This document provides supplementary material to the papers:

- *The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations*.
- *Agreement is overrated: A plea for correlation to assess human evaluation reliability*.

In case of acceptance, the supplementary material will be made available via github. The link will be inserted into the papers.

Criteria used for the papers selection

This document gives an explanation of the data used in both the papers *The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations* and *Agreement is overrated: A plea for correlation to assess human evaluation reliability*. The data are recorded in the spreadsheet named *Dataset for the analysis*. The aim of this document is to provide additional information on the dimensions used in our study that could not be accommodated within the page limits of the papers.

Before moving on to the descriptions of the dimensions, let us explain the criteria used for the selection of the papers.

The first problem we faced with this study, was to decide how to select the papers to be analysed, how to retrieve them and which span of time to select. Regarding the publication years, we decided that the interval from 2008 to 2018 would be a good span of time. Ten years allows for the collection of a good quantity of data and allows us to take into account the change marked by the neural networks revolution (usually considered 2012), which can be considered a watershed in the IA community in the use of methods and techniques.

For the source papers, we decided to use the ACL Anthology. It represents a very large dataset which collects published NLP papers. To keep the study focused on NLG, we decided to select the papers from the *Special Interest Group on Natural Language Generation (SIGGEN)* webpage hosted by the ACL Anthology website. A complete list of the venues is given in the section *Paper publication venue*. We are aware that such a choice cut out from our research NLG papers published in other important conferences -- for example, ACL, EMNLP, NAACL, COLING, etc. -- and journals. However, our study is based on 526 papers which give us a fair quantity of papers from which to draw a faithful snapshot of the use of Inter-Annotator Agreement (IAA) in the evaluation of NLG tasks. The corpus selected includes papers not only on end-to-end generation systems but also on components such as referring expression generation, surface realiser, etc.

Once we determined our source of papers, we had the problem to decide which papers to consider for our study. Because the aim of our work was to analyse the use of the IAA in the evaluation phase by the NLG community, we decided to use the following criteria:

1) the paper should include a study with human annotators/judges;
2) the study should be an evaluation study (we did not take into account other tasks involving human annotation)
3) the study should allow for measurement of the IAA (for example, we did not take into account either papers in which the human evaluation was done with open questions or papers whose

human evaluation consisted of an author manually inspecting outputs. Likewise we did not take into account papers that use extrinsic evaluation methodology. However, we included papers whose extrinsic evaluation methodology was followed by a survey which allows the study of the IAA, for example, surveys done with Likert scale questions.)

Taking into account the three criteria above we ended up with a corpus of 135 papers on which we performed our analysis.

Some papers were brief and imprecise in giving information about the human evaluation used. In these cases, we recorded only the data which could be safely inferred by the information reported in the papers.

Dimension used in the analysis

We decided to split our analysis into the following 15 dimensions:

1. Paper publication venue
2. Paper title
3. Paper topic
4. Methodologies used in the human study
5. Experiment design use for the evaluation
6. Scale interpretation
7. Methodology used for the results
8. Test used for the statistical significance testing
9. Number of annotators used in the human study
10. Subjects used in the evaluation
11. Criteria or questions used in the evaluation
12. Coefficient used to measuring Inter-Annotator Agreement
13. Inter-Annotator Agreement reported
14. Inter-Annotator Agreement interpretation scale used
15. Strategy to improve the Inter-Annotator Agreement

Before introducing a detailed description of the 15 dimensions, let us introduce some terminology.

The symbol " - " means that this information was not reported in the paper.

The symbol "**(?)**" means that this information was not clearly reported in the paper but could be deduced from some span in the text.

Any label before a description (for example the symbols "Rj " or "LK7") means that the description in play has to be regarded as performed under the assumption expressed by the label. The label is defined in the same row but some column before.

1. Paper publication venue

In this column we report the publication venue of the papers.

