TOWARDS BETTER EVALUATION METHODS FOR CONTROLLABLE AFFECTIVE NATURAL LANGUAGE GENERATION

Joris Brandt
j.m.brandt@students.uu.nl

Supervised by:
Kees J. van Deemter,
Lynda H. Hardman

08/12/2021

Abstract

In this thesis we discuss Natural Language Generation (NLG), and the more specific domain of Affective Natural Language Generation. We discuss the concepts of having Controllable Affective Natural Language Generation (CA-NLG), a specialized area focusing on controlling the emotions in generated language output. We also discuss various means of evaluating NLG systems and argue current research overfocuses on evaluating NLG systems intrinsically, and more extrinsic evaluations should be performed. We elaborate on specific means for measuring emotions in readers of NLG output texts, and related issues. We discuss various measuring systems and design an experiment to compare a recent proposal named the Geneva Emotion Wheel (GEW) against a well-known, more established method called the Positive Affect Negative Affect Scale (PANAS). We compare the two methods' test-retest reliability and find similar results for the former and a slight bias in favor of PANAS for the later. However due to sample size issues we cannot conclusively say which method performs better, objectively.

Acknowledgements

This work was supported by Utrecht University, as part of a master's degree program in Artificial Intelligence. We thank Professor C.J. van Deemter, Professor L. Hardman, and F. Qixiang for supervising this work, as well as Professors I. van Der Sluis, and C. Janssen for their assistance.

Table of Contents

1. Introduction

Affective Natural Language Generation is an area within the broader research area of Natural Language Generation (NLG) which focuses on producing texts intended to communicate a certain emotion or carry a certain "affective" message. Research in this area can focus on things such as word choice, lexical alignment with an interlocutor, or controlling the tone of automatically generated text. Within this area we have more recently seen the rise of a specific type of application, in which the affect or emotion of an NLG system's output is intended to be customizable and controllable by its user. We will refer to this as Controllable Affective Natural Language Generation (CA-NLG).

Evaluation is key in answering questions of effectiveness, naturalness of the outputted language, or the ability to generate consistent results.(van der Lee et al., 2019) Output texts can be evaluated in different ways, but among them two main categories can be defined. The first being intrinsic evaluation, which focuses on the properties and technical aspects of the system itself. This usually concerns itself with grammaticality, fluency, or the degree to which an outputted text resembles a 'gold standard' text. The second being extrinsic evaluation, which focuses more on the target audience of the text and the effects produced on them. For example the time it takes to read or how they judge a text based on their personal interpretation.

In recent years many evaluations strategies have focused on intrinsic evaluations, using metrics such as BLUE or ROUGE (van der Sluis & Mellish, 2010), or more recently MoverScore (Zhao et al., 2019). There is good reason for this: these evaluations are easy to reproduce, modify, and test (Gkatzia & Mahamood, 2015). Moreover, they often address a primary concern of any NLG pipeline: whether the text it produces is grammatical and fluent, which makes it easier to also use as an intermediate step in automated learning systems for AI-based NLG. After all, if the system can be given a concrete number by which to judge output it can more easily identify which of its results should be kept and which discarded.

Unfortunately, in our view, a bias towards intrinsic evaluation loses sight of the goal of NLG: to generate texts which are not only grammatical and fluent but are also useful towards accomplishing a certain goal. Of course, there exists a broad range of applications and so judging all applied NLG with the same standards would be impossible. But this should not be an impediment towards developing the underpinnings of evaluation methods which can then be modified for domain-specific approaches. More specifically: many CA-NLG systems today appear to focus mostly on generating texts which resemble those a 'gold standard' text or are judged favorably by experts, rather than measuring whether their intended purpose, to produce an emotional reaction in a subject, is achieved. The result is that we have a rich body of research and systems which claim to achieve controllable, customizable, affective text. But many of these are not actually ecologically valid or even effective at producing, or preventing, an emotional response.

