	个
真是没 <CL> 火啊	火	个	把	个	个
这两 <CL> 书不错, 希望好运, 呵呵!	书	本	本	本	本
年年不中奖, 只能各 <CL> 哼切呸	哼切呸	种	个	种	种
大年初二的早晨, 来一 <CL> 美味的心形糕点吧, 祝大家新年事业顺爱情顺事事顺!	美味	块	种	份	份
你儿媳妇会握手, 会上厕所, 厉害不我们才 2 <CL> 月大	月	个	个	个	个
四 <CL> 轮的车我们的宝牛	轮	个	个	个	个
烟可以不懂手的寂寞, 酒可以不懂喉的寄托, 泪可以不懂眼的脆弱, 不是每 <CL> 人都一定会快乐。	人	个	个	个	个
今晚熬 <CL> 小夜也无妨画了张小速写。	夜	个	个	个	个
都射了 6 <CL> 了, 这个女的还要 http t cn Sadr4V 雪豹追影恋人絮语我在伊朗长大野鸭子告密者能人冯天贵鸡犬不宁重生父亲大国之酿夜店倩女幽魂单身女王雪豹夜宴鬼域追影断刺神话潜伏中国远征军万有引力单身女王菊子	女	次	次	次	次
爷爷应该感到幸福, 四 <CL> 哥哥在服侍着!	哥哥	个	个	个	个
当一 <CL> 女人对你失望的时候, 即使最厉害的魔术师也变不回她的心!	女人	个	个	个	个
不知道为什么总是想起那 <CL> 话她难过所以我过去陪她可是你忘了更需要陪的我	话	句	句	句	句
每天一 <CL> 打底裤	打	条	次	条	条
我在你门口在你心门口徘徊徘徊着你心门紧闭我一直都进不去我在寻找寻找一 <CL> 开启你心门的钥匙又或者在等待等待你来为我开门	钥匙	把	把	声	把
第一 <CL> 吃驴肉, 我还不忍心吃的!	吃	次	次	次	次

巨蟹座天蝎座两<CL>适应力极强的星座在一起, 会很快适应对方的喜好, 好好相处, 白头到老。	适应力	个	个	个	个
有人用冲击波形容团购从网络向实体蔓延给国内既有商业模式造成的影响, 还有人将这<CL>现象归功于团购网购的种种先天优势。	现象	种	种	种	种
見識了比我还high还能唱地麥神被咬地還是心甘情願地下次繼續吧唯一遺憾地就是奇bi的新发型没有来<CL>特写。	特写	个	个	个	个
完全没有那<CL>实感。	实感	种	个	个	种

B Dataset Experiment 2

B.1 Random classifiers

We used these sentences to create our random dataset for Experiment 2. The column **Sentence** contains the sentence as it appears in the CCD. The location of the classifiers is indicated with <CL>. The column **Head noun** contains the head noun of the noun phrase the classifier belongs to. The column `CORPUS` contains the classifier as it appears in the CCD. The columns `GE`, `RULE`, and `BERT` contain the classifiers chosen by the respective models. This data can also be found here: https://github.com/AmberdeBruijn/Thesis-datasets.

Sentence	**Head noun**	CORPUS	GE	RULE	BERT
分享图片这是我爱拍照片的爸爸, 和一<CL>抱这金猪睡觉的妹妹!	妹妹	个	个	个	个
哈哈跳舞之前聚<CL>餐。	餐	个	个	个	个
爸妈在姑姑的委任下, 在圆满完成三<CL>月的陪读任务后, 带着表弟今天到家了。	月	个	个	个	个
活着很好, 嗯, 再吃一<CL>苹果。	苹果	个	个	个	个
嘉兴质监给大家拜<CL>早年, 祝大家龙年大吉, 新春快乐!	早年	个	个	个	个
林起之前微博胃痛痛到死左<CL>女仔, 不寒而栗	女仔	个	个	个	个
我从春节前两<CL>星期开始吃到现在胖了六斤继续!	星期	个	个	个	个
烧<CL>香这么不容易, 拔山涉水么?	香	个	个	个	个
我相信你绝对是因为找不到多人图了才勉为其难转了这<CL>图的对不对!	图	张	个	张	张
分享图片做了4<CL>不同速度的gif , 说快的比较好, 哈哈, 我觉得快的动起来好喜感呀女装版的锦渊, 娇柔却不乏英气, 女帝	速度	个	个	块	个
已经提前一<CL>多小时出来拉, 不会又迟到吧!	小时	个	个	个	个
求上海每<CL>区的中心位置可开银行的商铺。	区	个	个	个	个
对说昨天看了你主演的电视电影美人三嫁, 演绎的很到位, 塑造了一<CL>外表冰冷但内心善良的坚强女孩。	女孩	个	个	个	个
2011新款秋冬高腰裤休闲裤大码女裤子韩版显瘦小脚裤加厚铅笔裤售价8800元最近售出3085<CL>淘宝地址相关商品	地址	件	个	件	件
是<CL>冰山攻。	攻	个	个	个	个
http t cn z0kNWMT我可以不可跟我妈说我想买<CL>音箱回来学习模电。	音箱	个	个	个	个

