Evaluating the Evaluators: Subjective Bias and Consistency in Human Evaluation of Natural Language Generation

Thesis

The Open University

School of Computing and Communications

Evaluating the Evaluators

Subjective Bias and Consistency in Human Evaluation of Natural Language Generation

Jacopo Amidei

A Thesis submitted for the degree of

Doctor of Philosophy

September 2020

Abstract

The Natural Language Generation (NLG) community relies on shared evaluation techniques to understand progress in the field. Based on an analysis of papers published over 10 years (from 2008 to 2018) in NLG-specific conferences and on an observational study, this thesis identifies shortcomings with existing approaches to reporting the reliability of evaluation studies in NLG. It proposes a new set of methods for identifying judges' bias and reporting reliability, specifically for human intrinsic evaluation of NLG systems.

In this thesis, we propose to use the correlation statistic and Item Response Theory (IRT) to analyse judges' bias for cases that involve a high level of language variability. Both techniques provide insights about the trustability of human judgements. Whereas the correlation statistic offers an approach to measure judges' relative consistency, IRT provides a tool to identify judges' bias.

We found support for the use of the correlation statistic through three case studies that show the limits of considering agreement coefficients as the only criterion for checking evaluation reliability. Given the variability of human language – specifically variability in language interpretation and quality judgement – expecting judges to always arrive at exactly the same judgement seems both unrealistic and over-constrained. The correlation coefficients can be used to measure the extent to which judges follow a systematic pattern in their assessments, even when their individual interpretations of the phenomena are not identical.

Regarding IRT, we introduce a new interpretation and application of the technique to describe judges' bias. Using the QG-STEC evaluation dataset, and applying IRT to each judge, we show how to use IRT's probabilistic analysis to compare judges' bias and as result better characterize annotation disagreement. The new approach that we propose, can be used, for example,

to spot judges who are outliers, improve annotation guidelines and arrive at an improved interpretation of the agreement coefficients.

Acknowledgements

This page closes a chapter of my life that lasted three years. As with any long journey, during these years I had the luck to meet many people, who, one way or another, have been of help in achieving this goal. I will make sure to thank them one by one in person.

Here, I would like to write a few words to thank my supervisors: Paul Piwek and Alistair Willis.

I want to thank them for always being supportive and guiding my research during this journey. Not an easy thing, given my often unclear and fuzzy train of thought! Every time they were able to shape and contextualize my though, with constant constructive feedback, suggestions and corrections.

That said, I really want to thank Paul and Alistair not just for having been excellent supervisors but for being very understanding and supportive in a difficult period of my life due to personal reasons. They gave me the freedom to take the necessary steps in order to recover my personal balance. They allowed me to find the right harmony between academic demands and personal life. For this reason, I can only be extremely grateful for their understanding and sensibility.

Barcelona, September 2020.

Contents

Acknowledgements **i**

Introduction **vii**

I Background **1**

1 Diving into evaluation of NLG systems: An introduction to the main concepts **2**

1.1 NLG task definition 3

1.2 The importance of evaluation 4

1.2.1 The evaluation challenge 5

1.3 Evaluation types for NLG tasks 8

1.3.1 Intrinsic evaluation methods 8

1.3.2 Extrinsic evaluation methods 14

1.4 Annotation and annotation guidelines 15

1.4.1 The case of intrinsic human evaluation of NLG systems 16

1.4.2 Three concepts of data reliability 18

1.4.3 Kappa statistic and their scales of interpretation . . . 22

1.4.4 Problems with human agreement for the case of NLG 23

1.5 Conclusion 26

2 Analysis of the recent use of evaluation methodologies in

NLG **27**

2.1 Evaluation methodologies for AQG 29

2.1.1 Criteria for papers selection 29

2.1.2 A general overview 33

2.1.3 Intrinsic automatic evaluation 35

2.1.4 Human evaluation 37

2.1.5 Extrinsic evaluation 42

2.1.6 Preliminary conclusions 43

2.2 A deep analysis about the use of IAA in NLG evaluation task 45

2.2.1 Criteria for papers selection 45

2.2.2 10 years of IAA in evaluation of NLG systems 48

2.3 Conclusion 52

II Subjective Bias, Agreement and Consistency **55**

3 Agreement coefficients: Definition and use **56**

3.1 Preliminary concept and terminology 56

3.1.1 The agreement coefficient terminology problem 57

3.1.2 Observed agreement 58

3.1.3 Four interpretations and definitions of chance agreement 59

3.1.4 Type of data 61

3.2 Agreement coefficients 63

3.2.1 A common notation 63

3.2.2 Working with missing data 63

3.2.3 Weighting the agreement coefficients 64

3.2.4 A common definition of weighted percent agreement P(A) 67

3.2.5 1st approach to chance agreement: Equal probability . 69

3.2.6 2nd approach to chance agreement: Annotators probability . . . 71
3.2.7 3rd approach to chance agreement: Categories probability . . . 73
3.2.8 4th approach to chance agreement: Mixed probability 76
3.3 Agreement coefficients interpretation . . . 83
3.4 The prevalence paradox . . . 85
3.5 Agreement coefficients comparison: an experimental point of view . . . 87
3.6 How to perform an agreement study . . . 90
3.6.1 Choose the agreement coefficient . . . 90
3.6.2 Coefficient interpretation . . . 94
3.6.3 Agreement coefficient value reproducibility . . . 94
3.7 irrCAC a R library to measure agreement coefficients . . . 95
3.8 Conclusion . . . 101

4 Detecting sources of annotation disagreement 103
4.1 Description of the observational study . . . 104
4.2 A new taxonomy of divergences . . . 111
4.2.1 Style and taste . . . 112
4.2.2 Background knowledge . . . 113
4.2.3 Personal assumptions . . . 114
4.2.4 Use of common sense inferences . . . 115
4.2.5 Attention to detail . . . 116
4.2.6 Other sources of disagreement . . . 118
4.2.7 Concluding remarks . . . 119
4.3 Conclusion . . . 121

5 From agreement to consistency 123
5.1 Consistency as stability . . . 124

5.2 Relative consistency 127
5.2.1 The use of correlation coefficients for NLG human evaluation tasks 128
5.2.2 Datasets analysis 131
5.2.3 Interpretation of correlation coefficients case studies . 132
5.3 How to report correlation coefficient 140
5.4 Conclusion 144

6 Identifying judge bias **146**
6.1 Our proposal in a nutshell 147
6.2 Analyzing raters bias: From frequency to Item Response Theory 150
6.2.1 Frequency 150
6.2.2 Item Response Theory 151
6.3 The use of GRM to analyse judges' bias: An example from NLG evaluation 157
6.3.1 The ambiguity criterion 161
6.3.2 The variety criterion 164
6.3.3 The fluency criterion 167
6.3.4 The relevance criterion 171
6.4 How to report an IRT study 175
6.5 Conclusion 176

Conclusion and future direction **177**

III Appendix **184**

A Annotation guidelines for the iterations presented in Chapter 4 **185**
A.1 Question Generation evaluation guidelines: Iteration 1 (I1) . 185
A.2 Question Generation evaluation guidelines: Iteration 2 (I2) . 193

A.3 Question Generation evaluation guidelines: Iteration 3 (I3) . 198
A.4 Question Generation evaluation guidelines: Iteration 4 (I4) . 204

List of Figures **209**

List of Tables **211**

Bibliography **214**

Introduction

Let's suppose we ask two people, called A and B, to perform the following task:[1]

Please read the following paragraph before answering the question below.

Paragraph: *Frank Vincent Zappa (December 21, 1940 - December 4, 1993) was an American musician, composer, activist and filmmaker. [...] Zappa was born in Baltimore, Maryland. His mother, Rosemarie (née Collimore) was of Italian (Neapolitan and Sicilian) and French ancestry; his father, whose name was Anglicized to Francis Vincent Zappa, was an immigrant from Partinico, Sicily, with Greek and Arab ancestry.*

Q: *Did Frank Vincent Zappa die when he was 53 years old?*

Does **Q** have a clear answer within the input paragraph? Answer no if you can not find a clear answer in the text above.

[ ] YES [ ] NO

Let's call the box above an *item*, and let's suppose that A and B perform the

[1] The text was retrieved on the 15 of May 2020 from the Frank Zappa's English Wikipedia page: `https://en.wikipedia.org/wiki/Frank_Zappa`.

same task on 100 items.[2] In this situation we have an *annotation* and two annotators (A and B). More specifically, we are conducting an *evaluation* of natural language. Each item aims to evaluate if the question **Q** can be clearly answered by the input paragraph.

Regardless of the reasons that make someone collect this data — that is, to perform the annotation on the 100 items — the most vital question to ask is: *Is that annotation reliable?* So why is annotation reliability important?[3] As pointed out by Krippendorff (1980):

> [...] researchers need to demonstrate the trustworthiness of their data by measuring their reliability. If the results of reliability testing are compelling, researchers may proceed with the analysis of their data. If not, doubts as to what these data mean prevail, and their analysis is hard to justify. (Page, 212)

In other words, we can rephrase the question *Is that annotation reliable?* with the question: *Can we have trust in the annotators and confidently use the annotation?* Indeed, A and B can perform the annotation randomly. In this case, we shouldn't use the annotation.

The Content Analysis community (Krippendorff, 1980) developed precise methods (agreement coefficients) to answer our previous question — that is, to measure annotation reliability. Such methods are based on computing the extent to which annotators arrive at exactly the same annotation decision. The more the annotators reach the same annotation decision,[4] the higher the reliability. Nevertheless, as we will show in Chapter 2 and Chapter 4, the standard method used in Content Analysis falls short in the case of the evaluation of natural languages. The example we introduced at the

[2] More precisely, each item is a text passage and a question, **Q**, with A and B asked whether **Q** can be answered from the text.

[3] In Section 1.4.2, we we will cover this topic in detail.

[4] It is important that the annotators perform the annotation independently of each other.

beginning of this introduction can help demonstrate this point.

Let's return to our example. The item we presented aims to evaluate if the question **Q** can be clearly answered by the input paragraph. It can be a matter of disagreement whether **Q** can be clearly answered by the input paragraph. Indeed, for the annotators to answer **Q**, they require at least two pieces of background information. Firstly, they need to know that the information between the brackets that immediately follows the name Frank Vincent Zappa represents his date of birth and his date of death. Secondly, it requires understanding of the mathematical operation of subtraction in the domain of dates. Based on this observation, we can imagine a scenario where A answers "YES", but B answers "NO", because B thinks that the two implicit requests do not make it easy to answer **Q** clearly. We can also imagine a scenario where A answers "YES", but B answers "NO" because B has miscalculated Frank Zappa's age at the time of his death.

As the example has shown, non-linguistic aspects — for instance, preconceptions, background knowledge and inference ability — can affect the annotators' decisions, leading to a *disagreement* between annotators. Annotators' disagreements can be due to annotators' carelessness during the annotation task, but with the example in question it could also be *genuine disagreement* — that is, disagreement due to annotators' subjective bias. In annotation tasks where genuine disagreement has to be preserved, the methods developed by the Content Analysis community can be too restrictive. Indeed, excessively restrictive methods are in danger of discarding highly informative annotation which has a high level of genuine disagreement. The human evaluation of Natural Language Generation (NLG) tasks — that is, computational tasks that aim to produce natural language — is an instance of annotation that should take into account genuine disagreement.[5] Genera-

[5]Through the thesis when referred to "The human evaluation of Natural Language Generation (NLG) tasks" we refer to the human evaluation of NLG tasks' outputs.

tion system developers use evaluation results to help them understand how they can improve the communicative power of their systems. Levelling annotators' genuine disagreement runs the danger of biasing system developers towards ignoring important aspects of human language.

Our example suggests that, in cases where high variability in language interpretation is involved in the annotation, that expecting annotators to arrive at exactly the same annotation decision may be both unrealistic and overconstrained. As we will see in Section 4, genuine disagreement cannot be removed by improving the item question. Driven by such observations, this thesis aims to describe how to study reliability for an annotation task which involves a high level of language variability. Specifically, this thesis focus on human evaluation of NLG systems.[6]

More precisely, the following is the research question that has driven the development of this thesis:

> How should we carry out a reliability study for NLG human evaluation tasks, given their high level of language variability?

We believe that the thesis' outcome, and therefore its contribution, is twofold. First, it raises, in the NLG community, awareness about the need to handle the problem of human evaluation reliability. Secondly, it suggests a way to carry it out.

A glance at the chapters

This thesis is composed of three main parts:

I) Background;

[6] Although we tested our methods on human evaluation of NLG systems, we believe our proposals are also applicable to other annotation tasks with high levels of subjectivity such as discourse and dialogue annotation. Nevertheless, the applicability of our proposal to other areas of research has to be empirically tested.

II) Subjective Bias, Agreement and Consistency;

III) Appendix.

Background

The Background is made up of two chapters:

- In Chapter 1 we contextualise the work with reference to the literature, and present the preliminary concepts and terminology used in the thesis.
- In Chapter 2 we set out the evidence from two studies into evaluation practices which provide a justification for the research questions that are central to this thesis.

Subjective Bias, Agreement and Consistency

This part represents the main contributions of the thesis. It is based on four chapters:

- In Chapter 3 we present and discuss five popular coefficients of agreement and their interpretation. The aim is to limit, and hopefully eliminate, the trend we detected in Chapter 2: that is, the presence of shortcomings and oversights in reporting coefficients of agreement results in the NLG community.
- In Chapter 4 we perform an observational study which identifies annotators' subjective bias. Subjective bias seems resistant to improving the annotation guidelines, since realistic evaluation criteria need to be maintained.
- In Chapter 5 we argue for correlation — which can be used to measure the extent to which annotators follow a systematic pattern in their assessments, even in the absence of agreement — to assess human

evaluation reliability.

- In Chapter 6 we introduce an original bias identification method — based on a new interpretation and application of the Item Response Theory (IRT) (Gulliksen, 1950) — to detect annotators' bias.

Before the Appendix there is a conclusion chapter, in which we present a short example that summarises our proposal for performing a reliability study for NLG human evaluation tasks. We end the chapter with a discussion of some work to be undertaken in the future.

Appendix

In this part there is one appendix chapter:

- In Appendix A we collect the annotation guidelines used in the observational study presented in Chapter 4.

Sources

The main ideas of this thesis have been published in Natural Language Processing related conferences. More precisely, Chapter 2 is mainly based on the papers:

- J. Amidei, P. Piwek and A. Willis, Evaluation Methodologies in Automatic Question Generation 2013 - 1018, *In Proceedings of The 11th International Natural Language Generation Conference (INLG)*, Tilburg, The Netherlands, 5-8 November 2018, pages 307-317.

- J. Amidei, P. Piwek and A. Willis, Agreement is overrated: A plea for correlation to assess human evaluation reliability, *In Proceedings of The 12th International Natural Language Generation Conference (INLG)*, Tokyo, Japan, October 29 - November 1, 2019, pages 344-354.

Chapter 4 is mainly based on the paper:

- J. Amidei, P. Piwek and A. Willis, Rethinking the Agreement in Human Evaluation Tasks, *In Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, 20-26 August 2018, pages 3318-3329.

Chapter 5 is mainly based on the paper:

- J. Amidei, P. Piwek and A. Willis, Agreement is overrated: A plea for correlation to assess human evaluation reliability, *In Proceedings of The 12th International Natural Language Generation Conference (INLG)*, Tokyo, Japan, October 29 - November 1, 2019, pages 344-354.

Finally, Chapter 6 is mainly based on the paper:

- J. Amidei, P. Piwek and A. Willis, Identifying Annotator Bias: A new IRT-based method for bias identification, *In Proceedings of the 28th International Conference on Computational Linguistics (COLING), Barcelona, Spain, December 8-13, 2020, pages 4787–4797.*

Part I

Background

Chapter 1

Diving into evaluation of NLG systems: An introduction to the main concepts

The aim of this chapter is to introduce and define the relevant background from which we will develop our ideas. We present the preliminary concepts and terminology. While doing so, we contextualise them into the literature relevant to the present work. By outlining the context in which this thesis is placed, this chapter sets the present work inside a well-established research area. It is worth noting that this section does not aspire to – and given the vast research area could not be able to – give an exhaustive picture of the NLG evaluation landscape. The main goal is to contextualise the central ideas that are the background of this research.

In each section we introduce the main concepts we will work with in the following chapters. In Section 1.1 we supply a definition of NLG. In Section

1.2 we discuss the importance of the evaluation phase and the challenge that it raises. In Section 1.3 we present different methodologies used in the evaluation stage. In Section 1.4 we first introduce the concept of annotation and annotation guidelines and later we discuss the case of human evaluation in NLG systems. We present the Kappa statistic after we review three concepts of data reliability. Finally, we consider the limits of using the Kappa statistic as the only tool to measure data reliability.

1.1 NLG task definition

NLG is a computational process that converts an input source into words, sentences and more generally texts, in the form of natural language. Such a broad definition, which focuses on the aim of NLG, displays the difficulty of defining the boundaries of this discipline – see for example Evans et al. (2002) and Gatt and Krahmer (2018). Although there is agreement on what should be the output of an NLG system, the types of input source are still a matter of debate.

Defining the input types of input source is not just a taxonomy problem. The difference in input source can imply a change in the techniques used to produce the linguistic output. Based on the difference in the types of input source, to date, the main NLG tasks are *text-to-text* and *data-to-text* (Gatt and Krahmer, 2018). Although the text-to-text case specifies the linguistic nature of the input source, the data-to-text case can leave the boundaries on the input source nature quite blurred. Indeed, data-to-text can be considered a generalisation of text-to-text. As noted by Gatt and Krahmer (2018) textual input can be considered a form "in which input data might be presented" (Gatt and Krahmer, 2018, page 135).

In this thesis we are interested in the study of the evaluation of NLG systems. This put us in a position to focus on the output more than the kind of input

source. That is, regardless of what the type of input source is, when we refer to NLG tasks, we consider tasks that require generation of human language and whose output space is, as a result, very wide (given that a natural language often allows the same information to be expressed in many ways). Examples of these are: *dialogue generation* (Chen et al., 2017) in which a dialogue history is given as an input and the model must generate a response which follows up on the history. *Image caption generation* (Bai and An, 2018), in which an image is given as an input and the model generates an image description. *Question generation* (Rus et al., 2008), in which a text paragraph is given as an input and the model must generate a question about the paragraph.

1.2 The importance of evaluation

Over the past 10 to 15 years, the NLG community has come to rely increasingly on shared tasks and evaluation techniques to understand progress in the field. Evaluation has become a critical phase for the development of NLG systems. It helps to improve systems' performance, for example by showing systems' weaknesses. Furthermore, the use of shared evaluation techniques enables a potentially reliable comparison between systems. This provides a clear approach to measure the advances introduced by the systems.

As pointed out in Gatt and Krahmer (2018), an important event for the NLG evaluation tasks was the "establishment of a number of NLG shared tasks, launched in the wake of an NSF-funded workshop in Virginia in 2007 (Viethen and Dale, 2007)". In a Shared Task Evaluation Campaign (STEC) (Gatt and Belz, 2009) a task, a dataset – which is split into development and test set – and a set of evaluation methodologies are given. All the participants develop systems which are evaluated on the same test set and with the same methodologies. STEC help to understand the validity of different eval-

uation methods and the relationship between them. Some examples are the Question Generation Shared Task Evaluation Challenge (QG-STEC) (Rus et al., 2010), the TUNA Challenge (Gatt and Belz, 2009) or more recently the WebNLG Challenge (Colin et al., 2016) and the E2E NLG Challenge (Dušek et al., 2018).

1.2.1 The evaluation challenge

In NLG, the evaluation challenge arises mainly from the fact that human sentences can be generated, used and interpreted in many different ways. Indeed, humans can reach the same communicative goal by using many different expressions. This variability makes the effort of evaluating sentence quality extremely complicated. Given an NLG generation task T and a set of generated sentences S, how can we decide if a sentence fulfills the task T? Which sentence in S should be preferred? Let us be more precise using one example. Let's suppose that the task T in play is an Automatic Question Generation one. The task T is: "Given a piece of text, I, generate a question in natural language, S, such that S is fluent and it can be unambiguously answered from I".[1] Suppose that I is the following text:[2]

I: *Frank Vincent Zappa (December 21, 1940 - December 4, 1993) was an American musician, composer, activist and filmmaker. [...] Zappa was born in Baltimore, Maryland. His mother, Rosemarie (née Collimore) was of Italian (Neapolitan and Sicilian) and French ancestry; his father, whose name was Anglicized to Francis Vincent Zappa, was an immigrant from Partinico, Sicily, with Greek and Arab ancestry.*

Suppose that given I, a system $\hat{S}$ generates the following set S of outputs:

Q1: *Where Frank Vincent Zappa was born?*

[1] In Section 2.1.1 we will present three different kind of Automatic Question Generation tasks.

[2] We took this example from the Frank Zappa's English Wikipedia page `https://en.wikipedia.org/wiki/Frank_Zappa`. The example was retrieved the 15 of May 2020.

Q2: *Where was Vincent Zappa born?*

Given S, which question in S fulfills the task T? Which of them should be preferred? Both **Q1** and **Q2** aim to elicit the same information – that is the place where Frank Zappa was born. On one hand, **Q1** can be unambiguously answered from **I** – *Baltimore, Maryland* is the answer – nevertheless, it is not fluent English. On the other hand, **Q2** although fluent, cannot be unambiguously answered from **I** – both *Baltimore, Maryland* and *Partinico, Sicily* are answers for it. The example illustrates a simple instance of the evaluation challenge, which, as we will see in Section 4, can be more complex.

The ability to generate different sentences must be considered crucial in each generation system, since its aim is to produce human-like language. For example, to pass the Turing test (Turing, 1950), it becomes essential in NLG to develop evaluation methods for discerning good from bad automatically-generated sentences.

Any generation task, whose output space is very large, faces the problem of the absence of a comprehensive gold standard. A gold standard, usually obtained through human input, is thought to be the correct set of solutions for a particular task T, and it is used either for training or evaluating models for T's purpose. Given a task T, the gold standard evaluation approach assumes that we can define a precise and exhaustive (or near exhaustive) set of outputs for T. In NLG the gold standard, also called the reference, is a set of natural language words, sentences or more generally texts. Given the huge human ability to use many different expressions to reach the same communicative goal, language generation tasks cannot safely rest on the gold standard approach for evaluation purposes.

In NLP, the reference set approach has been used for evaluation through automatic metrics, which compare the automatically generated sentences against the references. Many recent studies in NLG have shed light on

the correlation between human judgement and metrics such as BLEU – see for example (Reiter and Belz, 2009). The results, which have shown this correlation to be weak – see for example (Reiter, 2018) –, cast doubt on the feasibility of using these metrics to evaluate the overall system quality. Indeed, a system may generate very high-quality sentences that are different from the ones in the reference set used for evaluation — this is especially possible if we train and evaluate a model with different corpus text. Let us return to the Automatic Question Generation example. Suppose that the reference set R is made by the questions:

- Where was Frank Vincent Zappa born?
- When was Frank Vincent Zappa born?
- Where was Francis Vincent Zappa born?

Suppose that a system $\hat{S}$ generates the questions:

- Did Frank Zappa have Italian, French, Greek and Arab ancestry?
- Did Frank Zappa die in 1993?

Using the gold standard approach and given the set R, $\hat{S}$ would be considered a bad generative system. Nevertheless, $\hat{S}$ can generate fluent questions which can be answered by **I**. We will examine the automatic metrics in Section 1.3.1.

For NLG tasks, the reference set approach which, given a task T, can supply an ideal solution for T, is unrealistic. For this reason, a referenceless quality estimator was proposed, see for example Dušek et al. (2017) and Groves et al. (2018). Such attempts, although of interest, are in the initial stages and more developments are needed.

As we will see throughout this thesis, the difficulty of building an exhaustive reference set is not just a problem for automatic evaluation metrics. Indeed,

it is reflected in the difficulty of reaching a high human evaluation agreement about sentence quality.

1.3 Evaluation types for NLG tasks

In the evaluation of NLG tasks we are interested in the NLG systems' outputs quality, that is, the quality of the generated sentences. Traditionally, two types of evaluation methodologies are used for NLG evaluation: *intrinsic evaluation* methods and *extrinsic evaluation* methods.

Gkatzia and Mahamood (2015) studied the use of evaluation methodologies for NLG. They performed their study analysing a corpus of 79 works, which was published in conference and journal papers between the years 2005-2014. Their results show the prevalence of intrinsic evaluation methods over extrinsic ones. Gkatzia and Mahamood (2015) also noticed that the evaluation approaches are correlated with the publication venue — that is, papers published in the same journal or conferences tend to use the same evaluation methodologies. These findings show the difficulty of, and emphasise the need for, standardising the evaluation methodologies. Chapter 2 represents a continuation and a refinement of Gkatzia and Mahamood's 2015 work.

1.3.1 Intrinsic evaluation methods

Intrinsic evaluation methods measure the performance of a system by evaluating the system's output "in its own right, either against a reference corpus or by eliciting human judgements of quality" (Gatt and Belz, 2009, page 264). For example, this could involve measuring the output's grammaticality and fluency. The prevailing intrinsic methods are *human evaluation* and *automatic evaluation*. In order to assess the quality of a generated sentence, whereas the first method uses human judgements, the latter applies an algorithm that automatically calculates a score (for example by checking the

similarity between the generated sentence and a set of reference sentences).

Human evaluation

In NLG, human evaluation is performed by asking people to assess the sentence quality either by rating (absolute annotations) or ranking (relative annotations) sentences against some criteria, for example fluency or naturalness. For each task T, the criteria – which are determined by an annotation scheme and described in an annotation guideline[3] – define the concept of quality the researchers are interested in for T.

> Before continuing with this section it is important to give a terminological clarification. In this thesis we will use the term *scale* and *item* in the following way. In the context of a statement, the term scale is the group of points making up the options offered to respondents. We refer to the combination of the statement and the scale as an item. In the case of an aggregate scale, such as the Likert scale, we use the term *scale* to indicate a collection of items. For more details about this point and the concept of scale we refer to Amidei et al. (2019).

Human annotators are usually asked to to annotate an item based on some default ranking or Likert-style scale in absolute annotations. For example, this could involve measuring the naturalness of a sentence associating with it a number between 1 to 5. Figure 1.1 presents a case in which human annotators are asked to assess the sentence quality on a five-point numerical scale based on the naturalness criterion.

Both rating and Likert scales are widely used in the human evaluation of NLG systems. However, their nature and their appropriate statistical analysis remain a matter of controversy. In particular, the statistical analysis

[3] We refer to Section 1.4 for more details.

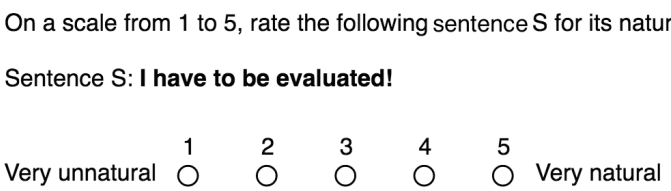


Figure 1.1: Example of an intrinsic human evaluation item based on a numerical rating scale.

of data from Likert scales is controversial. Some treat the data as interval, others treat the data as ordinal. The proper interpretation of a Likert scale is as an aggregate scale, that is, it is a composite of items which are summed or averaged all together to get an overall positive or negative orientation towards the object under examination in a survey. In this case, data from a Likert scale are considered interval. Accordingly, the use of parametric statistics are justified. Sometimes, however, individual items of a Likert scale are considered on their own by researchers. In this case, data from a Likert scale are considered ordinal. Under such an interpretation, the use of parametric statistics cannot be justified. For more details about this point, we refer to Amidei et al. (2019).

Alternative scale types can also be used. For instance, Belz and Kow (2011) show the viability of using *continuous scales* instead of a discrete one for NLG evaluation purpose. Alternatively, Siddharthan and Katsos (2012) used *magnitude estimation* for evaluating the readability of automatically generated texts. Magnitude estimation was introduced for linguistic judgment by Bard et al. (1996). Bard et al. used this technique for judging acceptability as a better alternative to judging grammaticality.[4] Both continuous scales and magnitude estimation were employed because they provide

[4]Grammaticality judgements aim to measure the extent to which a linguistic item (such as a sentence) meets a set of fixed rules regardless of the context in which the sentence is used. Acceptability judgements aim to measure the extent to which a sentence satisfying the grammatical rules is considered permissible, i.e. fluent, easily understandable and appropriate in a given context.

fine-grained measurements by capturing robust or subtle differences between sentences. Indeed, a limit of rating and Likert scales is to force judges to make a choice between a fixed number of possible distinctions. Magnitude estimation overcomes this problem by designing an evaluation task in which judges are asked to assign numbers to a series of sentences proportional to the sentence quality they perceive. Quality is defined by the criterion at play, for example, sentence acceptability. No fixed number is given. Judges are first exposed to a modulus sentence, to which they assign an arbitrary number. All the other sentences are rated proportionally to the modulus sentence — for example, if a sentence is perceived as twice as good as the modulus, it gets two times the number associated to the modulus sentence. As a result, magnitude estimation employs no fixed continuous numerical scale. After the annotation, the judges' individual scales are normalized by using the statistical z-score, in order to allow comparison between them.

In relative annotations, the human annotators are asked to assess the preference between a set of sentences based on some criteria. For example, this could involve choosing between two sentences, depending on the sentence naturalness. Relative and absolute annotations are compared in Belz and Kow (2010). Their experiment suggests that relative annotations reach a slightly better human agreement than absolute annotations. Such a result is confirmed in Novikova et al. (2018), where the authors introduced a new method (RankME) which combines the use of relative assessments and magnitude estimation (with continuous scales). Carterette et al. (2008) suggest that relative annotations also have the convenience of being more intuitive and quicker than absolute annotations. Nevertheless, as shown in Carterette et al. (2008), relative annotations have the drawback of the quadratic explosion of the possible alternatives. This makes the evaluation task quite expensive, both in an economic sense and in the expenditure of time (to be collected, the evaluation results could need a long time). On the other

hand, absolute annotations do not suffer from this problem and have the advantage of supporting a more fine-grained analysis. These allow for a better error analysis. Such evaluation can provide deep insight into the system weaknesses. For example, it can help us understand the degree to which the system is able to generate fluent sentences. Conversely, knowing that a system S_1 is able to generate sentences better (for example, from a fluency point of view) than the system S_2 does not means that S_1 is able to generate fluent sentences. Indeed, it is possible that both the systems S_1 and S_2 generate sentences that are not fluent, but that the ones generated by S_1 are slightly more fluent than the ones generated by S_2.

Regardless of whether absolute annotations or relative annotations are used, NLG human evaluation tasks are usually driven along two dimensions. A linguistic one, which aims to evaluate the sentence quality from a grammatical and idiomatic point of view, and an assignment-oriented one, which aims to check if the sentence fulfills the task for which it was generated. Gatt and Krahmer (2018) note that traditionally, in NLP, the linguistic dimension is defined by the *fluency* or *readability* criteria, while the assignment-oriented one is defined in term of *accuracy*, *adequacy*, *relevance* or *correctness* criteria. However, in Section 2 we will see that more criteria can be defined.

Automatic evaluation

Automatic evaluation metrics judge system quality by comparing the generated output against a set of references. Such references aim at fulfill the task goal for which the system was created. In NLG, sentence quality for automatic evaluation metrics is defined by the concept of humanlikeness. That is, the ability to automatically generate sentences as similar as possible to the reference ones. The assumption behind the humanlikeness is that, given the same input, the automatically generated sentence and the references' ones both aim to fulfill the same task goal. The more the automatically

generated sentence is similar to the reference sentences, the more it is good. In this context, the concept of similarity is defined by the automatic metric in play. There are diverse ways of cataloguing the automatic evaluation metrics in the literature because different metrics share the concept of similarity. Following Sharma et al. (2017), we can use the following categories: *Word-overlap based metrics* and *Embedding based metrics*.[5]

Word-overlap based metrics check the similarity between two sentences based on the *n-gram overlap*.[6] Some examples are *BLEU* (Papineni et al., 2002), *METEOR* (Banerjee and Laviel, 2005) and *ROUGE* (Lin and Och, 2004). Embedding based metrics rely on the idea that sentence similarity can capture semantic information that word-overlap based metrics cannot. [7] For example, suppose an NLG system generates the sentence "Trump speaks to the media in Illinois", where the reference set is composed of the sentences: "The President greets the press in Chicago", "The USA president holds a press conference in Chicago" and "The President entertains journalists at the Fountain of time". In a case like this, embedding metrics can detect the relevance of the generated sentence whereas word-overlap metrics cannot. Indeed, although there is no more than 1-gram overlap, sentences such as "Trump" and "The USA president" receive high word embedding similarity. This is the same for the words "Chicago" and "Illinois". Examples of embedding based metrics are *Vector extrema*, *Greedy matching* (Sharma et al., 2017) or BERTScore (Zhang et al., 2019).

Automatic evaluation metrics have the advantage of being fast, repeatable and cheap. However, it has been shown that they correlate poorly with

[5]For deep cataloguing, we refer to Gatt and Krahmer's paper (Gatt and Krahmer, 2018, page 126).

[6]n-gram are fixed length (n, for n natural number) consecutive sequences of words occurring in a text. For instance, $\{(\text{this is}), (\text{is an}), (\text{an example})\}$ is the set of 2-gram of the sentence "this is an example".

[7]A word (sentence) embedding is a vector-based representation of the word (sentence). It aims to model semantic relations between the embedded words in terms of mathematical operations on the vector representations of those words.

human judgments. For example, Reiter and Belz (2009), Liu et al. (2016), Novikova et al. (2017) Shimorina (2018) and Reiter (2018) show that existing automatic evaluation metrics are poor indicators of sentence quality as perceived by humans. This leaves the interpretation and the usability of the existing evaluation metrics as highly problematic.

1.3.2 Extrinsic evaluation methods

Extrinsic methods measure the performance of a system by evaluating the system's output with respect to its ability to carry out the task for which it was generated. An example of extrinsic evaluation methods is the one used to evaluate the STOP system developed by Reiter et al. (2003). STOP generates "short tailored smoking cessation letters, based on responses to a four-page smoking questionnaire" (Reiter et al., 2003, page 41) with the aim of helping people give up smoking. This system was evaluated "by recruiting 2553 smokers, sending 1/3 of them letters produced by STOP and the other 2/3 control letters, and then measuring how many people in each group managed to stop smoking" (Reiter, 2017). In this case the system was evaluated in the real world to see if it had the desired effect — that is, if it was able to fulfil the task goal for which it was developed.

In this context, the quality of a generated sentence is measured in terms of its ability to fulfil the task goal. On a task-based approach the quality of a generated sentence is measured by its capability in achieving a desired goal, where the definition of capability is based on the system's application and purpose.

As suggested in Reiter and Belz (2009), extrinsic evaluation methods "have traditionally been regarded as the most meaningful kind of evaluation in NLG" (page 531). The importance of this kind of evaluation is also defended by Reiter (2011). However, extrinsic evaluation methods are seldom used,

because they are expensive and time-consuming (Reiter and Belz, 2009). There has been no systematic comparison conducted between extrinsic evaluation methods and human judgement, unlike automatic evaluation metrics. The only study that we are aware of is that of Belz and Gatt (2008). Belz and Gatt (2008) shows no significant correlation between extrinsic evaluation methods and human judgement.

1.4 Annotation and annotation guidelines

Annotation is the task of augmenting corpus data with linguistic or other information. For the case of NLP, Pustejovsky and Stubbs (2013, page ix) define the annotation task as "the process of adding metadata information to the text". This operation can be done by machine or humans. The reasons for annotating a corpus data can be reduced to two main aims: a theoretical one and a practical one. Whereas the first aim is to develop and test linguistics theories, the latter is to train or evaluate systems for Artificial intelligence (AI) applications. Intrinsic human evaluation falls into the second category. They are an annotation task which aim to evaluate NLG systems.

Human annotations are driven by annotation guidelines, which are a direct manifestation of an annotation schemes. Whereas the latter characterise the criteria to be annotated, the former strictly define such criteria and suggest how they should be annotated. More specifically, annotation schemes determine the conceptual content of the annotation task, by identifying the set of legitimate alternatives the annotator can choose from. Annotation guidelines go with a particular annotation scheme and should strictly define the features that the annotator has to annotate, as well as how they should be annotated.

Although there is not a standard way to create annotation guidelines, some

common ground rules exist – see Palmer and Xue (2005) and Pustejovsky and Stubbs (2013). Usually a sound annotation guideline introduces at least some criteria which define the annotation task, alongside a description and some examples. Their aim is to help the annotator understand the criteria. Four examples of annotation guideline for the case of intrinsic human evaluation can be found in Appendix A.

1.4.1 The case of intrinsic human evaluation of NLG systems

The intrinsic human evaluation of NLG systems is an example of an annotation task. Two features that characterise such annotation are: i) the corpus dimension and ii) the (usually) elevated level of subjectivity involved in the annotation task. Unlike other annotation tasks, intrinsic human evaluations are performed with a limited set of sentences. Traditionally, a random set of system outputs, usually in the order of a few hundred or less, are used for such annotation tasks. The other feature is the subjectivity involved in the annotation. Given the wide use and interpretation of natural language, as it is subject to individual human differences, the definition of annotation guidelines becomes particularly difficult. There is a considerable risk of ending up with a set of poorly-defined criteria. We will show in Chapter 4 that, far from being due to a lack of precision in writing the guidelines, the problem seems to be intrinsic to the nature of human language. Sampson (2017) presented the following analogy to explain this point:

> Suppose we wanted to be able to say how large particular clouds are – what volume of space they occupy. Clouds are fuzzy things, so one problem would be what we mean by the volume of a cloud – what exactly should we count as its edge? But even if we adopted some precise definition of cloud boundaries, so that it became meaningful to say that this cloud is exactly N cubic yards in size, not N+1 or N-1, it might still be beyond mankind's

abilities actually to measure clouds so exactly.

An example of this phenomenon – from a study we will describe in detail in Chapter 4– concerns the evaluation of the following question:

Q: *Which NFL team represented the NFC at Super Bowl 50?*

Use the following paragraph to answer this question:

Paragraph: *The Panthers finished the regular season with a 15-1 record, and quarterback Cam Newton was named the NFL Most Valuable Player (MVP). They defeated the Arizona Cardinals 49-5 in the NFC Championship Game and advanced to their second Super Bowl appearance since the franchise was founded in 1995. The Broncos finished the regular season with a 12-4 record, and denied the New England Patriots a chance to defend their title from Super Bowl XLIX by defeating them 20-18 in the AFC Championship Game. They joined the Patriots, Dallas Cowboys, and Pittsburgh Steelers as one of four teams that have made eight appearances in the Super Bowl.*

In our experiment we found divergences in evaluations of the *pertinence* of the question **Q**. The pertinence criterion was defined to measure the degree to which a question has a clear and unambiguous answer within the reference paragraph.[8] Some participants judged **Q** as pertinent, others assessed it as not pertinent. Indeed, the latter thought that to answer **Q** we need to perform an inference, but that we do not have the adequate information to do so. Indeed, in the paragraph, there is no clear connection between the NFL and the NFC.