This is not without reason, nor the product of negligence on the part of researchers. Extrinsic evaluation can present many difficulties compared to intrinsic evaluation, and often requires an extremely fine-tuned approach. In terms of measuring emotions we can find multiple problems not just within the domain of NLG (van der Sluis et al., 2011) but more broadly in any kind of evaluation which aims to measure impact on emotions or

emotional responses (Mauss & Robinson, 2009). It is therefore the goal of this work to study the different ways in which extrinsic evaluation of CA-NLG system is performed, to note their strengths and weaknesses, and perhaps to propose alternatives or best practices to evaluate CA-NLG systems. Our purpose is to move towards the possibility of more reliably testing the effectiveness of such systems in producing emotional responses in audiences for generated texts. We believe that in improving the ways these systems are evaluated we can not only improve their design, but also provide insights into their functions. Which processes, or designs prove effective, and perhaps even what biological or psychological process underpins their effectiveness. Beyond proving that CA-NLG systems are effective, this could help understand and explain why they are, and what lessons or causal relations we can infer from them.

We will begin by further elaborating on the field of Affective NLG, it's challenges and recent examples. From there we will discuss several methods for evaluating Affective NLG systems, before expanding on the broader issues in measuring emotions. We will discuss methods such as the Positive Negative Affect Scale (PANAS), as well as possible alternatives. To validate one such method, the Geneva Emotion Wheel (GEW), we will detail an experiment and its results. Our focus will be on comparing the test-retest reliability, and the speed of completion, of the more recent GEW against that of PANAS. We will find that they perform similarly, with PANAS presenting a slight edge in terms of completion times. However due to a small sample size we will have limited ability to draw conclusions from said results.

2. Affective NLG

To better understand Affective NLG, it is perhaps wise to first elaborate on what Natural Language Generation (NLG) is. In previous surveys (Belz & Reiter, 2006; Reiter, 2001; Reiter & Dale, 1997) of NLG methods this was defined as *`the sub-field of artificial intelligence and computational linguistics that is concerned with the construction of computer systems than can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information' (Reiter & Dale, 1997, p. 1)*. This definition focuses mostly on generating text from data, also referred to as *data-to-text,* by the authors' admission (Reiter & Dale, 1997). Such applications could be used, for example, to summarize medical data in human-readable language (Reiter, 2007; Reiter et al., 2008). However, NLG systems can also be *text-to-text,* for example for use in summarizing existing texts or generating new texts from textual information (Haim & Graefe, 2017). More recent surveys (Gatt & Krahmer, 2018) also note the rise in systems which use images to generate texts (Huber et al., 2018).

While it can be argued that most of these examples are *data-to-text*, as the images or computer-readable will need to be converted to computer-readable data in the process, the key difference is that the data in question is generated within the system, as opposed to the previous examples of medical data, which already exists prior. Although it is difficult, therefore, to define NLG by its input it's largely agreed on what it is supposed to output: human-readable texts. While these can then be used to generate spoken utterances(Rieser & Lemon, 2010), the NLG part of such a system often still generates text to then be read by a different system which generates the sound. For this reason, the challenges of NLG are, by and large, linguistic. Transforming non-linguistic data to a textual representation requires various tasks including determining the content to include, aggregating different sentences to form a single cohesive text, and choosing which words, phrases, or expressions to use. But also, more complex tasks such as generating referring expressions(van Deemter, 2016).

Gatt and Krahmer identified three dominant approaches in creating NLG systems (Gatt & Krahmer, 2018). *Modular* approaches, which divide the different tasks into different systems, which can vary wildly in design and approach. *Planning* approaches, which use slightly more integrated systems in a series of planned actions. And *Integrated,* or *Global* approaches which seek to integrate all tasks into a single system. Within these approaches choices also must be made between rule-based and more stochastic or statistical approaches. The latter has recently been especially favored in research (Gatt & Krahmer, 2018) with recent developments in Machine Learning, Neural networks, and especially Deep Learning techniques (Otter et al., 2019).

Affective NLG takes its name from the field of "Affective Computing": computing which involves or influences emotions (Picard 1997). The goals of Affective NLG are similarly broad, they can involve tasks such as generating texts which "contain" a certain emotion, texts which induce a specific feeling in a reader, or even to the contrary: texts which carefully avoid doing this. There may be many different reasons as to why designing a system with emotion in mind might be desirable, it could be used to improve interactive learning systems by acting more efficiently on a learner's desire to be taught and motivation to continue (Lester, 2012). It could be used to generate texts in a professional environment which effectively communicate vital information without risking emotional

'noise' influencing understanding of said information. It can help in creating models aimed at better understanding human behavior by mimicking how emotions can influence said behavior. Affective NLG then becomes a specialized area within the domain of Affective Computing, broadly sharing the same interests but focused specifically on the process of creating expressions in natural language (Gatt & Krahmer, 2018).