Following the *Special Interest Group on Natural Language Generation (SIGGEN)* webpage on the ACL Anthology website we used the following conference venue from which to select the papers used in our study (we decided to focus our analysis from 2008 to 2018):

Conference Year	Conference venue	Number of selected papers
2008	INLG (41)	10
2009	ENLG (34)	3
2010	INLG (38)	7
2011	ENLG (53) UCNLG + Eval (8)	9 1
2012	INLG (28)	5
2013	ENLG (34)	10
2014	INLG (25) INLG + SIDDIAL (3)	11 1
2015	ENLG (27)	8
2016	INLG (44) WebNLG (13) Proceedings of the INLG 2016 Workshop on Computational Creativity in Natural Language Generation (9)	14 3 2
2017	INLG (42) Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms (10) XCI 2017 (4) LiRA@NLG (8) CC-NLG (5)	11 0 1 0 1
2018	INLG (63) ATA (6) Proceedings of the Workshop on NLG for Human–Robot Interaction (6) Proceedings of the First Workshop on Multilingual Surface Realisation (9) 2IS&NLG (10) CC-NLG 2018 (6)	29 1 1 1 5 1
Papers number	526	135

Table 1: Year and conference venue (number of papers in parentheses).

Table 1 shows the number of papers published in each conference and the number of papers selected for our analysis.

2. Paper title

This column contains the titles of the papers taken into account in our study. Each title is reported with the link to the paper in the ACL Anthology.

3. Paper topic

This column contains the topic of the papers.

4. Methodologies used in the human study

In our study, we divided the human methodologies used for the evaluation into three main categories: *absolute judgment, relative judgment* and *Turing Test-like evaluation*.

Absolute judgment: The human annotators are asked to assess the quality of a text based on some criteria and a scale (which can be a slider, a Likert or rating scale) or multiple choice questions. The scale can be continuous or discrete. For example, this could involve measuring the grammaticality of a sentence, associating it with a number between 1 to 5. Furthermore, in this class we also consider evaluation tasks that ask annotators to choose a category (or a text) from a set of categories (or texts). In this case, the choice is not made relative to other categories or texts.

Relative judgment: The human annotators are asked to state a preference between texts, based on some criteria. For example, this could involve either choosing a preferred sentence from two or more sentences, or ranking a set of sentences based in order of grammatical preference. Relative judgments can also involve a slider or interval scale where the annotator is asked to judge a text relative to other texts.

Turing Test-like evaluation: The human annotators are asked to decide whether a text is generated by a machine or by a human. Although this evaluation methodology can be considered as absolute judgments, we prefer to look at it as an independent class. This is motivated by the fact the Turing Test evaluation in artificial intelligence (AI) has a conceptual basis and tradition distinct from other kinds of evaluation that use, for example, ratings or sliding scales.

5. Experimental design used for the evaluation

This column contains the type of experimental design used in the human evaluations. We classify the experiments with the following labels:

Rating, Likert, Slider, Turing Test-like evaluation, Choose a category, Rank items, Relative selection and *Relative rating*.

Rating: This is an experiment that uses a rating scale. Rating scales are used in surveys to estimate feeling, opinions or attitudes of annotators. A rating scale is an n-point scale and can be both ordinal and an interval scale. The most commonly used are 3, 5, 7, 10 or 11 points. Rating scales can be both numerical and verbal. In a numerical scale a number is associated with each point. An example of a numerical scale is the following: “On a scale from 1 to 5, how fluent is the following sentence?” A variation of the numerical scale uses label words for the extreme values and assigns the intermediate values with numerical labels. An example is the scale “1-very bad, 2, 3, 4, 5-very good”, in which only the extreme values (1 and 5) are given a textual label. A rating scale that uses words as labels for the points is named a graphic rating scale. An example of points on a graphic rating scale are the following: “Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree”. It is common for the points of a graphic rating scale to also be labeled numerically, for example: “1-Strongly Disagree,

2-Disagree, 3-Neutral, 4-Agree, 5-Strongly Agree". Another kind of rating scale is the comparative rating scale. In these, scales are used to ask annotators to answer a question in terms of a comparison. An example might be: "Given the two sentences S1 and S2 which one do you prefer?" with an associated scale: "Strongly prefer S1, Prefer S1, No preference, Prefer S2, Strongly prefer S2".