In Chapter 4 we will present more examples that will make Geoffrey Sampson's analogy clearer.

As a result of the fuzziness of human language, the criteria defined in an in-

[8]The guideline for the pertinence criterion can be found in Appendix A.

trinsic human evaluation guideline can be quite vague and left to the annotators' interpretation. This can result in a low annotator agreement. Defining sentence quality is a difficult matter. Such a definition relies on several features, which can be biased through subjectivity and can change from task to task. Examples of these features are grammaticality, readability, suitability for the task and so on.

We will refer to annotators as judges throughout this thesis, whenever we discuss intrinsic human evaluation. We chose such terminology to emphasise the evaluative aim of the annotation.

1.4.2 Three concepts of data reliability

Once an annotation is performed, a pivotal step is to check the annotation validity. Validity concerns the extent to which the annotation captures what it is intended to capture. Precisely, validity concerns truths, more specifically the "truth" of the phenomenon which is studied.[9] Accordingly, validity allows the possibility of comparing the annotation with a given recognised true standard for that annotation. When (as in most cases of intrinsic human evaluation of NLG systems) a recognised true standard is missing, annotation reliability is measured in lieu of data validity. Reliability concerns the extent to which different annotators agree on the categories annotated. The higher the agreement the more reliable the data. For this reason, although reliability is considered a necessary but not sufficient condition for validity (see for example, Krippendorff (1980), Artstein and Poesio (2008) and Artstein (2017)), data reliability plays a pivotal role in human annotation efforts.

Based on how the agreement is tested, Krippendorff (1980) delineates three types of reliability, which are *stability*, *accuracy* and *reproducibility*.

[9] In this context the "truth" has to be thought "as speaking about the real world of people, phenomena, events, experiences, and actions."(Krippendorff, 1980, page 313).

Stability (or *Intraobserver Agreement* (IA)) is usually measured by the test-retest strategy, which is based on the resubmission, after some time, of some items to the original annotators. That is, annotators are asked to re-assess the same items after some time has elapsed. Comparing the annotations of the same items provides a measure of the annotators' consistency.

Accuracy is measured through calculating the deviations from a given standard, when one exists. More specifically, accuracy "compares the performance of one or more data-making procedures with the performance of a procedure that is taken to be correct"(Krippendorff, 1980, page 216). When the standard taken into account "is truth, or at least what is known to be true"(Krippendorff, 1980, page 216), accuracy turns into validity.

Reproducibility or *Inter-Annotator Agreement* (IAA) is a measure of the extent to which different annotators arrive at the same annotation when working independently. If different annotators, when independently performing the annotation task, consistently make the same annotation decision, then we have strong support for the belief that the phenomena to be annotated are well understood and shared across the annotators. The reproducibility of the annotation is dependent on a well-defined annotation scheme and clear annotation guidelines. Different annotators can perform the same annotation task reaching equivalent (or very similar) results. As shown in Hovy and Lavid (2010), Pustejovsky and Stubbs (2013), Artstein (2017) and Finlayson and Erjavec (2017), where general rules for annotation design are developed, this idea of reliability as reproducibility has become the predominant reliability concept used in any Computational Linguistics (CL) annotation task. Accordingly, guidelines and good practice descriptions for applying IAA in CL annotation tasks have been developed – for example, (Lombard et al., 2002; Artstein and Poesio, 2008; LeBreton and Senter, 2008; Kottner et al., 2011).

In CL, since the work of Carletta (1996), the most common way to measure reliability is using the coefficients of agreement. Specifically, reliability is measured by some form of Kappa statistic. In the next section we briefly present the Kappa statistic, which will be discussed in depth in Chapter 3.

Bayerl and Paul (2007) present a way to integrate the use of the Kappa statistic to the aim of measuring IA and IAA of human annotation data. The authors, providing an example of the case of phonetic transcriptions, suggest the use of Generalizability Theory (GT) (Cronbach et al., 1963) in manual annotation studies to identify "the main source responsible for poor annotation quality".

A further observation about reliability

Krippendorff (1980, page 211) introduced the concepts of stability, reproducibility, and accuracy as "three manifestations of reliability". They are three ways of testing the agreement between annotators, coders, judges or measuring instruments.

As we said before, since Carletta (1996) reliability has been mainly identified with reproducibility. Indeed, Carletta took inspiration from the content analysis community. This is clearly stated in the paper's abstract: "We discuss what is wrong with reliability measures as they are currently used for discourse and dialogue work in computational linguistics and cognitive science, and argue that we would be better off as a field adopting techniques from content analysis" (Carletta, 1996, page, 249). According to Krippendorff (2004) "reproducibility is arguably the most important interpretation of reliability" in content analysis. After Carletta, reproducibility became the main manifestation of data reliability in CL. Carletta suggests the use of some sort of coefficient of agreement as the standardised way to measure reliability, as practised by the content analysis community.

Let us pause for a moment and think about the aim of measuring reliability.

Reliability is measured in lieu of validity. The annotated data are telling us something about the world. If the data are valid, they can be considered true. By virtue of their validity, we can trust in that data and use them to draw conclusions: "Data, by definition, are the trusted ground for reasoning, discussion, or calculation" (Krippendorff, 1980, page 213). In this light, reliability has to be considered as a necessary step for data trustworthiness. For example, Krippendorff (2004) speaks about data reliability as "the sample of data from which the trustworthiness of a population of data is to be inferred". In the same vein Bayerl and Karsten (2011), referring to Artstein and Poesio (2008), wrote "The main reason for the analysis of annotation quality is to obtain a measure of the "trustworthiness" of annotations". Or, in other words, "The intended meaning of reliability should refer to the degree to which the data generated by coders applying a scheme can be relied upon" (Craggs and Wood, 2005, page 290).

Relative to the annotation tasks considered in the present thesis, **we assume that annotation trustworthiness has to be considered as the goal of the annotation reliability study.**

According to Krippendorff (2004) "Reproducibility is about data making, not about coders". However, coders (or annotators or judges) matter in the type of annotation tasks we consider in this thesis. In this kind of annotation, we argue that reproducibility is a too narrow as a principle. We propose to think of the trustworthiness of the annotation in term of whether there is no incompatibility between annotators judgements. Instead of checking if the annotators associate the same labels, we aim to measure the extent to which annotators show compatible annotation behaviours. What we want to avoid is an inconsistency between annotators.

1.4.3 Kappa statistic and their scales of interpretation

Traditionally IA and IAA are measured by some form of the Kappa statistic K. Here we are following the notation used by Carletta (1996). It is worth noting that the K formulation presented captured the most used agreement coefficients from the NLP community, such as Scott's π (Scott, 1955) Fleiss' generalisation of π (Fleiss, 1971) and Cohen's κ (Cohen, 1960).

The common theme of a variety of formulations is that K corrects annotators' agreement by the expected chance agreement:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ (the observed agreement or percent agreement) is the proportion of times the annotators agree, while $P(E)$ is the proportion of times the annotators would be expected to agree by chance.

Kappa statistics are used with some scales of interpretation. For example, Krippendorff (1980) (see Table 1.1) considers *good* any data annotation with agreement in the interval [0.8, 1], *tentative* any data annotation where agreement is in the interval [0.67, 0.8) and to *discard* any data annotation where agreement is below 0.67[10].

AV value	AV interpretation
AV < 0.67	Discard
$0.67 \leq$ AV < 0.8	Tentative
$0.8 \leq$ AV ≤ 1	Good

Table 1.1: Krippendorff scale of interpretation for the Kappa statistic. AV denotes agreement value as determined by some coefficients of agreement.

Given the arbitrary nature of the choice of the numerical intervals that make

[10] Arstein and Poesio (Artstein and Poesio, 2008, page 576) note that: *the description of the 0.67 boundary in Krippendorff (1980) was actually "highly tentative and cautious," and in later work Krippendorff clearly considers 0.8 the absolute minimum value of α to accept for any serious purpose: "Even a cutoff point of α = .800 ... is a pretty low standard".*

up a scale, we can find different scales to interpret K statistics. For example Table 1.2 presents the scale introduced by Landis and Koch (1977).

AV value	AV interpretation
$AV < 0$	Poor
$0 \leq AV \leq 0.2$	Slight
$0.2 < AV \leq 0.4$	Fair
$0.4 < AV \leq 0.6$	Moderate
$0.6 < AV \leq 0.8$	Substantial
$0.8 < AV \leq 1$	Almost Perfect

Table 1.2: Landis and Koch scale of interpretation for the Kappa statistics. AV denotes agreement value as determined by some coefficients of agreement.

Since Carletta (1996), the K statistic and the Krippendorff interpretation scale have been introduced in NLP. Carletta took these metrics from content analysis to "compare results in a standard way across different coding schemes and experiments and to evaluate current developments". Twelve years after Carletta's (1996) publication, the use of the coefficients of agreement in CL was systematically reconsidered and analysed by Artstein and Poesio (2008). Artstein and Poesio discuss the mathematics and interpretation of the K statistic and its use in several CL tasks. In Section 2.2.2 we will see that, more than twenty years after Carletta's work, and more than ten years after Artstein and Poesio's work, the use of the coefficients of agreement is still not an adopted standard in NLG.

More discussion of the agreement coefficients can be found in Chapter 3. More specifically, in Chapter 3, we will present and discuss five popular coefficients of agreement and their interpretation.

1.4.4 Problems with human agreement for the case of NLG

Given the wide use and interpretation of natural language, differences between judges' subjective preference can affect the judges' agreement.

Recent papers – for example, those of Sampson and Babarczy (2008), Lommel et al. (2014) and Joshi et al. (2016) – suggest the inadequacy of the Kappa statistic to analyse the reliability of human evaluation datasets. Such studies show that judges diverge in language annotation tasks due to a range of ineliminable factors such as background knowledge, preconceptions about language and general educational level. Their results suggest the inadequacy of a single number to analyse the reliability of human evaluation datasets. This point is developed by Artstein (2017), where the author argues that a single value can never capture the complexities of a full annotation task. Artstein demonstrates how to use measure of agreement to find data variation, which can be used to have a better understanding of the reliability of human annotations. This is when the data are subject to:

- Diversity in the underlying data.
- Similarity between the labels.
- Differences in the difficulty of individual items.
- Differences between individual annotators and annotators' populations.

The problem of reaching a high IAA in the evaluation phase was encountered in the Question Generation Shared Task Evaluation Challenge (QG-STEC) (Rus et al., 2010). In QG-STEC two tasks, task A and B, were defined. Whereas both shared the output type they partially diverged in the input[11]. Where task A took a paragraph and a target question type as an input, task B took a single sentence and a target question type as an input. Both the tasks were evaluated, through intrinsic human evaluation method, based on the 5 criteria: *relevance*, *syntactic correctness and fluency*, *ambiguity*, *ques-*

[11]The input was divided into two parts, a text type part and a question target type part. Whereas they shared the question target type part (that is, it was specified a target question type, e.g. who, what, where, when etc..), task A and B were different in the text type part.

tion type and *variety*. Quite interestingly the IAA reached in the evaluation phase was low. An attempt to improve the IAA for the task B was done by Godwin and Piwek (2016). Godwin and Piwek define an interactive process where the judges can discuss their opinions about the criteria used in the evaluation. At the end of the evaluation process, repeated three times with three judges, they achieved high IAA with a peak of 0.94 for one of the five criteria used in the evaluation. Nevertheless, their guideline was not tested with judges who were different from the ones used to define them. The result of Godwin and Piwek can be considered as the IAA upper bounds which can be achieved using the guideline defined in their paper.

In the same vein, Sampson and Babarczy (2008) investigate the upper bounds that can be achieved on IAA in the case of English grammar annotation. More precisely "the limits to the potential precision of English grammar annotation" (Sampson and Babarczy, 2008, page 471) are studied. The authors perform their experiment using the SUSANNE scheme (Sampson, 2002) for a parse-tree structure task. From their study they conclude that discrepancies in IAA emerge for three main reasons:

- Violation of an explicit feature of the annotation scheme.
- The lack of a single, unambiguous annotation decision yielded by the scheme, even though the meaning of the text is clear.
- Structural ambiguity in the text.

More generally, Bayerl and Karsten (2011) investigate several factors which affect the IAA score, performing a meta-analytic investigation that involves 96 annotation studies. From their analysis, the authors conclude that at least seven factors affect the IAA values. These factors are: "annotation domain, number of categories in a coding scheme, number of annotators in a project, whether annotators received training, the intensity of annotator training, the annotation purpose, and the method used for the calculation

of observed agreements" (Bayerl and Karsten, 2011, page 699).

1.5 Conclusion

Evaluation is a critical phase for the development of NLG systems. Two main methodologies, intrinsic and extrinsic, are used for this aim. Intrinsic automatic evaluation metrics are difficult to interpret and have been proven not to correlate with human judges. Intrinsic human evaluation and extrinsic methodologies are arguably the most suitable evaluation methods for the NLG efforts. Nevertheless, extrinsic methodologies often involve a prohibitive cost in time and money. This makes the intrinsic human evaluation the most relevant standard for NLG evaluation systems. However, human evaluation faces the problem of high variability in language interpretation, which can result in low agreement and lack of reliability. Since human evaluation is an annotation task, reliability plays a pivotal role in the validity of any human intrinsic evaluation efforts. Accordingly, the reliability of intrinsic human evaluations turns out to be an urgent topic to deal with in the field of evaluation of NLG systems. This thesis is dedicated to this issue.

Chapter 2

Analysis of the recent use of evaluation methodologies in NLG

The aim of this chapter is to set out the evidence from two studies into evaluation practices. These provide a justification for the research question that is central to this thesis.

The first study we present investigates the evaluation practices used in NLG. We performed this analysis in Automatic Question Generation (AQG) as a representative subtask of NLG. For this purpose, we selected a sample of 37 papers with a publication date between the years 2013 - 2018.[1] Our analysis shows a variegated evaluation landscape which highlights the lack of a standardised approach for evaluating AQG systems. Both intrinsic human and automatic methodologies (which are the most adopted) are used in such a way. This makes a systematic comparison across generation systems difficult. Indeed, we found an assorted use of automatic metrics, human

[1]The complete list of the papers analyzed can be found in Amidei's 2018 repository.

evaluation design and datasets adopted in the evaluation phase.

Notably, the problem of the reliability, measured by IAA, of the intrinsic human evaluation emerges from the analysis of the 37 papers. According to the Krippendorff scale for interpreting agreement coefficients (Table 1.1), most of the reported evaluations quoted IAA measures which should be categorised as "discard". Nevertheless, conclusions from them were drawn without a discussion of the reliability of the result.

The second study we present in this chapter delves deeper into existing practices of reliability reporting in NLG. We performed this study by analysing papers published in NLG specialist conferences between the years 2008 - 2018 (135 papers in total).[2] The main findings of our analysis are:

1. We found little use of reliability studies in the evaluation phase;
2. We found shortcomings and oversights in reporting the reliability studies, measured as IAA;
3. The majority of the papers that report the reliability studies reached a low value of IAA.

We explain how these points have been the basis for this thesis in the conclusion of the chapter.

This chapter is made up of three sections. In Section 2.1 we present the first study we performed. It stands for a general level analysis about the evaluation methodologies used in NLG. The analysis is based on a selection of papers about AQG published between January 2013 and June 2018, the latter being the month in which we carried out this research. The focus of this section is to understand the actual practices in using both intrinsic and extrinsic evaluation methodologies. It allows us to point out shortcomings and oversights which need to be addressed.

[2]The complete list of the papers analysed can be found in Amidei's 2019 repository.

From Section 2.1 the problem of intrinsic human evaluation reliability stands out. Given the priority of this problem, it is further investigated in Section 2.2. Indeed, this section presents the results of our analysis about the use of IAA in the intrinsic human evaluation of NLG systems.

In the last Section 2.3, we summarise the present chapter and we explain how it underpins the upcoming chapters.

2.1 Evaluation methodologies for AQG

In this section we present the findings of our analysis based on a sample of 37 papers about AQG. We focus our analysis on two dimensions: *intrinsic evaluation methodology* and *extrinsic evaluation methodology*.

2.1.1 Criteria for papers selection

For this task, we examined the papers in the ACL anthology[3] with a publication date between the years 2013-2018. Table 2.1 shows the distribution of the papers involved in the current study across this period.

Year of publication	Number of papers
2018 (January-June)	7
2017	13
2016	9
2015	5
2014	1
2013	2

Table 2.1: Number of papers per year describing question generation systems.

We used the single term *"question generation"* as the search term with the search engine provided in the ACL Anthology website. From the papers that were returned by this query, we focused only on those papers that were

[3] `http://aclweb.org/anthology/`.

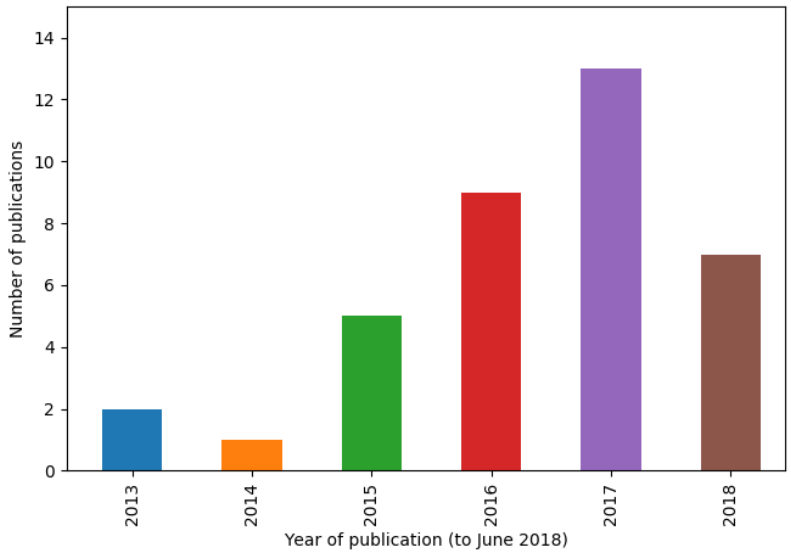


Figure 2.1: Number of papers on AQG published by year in the ACL anthology.

about question generation systems. This gave us 37 papers to analyse, of which 36 were published in conference proceedings and 1 was published in a journal. A complete list of papers used in this study can be found in Amidei's 2018 repository. The number of papers by year is given in Table 2.1 and illustrated in Figure 2.1. Figure 2.1 shows the rapid increase in publications in this area in recent years. Note that this study of the literature was carried out in June 2018, and so several major conferences in this area (including ACL, INLG, EMNLP and COLING) had not taken place.

Before looking more closely at the publications involved, let us introduce the AQG tasks studied in these papers. AQG is the task

> of automatically generating questions from various inputs such as raw text, database or semantic representation (Rus et al., 2008).

Publication type	Journal or conference name	Number of papers
Conference proceeding	INLG	7
	ACL	7
	NAACL-HLT	6
	Workshop on Innovative Use of NLP for Building Educational Application	6
	EMNLP	3
	IJCNLP	1
	EACL	1
	SIGDIAL	1
	COLING	1
	NLPTEA	1
	RANLP	1
	Workshop on Representation Learning for NLP	1
Journal	Computational Linguistics	1

Table 2.2: Number of papers per conference proceedings or journal.

The above definition, adopted by the AQG community, leaves room for researchers to decide what kind of questions and input to work with. Following Piwek and Boyer (2012) three aspects can characterise an AQG task: the form of the input, the form of the output, and finally the relationship between the input and the output. The 37 papers we analysed can be divided into the following three categories:

1. *Input*: text;
   Output: text;
   Relationship: the output question is answered by the input text or the output question asks a clarification question about the input text.

2. *Input*: knowledge-based structured data (for example triples ⟨subject, object, subject/object relation⟩);
   Output: text;
   Relationship: the output question is answered by the information

structure in the input.

3. *Input*: image, or image and text, or image segmentation annotations;
   Output: text;
   Relationship: the output question is answered by the information pictured in the input.

For the sake of simplicity we will denote with *Text2Text* the task expressed by category 1, *Kb2Text* the task expressed by category 2 and finally *Mm2Text* the task expressed by category 3, where *Mm* is short for "Multi-modal". Within each category, we find papers with different aims. We include these in the following list, where the number in brackets shows how many papers fall into that category:

1. *Text2Text* (30)
   - Web searching (1)
   - Chatbot component (1)
   - Creation of comparative questions related to the input topic (1)
   - Clarification questions (1)
   - Question Answering (5)
   - Dataset creation purposes (1)
   - Educational purposes (9)
   - AQG general purposes (11)
2. Kb2Text (4)
   - Question Answering (1)
   - Dataset creation purposes (1)
   - Educational purposes (1)

- AQG general purposes (1)

3. Mm2Text (3)
    - Data augmentation Visual Question Answering (VQA) purpose (1)
    - AQG general purposes (2)

Regarding the papers in the *Text2Text* category, we found some variety in the diverse types of output. Although in the majority of cases, the system's output was an interrogative sentence, there are 5 papers in which the output is a "fill the gap" question, 3 papers where output is a multiple choice question (with its associated set of distractors) and 3 papers in which the output is a question/answer pair. Also, in both the *Kb2Text* and *Mm2Text* categories, there is 1 paper each in which the output is a question/answer pair. We also note that 1 paper in the *Text2Text* category developed a question generator which takes a paragraph of text and an associated answer as input. In this case, the generated question must be answered by the answer given in the input. We conclude this section by specifying that *AQG general purposes* mean that the system was not tied to a particular domain or task-dependent setting, whereas *Question Answering* means that the AQG system was developed to be used in the Question Answering task.

2.1.2 A general overview

Table 2.3 shows the evaluation methodologies used in the papers that we examined. With respect to the frequency of the use of intrinsic compared to extrinsic methods, Table 2.3 confirms the trend identified by Gkatzia and Mahamood (2015). Gkatzia and Mahamood found that 74.7% of the papers used the intrinsic evaluation method. In our analysis we found that 83% of the papers used this methodology. However, we note that with respect to the results of Gkatzia and Mahamood (2015), we have an inverted trend

Evaluation methodologies	Number of papers			
	Text2Text	Kb2Text	Mm2Text	Total
Intrinsic human only	13	1	-	14
Intrinsic automatic only	9	-	1	10
Extrinsic (human) only	2	-	-	2
Intrinsic human & Intrinsic automatic	3	2	2	7
Intrinsic human & Extrinsic (human)	2	-	-	2
Intrinsic automatic & Extrinsic (automatic)	1	-	-	1
Intrinsic human & Intrinsic automatic & Extrinsic (automatic)	-	1	-	1

Table 2.3: Evaluation methodologies used.

between the use of an extrinsic method compared to both intrinsic and extrinsic. Indeed, Gkatzia and Mahamood found that 15.2% of the papers used extrinsic methods, versus the 6% we found in our analysis. 10.1% of the papers used both methodologies, whereas our analysis shows that 11% of the papers use a combination of both.

Furthermore, our analysis confirms the trend between the use of automatic compared to human intrinsic evaluation methodologies. Gkatzia and Mahamood (2015) report that in 45.4% of the cases human evaluation is used, whereas in 38.2% of the cases automatic evaluation was adopted. Similarly, our analysis shows that amongst the papers that prefer intrinsic evaluation methods, 45% used human evaluation, 32% used automatic evaluation and 23% used both human and automatic evaluation.

Table 2.1 shows that in the period since 2016, there has been a considerable increase in the number of publications in this area. It therefore makes sense to ask whether this increase has been accompanied by a change in the evaluation methodologies used. Table 2.4 shows how the range of evaluation methodologies used has changed. Between the years 2013 - 2015 only intrinsic evaluation methodologies were used – with 75% of papers using human evaluation, 12.5% using automatic evaluation and 12.5% using both methodologies. Between the years 2016 - 2018 extrinsic evaluation methods

Evaluation methodologies	Number of papers	
	2013-2015	2016-2018
Intrinsic human only	6	8
Intrinsic automatic only	1	9
Extrinsic (human) only	-	2
Intrinsic human & Intrinsic automatic	1	6
Intrinsic human & Extrinsic (human)	-	2
Intrinsic automatic & Extrinsic (automatic)	-	1
Intrinsic human & Intrinsic automatic & Extrinsic (automatic)	-	1

Table 2.4: Variation of the evaluation methodologies used between 2013 - 2015 and between 2016 - 2018.

were also introduced.[4] Indeed, although most of the papers in this period (79%) used intrinsic evaluation methods, 7% of papers used extrinsic evaluation methods and 14% used both the methodologies. We can also see a change in the tendency to use intrinsic methods. Between the years 2016 - 2018, 35% of the papers used human evaluation (a decrease of 40% from the years between 2013 - 2015), 39% of the papers used automatic evaluation (a 26.5% increase on the years between 2013 - 2015) and 26% of the papers used both methodologies (a 13.5% increase on the years between 2013 - 2015).

2.1.3 Intrinsic automatic evaluation

Table 2.5 presents a list of automatic metrics used in the papers studied in the present research.[5] From our analysis the most used automatic metric is BLEU (Papineni et al., 2002) followed by METEOR (Banerjee and Laviel, 2005). Note that Table 2.5 only describes those that use the specified metrics; other papers use metrics that are defined for the specific aims described

[4]This shows an interesting fact. As we have seen before, based on the analysis of Gkatzia and Mahamood (2015), between the years 2005 – 2014 both intrinsic and extrinsic evaluation were used in NLG. From our analysis, follow that the AQG community aligns with these results from 2016.

[5]Table 2.5 presents the number of time each metric was used. Accordingly, the papers that use more than one metric are counted more than one time.

in the paper that introduces them.

Evaluation methodologies	Number of papers			
	Text2Text	Kb2Text	Mm2Text	Total
BLEU (Papineni et al., 2002)	8	3	2	13
METEOR (Banerjee and Laviel, 2005)	4	2	1	7
ROUGE (Lin and Och, 2004)	3	1	-	4
Precision	4	-	-	4
Recall	4	-	-	4
F1	4	-	-	4
Accuracy	2	0	1	3
ΔBLEU (Galley et al., 2015)	-	-	1	1
Embedding Greedy (Rus and Lintean, 2012)	-	1	-	1
Others	5	-	-	5

Table 2.5: Automatic metrics used.

In our survey we found that 31% of the papers used just a single metric, whereas the other 69% used more than one. The mean is 2 metrics per paper, with a minimum of 1 metric (6 papers) and a maximum of 5 metrics (1 paper). In almost 50% of cases (9 papers), 3 metrics were used. We noticed that only a single paper used an embedding based metric. In a majority of studies, word-overlap based metrics were used.

To the best of our knowledge, the area of AQG is currently missing a study which aims to verify the correlation between human judgement and automatic metrics.[6] Such research would have two merits: on one hand, this kind of meta-evaluation study would give a better characterisation of the general problem. On the other hand, the research could provide guidance to

[6]Yuan et al. (2017) raise some doubts about the capacity of BLEU to effectively measure the quality of AQG systems used in *Text2Text* tasks.

researchers about which metric is most appropriate in evaluating a particular system. Research in AQG would benefit from a systematic study that aims to clarify the relationship between different evaluation methodologies.

2.1.4 Human evaluation

Among the various human evaluation methodologies, absolute annotations is most common. Only two papers used a relative annotation methodology. In one paper the human annotators are asked to assess pairwise preference between multiple questions. In the other paper they are asked to assess pairwise preference between a pair of questions. These are one human generated question and one automatically generated question, and the annotators are to assess which one is automatically generated (or which one is the human generated). The former paper also used absolute annotations.

Absolute annotation methodologies typically ask annotators to use rating or Likert-style scales to record their judgements. In our analysis, we found that 56% of the papers used some kind of numerical scale. For example, human judges were often asked to assess the grammaticality of a question on a numerical rating scale from 1 (worst) to 5 (best). On the other hand, 44% of the papers used a graphic rating scale. In these cases, human judges were typically asked to classify the questions in some categories such as coherent, somewhat coherent or incoherent.[7] The number of categories used in the rating or Likert-style scales by the papers that adopted absolute annotations methodologies are shown in Table 2.6.

Only 3 papers used more than 1 type of scale in the evaluation. One of these uses a free scale in which the annotators have to choose a positive integer to count the inference steps necessary to answer a question.

Table 2.6 shows that the two most common number of categories used in the

[7]For more details about the difference between numerical rating scale and graphic rating scale we refer to Amidei et al. (2019).

Number of categories	Number of papers			
	Text2Text	Kb2Text	Mm2Text	Total
2	6	-	-	6
3	6	-	2	8
4	1	1	-	2
5	8	1	-	9
7	-	1	-	1

Table 2.6: Number of categories used in the Likert or rating scales.

rating or Likert-style scales are 3 and 5. In a recent paper, Novikova et al. (2018) suggest that the use of a continuous scale and relative annotations can improve the quality of human judgments. Although we found 2 papers that used relative annotations, we did not find any papers that use a continuous scale.

Another interesting point is the number of judges used in the evaluation. This number varies a lot from paper to paper. We found a minimum of 1 judge (2 papers) to a maximum of 364 judges (1 paper). Taking the papers which provided information on the number of judges used (24 papers), and removing five papers that used 53, 63, 67, 81 and 364 judges, we found out that the mean number of judges used was almost 4. The most common number was 2 judges, used by 29% (7 papers) of the papers. 3 judges were used by 17% (4 papers) and 4 judges were used by 13% (3 papers). The others paper used 5, 7, 8 or 10 judges.

There is a similar breadth to the number of output questions used (that is, the questions generated by the systems), and the criteria (that is, the question features to be checked) used in the evaluation. The number of questions ranged from a minimum of 60 questions (1 paper) to a maximum of 2186 (1 paper). Amongst those papers which provide this information (17 papers out of 28), we found that the mean number of questions used per paper was almost 493. 7 papers did not report this information, whereas

2 papers reported information about the amount of data from which the questions were generated, without giving the exact number of questions used for the evaluation.

Regarding the criteria used, we noticed that 35% of the papers (8 studies) used an overall quality criterion, that is, a single criterion which was used to evaluate the questions' overall quality. On the other hand, 52% of the papers (12 studies) used specific criteria, for example, question grammaticality, question answerability, etc. A full list of these criteria is shown in Table 2.7. 13% of the papers (3 studies) used both specific criteria and an overall criterion. As Table 2.7 shows, there is a wide assortment of criteria used across the set of collected papers.

As we can see from Table 2.7, the specific criteria are mainly used in the *Text2Text* task. Just two criteria are used in the *Kb2Text* task and none in the *Mm2Text*, where an overall quality criterion was preferred. We note that some criteria, for example timing or importance, are specific to one of the aims of the paper in which they are used. Indeed, as shown in the Section 2.1.1, we can find different aims behind the papers' motivations. We note that among the papers analysed here, often only little information is provided about the evaluation guidelines. We cannot exclude the possibility that, given the evaluation guidelines, some of the criteria presented in Table 2.7 could collapse together. That is, it is possible that different researchers use different names to check the same question feature.

Table 2.8 supplies an overview about the IAA reached in the human evaluations. We note that 54% of the papers (14 studies) did not supply this information. Only one of the two papers that used relative annotations reported the agreement between judges. In that paper, Fleiss' κ (Fleiss, 1971) was used to measure the IAA reached between 3 to 5 judges. The results, for 3 batches with different judges and questions, were 0.242, 0.234

Criterion used	Number of papers			
	Text2Text	Kb2Text	Mm2Text	Total
Grammaticality	7	-	-	7
Semantic correctness	4	-	-	4
Answer existence	3	-	-	3
Naturalness	2	1	-	3
Question type	3	-	-	3
Clarity	3	-	-	3
Discriminator quality	3	-	-	3
Relevance	2	-	-	2
Correctness	2	-	-	2
Well-formedness	1	-	-	1
Key selection accuracy	1	-	-	1
Corrected retrieval	1	-	-	1
Fluency	1	-	-	1
Coherence	1	-	-	1
Timing	1	-	-	1
Inference step	1	-	-	1
Question diversity	1	-	-	1
Importance	1	-	-	1
Specificity	1	-	-	1
Predicate identification	-	1	-	1
Difficulty	1	-	-	1
Overall criterion	7	2	2	11

Table 2.7: Criteria used.

and 0.182. Table 2.8 presents the IAA results reported by the papers that used absolute annotations methods. Between the papers that reported this information, we found that the IAA was measured in 26 cases and 9 of these were measured with two different coefficients, for a total of 35 IAA values. The agreements were measured for specific criteria or for the overall quality criterion. In one case the agreement over all the criteria was reported. It is notable that the agreement reached in the various evaluations is generally quite low. Indeed, according to existing scales of IAA interpretation most evaluations fail the reliability test. For example, taking into consideration the Krippendorff scale of interpretation (Table 1.1), among the papers that

Coefficient used for calculate IAA	Number of criteria	Mean	Min.	Max.
Cohen's κ	14	0.46	0.10	0.80
Krippendorff's α	2	0.14	0.05	0.23
Fleiss's κ	4	0.45	0.33	0.62
Pearson's r	4	0.71	0.47	0.89
Percent agreement	9	0.80	0.50	0.91
κ, type no specified	2	0.08	0.08	0.09

Table 2.8: Measures of IAA.

report IAA, very few evaluations should be considered reliable. More specifically, 43% (15 out of 35) of the IAA values is higher than 0,67 (which is the threshold set by Krippendorff for tentative conclusions) and only 23% (8 over 35) of the IAA values were greater than or equal to 0.8 (which is the threshold set by Krippendorff for reliable conclusions).

Checking the agreement for number of judges we found that in the case with 364 judges the IAA, measured for two criteria and a κ which type was not specified, was between 0.08 and 0.09. We found only 2 cases for 5 judges, which reported a value of 0.05 for Krippendorff's α (Krippendorff, 1980) and a percent agreement of 0.89. Two papers used 4 annotators altogether: one reported a value of Krippendorff's α of 0.236, with the other reporting a Pearson's r (Witte and Witte, 2017) of 0.71. Another paper used 3 evaluators and the Fleiss's κ to measure the IAA for 4 criteria. The results are reported in Table 2.8. All other papers reporting an IAA measure were in evaluations that used 2 judges.

There are sometimes attempts to design the experimental methodology to improve the level of IAA. In order to improve the agreement, one paper collapsed two score classes into one, whereas two papers allowed a difference of one score between the annotators' rating. Two examples of the latter case are the maximum value for Cohen's κ and the maximum value for the percent agreement reported in Table 2.8.

2.1.5 Extrinsic evaluation

As shown in Table 2.3, extrinsic evaluation methodologies are rare in the area. As reported by Gkatzia and Mahamood (2015) this is generally true for NLG tasks. Amongst the papers that have chosen to use this kind of evaluation technique, human participants were used in 4 times out of the 6. In the papers where human participants were not used, the Question Generation (QG) system was tested as a component of a Question Answering (QA) system. The performance was evaluated by checking the difference between two implementations of the QA system. One was without the use of the QG system and the other with the use of the QG system. The aim of those papers was to improve QA systems by creating more accurate question/answer pairs to be used for training purposes.

Because of the different tasks at play, the other papers used humans in different ways. We can find tasks such as: "Answer the generated questions or use the generated questions in a web page and then answer a survey about the utility of those questions". Or also: "Engage in a conversation with a chatbot which involves a question-based dialogue, and then rate the conversations".

Also in this case, the number of humans involved in the evaluation varies from paper to paper, ranging from 2 to 81. In contrast to the case of intrinsic human evaluation, in these cases the IAA is not reported. We note that human agreement in extrinsic evaluation is not as relevant as in the case of intrinsic evaluation. In the case of extrinsic methods, the evaluation aim is to check whether the generated questions fulfil the task for which they were generated. To test this, humans need to use those questions in real contexts. Now, humans make use of questions in several ways, and similarly, they answer questions in diverse ways. For this reason, humans are unlikely to reach equivalent results in a real context of language use.

2.1.6 Preliminary conclusions

Although systems and tools have been developed in the AQG area over the last few years (as illustrated by Figure 2.1), this has not been accompanied by similar improvements in evaluation methodologies. Indeed, with the exception of the Shared Task Evaluation Challenge (QG -STEC) (Rus et al., 2010), no attempts have been undertaken to introduce a common framework for evaluation that allows for comparisons between systems. The variety of evaluation methodologies, as brought to light by the present work, demonstrates how difficult it is currently to check question quality across generation systems. This prevents us from understanding the actual contributions that are made by new generation systems, which are being introduced ever more frequently.

The problem of having a high degree of variation in methodologies is compounded by the use of different datasets in the evaluation phase (see Table 2.9). The use of a common dataset for evaluation — as suggested for the NLG Shared Task Evaluation Campaign STEC (Gatt and Belz, 2009) — could remove bias coming from the training phase. This is particularly true for generation systems that use machine learning techniques.[8] If we want to understand the degree to which a system advances the state of the art, we need to compare different systems on the same dataset, or, better yet, a set of datasets that use the same evaluation methodologies.

In the area of AQG, both intrinsic and extrinsic methodologies are used in such a way that prevent a sensible comparison across generation systems. We found a minimal and variegated use of extrinsic methodologies, as well as an assorted use of automatic metrics, human evaluation design and datasets adopted in the evaluation phase.

[8] We note that the high variability in the dataset used in the evaluation phase is also due to the variation in the papers' motivations.

Tasks	Dataset or source of test articles
Text2Text	SQuAD; MS-MARCO; WikiQA; TriviaQA; TrecQA; Wikinews; Penn Treebank; QG-STEC datasets; StackExchange; Wikipedia; OMG! website; Project Gutemberg; ReadWorks.org; Engarde corpus; CrunchBase; Newswire (PropBank); textbook from OpenStax and Saylor; not specified TOEFL book; not specified science text books; not specified course Web page; not specified news articles; not specified teachers articles; 40 people's personal data.
Kb2Text	Ontology documenting K-12 Biology concepts; SimpleQuestions; Freebase; WikiAnswers.
Mm2Text	COCO-QA; COCO-VQA; IGC_{Crowd}; Bing; COCO; Flickr.

Table 2.9: Dataset used.

There is scope for more extrinsic evaluation, which can "provide useful insight of domains' need, and thus they provide better indications of the systems' usefulness and utility" (Gkatzia and Mahamood, 2015, page 60). Unfortunately, as we have seen in our analysis, extrinsic evaluations are not yet widely used. Moreover, as suggested in Reiter and Belz (2009) and Reiter (2011), extrinsic methodologies are often accompanied by a prohibitive cost in time and money.