In practice this says very little about the actual approach chosen by researchers, as the idea is broadly compatible with classical NLG pipelines or more modern Machine Learning-based systems. It is not out of the question to design an Affective NLG system which does not offer a great deal of control over the type or intensity of emotion expressed. But in some cases, having this control could be considered vital for the proper use of the system, either as part of its goals or as to avoid interfering with them. In these cases, one could speak of Controllable Affective NLG (CA-NLG), as referring to systems generating natural language expressions with a certain affective load. As an example, let us consider previous work on generating summaries of medical data concerning neonates in care at a hospital (Portet et al., 2009). Part of the requirement of this application was to provide parents with summarized updates of the neonate's treatment, one can imagine that, in this situation, it is important they not be made to worry unnecessarily nor to downplay the gravity of the situation. In these cases, a method to ensure the proper affect in the messages could potentially be very useful (Mahamood & Reiter, 2011).

In another example, the project "blogging birds" focused on generating summarized stories of the movement of birds across the northern United Kingdom, with the intended effect of generating more interest in the conservation efforts of these birds (Siddharthan et al., 2019). For this purpose, texts were generated automatically and later modified by humans to create engaging narratives. Such a system could benefit from a way to control the affective charge of the outputted, generated text to increase the scale of operations or reduce the workload of the human editors.

It is perhaps also wise to mention that there are potential abusive applications to such a system. Ethical concerns already exist for the use of NLG systems to create misleading information (Dörr & Hollnbuchner, 2017; Smiley et al., 2017). One could imagine, for example, a system designed to create targeted disinformation campaigns aimed at provoking certain sentiments to further polarize political discourse or destabilize democratic processes. Automation of the generation of incensing or provocative media could make the tasks of countering misinformation harder at scale. On the other hand, developing an understanding of how such media is generated could also provide avenues of creating effective detection and/or countermeasures at a similar scale. In any case we believe it wise to approach this topic with a certain amount of caution as to the effects of real-world applications of such systems.

Various examples of existing CA-NLG systems exist. Often diverging in how to approach the basic question of how affect must be included in generated natural language. Hinojosa, Moreno, and Ferré argue for the use of models based on neurocognitive processes which they suggest represent emotion in the human brain as a set of semantic features (Hinojosa et al., 2020). In their paper they argue that the process by which the brain processes emotional information is like, or possibly activated by, the same process by which the brain processes semantic information but is also influenced by other factors such as lexical information, temporal variation such as time taken to read, and even at what age certain

emotional words were first learned. The research further suggests, that processing this emotional information influences understanding of sentences (Hinojosa et al., 2020) which would further underscore the potential usefulness of controlling the outputted emotions.

Approaches such as Affect-LM (Ghosh et al., 2017), Emotional Chatting Machine (Zhou et al., 2018), and one approach by Kezar (Kezar, 2018) rely on *Sequence to Sequence* neural network architectures to generate text, due to recent interest in such systems and their proven effectiveness (Mikolov et al., 2010). *Sequence to Sequence* systems, or seq2seq for short, essentially use two separate neural networks, called an *Encoder* and a *Decoder*. The Encoder takes input data, encodes it into a smaller dimensional representation, then passes it on to a Decoder which will create a textual output based on said representation. Advantages of such systems include their ability to be language-agnostic depending on the data used to train the system. For example, a seq2seq system trained on an English corpus will comprehend and output texts in English, but the same system trained on Mandarin data will operate in Mandarin.