大半夜一 <CL> 人都能咯咯咯地笑出声来, 因为此刻的我想起了某人傻笑的样子。	人	个	个	个	个
姐姐从英国回来越发漂亮自信, 而我却还是 <CL> 卑微的丑小鸭。	鸭	个	个	只	个
看人人, 看到一 <CL> 谁 TM 能这么爱我的日志。	日志	个	个	篇	个
相识就是缘, 朋友们牢牢铭记, 凡尘过后终了无牵过, 一 <CL> 今生缘, 让我们铭记, 屏住呼吸, 静静聆听川子今生缘	今生缘	首	个	个	句
感情没有底线, 人格尊严却是有界限, 不算是等价交换, 这 <CL> 生意一辈子我都不会做。	生意	笔	个	笔	种
阳光明媚, 心情大好, 今儿是 <CL> 好日子!	日子	个	个	个	个
年集镇 2 月房价环比涨幅折半调控效劳显示本报讯文表记者张忠安有数据显示, 全国 100 <CL> 城市房价环比涨幅放缓, 2 月份仅上涨 048 , 十大主要城市房价涨幅也几乎减半广州楼市更是出现内冷外热局面。	涨幅	个	个	家	个
有申请普度的童鞋们来吱 <CL> 声啊	声	个	个	个	个
他在序里说, 中共中央组织部把雷锋、焦裕禄、王进喜、史来贺跟钱学森这 5 <CL> 人作为解放 40 年来在群众中享有崇高威望的共产党员的优秀代表。	人	个	个	个	个
终于又到达一 <CL> 平衡的状态!	状态	个	个	种	个
, 我投给了杜荷这 1 <CL> 选项。	选项	个	个	个	个
亲爱的们, 要的太多了, 一两 <CL> 微博搞不定啊!	微博	条	个	条	个
淘宝客服很亲切的说, 亲, 我们放假了哦我们不保证年前发货哦你可以等我们年后上班再来买哦亲, 完了还加 <CL> 笑脸你妹的!	你妹	个	个	个	个
大年初一 <CL> 一天啊龙年大吉阿祝	大吉	头	个	个	头
真的是 <CL> 太监	太监	个	个	个	个
从几天前妈妈就开始在盘算给我寄 <CL> 多大的包裹到上海去呢, 她恨不得能把家里所有好吃的都装到包裹里。	包裹	个	个	个	个
上传了 1 <CL> 照片到相册墨西哥	照片	张	个	张	张
春运, hold 住刚刚有 <CL> 哥们订票回家, 我瞅了一眼, 他订广州曼谷西双版纳的机票!	哥们	位	个	个	个
好想给自己添双三叶草, 最近没空跑市区阿, 快点打折打折打折打折打折打折打折打折打折打折打折打折打折, 雪地靴下 <CL> 月去那边买, 一定便宜极了	月	个	个	个	个
洪老师对我的美好祝愿金石为开, 真情所至, 遇的佳偶, 好运多多人生的第一 <CL> 藏头诗啊!	诗	首	个	首	首
周三多只软件板块个股涨停, 成为震荡市一 <CL> 明显亮点。	亮点	个	个	个	个
三 <CL> 下铺太神奇了!	下铺	张	个	个	个
另一 <CL> 小巧的音箱设计是迷你音箱, 相比之下这个耳机更能搏得女生的欢心。	小巧	个	个	款	个
下期变形记的主角貌似比易虎臣更叛逆, 还是 <CL> 女生	女生	个	个	个	个
人生总会充满不尽人意, 每 <CL> 人都会有自己的故事, 到底其中有几多人会知, 诉说依然是无药可医。	人	个	个	个	个
RCECI 原创设计谁的衣柜里还没有一 <CL> 牛仔布衣物?	衣物	件	个	件	件