Likert: This is an experiment that uses a Likert scale or a Likert item. A Likert *scale* is a summated graphic rating scale, where each graphic scale is called a Likert *item*. This means that a Likert scale is composed of a number of itemised statements. However, an individual item in itself should not be considered as a single scale. The items that make a Likert scale should not be considered in isolation and they should be summed to produce a total score. Usually, the items of a Likert scale are built in such a way that annotators express their level of agreement or disagreement with the sentence expressed by the individual Likert items. The nature of the Likert scale is highly controversial. Researchers are split between those who consider it as an interval scale and whose consider it as an ordinal scale. In fact, the key determiner of its status is usually the experiment design; that is, how the items are presented in the experiment.

Slider: This is an experiment that uses a slider scale. Annotators use a slider scale to quantitatively judge a statement or a question in a continuous setting. Although the slider scale can be both worded or numerical, in most cases the extreme points are labelled with numbers, for example from -100 to 100. Sometimes, it is possible to find a label in the middle of the scale. In some experiments, neither numbers nor words are visible to the annotators.

Multiple-choice questions: This is an experiment that uses multiple choice questions.

Turing Test-like evaluation: In this kind of experiment, annotators are asked to decide whether a sentence was generated by a machine or by a human.

Choose a category: In this kind of experiment, annotators are asked to choose one from a set of categories. We also put yes/no questions in this class. Sometimes, annotators were asked to choose one or more sentences rather than categories.

Rank items: In this kind of experiment, annotators are asked to rank a set of texts based on some criterion. For example, suppose that the criterion being judged is sentence fluency. Then, given a set of sentences, the annotators have to order the sentences from the most fluent to the least fluent.

Relative selection: In this kind of experiment, annotators are asked to select a text from a set of texts based on some criterion**.** For example, suppose that the criterion being judged is sentence fluency. Then, given a set of sentences, the annotators have to select the sentence (or *n* sentences, where *n* is an integer number given in the human instruction) that they consider most fluent in relation to the other sentences in the set. Sometimes, the preference for one text over another is expressed through a n-point comparative rating scale.

Relative rating: In this kind of experiment, annotators are asked to judge a text relative to other texts based on some criterion. For example, suppose that the criterion being is sentence fluency. Then, given a set of sentences, the annotators are asked to rate the fluency of each sentence in relation to the fluency of the other sentences in the set. Sometimes the rating is performed with a slider scale.

Magnitude estimation: In this kind of experiment, the magnitude estimation technique is used to analyse the annotators' ratings.

Sometimes, the type of the rating scale used was not clear and we could not infer a possible scale from the little information provided in the paper. In addition, sometimes the name used for the scale was either imprecise or wrong. When this happened we wrote what we think would be the correct scale name, and introduce it with the wording: "**Alternative name**".

6. Scale interpretation

This column contains the measurement scale used to gather the evaluation data. Each evaluation was performed by a survey or questionnaire. Each question uses a measurement scale to either categorize or quantify variables (or both). Eventually, each scale provides a particular type of data set, and this determines the type of statistics to be used.

Our study takes account of the *Nominal, Ordinal*, and *Interval* scales which are used in the collected papers.

Nominal scale: Nominal scales are used to ask annotators to choose a category for a set of (non-overlapping) categories. It is possible to associate to each category with a number. In a nominal scale, numbers should be considered as simple labels that do not represent a strict mathematical relation to the scale points. Such numbers cannot be used by the annotators to express a degree in their answers. The possible statistics that can be done on nominal data are either the percentages, for example the percentage of times that a category is chosen, or the mode, which indicates the most frequent value in the data.