These systems are then supplemented with additional systems to insert a desired emotion in a controlled fashion. In Affect-LM this is done by including emotional annotation in the training corpus and adding additional weights to the LSTM model which bias towards words categorized with a certain emotion, and determine their frequency based on a desired "strength" of affect provided by the operator (Ghosh et al., 2017). However, this approach is limited in its effectiveness since it cannot balance grammaticality and the expression of the desired emotions and will often generate correct but hard-to-perceive emotional expressions, relying heavily on commonly understood words which may or may not contain a degree of ambiguity. The ECM system was therefore designed with a more dynamic generation mind, relying on the fact that an emotional expression might change as more information is added to a sentence (Zhou et al., 2018) However, as noted by Kezar in a 2018 paper, these models are limited in that they can only accept one emotion as input, excluding the possibility of having multiple emotions or affective loads present in a same text or sentence. Kezar argues that this is an inherently limiting approach which does not reflect how humans communicate in the real world, nor the idea that we can experience more than one discreet emotion at a time. This also limits the system's capacity for nuance as it cannot balance its emotional output, it can only increase or decrease it as desired by the operator. Kezar proposed an extension of the system used in Affect-LM to allow for the inclusions of multiple emotions, while also incorporating some of the advances proposed in ECM to avoid over-emphasizing the emotional content of a message and to retain grammaticality (Kezar, 2018). Similarly, work by Xi and Sun (J. Li & Sun, 2018) sought to alleviate these issues by creating a syntactically constrained model. Like ECM and Affect-LM this is a conversational agent system which puts more emphasis fluency and logical soundness of generated text to ensure replies stay on topic whilst still containing the desired emotion (J. Li & Sun, 2018).

As an alternative to the above approaches, Santhanam and Shaikh designed a model based on a *transfer learning* system. Transfer learning allows one to use a pre-trained "base" model such as GPT-2 or BERT, which drastically reduces the need for additional configuration or training data. Transformer models such as GPT-2 have performed extremely well in generating fluent, logical texts which preserve context (Vaswani et al., 2017). Santhanam and Shaikh argue these systems can be easily re-trained to include

more parameters such as emotion, and provide output comparable to, or even better than, approaches such as Affect-LM or ECM (Santhanam & Shaikh, 2019).

However, all previously mentioned research presented some other, in our view distinctive, flaws: most only evaluated the grammaticality of the output, compared it to a “gold-standard” text, or evaluated the emotionality using a small number of human judges. In the following part we will explain why we believe this is an inherent limitation in researching CA-NLG.

3. Current trends in evaluating NLG systems

Evaluations can be divided into several distinct categories, however two often recurring categories are *Intrinsic* and *Extrinsic* evaluations(Belz & Reiter, 2006; Celikyilmaz et al., 2021). *Intrinsic* evaluations focus on the intrinsic properties of the model itself. Such evaluations involve having human subjects rate texts for the presence of a desired property. Evaluating a model this way can still be costly and time consuming, and does not always provide a straightforward or consistent results (Amidei et al., 2019). But this can be improved by using a ranking system, for example asking judges to rank different output texts in order of preference using a system like RankME (Celikyilmaz et al., 2021).

Alternatively, texts could be compared to a baseline or "gold standard", then rated on their similarities using a metric such as BLUE, NIST, or ROUGE. BLUE had become particularly popular due to its easy implementation and high rate of use in Natural Language Processing evaluations. However, metrics such as BLUE are somewhat limited in their usefulness, and their validity as an evaluation tool has been called into question (Reiter, 2018). More recently other metrics such as MoverScore (Zhao et al., 2019), Sentence Mover's Distance, Word Mover's Distance, or PYRAMID have also gained in popularity, as many seek to fix shortcomings of older metrics(Celikyilmaz et al., 2021). A big advantage of using these metrics being their capacity to be automated or performed by a computer, which greatly cuts down on both time needed and manpower required to perform such evaluations(Belz & Reiter, 2006; Celikyilmaz et al., 2021), and allowing for more rapid iteration in designing NLG systems.

On the flip side *extrinsic* evaluations, which focus more on measuring the impact of a system and/or its effectiveness at a given task. This can be accomplished in different ways, for example by comparing performance of human subjects on a given task before or after being presented with a generated expression or measuring reported emotions by human readers (Belz & Reiter, 2006). Such task-based evaluations are expensive and complicated to perform (Reiter, 2011), especially in the realm of Affective NLG (Strapparava & Mihalcea, 2007). And are not especially favored for this reason (Celikyilmaz et al., 2021; Gkatzia & Mahamood, 2015). Alternatively, success-based extrinsic evaluations can be used to measure a system's ability to complete a task for which it was intended. For example a 2003 paper by Reiter, Robertson, and Osman designed an NLG system to help people quit smoking by generating personalized messages encouraging them to do so, and measured it's success by the rate at which participants effectively quit smoking (Reiter et al., 2003).