Automatic evaluation metrics can be thought of as a technique to provide a way to standardise the evaluation. Nevertheless, they are difficult to interpret and do not correlate well with human judges (Novikova et al., 2017; Reiter, 2018). This is especially true in areas such as AQG, where a systematic comparison of human and automatic evaluation is missing. This makes it difficult to understand the extent to which automatic metrics capture the systems' quality.

Our analysis shows some shortcomings regarding intrinsic human evaluation. This includes the lack of a shared use of criteria and scales/categories. Lately, similar findings were presented in Van der Lee et al. (2019). More importantly, the problem of evaluation reliability emerges from the papers examined here. In those studies where it has been reported, the reliability, measured as IAA, is generally low. Since human evaluation is an example of an annotation task, reliability plays a pivotal role in the validity of any human intrinsic evaluation efforts. Because of the importance of this point, we decided to analyse the use of reliability study in NLG in depth. To do so, we focused our attention on the use of IAA in papers published in NLG specialist conferences over the years 2008 - 2018 (135 papers in total). We present this study in the next Section 2.2.

2.2 A deep analysis about the use of IAA in NLG evaluation task

In this section we present the findings of our analysis, which aims to check how IAA is used in the human evaluation of NLG systems. The analysis is based on 135 papers.[9] Let us begin by explaining the criteria used for the selection of the papers.

2.2.1 Criteria for papers selection

The first decisions we faced when we began this study was how to select the papers to be analysed, how to retrieve them and which timespan to select. Regarding the publication years, we decided that the interval from 2008 to 2018 would be a good timespan. Indeed, ten years allows the collection of a good quantity of data and allows one to take into account the change marked by the neural networks revolution — usually considered 2012 (Parloff, 2016)

[9] A complete list of the papers can be found in in Amidei's 2019 repository.

—, which mark a watershed in the IA community in the use of methods and techniques.

We decided to use the ACL Anthology to retrieve the papers (`https://www.aclweb.org/portal/`). Although the ACL Anthology does not include many relevant studies, it is a very large dataset which collects published NLP papers. Regarding the first problem, we thought that a better solution could be to select the papers published in the proceedings of the NLG community conferences. Indeed, we thought that an NLG task-based research was in danger of giving too much weight to some tasks and not enough weight to other tasks. However, selecting the papers from the NLG conference proceedings allows us to consider specific tasks which are considered by the community as a proper NLG task. Eventually, we decided to select the papers to be analysed from the Special Interest Group on Natural Language Generation (SIGGEN) webpage (`https://aclweb.org/anthology/venues/inlg/`) hosted by the ACL Anthology website. A complete list of the venues is given in Table 2.10, which shows the number of papers published in each conference and the number of papers selected for our analysis.

We are aware that such a choice excluded certain NLG papers published in other important conferences – for example, ACL, EMNLP, NAACL, COLING, etc. – and journals. However, our study is based on 526 papers. This gives us a fair quantity of papers from which to draw a faithful snapshot of the use of the IAA in the evaluation of NLG tasks. Indeed, the corpus selected focuses not only on end-to-end generation systems but also on its components, such as referring expression generation, surface realiser, etc.

Once we had selected the papers to analyse, we had the problem of deciding which papers could be considered in our study. Because the aim of our work was to analyse the use of the IAA in the evaluation phase by the NLG

Conference Years	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 + SIGDIAL (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 2.10: Years and conference venue (number of papers in parentheses).

community, we decided to focusing on the following features:

1. the paper needs to have a human study;

2. the study must be an intrinsic human evaluation study (we did not

take into account different annotation tasks);

3. the study must allow the measure of the IAA (for example, we did not consider papers in which the human evaluation was done with open questions or papers whose human evaluation was a manual author inspection, unless such an inspection allowed the study of the IAA. Likewise, we did not take into account papers that use extrinsic evaluation methodology. However, we considered papers whose extrinsic evaluation methodology was followed by a survey which allows the study of the IAA, for example, surveys done with rating scale questions).

Considering the three features above we ended up with a corpus of 135 papers on which we performed our analysis.

2.2.2 10 years of IAA in evaluation of NLG systems

The main findings of our analysis are:

1. We found minimal use of reliability studies in the evaluation phase.
2. We found shortcomings and oversights in reporting the IAA studies, and consequent lack of a common practice in the use of IAA.
3. Generally, the level of IAA reached is low.

Point 1: Minimal use of reliability studies

The first thing that stands out in our analysis is the small number of papers which compute IAA to validate the evaluation results. Indeed, of the 135 papers in our study, just 18% (24 papers) report information about the IAA. While 20 papers use one coefficient to measure the IAA, 4 papers use two different coefficients to fulfill this aim. Most of the papers reporting IAA (i.e. 67%) were published in the final two years (2016-2018) of the period we covered. This may signal that there is a positive trend towards more

reporting of IAA values.

Point 1 underlines a shortcoming of NLG human evaluation tasks. When human evaluations are performed, it is good practice to verify the reliability of the evaluations. Without a reliability study there are no solid reasons to accept the conclusions from an evaluation. In Chapter 5 we suggest that correlation coefficients and agreement coefficients should be used together to obtain a better assessment of the evaluation data reliability.

Point 2: Shortcomings with reporting IAA studies

Regarding point 2, the following shortcomings have been identified. Papers often:

- do not report the names of the coefficients used;
- do not report sufficient detail about the experiments used to collect the data;
- use a coefficient that is not suitable for the data collected;
- do not report the number of items on which the IAA study is performed;
- do not report whether the annotators were performing the evaluation independently or not;
- do not report the scale used to interpret the IAA values, and when reported do not discuss the results accurately.

More specifically, between the papers that report the IAA, 37% of the papers (9 works) use an IAA coefficient that is not suitable for the data collected. For example, the use of Fleiss' κ coefficient for data whose level of measurement is interval. Related to this point, we note that often the researchers do not report in sufficient detail the experiment used to collect the data, which

can also give information about the data's level of measurement – that is, whether the data are nominal, ordinal, intervals or ratios.[10] Across the papers we studied, such information had to be deduced from the statistic used for analysing the data.[11]

From Table 2.11 we can see that although discouraged by previous work – see for example, Krippendorff (1980), Craggs and Wood (2005) and Artstein and Poesio (2008) – percent agreement is the coefficient used the most. Indeed, it is reported for 25% of the works (7 papers). It is followed by Krippendorff's α (Krippendorff, 1980) and Fleiss's κ (Fleiss, 1971). Both coefficients were used in 5 papers each. Three papers do not report the name of the Kappa statistic used. Because each metric is different, reporting the exact coefficient used in the analysis would help the readers to better understand the data reliability and the evaluation results.

Few papers discuss the interpretation of the IAA for their evaluation. Between the papers that report the IAA, just 20% of the papers (5 works) make implicit or explicit reference to the interpretation scales used. The IAA interpretation scales reported by these papers are the Krippendorff scale (Krippendorff, 1980, see Table 1.1) and the Landis and Koch scale (Landis and Koch, 1977, see Table 1.2).

In almost every paper we analysed, the number of items used for the IAA studies was not reported. Likewise, there were few cases in which it was reported whether the annotators worked independently.

Finally, we also note that the terminology used is not shared across the analysed papers. Some examples are: reliability, agreement, inter-evaluator agreement, pair-wise agreement, inter-annotator agreement, inter-assessor

[10] For more details about these concept we refer to Appendix 3.1.4.

[11] We note that this is an imperfect, although sometimes the only possible, way to deduce the data level of measurement. Indeed, researchers can use the wrong statistic to analyse the data, which results in a distorted image of the data level of measurement.

agreement, inter-rater reliability and inter-coder agreement.

Chapter 3 is devoted to best practice in performing and reporting a reliability study as measured using coefficients of agreement.

Point 3: Low IAA values

Table 2.11 shows a tendency also found in other work; see for example (Craggs and Wood, 2005; Lommel et al., 2014; Liu et al., 2016; Sedoc et al., 2018) and the supplementary material of (Reiter, 2018).[12] The trend is that in human evaluation of NLG systems the IAA values reached are relatively low. Following the Krippendorff scale of IAA interpretation (Krippendorff, 1980) – which considers the threshold 0.67 as the minimum to be reached to get a reliable set of data (see Table 1.1) – most of the evaluations should be discarded. The problem of how to interpret IAA values is

Coefficient	# used	Mean	Min.	Max.
Percent agreement	7	0.69	0.44	0.94
Cohen's κ	4	0.40	0.10	0.88
Krippendorff's α	5	0.62	0.37	0.90
Fleiss's κ	5	0.53	0.29	0.78
Pearson's r	2	0.42	0.20	0.71
Kendall's W	1	0.61	0.47	0.76
Weighted κ	1	0.07	0.07	0.07
κ, type no specified	3	0.57	0.32	0.77

Table 2.11: Mean, minimal and maximum IAA value per coefficient. *# used* means the number of times that a coefficient was used in total across the papers. In each paper each coefficient was used to measure the annotators' agreement about one or more questions or criteria.

an intriguing and complicated one. Artstein and Poesio (2008) describe this as "the most serious problem with current practice in reliability testing". As noted by Krippendorff (1980, 2004), Craggs and Wood (2005) and Hovy

[12] We note that for the κ coefficients which are "no better specified" the mean measure is not appropriate. Indeed, they could be different κ coefficients. However, we chose to report the mean for uniformity reasons. It worth saying that such a choice does not affect the theoretical point here presented.

and Lavid (2010), the choice of IAA interpretation scale is arbitrary and task-dependent. The reduction of a statistical test interpretation to a simple number, whilst common, can be arbitrary and accordingly give us little information.[13] For example, Artstein (2017) shows that a single label is not sufficient to give a deep understanding of the reliability of an annotation. In Chapter 3 we will present a new interpretation approach to the scale used for interpreting the coefficients of agreement as introduced in Gwet (2014).

Point 3 also reveals a big issue in the area. Indeed, the main purpose of IAA is to check the reliability of the annotated data. Following the existing scales of IAA interpretation, for example those of Krippendorff (1980) and Landis and Koch (1977), most of the evaluations should be discarded because they are unreliable.

In Chapter 4 we will analyse in detail the phenomena of low IAA values for the case of intrinsic evaluation of NLG systems.

2.3 Conclusion

In this chapter we presented two studies. Both are systematic literature reviews of NLG papers.

The first study investigates the evaluation methodologies used in AQG in the years between 2013 - 2018. Our overview shows a variegated evaluation landscape which illustrates the lacking of a shared approach to the evaluation phase. Both intrinsic and extrinsic evaluation methodologies were analysed and some of their shortcomings were presented and discussed. Between these, the problem of the reliability of the intrinsic human evaluation is arguably the most urgent to be addressed. To date, intrinsic human evaluation methodologies are the most adopted evaluation methodologies in the area, and they can be considered as the standard for NLG evaluation sys-

[13] Lately, this point has been raised also for the $p - value$ (Wasserstein et al., 2019).

tems. Nevertheless, without a proper reliability study they should not be used. The investigation of the reliability of intrinsic human evaluation was the focus of our second study.

In the second study of this chapter, we presented a snapshot of the reliability studies (as measured by the IAA) of intrinsic human evaluation tasks in the NLG area. Our investigation was based on an analysis of papers published over the last 10 years in NLG-specific conferences (in total 135 papers). The main findings of our second study can be summarised in the following three points:

1. We found little use of reliability studies in the evaluation phase.
2. We found shortcomings and oversights in reporting the reliability studies, measured as IAA.
3. The majority of the papers that report the reliability studies reached a low value of IAA.

Point 1 underlines a common shortcoming of NLG intrinsic human evaluation tasks — indeed, without a reliability study there are no solid reasons to accept the conclusions from an evaluation. Such a deficiency legitimates the aim of this thesis: *to enhance, in the NLG community, awareness about the need to handle the problem of intrinsic human evaluation reliability, and suggest a way to carry it out.*

We will provide an in-depth analysis in the following chapters, and we will suggest solutions for the issues raised in points 2 and 3. Chapter 4 is dedicated to investigating the reasons for point 3, i.e. the generally low IAA found in NLG evaluations. The result of our investigation of Chapter 4, will drive Chapter 5 and Chapter 6. In these two chapters we propose a new set of methods for identifying judges' bias and reporting reliability for the case of human intrinsic evaluation of NLG systems.

Regarding point 2, much has been said in previous work about the development of guidelines and good practice descriptions for applying IAA, for example Krippendorff (1980), Artstein and Poesio (2008), LeBreton and Senter (2008), Kottner et al. (2011), Gwet (2014) and Artstein (2017). To make this thesis self-contained, in Chapter 3, we suggest good practices for using and reporting coefficients of agreement based mainly on Artstein and Poesio (2008) and Gwet (2014).

Part II

Subjective Bias, Agreement and Consistency

Chapter 3

Agreement coefficients: Definition and use

One of the main problems we detected in Chapter 2 was the presence of shortcomings and oversights in reporting the reliability studies, measured as IAA. This Chapter aims to limit, and hopefully, eliminate this trend. To do so, we will present and discuss five popular coefficients of agreement and their interpretation.

The present Chapter is mainly based on:

- Artstein and Poesio (2008) and
- Gwet (2014).

3.1 Preliminary concept and terminology

As we have seen in Section 1.4.3 a general formulation for describing several agreement coefficients, presented by Carletta (1996), is the following:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$, the observed agreement (or percent agreement), is the proportion of times the annotators agree, whereas $P(E)$ is the proportion of times the annotators would be expected to agree by chance.

Based on the interpretation of $P(E)$, the number of annotators involved in the annotation, the type of data annotated, the K equation can take a different mathematical formulation and meaning. In order to deeply understand the different mathematical formulations that arise from the general equation K, we follow the elegant framework introduced by Gwet (2014).

Before further developing our analysis, some clarifications regarding the following points are required:

- The agreement coefficient terminology;
- The Observed agreement;
- The interpretation and definition of chance agreement;
- The type of data involved in the annotations.

3.1.1 The agreement coefficient terminology problem

The literature on agreement coefficients is characterised by a high degree of inconsistency about the terminology used for the agreement coefficients. The general presentation in the CL community, proposed in Carletta (1996), does show the common point between the different agreement coefficients, but does not resolve this confusion.

For this reason we will stop using the term K as a general term for the agreement coefficient. Every time we introduce an agreement coefficient we specify its name. Where necessary, we will examine that name in order to resolve its terminological confusion.

3.1.2 Observed agreement

In human annotation efforts, an agreement coefficient measures the extent to which different annotators can make the same annotation decision when annotating the same data independently. The first natural way to measure the level of agreement between human annotators is the observed agreement. That is, to calculate the number of times the annotators agree divided by the total number of annotations.[1] Although it is easy to calculate and interpret the observed agreement it incurs other problems that make it less than optimal as a measure of annotators' agreement. The main problem with the observed agreement is that it does not take into account situations in which annotators can reach the same annotation decision by chance (this is particularly possible in annotations where few categories are used). In other words, the observed agreement is not corrected for chance agreement.

During an annotation, it is possible that annotators select the same category by chance. A high number of chance annotations decision can increase the level of annotator agreement. However, it can frustrate the annotation aim. In other words, the chance agreement is not related to the annotation features of interest, and as such they cannot be used to demonstrate annotation reliability. Nevertheless, observed agreement is not able to distinguish agreement by chance from genuine agreement. This incapacity can be a source of annotation reliability overestimation. In addition, the fewer the number of categories, the higher the probability of overestimation.

Coefficients of agreement that correct for chance agreement were developed in order to overcome this problem.

[1] A formal definition is presented in Section 3.2.4.

3.1.3 Four interpretations and definitions of chance agreement

From a formal point of view chance agreement is a probability, that is, the probability that annotators make the same annotation decision by chance. This is a prior probability. All the definitions of chance agreement try to answer the question: how can we define a prior probability? The answer to this question determines the definition of chance agreement and the interpretation and use of the agreement coefficients that apply to it. In this chapter we present the four following approaches that define chance agreement.

1. All the categories are equally probable (Equal probability).
2. Each annotator has his(her) own probability distribution over the categories in play (Annotators probability).
3. Each categories have their own probability distribution, which is defined based on the annotators' choice through the annotation (Categories probability).
4. A combination of 1 and 3 (Mixed probability).

Between the four approaches we presented, the first can be considered the only genuine prior probability. Indeed, the others, in order to be defined, need the annotation performed. That is, the other approaches define a prior probability based on a post annotation, that is, on the actual annotation. In Section 3.2 we formally present all these types of chance agreement. In this section we present them informally.

Equal probability: The definition of chance agreement in point 1 is the most straightforward and does not involve the actual annotation. It is based on the assumption that there are no reasons to prefer a category over another. The example is that of one no bias draw. Suppose we have a box full of one hundred balls numbered from 1 to 100. Suppose we want to draw

one of these balls. If there is no bias there is no reason to think that the probability to draw, for example, the ball number 5 is higher or lower than the probability to draw the ball, for example, number 23. In this case we can suppose that the probability of drawing a ball is the same for each ball. In the example such probability is $1/100 = 0.01$.

Annotators probability: The definition of chance agreement in point 2 is based on the assumption the each annotator has its own probability distribution. Such a definition takes into account the annotators' subjective bias. Annotators can have personal reason to consider a category more probable than another one. They can reveal such bias through the annotation. That is, after the annotation we can know if an annotator prefers one category over another and based on such an annotation we can define a category's probability distribution for that annotator.

Let us return to the balls draw examples. This time we suppose that the balls are made in 5 different colours, 20 balls for each colour. Let's say, red, black, purple, green and pink. Suppose we are observing one person P drawing the balls. Suppose also that the ball is put back in the box after it was drawn. After we have seen P drawing a ball, let's say 125 times, we can ask: what is the probability that P will draw a red ball the next time? Following the approach presented in point 2, we need to count the times P drew the red ball over the times P drew a ball. Supposing that P drew the red ball 76 times, the probability that P will draw a red ball in the 126th draw is $76/125 = 0.6$.

Categories probability: The definition of chance agreement in point 3 is based on the assumption that each category has its own probability to be chosen. Such probability can be discovered by looking at the annotation trend.

Let us come back to the balls draw examples one again. Also in this case

we suppose that the balls are made in 5 different colours, 20 balls for each colour. Let's say, red, black, purple, green and pink. As before, the balls picked are put back in the box after they are drawn. This time, let's suppose that three people P_1, P_2 and P_3 have drawn 35 balls each. That gives us a sample of 105 drawn balls. We can ask what is the probability that a fourth person P_4 will draw a red ball? Let's suppose that in the 105 draw, P_1 drew a red ball 5 times, P_2 drew a red ball 15 times and P_3 drew a red ball 19 times. Following the approach presented in point 3, we can say that the probability of P_4 drawing a red ball is $(5 + 15 + 19)/105 = 0.37$.

In Section 3.2 we will see how this probability distribution is used to define the expected probability of agreement for each category.

Mixed probability: The definition of chance agreement in point 4 combines the approaches presented in point 1 and point 3. In Section 3.2 we will see how this is formally done.

3.1.4 Type of data

The collection of data, originating from a human annotation, is usually created by answering a questionnaire or by following specific instructions that aim to isolate a particular phenomenon from some data. For example, an intrinsic human evaluation can be performed by answering a questionnaire that uses items made with a five point numeric scale. Eventually, each annotation provides a particular type of dataset, and this determines the kind of statistics to be used. In Statistics four types of data are considered (Stevens, 1946): *Nominal (or Categorical)*, *Ordinal*, *Interval* and *Ratio*.

Nominal: Nominal data are constituted by a set of non-overlapping categories. Each category represents a specific characteristic which is of interest for the annotation. For example, male/female or yes/no are nominal data. Although categorical data can be presented with numerical labels, for ex-

ample 1 for yes and 0 for no, they do not have any mathematical meaning. That is, such numbers do not satisfy any mathematical property. For example, let's give 1 for yes and 0 for no. In this case, $1 + 0$ does not make sense in the same way that $yes + no$ does not.

Ordinal: Ordinal data represent a step forward from the nominal data. They are used when the order between the categories used in the annotation is considered meaningful. Also in this case ordinal categories can be presented with numerical labels. However, such numbers measure relative values between them and arithmetical calculations cannot be significantly done. For instance, the distance between two categories cannot be calculated. More specifically, suppose an annotation was performed with a Likert item that used the five categories "Strongly Disagree" (1), "Disagree" (2), "Neutral" (3), "Agree"(4) and "Strongly Agree"(5). In this case it is not possible to say that he difference between "Strongly Disagree" and "Disagree" is the same as the difference between "Neutral" and "Agree". Using the numerical categories we can express it by the expression $2 - 1 \neq 4 - 3$. In ordinal data, the order between the categories is what matters, but the differences between them are not known.

Interval: Interval data represents a step forward from the ordinal data. The categories (numerical value) of interval data have these properties: 1) they are ordered; 2) the difference between any two values is known. Within interval data it is possible to perform algebraic operations such as addition and subtraction. However, algebraic operations such as multiplication, division or the ratios' calculation cannot be performed. An example of interval data is the temperature.

Ratio: Ratio data represents a step forward from the interval data. Differently from interval data, the ratio data have the *true zero*, that is, a meaningful zero. Coming back to the temperature example, the value 0

does not mean absence of temperature. An example of ratio data is the length. 0 length means no length. Multiplication, division and the ratios' calculation can also be performed within the categories (numerical value) of ratio data.

3.2 Agreement coefficients

In this section we present some of the agreement coefficients handled in Gwet (2014). Although the agreement coefficients we present do not exhaust all the coefficients developed in the literature, we believe they provide a large set of possibilities to be used in the annotation effort of NLG tasks.

3.2.1 A common notation

Let us start by giving some notation. In what follows we assume n annotators, $a_1, \ldots, a_n$ (for $n \geq 2$), q rating categories[2] $c_1, \ldots, c_q$ (for $q \geq 2$) and i items $i_1, \ldots, i_i$. When a variable (we will mainly use the variable m) range from 1 to n, it is ranging across the annotators. When a variable (we will mainly use the variable k and l) range from 1 to q, it is ranging across the categories. Finally, when a variable (we will mainly use the variable j) range from 1 to i, it is ranging across the items. So, the letters n and m are used for annotators, the letters q, k and l are used for categories, whereas the letters i and j are used for items.

3.2.2 Working with missing data

Some annotation can have *missing data*, that is data annotated by only one annotator. Gwet (2014) generalises S, κ, π and AC_2 coefficients in order to handle annotations with missing data. To this aim Gwet (2014) uses the following strategy:

[2]The rating categories are possible labels from which annotators can choose from.

- To measure the observed agreement $P(A)$ only the items annotated by at least two annotators are used. Accordingly, all the items annotated by only one annotator are discarded.

- To measure the chance agreement $P(E)$ all the items are used. Those items annotated by only one annotator are also included in this count.

Gwet (2014) justifies the choice of taking into account all the annotated items in the calculation of $P(E)$, because it allows a more accurate measurement of either the annotators probability interpretation or the categories probability interpretation. On the other hand, in order to measure the observed agreement $P(A)$, items annotated by only one annotator cannot be taken into account. Items annotated by just one annotator do not allow us to count annotators' agreement.

Krippendorff's agreement coefficient α works for missing data as well. Nevertheless, the strategy used by Gwet (2014) is different from the one used by Krippendorff (1980). As we will see in detail in Section 3.2.7, Krippendorff's α removes all the items annotated by only one annotator, as well as for computing $P(E)$.

3.2.3 Weighting the agreement coefficients

Weights are introduced in order to deal with different data types. When nominal data are used, because there is no structured connection between categories, disagreement and agreement are two well-defined and separate concepts. Annotators either agree or they disagree. Nevertheless, when some structure is involved between categories then agreement and disagreement are no longer two clear-cut and distinct concepts. For instance, suppose we are working with ordinal data. Let's say it is data collected from a Likert item that uses the following five categories: "Strongly Disagree", "Disagree", "Neutral", "Agree" and "Strongly Agree". In this case a disagreement be-

tween “Strongly Disagree” and “Disagree” is less severe than a disagreement between “Strongly Disagree” and “Agree”. Because they are ordered, the categories “Strongly Disagree” and “Disagree” are closer than the categories “Strongly Disagree” and “Agree”. In other words, suppose that the annotator a_1 choose the category “Strongly Disagree”, the annotator a_2 chose the category “Disagree” and the annotator a_3 chose the category “Agree”. We can say that the annotators a_1 and a_2 disagree less then the annotators a_1 and a_3. In cases like this the concept of disagreement becomes unclear, leaving room for the concept of *partial agreement*. Weighted agreement coefficients were introduced to deal with partial agreement.

Gwet (2014) suggests the following set of seven weights: *ordinal*, *linear*, *quadratic*, *radical*, *ratio*, *circular* and *bipolar*. The set of weights introduced by Gwet (2014) allows using the agreement coefficients for any data type. All the weights take a value in the interval $[0, 1]$. 1 indicates perfect agreement and 0 complete disagreement. All the values in the middle represent the degree of partial agreement. In order to have an unweighted coefficient the weights w_{kl} are defined in the following way:

$$w_{kl} = \begin{cases} 1, \text{ if } c_k = c_l \\ 0, \text{ if } c_k \neq c_l \end{cases}$$

Such a weights system is called *identity weights*.

In order to present the ordinal weights let’s provide some notation. Let k, l, r and s be variables for categories. We can define M_{kl} as the number of pairs of categories (c_r, c_s), with $c_r < c_s$. Formally M_{kl} is defined as follows:

$$M_{kl} = \#\{(c_r, c_s) : min(c_k, c_l) \leq c_r < c_s \leq max(c_k, c_l)\}$$

Where the symbol # has to be read as “number of elements in the set”,

$min(x, y)$ means the smaller number between x and y, and $max(x, y)$ means the bigger number between x and y. Then, denoting with s and q respectively the lowest and highest categories used in the annotation, the ordinal weights are defined by the following equation:

$$w_{kl} = \begin{cases} 1 - M_{kl}/M_{sq}, \text{ if } c_k \neq c_l \\ 1, \text{ if } c_k = c_l \end{cases}$$

Linear, quadratic, and radical weights are special cases of the following equation: For each $k \neq l$

$$w_{kl} = \frac{|c_k - c_l|^z}{|c_{max} - c_{min}|^z}$$

Where c_k and c_l are respectively the kth and the lth sorted categories and c_{max}, c_{min} are respectively the maximum and minimum of all categories. w_{kl} is the partial agreement between the kth categories and the lth categories.

By setting the variable z equal to 1, linear weights are obtained. If $z = 2$ quadratic weights are obtained. Finally, radical weights are obtained by setting $z = 0.5$.

By using the same notation as before, ratio weights are defined by the following equation: For each $k \neq l$

$$w_{kl} = 1 - \frac{[(c_k - c_l)/(c_k + c_l)]^2}{[(c_{max} - c_{min})/(c_{max} + c_{min})]^2}$$

Circular weights are instead defined by the following equation: For each $k \neq l$

$$w_{kl} = 1 - \frac{sin[an(c_k - c_l)/(c_{max} - c_{min} + 1)]^2}{max(w)}$$

Where an is an angle, for example 180°, and $max(w)$ is the maximum of all weights.

Bipolar weights are instead defined by the following equation: For each$k \neq q$

$$w_{kl} = 1 - \frac{(c_k - c_l)^2}{max(w)(c_k + c_l - 2c_{min})(2c_{max} - c_k - c_l)}$$

Gwet (2014) suggests the use of ordinal weights in those cases where ordinal data are used. Quadratic, linear and radical weights should be used in those cases where interval data are involved. Finally, in those cases where rational data are used all but the ordinal weights are suggested.

In Section 3.5 we will discuss further the difference between weights based on some experiments.

3.2.4 A common definition of weighted percent agreement P(A)

The strategy used to measure the agreement when more than two annotators are involved is to consider all the possible pairs of annotators. Given n annotators, there are $n(n-1)/2$ possible annotators' pairs. For each pair it is possible to measure the $P(A)$ and $P(E)$ and use the average of them.

The observed agreement $P(A)$ is then defined by the following equation:

$$P(A) = \frac{1}{i'} \sum_{j=1}^{i'} \sum_{k=1}^{q} \frac{A_{jk}(\sum_{l=1}^{q} w_{kl} A_{jl} - 1)}{A_j(A_j - 1)} \qquad (1)$$

Where i' denotes the number of items that are annotated by two or more annotators. This allows us to handle annotations with missing data. If all the items i are annotated by more than one annotator, then $i' = i$. A_{jk} denotes the number of annotators that agree to give to the item j the category k, whereas A_j denotes the number of annotators that annotate the item j. If there are not missing data, that is, all n annotators annotate all the items, then $A_j = n$.

Let's further discuss the equation (1). For the sake of simplicity let's consider the unweighted case. The unweighted equation for $P(A)$ is the following:

$$P(A) = \frac{1}{i'} \sum_{j=1}^{i'} \sum_{k=1}^{q} \frac{A_{jk}(A_{jk} - 1)}{A_j(A_j - 1)} \qquad (2)$$

Let's suppose we are interested in knowing the percent of observed agreement for a given item j to be annotated with the category k. To do so we need to perform the following steps:

(i) know the number of couples of annotators that agree in associating the category k with the item j.

(ii) know the number of couples of annotators that annotate the item j.

(iii) divide the number obtained in (i) by the number obtained in (iii).

To calculate the step (i) we first need to count the number of annotators that associate the category k with the item j. In the notation of (2) this number is denoted by A_{jk}. From this number we can know the number of couples of annotators that agree in giving the category k to the item j by the equation:

$$\frac{A_{jk}(A_{jk} - 1)}{2}$$

To calculate the step (ii) we need first to know the number of annotators that annotate the item j, that is, using the notation of (2), A_j. Ones A_j has been calculated, the number of pairs of annotators that annotate the item j is determined by the equation:

$$\frac{A_j(A_j - 1)}{2}$$

At this point the step (iii) is immediate. In conclusion, the percentage of observed agreement for a given item j to be annotated with the category k

is given by the equation:

$$\frac{\frac{A_{jk}(A_{jk}-1)}{2}}{\frac{A_j(A_j-1)}{2}} = \frac{A_{jk}(A_{jk}-1)}{2} * \frac{2}{A_j(A_j-1)} = \frac{A_{jk}(A_{jk}-1)}{A_j(A_j-1)}$$

Because we are interested in calculating the percentage of observed agreement for all annotation, and not just for one item and one category, we need to sum up all the items (which are annotated by more than one annotator, in the notation of (1), i') and all the categories. This explains the double summation in equation (2). Finally, in order to be normalised, the final summation is multiplied by $\frac{1}{i}$.

In equation (1), the pairwise agreement between annotators are weighted.

If the number of annotators n reduces to 2, then the equation (1) reduces to the equation:

$$\sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}\frac{A_{kl}}{i}$$

or in the case where weights are not involved:

$$\sum_{j=1}^{i}\frac{A_{jj}}{i}$$

In both cases we assumed that $i' = i$. A_{kl} is the number of items that one annotator annotates into category k whereas the other annotator annotates into category l. Accordingly, A_{jj} is the number of items that both annotators annotate into category j.

Regarding the chance agreement $P(E)$, as we said in 3.1.3, four different approaches can be defined. We are going to formally present all of them.

3.2.5 1st approach to chance agreement: Equal probability

In this subsection we present the Brennan and Prediger (Brennan and Prediger, 1981) agreement coefficient. When used with two annotators it reduces

to Bennett, Alpert and Goldstein's S coefficient (Bennett et al., 1954). In (Gwet, 2014) the Brennan and Prediger agreement coefficient is denoted by $\hat{\kappa}_q$, and in the more general cases by $\hat{\kappa}_{BP}$. For example $\hat{\kappa}_2$ in used in the case of two annotators and two categories. Gwet (2014) refers to this coefficient as the Holley and Guilford's G-index (Holley and Guilford, 1964), of which Brennan and Prediger's agreement coefficient is a generalisation. Guilford's G-index works for two annotators and two categories. Alternatively, Bennett, Alpert and Goldstein's S works for two annotators and two or more categories. In (Artstein and Poesio, 2008), where Bennett, Alpert and Goldstein's coefficient for two annotators is presented, it uses the terminology S.

We refer to the Brennan and Prediger agreement coefficient as S, in the hope of reducing confusion about the notation. We do so because the Brennan and Prediger coefficient is a generalisation of Bennett, Alpert and Goldstein's more popular coefficient. Furthermore, the S coefficient we present here is the most general. Indeed, Brennan and Prediger's coefficient can be obtained when weights are not used. Similarly, Bennett, Alpert and Goldstein's coefficient can be obtained when weights are not used and just two annotators are involved. Whether the S used is Brennan and Prediger's coefficient or Bennett, Alpert and Goldstein's coefficient will be determined by the context of use.

The S coefficient

S is defined by the following equation:

$$S = \frac{P(A) - P(E)}{1 - P(E)}$$

where $P(A)$ is given by the equation (1) and $P(E)$ is defined in the following way:

$$P(E) = \frac{1}{q^2} \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} \qquad (3)$$

We remind that q is the number of categories used in the annotation. In the case where weights are not involved $P(E)$ reduces to $1/q$ and $P(A)$ is the equation (2). Indeed, we note that in the unweighted case $\sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl} = q$.

The assumption behind the definition of chance agreement in S is to consider the annotation a random process. In such an interpretation, each item has equal probability of being associated with each of the categories. Such probability is expressed by the equation $1/q$.

3.2.6 2nd approach to chance agreement: Annotators probability

The most popular agreement coefficient that uses this approach was developed by Cohen (1960). Cohen's κ was so influential that the coefficient agreements are often identified with the name κ.[3] This is a reason for the terminological confusion that we mentioned in Section 3.1.

Here we present Conger's κ generalisation Conger (1980) of the coefficient introduced by Cohen (1960). The first weighted version of Cohen's κ was introduced in Cohen (1968).

In (Artstein and Poesio, 2008) Cohen's coefficient is denoted by κ and κ_w in cases where weights are used. Their generalisation of the case with more than two annotators is denoted as *multi-*κ. Gwet (2014) denote Cohen's coefficient and its Conger generalisation as $\hat{\kappa}_C$.

As in the case of the S coefficient, the weighted version we present here is the most general. Indeed, Conger's coefficient is its unweighted version and Cohen's weighted coefficient is obtained when just two annotators are

[3]For example, Carletta (1996) collects several coefficient agreements under the name Kappa statistic. In the analysis we presented in Section 2.2.2 between the 24 papers that use the coefficients agreement, three papers report the coefficients agreement as κ, without providing further information.

involved. The original Cohen coefficient is obtained when the weights are not used and two annotators are involved. For this reason, in the hope of limiting the notation confusion, we will use the symbol κ. Also in this case, if the κ used is Cohen's coefficient, Cohen's weighted coefficient or Conger's coefficient will be determined by the context of use.

The κ coefficient

κ is defined by the following equation:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

$P(A)$ is expressed by equation (1). To define $P(E)$ firstly we need some notation. Let n_{mk} be the number of items annotate with the category k by the annotator a_m, and let n_m be the total number of items annotated by the annotator a_m. Then $p_{mk} = n_{mk}/n_m$ denotes the proportion of items that the annotator a_m associates with the category k. Let $\overline{p_k} = \frac{\sum_{m=1}^{n} p_{mk}}{n}$ be the mean value of the probability for the category k. Finally, let $s_{kl}^2 = \frac{1}{n-1}\sum_{m=1}^{n}(p_{mk}p_{ml} - \overline{p_k p_l})$ denote the variance of the paired proportions $p_{1k}p_{1l}, \ldots, p_{nk}p_{nl}$, where $l, k \in \{1 \ldots q\}$. We remind that n denotes the number of annotators and q the number of categories.

We can now define $P(E)$ by the following equation:

$$P(E) = \sum_{k=1}^{q}\sum_{l=1}^{q} w_{kl}(\overline{p_k p_l} - \sum_{l=1}^{q} s_{kl}^2/n) \qquad (4)$$

When the weights of equations (4) are the identity weights, the unweighted κ, which is suitable for measuring categorical data, is obtained. In this case $P(A)$ will be the equation (2) and the equation (4) reduces to:

$$P(E) = \sum_{k=1}^{q} \overline{p_k}^2 - \sum_{k=1}^{q} s_k^2/n$$

In this case the variance of the paired proportions $p_{1k}p_{1l}, \ldots, p_{nk}p_{nl}$ (s^2_{kl}) reduces to the variance of the proportions $p_{1k}, \ldots, p_{nk}$ (s^2_k) by the following equation $s^2_k = s^2_{kl} = \frac{1}{n-1}\sum_{m=1}^{n}(p_{mk}p_{mk} - \overline{p_k p_k}) = \frac{1}{n-1}\sum_{m=1}^{n}(p^2_{mk} - \overline{p_k}^2) = \frac{1}{n-1}\sum_{m=1}^{n}(p_{mk} - \overline{p_k})^2$.

$\sum_{k=1}^{q} p_{mk} = 1$ represents the probability distribution over the q categories in play as used by the annotator a_m. p_{mk} takes into account the personal bias of the annotator a_m, measuring the extent to which he(she) uses the category k. With this strategy a personal probability distribution can be defined for each annotator who takes part in the annotation. Consequently, the value obtained from an agreement coefficient that defines the chance agreement as "Annotators probability" is strictly tied to the annotators in play, and is less suitable to be generalised to other annotators who did not take part in the annotation. In this respect such coefficients are not fully suitable for measuring the reproducibility of one annotation. Nevertheless, they are suitable for measuring the trustworthiness of one annotation.[4]

3.2.7 3rd approach to chance agreement: Categories probability

The most popular coefficient of agreement that uses this approach was developed by Scott (1955). Scott's π was generalised by Fleiss's coefficient (Fleiss, 1971). Fleiss called his coefficient κ, as he was aiming to generalise Cohen's κ. Nevertheless, as a matter of fact, Fleiss's coefficient in the case of two annotators reduces to Scott's π. Another very popular coefficient that uses this approach to the definition of chance agreement is Krippendorff's α (Krippendorff, 1980).

In Artstein and Poesio (2008), Scott's coefficient is denoted by π. Its generalisation of the case with more than two annotators is denoted as *multi-*π.

[4] We will come back to this point in Section 3.6.1.

Gwet (2014) denote Cohen's coefficient differently as $\hat{\kappa}_S$ and Fleiss's generalisation of $\hat{\kappa}_S$ as $\hat{\kappa}_F$.

In this case we prefer to use only the symbol π for the same reasons we presented in the case of S and κ coefficients.

Regarding Krippendorff's coefficient, in Artstein and Poesio (2008) it is denoted as α, whereas Gwet (2014) uses the notation $\hat{\alpha}_\kappa$.

In this thesis we preferred to use the symbol α.