对说 j 今天拿着满满的东西去看望外婆, 透过窗子看着她一 <CL> 人静静的似乎在想什么, 我觉得很酸楚, 同时也愧疚自己给她的爱太少。	人	个	个	个	个
那 <CL> 酒喝多的人准备罚款吧, 哈哈哈	酒	个	个	杯	个
鲭鱼偶像我竟然没有发现有哪 <CL> 字幕组在做	字幕	个	个	个	个
发表了一 <CL> 转载博文转载测测您的脑年龄	博文	篇	个	篇	篇
骂几 <CL> 我这人脸皮比较厚。	脸皮	声	个	张	句
4 早晚各 30 <CL> 仰卧起坐, 最好结合跑步或者游泳等有氧运动。	仰卧	个	个	个	个
热卖厚底鞋高跟短靴坡跟亚历山大王鞋女靴马丁靴松糕尖头女鞋售价 17800 元最近销量 686 <CL> 地址相关	地址	件	个	件	件
这是我家两 <CL> 可爱的宝贝。	宝贝	个	个	件	个
困得东倒西歪到家, 伸了 <CL> 懒腰。	懒腰	个	个	个	个
人生难道就是一 <CL> 不断退而求其次的过程么?	程么	个	个	个	个
如果不是, 那就很悲哀, 作为一 <CL> 内地名校名师, 竟然能说出某某地方上的人是狗这样的没素质的话来, 我真替国人感到悲哀	名校	位	个	类	个
如果你的领导不喜欢你, 如果这是人生中必须要面对的一 <CL> 题, 我愿意现在开始尽力解决, 拒绝抱怨。	题	道	个	道	个
都没几 <CL> 人找我。	人	个	个	个	个
祝福中国 2012 凡事要尽心捧着一 <CL> 心来不带半根草走!	心	颗	个	颗	颗
找 <CL> 帅哥美女跟我放鞭炮	美女	个	个	个	个
是因为今年有多 <CL> 黑色星期五吗?	星期五	个	个	个	个
今年我去拜年一 <CL> 红包都没有拿到不爽	红包	个	个	个	个
爱一 <CL> 人, 就是在漫长的时光里和他一起成长, 在人生最后的岁月一同凋零。	人	个	个	个	个
一 <CL> 人在宿舍, 可以不客气的打开音响定时播放, 放着阿妹的歌。	人	个	个	个	个
是的这么多年没变的就是难忘今宵我那 <CL> 内牛满面	满面	个	个	个	个
因为某 <CL> 方向同学会一直关注望着你。	方向	个	个	个	个
假如韩寒不能证明, 起码在很大一部分人, 比如我, 的意识中, 会觉得谁读韩寒, 起码是一 <CL> 比较幼稚的人。	人	个	个	个	个
附赠一 <CL> 相片	相片	张	个	张	张
图为一 <CL> 僧人从广场走向天安门城楼。	僧人	位	个	个	位
这个年过的不太平, 婆媳问题, 永远的难题, 两 <CL> 人都不怎么懂事, 我也搞不定了!	人	个	个	个	个
对说回家给你 <CL> 惊喜	惊喜	个	个	个	个
今天陪家人逛花市自己捞了一 <CL> 风车, 转运转运。	风车	个	个	个	个
发表了一 <CL> 转载博文转载北京西山汉德法官在转移视线扭转方向	博文	篇	个	篇	篇
对待一 <CL> 毫无过错的孩纸, 太残忍了!	孩纸	个	个	个	个
盘点女孩最容易失初吻的十 <CL> 地方	地方	个	个	个	个