The vast majority of NLG systems are evaluated using said intrinsic evaluations, whether it be by human judges or using automated systems (Belz & Reiter, 2006; Celikyilmaz et al., 2021). In a 2015 meta-analysis of almost a decade of NLG papers more than 74% of all papers reported intrinsic evaluations, versus only 15% extrinsic and 10% using both (Gkatzia & Mahamood, 2015). This bias towards intrinsic evaluation is easy to explain. Intrinsic evaluations are widely known and used. Moreover, they often involve metrics that rely on real data and on a system's actual properties, rather than subjective judgements from human subjects.

Such metrics also opened the possibility of performing these evaluations automatically allowing for faster design iteration and removing the need for human subjects. Meaning

research using such evaluations can be faster and cheaper (Belz & Reiter, 2006). They are also easier to translate outside the realm of academic research, for example in the corporate sector. Having a singular number or "score" on which the model is judged can provide potential investors with a concrete number to aid them in deciding which system will be more performant in their application.

While such concerns are valid, it has been argued they inherently limit the evaluation of research and NLG models(van der Lee et al., 2019), as they do not take into account how a model will then perform in a real-world setting. In such a context its output is no longer judged by qualified judges or automated systems, but by everyday people who may approach it with their own set of biases and may not be inclined to judge the model on its own merits but rather on the value it provides to them. Reiter also expressed such sentiment in a 2011 paper arguing there had been an overfocus on control groups, and not enough focus on task evaluation for NLG in real-world scenarios (Reiter, 2011).

We agree with these positions and believe that, to advance the field of CA-NLG, it is necessary to put more focus on extrinsic evaluations. As they give the clearer picture of how a system performs at the actual task for which it is designed. Knowing the effectiveness at accomplishing said tasks should not be a secondary concern. As the field of NLG grows, and more and more practical applications are found for NLG systems, practical and ethical risks arise (Dörr & Hollnbuchner, 2017), risks which are more likely to be captured in task-based extrinsic evaluations. Omitting such evaluation, while understandable, deprives us of valuable insights into how humans interact with NLG systems, which in many ways should be the end goal of many such systems. Intrinsic evaluations do not suffice on their own as the evidence does not suggest the score of intrinsic evaluation methods correlate to extrinsic evaluation scores (Reiter & Belz, 2009). Ideally both types of evaluation should be performed where possible.

However, in the case of CA-NLG, it is not yet entirely clear what tools or methods are preferred to measure the effectiveness of a system. While this is also heavily dependent on the context and nature of the task, one must also clearly state the goal of the system in question (van der Sluis et al., 2011; van der Sluis & Mellish, 2010). In the case of the BABYTALK system the goal had been established clearly as providing meaningful, correct texts to communicate information to different groups (Portet et al., 2009). Or, for example, in the case of the research by Reiter at al. into using NLG systems to assist in giving up smoking (Reiter et al., 2003), where the system could be partially judged by the number of people who did effectively give up smoking. But what off the case of an affective CA-NLG system? What if the intent is to induce an emotional response, or to avoid doing so? In such cases two possibilities are open: either the subjects judge the emotional qualities of the generated texts, and this is used as a metric of effectiveness, or the emotional state of the subjects is measured to determine if the reaction is as expected in both the type of emotion and/or it's intensity. The latter, however, presents several problems, as we will explore.

3.1 Standard Methods in Measuring emotions

Extrinsic evaluation, specifically measuring emotions, presents a host of problems and complications making it hard to argue conclusions or difficult to set up a proper evaluation.

In a 2009 review of several studies Mauss and Robinson noted several of these issues, concluding that there was no “gold standard” of emotional measurement which could be universally applied across studies, and that even between methods enough variance existed that convergence was often not possible (Mauss & Robinson, 2009).

It should also be noted that, while there exists literature arguing in favor of conceptualizing emotions as discrete and tied to unique physiological processes, it seems much more likely now that emotions do generally exist as a continuous measure. Although more discreet, specific emotions might still be isolated within such data (Mauss & Robinson, 2009).

One very popular instrument is self-reporting methods for emotions: simply asking participants to judge their own emotional state. However participants may not always be the best judges of said emotional state as said judgement can be subconsciously influenced by the framing of the task, the question, or even the conditions in which it is performed (Tversky & Kahneman, 1981). Alternatively, measurements can be taken by asking participants to perform a task and judging variations in results, however this too is prone to confounds introduced by framing, conditions, or simply by pre-existing biases between or particular to participants (van der Sluis et al., 2011).