Ordinal scale: Ordinal scales represent a step forward from the nominal scales. They are used when the order between the scale points is considered meaningful. However, the distance between two points cannot be calculated. Although ordinal scales can use numerical labels, they measure *relative* values between its points and arithmetical calculations cannot be significantly done between them. For example, it is not usually possible to say that, on a 5 point scale, the difference between "Strongly Disagree" and "Disagree" is the same as the difference between "Neutral" and "Agree". On an ordinal scale, the order between the points is what matters, but the differences between them are not known. On an ordinal scale, it is meaningful to calculate, besides percentages and mode also the median, which gives the middle value that separates the higher half from the lower half of the data. However, the arithmetic mean, such as by measuring the average value of the items, cannot be justified in the analysis of ordinal scales.

Interval scale: Interval scales represents a step forward from the ordinal scales. An interval scale allows meaningful sums and subtractions to be made between the points on the scale. By definition, the points in an interval scale are equally spread, so that the distance between any two adjacent points is the same. Within interval data it is possible to calculate the arithmetic mean of items, as well as percentages, the mode and the median.

We decide to consider the scale used as ordinal or interval depending on the statistical analyses performed on the data; that is, whether the authors used parametric or nonparametric statistics. This study takes into consideration whether the scale used for the evaluation was interpreted as nominal, ordinal or interval by the authors of the papers.

In this column we also reported whether we considered the use of interval scale to be justified or not. In the case the authors give a justification for the use of parametric statistics, we report the author's justification.

7. Methodology used for the results

This column reports how the results of the evaluation experiments were presented. In our study we found the following 18 methodologies were used. Each methodology, based on the associated experimental design, can be applied to scores, preferences or categories. The analysis can be both item-by-item or aggregated. Where the results are not specified, we mean that the results are based on a item-by-item analysis. The following methodologies are used:

Mean, Standard Deviation, Assessments addition, Frequency, Counting, Percentage, Min, Max, Mode, Median, z-score, Ordinal logistic regressions, Accuracy, Precision, Significance analysis, Categories correlation, Ratio, TrueSkill.

Although the majority of these are self-explanatory, some of these require further explanation:

Accuracy: reports the accuracy between the annotator ratings and the gold standard.

Precision: reports the number of times that annotators' judgements were the same as the gold standard.

Significance analysis: reports a significance analysis between categories used in the evaluation.

Categories correlation reports a correlation study between categories used in the evaluation.

Ratio: reports the ratio at which a system is preferred to other systems based on some criterion.

TrueSkill: reports the ranking systems based on annotators ratings.

8. Tests used for the statistical significance testing

This column contains the statistic used to measure the variance of the data collected in the evaluation study. In our study we found the following 15 statistical tests were used:

One-way ANOVA, Two-way ANOVA, One tailed sign test, t-test, Wilcoxon rank sum test, χ^2, Post-hoc Tukey, Mann-Whitney, z-test, Friedman test, Kruskal-Wallis, Wilcoxon signed-rank, Binomial test, Two-tailed Welch's t-test, Least Significant Difference test.

9. Number of annotators used in the human study

This column contains the number of people used in the evaluation study.

10. Subjects used in the evaluation

This column contains details about the subjects used in the evaluation study. Although it is not always possible to define the concept of an expert annotator in the task of evaluating human language, we decided to split this dimension into two main categories: *expert, not expert*. We decided to do so because there are cases in which it is possible to decide whether the annotators are expert (for example, meteorologists in the evaluation of weather report).

We report the information given in the papers about who the subjects were, alongside the **expert/not expert** label.

11. Criteria or questions used in the evaluation

This column contains details about the criteria or the questions used in the evaluation study.

12. Coefficient used to measuring Inter-Annotator Agreement

This column contains the coefficient used to measure the IAA reached in the evaluation study. In our study we found that the following coefficients were used:

Percent agreement, Krippendorff's α, Fleiss's k, Cohen's k, Cohen's Weighted k, Pearson's r, Kendall's W.

13. Inter-Annotator Agreement reported

This column contains the IAA measured for each criteria or questions used in the evaluation study.