The π coefficients

As in the case of S and κ, π is defined by the following equation:

$$\pi = \frac{P(A) - P(E)}{1 - P(E)}$$

$P(A)$ is given by the equation (1). $P(E)$ is defined in the following way:

$$P(E) = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} \pi_k \pi_l \qquad (5)$$

where:

$$\pi_k = \frac{1}{i} \sum_{j=1}^{i} \frac{A_{jk}}{A_j}$$

It is worth reminding that i is the number of items that take part in the annotation, A_{jk} denotes the number of annotators who agree in giving category k to the item j, whereas A_j denotes the number of annotators that annotate the item j.

In the unweighted case $P(A)$ reduces to (2) and $P(E)$ to $\sum_{k=1}^{q} \pi_k^2$.

The α coefficients

In this subsection we present the α coefficient as defined in (Gwet, 2014). Such a presentation is different from the original given by Krippendorff

(1980). We prefer to present Gwet's version to retain uniformity with the presentation of the other coefficients. As stated by Gwet (2014) (page 87) Krippendorff's version and Gwet's version yield the exact same result.

We saw in Section 3.2.2 that the α coefficients deal with missing data in a different way. Only the items that are annotated from two or more annotators are taken into account in the calculation of α. That is, all items that are ranked by a single annotator are omitted from both the calculation of $P(A)$ and the calculation of $P(E)$. For this reason the $P(A)$ formulation in equation (1) has to be reconsidered. Nevertheless, the α coefficient is based on the general formulation:

$$\alpha = \frac{P_\alpha(A) - P(E)}{1 - P(E)}$$

In order to define $P_\alpha(A)$ and $P(E)$, let's start by giving some notation. Let i' be the number of items annotated by two or more annotators. Let A_{jk} denote the number of annotators that agree in giving the category k to the item j, whereas A_j denotes the number of annotators that annotate the item j. Let $\overline{A}$ be the average of the A_j for $j, \in \{1, \ldots, i\}$. We remind that i is the number of items used in the annotation. Let $\epsilon_i = 1/(i'\overline{A})$. In order to define $P_\alpha(A)$ we need to define $P(A)'$:

$$P(A)' = \frac{1}{i'} \sum_{j=1}^{i'} \sum_{k=1}^{q} \frac{A_{jk}(\sum_{l=1}^{q} w_{kl} A_{jl} - 1)}{\overline{A}(A_j - 1)}$$

Then $P_\alpha(A)$ can be defined by the following equation:

$$P_\alpha(A) = (1 - \epsilon_i)P(A)' + \epsilon_i$$

In order to define $P(E)$ it is necessary to slightly modify the π_k as defined in the π coefficient. Such modification, here denoted by π_k^α, takes into ac-

count the procedure to deal with missing data as introduced by Krippendorff (1980). The π_k^α is defined as follows:

$$\pi_k^\alpha = \frac{1}{i'} \sum_{j=1}^{i'} \frac{A_{jk}}{\overline{A}}$$

In the cases where there are no missing annotations, then $\pi_k^\alpha = \pi_k$.

Finally, $P(E)$ for the α is defined by the following equation:

$$P(E) = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl} \pi_k^\alpha \pi_l^\alpha \qquad (6)$$

In the unweighted case, $P_\alpha(A) = (1 - \epsilon_i)P(A) + \epsilon_i$ where $P(A)$ is equation (2) and $P(E)$ reduces to $\sum_{k=1}^{q} (\pi_k^\alpha)^2$.

Both π_k and π_k^α represent the single probability distribution for the category k. As such, $\sum_{k=1}^{q} \pi_k = 1$ ($\sum_{k=1}^{q} \pi_k^\alpha = 1$) represent the probability distribution over the q categories in play as used by the annotators that take part in the annotation. This strategy defines a single category probability distribution for all the annotators. Accordingly, the value obtained from an agreement coefficient that define the chance agreement as "Categories probability" is suitable to be generalised to other annotators that did not take part in the annotation. In this respect such coefficients are suitable for measuring the reproducibility of one annotation. Nevertheless, they are less suitable for measuring the trustworthiness of one annotation.[5]

3.2.8 4th approach to chance agreement: Mixed probability

To the best of our knowledge, the only coefficients that falls into this category are the AC_1 coefficient introduced by Gwet (2008) (and its generalisation AC_2) and the Aickin's α (Aickin, 1990). In this Section, we will deeply

[5] We will came back to this point in Section 3.6.1.

consider only the coefficient defined by Gwet. We use the terminology AC_2 in this thesis because the coefficient AC_2 reduces to AC_1 in the absence of weights.

AC_2 considers two possible sources of disagreement. Disagreement can be either due to chance annotation or for other reasons. The aim of AC_2 is to measure the agreement once the random chance agreement is removed from consideration. In order to do so, AC_2 is applied to a sub-population of items I'. I' is obtained by removing, from the initial items population, the items which can lead to random chance agreement.

The strategy used is to define, from a theoretical point of views, the H-subject and the E-subject. H-subject are hard to score items, whereas E-subject are easy to score items. In order to define the H-subject, Gwet (2014) (Page 103) introduces the concept of *nodeterministic* as follows:

> the process of rating [annotating in our notation] a subject [an item in our notation] is considered *nodeterministic* if it has no apparent connection with the subject's characteristic.

To move from the theoretical level to the practical one, the following assumptions are made (Gwet, 2014):

(I) Each annotator has his(her) group of H-subject and his(her) group of E-subject;

(II) H-subject are nodeterministic and chance agreement on them is considered random.

H-subject and E-subject cannot be identified, so a probabilistic model is needed in order to link the annotation to the theoretical concept of H-subject and E-subject. Both in (Gwet, 2008) than (Gwet, 2014), the rationale behind the implementation of the agreement coefficient AC_2 is presented for the case of two annotators, let's say a_1 and a_2. The first step is to identify

the following events (Gwet, 2014, Page 115):

- $\mathcal{R}$: The selected subject is an H-subject *(i.e one of the two raters or both will perform a nondeterministic rating when classifying this subject).*
- A: both raters a_1 and a_2 agree on the classification of the selected subject.
- $\mathcal{C} = A \cap \mathcal{R}$: Represent an agreement by chance *(i.e. the selected subject is an H-subject, and both raters a_1 and a_2 agree about its classification).*

In the definition of $\mathcal{R}$ the assumption (I) takes place.

Given two categories l and k, let P_{kl} denote the probability that an item will be associated with the category l by one annotator and with the category k by the other annotator. By using the rule of probability P_{kl} can be defined as follows:

$$P_{kl} = P(kl \cap \mathcal{C}) + P(kl \cap \overline{\mathcal{C}}) \qquad (7)$$

Where $\overline{\mathcal{C}}$ is the event *no chance agreement*, $P(kl \cap \mathcal{C})$ is the probability that an item will be associated with the category l by one annotator and with the category k by the other annotator by chance. However, $P(kl \cap \overline{\mathcal{C}})$ is the probability that an item will be associated with the category l by one annotator and with the category k by the other annotator not by chance.

The aim is to quantify $P(kl \cap \overline{\mathcal{C}})$ when $k = l$. Indeed, $P(kk \cap \overline{\mathcal{C}})$ is the probability of agreement reached not by chance on the categories k and l.

By Bayes's rule:

$$P(kl \cap \mathcal{C}) = P(\mathcal{C})P(kl|\mathcal{C}) \qquad (8)$$

and

$$P(kl \cap \overline{\mathcal{C}}) = P(\overline{\mathcal{C}})P(kl|\overline{\mathcal{C}}) \qquad (9)$$

Substituting the equations (8) and (9) in equation (7) it follows:

$$P_{kl} = P(\mathcal{C})P(kl|\mathcal{C}) + P(\overline{\mathcal{C}})P(kl|\overline{\mathcal{C}}) \qquad (10)$$

By again applying Bayes's rule on $P(\mathcal{C})$ ($P(\mathcal{C}) = P(\mathcal{R})P(A|\mathcal{R})$)[6], and considering the fact that $P(\overline{\mathcal{C}}) = 1 - P(\mathcal{C})$, it follows:

$$P_{kl} = P(\mathcal{R})P(A|\mathcal{R})P(kl|\mathcal{C}) + 1 - P(\mathcal{R})P(A|\mathcal{R})P(kl|\overline{\mathcal{C}}) \qquad (11)$$

From assumption (II) it follows that $P(A|\mathcal{R}) = 1/q$, where, we recall, q is the number of categories used in the annotation. Similarly, $P(kk|\mathcal{C}) = 1/q$ when $k = l$. By defining the variable:

$$d_{kl} = \begin{cases} 1, \text{ if } k = l \\ 0, \text{ otherwise} \end{cases}$$

it follows that $P(kl|\mathcal{C}) = d_{kl}/q$. Taking these into account we can rewrite equation (11) as follows:

$$P_{kl} = P(\mathcal{R})\frac{1}{q}\frac{d_{kl}}{q} + 1 - P(\mathcal{R})\frac{1}{q}P(kl|\overline{\mathcal{C}}) = \frac{P(\mathcal{R})d_{kl}}{q^2} + 1 - \frac{P(\mathcal{R})}{q}P(kl|\overline{\mathcal{C}}) \qquad (12)$$

By summing up all the categories, it follows:

$$\sum_{k=1}^{q} P_{kl} = \frac{P(\mathcal{R})\sum_{k=1}^{q} d_{kl}}{q^2} + \frac{1 - P(\mathcal{R})}{q}\sum_{k=1}^{q} P(kl|\overline{\mathcal{C}}) \qquad (13)$$

Considering the case where $k = l$, it follows:

$$\sum_{k=1}^{q} P_{kk} = \frac{P(\mathcal{R})}{q} + \frac{1 - P(\mathcal{R})}{q}P(kk|\overline{\mathcal{C}}) \qquad (14)$$

[6] $P(A|\mathcal{R})$ is the conditional probability that the agreement is reached, given that one annotator between a_1 and a_2 (or both) has performed a random annotation.

At this point it is possible to measure $P(kk|\overline{\mathcal{C}})$ by the following equation:

$$P(kk|\overline{\mathcal{C}}) = \frac{P(A) - P(\mathcal{R})/q}{1 - P(\mathcal{R})/q} \qquad (15)$$

Where $P(A) = \sum_{k=1}^{q} P_{kl}$. The standard equation for agreement coefficient is found by defining $P(E) = P(\mathcal{R})/q$. The problem now is how to define $P(\mathcal{R})$. In order to do so, Gwet (2008) (and Gwet (2014)) made the following assumption:

> Subjects [items in our notation] distributed more uniformly across categories are more likely to contain H-subject. (Gwet, 2014, Page 116)

Gwet (2008) (and Gwet (2014)) define $P(\mathcal{R})$ by the following equation:

$$P(\mathcal{R}) = \frac{\sum_{k=1}^{q} \pi_k(1 - \pi_k)}{1 - 1/q}$$

π_k is the probability that an annotator (randomly selected from the population of annotators) will classify an item (randomly selected from the population of items) with the category k. With a further assumption about the nature of the change agreement, π_k is defined as in the case of the π coefficient.

Finally, with a few arithmetical steps, $P(E) = P(\mathcal{R})/q$ can be defined as

$$P(E) = \frac{1}{q-1} \sum_{k=1}^{q} \pi_k(1 - \pi_k)$$

In the following section we present the generalisation given by Gwet (2014) from the case of more than two annotators.

The AC_2 coefficient

As we have seen in the previous section, AC_2 is defined by the following equation:

$$AC_2 = \frac{P(A) - P(E)}{1 - P(E)}$$

Also in this case, $P(A)$ is expressed by the equation (1). $P(E)$ is defined in the following way:

$$P(E) = \frac{T_w}{q(q-1)} \sum_{k=1}^{q} \pi_k(1 - \pi_k) \qquad (16)$$

Where $T_w = \sum_{k=1}^{q} \sum_{l=1}^{q} w_{kl}$ is the sum of all weight w_{kl} and π_k is defined as in the case of the π coefficient. Finally, we remind that q is the number of categories used in the annotation.

In the unweighted case, $P(A)$ reduces to (2) and $P(E)$ to:

$$P(E) = \frac{1}{q-1} \sum_{k=1}^{q} \pi_k(1 - \pi_k)$$

In this case, T_w in equation (16) reduces to q.

Divergences with the Aickin's α

A similar approach to the one proposed by Gwet (2008) was presented by Aickin (1990) who introduced Aickin's α.

As pointed out by Gwet (2014) the main differences between Aickin's α and Gwet's AC_2 can be summarised by the following points:

- Aickin's α is defined for two annotators. In contrast, Gwet's AC_2 is defined for any number of annotators.
- In Aickin's proposal the annotators share the same H-subject and E-subject. In contrast, Gwet's AC_2 assumes that each annotator has

his(her) group of H-subject and his(her) group of E-subject.

- Aickin's α is applied on all the initial items population. In contrast, Gwet's AC_2 is applied to a sub-population of items I' (I' is obtained by removing, from the initial items population, the items which can lead to random chance agreement).

From a mathematical point of view, Aickin's α is defined by the following equation:

$$\frac{\sum_{k=1}^{q} P_{kk} - \sum_{k=1}^{q} P_{k|H}^{A} P_{k|H}^{B}}{1 - \sum_{k=1}^{q} P_{k|H}^{A} P_{k|H}^{B}}$$

Where A and B are two annotators, P_{kk} is the probability that A and B associate the category k to a randomly selected item, and finally $P_{k|H}^{X}$ is the probability that the rater X (for $X \in \{A, B\}$) associates the category k to a H-subject.

Aickin's α takes the same general shape of the K equation (see Section 3.1). Nevertheless, the chance agreement $P(E)$ is defined only for the H-subject.

For this reason, Gwet (2014) argues that:

> by excluding subjects [items in our notation] that are susceptible to chance agreement from the numerator while leaving them in the denominator, Aickin makes it difficult if not impossible for its coefficient to reach the perfect value of 1. This is particularly the case when "Hard" subjects are present in the subjects population. Consequently, Aickin's alpha coefficients could be artificially low for some subject population. (Gwet, 2014, page 114)

3.3 Agreement coefficients interpretation

Given its numerical nature, each agreement coefficient needs to be interpreted in order to be used. The interpretation of the agreement coefficients is what determines the usability of the annotated data. For this reason it is an important step in each annotation task. Nevertheless, to date, there is a lack of consensus about the interpretation of agreement coefficients. Artstein and Poesio (2008) refer to such a deficit "as the most serious problem with current practice in reliability testing, and one of the main reasons for the reluctance of many in CL to embark in reliability studies". Indeed, as noted by Eugenio and Glass (2004), Krippendorff (2004), Craggs and Wood (2005) and Hovy and Lavid (2010), the choice of an interpretation scale is arbitrary and task dependent. For example, Krippendorff (1980) (see 3.1)

AV value	AV interpretation
$AV < 0.67$	Discard
$0.67 \leq AV < 0.8$	Tentative
$0.8 \leq AV \leq 1$	Good

Table 3.1: Krippendorff scale of interpretation for the Kappa statistic. AV denotes agreement value as determined by some coefficients of agreement.

considers good any data annotation with agreement in the interval [0.8, 1], tentative any data annotation where agreement is in the interval [0.67, 0.8] and to discard any data annotation where agreement is below 0.67.[7]

Another example of interpretation scale is the one introduced by Landis and Koch (1977) and shown in Table 3.2. Since Carletta (1996) the agreement coefficients and the Krippendorff interpretation scale have been introduced in the area of CL. Carletta took these metrics from content analysis in order to "compare results in a standard way across different coding schemes and

[7] Artstein and Poesio (2008) (page 576) note that: *the description of the 0.67 boundary in Krippendorff (1980) was actually "highly tentative and cautious," and in later work Krippendorff clearly considers 0.8 the absolute minimum value of α to accept for any serious purpose: "Even a cutoff point of $\alpha = .800$ ... is a pretty low standard".*

AV value	AV interpretation
$AV < 0$	Poor
$0 \leq AV \leq 0.2$	Slight
$0.2 < AV \leq 0.4$	Fair
$0.4 < AV \leq 0.6$	Moderate
$0.6 < AV \leq 0.8$	Substantial
$0.8 < AV \leq 1$	Almost Perfect

Table 3.2: Landies and Koch scale of interpretation for the Kappa statistic. AV denote agreement value as determined by some coefficients of agreement.

experiments and to evaluate current developments" specifying that "whether we have reached (or will be able to reach) reasonable level of agreements in our work as a field remains to be seen".

Gwet (2014) presents a new approach to the scale interpretation problem. Gwet (2014) start from the observation that the interpretation scales developed are not static. They are fixed once and do not take into account several factors that can be involved in one annotation. Gwet (2014) notes that the number of subjects, the number of raters and the number of categories used are determining factors that have to be taken into account in the interpretation step. Consequently, deterministic scales of interpretation, for example 3.1 or 3.2, have to be reconsidered as probabilistic. The strategy used in (Gwet, 2014, page 174) involved three steps.

(A) Given a scale of interpretation, compute the probability for a coefficient to fall into each of the threshold intervals described by the scale.

(B) Compute the cumulative probability, starting from the highest threshold level.

(C) Select the first threshold interval for which the cumulative probability exceeds a given level (Gwet (2014) uses the threshold of 95%).

To perform point (A) two steps are needed.

(A.1) Given an agreement coefficient Agr (for example κ or α), measure the standard error (SE) associated to Agr.

(A.2) Given a chosen scale of interpretation (SI), for each threshold intervals $(a, b)_{SI}$ of SI, compute the *Interval Membership Probability* (IMP) by the following equation:

$$IMP = P\left(\frac{Agr - b}{SE} \leq Z \leq \frac{Agr - a}{SE}\right)$$

where Z is a standard Normal variate.

Regarding point (C), the threshold of 95% is suggested in Gwet (2014) as a good level to minimise the risk of error associated with the conclusions drawn from the agreement coefficient. That said, other thresholds can be used.

The framework proposed by Gwet (2014) places the probabilistic and statistical nature of the agreement study at the centre of the analysis. As we saw in the previous section, all the coefficients are based on some probabilistic definition of the change agreement. Changing the number of items, annotators or categories can make the final agreement value subject to different interpretations. The method of Gwet (2014) allows a more meaningful comparison between studies that take into account different experiment designs.

3.4 The prevalence paradox

Sometimes, when a measure of agreement is performed, unexpected results can be produced. In literature such results are referred to as the *kappa paradoxes*. The most popular is the *prevalence paradox*. The prevalence paradox can cause paradoxical results in which a high observed agreement value is associated with a low agreement coefficient value. Gwet (2014) considers such results highly problematic. Conversely, Artstein and Poesio

(2008) consider such surprising results as "correct and justified" (Artstein and Poesio, 2008, page 26), arguing that agreement coefficients (for example κ) measure the ability to agree on the rare categories.

The prevalence paradox occurs when an annotation is unbalanced towards one category. In this case, it becomes difficult to distinguish between the rare categories. At the same time, random annotation on the common category can result in high agreement. For example, suppose an annotation is performed by two annotators a_1 and a_2 on the categories c_1 and c_2. Suppose that 90% of the data falls into the category c_1. Then the probability that a_1 and a_2 randomly agree in annotating an item under the category c_1 is $0.9 \times 0.9 = 0.81$. At the same time the probability that the annotators randomly agree in annotating an item with the category c_2 is $0.1 \times 0.1 = 0.01$. In the case of the prevalence paradox, Artstein and Poesio (2008) suggest using agreement coefficients such as κ, π and α. Such coefficients can be accompanied by the observed agreement. Nevertheless, researchers must emphasise that the observed agreement is not corrected by chance agreement and that it is inadequate to use as a measure of data reliability.

On the other hand, Gwet (2014) argues that agreement coefficients such as κ are inadequate for measuring the agreement in an imbalanced annotation. Imbalanced annotations raise questions about the nature of the chance agreement, as defined by agreement coefficients that suffer from the prevalence paradox. Gwet (2014) (page 60) claims that, for such agreement coefficients, the main problem with the definition of chance agreements is that they assume "all or most rating associated with a category could be used in the calculation of p_e [P(E) in our notation] as if they were all assigned randomly". In defining the AC_2 agreement coefficient, Gwet (2008), made the following assumptions (Gwet, 2008, page 35):

(a) Chance agreement occurs when at least one rater rates an individual

randomly.

(b) Only an unknown portion of the observed ratings is subject to randomness.

In section 3.6 we will come back to the prevalence paradox. We will argue that the agreement coefficient to be used in an imbalanced annotation depends on the annotation aim.

3.5 Agreement coefficients comparison: an experimental point of view

In Section 3.2 we presented five different agreement coefficients. We now try to test the difference between them by comparing the values they reach in some annotations in the area of NLG. The aim is to check the differences between such coefficients. More precisely we want to check how they work in the case of the prevalence paradox and if they are robust, based on a different use of the weights.

Dataset used

In order to perform our experiment we used the QG-STEC evaluation dataset (Rus et al., 2010).[8] The use of QG-STEC evaluation dataset allowed us to collect data based on datasets with a different number of evaluated sentences — from a minimum of 67 to a maximum of 158.

For each dataset we measured the S, κ, π, α and AC_2 agreement coefficients. For each coefficient we also measured their weighted version. More precisely we measured the ordinal, linear, quadratic, radical, ratio, circular and bipolar weights.

All the results from the experiments are available online (Amidei, 2020b).

[8] More details about the QG-STEC evaluation dataset can be found in Section 5.2.3.

Together with the agreement values we reported the confidence interval and the interpretation categories as defined by Landis and Koch (1977) (see Table 3.2). We used the Landis and Koch interpretation scale because it allows a fine grain analysis for low agreement values. For example the Krippendorff (1980) scale of interpretation (Table 3.1) collapses all the values less than 0.65 together. Conversely, the Landis and Koch (1977) scale of interpretation subdivides the values less than 0.65 into four categories.

Agreement coefficient comparison

We found some regularity in our analysis. For example, in the case of unbalanced annotations the following differences are generally true $\pi \leq \alpha \leq \kappa < S << AC_2$. Nevertheless, this is not always the case. For example, if the annotation is unbalanced in two categories[9] with the ordinal, quadratic, linear and bipolar weights, the order $S < AC_2 < \pi \leq \alpha < \kappa$ is detected. For instance, the fluency criterion judged by Judge 1 and Judge 2, where the categories frequency are 67 (category 1), 24 (category 2), 12 (category 3) and finally 57 (category 4). The differences in value between the agreement coefficients reduces with the diminution of the categories imbalance. In the case of balance annotation we found that $\pi \leq \alpha \leq \kappa \leq S \leq AC_2$. For example, the variety criterion judged by Judge 3 and Judge 5, where the categories' frequencies are 71 (category 1), 98 (category 2) and 81 (category 3).

The fluency criterion judged by Judge 1 and Judge 4 present the only case in which $S \leq \pi \leq \alpha \leq \kappa \leq AC_2$. In this case the frequencies are 61 (category 1), 19 (category 2), 48 (category 3) and 34 (category 4).

There are cases in which the difference between coefficients is characterised by a difference in the categories' coefficient interpretation. This fact is more

[9]This means that there are two categories that are selected more times than the other categories.

marked between AC_2, S and the other coefficients, especially in the unbalanced annotations. For example, in the case of the relevance criterion judged by Judge 1 and Judge 2 the ordinal κ reaches a value of 0.14, the ordinal S reaches the value 0.48 and the ordinal AC_2 reaches the value of 0.75. In term of the scale of interpretation, this variance means a jump from the category *Poor* (κ) to the category *Fair* (S) to the category *Substantial* (AC_2). In our experiment, we did not find significant differences between the coefficients π, α and κ, which are generally similar.

Regarding the weights, we generally found these inequalities: $circular \leq unweighted < ratio \leq ordinal \leq bipolar < linear < radical < quadratic$. Differences in weights can create differences in the coefficient interpretation categories. For example, in the case of the fluency criterion judged by judges 1 and 3, relative to the κ coefficient, the ordinal weight reaches the value of 0.51, the quadratic weight 0.55 and the unweighted 0.31. Such differences, from a scale of interpretation point of views, move from *Slight* to *Fair*. Similar results are true also for the other coefficients. These results suggest the importance of using the correct weights for the data. Indeed, measuring ordinal data with a coefficient of the agreement created for nominal data can lead to an underestimation of the data reliability. Similar unwanted results can be obtained by using radical or circular weights, which are more adequate for ratio or interval data. Likewise with the use of ratio weight, recommended by Gwet (2014) for ratio data. For example, in the case of the relevance criterion judged by judges 1 and 4 the ordinal score for AC_2 is 0.71, whereas the ratio score is 0.59. In terms of scale interpretation, this means a drop from *Substantial* (ordinal) to *Moderate* (ratio).

Our results suggest that quadratic or linear weights, although recommended by Gwet (2014) for interval data, can be adequate for measuring ordinal data. The same suggestion is also true for bipolar weights.

3.6 How to perform an agreement study

3.6.1 Choose the agreement coefficient

Once the annotation data are collected some variables have to be checked in order to justify the choice of the agreement coefficient.

In this thesis we do not take into account problems linked to the number of annotators or the presence of missing data. Indeed, by using the irrCAC R library (see Section 3.7 for more details), all the S, κ, π, α and AC_2 coefficients work for all numbers of annotators and either for datasets with no missing data or datasets with missing data.

In cases where researchers prefer to use different statistical software to measure the coefficient of agreement or coefficients different from S, κ, π, α and AC_2, then the number of annotators or the presence of missing data become variables to be taken into account for choosing the agreement coefficient.

The data type

The first variable to be checked is the data type. As we saw in Section 1, Gwet (2014) suggests the use of ordinal weights in cases where ordinal data are used. He suggests using quadratic, linear and radical in cases where interval data are involved. Finally, in the case of rational data, all but the ordinal weights are suggested.

Our experiments suggest that quadratic or linear weights can be adequate for measuring ordinal data. The same suggestion is true also for bipolar weights. The use of one weight over another depends on the distance that researchers want to emphasise between the annotation categories in play.

Aim of the annotation

In CL annotation tasks, we can delineate at least the two following purposes for using an agreement coefficient:

P1 Validating and improving annotation scheme and guideline;

P2 Validating the final human annotation.

Although in both cases, agreement coefficients are used for validation purposes, the aims behind them are quite different. In the first case (P1), validation is sought to test the quality of the data to be annotated, develop an annotation scheme and test guideline reliability as reproducibility. In the second case (P2), the annotation validity or trustworthy is sought.

More specifically, the main aim of P1 is to create an annotation guidelines which is reproducible. A fundamental step in the goal of developing an annotation guideline is to cyclically test its reliability (Artstein, 2017). This practice, which uses agreement coefficients as an evaluation metric, is based on a loop of guidelines testing and guidelines (or annotation scheme) improvement that can last a long time. The loop ends when a reasonable level of the agreement coefficient used is reached.[10]
In P1, the use of the agreement coefficient aims to answer the following question. Is the annotation generalisable? That is, if we use a different set of annotators, are we going to reach the same agreement coefficient value?

The main aim of P2 is to check to which extent the annotation is affected by the raters' bias. In P2, the use of the agreement coefficient aims to answer the following question. Is the annotation trustworthy? That is, can we safely use it to make inferences?

P1 and P2 have different purposes, and as such different agreement coeffi-

[10]The threshold considered as a reasonable level of the agreement coefficient used changes from task to task. Nevertheless, with due clarification, Artstein and Poesio (2008) suggests the value 0.8.

cient should be used (Artstein and Poesio, 2008; Craggs and Wood, 2005).

From a theoretical point of view:

- When the purpose P1 is involved agreement coefficients that interpret the chance agreement as "Categories probability"should be used — that is, each category has its own probability distribution, which is defined based on the annotators' choices through the annotation. For example, the π, α or AC_2 are adequate for such a purpose. In this case, a probability distribution is suitable to be generalised with other annotators because it is calculated for each category by using the annotation of different annotators.

- When the purpose P2 is involved agreement coefficients that interpret the chance agreement as "Annotators probability"should be used — that is, each annotator has his(her) own probability distribution over the categories in play. For example, κ is suitable for such a purpose. Because, in this case, a probability distribution is defined for each annotator, such probability distribution exhibits the annotator bias. For these reasons it is less suitable to be generalised with other annotators.

From an empirical point of view Artstein and Poesio (2008) note that the difference between π, α and κ is often not that high. Such a fact is confirmed from our experiments, in which the difference between these agreement coefficients is negligible. Nevertheless, this is not the case for the difference between AC_2 and κ, which is often remarkable. In this case the use of an agreement coefficient not suitable for the annotation aim in play can cause an underestimation or overestimation of the annotation agreement.

The case of imbalanced annotation

In the case of imbalanced annotation — which can cause the prevalence paradox for π, α or κ — the choice of the agreement coefficient can change,

based on the annotation aim in play.

As noted in Artstein and Poesio (2008, page 26) "Reliability implies the ability to distinguish between categories, but when one category is very common, high accuracy and high agreement can also result from indiscriminate coding. The test for reliability in such cases is the ability to agree on the rare categories (regardless of whether these are the categories of interest). Indeed, chance-corrected coefficients are sensitive to agreement on rare categories." In a case where purpose P1 is involved, agreement coefficient as π and α looks to be adequate. Indeed, they can provide useful information on how to improve the annotation guidelines for the case of rare categories.

In a case where the purpose P2 is involved, the situation is more complicated. Indeed, although κ can give useful information about the agreement on the rare categories, it can lose information about the genuine agreement — that is, agreement that is not due to chance — for the non-rare categories. In this case, observed agreement can be reported.[11]

Following Gwet (2014), the AC_2 coefficient can be use together with the κ value. Although AC_2 is more adequate for purpose P1 than for purpose P2, in the case of imbalanced annotation it can be highly informative also for the purpose P2. Indeed, it can provide insight into the agreement of the non-rare categories. When the purpose P2 is in play, we are interested not only in the level of agreement between the rare categories — which can be measured with κ — but we are also interested in the level of agreement between the non-rare categories — which can be measured with AC_2. From a theoretical point of view, the use of AC_2 for the purpose P2, can be justified, because the annotators' probability distribution on the non-rare categories tend to be the same.

[11] This is the general suggestion given in (Artstein and Poesio, 2008). Artstein and Poesio (2008) suggest also that researchers must notice that the observed agreement is not corrected by chance agreement and it is inadequate to be used as a measure of data reliability.

3.6.2 Coefficient interpretation

As we saw in Section 3.3, Gwet (2014) introduced a statistical method to interpret the agreement coefficient value, that takes into account possible variation in the number of items, annotators or categories. Such a model represents an improvement with respect to the static interpretation of the agreement coefficient. Nevertheless, in the case of NLG intrinsic human evaluation the threshold defined from a interpretation scale can still be too restrictive. Given the result presented in Section 4, we suggest reporting the confidence interval associated with the agreement value. The confidence interval allows us to consider the uncertainty linked to the agreement measure. We should expect the agreement value to fall into the confidence interval in cases of the annotation reproducibility.

In case the interpretation of the agreement coefficient value is reported we suggest using the Gwet (2014) method instead of the static one – that is, reporting a category's value as directly determined by the scale of interpretation. Any scale of interpretation used should be justified.

3.6.3 Agreement coefficient value reproducibility

Given the language variability we describe in Chapter 4, the reproducibility of a NLG intrinsic human evaluation can be difficult to obtain. That is, given the same set of generated sentences to two (or more) different sets of annotators, can we reach the same results? From an agreement point of view, we can ask if we can reach the same agreement values. Researchers should pursue the following best practices when the agreement of an intrinsic human evaluation is performed, in order to make such questions meaningful.

- **Provide information about the agreement coefficient used:**
  - The name of the agreement coefficient used;

  - If the agreement coefficient use weights, which kind of weights are used;
  - The software used to measure the agreement coefficient;
  - The value of the agreement coefficient.

- **Provide information about the interpretation of the agreement coefficient:**
  - The confidence interval associated with the agreement value;
  - If a scale of interpretation is use, the name of the interpretation scale used.

- **Provide information about the dataset on which the evaluation was performed:**
  - Where to retrieve the dataset.

- **Provide information about the evaluation guideline:**
  - Where to retrieve the evaluation guideline;
  - The number of annotators for item used in the evaluation.

All the choices made by the researchers, for example the agreement coefficient used, the weight used or the scale of interpretation used, should be justified from a theoretical point of view.

3.7 irrCAC a R library to measure agreement coefficients

In this section we present the *irrCAC* library (`https://rdrr.io/cran/irrCAC/`) developed in the statistical software *R*. irrCAC provides a friendly framework which measures and interprets the coefficients of agreement presented in (Gwet, 2014).

The library irrCAC allows us to measure the coefficient of agreement S, π, α, κ and AC_2. Each coefficient allows for a weighted analysis with the seven weights: ordinal, linear, quadratic, radical, ratio, circular and bipolar. The library measures the agreement on 3 types of input data (`https://rdrr.io/cran/irrCAC/f/vignettes/overview.Rmd`):

- contingency table;
- the distribution of raters by subject and by category;
- the raw data, which is essentially a plain dataset where each row represents an item and each column represents the ratings associated with one rater.

In what follows we present examples based on the raw data type. We will use the QG-STEC evaluation dataset. More specifically the batch of data evaluated by judge 1 and judge 3 on the category ambiguity. The raw data are depicted in Figure 3.1

Judge.1	Judge.3
1	2
1	1
1	1
2	1
1	1
1	2

Figure 3.1: Raw data of the ambiguity evaluation by judge 1 and judge 3.

Once the data are collected in a CVS file, the data can be easily uploaded with the command:

read.csv(‘`data_path`’)

Let's call the upload data `ambiguity_j1andj3`. That is:

`ambiguity_j1andj3` = read.csv(‘`data_path`’).

With this notation, the data depicted in Figure 3.1 can be printed with the command:

head(`ambiguity_j1andj3`)

Once the data are uploaded, it is possible to measure the agreement coefficients that we need. The first step is to install the irrCAC packages. This can be done with the command:

install.packages(“irrCAC”)

Once the irrCAC packages are installed, the next step is to import the irrCAC library. The following command does this job:

library(irrCAC)

Now we are ready to measure the agreement coefficient we need. Let’s see a few examples. In order to measure the percentage of agreement (or observed agreement) the following command has to be used:

```
a.coeff.raw(ambiguity_j1andj3)
```

The print of such a command is depicted in Figure 3.2

coeff.name	pa	pe	coeff.val	coeff.se	conf.int	p.value	w.name
Percent Agreement	0.5522388	0	0.5522388	0.06121	(0.43,0.674)	3.970158e-13	unweighted

$weights

1	0	0
0	1	0
0	0	1

$categories

1 2 3

Figure 3.2: Example of unweighted percentage of agreement with the visualisation of weights and categories.

The printing shown in Figure 3.2, has the same format for each coefficient.

The labels have the following intuitive meaning:

- **coeff.name:** Is the name of the coefficient of agreement used;
- **pa:** is the value of the observed agreement (or percentage agreement);
- **pe:** is the value of the expected agreement;
- **coeff.val:** is the value of the coefficient of agreement used;
- **coeff.se:** is the value of the standard deviation;
- **conf.int:** is the confident interval for the value reported in **coeff.val**;
- **p.value:** is the p-value;
- **w.name:** is the name of the weight used;
- `$weights`**:** is the visualisation of the weight used;
- `$categories`**:** report the categories used in the annotation.

We note that in the case of Figure 3.2, **pa** and **coeff.val** report the same value. This is due to the fact that, as reported by **coeff.name**, the coefficient of agreement used is the percentage agreement.

In order to remove the visualisation of weights and categories the variable `$est` has to be added to the end of the command `a.coeff.raw(ambiguity_j1andj3)`. For example, in the case of the percent agreement, the command:

```
a.coeff.raw(ambiguity_j1andj3)$est
```

will print Figure 3.3.

coeff.name	pa	pe	coeff.val	coeff.se	conf.int	p.value	w.name
Percent Agreement	0.5522388	0	0.5522388	0.06121	(0.43,0.674)	3.970158e-13	unweighted

Figure 3.3: Example of unweighted percent of agreement without the visualisation of weights and categories

Let's use an example to understand the difference in weight. Let's use Conger's κ. In our case, the basic command to calculate Conger's κ is the following:

```
conger.kappa.raw(ambiguity_j1andj3, weights = "")
```

In the case when unweighted weight wants to be used, the variable weights has to be set as "unweighted". That is, the command

```
conger.kappa.raw(ambiguity_j1andj3, weights = "unweighted")
```

will print the Figure 3.4.

coeff.name	pa	pe	coeff.val	coeff.se	conf.int	p.value	w.name
Conger's Kappa	0.5522388	0.4571174	0.17522	0.1022	(-0.029,0.379)	0.09113209	unweighted

$weights

1	0	0
0	1	0
0	0	1

$categories

1 2 3

Figure 3.4: Example of unweighted weights for κ

In the case when ordinal weights want to be used, the variable weights have to be set as "ordinal". That is, the command

```
conger.kappa.raw(ambiguity_j1andj3, weights = "ordinal")
```

will print the Figure 3.5.

Let's assume we are interested in printing the interpretation value for Landis and Koch's (1977) scale (Table 3.2) of interpretation for Conger's κ, which uses ordinal weight. For this aim, let's initialise the variable `conger_ordinal` with the command `conger.kappa.raw(ambiguity_j1andj3, weights = "ordinal")`. This operation can be done with the following command:

```
conger_ordinal <- conger.kappa.raw(ambiguity_j1andj3, weights
```

coeff.name	pa	pe	coeff.val	coeff.se	conf.int	p.value	w.name
Conger's Kappa	0.7810945	0.7204277	0.217	0.12372	(-0.03,0.464)	0.08407895	ordinal

$weights

1.0000000	0.6666667	0.0000000
0.6666667	1.0000000	0.6666667
0.0000000	0.6666667	1.0000000

$categories

1 2 3

Figure 3.5: Example of ordinal weights for κ.

```
    = "ordinal")
```

Landis and Koch's (1977) scale of interpretation can be printed with the command:

```
landis.koch.bf(conger_ordinal$coeff.val, conger_ordinal$coeff.se)
```

coeff.name	pa	pe	coeff.val	coeff.se	conf.int	p.value	w.name
Conger's Kappa	0.7810945	0.7204277	0.217	0.12372	(-0.03,0.464)	0.08407895	ordinal

	Landis-Koch	CumProb
(0.8 to 1)	Almost Perfect	0
(0.6 to 0.8)	Substantial	0.00098
(0.4 to 0.6)	Moderate	0.06955
(0.2 to 0.4)	Fair	0.55465
(0 to 0.2)	Slight	0.96028
(-1 to 0)	Poor	1

Figure 3.6: Example of scale of interpretation in the case of ordinal weights for κ.

The Library irrCAC also implements the Altman scale of interpretation (Altman, 1990) with the command `altman.bf()` and Fleiss scale of interpretation (Fleiss, 1981) with the command `fleiss.bf()`.