各国年夜饭也讲老礼虽然每 <CL> 国家有各自风俗习惯, 但准备丰盛年夜饭的习惯大致相当, 新旧日子的交替, 吃上丰足的食物预示吃掉过去的不快来年大丰收。	国家	个	个	个	个
天生就有好运的星女第一 <CL> 射手座。	射手座	名	个	名	名
约时间碰 <CL> 面吧, 不然一年到头没也多少机会	面	个	个	个	个
终于回到家了 2300 华农 130 江门开平, 220 阳江, 320 茂名, 420 湛江, 510 雷州, 600 徐闻全程不跑高速哥走了三分之二 <CL> 广东省 660 公里才能回到家家远的孩子真的伤不起	孩子	个	个	个	趟
黄云珍 8894 又完成了一 <CL> 市民的心愿。	心愿	个	个	个	个
从现在开始, 我的生活只有两 <CL> 事可做学习!	事	件	个	件	件
中午起来的时候雪已经被铲开了少 <CL> 体力活	活	件	个	个	个
这不是一 <CL> 女人该有的行为啊!	女人	个	个	个	个
貌似账号被盗了突然间冒出来 200 多 <CL> 关注	关注	个	个	个	个
快去看吧, 你家霆锋不会让你失望的, 周董也是 <CL> 惊喜哦!	惊喜	个	个	个	个
因果关系, 每 <CL> 选择都有其后果。	选择	个	个	个	种
你没衣服穿是吗没就买呗你就那么喜欢穿人家衣服吗真的很讨厌人家就剩下那两 <CL> 衣服还要拿去真的想骂人啊啊啊啊啊啊啊啊啊啊啊	衣服	件	个	件	件
律师是诡辩家, 不会诡辩就难成律师, 我以前的一 <CL> 同事, 因为能诡辩, 后来当上了律师, 而且还生意兴隆呐!	同事	个	个	个	个
我想听这 <CL> 专辑輕快的生活以莉高露	专辑	张	个	张	张
年前和霆哥去剪 <CL> 头发吧?	头发	个	个	个	个
现在只剩下我一 <CL> 学生在画室了	学生	个	个	个	个
当面对两 <CL> 选择时, 抛硬币总能奏效, 并不是因为它总能给出对的答案, 而是在你把它抛向空中的那一刹那, 你突然知道你希望它是什么。	选择	个	个	个	个
永远都是最快乐的那 <CL> 大头!	头	个	个	个	个
分享了一 <CL> 文章为什么他们总忽悠你憎恨美国	文章	篇	个	篇	篇
你不成为一 <CL> 白痴的原因是什么?	白痴	个	个	个	个
那 <CL> 你说越长越的同学今儿又被姐遇上了	同学	个	个	个	个
缝了一下午的哆啦 A 梦扣子被我弄坏了要拆掉重新换新的扣子一 <CL> 下午的杰作都木有了	杰作	个	个	个	个
原来我也有过在某 <CL> 网游里给别人 100w 喊人家去买装备的豪迈事迹年轻的岁月真是不复返啊	网游	个	个	个	个
一 <CL> 楚楚怜人的女孩子, 纵然不是美女才女, 做事情一塌糊涂, 但是能勾起男人的保护欲。	楚楚	个	个	个	个
原来我还是 <CL> 文艺青年呀!	青年	个	个	个	个
, 我投给了智能手機这 1 <CL> 选项。	选项	个	个	个	个
哈哈, 电视里男人总说的一 <CL> 话就是我现在正在事业的上升期, 你自己看着办吧	话	句	个	句	句
图中的差别仅仅是 <CL> 开始, 你想象中的 XXOO 是怎样的?	开始	个	个	个	个

B.2 Infrequent classifiers

We used these sentences to create our infrequent dataset for Experiment 2. The column **Sentence** contains the sentence as it appears in the CCD. The location of the classifiers is indicated with <CL>. The column **Head noun** contains the head noun of the noun phrase the classifier belongs to. The column `CORPUS` contains the classifier as it appears in the CCD. The columns `GE`, `RULE`, and `BERT` contain the classifiers chosen by the respective models. This data can also be found here: https://github.com/AmberdeBruijn/Thesis-datasets.