14. Inter-Annotator Agreement interpretation scale used

This column contains the scale used to interpret the IAA reached in the evaluation study.

15. Strategy to improve the Inter-Annotator Agreement

This column contains details about the strategy used to improve the IAA reached in the evaluation study.

Basic concepts from statistics

For the convenience of the reader, the final section of this document provides a brief introduction to the statistical concepts that are used in this document. For an in-depth account of such concepts we refer to:

- Robert S. Witte and John S. Witte. 2017. *Statistics*, Eleventh Edition. Wiley.
- Robert L. Johnson and Grant B. Morgan. 2016. *Survey Scales. A Guide to Development, Analysis, and Reporting*. The Guilford Press.

Parametric and nonparametric statistics: Parametric and nonparametric statistics are used to analyse differences between two or more populations of interest – where a *population* is the set of subjects of a group who share one or more features which are the object of study. The term *parameter* refers to the numerical values used to describe a population. The difference between parametric and nonparametric statistics rest on the parameters' nature. If the parameters are fixed we speak of parametric statistic otherwise of nonparametric statistic. Let us now look in some more detail at these important concepts. In order to do so, we first need to present some basic statistical concepts.

Descriptive statistics: Descriptive statistics is used to summarise, through quantitative measures, features of a *sample*, where a sample is a subset of members of a given population. Given a sample from a population, the scores obtained through the use of surveys with the members of the sample

make up a *distribution*. Descriptive statistics are used to describe such distributions. *Measure of central tendency* and *measures of variability* are the most-used tools in descriptive statistics. Whereas, measures of central tendency calculate where the center of the typical value of a set of scores lie, the measure of variability calculates how similar or different the scores are.

Popular measures of central tendency are the *mode*, the *median* and the *mean*. The mode indicates the most frequent score in the collected data. The median gives the middle value that separates the higher half from the lower half of the data. That is the median is the middle value in an ordered set of scores. Finally, the mean is measured by adding all scores together and then dividing the result by the number of scores.

Prominent measures of variability are the *range*, the *variance* and the *standard deviation*. Given a set of scores, the range measures the difference between the smallest and the largest scores in the set. The variance is a type of mean. More specifically, it is the mean of the distance or deviation of each score from the distribution mean. Given the deviation score, that is the difference between a single score and the scores' mean, the variance is the mean of the squared deviation scores. The standard deviation measures the amount of deviation that a score from a distribution typically has from the sample mean. It is the square root of the variance of a set of scores.

Inferential statistics: Inferential statistics are used to draw conclusions about a population. Whereas descriptive statistics are applied to a sample of a given population, inferential statistics apply to the population. More precisely, inferential statistics use the measure of central tendency and the measure of variability between sample. In this case we refer to the set of measures performed, for example mean or variance, on each sample as *sampling distribution*. The standard deviation for a sampling distribution takes the name of *standard error*. Roughly, the standard error measures how distant each sample mean is from the population mean. Inferential statistics use data from descriptive statistics applied to samples and standard error to make decisions about population parameters. To this end, both *hypothesis testing* and *statistical significance testing* are essential.

Hypothesis testing is used to check if the data collected from a sample agrees with certain predictions or hypotheses about some measured variable. A *variable* is a feature that changes from an individual of a population to another. Significance tests are used to compare the real data collected to the data predicted by the hypothesis. If the real data significantly diverge from the hypothesis then the hypothesis can be discarded. The *null hypothesis* (H_0) and the *alternative hypothesis* (H_a) are the two mutually exclusive hypotheses used in such tests. A null hypothesis is a statement about some parameters of a given population that researchers desire to reject or nullify based of the sample data. When the null hypothesis is rejected then the alternative hypothesis is accepted. The null hypothesis declares that the differences in a set of observations from a sample result purely from chance, that is there is no significant variation between variables. Using the concept of distribution mean, the null hypothesis assert that the variation of each variable from the population mean is due to chance. Similarly, given a population and two samples the null hypothesis affirms that there is no difference between the distribution mean of the two samples. The alternative hypothesis contradicts the null hypothesis.