The commands for measuring the S, π, α and AC_2 agreement coefficient in the case of raw data are:

- S: `bp.coeff.raw()`
- π: `fleiss.kappa.raw()`
- α: `krippen.alpha.raw()`
- AC_2: `gwet.ac1.raw()`

As in the example we show that the κ coefficient can be measured with the command

`conger.kappa.raw()`.

For further details we refer to `https://rdrr.io/cran/irrCAC/` and (Gwet, 2014).

3.8 Conclusion

Both the minimal use of reliability studies in the evaluation phase and the lack of a common practice in the use of coefficients of agreement emerge from the analysis we presented in Section 2.2.2. From that analysis we can conclude that, more than twenty years after Carletta's (1996) work, and more than ten years after Artstein and Poesio's work, the use of the coefficients of agreement in the NLG community is still not adopted as a standard.

In order to limit, and hopefully, eliminate this trend in this Chapter we have presented five coefficients of agreement and their interpretation. We compared all the five coefficients with the QG-STEC evaluation dataset. We presented a new approach to interpret the coefficients of agreement as proposed by Gwet (2014) and we indicated how to perform an agreement study. Finally, we presented the *irrCAC* library developed in the statistical software *R*.

Our desire is that this Chapter will create awareness in the NLG commu-

nity about how to handle coefficients of agreement in the intrinsic human evaluation phase.

Chapter 4

Detecting sources of annotation disagreement

In this chapter we investigate possible reasons for low levels of IAA that are reported in most papers we surveyed, following up on the findings from Chapter 2. For this purpose, we performed an observational study to uncover sources of annotation disagreement. The study was conducted on data from intrinsic human evaluation of AQG systems.[1]

Our strategy was to look for sources of disagreement by making judges discuss (based on some examples) their perception of a number of criteria used for question evaluation. For this aim, we started by defining an evaluation guideline and tried to refine the criteria chosen through four iterations of discussions and evaluations. Each iteration allowed us, based on the evaluation example, to track sources of disagreement raised by the discussion. During these iterations we noticed that regardless of how many changes we made,

[1] More specifically, we perform our study on the Text2Text task. We recall that in this thesis with Text2Text we refer to the task of generating a question in relation to a given input paragraph. For more details, we refer to Section 2.1.1. Although we focused on the AQG task, we believe the causes we discovered behind judges' disagreement are more general, and they are likely to apply to different NLG tasks.

there remained a divergence in the judgements that we could not reduce by modifying the guidelines, whilst retaining the spirit of the exercise.

Such divergences — which we describe in Section 4.2 — show that there are factors inherent to language evaluation that set limits to the degree of human agreement. Our study suggests that subjective interpretation bias has to be to take into account in the study of the reliability of intrinsic human evaluation of NLG systems. In Chapter 5 we propose a way to do so.

4.1 Description of the observational study

Our observational study was performed with four annotation iterations. In this section, we will refer to them as I1 (iteration 1), I2 (iteration 2), I3 (iteration 3) and I4 (iteration 4). Each iteration aims to mimic the way annotation guidelines are refined. In annotation efforts, annotation guidelines are defined based on an iterative process — see, for example, Pustejovsky and Stubbs (2013) and Artstein (2017). This process, which uses IAA as an evaluation metric, is based on a loop of evaluation of guidelines and guidelines improvement which can last a long time. The loop ends when a reasonable level of IAA is reached. The assumption here is that the higher the IAA the more reliable the annotation guidelines. Consequently, the loop will end when a high level of IAA is reached.

Method

All the iterations were performed independently by the judges. The judges were asked to perform the annotation by filling in a paper survey provided with the guidelines. At the end of each iteration, I discussed the adequacy of the criteria with the judges involved in the evaluation. This enabled us to understand and reduce the sources of disagreement. The discussions

were initially more informal but became more structured throughout the iterations. This was due to the difficulty in reaching agreement between the judges. In the end, the judges were asked to discuss their responses for each criterion and for each question, one by one. After each iteration we attempted to improve the criteria by clarifying their descriptions, adding clarifying examples and splitting them into sub-criteria.

Participants

The participants of the study were nine volunteers. All of them were studying or employed at The Open University. Among them, five were PhD students and four lecturers. Only three of the judges had a background in linguistics. Among the nine judges, three were native English speakers, the others were proficient in English. We decided to perform the iterations with a mixed set of judges (English speakers and non-native English speakers) to take into account the variation of English use. The use of judges without linguistic background allows us to take into account judgement of people that were not aware of problems linked to human evaluation of sentence quality. This allowed us to mimic evaluations done with non-experts and with non-native speakers of the language under examination. Such circumstances can be often encountered, for instance, in evaluations that use online platforms (e.g. Amazon Mechanical Turk).

I1, I2 and I3: We asked two volunteers to join us in the process of annotating the questions. Both volunteers were Spanish, who had lived several years in the UK and US, but who had not formally studied linguistics. Each iteration was performed with four people: the two supervisors of this research and the two volunteers.

I4: This time seven judges, including the two supervisors, engaged in the annotation task. Between the seven judges, three were native speakers of

English. The other five were proficient in English. Three of the judges had a background in linguistics.

Questions selection

Each iteration was performed with a new set of questions. All the questions were selected or modified by myself (I did not take part in any iteration). The participants were not aware of any kind of decision that was made in the questions selection.

We tested our evaluation guidelines, taking random input paragraphs and questions from the SQuAD dataset (Rajpurkar et al., 2016). The SQuAD dataset is a Reading Comprehension dataset.[2] It includes 107785 question/answer pairs from individual paragraphs extracted from 536 English Wikipedia articles. All the questions are posed by crowdsourced workers, whereas the answers are spans or text segments from the paragraphs from which the question is taken.

I1, I2 and I3: In each iteration we took five paragraphs and for each paragraph six questions. All the questions were human generated. Of the six questions, three were taken from the reference questions of the input paragraph (for one of these we automatically swapped some words in order to change the grammaticality), and three were about the topic of the input paragraph, but originally written for a different paragraph as an input (but with similar topic). After these three iterations, we concluded with a main iteration which involved a larger number of annotators.

I4: Once again we used the input paragraphs from the SQuAD dataset. This time we took two paragraphs and five questions for each paragraph. One of the questions was automatically generated by the Karen Mazidi's

[2]The dataset is available at: `https://rajpurkar.github.io/SQuAD-explorer/`.

question generator algorithm (Mazidi, 2018). Three were human generated questions from the SQuAD dataset: two were taken from the reference questions of the input paragraph (for one of these we automatically swapped some words in order to change the grammaticality), and one was about the topic of the input paragraph, although generated with a different input paragraph. I wrote the last question (I did not personally take part in the evaluation). The last question aims to verify whether the judges were paying attention to the guidelines. Indeed, this question was a clear violation of an example used to define the criteria.

Criteria selection

We started by choosing the criteria to use. For this aim, we took inspiration from the evaluation guidelines defined in Godwin and Piwek (2016). We decided to study the evaluation along two dimensions, a linguistic one, which aimed to evaluate the question quality from a grammatical and idiomatic point of view, and a task-oriented one, which aimed to evaluate how well the questions fulfil the task for which they were generated. In the task-oriented dimension we were interested in the degree to which the question was answered by the input text. The annotation criteria attempted to characterise these two dimensions.

The guidelines used in the four iterations are reported in Appendix A. In this subsection we present the criteria used in the iterations and the main rationale that brings us from one iteration to the next one. Once again we refer to Appendix A to explore in detail the changes in the criteria definition and examples.

I1: For this guideline (see Section A.1) we define the following criteria:

- Syntactic correctness and fluency;

- Specificity;
- Relevance.

The *relevance* criterion was for the task-oriented dimension. It was ranked on a scale from 0 to 3. The other two criteria were defined for the linguistic dimension. The *syntactic correctness and fluency* was ranked on a scale from 0 to 3. The *specificity* was ranked on a scale from -2 to 0.[3]

I2: After the discussion relative to I1 we defined the following criteria (see Section A.2):

- Syntactic correctness and fluency;
- Specificity;
- Pertinence.

We changed the name *relevance* to *pertinence* in order to underline the need for the question to be directly and strictly related to the paragraph. The judges felt that the name *pertinence* described the criterion aim more accurately than *relevance*. This time, the *specificity* criterion was ranked on a scale from 0 to 2. The judges found the negative scale confusing. Regarding the *syntactic correctness and fluency*, we reduced the scale categories from 0 to 2. The judges were confused about the difference between the categories 2 and 1.[4] For I2 we have joined them together in category 1 by giving a new definition for that category.

[3]We use negative numbers because we thought of the *specificity* criterion as a refining of the *relevance* criterion. The idea was to catch the following: "it could happen that, although the question is relevant to the input paragraph, we can use the same question in more than one paragraph. That is, the question is so general that can be correctly answered in several, or without, contexts." For this reason, we consider the value associated with *specificity* as a value to be subtracted from the value associated with the question *relevance*.

[4]Category 2 was described as " The question is grammatically correct but does not read as fluently as we would like". Category 1 was described as "There are some grammatical errors in the question". The judges found it difficult to distinguish the presence of "some grammatical errors" from "the question did not read as fluently as we would like".

I3: The discussion raised by I2 led us to define the following criteria (see Section A.3):

- Syntactic correctness;
- Comprehensibility;
- Fluency;
- Pertinence.

The judges suggested that the *syntactic correctness and fluency* could be better characterised if split into three binary criteria:

1. *Syntactic correctness*;
2. *Comprehensibility*;
3. *Fluency*

The *syntactic correctness* is intended to check possible grammatical errors in the questions. *Comprehensibility* aims to better frame the grammaticality judgement. Indeed, it can happen that a question, although ungrammatical, is perfectly understandable. That is, it can be possible to work out what the question is asking, even if it is ungrammatical.[5] Finally, the *fluency* criterion is intended to judge whether the question is idiomatic/natural.

Regarding the *specificity* criterion the judges found it redundant, in the sense that it can be covered by the *pertinence* criterion. For this reason we decided to remove it.

I4: In the last version of our evaluation guidelines (see Section A.4), we had the following four criteria:

- Pertinence;
- Grammaticality;

[5] In Section 4.2 we will present some examples of these cases.

- Comprehensibility;
- Fluency.

The name *syntactic correctness* was changed to *grammaticality*. The judges found this a better description of the criterion aim. The judges also suggested moving the numerical categories 0/1 to the linguistic category *no/yes*. They considered such categories more intuitive. Furthermore, we decided to evaluate the linguistic dimension (*grammaticality*, *comprehensibility* and *fluency*) without providing the paragraph from which the questions were generated. However, the *pertinence* was judged after the paragraph was provided to the judges. Indeed, we believe that the *grammaticality*, *comprehensibility* and *idiomaticity* of a question can be judged independently of the paragraph from which it was generated. So each annotation was carried out in two stages. First, the judge was presented with the question in isolation, and asked to provide a judgement on the questions of *grammaticality*, *comprehensibility* and *fluency*. Next, the judge was presented with both the question and the input text, and asked to provide a judgement on the *pertinence*.

Results

The aim of each annotation was to find annotation divergences to be discussed in order to reduce them. In our study, of the 100 questions we analysed, the judges discussed a total of 55 divergences in the linguistic dimensions and a total of 78 divergences in the task-oriented dimension. After each iteration we checked the agreement by the use of Conger's κ (Conger, 1980). We used Conger's κ because we aimed to check to what extent each annotation was affected by the judges' bias. To this aim, the κ-style agreement coefficients are more adequate. [6]

[6]For further details, we refer to Section 3.6.1. We note that for the aim of improving annotation scheme and guideline the π-style agreement coefficients are usually used. Nev-

The higher level of agreement was reached in I3 for the syntactic correctness criterion. For that annotation the κ was 0.64. The lower level of agreement was reached in I1 for the specificity criterion. For that annotation the κ was 0.15. In all the annotations the average values reached by all the criteria were: 0.36 for syntactic and fluency, 0.46 for relevance (pertinence), 0.16 for specificity and finally 0.39 for comprehensibility.

Because of the small number of sentences in each annotation we were aware of the low statistical significance of the measurement. Nevertheless, it gave us some quantitative information to be added to the qualitative analysis of the annotation divergences that we were interested in.

The main results of our study identified a series of differences. As we will see in the next sections, they were based on judges' subjective taste and experiences.

4.2 A new taxonomy of divergences

We classify the main sources of disagreement we found through our study in five categories:

1. *Style and taste*;
2. *Background knowledge*;
3. *Personal assumptions*;
4. *Use of common sense inferences*;
5. *Attention to detail.*

ertheless, κ-style agreement coefficients suit better the aim of our experiment — that is, understanding the source of annotation disagreement by discussing annotation divergences among the judges. In our experiment, we were interested in an agreement measure that takes into account the fact that each judge has his(her) own probability distribution over the categories in play.

We present some examples of these divergences in order to explain how the above categories were chosen.

4.2.1 Style and taste

We found some divergences were related to the judge's taste and their writing style. This kind of divergence emerged in the question on idiomaticity judgements. For example, we notice that American and British English differences played against question idiomaticity. Given the question:

Jean De Rely's illustrated French-language scriptures were first published in what city?

we found divergent judgements because the question sounded awkward to some British judges, who preferred the use of "which" rather than "what". Similarly, we had divergent judgements with the question:

The adaptive immune system must distinguish between what types of molecules?

where one of the British judges not only preferred the use of "which" instead of "what", but also marked the question as not idiomatically natural because he preferred the question written in the reverse order: "Which types of molecules must the adaptive immune system distinguish between?" In a case like this, we can see how the personal writing style influences the evaluation. This eventuality is a result of the fact that, in a generation task, we can use very different sentences, in this case questions, in order to reach the same communicative goal.

4.2.2 Background knowledge

Another issue we found concerned the amount of background knowledge that is needed in order to understand a question. Where judges did not share relevant background knowledge, they often gave different judgements for a question's comprehensibility and pertinence. For example the question:

How many Time incarnations can a Lord have?

even if not grammatical (the original question is: How many incarnations can a Time Lord have?), was recognised as comprehensible by the judges who were aware of the 'Doctor Who' television programme, whereas it was marked as not comprehensible by the ones who did not know the show. In the same way, given the paragraph:

Microorganisms or toxins that successfully enter an organism encounter the cells and mechanisms of the innate immune system. The innate response is usually triggered when microbes are identified by pattern recognition receptors, which recognize components that are conserved among broad groups of microorganisms...

the question:

What part of the innate immune system identifies microbes and triggers immune response?

raised divergent opinions for two reasons: on the one hand, it was necessary to know that microorganisms are the same as microbes. On the other, it

was necessary to know that the pattern recognition receptors are part of the innate immune system. The lack of this knowledge from some judges resulted in a divergence in ranking the pertinence criterion. In these examples, we can see how the judges' knowledge determines the evaluation. Indeed, personal knowledge and experiences are reflected in the way we generate and understand sentences. So, divergences in knowledge can be reflected in evaluation divergences.

4.2.3 Personal assumptions

Other divergences arose from different judge assumptions. Also in these cases the divergences emerged predominantly in the question's comprehensibility and pertinence judgements. For example the question:

Which team finished the regular season?

was considered not comprehensible by the judges who assumed that all teams would finish the season. These judges considered the question meaningless. In contrast, other judges considered a scenario in which some teams could not finish the season. In this case the question was marked as comprehensible. A similar problem was found with the question:

What was the win/loss ratio in 2015 for the Carolina Panthers during their regular season?

Given the paragraph:

The Panthers finished the regular season with a 15-1 record, and quarterback

one judge noticed that the question can be marked pertinent under the assumption that the Carolina Panthers in the question and the Panthers in the paragraph are referring to the same team. He did not make this assumption. Other judges did make this assumption and ranked the question as pertinent. In a similar way to background knowledge and experiences, personal assumptions are reflected in the way we generate and understand sentences. Again, divergences in assumptions can be reflected in divergences in evaluation.

4.2.4 Use of common sense inferences

This kind of divergence emerged predominantly in the question pertinence judgements. For example, given the paragraph:

The availability of the Bible in vernacular languages was important to the spread of the Protestant movement and development of the Reformed church in France. The country had a long history of struggles with the papacy by the time the Protestant Reformation finally arrived. Around 1294, a French version of the Scriptures was prepared by the Roman Catholic priest, Guyard de Moulin. A two-volume illustrated folio paraphrase version based on his manuscript, by Jean de Rély, was printed in Paris in 1487.

the question:

Jean De Rely's illustrated French-language scriptures were first published in what city?

was marked as not pertinent by one judge who noted that the paragraph provides information about the place where Jean De Rely's scriptures were printed and not where they were published. To consider this question as unambiguously answered by the paragraph it is necessary to assume that the place where the Jean De Rely's illustrated scriptures were printed is the same where they were published. Although this inference goes much further than the information presented in the text, other judges made it and marked the question as pertinent. Likewise, the question:

When did the first French language bible appear?

was marked as pertinent by one judge, who inferred the answer 1294 from the paragraph. Other judges considered the question as not pertinent because the answer cannot be correctly inferred from the paragraph. As a matter of fact, in the text there is no mention of the fact that the Guyard de Moulin version of the Scriptures was the first French language bible to appear and at the same time it is not possible to rigorously reach this conclusion by the information provided by the text. Here we faced another problem: people reason in different ways. Obviously this divergence also emerges in the way people generate, understand, and in this case evaluate, sentences.

4.2.5 Attention to detail

We noticed that in some cases, judges overlooked some question details. Also in these cases, the divergences emerged predominantly in the question pertinence judgements. For example, given the paragraph:

The most impressive examples of rococo architecture are Czapski Palace (1712-1721), Palace of the Four Winds (1730s) and Visitationist Church

(façade 1728-1761).

the question:

What type of architecture is the Palace of Four Windows an impressive example of?

was ranked as pertinent by some judges who did not notice that the question uses the word "Windows" instead of "Winds", but it was ranked as not pertinent by the judges who did notice this detail. Another example of this kind of rank divergences is the following. Given the paragraph:

Victorian farms produce nearly 90% of Australian pears and third of apples. It is also a leader in stone fruit production. The main vegetable crops include asparagus, broccoli, carrots, potatoes and tomatoes. Last year, 121,200 tonnes of pears and 270,000 tonnes of tomatoes were produced.

the question:

How many tonnes of tomatoes does Victoria produce?

was ranked as pertinent by some judges who did not notice that the paragraph was speaking about "last year" production, whereas the question is more general (maybe it is asking for an annual average, which is not mentioned in the paragraph). Other judges did notice this detail and ranked the question as not pertinent. This problem is linked to the fact the people can generate and interpret sentences to different levels of generality and detail.

This is reflected in the way people understand sentences, in this case the questions, in the context from which they are generated.

4.2.6 Other sources of disagreement

We conclude this section by observing that together with the five sources of disagreement we have just presented, we found other sources of discrepancy between judges — which we suppose are present in each annotation task — related to distractions, misunderstanding the guidelines, or forgetting how to apply the guidelines.

In our study, we found such sources of disagreement impact grammatical judgements the most. For instance the question:

Whic NFL team represented the NFC at Super Bowl 50?

was marked as grammatical by one judge who did not pay attention to the explicit request of considering ungrammatical the questions which contain those errors that can look like a typo, in this case the word "Whic".

In another case, one judge misunderstood the guidelines in two respects and marked as ungrammatical the following question:

What was the win/loss ratio in 2015 for the Carolina Panthers during their regular season?

In this case, the judge mixed up the instructions for grammaticality and fluency. Furthermore, the judge misunderstood the guideline in another respect. Indeed, the instruction for assessing the fluency criterion explicitly stated that American and British English differences don't count against question fluency. Nevertheless, in the discussion that followed the annotation, the judge cited his(her) preference for British English as the reason for considering the sentence ungrammatical.

4.2.7 Concluding remarks

Our study aimed to detect sources of annotation disagreement by discussing with the judges the divergences that emerged during four annotations. Each annotation was part of an iterative process that mimics the way annotation guidelines are refined.

In I1 we found that the main source of disagreement was linked to the "specificity" criterion (the judges were in disagreement about this criterion for 50% of the questions) and for the "syntactic correctness and fluency" criterion (the judges were in disagreement about this criterion for 46% of the questions). For the "specificity" criterion the main source of disagreement was linked to background knowledge, attention to detail, personal assumptions and use of common sense inferences. Such sources of disagreement were present also for the "relevance" criterion (the judges were in disagreement about this criterion for 30% of the questions). Regarding the "syntactic correctness and fluency" criterion it emerged that style and taste were the main sources of disagreement.

In I2 we did not detect anything different from I1. Our revised guidelines were able to reduce disagreement only slightly. More specifically, the judges were in disagreement about the "specificity" criterion for 46% of the questions, 40% for the "syntactic correctness and fluency" criterion and 23% for the "pertinence" criterion.

More interesting was I3. In this iteration we understood better the source of disagreement related to the "syntactic correctness and fluency" criterion by splitting it in three criteria: "syntactic correctness", "comprehensibility" and "fluency". From this iteration it emerged that "fluency" was the most problematic criterion. The judges were in disagreement about the "fluency" criterion for 43% of the questions, 20% for the "comprehensibility" criterion and 16% for the "syntactic correctness" criterion. Whereas disagreement on

the "comprehensibility" criterion was mainly due to background knowledge and personal assumptions, the disagreement on "fluency" was mainly linked to style and taste. Regarding the "syntactic correctness", we found that the main source of disagreement was linked to distraction and misunderstanding of the guidelines. Also in this case, we found that for the "pertinence" criterion the main sources of disagreement were due to background knowledge, attention to detail, personal assumptions and use of common sense inferences.

The sources of disagreement found in I3 were confirmed in I4.

Summing up, table 4.1 summarises our main results. As expected the most subjective criteria, pertinence and comprehensibility, are the most affected by judges' bias.

	Pertinence	Grammaticality	Comprehensibility	Fluency
S&T	no	no	no	YES
BK	YES	no	YES	no
PA	YES	no	YES	no
CI	YES	no	no	no
AD	YES	no	no	no

Table 4.1: Main source of disagreement for criteria found in our observational study. The label in the first column refers to the five categories delineated in this section: S&T (Style and taste), BK (Background knowledge), PA (Personal assumptions), CI (Use of common sense inferences), AD (Attention to detail).

The importance of style and taste in the perception of sentence fluency emerges from our study. This shows the difficulty in defining the concept of fluency, which, in a way, can be considered an aesthetic issue.

Regarding the grammaticality criterion, we found that the main sources of disagreements were linked to distraction and misunderstanding the guidelines. This finding is in line with Sampson and Babarczy (2008)'s study. Although we note that the background knowledge can also have an impact

on grammatically, especially with non-native speakers. Style and taste can be another source of disagreement with guidelines that merge grammaticality and fluency.

4.3 Conclusion

Many studies — see for example: Sampson and Babarczy (2008) and Bayerl and Karsten (2011) — have shown IAA appearing to depend on several factors, for example: data typology, data ambiguity, the number of categories used in the annotation, the number of annotators, the use of expert or non-expert annotators, and annotators' attention, skills, memory and training. In this chapter we have shown that, in a language generation task such as Text2Text, we also need to add the following factors: the annotators' background knowledge, their personal taste and personal assumptions, their attention to detail and their inferential skills.

The conclusion we draw is that judges' disagreements are part of the nature of language. We believe that attempts to create evaluation guidelines that greatly reduce disagreement among annotators are in danger of missing the goal of producing human language variability in NLG intrinsic human evaluation tasks. Rather than regarding the evaluation disagreement as a problem to be fixed, we suggest that it should be thought of as an ineliminable feature of generation tasks, which reflects the variety of human languages and their uses. This has to be taken into account by every intrinsic human evaluation reliability study.

As we said before, since Carletta (1996)'s paper, the standard measure to calculate reliability in annotation efforts is the measure of IAA. Traditionally, the Kappa statistics are used to meet this aim. Although such statistical measures take into account some level of randomness introduced by annotators, it might be that they are not generally adequate for NLG evaluation

purposes, where a high level of subjectivity is involved. Most data collected for NLG evaluation fail the reliability test, according to the existing scales of IAA interpretation presented in the previous chapter.

In the next chapter, given human language variability, we propose that for human evaluation of NLG the concept of judge consistency should be used together with the concept of agreement to obtain a better assessment of the evaluation data reliability.

Chapter 5

From agreement to consistency

The observational study we presented in Chapter 4 shows us that the tasks of intrinsic human evaluation of NLG systems suffer from subjective bias. Such biases are part of the way humans interpret and use language. This has to be taken into account in the reliability study of human evaluation. Evaluation results have to inform generation system developers of the extent to which they can improve the communicative power of their systems. As such, attempting to constrain human language use and interpretation runs the risk of biasing system developers against important aspects of human language.

The measuring of judges' agreement gives information about the extent to how similar (or dissimilar) their interpretations of the phenomena annotated are. However it cannot be exhaustive for a reliability study in tasks in which a high level of subjectivity is involved. In this case, the low level of agreement between judges cannot in itself make an evaluation lose reliability. We suggest that the concept of judge consistency plays a pivotal role in

the annotation reliability analysis in annotation tasks where a high level of subjectivity is involved.

At least two kinds of judges' consistency can be outlined:

- Consistency as stability (or Intraobserver Agreement (IA)).
- Relative consistency.

In this chapter, we discuss these concepts. The main contribution of the chapter is the introduction of the idea of relative consistency.

5.1 Consistency as stability

We recall that stability aims to measure the extent to which an annotation remains unchanged over time. More precisely, stability measures whether judges annotate an item in the same way at different times. Stability involves annotating the same items after some time has elapsed. Such a procedure requires judges to annotate the same items in different trials. Krippendorff (1980) presents Stability as the weakest measure of data reliability.

In order to measure Stability, at least two procedures can be carried out:

- IR1: Annotating the same item multiple time in the same session
- IR2: Annotating the same item in two sessions, with the second happening some time after the first.

Both the procedures IR1 and IR2 have pros and cons. Which of the two used depends on the task at hand.

Because the annotation is performed just once in the case of IR1, the result can be obtained quickly. Nevertheless, one can encounter a memory problem when IR1 is performed (Streiner et al., 2015). Indeed, judges might remember the annotation performed some steps before, thus weakening the aim of measuring the IA. Such problems can be mitigated by the use of several

items. The more items are seen by the judges, the lower the probability of remembering the items seen before. However, such a strategy raises other problems. The more items, the longer the time needed to perform the annotation. This, for instance, can be a cause of carelessness, distractions and fatigue in the judges, resulting in a loss of concentration and a relaxing of performance standards. Another strategy to mitigate the memory problem is to test the same phenomena twice by using two different items placed in a different point in the annotation. That is, the two items have to be considered in alternative forms to reach the same information. For example, suppose we are interested in the fluency of a sentence "S". Then the following items can be used:

On a scale from 1 (no fluent) to 5 (fluent), is the sentence "S" fluent?

1 2 3 4 5

On a scale from 1 (non native speaker) to 5 (native speaker), could the sentence "S" have been produced by a native speaker?

1 2 3 4 5

When this strategy is performed at least two points have to be taken into account.

Firstly, the assumption that the items are measuring exactly the same phenomena is vital. This is something that has to be clearly stated. For instance in the example above, it is assumed that native speakers can produce fluent sentences.

Secondly, researchers have to think about the phenomena they are interested in annotating. Indeed, the time used to annotate a phenomenon with a second description could instead be used for annotating another phenomenon. However, other methods can also be used. For instance, in a case in which several phenomena have to be annotated, the IA can be measured only for

one or two of them. In this case, it is assumed that the judges annotate all the phenomena, and the phenomena used to measure the IA are to be considered a general measure of the annotation' stability.

The performance of IR2 can reduce the memory problem. The greater the time between one annotation and the other, the greater the probability that the judges will not remember the items annotated previously. Nevertheless, other problems can emerge. Indeed, the time interval between tasks has to take into accounts two factors: i) the memory problem; ii) the problem of the change of the underlying trait.
On one hand, the judges shouldn't recall the annotation performed previously. If there is a short time interval between the two annotations, judges can remember the answer used in the annotation performed previously.
On the other hand, the judges should not change the trait which is at the base of their annotation decision. If there is a long time interval between the two annotations, the judges may change their understanding and interpretation of the phenomena annotated. For instance, judges could learn something new about the phenomena annotated. Likewise, the greater the time interval the higher the probability that the judges used in the annotation can speak amongst themselves.[1] In this eventuality, their underlying trait can be changed because of that conversation — one judge can influence the interpretation that the other judge has of one phenomenon annotated.
When the IR2 procedure is performed, researchers should justify the interval time chosen for reassessing the annotation.

We suggest using an agreement coefficient (for instance, some sort of kappa statistic) to measure the IA, regardless of the procedure used. For each judge, the IA can be measured by computing an agreement coefficient between two scores for the same item. We suggest using the agreement coef-

[1]This eventuality is particularly true in the case of colleagues or students used for the annotation.

ficients because we are interested in knowing the judges' consistency over time. As long as the annotations we are considering in this thesis are NLG evaluations, where criteria such as fluency are taken into account, we require that the judges have the same criteria interpretation, at least in a short period of time. In other words, we expect that judges provide strictly identical results on the two annotation occasions. Nevertheless, if researchers are performing an evaluation in which the IA takes into account possible differences between the two annotations — that is, the annotation allowed that judges change their interpretation of the criteria at play from an annotation to the other one —, then the correlation coefficients (for instance, Pearson's r and Spearman's ρ) will be a suitable choice.

In conclusion, the choice between agreement coefficients or correlation coefficients depends on the task at hand. The difference between agreement coefficients and correlation coefficients will be presented in Section 5.2.1. Also, in this case, researchers should justify the coefficient chosen to measure the IA.

5.2 Relative consistency

The observational study we presented in Chapter 4 gives us an insight into the reason why many reliability studies of NLG human evaluation tasks that use kappa statistics fail the reliability test. Kappa statistic coefficients are designed to get high value when the judges choose the same category. However, given the high level of subjectivity involved in the evaluation, such a requirement looks to be too demanding.

In this section we are going to further develop this point and suggest a way to address this problem. The main aim of this section is to suggest that for human evaluation of NLG, judges' relative consistency and judges' agreement should be used together to obtain a better assessment of the

evaluation's reliability.

We propose the use of correlation coefficients in order to measure judges' relative consistency.[2]

5.2.1 The use of correlation coefficients for NLG human evaluation tasks

Correlation coefficients are generally considered inappropriate for measuring the reliability of annotated data; see for example, Lombard et al. (2002), Krippendorff (2004) and Artstein and Poesio (2008). The main concern about the use of correlation coefficients for reliability studies is well expressed by the following quotation:

> [Correlation coefficients, for example Pearson's r] measure the extent to which two logically separate interval variables, say X and Y, covary in a linear relationship of the form $Y = a + bX$. They indicate the degree to which the values of one variable predict the values of the other. Agreement coefficients, in contrast, must measure the extent to which $Y = X$. (Krippendorff, 1980, p. 244)

Indeed, the rationale behind agreement coefficients, such as the Kappa statistic, is to catch the extent to which judges rank a given item equally. When judges rank a given item in the same way, it is assumed that the judges share the same interpretation and understanding of the schema and guidelines used in the annotation task. When this happens, given the fact that the annotation is reached with judges that work independently, the concept of reliability as reproducibility[3] suggests that the same annotation can be reached with other judges. This makes the annotation repeatable

[2]Some examples of correlation coefficients are Kendall's τ, Pearson's r, Spearman's ρ and Goodman and Kruskall's Gamma.

[3]For more detail we refer to Section 1.4.2.

and consequently reliable.

Although such a concept of reliability as reproducibility is well-founded in cases where the phenomenon under investigation has some objective meaning, for example in the case of CL annotation tasks where the gold standard is available (for instance part-of-speech tagging), it falls short in the case of NLG evaluation tasks. In the case of NLG, where the existence of a gold standard is mostly not available – for example, criteria such as ambiguity, relevance, usefulness or overall quality – the concept of reliability as reproducibility can hide some pitfalls. For evaluation tasks that aim to evaluate semantic or pragmatic language aspects – for instance concepts such as text usability, fluency, comprehensibility etc. – two people can entertain different, although equally valid, opinions. In cases such as these, given the variability of human language – specifically variability in language interpretation and quality judgement – expecting judges to always arrive at exactly the same judgement may be both unrealistic and over-constrained. Variation in language interpretation and use makes strict agreement unsuitable for measuring human evaluation reliability.

In this thesis, we argue that, from an evaluation point of view, what is important, more than the fact that judges have the same interpretation of the phenomena studied, is to know whether the judges are consistent relative to each other. A possible first step to test this is checking judges' relative consistency, that is checking whether the judges follow a systematic pattern in their assessments.

A feasible strategy to frame this problem is the following. Expecting judges to always arrive at exactly the same judgement may be unrealistic. For instance, one judge may be stricter than another one. However, in such situations the judgements would still covary. In other words, we can ask: Is it possible to predict J_a's judgements based on J_b's judgements, where J_a

and J_b are two judges who are judging the same set of sentences?

Correlation coefficients can be used to answer this question. Such coefficients measure to what extent a variable changes, in a way not expected by chance alone, in relation to the change of another variable. That is, they measure the covariance of two variables. The change can be either in the same (positive correlation) or in the opposite (negative correlation) direction. In the presence of correlation, given a judges' annotation, it is mostly possible to predict the annotation of another judge. Correlation coefficients, measuring the judges' covariance, can give an insight into to the extent different judges are consistent relative to each other when annotating the data, even when their individual interpretations of the phenomena are not identical but following a consistent pattern, see for example Stemler and Tsai (2008, page 38) and Gisev et al. (2013, page 331).

To test such an interpretation of correlation coefficients, we use the following three datasets:

- the data collected in the last iteration (I4) we used in the observational study of the Chapter 4;[4]
- the evaluation QG-STEC datset;
- the Flickr-8k datset.

We recall that the observational study aimed to find the sources of judges' disagreement. Our study was based on a number of criteria used for question evaluation. The methodology we used was that of refining the criteria chosen through several iterations of discussions and evaluation. During these iterations we noticed that regardless of how many changes we made, there remained a divergence in the judgements that we could not reduce by modifying the guidelines. Nevertheless, we realized that such divergences showed

[4]The dataset can be found online (Amidei, 2019). The evaluation guideline can be found in the Appendix A.

an interesting degree of consistency, due to the fact that the judges were consistent in following their interpretation of the criteria at play. The iteration study (I4) we use in this chapter, although consisting only of ten items, helps to formalize the problem and makes it clear from a visual point of view. Indeed, the use of ten items allows a clear visualization of the data. Although judgements are different in values, they show a clear pattern – see Figure 5.2 and Figure 5.1. Once we test the use of correlation coefficients in the iteration study (I4) we scale the experiment by the use of larger datasets, QG-STEC and Flickr-8k, that allow significant statistical conclusions.

5.2.2 Datasets analysis

Following Siegel and Castellan (1988) and Singh (2007) we use Goodman and Kruskal's Gamma as a correlation coefficient (Goodman and Kruskal, 1954). Goodman and Kruskal's Gamma is the most adequate coefficient for ordinal data with many ties which is exactly our case.[5] For binary categorical data we use Yule's Q (Yule, 1912), which is a special case of Goodman and Kruskal's Gamma.

Goodman and Kruskal's Gamma was measured with the *GoodmanKruskalGamma* function supplied by the *R* software.[6] Yule's Q was measured with the *pycm* library.[7]

We interpreted the correlation values by the use of the scale for correlation coefficient introduced by Rosenthal (1996) (see Table 5.1). We chose this scale because it extends Cohen's popular scale (Cohen, 1988). More precisely, it allows a more fine grained value distinction for the interval $[0.50, 1]$ – in particular, Rosenthal's scale specifies Cohen's "large" interval $[0.50, 1]$

[5]Data with ties are data with value repetition.

[6]The documentation for this function can be found at: `https://www.rdocumentation.org/packages/DescTools/versions/0.99.19/topics/GoodmanKruskalGamma`.

[7]For further details, we refer to `https://pypi.org/project/pycm/`.

into the two intervals "large" $[0.50, 0.7]$ and "very large" $[0.70, 1]$.[8] Because Goodman and Kruskal's Gamma (Yule's Q) tends to give higher values than other correlation coefficients, such a choice allows a finer-grained analysis.

Correlation value into the interval	Value interpretation
(-0.1, 0] \ [0, 0.1)	Negligible
(-0.3, - 0.1] \ [0.1, 0.3)	Small
(-0.5, - 0.3] \ [0.3, 0.5)	Medium
(-0.7, - 0.5] \ [0.5, 0.7)	Large
[-1, - 0.7] \ [0.7, 1]	Very large

Table 5.1: Rosenthal (1996) scale for the interpretation of correlation coefficient.

Regarding the coefficient of agreement we used Conger's κ (Conger, 1980). This coefficients is a generalization of Cohen's κ (Cohen, 1960). We used a Cohen's κ style coefficient, because our aim was that of validating the final human annotation.[9]

To measure Conger's κ we used the R software. More precisely we used the function *conger.kappa.raw()* supplied by the *irrCAC* library. We set the variable *weights* as "ordinal" for ordinal data and as "unweighted" for categorical data.[10]

In order to interpret the values obtained in the analysis, we use the Krippendorff scale of interpretation for agreement coefficients (Krippendorff, 1980, see Table 1.1). Since Carletta (1996), the Krippendorff scale of interpretation has become the standard for CL annotation tasks.

5.2.3 Interpretation of correlation coefficients case studies

Iteration (I4): Table 5.2 reports the results for the iteration study (I4). We observe that the native English speakers get higher IAA value and cor-

[8] For more details we refer to Section 5.3.
[9] For further details we refer to Chapter 3.
[10] For further details we refer to Chapter 3.

relation results than non-native English speakers. Quite interestingly, although for the fluency criterion, the IAA value reached by the native English speakers is below 0.4, they get a perfect correlation. Differently, non-native English speakers reached a correlation value of 0.27. This can be an indication that native judges have a different but strong interpretation of the concept of fluency[11]. Figure 5.1, which depicts the evaluation of question

Criteria	Coefficient	Dataset		
		All	Native	Non-Native
Syntactic	Conger's k	0.36	0.59	0.21
	Yule's Q	0.56	0.86	0.37
Comprehensibility	Conger's k	0.55	0.63	0.48
	Yule's Q	0.92	1	0.91
Fluency	Conger's k	0.30	0.39	0.17
	Yule's Q	0.56	1	0.27
Pertinence	Conger's k	0.20	0.22	0.07
	Goodman Kruskal's Gamma	0.57	0.47	0.48

Table 5.2: Results of Conger's κ and Yule's Q/Goodman and Kruskal's Gamma in the Iteration dataset. *All*, *Native* and *Non-native* indicate the measure performed respectively over the seven judges, over the three English native speaker judges and over the four no English native speaker judges.

fluency, can help to better understand this phenomenon.[12]

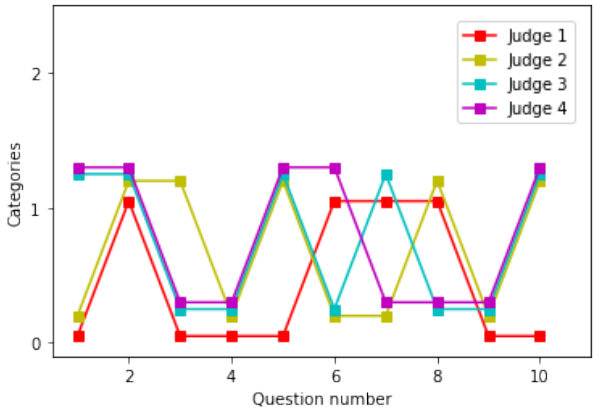
Regarding the case of native English speakers, Figure 5.1 (a) shows that Judge 5 systematically ranks with a value that is equal or less than the value given by Judges 6 and 7. In contrast, Figure 5.1 (b) shows that the ranks provided by non-native English speakers lack systematicity.