Sentence	**Head noun**	`CORPUS`	`GE`	`RULE`	`BERT`
有钱人简直伤不起出去买车大众的某 <CL> 然后看到车说哎呀这个车的内置杂和家里的宾利一样!	内置杂	辆	个	个	款
节日一如平日, 只是少了一点点压力, 少了一点点紧迫, 但是谁都知道, 一旦节日的喜庆褪去颜色, 🄵个人都要背着 N <CL> 大山前行可是, 是否🄵个人都知道前进和生存的意义呢	大山	座	个	座	座
独自行走于醉人的月色下, 体会着清风的律动, 那模糊的旋律, 好似正在演奏着一 <CL> 梦幻般的风月, 不胜的伤感而又优美。	梦幻般	曲	个	层	场
992 城区路况赣州京九路, 红辣椒酒店附近, 2 <CL> 小车相挂。	车	辆	个	辆	辆
穿一 <CL> 素色白裙, 走在人间四月, 等待一树又一树的花开。	素色	袭	个	个	件
摇钱树看到这 <CL> 下钱雨的摇钱树, 转发越快, 钱来的越快。	树	棵	个	棵	棵
增高鞋可调 3 高弹力内增高鞋女式男式硅胶男士 5cm 增高垫购买 600 最近成交 3030 <CL> 上海健足乐旗舰店	旗舰店	笔	个	笔	笔
她的面前, 是一 <CL> 黑红色的裸体。	裸体	架	个	个	副
法新社报道, 农民联合会和义勇割除队的这些成员随后在散落一地的玉米堆上拉开了一 <CL> 横幅, 上面写着基因改造地带。	横幅	面	个	条	条
这 <CL> 话特想告诉一人, 现在还不能跟此人翻脸, 我受教育好, 从小知道什么叫识时务, 小太爷来日跟你丫玩儿命!	话	席	个	句	句
晒晒很精致的一 <CL> 台历, 制作精良的插页, 时刻提醒我重要特殊的日子, 感谢, 感谢小编,	台历	本	个	本	本
更令人无法理解的是, 这 <CL> 飞机看来就像它失踪时一样簇新。	飞机	架	个	架	架
小镇上的第一 <CL> 商品房, 平均不到 20w 一套, 开卖就销售一空, 现在二期三期都跟上来, 好牛好强大。	商品房	处	个	个	套
刚才坐电梯时, 发现我兜兜里有 <CL> 大白兔!	白兔	颗	个	只	个
好歹同学两年半, 本不想说什么, 妈的最后一 <CL> 烟也偷	烟	根	个	支	根
演唱会于晚上七点五十正式拉开帷幕, 一 <CL> 火红装束亮相的王力宏瞬间让现场歌迷的尖叫声响彻整个温州的夜空。	火红	袭	个	部	身
辞旧迎新之际, 派发一 <CL> 鞭炮丁俊晖, 你喝蒙了吗?	鞭炮	枚	个	个	堆

写这 <CL> 作品前, 我做了祷告, 其中路字右边的各字很细, 出乎我的意料, 但一想你们要行窄路, 便恍然大悟。	作品	幅	个	部	篇
不识字的老婆给老公写的一 <CL> 信, 你能看懂吗?	信	封	个	封	封
翰创分享设计师 Cho Hyung Suk 从自行车的链条获得灵感, 设计了这 <CL> 链条台灯 The B Chain Lamp 。	链条	盏	个	个	款
刚才打你电话没通, 提示说你拨打的用户是猪, 过了会再打, 居然提示你拨打的那 <CL> 猪已经屠宰。	猪	头	个	只	头
一 <CL> 要 1290 , 好想買喔, 可是🄵錢, 等過年時再🄵吧, 感覺好棒, 我喜歡, 能讓臉作深層的潔🄵, 比什🄵都好, 因🄵我平常不用保養品, 相對的, 臉務必要徹底清潔到底好想買	要	枝	个	次	件
有一 <CL> 牙签走在路上, 突然走累了。	牙签	根	个	根	根
初五停电了, 我们这儿 <CL> 因为变电站家里忽然停电了, 我在蜡烛昏暗下吃完了晚饭, 不寻常的一天。	停电	幢	个	次	次
这时间睡醒了哪能办想到 6 点半要起床上班我就像挨一 <CL> 闷棍	闷棍	记	个	个	根
小区里面隔壁那 <CL> 房子着火了, 火势好大, 祈祷祈祷, 这大过年的, 放烟火真的要注意呀	房子	幢	个	个	个
我在这里东门步行街西华宫西南自己跑出来老街吃丸仔这 <CL> 店的好多年了	店	间	个	家	家
白蛇回到紫竹林, 却发现清风洞里躺着一 <CL> 尸体, 那便是许仙。	尸体	具	个	具	具
一 <CL> 奇葩昨日我逛温州大世界超市, 哇噻!	奇葩	朵	个	朵	朵
你不喜欢人家就不要同住一 <CL> 房间!	房间	间	个	个	个
谁家的酒在楼梯间打破了, 整 <CL> 楼弥漫着酒香。	楼	栋	个	层	栋
第三 <CL> : 魔羯座。	魔羯座	名	个	名	名
二十多 <CL> 窗户, 擦到姐手软 T T	窗户	扇	个	扇	个
webber , 哥昨晚梦到和你一 <CL> 教室在那做不知道是什么科目的题。	教室	间	个	级	个
今夜, 我想舒展红袖, 鸣一 <CL> 洞箫, 借蒹葭苍苍的方向, 轻敲你灵魂横渡的心窗, 走进你内心的殿堂。	洞箫	管	个	个	声
坦然我的忧伤和甜蜜似一 <CL> 温婉的歌赋缀入心间看到迁徙鸟的博文傲梅的春日原创现代诗卉有感而发的评论。	温婉	曲	个	份	首
而这里更是有 19 <CL> 形态各异的瀑布, 其中最有吸引力的当属彩虹瀑布。	形态	处	个	种	个
10 年后, 她生前衣物慈善义卖, 87 岁的他拄着拐棍, 买回那 <CL> 胸针, 不久与世长辞。	胸针	枚	个	个	支
老师一 <CL> 肺腑之言我要感谢孩子学习不用愁特聪明, 英语要勤练, 好规矩要养成。	肺腑之言	席	个	个	句
买 <CL> 仙人掌给你	仙人	棵	个	个	个
供着一 <CL> 不爱她的男神, 唉, 人生啊!	神	尊	个	个	个
初一第一 <CL> 香	香	株	个	个	支
我的第一 <CL> 学校的	学校	所	个	个	个
你关闭了别人一道门, 同时也就关闭自己的一 <CL> 门张邵刚是最好的例子!	门	扇	个	个	扇