In order to discard the null hypothesis, the probability of collected sample observations given the null hypothesis is calculated, that is considering the null hypothesis true. Such probability takes the name of *p-value*. The smaller the p-value the stronger the evidence against the null hypothesis. But which threshold to chose to declare the null hypothesis false? In social science the threshold is set to 0.05. That is, one is permitted to reject the null hypothesis when the p-value ≤ 0.05. This assumption is called the *Type I error* rate and is denoted by α. This mean that the chance for the researchers to reject the null hypothesis is setted to the 5%. Because such an assumption is a taxonomic one, it is

possible that researchers conclude that the set of observations from a sample are statistically significant when they are not. In this case it is said that the researchers committed the Type I error. Vice versa, it is possible that researchers consider a set of observations as statistically insignificant when they are statistically significant. When this happen, it is said that the researchers committed the Type II error, which is denoted by the letter β.

We are now in a position to introduce the difference between parametric and nonparametric statistics.

Parametric statistics: Parametric statistics is a branch of inferential statistic. It assumes that the samples analyzed have a fixed set of parameters, for example that the sample distribution be known. Popular parametric statistical significance tests are: *t-test*, *z-test*, *analysis of variance* (*ANOVA*) and *post-hoc Tukey tests*.

The t-test is used to either measure if there is a significant distance between a sample variable from a determinate parameter (one-sample t-test) or if there is a significant distance between two samples variable. The aim of the t-test is to measure if such a distance is significant or is due to chance. In order to perform the t-test some assumptions have to be satisfied:

- The sample(s) have to be randomly drawn from the population.
- The observations have to be independent to each other.
- The data have to be normally distributed.

A normal distribution is a probability function. Consequently, it outlines how the values of a variable are distributed. In a normal distribution, most of the observations cluster around the central mean. The other values spread equally from the mean in both directions. For this reason, the normal distribution is considered symmetric.

In a t-test is assumed that the mean is known but the variance is unknown. In the case where the variance is know the z-test can be used instead of the t-test.

Differently from the t-test, ANOVA measures significant distance also for more than two samples variables. It can be seen as a generalization of the t -test. There are two type of ANOVA, *one-way ANOVA* and *two-way ANOVA*. Whereas in one way ANOVA only one *independent variable* is considered, in the case of two-way ANOVA two independent variables are considered. The following are the assumptions ANOVA requires to be satisfied:

- The samples are drawn from normally distributed populations.
- The sample variances have to be equal.
- The samples have to be independent of each other.

Sometimes the ANOVA test is followed by the post-hoc Tukey tests. The Tukey tests compare pairwise sample means, and besides determining if there is a significant difference between sample means, it also specifies such a difference when it is greater than the standard error.

Nonparametric statistics: Differently from parametric statistics the nonparametric statistic does not assume a fixed set of parameters. For example in nonparametric statistics sample distribution cannot be known. For this reason, nonparametric tests are also called distribution-free tests. Popular nonparametric statistical significance tests are: *Mann-Whitney U Test*, *Kruskal-Wallis one-way analysis of variance*, the *Friedman test* and the χ^2 test.

The Mann-Whitney U Test, Kruskal–Wallis one-way analysis of variance and the Friedman test are used with ordinal data. All of them measure significant distance from sample variables. Mann-Whitney U Test, Kruskal–Wallis one-way analysis of variance and the Friedman test can be considered as the counterpart, for ordinal data, of the t-test, one way ANOVA and ANOVA respectively. However, in contrast with the t-test and ANOVA, the Mann-Whitney U Test, Kruskal–Wallis one-way analysis of variance and the Friedman test do not require the sample distribution to be normal.

The χ^2 test is used when the data are collected with nominal measurements. The χ^2 test focus on the concept of frequencies. It measures the difference between the observed frequencies and the expected frequencies as declared by the null hypothesis.