It is also worth noticing that for the cases of comprehensibility and perti-

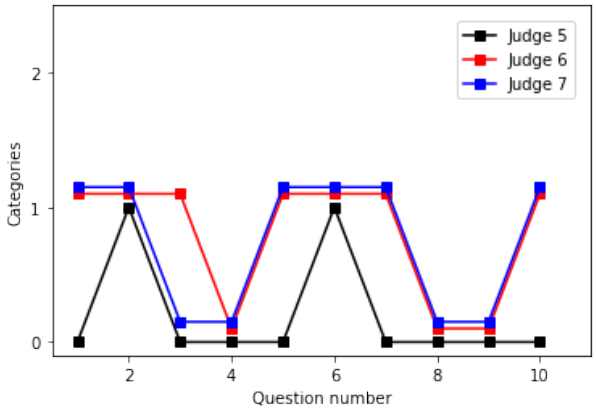
[11] It is worth noticing that in the case of no English native speaker, the Yule's Q measured on triple of judges reached the following values: 0.54, 0.29, 0.12, 0.12.

[12] In both Figure 5.1 and Figure 5.2 the x-axis represent the question number we used in the evaluation. We note that a different questions' numeration will result in a different graph representation. Nevertheless, the main point of the discussion is not affected by this. Indeed, the aim of the graph representation is to show the presence of possible patterns in the evaluation. Changing the question order does not change these patterns. Figure 5.1 (b) is enlightening in this respect. As we can see, Judge 5 (black line) constantly annotates equal to or lower than Judge 6 and Judge 7. Such a phenomenon will be present in every possible permutation of the questions.

nence, in the case of non-native English speakers, there is an interesting gap (more than 0.4) between Conger's κ value and the correlation coefficient value.



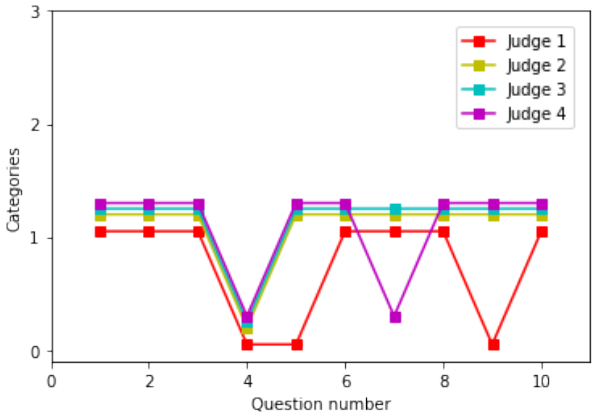
(a)



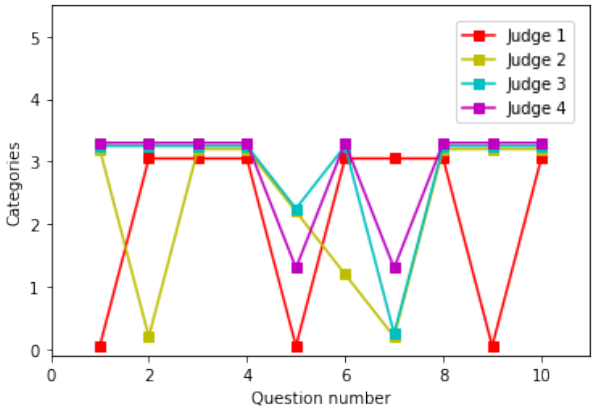
(b)

Figure 5.1: Plots of the evaluation of question fluency for non-native English speakers (a) and for native English speakers (b). For better readability, the scores are shifted upward slightly.

Figure 5.2 shows the judges' ranks are different in value, which explains the low Conger's κ for comprehensibility and pertinence criteria. However, there is systematicity in the judges' rankings – it is really clear in the case of comprehensibility (Figure 5.2 (a)), and less accentuated in the case of pertinence (Figure 5.2 (b)).



(a)



(b)

Figure 5.2: Evaluation of question comprehensibility (a) and pertinence (b). Both the annotation were performed by non-native English speakers. For better readability, the scores are shifted upward slightly.

Table 5.2 can be used to attempt a conclusion about the reliability of the dataset. Following the Krippendorff (1980) scale of interpretation, the evaluation data should be discarded because the IAA is below the threshold of 0.67. However, following the scale of interpretation for non-parametric correlation coefficients introduced in Rosenthal (1996), the data reach almost everywhere a large correlation, and a very large correlation in the case of native English speakers. Taking into account the interpretation we gave in Section 5.2.1, although the judges use different values in the evaluation,

their interpretations are constant in relation to each other: their judgements covary systematically with each other. This interpretation suggests that the data are reliable.

Flickr-8k: The Flickr-8K dataset contains quality judgements for 5,822 sentences.[13] Each sentence was a description of an image. The annotation was carried out by three human experts who judged the sentence semantic correctness on a scale from 1 to 4.

Because we don't have the information about how the data were collected, in order to decide which kind of analysis to carry out on the Flickr-8k dataset we plot the distribution of the categories used by the judges. Figure 5.3 suggests that the data do not have a normal distribution, and so we opt for the use of nonparametric statistics. We used Goodman and Kruskal's Gamma and Conger's κ to carry out our analysis. For Goodman and Kruskal's Gamma, we report the average results of the pairwise measure between the judges. This method is suggested by Siegel and Castellan (1988) for the case of Kendall τ correlation coefficient, which is a variant of the Goodman and Kruskal's Gamma.

The measurements give a κ value of 0.52 and a Gamma value of 0.98 (p-value < 0.05). Following the Krippendorff interpretation of IAA, the annotation has to be considered not reliable. However, the annotation achieves a very high correlation, which suggests a high relative consistency between the judges. Indeed, when they are in disagreement, Judge 2 ranks systematically higher than Judge 1, and Judge 3 ranks systematically higher than Judge 2. Although judges rank the items with different magnitude their judgement covary systematically. Also in this case, the correlation coefficient suggests the evaluation data are reliable.

[13]The dataset is available at: `https://github.com/elliottd/compareImageDescriptionMeasures`. For details on the original dataset we refer to Hodosh et al. (2013).

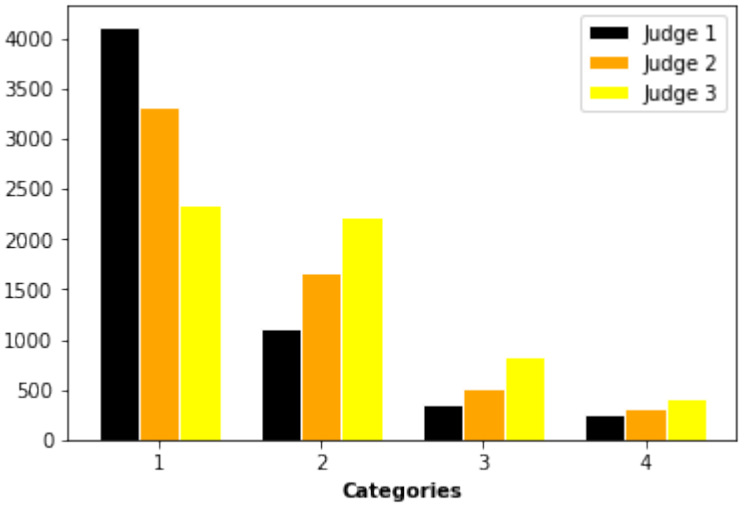


Figure 5.3: Distribution of the categories used by the judges in the Flickr-8k dataset.

QG-STEC dataset: The QG-STEC evaluation dataset is composed of questions generated from four systems that participated in the QG-STEC (Rus et al., 2010) Task B, that is, the task to generate a question from an input sentence.[14] Each question is evaluated based on five criteria: Relevance (on a scale from 1-4), Question Type (on a scale from 1 to 2), Syntactic Correctness and Fluency (on a scale from 1-4), Ambiguity (on a scale from 1-3) and Variety (on a scale from 1-3). The evaluation guidelines can be found via the link `http://computing.open.ac.uk/coda/resources/qg_form.html`.

Six judges took part in the evaluation. They judged batches of sentences independently. Table 5.3 shows the batch of questions judged and independent judges for that batch.

Table 5.4 shows the result of the analysis carried out for the QG-STEC dataset. Also in this case an interesting discrepancy between the IAA values and the correlation values is measured. Although IAA values are low almost

[14]The QG-STEC evaluation dataset is aviable at: `http://computing.open.ac.uk/coda/resources/qg_form.html`.

Judges	Batches of question
J1 and J2	80
J1 and J3	67
J1 and J4	81
J1 and J5	7
J1 and J6	106
J2 and J5	158
J3 and J5	125
J4 and J5	142
J5 and J6	129

Table 5.3: Batches of question with independent judges assigned to them. For $i = 1, \ldots, 6$, J_i means judge i.

everywhere, judges reach large (almost everywhere), and in two cases, very large Gamma correlation values.

Criteria	Coefficients	Coefficients values for pair of judges							
		J1 and J2	J1 and J3	J1 and J4	J1 and J6	J2 and J5	J3 and J5	J4 and J5	J5 and J6
Ambiguity	k	0.12	0.21	0.41	0.09	0.61	0.25	0.24	0.38
	Gamma	0.33 *	0.33 *	0.61	0.21 *	0.84	0.51	0.39	0.75
Fluency	k	0.42	0.51	0.58	0.07	0.33	0.51	0.46	0.15
	Gamma	0.67	0.63	0.71	0.39	0.57	0.62	0.57	0.39
QuestionType	k	0.32	-0.06	0.22	0.37	1	0.14	0.15	0.52
	Gamma	0.91	-1 *	1	0.89	1	0.45 *	0.80	0.95
Relevance	k	0.14	0.17	0.26	0.08	0.25	0.39	0.41	0.01
	Gamma	0.41	0.63	0.67	0.76	0.58	0.79	0.63	1 *
Variety	k	0.43	0.93	0.06	0.04	0.66	0.49	0.22	0.41
	Gamma	0.81	0.99	0.08 *	0.15 *	0.82	0.73	0.43	0.54
Average	k	0.28	0.35	0.30	0.13	0.57	0.35	0.29	0.29
	Gamma	0.62	0.31	0.61	0.48	0.76	0.62	0.56	0.72

Table 5.4: Conger's κ and Goodman and Kruskal's Gamma values reached in the QG-STEC dataset. For $i = 1, \ldots, 6$, J_i means judge i. The symbol $*$ indicate a p-value higher than 0.05.

As in the previous cases, the Gamma coefficient suggests that all the batches are annotated by judges that show a relative consistency and suggest data reliability.

Each pair of judges evaluated different batches of questions, which were generated from 4 different systems. Consequently, given the variance in

question quality, a deeper analysis is complicated. However, we can see that Judge 5 gets good correlation in any batch, which is also the case for Judge 2. This fact allows us to consider the batch they annotated together as the more reliable one. As we can see form the average values reported in Table 5.4, this is confirmed by the κ and Gamma values reached in the evaluation.

We can also notice that regarding the variety criterion, it is arguable that judge 4 and judge 6 miss a sound interpretation of it. Indeed, both of them get low correlation with judge 1. Judge 1, on the other hand, gets really high correlation with both Judge 3 and Judge 2. At the same time, Judge 5 gets high correlation with both Judge 2 and Judge 3 and lower correlation with Judge 4 and Judge 6. This evidence suggests that, in the case of the variety criterion, care must be taken with the data collected by Judges 4 and 6.

Concluding remarks The examples we presented in this section show that the use of correlation coefficients can help in obtaining a better assessment of the evaluation's reliability. For annotation tasks that involve a high level of subjectivity, a low level of agreement can be still acceptable if it is accompanied by a high level of correlation.

Agreement coefficients give us information about the extent to which judges share a common interpretation of the instructions used to annotate the phenomenon of interest. However, given the high level of subjectivity involved in the evaluation, judges can entertain different, although equally valid, interpretations of the instructions.

Correlation coefficients can then be used to take into account this valid variation in the interpretation. Correlation coefficients can measure the extent to which judges are consistent with each other in the annotation.

Taking into account our interpretation, a low level of agreement in conjunc-

tion with a high level of correlation suggests that the judges, although not sharing the exact interpretation of the phenomenon to be annotated, have a compatible interpretation of that phenomenon. In other words, given a ranking scale judges can give different scores. However, some judges may be more lenient or severe than others but when compared with each other they are consistently more lenient or severe. Because language variability can bring judges to have a (more or less) different interpretation of the phenomena to be annotated, the fact that they are consistent in their interpretation throughout the annotation suggests that the annotation is trustworthy.

In the case of a low level of correlation, our interpretation suggests that there is potentially something problematic with the annotation. By showing inconsistent annotation behaviours, judges display a non-trustworthy annotation performance which has a negative impact on the annotation trustworthiness.[15]

5.3 How to report correlation coefficient

In this section, we report some suggestions on how to report a correlation study when used to measure data reliability. It is important to stress that the information reported has to be considered as a suggestion. Other choices are possible. Nevertheless, it is important that researchers explain and justify their choices, giving all the possible information about the annotation. In this way, the readers can understand and evaluate the researchers' conclusions about data reliability themselves.

In order to measure the correlation for evaluations that involve more than two judges, we suggest reporting the average results of the pairwise measure

[15]In this case, when it is possible, we suggest trying to investigate the source of inconsistency by talking to the judges. This action can help in understanding the judges' interpretation of the phenomena to be annotated and their annotation behaviours (maybe it can also be used to improve the guidelines).

between the judges (Siegel and Castellan, 1988).

Regarding the correlation coefficient to be used, Table 5.5 suggests some possible choices. Table 5.5 is not intended to be exhaustive. It presents the most popular correlation coefficients and it gives at least a suggestion for each level of measurement and number of categories used in the evaluation. Table 5.5 has to be considered as a suggestion and other coefficients can be used. At the same time, the software and the relative library reported in Table 5.5 are to be considered as possible suggestions. Also in this case other software and libraries can be used.

Coefficients Name	Nominal	Ordinal	Interval/Ratio	> 2 categories	Software, Library
Yule's Q	✓	✗	✗	✗	Python, pycm / R, DescTools
Goodman Kruskal's Lambda	✓	✗	✗	✓	R, DescTools
Kendell's τ	✗	✓	✗	✓	Python, scipy.stats / R, DescTools (or stats)
Spearman's ρ	✗	✓	✗	✓	Python, scipy.stats / R, DescTools (or stats)
Goodman Kruskal's Gamma	✗	✓	✗	✓	R, DescTools
Pearson's *r*	✗	✗	✓	✓	Python, scipy.stats / R, stats

Table 5.5: Some correlation coefficients. Nominal, ordinal and interval/ratio represent level of measurement.

Between the possible scales of interpretation, it is arguable that the most popular is Cohen's (1988) one (Table 5.6).

Correlation value into the interval	Value interpretation
(- 0.3, 0] \ [0, 0.3)	Small
(-0.5, - 0.3] \ [0.3, 0.5)	Moderate
[-0.5, - 1] \ [0.5, 1]	Large

Table 5.6: Cohen's (1988) scale for the interpretation of correlation coefficient.

A variation of Cohen's (1988) scale is introduced by Rosenthal (1996) (Table 5.7). The main difference between the two scales lies in the high values. Indeed, Rosenthal (1996)'s scale allows for a finer-grained analysis of such values.

Both Cohen's (1988) and Rosenthal's (1996) scales are more adequate for

Correlation value into the interval	Value interpretation
(- 0.1, 0] \ [0, 0.1)	Negligible
(-0.3, - 0.1] \ [0.1, 0.3)	Small
(-0.5, - 0.3] \ [0.3, 0.5)	Medium
(-0.7, - 0.5] \ [0.5, 0.7)	Large
[-1, - 0.7] \ [0.7, 1]	Very large

Table 5.7: Rosenthal's (1996) scale for the interpretation of correlation coefficient.

non-parametric correlation coefficients (Rosenthal, 1996; Corder and Foreman, 2011). On the other hand, the Political Science Department at Quinnipiac University propose a scale (Table 5.8) for parametric correlation coefficients, more specifically for the Pearson's r correlation coefficient (Glen, 2020).

Correlation value into the interval	Value interpretation
0	None
(-0.2, 0) \ (0, 0.2)	Negligible
(-0.3, - 0.2] \ [0.2, 0.3)	Weak
(-0.4, - 0.3] \ [0.3, 0.4)	Moderate
(-0.7, - 0.4] \ [0.4, 0.7)	Strong
[-1, - 0.8] \ [0.7, 1]	Very strong

Table 5.8: Political Science Department at Quinnipiac University scale for the interpretation of correlation coefficient.

As in the case of the correlation coefficients of Table 5.5, the scales we presented should be considered a suggestion. As a matter of fact other scales can be found in the literature, for example Hinkle et al. (2003). Nevertheless, we stress that, regardless of the scale used, it is important to justify the choice of the scale.

Where possible, confidence intervals can be reported. Such intervals can be used for reproducibility aims. Indeed, they provide the range of plausible coefficient values which we should expect in case a new evaluation is carried out on the same items, but with the same number of new judges.

The matter of significance test is controversial. For example, without arguing in favour of stopping the use of statistical methods, Koplenig (2019) provides reasons to abandon statistical significance testing in corpus linguistics.[16] Koplenig (2019)'s argument moves from the impossibility of corpus linguistics to satisfy the independence assumption.[17] Accordingly, statistical approaches (for example, significance testing) that assume data independence cannot be applied.

Without taking the extreme position of Koplenig (2019), we follow the more moderate position presented in McShane et al. (2019):

> [...] the p-value be demoted from its threshold screening role and instead, treated continuously, be considered along with the currently subordinate factors as just one among many pieces of evidence. First, we recommend authors use the currently subordinate factors to motivate their data collection, statistical analysis, interpretation of results, writing, and related matters; we also recommend they analyze and report all of their data and relevant results. Second, we recommend editors and reviewers explicitly evaluate papers with regard to not only purely statistical measures but also the currently subordinate factors.[18] [Page 239]

In conclusion, when the correlation coefficients are used, we suggest researchers report and justify at least the following information:

- Number of judges.

[16] Lately, this point has been raised for *p-value* more in general (Wasserstein et al., 2019).

[17] The independence assumption requires that the data under examination is collected from a representative, randomly selected, portion of the total population.

[18] Examples of subordinate factors are: "related prior evidence, plausibility of mechanism, study design and data quality, real world costs and benefits, novelty of finding, and other factors that vary by research domain." (McShane et al., 2019, page 239)

- Number of items judged.
- The provenance of the sentences judged.
- The criteria used for the evaluation (for the purpose of reproducibility, make the guidelines used for the evaluation available).
- The coefficient value.
- The *p-value* (or other significance test).
- The scale used for the interpretation.
- The software used for the analysis.

Furthermore, as said above, to promote reproducibility, we suggest to report the confidence intervals, and make available the items used in the evaluation.

5.4 Conclusion

Using three case studies, we showed the limitations of using the IAA as the only criterion for checking the reliability of an evaluation. Given the variability of human language, we suggest that correlation coefficients and agreement coefficients, such as the Kappa statistic, can be used together to have a better picture of the evaluation data reliability in human evaluation of NLG. Agreement coefficients can be used both in pilot studies to improve annotation schemes and guidelines, and for data analysis to give a picture of how dissimilar the judges' interpretation of the phenomena is. Correlation coefficients can instead tell us to what extent judges are consistent with each other. As we have seen in Section 5.2.1, a low agreement coefficient value can hide a consistent pattern in the annotation, which is captured by high value for the correlation coefficient. Although judges have different opinions about the quality of a generated text, which is a result of the language variability, they entertain consistent relative interpretations. Consequently,

their judgments may still be considered reliable.

Chapter 6

Identifying judge bias

In Chapter 4, we saw the role of human biases in an intrinsic evaluation of NLG systems. We saw how subjective bias heavily determines the final annotation agreement. Our observational study shows that such biases are ineliminable and they have to be taken into account in each evaluation task. For this reason, statistical analyses that allow for better bias identification and understanding can be a valuable tool in evaluation efforts.

In this chapter we introduce a new interpretation and application of the Item Response Theory (IRT) (Gulliksen, 1950) to detect judges' bias. Our interpretation of IRT offers an original bias identification method that can be used to compare judges' bias and characterise annotation disagreement. Our method can be used to spot outlier annotators, improve annotation guidelines and provide a better picture of the annotation reliability. Additionally, because scales for IAA interpretation are not generally agreed upon, our bias identification method can assist with an understanding of the IAA value, which in turn can help with understanding the annotation disagreement.

6.1 Our proposal in a nutshell

As we have seen in Chapter 4, an important factor that can affect the annotation reliability is the presence of judge bias – that is, differences between judge preferences for subjective reasons. Our observational study shows that judges diverge in language annotation tasks due to a range of ineliminable factors such as background knowledge, preconceptions about language and general educational level. Although clear annotation schemes with effective guidelines and judge training aim to reduce annotation bias, some individual differences persist. A method that allows for a better identification and understanding of individual bias could be valuable in annotation efforts. For example, it could be used to:

- display differences in judges' behaviour. Annotations that show a markedly different pattern from the other annotations can either be removed or further analysed. This can help to spot outlying judges and help with improving annotation guidelines and reduce annotation disagreement.[1]

- provide a better picture of the annotation reliability. Indeed, once identified, the judge bias could be used to explain and understand annotation disagreement and accordingly the IAA values. For instance, as we will see in Section 6.3, such a method could be used to show in which respect a judge shows more strict annotation behaviour than another judge.

In some human annotation tasks, where a high level of subjectivity is involved — for example annotation that concerns the quality of a generated sentence (on a range of dimensions including syntax, semantic and pragmat-

[1] It is important to note that for annotation tasks which involve a high level of subjectivity, there is a limit to the reduction of annotation disagreement. Chapter 4 presents evidence for such limits.

ics) — what determines the judges' decisions cannot be measured directly. It is not a physical dimension, as for example weight or distance. Nevertheless, we can study the annotation behaviour, as directly observable, in order to have a better understanding of the unobservable decision process behind the judges' annotation. A first step in this direction is to analyse the frequency of the categories used by the judges. The frequency gives us a first approximation of the judges' behaviour. Nevertheless, the frequency of the categories used is a raw image. Ideally, we would like to be able to compare the judges' annotations in terms of a more fine-grained approximation of their perception of the phenomena being annotated. For this reason, in this chapter we introduce a way to analyse the unobservable decision process behind the judge annotation.

Our method is based on a new interpretation and use of the Item Response Theory (IRT). IRT is a psychometric theory used for analysing and designing tools — for example, surveys and tests — for measuring abilities or attitudes. We will use two examples to explain the novelty of our interpretation: mathematical ability (a traditional use of IRT), and natural language generation (our novel use).

IRT has traditionally been used, for example, to determine the validity of a test, say a test of mathematical ability. Such a test is administered to a number of students. On each test item, each student will achieve a certain score. Ideally, students with a strong mathematical ability should receive high scores on the test, and students with weak mathematical ability should receive low scores. There should also be a range in between these extremes. Based on data of a large group of students completing such a test, IRT can extract from the data a model that gives us, for a hypothetical student with a specific level of mathematical ability, the probability of specific test scores. A high level of mathematical ability should be associated with a high probability for a high test score. Importantly, mathematical ability is not

observed directly. It is only available via the students' performance on the test. And this coupling of the trait and the test performance is not going to be perfect (some good students may on occasion not do well, for this reason the IRT uses a probability function[2]).

In IRT, mathematical ability is modelled as a latent (not directly observable) trait of individuals. It is essential to note that a latent trait assumes a context, e.g. that of a conventional mathematical education. It exists against a background of assumptions about culture, regional variance, personal preferences, schooling/training, etc.

In our IRT-based method to detect annotator bias, the annotation exercise involves annotating corpus data (for example a set of sentences or images) with linguistic or other information. In this thesis we are focusing on evaluation of NLG systems. In this case the annotation exercise involves a set of sentences. Judges are asked to judge each of these sentences according to one or more criteria. Let's assume the criterion in question is the fluency of the sentence. *Our proposal is to treat fluency as a latent trait of sentences: it cannot be directly observed or measured, but we can get hold of it via the judgements of the annotators.*

In an annotation, the judges will not always agree: for instance, some may be more severe in their judgements than others. We can now apply IRT for individual judges. This time, IRT can extract from an individual's annotation the probability of a score's answer given a hypothetical sentence of a certain level of fluency. For each of the fluency scores which the judge can choose from when annotating, IRT gives us the probability that that judge will choose that particular of fluency score. This allows us to see how the judges' behaviour (the score they assign) relates to the latent trait (the level of fluency of a hypothetical sentence).

[2]More details are presented in Section 6.2.2.

Also in this case the latent trait assumes a context. It is a standard of fluency which is implicit in the judgements of a (more or less homogeneous) language community. The language community does distinguish between more and less fluent sentences. However, on specific occasions, individual members (or subgroups) will display smaller or larger differences in judgement. IRT allows us to quantify and identify these differences.

6.2 Analyzing raters bias: From frequency to Item Response Theory

By way of example, in this chapter, we use the QG-STEC evaluation dataset (Rus et al., 2012).[3] More specifically, we analyse the data collected from Judge 1 and Judge 3. This pair was randomly selected from all pairs of judges. We limit our investigation to the criteria ambiguity and variety (on a scale from 1 to 3) and syntactic correctness and fluency (for the sake of simplicity, in what follows, we refer to this criterion as fluency) and relevance (on a scale from 1 to 4).

6.2.1 Frequency

We are looking for a way to characterise the annotation behaviour of individual judges. This can help us to understand better the source of disagreement in an annotation. A way to do this is by analysing the frequency of the categories used by the judges. Indeed, the frequency gives us a first approximation of the judges' annotation behaviour. The comparison between different judges' decisions can show sources of difference that explain the disagreement in the annotation. Table 6.1 shows the frequency of the scores used by Judge 1 and Judge 3. It shows that Judge 3 tends to give slightly higher scores than Judge 1 for the ambiguity, fluency and relevance criteria.

[3]For more details about the dataset we refer to Section 5.2.3.

	Ambiguity			Variety			Fluency				Relevance			
	1	2	3	1	2	3	1	2	3	4	1	2	3	4
J1	0.68	0.19	0.11	0.20	0.04	0.74	0.43	0.26	0.22	0.07	0.91	0.01	0.07	0
J3	0.55	0.32	0.11	0.23	0.04	0.71	0.40	0.29	0.14	0.14	0.83	0.05	0.07	0.02

Table 6.1: Frequency of the scores used by Judge 1 (J1) and Judge 3 (J3).

Indeed, Judge 3 tends to be more cautious in giving the extreme score 1 than Judge 1, preferring the more neutral score 2. This trend is inverted for the extreme score 4. In this case Judge 1 tends to be more cautious and prefers lower scores than score 4. For example, we note that Judge 1 does not use the score 4 in the relevance criterion. In the case of variety, Judge 1 tends to score higher than Judge 3, preferring high scores whereas Judge 3 prefers low scores.

The frequency gives us a first approximation of the judges' annotation behaviour. Nevertheless, it is a raw image. Ideally, we would like to have a fine-grained picture of the judges' annotation behaviour. We propose the use of IRT for this aim. IRT provides a probabilistic analysis which allows inferences to be drawn about the judge's annotation behaviour.

6.2.2 Item Response Theory

IRT is used to measure various types of latent trait which are investigated by the use of item tests. For example it can be abilities, such as mathematical ability, or it can be a behavioral attitude, such as tendency to make particular purchases. The main aim of the theory is to evaluate and adjust test items and score examinees based on their latent traits such as abilities or attitudes.

In this section we are going to use IRT to describe judges' annotation behaviour in order to better understand the disagreement that arises from their annotation. We will see that IRT, as well as the frequencies, can be used to study the judges' preference of the scores. Furthermore, in contrast

to the frequency analysis, the IRT analysis uses a probabilistic model that allows us to have an insight where these differences take place in relation to the latent trait of a sentence.

To do so, we are going to give a new interpretation of the IRT. For this reason, it is important to explain the traditional use of IRT first. We will then present the ways in which our interpretation deviates from the traditional use.

The traditional use of IRT

Traditionally a test (or a survey) is designed in such a way that there are few items and many respondents. IRT is based on the assumption that each respondent answers each item in line with their level of the latent trait. In IRT it is assumed that the latent trait can be measured on a scale having a midpoint of zero and it can take any real number from $-\infty$ to $+\infty$. For practical considerations usually the range is limited to the interval [-4, 4]. It is important to keep in mind that this is just for simplicity and other intervals can be used.

Once a scale of measurement is given, it is possible to define a probability function of the possible answer as a function of the latent trait. Standard IRT uses a logistic model for this purpose. Logistic models have an S-shaped probability function such as the one depicted in Figure 6.1.

From a mathematical point of view the model, and more specifically Rasch's logistic model (Rasch, 1960), can be expressed by the following equation:

$$P(x_{im} = 1|z_m) = \frac{e^{(z_m - \beta_i)}}{1 + e^{(z_m - \beta_i)}}$$

where 1 is a label that represents the correct response, x_{im} represents the response of the mth respondent for the ith item and z_m represents the

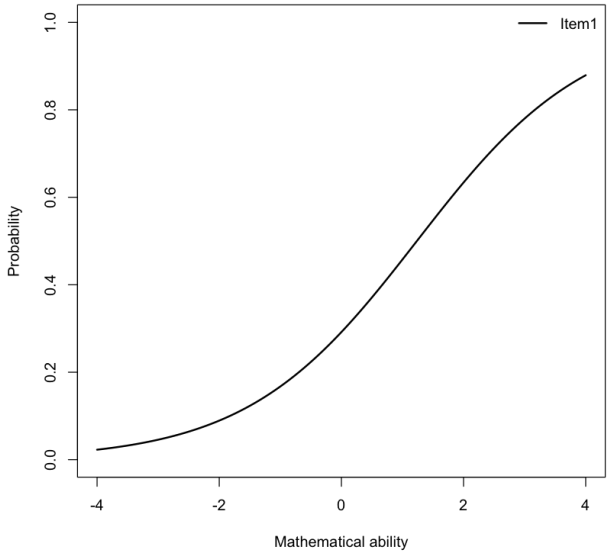


Figure 6.1: IRCCC example for a mathematics exam. On the x-axis is reported the latent trait. On the y-axis is reported the probability of a correct answer for item 1.

latent trait of the mth respondent. Finally, β_i is a parameter that takes into account the difficulty of the ith item. In other words, $P(x_{im} = 1|z_m)$ is the probability that respondent m correctly answers the item i given his/her latent trait z_m. Intuitively, the higher the latent trait z_m the higher the probability of correctly answering the ith item.

Each IRT model allows an Item Response Category Characteristic Curve (IRCCC) to be defined for each item (Figure 6.1 depicts an example of an IRCCC for Rasch's model). The IRCCC depicts the relationship between the latent trait and the scores (in the case of ability, for example mathematical ability) or each chosen item category (in the case of behaviour attitude, such as in a Likert scale survey). It shows the likelihood of a respondent receiving a score (in the case of ability) or selecting a certain category (in the case of behaviour attitude) at various levels of the latent trait. Concretely, the IRCCC gives a graphic representation of the probability function determined by an IRT model.

Let us give an example that involves a mathematics exam and Rasch's IRT model. In this case there are few items, let's say 4 mathematical questions and many students, let's say 35 students. All the students have to answer the 4 items. We assume that each student has a mathematical ability which contributes to the answers that they give. Based on the students' answers, both the items' difficulty and the students' mathematical ability can be measured. Once the test is done, all the answers are marked as correct or incorrect. Based on such binary test scoring Rasch's model can be used to define a probability function that describes the expected score the respondents should receive based on their mathematical ability. As said before, the probability function can be graphically represented by the IRCCC.

Figure 6.1 depicts the IRCCC of the first item (item 1). Accordingly, Figure 6.1 shows the probability that respondents will answer item 1 correctly, based on their latent trait, which represents their mathematical ability. Suppose for example that the mathematical ability of a respondent E_0 is 0. In this case Figure 6.1 suggests that E_0 has a low probability to answer correctly. To check this we imagine a straight line from point 0 of the latent trait to identify at which point it intersects with the line depicted in the graph. Figure 6.1 show that this happens for a probability value which is slightly higher than 0.2. In other word, the probability of E_0 to answer correctly is slightly higher than 0.2. Now, suppose that we have a respondent E_4 whose mathematical ability is 4. In this case the IRCCC suggest that E_4 has a high probability of answering item 1 correctly. Indeed, following the same procedure as we did for E_0, we can see that the probability of E_4 to answer item 1 correctly is slightly higher than 0.8.

In the mathematical test example, we were working with a binary or dichotomous setting. We assumed that the answers to the items could be either correct or incorrect. However, there are also IRT models for non-dichotomous ordinal category tests. One of these models is the Graded

Response Model (GRM) (Samejima, 1969). GRM is a variant of IRT developed to analyse tests that use polytomous categories, that is tests that use more than two categories. This is particularly appropriate for the analysis of rating and Likert scales. GRM can be applied to gather information about how change in the latent trait affects observed item responses.

IRT for judges' bias detection

The traditional application of IRT focuses on modelling the relationship between the observable respondents' answers and their unobservable latent trait, for example mathematical ability or behavioural attitude. In the case of a mathematical test, the respondents' performance on the test items should vary according to their mathematical ability: ideally, higher mathematical ability results in higher item test scores. The IRCCC is used to visualise the relationship between item test scores and the latent trait for a specific test item. It can tell us whether the test item indeed has the property that higher mathematical ability (of the person taking the test item) results in higher scores.

In order to apply IRT for identifying bias among annotators, we introduce the following twist in the application of IRT: rather than model a latent trait of the respondents, the latent trait in question now is conceived of as a property of linguistic items, that is sentences. One such property is sentence fluency: one sentence can be more fluent than another, but this is not a property that can be measured directly. Rather, it is a latent trait of sentences that we can uncover only via the judgements of language users. The IRT helps with modelling how different judges respond to different levels of a latent trait, such as fluency. Individual biases, which may be due to a wide variety of factors (e.g. regional variance, personal preferences, schooling/training) that skew application of the shared norms or standards (that the language community as a whole has adopted) can thus be laid bare.

The IRCCC is now used to visualise for individual judges the relationship between fluency scores and the latent trait (fluency) of sentences. Judges can be compared with each other by putting their IRCCCs next to each other.

The latent trait introduces a new ingredient with respect to other probabilistic analyses of judges' bias, for example the model introduced by Dawid and Skene (1979) for the case of medical diagnosis.[4]

The model defined by Dawid and Skene, is based on an estimation of the judges accuracy at identifying the true scores (where the set of the scores is finite). Given a judge and a potential true score[5], the model gives us a probabilistic estimation that the judge will choose that score. Because such estimation is done for each judge, the model allows for a comparison among the judges.

In contrast, our IRT-based model for bias detection defines a probabilistic estimation of judge choice of scores given a hypothetical sentence of a certain latent trait (which by definition is a continuum). Also in this case, as the estimation is done for each judge, the model allows for a comparison among the judges. For example, let's consider once again a human evaluation about fluency. In this case the latent trait is the sentence fluency. As the fluency of the sentence varies, the model gives a probabilistic estimation of how the judge will annotate that sentence.

For each judge and for each criterion, the model of Dawid and Skene (1979) provides a confusion matrix which describes the judge responses by the estimated true scores. In contrast, for each judge and for each criterion, the model we propose provides a probabilistic estimation of how the judge

[4]The same methods was applied for the case of word sense annotation by Passonneau and Carpenter (2014).

[5]The model of Dawid and Skene considers both the cases where the true scores are given, and the case in which the true scores are missing. In the last case, the true scores are inferred from the annotation as an estimate of the prevalence of each score.

will annotate a sentence as the latent trait changes. In our model, such information is provided with both the IRCCC and the extremity parameters.

6.3 The use of GRM to analyse judges' bias: An example from NLG evaluation

In this section, we present our IRT-based method. We use GRM to analyse the judges' decision bias. For each criterion studied, we present two analyses.

The first analysis is based on the IRCCC graphs. The IRCCC graphs provide an informal analysis of the judges decision behaviour through the evaluation. They have the advantage of being easy to interpret and they allow for a quick insight about the judges' disagreement. Nevertheless, in order to have a more refined analysis we will introduce the concept of an extremity parameter.

The extremity parameters show the latent trait score at which judges have a 50% chance of selecting certain categories and 50% chance of selecting the remaining categories. For this reason they can be used to suggest the spectrum of the latent trait where we can find the main source of disagreement.