新闻摄影是以附有简短文字说明的新闻照片形式同读者见面的, 它不同于电影和电视上的活动形象, 而是以静止的形象, 即将新闻自身的形象瞬间定格在一 <CL> 画面上。	画面	幅	个	个	个
吴清源围棋全集第四 <CL> 中盘战术死活和收官	死活	卷	个	种	集
我看不到这 <CL> 城市的明天, 我很怕自己最终是温水里的青蛙。	城市	座	个	个	座
今天发了薪水, 还了贷款, 交了房租水电煤气费, 买了油米和泡面, 摸摸口袋剩下的钱, 感叹一 <CL> 这月工资又白领了!	工资	声	个	份	声
那天恰巧有几 <CL> 坐着妓女的轿子, 朱良才混在抬轿人中逃离虎口。	轿子	顶	个	个	辆
今早发现单位卫生间里竟然放了一整 <CL> 的卫生纸。	卫生纸	卷	个	卷	包
学习时两 <CL> 灯同时照明。	灯	盏	个	盏	盏
去图书馆寻找一 <CL> 书籍。	书籍	册	个	本	本
第一 <CL> 官方认可的满分杆 147 戴维斯 VS 斯宾塞, 确实激动人心	满分杆	杆	个	个	个
2012 年农历春节之际, 中国海军第四 <CL> 071 级综合船坞登陆舰在上海沪东造船厂下水。	船坞	艘	个	个	艘
但如果我們始終深信不疑, 有一扇門就會向我們打開, 它或許不是我們曾經想到的那扇門, 但我們始終會發現, 它是一 <CL> 有益的門。	門	扇	个	个	扇
国家扶贫开发重点县山西兴县, 2010 年盖起了一 <CL> 廉租房, 但建而不分闲置至今。	房	栋	个	间	栋
您有一 <CL> 从现在移植到未来的财产吗?	财产	笔	个	个	笔
真闹心啊一抽血就抽四 <CL> 血	血	管	个	滴	次
年前的聚会太多, 趁我还能清醒之时, 真心道一 <CL> 众亲晚安!	众亲	声	个	个	声
2011 年, 中国巡逻艇剪断了一 <CL> 越南调查船的电缆。	船	艘	个	艘	艘
而你是深处倏然而过的一 <CL> 魅影, 明明灭灭在我的生命里生长, 绽放, 最后一转眼就消失不见。	魅影	道	个	部	道
给我一 <CL> 骏马我就可以驰骋	骏马	匹	个	首	匹
还骑了 <CL> 马去欧尚!	马	匹	个	匹	个
谁回眸时不经意的惊鸿一瞥, 牵动着一 <CL> 心到支离破碎。	心到	颗	个	个	颗
心理提示四 <CL> 1 活鱼会逆流而上, 死鱼才会随波逐流 2 逆风的方向更适合飞翔 3 走在光滑的冰面上容易摔倒, 是因为上面没有坎坷 4 环境永远不会十全十美, 消极的人受环境控制, 积极的人却控制环境。	方向	则	个	个	条
受不了, 因为一 <CL> 头发, 一回家, 就被人说沧桑, 老了。	头发	顶	个	个	根
泛有绸缎光泽的 <CL> 红色窗帘和米白色的壁纸, 本身已构成整个空间的背景色彩, 再用适当的米白色作为沙发区的调节色, 让客厅更加温暖舒适。	窗帘	枚	个	个	点
不料, 时过八年, 这 <CL> 船在百慕大原海区又奇迹般地出现了!	船	艘	个	艘	艘
灯一 <CL> 就够了!	够	盏	个	个	盏