The extremity parameters: an explanatory example

Suppose that two Judges (let's say $J1$ and $J2$) are performing an evaluation about the fluency of a set of sentences, and assume that the evaluation is based on the four scores: 1, 2, 3, and 4. Suppose we're interested in the extremity parameter that suggests the 50% chance of selecting scores 1 and 2 (for sake of simplicity let us denote this as $Ex2$). $Ex2$ expresses the latent trait level at which $J1$ and $J2$ have a probability of 0.5 of selecting either score 1 or score 2 and a probability of 0.5 of selecting either score 3 or score 4. Let's suppose the $Ex2$ level for $J1$ is 0.1, whereas the $Ex2$ level for $J2$ is 0.6 (see the green arrows of Figure 6.2). Under our interpretation of IRT, this means that, given a sentence of latent trait (fluency in our example)

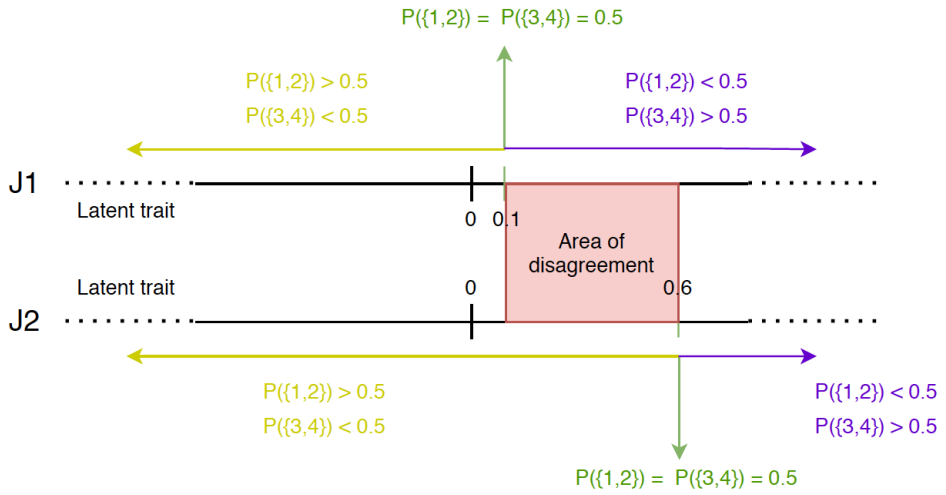


Figure 6.2: A graphical description of the explanatory example. $P(\{x, y\}) < 0.5$ means that the probability of selecting score x or score y is smaller than 0.5, for $x, y \in \{1, 2, 3, 4\}$. The same interpretation applies also for the symbol $>$ (higher than) and for the symbol $=$ (equal to).

of 0.1 the probability that $J1$ selects categories 1 or 2 is 0.5. Likewise, the probability that $J1$ select categories 3 or 4 is 0.5. On the other hand, $J2$ has a probability of 0.5 of selecting categories 1 or 2 for sentences of latent trait of 0.6. Likewise, for sentences of latent trait of 0.6 the probability that $J2$ select categories 3 or 4 is 0.5.

As shows by Figure 6.2, the interval $[0.1, 0.6]$ (the red rectangle) defines a probabilistic analysis of the disagreement between $J1$ and $J2$. In such an interval, the IRT analysis suggests the disagreement mainly comes about because $J2$ tends to give lower scores than $J1$. Indeed, approaching the latent trait score 0.6, $J1$ has a higher probability of selecting scores 3 and 4 than scores 1 and 2, whereas $J2$ has a slightly higher probability of selecting scores 1 and 2 than scores 3 and 4. Figure 6.2 shows that for Judge 1 the interval $[0.1, 0.6]$ is dominated by the purple arrow ($P(\{1, 2\}) < 0.5$ and $P(\{3, 4\}) > 0.5$). In contrast, for Judge 2 the interval $[0.1, 0.6]$ is dominated by the yellow arrow ($P(\{1, 2\}) > 0.5$ and $P(\{3, 4\}) < 0.5$).

The intervals detected by the extremity parameters analysis tell us where the major areas of disagreement lie. Where the IRCCC gives us a friendly, although crude, picture of these areas of disagreement, the extremity parameters give us the exact points in the latent trait, where a different annotation behaviour among the judges emerges. The divergence between judges made explicit by the extremity parameters give us some information about the discrepancy between the judges interpretation and understanding of the phenomena to be annotated.

We hypothesize that the main reason for these divergences are due to judges' bias. The link between judges' bias (determined by the intervals detected by the extremity parameters) and the latent trait (where the intervals lie) allows us to better understand judges' bias.

In the following analysis, when using the extremity parameters, we use the definitions from the annotation guidelines and concrete examples from the dataset to show how to interpret the main differences among judges.

GRM-based method for our analysis

As we have seen in the traditional use of IRT, the test is designed in such a way that there are relatively few items and many respondents. In this case, for each item the IRCCC and the extremity parameters are defined based on the respondents' answers to that item. In the case of intrinsic human evaluation of NLG systems we have few judges and several items. Usually each item is a different instance of few criteria that are analysed. For example, given the fluency criterion, there are several items consisting of the same question about fluency but with different generated sentences to be evaluated.

In the present chapter we used the GRM model in the following way. Given a judge J and a criterion C, we collect all the items annotated by J aimed

at evaluating C. Let's call I_C^n the nth item that aims to evaluate C. In this case, given a judge J, for each criterion C the IRCCC and the extremity parameters are defined based on J's answers to I_C^n for $n = 1 \dots m$, where m is the number of sentences to be evaluated. In what follows, given a criterion C we collect all the items I_C^n under the name of C.

Our GRM-based method provides a probabilistic analysis of the judges' annotation behaviour for each criterion. Because our method analyses one judge at a time, it allows us to compare different judges' annotation behaviour. Such comparison, allows us to determine the judge's bias and to better understand the disagreement between judges.

Interpreting the criteria

GRM considers the scores in increasing order. Conversely, in the QG-STEC evaluation dataset the scores are considered in decreasing order. For this reason, in what follows, we have to pay attention to the interpretation of the IRCCC. More specifically, the relevance, the fluency and the variety criteria consider 1 as the best score — best from a criterion quality point of view. For this reason we have to interpret the positive latent trait as the sentence irrelevance, non-fluency and non-variety. The situation for the ambiguity criterion is different. It is stated in terms of non-ambiguity (that is, lack of ambiguity). In this case we have to interpret the positive latent trait for the ambiguity criterion as the sentence ambiguity.

Software used for the analysis

In what follows, we carry out our IRT analysis with the statistical software *R*. We used the library *ltm* (Rizopoulos, 2018). This library was developed to provide researchers with a flexible framework to perform IRT analyses. More specifically, we used the function *grm()*.

Regarding the use of the coefficient of agreement, as done in Chapter 5, we use the Conger's κ (Conger, 1980) with ordinal weight, as defined by Gwet (2014). To measure Conger's κ we used the library *irrCAC* provided by the *R* software. More specifically, we used the function *conger.kappa.raw()* with the variable weights set to "ordinal".[6]

6.3.1 The ambiguity criterion

The ambiguity criterion reaches a Conger's κ of 0.21, indicating a low level of agreement. Such a value suggests that the judges rarely make the same score decision.

Table 6.2 reports the scores and their description for the ambiguity criterion.[7]

Score	Description
1	The question is unambiguous.
2	The question could provide more information.
3	The question is clearly ambiguous when asked out of the blue.

Table 6.2: Scores and scores description for the ambiguity criterion.

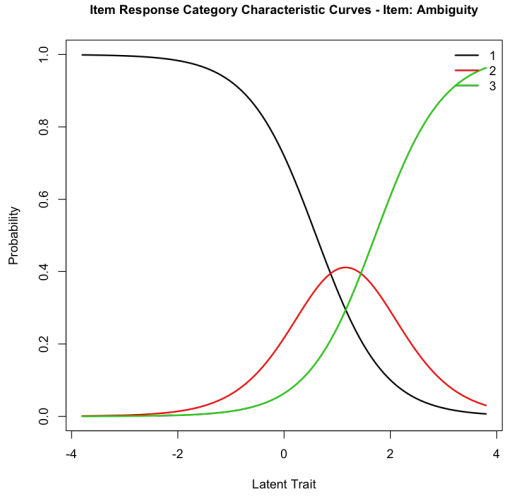
The IRCCC graphs analysis

The scores frequency analyses in Section 6.2.1 suggest that the disagreement emerges mainly due to the fact that Judge 1 prefers score 1 whereas Judge 3 prefers score 2. We can use GRM to investigate such a divergence in more depth.

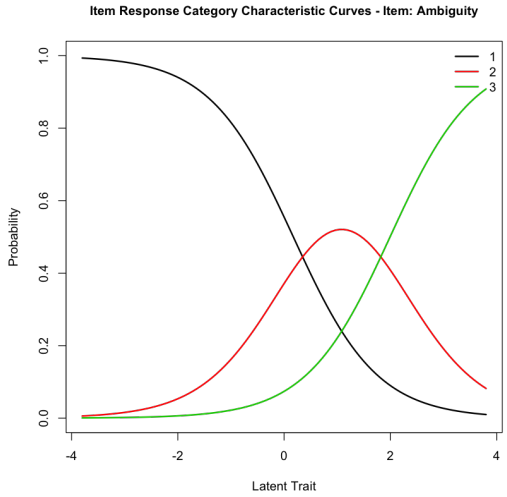
The IRCCC for the ambiguity criterion shows that Judge 1 (see Figure 6.3(a)) was less cautious in choosing score 1 than than Judge 3 (see Figure 6.3(b)). Indeed, we can see that:

[6]For more detail we refer to Chapter 3.

[7]The complete evaluation guidelines are available at `http://computing.open.ac.uk/coda/resources/qg_form.html`.



(a)



(b)

Figure 6.3: IRCCC for the ambiguity criterion for Judge 1 (a) and Judge 3 (b). The graphs show that the main source of disagreement can be found in the latent trait interval [0, 2] mainly relative to score 2 (red line).

- For positive levels of the latent trait (approximately the interval [0, 2]) the peak of the curve of score 2 (red line) is higher for Judge 3 than for Judge 1.

- At the same time, we can see how in the case of Judge 1 the curve

of score 2 (red line) intersects that of score 1 (black line) for higher latent trait levels than is the case for Judge 3.

These facts together suggest the following. On one hand, for the latent trait levels that are approximately in the interval [0, 2], Judge 3 has a higher probability of selecting score 2 than Judge 1. On the other hand, for the latent trait levels that are approximately in the interval [0, 1], Judge 1 has a higher probability of selecting score 1 than Judge 3.

From the IRCCC graphs, we can conclude that the main divergence between Judge 1 and Judge 3 takes place approximately in the interval [0, 2] of the latent trait. We can now use the extremity parameter to refine such analysis.

The extremity parameter analysis

The extremity parameter for the ambiguity criterion suggests that:

- Judge 1 has a 50% chance of selecting the score 1 with a latent trait level of 0.601 and a 50% chance of selecting the scores 2 or 1 with a latent trait level of 1.717.
- On the other hand, Judge 3 has a 50% chance of selecting the score 1 with a latent trait level of 0.173 and a 50% chance of selecting the scores 2 or 1 with a latent trait level of 1.993.

From these levels, it follows that:

- For the latent trait levels between the interval [0.173, 0.601], Judge 1 has a higher probability of selecting score 1 than Judge 3. Whereas for the same interval, Judge 3 has a higher probability of selecting scores 2 and 3 than score 1.
- At the same time we can see that between the latent trait of [1.717, 1.993] Judge 1 has a higher probability of selecting score 3 than score 1 and 2 whereas Judge 3 has a higher probability of selecting score 1

and 2 than score 3 (this is due to the fact that for 3 sentences Judge 3 gives score 2 whereas Judge 1 selects score 3).

The IRT analysis suggests that the main source of disagreement can be found in the latent trait intervals [0.173, 0.601] (let's denote it as I^1_{am}) and [1.717, 1.993] (let's denote it as I^2_{am}).

Based on the evaluation guideline provided for the QG-STEC task B for the ambiguity criterion, and under the assumption that a question can be more ambiguous than another one, we can draw the following conclusions.

The questions that fall in I^1_{am} can be interpreted as questions that, if stated out of the blue, are slightly ambiguous in that they are missing some information. A couple of examples from the dataset are:

> "How many gunners who died should there be a fitting public memorial to?" or "Where did the trust employ over 7,000 staff and manage another six sites?".

In I^1_{am}, we can expect Judge 3 to exhibit a more strict annotation behaviour than Judge 1. The questions that fall in I^2_{am} can be interpreted as questions that, if stated out of the blue, have high probability of being perceived as ambiguous. A couple of example from the dataset are:

> "What is the axiom in Euclidian Geometry?" or "Why tend accidents to be relatively minor ?".

In I^2_{am}, we can expect Judge 1 to present a more strict annotation behaviour than Judge 3.

6.3.2 The variety criterion

The variety criterion aims to measure, given two questions, the extent of their difference. This measures the ability of a system to generate a variety of different questions given the same input. The variety criterion reaches a

high Conger's κ agreement coefficient of 0.93. This value shows a tendency of Judges 1 and 3 to reach the same score decision.

6.3 reports the scores and their description for the variety criterion.[8]

Score	Description
1	The two questions are different in content.
2	Both ask the same question, but there are grammatical and/or lexical differences.
3	The two questions are identical.

Table 6.3: Scores and scores description for the variety criterion.

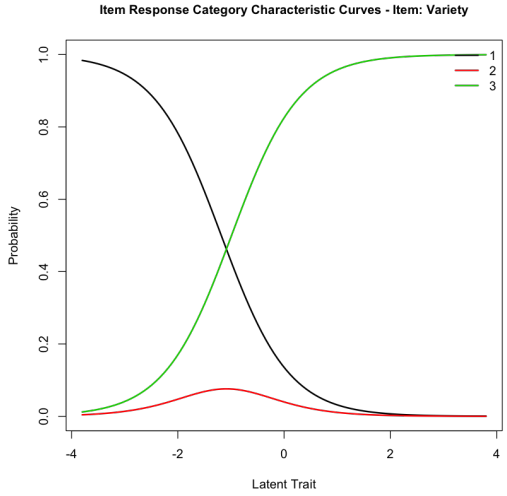
The IRCCC graphs analysis

The scores frequency analysis in Section 6.2.1 shows that there is little difference between the judges' annotation. A deeper analysis shows that the disagreement emerges in three pairs of questions where Judge 1 chose the scores 2, 3, 3 whereas Judge 3 chose the scores 1, 1, 2. This fact is captured by the IRCCC depicted in Figure 6.4. More precisely, Figure 6.4 shows that:

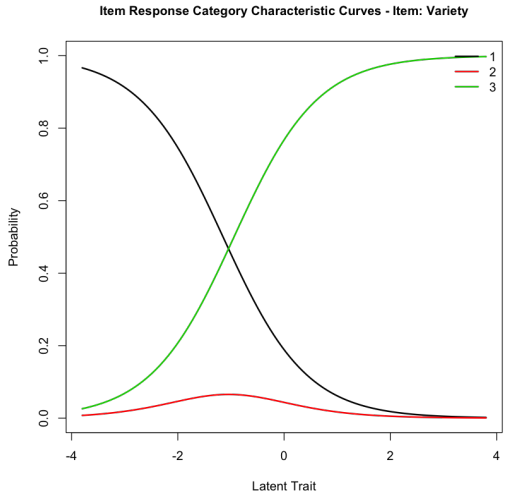
- The curves in both the figures are highly similar.
- The main difference is about the score 1 (black line). We notice that the line moves slightly towards the positive trait (that is, it meets the green line slightly towards the positive trait) for Judge 3, who indeed chose such scores more than Judge 1.

We note also that the low peak for the score 2 (red line) shows that both the judges tend to avoid the score 2 in favour of the extreme scores 1 and 3. Indeed, the frequency for the variety criterion depicted in Table 6.1 shows that both the judges choose the score 2 4.5% of the time.

[8]The complete evaluation guidelines are available at `http://computing.open.ac.uk/coda/resources/qg_form.html`.



(a)



(b)

Figure 6.4: IRCCC for the variety criterion for Judge 1 (a) and Judge 3 (b). The graphs are very similar which shows a low level of disagreement between the judges.

The extremity parameter analysis

The extremity parameters for the variety criterion show that:

- Judge 1 has a 50% chance of selecting the score 1 with a latent trait

level of -1.183 whereas Judge 3 with a latent trait level of -1.150.

- Similarly, Judge 1 has a 50% chance of selecting the scores 1 and 2 with a latent trait level of -0.990 whereas Judge 3 with a level of -0.944.

The small differences between the extremity parameters explain the small disagreement between Judge 1 and Judge 3. In this case, the GRM shows that Judge 3 was slightly more cautious (indeed this happens in just 3 cases) in giving high scores than Judge 1 for a latent trait level at approximately the levels -1 and 0.

6.3.3 The fluency criterion

For the fluency criterion the judges reach a Conger's κ value of 0.51.

Table 6.4 reports the scores and their description for the fluency criterion.[9]

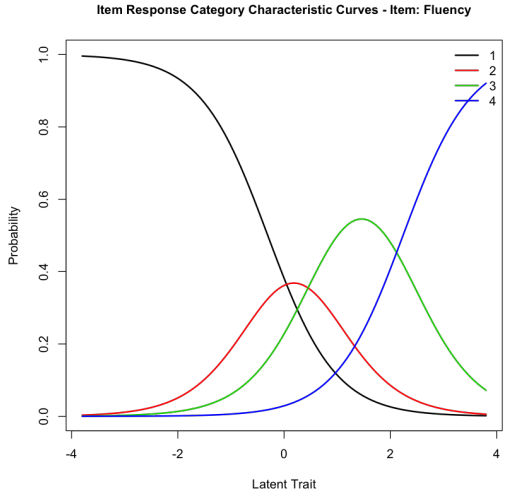
Score	Description
1	The question is grammatically correct and idiomatic/natural.
2	The question is grammatically correct but does not read as fluently as we would like.
3	There are some grammatical errors in the question.
4	The question is grammatically unacceptable.

Table 6.4: Scores and scores description for the fluency criterion.

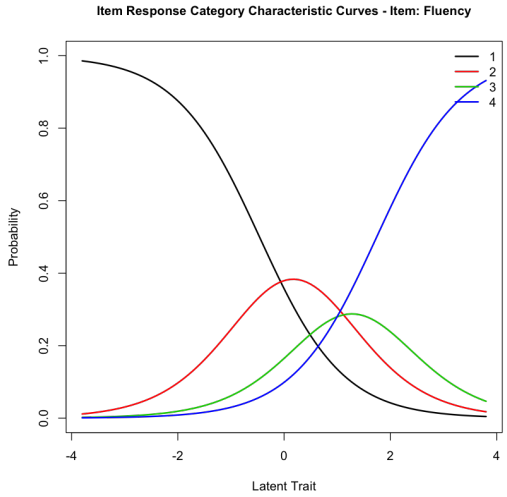
The IRCCC graphs analysis

In the case of the fluency criterion, the frequency analyses in Section 6.2.1 suggest that Judge 1 prefers to use scores 1 and 3, whereas Judge 3 prefers to use scores 2 and 4. This is illustrated in Figure 6.5, which clearly shows the phenomenon in the latent trait interval that is approximately between -2 and 2. More specifically, for Judge 1 we can see that:

[9]The complete evaluation guidelines are available at `http://computing.open.ac.uk/coda/resources/qg_form.html`.



(a)



(b)

Figure 6.5: IRCCC for the fluency criterion for Judge 1 (a) and Judge 3 (b). The graphs show that the main sources of disagreement can be found in the latent trait interval [(±)1, (±)2.2] mainly relative to score 3 (green line) and score 4 (blue line).

- The probability of selecting the score 3 (green line) is higher than selecting the score 4 (blue line).

- At the same time, the probability of selecting the score 2 (red line)

is slightly higher than selecting the score 3 (green line) in the latent trait interval that is approximately between -2 and 0.

- On the other hand, from level 0 to level 2 of the latent trait, the probability of selecting the score 3 (green line) is higher than selecting the score 2 (red line).

Conversely, for Judge 3:

- The probability of selecting the score 3 (green line) is slightly higher than selecting the score 4 (blue line) up to the latent trait level of approximately 1.
- For values of the latent trait level higher than 1, the probability of selecting score 4 increases dramatically.
- At the same time, the probability of selecting the score 2 (red line) is higher than selecting the scores 3 (green line) and 4 (blue line) in the latent trait interval that is approximately between -3 and 1.

The extremity parameter analysis

The extremity parameters for the fluency criterion show that:

- Judge 1 has a 50% chance of selecting score 1 with a latent trait level of -0.309, whereas Judge 3 has a 50% chance of selecting score 1 with a latent trait level of -0.464.
- At the same time, Judge 1 has a 50% chance of selecting the scores 1 and 2 with a latent trait level of 0.677 whereas Judge 3 has a 50% chance of selecting the scores 1 and 2 with a latent trait level 0.809. In this part of the latent trait, the judges show a similar annotation trend.
- A very interesting divergence can be found for the high scores. Here,

> Judge 1 has a 50% chance of selecting the scores 1, 2 and 3 with a latent trait level of 2.238 whereas Judge 3 has a 50% chance of selecting the scores 1, 2 and 3 with a latent trait level of 1.742.

The analysis of the extremity parameters suggests that the main divergence between Judge 1 and Judge 3 takes place in the latent trait that is in the interval [1.742; 2.238] (let's denote it as I_f^1). Based on the evaluation guidelines provided for the QG-STEC task B for the fluency criterion, and under the assumption that one question can be more fluent than another one, we can draw the following conclusions:

The questions that fall in I_f^1 can be interpreted as questions that present clear grammatical errors, and so for which the fluency is problematic. A couple of examples from the dataset are:

> "Was the British information on Dean Thomas was left in the US version?" and "To what is dating of prehistoric materials particularly crucial? ".

In I_f^1, Judge 3 shows more strict annotation behaviour than Judge 1, tending to give the score 4 whereas Judge 1 scores 3 or 2.

The same strict behaviour of Judge 3 can be detected in the latent trait interval [-0.464, -0.309] (let's denote this as I_f^2). The range I_f^2 contains questions which are slightly lacking fluency, or questions with minor grammatical errors. A couple of examples from the dataset are:

> "The son purchased which company?" and "What apply to a range of social care professionals?".

In I_f^2, Judge 3 tends to give scores higher than 1, whereas Judge 1 tends to give the score exactly 1.

6.3.4 The relevance criterion

The relevance criterion reaches a low Conger's κ value of 0.17. We can see from the actual scores frequency (provided in Section 2.1) that the frequency of score 1 is much higher than that of the other scores. As explained in Artstein and Poesio (2008) and Gwet (2014), this gives rise to the *prevalence paradox*: a high degree of observed agreement is associated with a low agreement coefficient. Artstein and Poesio (2008) suggest that to address the prevalence paradox (if necessary) it may be best to also report the observed agreement.[10] In this case the observed agreement is 80%.

Table 6.5 reports the scores and their description for the relevance criterion.[11]

Score	Description
1	The question is completely relevant to the input sentence.
2	The question relates mostly to the input sentence.
3	The question is only slightly related to the input sentence.
4	The question is totally unrelated to the input sentence.

Table 6.5: Scores and scores description for the relevance criterion.

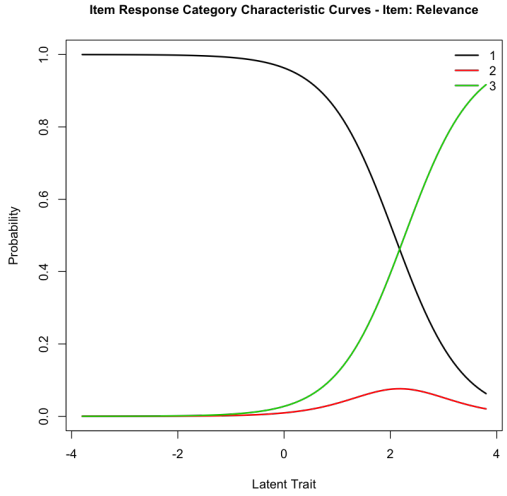
The IRCCC graphs analysis

The scores frequency analyses in Section 6.2.1 show for relevance that the Judges 1 and 3 tend to make most use of score 1.

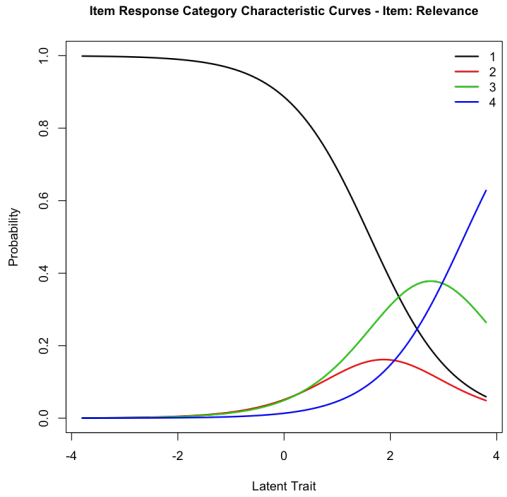
A first cursory look at the IRCCCs for relevance in Figure 6.6 suggest a substantial difference between the graph for Judge 1 (a) and Judge 3 (b). However, closer inspection suggests that the differences may not be that substantial: The curve for score 1 (black line) and score 2 (red line) are

[10] It is important to remember that, although the observed agreement provides information about the Judges' agreement, it shouldn't be used as a measure of reliability, see for example Krippendorff (1980) and Artstein and Poesio (2008).

[11] The complete evaluation guidelines are available at `http://computing.open.ac.uk/coda/resources/qg_form.html`.



(a)



(b)

Figure 6.6: IRCCC for the relevance criterion for Judge 1 (a) and Judge 3 (b). The graphs show as the small source of disagreement resides in the extreme part of the latent trait.

quite similar — the differences are:

- Judge 1 (Figure 6.6 (a)) does not use the category 4.
- Graph 6.6(b) shows a slightly higher peak for score 2 (showing that

Judge 3 uses score 2 slightly more than Judge 1).

- The curve for score 1, goes down more slowly for Judge 1 than for Judge 3 (showing that Judge 1 uses score 1 slightly more than Judge 3).

- The curve for the score 3 (green line) is quite similar for both graphs till the latent trait level reaches about 2. At this point the curve increases for Judge 1 — see Graph 6.6(a) — and decreases for Judge 3 — see Graph 6.6(b). For Judge 3, at that level of the latent trait, the score 3 (green line) gives way to the score 4 (blue line). In contrast, Judge 1 never selects score 4 for higher levels of the latent trait, preferring score 3 instead.

The small differences between the graphs, particularly for scores 1 and 2, explain the high level of observed agreement. The marked difference between use of scores 3 and 4 explain the low κ value.

The extremity parameter analysis

The extremity parameters for the relevance criterion show as the divergence in annotation between Judges 1 and 3 that can be found at high levels of the latent trait:

- Judge 1 has a 50% chance of selecting the score 1 with a latent trait level of 2.076, whereas Judge 3 has a 50% chance of selecting the score 1 with a latent trait level of 1.618.

- Judge 3 has a 50% chance of selecting the scores 1, 2 or 3 with a latent trait level of 3.387, whereas for Judge 1 such a level is not defined because they never use score 4.

- The Judges behave similarly regarding the 50% chance of selecting the scores 1 and 2. In this case, for Judge 1 the latent trait level is 2.271

and for Judge 3 the latent trait level is 2.132.

The extremity parameters suggest that the main difference between Judges 1 and 3 resides in the latent trait interval [1.618, 2.076] (let's denote it as I_r). Based on the evaluation guidelines provided for the QG-STEC task B for the relevance criterion, and under the assumption that, given an input text T, one question can be more relevant to T than another one, we can draw the following conclusions:

The questions that fall in I_r can be interpreted as questions that are somewhat related to the input sentence. A couple of examples from the dataset are:

> *Input sentence A*: Women tend to cover shorter urban journeys and therefore their driving is slower and accidents tend to be relatively minor.
>
> *question A*: What cover shorter urban journeys ?
>
> *Input sentence B*: In Philosopher's/Sorcerers Stone, information on Dean Thomas was left in the US version, but not the British,
>
> *question B*: who was left in the US version, but not the British?

In I_r, Judge 1 shows a more lenient annotation behaviour than Judge 3. Indeed, we can expect Judge 3 to give higher scores then Judge 1 for these and similar items.

Because Judge 1 does not use the score 4, for latent trait levels higher than 3.387, the model predicts a persistent difference in annotation behaviour between the two Judges: Where Judge 1 uses score 3, Judge 3 uses score 4.[12]

[12]We have to keep in mind that IRT provides a probabilistic analysis based on a set of effective annotations. Indeed, the analysis we presented allows for an analysis of Judges' bias based on the actual annotations.

6.4 How to report an IRT study

When an IRT study is performed we suggest to report both the IRCCC graphs analysis and the extremity parameters analysis. Both the analyses should be accompanied by an explanation of the results which justifies their use.

As in the case of the correlation coefficient, general information about the evaluation should be reported.

In conclusion, when an IRT study is performed, we suggest reporting and justifying at least the following information:

- Number of judges.
- Number of items judged.
- The criteria used for the evaluation (for the benefit of reproducibility, make available the guideline used for the evaluation).
- The IRCCC graphs.
- The extremity parameters. Based on the extremity parameters analysis, use the definitions from the annotation guidelines and concrete examples from the dataset to show how to interpret the main differences between Judges.
- The software used for the analysis.

Furthermore, for the purpose of reproducibility we suggest that the items used in the evaluation should be made available.

From a practical point of view, where space is limited because of page restrictions or similar, we suggest that the IRT analysis should be presented in an appendix, or in supplementary material which can be published, for example, in an online repository. In order to compress the presentation of

the extremity parameters, they can be reported in a table.

6.5 Conclusion

In this chapter we have introduced a way to visualise and identify judges' bias by the use of IRT, more specifically the GRM, going beyond the analysis of category frequency. Our interpretation allows us to use IRCCC and the extremity parameters to gain an insight into the judges' annotation bias, based on a common latent trait scale of measurement. The use of GRM sheds light on the annotation disagreement. It can also be used to spot annotator outliers, improve annotation guidelines or to have a better picture of the annotation reliability. In this chapter we have showed how it can be used to accompany IAA values in order to explain the annotation disagreement. For instance, IRT can show in which respect one judge shows a stricter or a more lenient annotation behaviour than another judge.

Conclusion and future directions

The NLG community relies on shared evaluation techniques to understand progress in the field. By analyzing papers published over 10 years (from 2008 to 2018) in NLG-specific conferences we found that:

- Human intrinsic evaluation is the most-used evaluation method in NLG.
- Human intrinsic evaluation suffers from shortcomings in existing approaches to measuring and reporting the reliability.

Driven by these findings, in this thesis we worked to answer the following question:

How should we carry out a reliability study for intrinsic human evaluation of NLG systems?

To answer this question we proposed a new set of methods for identifying judges' bias and reporting reliability, specifically for human intrinsic evaluation of NLG systems.

The methods we suggested were informed by the results of an observational study. Our studies show that judges' bias is an important aspect of human intrinsic evaluation of NLG systems, to such an extent that it has to be

taken into account in the reliability study. We propose utilising the concept of relative consistency, and its formalisation through the use of correlation coefficient, for checking the reliability of intrinsic human evaluation. At the same time the need for better bias identification and understanding led us to introduce a new interpretation and application of IRT to identify judges' bias.

Looking back: collecting the thesis contributions

After analysing the results presented in this thesis, we can summarise our main contributions as follows:

- We offered an overview of the evaluation landscape of the NLG evaluation methodologies.
- We identified shortcomings in existing approaches to reporting the reliability of intrinsic human evaluation studies in NLG.
- We showed that there are factors inherent to language evaluation which set limits of how high we can expect human agreement to be.
- We defined a new taxonomy of subjective annotator bias in language evaluation in the evaluation of NLG systems.
- We suggested a new additional way of reporting reliability based on the use of correlation coefficients.
- We introduced a new method for better identifying and understanding judges' bias.

We believe the overall thesis' contribution is threefold. Firstly it enhances, in the NLG community's, awareness of the need to handle the problem of human evaluation reliability. Secondly it suggests a way to carry it out. Finally, the thesis suggests an original way to identify the judges' bias.

Returning to the original question behind this thesis: *How can we carry out a reliability study for intrinsic human evaluation of NLG systems?*, let's use an example of a report of a reliability study to illustrate the results and recommendations we have put forward in this thesis.

Reporting a reliability study

In Section 5.3 and Appendix 3.6 we suggest how to report correlation coefficients and agreement coefficients.

Once again we use the QG-STEC evaluation dataset. More specifically the dataset collected by Judge 1 and Judge 3 in order to evaluate the Syntactic Correctness and Fluency.

Let's suppose we build an AQG system and we aim to evaluate its ability in generating fluent questions. To do so, we hire two people with the task of judging 67 outputs of our system based on a guideline we provide to them. Let's suppose that the evaluation we performed is the one done by Judge 1 and Judge 3 in order to evaluate the Syntactic Correctness and Fluency.

Once the evaluation is performed we suggest reporting the reliability study as follows:

> 67 questions were independently judged by two judges. The judges were native English speakers and they were not trained for the evaluation task. Furthermore, they were different from the system developers. The evaluation was performed based on one criterion that aimed to verify the fluency of the question. The fluency criterion was rated on a 4-point ranking scale from 1 (the question is fluent) to 4 (the question is not fluent). The dataset and the evaluation guidelines can be found at: `http://computing.open.ac.uk/coda/resources/qg_form.html`.

Given the ordinal nature of the data we used the Conger's κ (Conger, 1980) with ordinal weight, as defined by Gwet (2014).[13] We interpreted Conger' κ based on Landis and Koch's (1977) scale of interpretation. Regarding the correlation coefficient we used the Goodman and Kruskal's *Gamma* (Goodman and Kruskal, 1954), which is an adequate coefficient for ordinal data with many ties.[14] We interpret *Gamma* based on the scale for correlation coefficients introduced by Rosenthal (1996). We chose this scale because it allows a more finer-grained analysis than the more popular Cohen scale (Cohen, 1988), specifically for the interval value $[0.50, 1]$.

Coefficient Name	Coefficient Value	Confidence Interval	Scale Name	Scale interpretation
Conger's k	0.51	(0.32, 0.71)	Landis-Koch (1977)	Fair
Goodman Kruskal's Gamma	0.63 (p < 0.05)	(0.41, 0.85)	Rosenthal (1996)	Large

Table 6.6: Reliability study result from our example.

Table 6.6 provides the results from the reliability study. The *fair* level of κ suggests a partial agreement between Judge 1 and Judge 3. This means that Judge 1 and Judge 3 hold different opinions about the fluency of some of the questions they judged. Nevertheless, the large value of Gamma suggests that such divergent opinions are quite constant throughout the evaluation. This fact suggests a high level of relative consistency between Judge 1 and Judge 3, which invalidates the hypothesis of a random evaluation. In conclusion, our analysis suggests that the data is reliable and it justifies its use.

[13]To measure Conger' κ we used the library *irrCAC* provided by the *R* software. More specifically, we used the function *conger.kappa.raw()* with the variable weights set to "ordinal".

[14]Goodman and Kruskal's *Gamma* was measured with the *GoodmanKruskalGamma()* function supplied by the *R* software.

We note that the description of our reliability study above Table 6.6 is mainly important for evaluation reproducibility. If there is a lack of space, due to conference or journal demand, we suggest reporting the following information in the appendix or in an online repository:

- The software used for the analysis;
- The information about the judges;
- The criteria used for the evaluation;
- The information about where to retrieve the dataset and the guidelines;
- The reason for choosing a coefficient and a scale of interpretation.

Likewise, the results from the IRT method we proposed can be reported in the appendix or in an online repository. For instance, returning to our example, we could use the IRT method to accompany the κ values in order to explain the annotation disagreement. This was performed in Section 6.3.3.

Looking ahead: What comes next?

The research presented in this thesis suggests some lines of future work. We can catalogue them in two main categories. One, titled "Reliability and bias identification", is strictly bound to the main topic of the thesis. The other one, titled "Evaluation of NLG systems", is more general, and it concerns topics that we have touched on throughout the development of the thesis.

Reliability and bias identification

In this thesis, we proposed using the correlation statistic and IRT to analyse judges' bias and reporting reliability for cases that involve a high level of language variability.

Another way to accomplish the goal that guided this thesis could be using Generalisability Theory (GT) (Cronbach et al., 1963). Such a method was proposed in Bayerl and Paul (2007). To the best of our knowledge it has not been used for NLG evaluation tasks. As demonstrated by Cronbach et al. (1963) GT can be a valuable tool, to be integrated with correlation coefficient and agreement coefficient, in order to assess data reliability.

Likewise, regression analysis methods (Freedman, 2009) (for instance ordinal regression (Winship and Mare, 1984)) could be used for the same aim. They can be adopted to formalise the concept of relative consistency. Indeed, regression analysis allows estimating the relationships between two variables.

A possible further development is linked to the concept of stability. In Chapter 5.1 we presented two procedures that can be used to measure stability. If relative annotations are involved, a third procedure can also be considered. Such a procedure, which involves the concept of transitivity, is explored by Amidei (2020a). Nevertheless, Amidei presents just the theoretical construction of the idea. Further investigation could involve: i) an extensive study and evaluation of the theoretical idea and ii) the use of possible weight in the preference judgement, representing the intensity of judges' preference.

Evaluation of NLG systems

Other topics were briefly considered in this thesis. All of them are important and they deserve further analysis.

Amongst these, the definition and analysis of significance tests for NLP tasks is one of the most important. This problem is quite delicate and urgent because significance tests assess the evidence provided by data about some claim concerning a population of interest. As showed by Koplenig (2019) and McShane et al. (2019), the matter of significance testing is coming under increasing scrutiny, and as such, it needs further investigation.

Another topic that needs further investigation is the development of cheaper extrinsic evaluation methods. As suggested in Reiter and Belz (2009), extrinsic evaluation methods "have traditionally been regarded as the most meaningful kind of evaluation in NLG" (page 531). Nevertheless, extrinsic evaluation methods are seldom used, because they are expensive and time-consuming. For this reason, the NLG community would benefit from new extrinsic evaluation methods. A possible way, also suggested by Gkatzia et al. (2015), could be the development of game-based evaluation setups.

As showed from the analysis we presented in Section 2.1.4 there is a wide assortment of criteria used across NLG papers. Also in this case, the NLG community would benefit from the definition of a basic set of criteria to be shared for intrinsic human evaluation study. A shared definition of such criteria could indeed bring consistency in the intrinsic human evaluation of NLG systems.

Finally, a topic which deserves great attention from the NLG community is the development of intrinsic automatic evaluation metrics which correlate well with human judgements. The advantage of automatic evaluation metrics is that of being cheap and repeatable. But without a strong correlation with human judgements, automatic metrics cannot be interpreted and their use would be difficult to justify.

Part III

Appendix

Appendix A

Annotation guidelines for the iterations presented in Chapter 4

In this Appendix we report the annotation guidelines for the text2text task we used in the four iterations presented in Chapter 4.

A.1 Question Generation evaluation guidelines: Iteration 1 (I1)

Task introduction

The purpose of this evaluation is to assess questions that are about a given English paragraph. You have to evaluate the goodness of a question following the criteria presented bellow.

Two pieces of information will be provided to you:

1. The input paragraph, and;

2. the question.

The questions are to be evaluated using the three criteria: *Syntactic correctness and fluency*, *Relevance* and finally *Specificity*. To each of these will be assigned a rank, with **-2** begin the **worst case**. We are using the principle: The higher the rank is the better the question is!

When a human evaluates the quality of a question (s)he takes a judgement based on some internal, maybe not defined, criteria. Here we want to make explicit some criteria that, from our point of view, can be useful in order to evaluate question quality. It must be stressed that the goal here is not to find all possible criteria usable in question evaluation, also because, they can change by task to task — in any task we could use a criteria that is not interesting in another task. So we would like to find some very general criteria that we expect be satisfied by any question evaluation assess.