全国每 <CL> 城市都在不断发展, 发展到当你走在他乡街上却感受不出你身在何处。	城市	座	个	个	个
超過八千 <CL> 工人前日發動罷工抗議, 要求資方改善待遇。	工人	名	个	个	名
曾经有一 <CL> 非常耀眼的 UFO 来地球上接我, 我没有珍惜。	UFO	架	个	次	个
身为韩国传统形式下的组合, KKB 却有能力在西楚洞的时尚摩纳哥美术馆举办的画展中展出自己的独立原创的 16 <CL> 画。	画	幅	个	幅	幅
为啥格尔木地下室会有一 <CL> 棺材哇?	棺材	具	个	个	个
前辈的话犹如一 <CL> 镜子, 照出了我的不足和短处, 革命尚未成功, 小辉辉仍需努力。	镜子	面	个	面	面
今早三 <CL> 土匪和德迅中国货运代理有限公司大中华区销售副总赵总吃饭。	德迅	则	个	个	个
用篮球术语讲, 吃了一 <CL> 火锅而已啦!	火锅	记	个	个	顿
活在当下, 必须要有一 <CL> 强壮的心脏。	强壮	颗	个	只	颗
全家人一起看动物世界, 哥哥突然指着电视上一 <CL> 猪对我说咦!	猪	头	个	只	只
当你发现你被戴了一 <CL> 绿帽子的时候也许你已经被戴了 N 顶了	帽子	顶	个	个	个
海迪住在济南市东北方位的花园小区一 <CL> 普通得连暖气都没有的居民楼内。	居民楼	幢	个	个	栋
大家都回家过年了这 <CL> 大厦突然变得好冷清	大厦	栋	个	号	座
它的美丽与否, 要看人们对这 <CL> 植物养护得如何。	植物	株	个	种	种
一 <CL> 对话我恍然了, 我们的梦有差。	对话	席	个	段	段
欣赏又信赖的感觉, 放在心底, 以后就成了一 <CL> 往事。	往事	则	个	段	段
新概念英语名师精讲第一 <CL>	精讲	册	个	个	集
向霍金致敬, 愿他今后能一如既往地苦思冥想, 努力破解这 <CL> 大难题。	难题	道	个	个	道
缘, 不可说, 而我, 像一 <CL> 仙人掌傲立于沙漠中, 静静地望着它, 看缘起缘灭。	仙人掌	株	个	棵	棵
其实适合来自不同体制间的学校整合过去是一 <CL> 所独立学校校情差异各不相同现在要整合在一起校区之间要有共识把争议放下来凝聚一下共识相信群众的智慧放下小集体的利益从更长的角度来思考学校的发展。	发展	所	个	个	个
就一 <CL> 江南调, 看书去。	江南调	曲	个	个	首
六辔是六 <CL> 马的车, 因马多且各有想法, 驾驭极难	马	匹	个	匹	匹
案情分析会, 几 <CL> 老烟枪薰得人好晕, 等下突审, 佛祖保估顺利, 不然回不了家过年了。	烟枪	杆	个	支	支
宝贝, 我在听无眠, 很舒服很适合夜晚听, 也很适合此刻心情今夜的月光超载太重, 照著我一夜哄不成梦, 每 <CL> 头发都失眠。	头发	根	个	个	根
我也算是一 <CL> 奇葩了。	奇葩	朵	个	朵	朵
浙江卫视中国蓝年会结束, 虽然没中奖但觉得为自己是中国蓝一 <CL> 骄傲。	骄傲	员	个	个	份

看烟花灿烂的同时, 看着 3 <CL> 消防车闪着灯的就奔北去了	消防车	辆	个	辆	辆
如果是 <CL> 小草, 即使在最好的企业里, 你也长不成大树。	小草	棵	个	棵	棵
五眼潭如五 <CL> 明镜, 镶嵌谷中, 人称京东一奇。	明镜	封	个	面	面
昨晚 10 点被老姐一 <CL> 电话给弄醒, 因为大批的完全不熟的探亲客要落户他们家, 听她电话里清晰有力的声音, 哈哈笑声, 突然醒来, 然后觉得轻松了。	电话	记	个	个	个
灰白色的沉重的晚云中间时时发出闪光, 接着一 <CL> 钝响, 是送灶的爆竹近处燃放的可就更强烈了, 震耳的大音还没有息, 空气里已经散满了幽微的火药香。	钝响	声	个	个	声

C Statistical analyses

This Appendix states the settings and/or calculations with which we acquired our results. We used IBM SPSS Statistics for Windows (Version 28) and Microsoft Excel (Microsoft Office 365) to perform our statistical analyses. For each test, we specify which program was used.