Criteria description and examples

Language is complex and not all evaluations will be so clear cut as the examples given in this section. When dealing with such an evaluation it can be helpful to think of a it as a puzzle. A sensible approach would be to eliminate any obviously incorrect-ranked criteria first and then attempt to make an argument for the remain criteria. To assist with this approach, where appropriate, each of the criteria identifies a key point to consider when making the evaluation. Furthermore it is worth remember that here we are interested in evaluate a question based of its relation with the input paragraph. **Remember:** The input paragraph is the only knowledge required to answer the question.

Syntactic correctness and fluency

Description: The syntactic correctness is rated to ensure that the question is correct from a syntactical point of view. Please rank question higher if you fell the question is fluent to you.

Consideration: When assessing this criteria ask yourself the following question: **Can the question be comprehend?** If you cannot work out what the question is asking or what the answer is supposed to be then the criteria must be rank 0.

Example:

Input paragraph: World War I (WWI or WW1), also known as the First World War, the Great War, or the War to End All Wars, was a global war originating in Europe that lasted from 28 July 1914 to 11 November 1918. More than 70 million military personnel, including 60 million Europeans, were mobilised in one of the largest wars in history. Over nine million combatants and seven million civilians died as a result of the war (including the victims of a number of genocides), a casualty rate exacerbated by the belligerents' technological and industrial sophistication, and the tactical stalemate caused by gruelling trench warfare.

Rank	Description	Example
3	The question is grammatically correct and idiomatic/natural.	How long did the World War I last? How many civilians died as the result of the World War I?
2	The question is grammatically correct but does not read as fluently as we would like.	What does the belligerents' technological and industrial sophistication, as well as the tactical stalemate caused by gruelling trench warfare, involve?
1	There are some grammatical errors in the question.	Where was World War I originated?
0	The question is grammatically unacceptable.	Were where originating World War I?

If you have ranked 0 the Syntactic correctness and fluency then assess 0 the following criteria.

Relevance

Description: Question should be relevant to the input paragraph. This criteria measures how suitably the question can be answered based on what the input paragraph says.

Consideration: When assessing this criteria remember that: Only the information provided in the input paragraphs is important. **A question where the answer cannot be found in the input paragraph is not relevant. Could also be helpful to ask the following question: Can more information be added from the input text to make the question less general and more specific?**

Example:

Input paragraph: World War I (WWI or WW1), also known as the First World War, the Great War, or the War to End All Wars, was a global war originating in Europe that lasted from 28 July 1914 to 11 November 1918. More than 70 million military personnel, including 60 million Europeans, were mobilised in one of the largest wars in history. Over nine million combatants and seven million civilians died as a result of the war (including the victims of a number of genocides), a casualty rate exacerbated by the belligerents' technological and industrial sophistication, and the tactical stalemate caused by gruelling trench warfare. [...] In the introduction to his book, Waterloo in 100 Objects, historian Gareth Glover states: "This opening statement will cause some bewilderment to many who have grown up with the appellation of the Great War firmly applied to the 1914-18 First World War. But to anyone living before 1918, the title of the Great War was applied to the Revolutionary and Napoleonic wars in which Britain fought France almost continuously for twenty-two years from 1793 to 1815." In 1911, the historian John Holland Rose published a book titled William Pitt and the Great War. [...] . Russian political manoeuvring in the region destabilised peace accords that were already fracturing in the Balkans, which came to be known as the "powder keg of Europe." In 1912 and 1913, the First Balkan War was fought between the Balkan League and the fracturing Ottoman Empire.

Rank	Description	Example
3	The question can be unambiguously answered by the input paragraph.	How long did the World War I last? How many civilians died as the result of the World War I?
2	The input paragraph supplies more than one correct answer to the question.	How long did the Great War last? How long did the First War last?
1	The question is partially answered by the input paragraph.	Was the World War I a trench and guerrilla warfare fight?
0	The question cannot be answered by the input paragraph or the Syntactic correctness and fluency criteria was ranked as 0.	Was Portugal involved in the World War I? Was the World War I sad?

Specificity

Description: It is important to note that you are assessing a question relatively to a input paragraph, so sometimes, can happen that a question is relevant, because you are judging it with the paragraph as a contest. This is fine, but we are asking you to go further and analyze the relevance of the question by using the Specificity criteria. Indeed could happen that, although the question is relevant to the input paragraph, we can use the same question in more than one paragraph. That is, the question is so general that can be correctly answered in several, or without, contexts (that is different input paragraphs).

Consideration: When assessing this criteria ask yourself the following questions: **Can more information be added from the input text to make**

the question less general and more specific? and **Can I simply imagine a different context in which the question can be correctly answered?** and **Can I answer the question using very simply common sense, non specific, knowledge?** Could be also worth ask yourself the following question: **Can the question be answered without the input text?**

Example:

Input paragraph: World War I (WWI or WW1), also known as the First World War, the Great War, or the War to End All Wars, was a global war originating in Europe that lasted from 28 July 1914 to 11 November 1918. More than 70 million military personnel, including 60 million Europeans, were mobilised in one of the largest wars in history. Over nine million combatants and seven million civilians died as a result of the war (including the victims of a number of genocides), a casualty rate exacerbated by the belligerents' technological and industrial sophistication, and the tactical stalemate caused by gruelling trench warfare.

Rank	Description	Example
0	None of the following descriptions are applicable.	
-1	The question can be answered by the input paragraph as well as other paragraphs whose subject is different from the input one (i.e there are other paragraphs that can supply a correct answer).	How long did the World War I last? *[Think about the Wikipedia entry of World War II, note that this time the subject is not the World war I but the World War II]*
-2	The question can be answered by the input paragraph as well as by non specific knowledge. (The question can also be answered without any input paragraph).	How long did the War last? *[Think about your personal historical knowledge abut any War]*

A.2 Question Generation evaluation guidelines: Iteration 2 (I2)

Task introduction

The purpose of this evaluation is to assess the quality of questions which are relative to some English paragraphs. In this questionnaire you will be rating the quality of a number of questions using three criteria: *Syntactic correctness and fluency*, *Specificity* and *Pertinence*, the latter based on its relation to the input paragraph. To each of these is assigned a rank with **0** being the **worst case**. We are using the principle: the higher the rank the more acceptable the question.

Language is complex and not all evaluations will be so clear cut as the examples we are going to give in the next section. When dealing with such an evaluation, it can be helpful to think of a it as a puzzle and keep in mind that each of the criteria mentioned above identifies a key point to consider. When you are evaluating a question through the criteria a sensible approach could be as follows: for each criterion you could start by eliminate any rank that you feel as obviously incorrect and then attempt to make an argument for the remaining ranks. To assist with this approach we will give you some examples for each of the ranks of each criterion.

Criteria description and examples

(1) Syntactic correctness and fluency

Criterion description: The syntactic correctness is rated to ensure that the question is correct from a syntactical point of view. Apart from syntactical correctness or grammaticality, you will also be asked to judge how natural or fluent the sentence is. You are going to assess the question before reading the input paragraph.

Consideration: When assessing this criteria ask yourself the following question: **Can the question be comprehended?** If you cannot work out what the question is asking or what the answer is supposed to be then the criterion must be ranked 0.

Rank description:

Rank	Description	Example
2	The question is grammatically correct and idiomatic/natural.	How long did World War I last? How many civilians died as the result of World War I?
1	There are some grammatical errors in the questions or it sounds unnatural when read aloud, but you can still work out what the question is asking.	Where was World War I originated? What does the belligerents' technological and industrial sophistication, as well as the tactical stalemate caused by gruelling trench warfare, involve?
0	The question is grammatically unacceptable.	Originating were World War I where?

(2) Specificity

If you have ranked 0 the Syntactic correctness and fluency then assess 0 the Specificity criteria.

Criterion description: Specificity rate the question's degree of generality, that is we are trying to measure at which extend a question can be answered correctly in input paragraphs whose subjects are different. As the syntactic correctness and fluency criteria you will assess the question before reading the input paragraph.

Consideration: When assessing this criteria ask yourself the following questions: **Can I navigate through internet and find a clear and unambiguous answer to the question?** and **Can I imagine more than one input paragraph about a different subjects that could answer the question correctly?**

Rank description:

Rank	Description	Example
2	The question is very specific.	How long did World War I last? How many civilians died as the result of World War I?
1	The question is little specific.	How long did the War last? *[Note that any input paragraph whose subject is a War, for example a paragraph about the Vietnam War, can supplies an answer to the question.]*
0	The question is not specific and very general or the Syntactic correctness and fluency criteria was ranked 0.	How long it last? *[Note that any input paragraph whose subject is an event placed in a time scale, for example a paragraph about a concert or a box match, can supplies an answer to the question.]*

(3) Pertinence

If you have ranked 0 the Syntactic correctness and fluency then assess 0 the Pertinence criteria.

Criterion description: Question should be pertinent to the input paragraph. So Pertinence rate the degree to which the question is answered by the input

paragraph. If a question has not a clear answer in the input paragraph then it is not pertinent. Unlike the others criteria, you will be asked to assess the question in relation to the input paragraph.

Consideration: When assessing Pertinence remember that: Only the information provided in the input paragraph is relevant. **A question where the answer cannot be found in the input paragraph is not pertinent**.

Input paragraph:

Input paragraph: World War I (WWI or WW1), also known as the First World War, the Great War, or the War to End All Wars, was a global war originating in Europe that lasted from 28 July 1914 to 11 November 1918. More than 70 million military personnel, including 60 million Europeans, were mobilised in one of the largest wars in history. Over nine million combatants and seven million civilians died as a result of the war (including the victims of a number of genocides), a casualty rate exacerbated by the belligerents' technological and industrial sophistication, and the tactical stalemate caused by gruelling trench warfare. [...] In the introduction to his book, *Waterloo in 100 Objects*, historian Gareth Glover states: "This opening statement will cause some bewilderment to many who have grown up with the appellation of the Great War firmly applied to the 1914-18 First World War. But to anyone living before 1918, the title of the Great War was applied to the Revolutionary and Napoleonic wars in which Britain fought France almost continuously for twenty-two years from 1793 to 1815." In 1911, the historian John Holland Rose published a book titled William Pitt and the Great War. [...] Russian political manoeuvring in the region destabilised peace accords that were already fracturing in the Balkans, which came to be known as the "powder keg of Europe." In 1912 and 1913, the *First Balkan War* was fought between the Balkan League and the fracturing Ottoman Empire.

Rank description:

Rank	Description	Example
3	The question has a clear and unambiguous answer in the input paragraph.	How long did World War I last? How many civilians died as the result of World War I?
2	The question can be answered correctly by one or more incompatible responses given by the input paragraph.	How long did the Great War last? How long did the First War last? *[Note for both the questions the paragraph supply two correct answers.]*
1	The question is partially answered by the input paragraph, that is the input paragraph supplies information for answering only part of the question.	Was the World War I a trench and guerrilla warfare fight? *[Note the paragraph supplies informations about the trench warfare but there are not mentions about the guerrilla warfare.]*
0	The question cannot be answered by the input paragraph or the Syntactic correctness and fluency criteria was ranked 0.	Was Portugal involved in the World War I? Was the World War I sad?

A.3 Question Generation evaluation guidelines: Iteration 3 (I3)

Task introduction

The purpose of this evaluation is to assess the quality of questions which are relative to some English paragraphs. In this questionnaire you will be rating the quality of a number of questions using four criteria: *Syntactic correctness*, *Comprehensibility*, *Fluency* and *Pertinence*, the latter based on its relation to the input paragraph. To each of these is assigned a rank with **0** being the **worst case**. We are using the principle: the higher the rank the more acceptable the question. Language is complex and not all evaluations will be so clear cut as the examples we are going to give in the next section. When dealing with such an evaluation, it can be helpful to think of it as a puzzle and keep in mind that each of the criteria mentioned above identifies a key point to consider. When you are evaluating a question through the criteria a sensible approach could be as follows: for each criterion you could start by eliminate any rank that you feel as obviously incorrect and then attempt to make an argument for the remaining ranks. To assist with this approach we will give you some examples for each of the ranks of each criterion.

Criteria description and examples

(1) Syntactic correctness

Criterion description: The syntactic correctness is rated to ensure that the question is correct from a syntactical point of view.

Consideration: When assessing this criterion keep in mind you are just judging the question's grammaticality. If a question has some grammatical errors, although you can work out what the question is asking, then the

criterion must be ranked 0. Furthermore contextual informations such as anaphora, dates, name, places etc. don't count against question's grammaticality.

Rank description:

Rank	Description	Example
1	The question is grammatically correct.	How long did World War I last? How long did it last? How many civilians died as the result of World War I?
0	The question is grammatically incorrect.	Originating were World War I where?

(2) Comprehensibility

Criterion description: Comprehensibility is closely related to the previous criterion. Indeed, can happen that a question, although ungrammatical, is perfectly understandable that is it is possible working out what the question is asking (think about non native English speaker).

Consideration: When assessing this criterion ask yourself the following question: **Can the question be comprehended?** If you cannot work out what the question is asking or what the answer is supposed to be, then the criterion must be ranked 0.

Rank description:

Rank	Description	Example
1	It is possible work out what the question is asking.	How long did last World War I ? How many civilians, as the result of World War I, died?
0	It is not possible work out what the question is asking.	War did World I How?

(3) Fluency

Criterion description: Fluency rates how natural or fluent the question is.

Consideration: When assessing this criterion, if you feel that the question is awkward, ask yourself the following question: **Is there an obvious way to rewrite the question to improve its fluency?** If the answer is yes then the criterion must be ranked 0.

Rank description:

Rank	Description	Example
1	The question is idiomatic/-natural.	How long did World War I last? How many civilians died as the result of World War I?
0	The question is awkward or sounds unnatural when read aloud or the Syntactic or the Comprehensibility criteria were ranked 0.	What does the belligerents' technological and industrial sophistication as well as the tactical stalemate caused by grueling trench warfare involve?

(4) Pertinence

Criterion description: Because the question should be pertinent to the input paragraph, unlike the other criteria, you will be asked to assess the question

in relation to the input paragraph. So Pertinence rates the degree to which the question is answered by the input paragraph. If a question has no a clear answer in the input paragraph then it is not pertinent.

Consideration: When assessing Pertinence remember that: Only the information provided in the input paragraph is relevant. **A question where the answer cannot be found in the input paragraph is not pertinent**. When assessing this criterion ask yourself: **Is there a clear and unambiguous answer in the input paragraph?** If the answer is yes the criterion must be ranked 3.

Input paragraph:

Input paragraph: World War I (WWI or WW1), also known as the First World War, the Great War, or the War to End All Wars, was a global war originating in Europe that lasted from 28 July 1914 to 11 November 1918. More than 70 million military personnel, including 60 million Europeans, were mobilised in one of the largest wars in history. Over nine million combatants and seven million civilians died as a result of the war (including the victims of a number of genocides), a casualty rate exacerbated by the belligerents' technological and industrial sophistication, and the tactical stalemate caused by gruelling trench warfare. [...] In the introduction to his book, *Waterloo in 100 Objects*, historian Gareth Glover states: "This opening statement will cause some bewilderment to many who have grown up with the appellation of the Great War firmly applied to the 1914-18 First World War. But to anyone living before 1918, the title of the Great War was applied to the Revolutionary and Napoleonic wars in which Britain fought France almost continuously for twenty-two years from 1793 to 1815." In 1911, the historian John Holland Rose published a book titled William Pitt and the Great War. [...] Russian political manoeuvring in the region destabilised peace accords that were already fracturing in the Balkans, which came to be known as the "powder keg of Europe." In 1912 and 1913, the *First Balkan War* was fought between the Balkan League and the fracturing Ottoman Empire.

Rank description:

Rank	Description	Example
3	The question has a clear and unambiguous answer in the input paragraph.	How long did World War I last? How many civilians died as the result of World War I?
2	The question can be answered correctly by one or more incompatible responses given by the input paragraph.	How long did the Great War last? How long did the First War last? *[Note for both the questions the paragraph supply two correct answers.]*
1	The question is partially answered by the input paragraph, that is the input paragraph supplies information for answering only part of the question.	Was the World War I a trench and guerrilla warfare fight? *[Note the paragraph supplies information about trench warfare but there are no mentions of guerrilla warfare.]*
0	The question cannot be answered by the input paragraph.	Was Portugal involved in the World War I? Was the World War I sad?

A.4 Question Generation evaluation guidelines: Iteration 4 (I4)

Thank you for participating in this study. You are free to stop participating in the study at any time you want.

You will be presented with a number of instances of questions. You will then be asked for your judgement about these questions. Before starting to judge questions, please read the description and the examples in this guidelines document carefully. Do also feel free to refer back to this document at any time during the judgement process. Indeed, we encourage you to read this guidelines document anytime you have some doubt. There is no time limit to finish. You will first be asked about the grammaticality, fluency, and comprehensibility of a question. You will then be given a paragraph of text, and asked to judge whether the question is answered by that specific paragraph. You should only read this paragraph when prompted to so and do **not** change your answers regarding grammaticality, fluency, and comprehensibility after reading the paragraph.

1) As a first query we are going to ask you to evaluate the **grammaticality** of a question. If a question has some grammatical errors, although it is possible to work out what the question is asking, it has to be marked as ungrammatical. That means that, for instance, typos are considered errors. For example the questions below should be marked as ungrammatical:

i) World War I began in whic year? (whic instead of which):

ii) Who fought the in war of 1812? (inversion between the words "the" and "in"):

iii) Which general famously stated 'I shall return' (missing the question

mark "?").

Furthermore, if a question depends on contextual information such as anaphora, dates, names, places etc., then **don't** count this against question grammaticality. For example the questions below should be marked as grammatically correct:

i) i) What is the present day location of this church? (We don't have any idea about the question referent, nevertheless the question is correct from a grammatical point of view):

ii) When did it happen? or How many people were killed? (We don't have any idea about the event that the questions are about, nevertheless the questions are correct from a grammatical point of view):

iii) When was Leonardo born? (We don't have any idea who Leonardo is, nevertheless the question is correct from a grammatical point of view).

2) The second query is intended to see if the question is **comprehensible**. Indeed, it can happen that a question, although ungrammatical, is perfectly understandable. That is, it can be possible to work out what the question is asking, even if it is ungrammatical (think about non-native English speakers). The following are examples of this kind of question:

i) The Battle of Hastings in 1066 was in fought which country?

ii) Magna Carta were published by the King of which country?

iii) Which famous 5th century A.D conqueror were know as 'The Scourge of God'?

3) The aim of the third query is to judge whether the question is **fluent**. When considering a question, if you feel that it is awkward, ask yourself: Is there an obvious way to rewrite the question to improve its

fluency? If the answer is yes then consider the question as not fluent. For example:

i) Who the first was Western explorer to reach China? (We can rewrite the question in a more fluent way: Who was the first Western explorer to reach China?)

ii) The number of new Huguenot colonists declined after what year? (We can rewrite the question in a more fluent way: In which year did the number of new Huguenot colonists decline?)

iii) What nationality Hoesung Lee is? (We can rewrite the question in a more fluent way: What is the nationality of Hoesung Lee?)

American and British English differences don't count against question fluency. For example the question:

(iii) Jean De Rely's illustrated French-language scriptures were first published in what city?

may sound awkward to a British speaker, who could prefer the use of "which" instead of "what". However, "what" is more acceptable than "which" in American English. So in cases like this, please, consider the question as fluent. Furthermore note also that a typo could play against fluency. For example the question "World War I bugan in which year?" sounds more awkward than "World War I began in which year?" and indeed the second is an obvious way to rewrite the original question to improve its fluency. Also in cases like this, please, consider the question as fluent but mark it as ungrammatical.

4) Finally, we ask you to judge a question in the context of one or more paragraphs of text. You are asked to judge **the degree to which the question is answered by the text**. If you judge that the text does **not** provide a clear and unambiguous answer to the question, we ask

you the reason for your judgement. For example given the following paragraph:

Frank Vincent Zappa (December 21, 1940 - December 4, 1993) was an American musician, composer, activist and filmmaker. His work was characterized by non- conformity, free-form improvisation, sound experiments, musical virtuosity, and satire of American culture. Zappa was born in Baltimore, Maryland. His mother, Rosemarie (née Collimore) was of Italian (Neapolitan and Sicilian) and French ancestry; his father, whose name was Anglicized to Francis Vincent Zappa, was an immigrant from Partinico, Sicily, with Greek and Arab ancestry.

The question:

1) Where was Vincent Zappa born?

is **not** unambiguously answered, because the question can be answered correctly by more than one incompatible response. Indeed, based on the information given by the paragraph we can correctly answer this question by Baltimore (if the question is referring to Frank Vincent Zappa) or by Sicily (if the question is referring to Francis Vincent Zappa).

Given the same paragraph, the question:

ii) Was Frank Vincent Zappa a musician and a teacher?

is **not** clearly answered because the paragraph supplies information about the fact that Frank Zappa was a musician but there is no mention about the fact that he was a teacher.

We can also have questions like the following:

iii) Was Frank Vincent Zappa a tall man?

also this question is **neither** clearly **nor** unambiguously answered, because the paragraph does not supply the kind of information required by the ques-

tion.

Finally, there may be other reasons why the text does not clearly and unambiguously answer the question. In this case, please provide your reasons.

Sometimes the correct answer may not be explicit, but can be worked out from the information that is provided in the text. For example, coming back to the Frank Zappa question and associated paragraph:

Frank Vincent Zappa (December 21, 1940 - December 4, 1993) was an American musician, composer, activist and filmmaker. His work was characterized by non- conformity, free-form improvisation, sound experiments, musical virtuosity, and satire of American culture. Zappa was born in Baltimore, Maryland. His mother, Rosemarie (née Collimore) was of Italian (Neapolitan and Sicilian) and French ancestry; his father, whose name was Anglicized to Francis Vincent Zappa, was an immigrant from Partinico, Sicily, with Greek and Arab ancestry.

iv) Did Frank Vincent Zappa die when he was 53 years old?

can be unambiguously answered by the text. Indeed, we can use the paragraph information, specifically the date of birth and that of death, to infer the answer. When you are judging the last query use the following rule of thumb: only the information provided by the paragraph is relevant to answer the question!

List of Figures

1.1 Example of an intrinsic human evaluation item based on a numerical rating scale. 10

2.1 Number of papers on AQG published by year in the ACL anthology. 30

3.1 Raw data of the ambiguity evaluation by judge 1 and judge 3. 96

3.2 Example of unweighted percentage of agreement with the visualisation of weights and categories. 97

3.3 Example of unweighted percent of agreement without the visualisation of weights and categories 98

3.4 Example of unweighted weights for κ 99

3.5 Example of ordinal weights for κ. 100

3.6 Example of scale of interpretation in the case of ordinal weights for κ. 100

5.1 Plots of the evaluation of question fluency for non-native English speakers (a) and for native English speakers (b). For better readability, the scores are shifted upward slightly. . . . 134

5.2 Evaluation of question comprehensibility (a) and pertinence (b). Both the annotation were performed by non-native English speakers. For better readability, the scores are shifted upward slightly. 135
5.3 Distribution of the categories used by the judges in the Flickr-8k dataset. 137

6.1 IRCCC example for a mathematics exam. On the x-axis is reported the latent trait. On the y-axis is reported the probability of a correct answer for item 1. 153
6.2 A graphical description of the explanatory example. $P(\{x, y\}) < 0.5$ means that the probability of selecting score x or score y is smaller than 0.5, for $x, y \in \{1, 2, 3, 4\}$. The same interpretation applies also for the symbol $>$ (higher than) and for the symbol $=$ (equal to). 158
6.3 IRCCC for the ambiguity criterion for Judge 1 (a) and Judge 3 (b). The graphs show that the main source of disagreement can be found in the latent trait interval [0, 2] mainly relative to score 2 (red line). 162
6.4 IRCCC for the variety criterion for Judge 1 (a) and Judge 3 (b). The graphs are very similar which shows a low level of disagreement between the judges. 166
6.5 IRCCC for the fluency criterion for Judge 1 (a) and Judge 3 (b). The graphs show that the main sources of disagreement can be found in the latent trait interval [(±)1, (±)2.2] mainly relative to score 3 (green line) and score 4 (blue line). 168
6.6 IRCCC for the relevance criterion for Judge 1 (a) and Judge 3 (b). The graphs show as the small source of disagreement resides in the extreme part of the latent trait. 172

List of Tables

1.1 Krippendorff scale of interpretation for the Kappa statistic. AV denotes agreement value as determined by some coefficients of agreement. 22

1.2 Landis and Koch scale of interpretation for the Kappa statistics. AV denotes agreement value as determined by some coefficients of agreement. 23

2.1 Number of papers per year describing question generation systems. 29

2.2 Number of papers per conference proceedings or journal. . . . 31

2.3 Evaluation methodologies used. 34

2.4 Variation of the evaluation methodologies used between 2013 - 2015 and between 2016 - 2018. 35

2.5 Automatic metrics used. 36

2.6 Number of categories used in the Likert or rating scales. . . . 38

2.7 Criteria used. 40

2.8 Measures of IAA. 41

2.9 Dataset used. 44

2.10 Years and conference venue (number of papers in parentheses). 47

2.11 Mean, minimal and maximum IAA value per coefficient. # *used* means the number of times that a coefficient was used in total across the papers. In each paper each coefficient was used to measure the annotators' agreement about one or more questions or criteria. 51

3.1 Krippendorff scale of interpretation for the Kappa statistic. AV denotes agreement value as determined by some coefficients of agreement. 83

3.2 Landies and Koch scale of interpretation for the Kappa statistic. AV denote agreement value as determined by some coefficients of agreement. 84

4.1 Main source of disagreement for criteria found in our observational study. The label in the first column refers to the five categories delineated in this section: S&T (Style and taste), BK (Background knowledge), PA (Personal assumptions), CI (Use of common sense inferences), AD (Attention to detail). . 120

5.1 Rosenthal (1996) scale for the interpretation of correlation coefficient. 132

5.2 Results of Conger's κ and Yule's Q/Goodman and Kruskal's Gamma in the Iteration dataset. *All*, *Native* and *Non-native* indicate the measure performed respectively over the seven judges, over the three English native speaker judges and over the four no English native speaker judges. 133

5.3 Batches of question with independent judges assigned to them. For $i = 1, \ldots, 6$, J_i means judge i. 138

5.4 Conger's κ and Goodman and Kruskal's Gamma values reached in the QG-STEC dataset. For $i = 1, \ldots, 6$, J_i means judge i. The symbol $*$ indicate a p-value higher than 0.05. 138

5.5 Some correlation coefficients. Nominal, ordinal and interval/ratio represent level of measurement. 141
5.6 Cohen's (1988) scale for the interpretation of correlation coefficient. 141
5.7 Rosenthal's (1996) scale for the interpretation of correlation coefficient. 142
5.8 Political Science Department at Quinnipiac University scale for the interpretation of correlation coefficient. 142

6.1 Frequency of the scores used by Judge 1 (J1) and Judge 3 (J3).151
6.2 Scores and scores description for the ambiguity criterion. . . . 161
6.3 Scores and scores description for the variety criterion. 165
6.4 Scores and scores description for the fluency criterion. 167
6.5 Scores and scores description for the relevance criterion. . . . 171
6.6 Reliability study result from our example. 180

Bibliography

Aickin, M. (1990). Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to cohen's kappa. *Biometrics*, pages 293–302.

Altman, D. G. (1990). *Practical statistics for medical research*. CRC press.

Amidei, J. (2018). Supplementary material for the paper: "Evaluation methodologies in Automatic Question Generation 2013-2018". Retrievable from: `https://bit.ly/2IuPJIa`.

Amidei, J. (2019). Supplementary material for the INLG 2019 submission. Retrievable from: `https://bit.ly/2lKL516`.

Amidei, J. (2020a). Aligning intraobserver agreement by transitivity. `https://arxiv.org/pdf/2009.13905.pdf`.

Amidei, J. (2020b). Supplementary material for the chapter "Agreement coefficients comparison an experimental point of view". Retrievable from: `https://bit.ly/2CJZd3L`.

Amidei, J., Piwek, P., and Willis, A. (2019). The use of rating and likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations. In *Proceedings of The 12th International Natural Language Generation Conference, Tokyo, Japan, October 29 - November 1*.

Artstein, R. (2017). *Inter-annotator agreement*. In Handbook of Linguistic Annotation, (Eds.) Nancy Ide and James Pustejovsky, pages 297-313. Springer, Dordrecht.

Artstein, R. and Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Bai, S. and An, S. (2018). A survey on automatic image caption generation. *Neurocomputing*, 311:291–304.

Banerjee, S. and Laviel, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Bard, E. G., Robertson, D., and Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, pages 32–68.

Bayerl, P. S. and Karsten, P. I. (2011). What determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Computational Linguistics*, 37(4):699–725.

Bayerl, P. S. and Paul, K. I. (2007). Identifying sources of disagreement: Generalizability theory in manual annotation studies. *Computational Linguistics*, 33(1):3–8.

Belz, A. and Gatt, A. (2008). Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of ACL-08: HLT, Short Papers*, pages 197–200.

Belz, A. and Kow, E. (2010). Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the Sixth International Natural Language Generation Conference*, pages 7–9.

Belz, A. and Kow, E. (2011). Discrete vs. continuous rating scales for lan-

guage evaluation in NLP. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 230–235.

Bennett, E. M., Alpert, R., and Goldstein, A. (1954). Communications through limited-response questioning. *Public Opinion Quarterly*, 18(3):303–308.

Brennan, R. L. and Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement*, 41(3):687–699.

Carletta, J. (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Carterette, B., Bennet, P. N., and Chickering, D. M. (2008). Here or There: Preference Judgments for Relevance. *Computer Science Department Faculty Publication Series*, 46.

Chen, H., Liu, X., Yin, D., and Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. West Publishing Company, USA.

Colin, E., Gardent, C., M'rabet, Y., Narayan, S., and Perez-Beltrachini, L. (2016). The WebNLG Challenge: Generating Text from DBPedia Data.

In *Proceedings of the 9th International Natural Language Generation conference*, pages 163–167.

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88(2):322.

Corder, G. W. and Foreman, D. I. (2011). *Nonparametric statistics for non-statisticians*. John Wiley & Sons, Inc.

Craggs, R. and Wood, M. M. (2005). Evaluating discourse and dialogue coding schemes. *Computational Linguistics*, 3(3):289–296.

Cronbach, L. J., Rajaratnam, N., and Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 16(2):137–163.

Dawid, A. P. and Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Dušek, O., Novikova, J., and Rieser, V. (2017). Referenceless Quality Estimation for Natural Language Generation. *arXiv preprint arXiv:1708.01759*.

Dušek, O., Novikova, J., and Rieser, V. (2018). Findings of the E2E NLG Challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, Tilburg, The Netherlands. arXiv:1810.01170.

Eugenio, B. D. and Glass, M. (2004). The Kappa Statistic: A Second Look. *Computational linguistics*, 30(1):95–101.

Evans, R., Piwek, P., and Cahill, L. (2002). What is NLG? In *Proceedings of the International Natural Language Generation Conference*, pages 144–151.

Finlayson, M. A. and Erjavec, T. (2017). *Overview of Annotation Creation:*

Processes and Tools. in Handbook of Linguistic Annotation, (Eds.) Nancy Ide and James Pustejovsky, pages 167-911. Springer, Dordrecht.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Jhon Wiley and Sons, 13: 25.

Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Cambridge University Press.

Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., and Dolan, B. (2015). ΔBLEU: A Discriminative Metric for Generation Tasks with Intrinsically Diverse Targets. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 26–31.

Gatt, A. and Belz, A. (2009). Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In *Empirical methods in natural language generation*, pages 264–293. Springer.

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.

Gisev, N., Bell, J. S., and Chen, T. F. (2013). Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, 9(3):330–338.

Gkatzia, D., Curry, A. C., Rieser, V., and Lemon, O. (2015). A game-based setup for data collection and task-based evaluation of uncertain information presentation. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 112–113.

Gkatzia, D. and Mahamood, S. (2015). A snapshot of NLG evaluation practices 2005-2014. In *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, pages 57–60.

Glen, S. (July, 2020). Correlation Coefficient: Simple Definition, Formula, Easy Steps, Retrieved from: `https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/`.

Godwin, K. and Piwek, P. (2016). Collecting Reliable Human Judgements on Machine-Generated Language: The Case of the QG-STEC Data. In *The 9th International Natural Language Generation conference*, pages 212–216.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49(268):732–764.

Groves, I., Tian, Y., and Douratsos, I. (2018). Treat the system like a human student: Automatic naturalness evaluation of generated text without reference texts. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 109–118.

Gulliksen, H. (1950). *Theory of mental tests*. Wiley: New York.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Hinkle, D. E., Wiersma, W., and Jurs, S. G. (2003). *Applied statistics for the behavioral sciences*, volume 663. Houghton Mifflin College Division.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Holley, J. W. and Guilford, J. P. (1964). A note on the G index of agreement. *Educational and psychological measurement*, 24(4):749–753.

Hovy, E. and Lavid, J. (2010). Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*, 22(1):13–36.

Joshi, A., Bhattacharyya, P., Carman, M., Saraswati, J., and Shukla, R. (2016). How do cultural differences impact the quality of sarcasm annotation?: A case study of indian annotators and american text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99.

Koplenig, A. (2019). Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*, 15(2):321–346.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B. J., Hróbjartsson, A., Roberts, C., Shoukri, M., and Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *International journal of nursing studies*, 48(6):661–671.

Krippendorff, K. (1980). *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Beverly Hills, CA.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

LeBreton, J. M. and Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational research methods*, 11(4):815–852.

Lin, C.-Y. and Och, F. J. (2004). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 605–612.

Liu, C. W., Lowe, R., Serban, I. V., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Lombard, M., Snyder-Duch, J., and Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human communication research*, 28(4):587–604.

Lommel, A., Popović, M., and Burchardt, A. (2014). Assessing Inter-Annotator Agreement for Translation Error Annotation. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), 26-31 May. Reykjavik, Iceland.*

Mazidi, K. (Last use: March 2018). Question generator, freely available at: https://kjmazidi.pythonanywhere.com/.

McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1):235–245.

Novikova, J., Dušek, O., Curry, A. C., and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In *In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Novikova, J., Dušek, O., and Rieser, V. (2018). RankMe: Reliable human ratings for natural language generation. *arXiv preprint arXiv:1803.05928*.

Palmer, M. and Xue, N. (2005). Linguistic Annotation. *Computational Linguistics*, 31(1):71–106.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Parloff, R. (2016). Why deep learning is suddenly changing your life. *Fortune. New York: Time Inc.*

Passonneau, R. J. and Carpenter, B. (2014). The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Piwek, P. and Boyer, K. E. (2012). Varieties of question generation: introduction to this special issue. *Dialogue & Discourse*, 3(2):1–9.

Pustejovsky, J. and Stubbs, A. (2013). *Natural Language Annotation for Machine Learning*, volume 1. Published by O'Reilly Media, Gravenstein Highway North, Sebastopol, CA.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, November 1-5*, pages 2383 – 2392.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Reiter, E. (2011). Task-Based Evaluation of NLG Systems: Control vs

Real-World Context. In *Proceedings of the UCNLG+ Eval: Language Generation and Evaluation Workshop*, pages 28–32.

Reiter, E. (2018). A structured Review of the Validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Reiter, E. (January 2017). Types of NLG Evaluation: Which is Right for Me? [Blog post], Retrieved from: https://ehudreiter.com/2017/01/19/types-of-nlg-evaluation/.

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, E., Robertson, R., and M.Osman, L. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, 144(1-2):41–58.

Rizopoulos, D. (2018). Latent Trait Models under IRT, Retrieved from: `https://cran.r-project.org/web/packages/ltm/ltm.pdf`.

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of social service Research*, 21(4):37–59.

Rus, V., Cai, Z., and Graesser, A. (2008). Question generation: Example of a multi-year evaluation campaign. *Proc WS on the QGSTEC*.

Rus, V. and Lintean, M. C. (2012). A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In *In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*.

Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., and Moldovan, C. (2010). The first question generation shared task evaluation challenge. In *Proceedings of the Sixth International Natural Language Generation*

Conference (INLG 2010), 7-9 Jul 2010, Trim Castle, Ireland, pages 251 – 254.

Rus, V., Wyse, B., Piwek, P., Lintean, M., Stoyanchev, S., and Moldovan, C. (2012). A detailed account of the first question generation shared task evaluation challenge. *Dialogue & Discourse*, 3(2):177–204.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Sampson, G. (2002). *English for the computer: The SUSANNE corpus and analytic scheme*. MIT Press.

Sampson, G. (November 2017). The SUSANNE Analytic Scheme [Blog post], Retrieved from: https://www.grsampson.net/rsue.html.

Sampson, G. and Babarczy, A. (2008). Definitional and human constraints on structural annotation of english. *Natural Language Engineering*, 14(4):471–494.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly*, pages 321–325.

Sedoc, J., Ippolito, D., Kirubarajan, A., Thirani, J., Ungar, L., and Callison-Burch, C. (2018). ChatEval: A Tool for the Systematic Evaluation of Chatbots. In *Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG)*, pages 42–44.

Sharma, S., Asri, L. E., Schulz, H., and Zumer, J. (2017). Relevance of unsupervised metrics in task-oriented dialogue for evaluating natural language generation. *arXiv preprint arXiv:1706.09799*.

Shimorina, A. (2018). Human vs automatic metrics: on the importance of correlation design. *arXiv preprint arXiv:1805.11474*.

Siddharthan, A. and Katsos, N. (2012). Offline sentence processing measures

for testing readability with users. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 17–24.

Siegel, S. and Castellan, N. J. J. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw-hill, New York.

Singh, K. (2007). *Quantitative social research methods*. Sage, New Delhi.

Stemler, S. E. and Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. *In, Osborne, J. W., (ed.) Best practices in quantitative methods*, pages 29–49. Sage, California.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680.

Streiner, D. L., Norman, G. R., and Cairney, J. (2015). *Health measurement scales: a practical guide to their development and use*. Oxford University Press, USA.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind*, 59(236):433.

Van der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., and Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Viethen, J. and Dale, R. (2007). Evaluation in natural language generation: Lessons from referring expression generation. In *TRAITEMENT AUTOMATIQUE DES LANGUES 48 (1)*, pages 141–160.

Wasserstein, R. L., Schirm, A., and A., L. N. (2019). *Statistical Inference in the 21st Century: A World Beyond $p < 0.05$*, volume 73, sup1. Taylor & Francis.

Winship, C. and Mare, R. D. (1984). Regression models with ordinal variables. *American sociological review*, pages 512–525.

Witte, R. S. and Witte, J. S. (2017). *Statistics, Eleventh Edition.* John Wiley & Sons, Inc., LaVergne, Tennessee, USA.

Yuan, X., Wang, T., Gulcehre, C., Sordoni, A., Bachman, P., and Zhang, S. (2017). Machine Comprehension by Text-to-Text Neural Question Generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, 75(6):579–652.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675*.