C.1 Chi-Squared Test

The Chi-Squared Test was used to test:

- Post hoc Experiment 1: adding to Mood's Median Test (sections 5.5.3 and 7.1.1).

Calculations for the Chi-Squared Tests were performed in Microsoft Excel (Microsoft Office 365). For each cut-off point (i.e. 1, 2, 3, 4, and 6), we created a 2x2 contingency table; one containing the observed frequencies, the other containing the expected frequencies. We then calculated the p-value using the `CHISQ.TEST` function. From this p-value, for completeness, we computed the χ^2 value using the function `CHISQ.INV.RT`, with the p-value and the degrees of freedom (which is 1 in all cases because we used 2x2 contingency tables) as its parameters.

C.2 Mann-Whitney U Test

The Mann-Whitney U Test was used to test:

- Experiment 2: Hypothesis 1 (sections 6.5 and 7.1.2).

Calculations for the Mann-Whitney U Test were performed in IBM SPSS Statistics for Windows (Version 28). We made two variables: one nominal variable indicating if a value belongs to the random or the infrequent group, and one ordinal variable representing the participants' answers. On every row, each first cell denoted the group, and each second cell contained one value from 1 to 7, corresponding to the answer given by a participant for a given classifier. To analyse the data, we selected *Analyze → Nonparametric Tests → Independent Samples*. In *Fields*, we selected the nominal variable containing the random-infrequent groups as *Groups* and the ordinal variable containing the Likert scores as *Test Fields*. In *Settings*, we selected the test *Mann-Whitney U (2 samples)* and for the test options we select a significance level of 0.01 and a confidence interval of 99.0.

C.3 Mood's Median Test

Mood's Median Test was used to test:

- Experiment 1: Hypothesis 1 (sections 5.5.1 and 7.1.1).

Calculations for Mood's Median Test were performed in IBM SPSS Statistics for Windows (Version 28). We made two variables: one nominal variable indicating if a given classifier occurred in the corpus, and one ordinal variable representing the participants' answers. On every row, each first cell denoted the corpus match. Each second cell contained one value from 1 to 7, corresponding to the answer given by a participant for a given classifier. To analyse the data, we selected *Analyze → Nonparametric Tests → Independent Samples*. In *Fields*, we selected the nominal variable containing the corpus match distinction as *Groups* and the ordinal variable containing the Likert scores as *Test Fields*. In *Settings*, we selected the test *Median test (k samples)* and for the test options we select a significance level of 0.01 and a confidence interval of 99.0.

C.4 Spearman's Rank Correlation

Spearman's Rank Correlation was used to test:

- Post hoc Experiment 1: Clarity and Fluency (sections 5.5.4 and 7.2.1).
- Post hoc Experiment 2: Clarity and Fluency (sections 6.5.1 and 7.2.1).

Calculations for Spearman's Rank Correlation were performed in IBM SPSS Statistics for Windows (Version 28). For each test, we compared two ordinal variables, Clarity and Fluency. Each cell contained one value from 1 to 7, corresponding to the answer given by a participant. Values sharing a row are the Clarity and Fluency scores provided by the same participant for the same classifier within the same context. To analyse the data, we selected *Analyze → Correlate → Bivariate*. We then selected the two relevant fields we wanted to test. Under *Correlation Coefficients* we selected *Spearman* and make it two-tailed.

C.5 Wilcoxon's Signed-Rank Test

Wilcoxon's Signed-Rank Test was used to test:

- Experiment 1: Hypotheses 2 and 3 (sections 5.5.2 and 7.1.1).
- Experiment 2: Hypotheses 2 to 5 (sections 6.5 and 7.1.2).
- Post hoc Experiment 1: differences between individual models, focus on `CORPUS` and `BERT` (sections 5.5.5 and 7.2.2).
- Post hoc Experiment 2: differences between `CORPUS` and `BERT` (sections 6.5.2 and 7.2.2).

Calculations for Wilcoxon's Signed-Rank Test were performed in IBM SPSS Statistics for Windows (Version 28). For each test, we compared two ordinal variables. Each cell contained one value from 1 to 7, corresponding to the answer given by a participant for a given classifier. Values on the same row are all provided by the same participant in the same context. To analyse the data, we selected *Analyze → Nonparametric Tests → Related Samples*. We then selected the two relevant fields we wanted to test. In *Settings*, we selected the test *Wilcoxon matched-pair signed-rank (2 samples)* and for the test options we select a significance level of 0.01 and a confidence interval of 99.0.