These effects can be compounded by a difference between emotional experience, which happens in the moment, and emotional recall, which occurs later as the participants is asked to describe a prior emotional state. Such effects have already been proven to influence people to judge their past emotional states more positively, sometimes colloquially referred to as the “rose-tinted glasses effect” (Mitchell et al., 1997). Research has found that, the further temporally removed a person is from their experience of an emotion, the more their account starts to differ from what may have been recorded at the time (Robinson & Clore, 2002). Additionally, it has been suggested that, even in cases where subjects are asked to report in a shorter period after experiencing the emotion, the act of active cognition regarding their experience alters their memory of it, as people are more likely to report an emotional reaction which they are “supposed” to feel (Chan, 2008). Such effects even see differences influenced by gender, age, and level of education (Barrett et al., 1998), as people will heuristically remodel their memory of an experienced emotion to be in-line with their self-image. Such findings suggest that, in the case of self-reporting, it is paramount to clearly define what is exactly being measured: the experience of the emotion “in the moment” or the memory of said experience. In the case of the former it is then suggested the measurement must take place as soon as possible, leaving the subject as little time as possible to think about and intellectualize their reaction (Mauss & Robinson, 2009). But doing this requires a self-reporting system that is both intuitive to use and abstract enough that the subject does not additional time thinking about their answer before giving it. In some cases, it could even be argued that it might be more fruitful to have the subject merely state their emotions and have an observer note this, rather than asking the subject to switch their attention to their reporting form. However, this may then introduce additional issues and observer effects such as subjects’ experiences being affected by the presence of the observer, or observers having to interpret answers leading to perhaps a more subjective reporting of the subject’s emotional experience (Robinson & Clore, 2002).

An alternative method then, could be not to directly ask a subject to state their emotions, but rather to give them another task, such as writing a text, and performing a sentiment

analysis to determine if any effects can be observed. The theory being that a subject would, subconsciously, allow their emotions to “slip in” into the task, as it were, which would not require them to actively think about said emotions. However, this method, too, presents several issues. Primarily in that it once again asks a subject to shift attention away from another experimental task if one is present, or in the case of a delayed report, asking a subject to recall their thoughts and feelings instead of measuring them in the moment. Additionally, systems used to perform sentiment analysis must be carefully chosen, as many popular and easy-to-implement ones are based on neural network architectures, care must therefore be taken to choose a system which is proven to handle the task well in the chosen context and deals well with outliers or unexpected data. Sadly, also, these systems suffer from a familiar “black box” issue where it is hard to understand why the system makes it judgements or which data points it favors in doing so(Gatt & Krahmer, 2018). This can make it difficult to be sure the system is adequately picking up emotions, as it offers no sets of verifiable rules by which it does so.

Another proposed method then could be to rely on physiological measurements such as skin conductance or neural activity. But this presents both logistical challenges, as it requires specialized equipment and/or laboratory space (van der Sluis et al., 2011), and practical challenges, since researchers with expertise in operating said equipment are required during the experiment and experience with interpreting results is required afterwards (Mauss & Robinson, 2009). To offer a specific example: in the case of measuring skin conductance researchers must account for a small period of time in which values must “reset” to nominal values, failing to do so may result in the measurements presenting irregularities or introducing “noise” in the data which will make it harder to extract relevant information (Bakker et al., 2021). Neural activity such as Electroencephalography or fMRI scans also present these issues, in addition to requiring even more sensitive and costly equipment to perform.

Lastly then it is possible to film or observe a subject’s body language, including facial expressions, tics, and movement of the head, to attempt and interpret an emotional response behind it.(Scherer, 2005) While this is sometimes less costly than previously presented methods it is still subject to a lot of subjective interpretation, and may offer wildly varying results between subjects, making it more difficult to interpret the data.

It should be noted that physiological measurements can be taken in tandem with, for example, self-reports, to generate a richer dataset (van der Sluis et al., 2011). Doing so could arguably supplement the more abstract and subjective data with more objective physical measurements. Other researchers have had some measure of success in applying this theory (Bakker et al., 2021) although they note that the two types of measurement do not converge or map onto each other, as other research similarly affirms (Mauss & Robinson, 2009). While this doesn’t mean the technique is therefore useless it does not provide a way to simply reaffirm the results of one test using the other, instead providing an additional datapoint which can be instructive but must be interpreted with care (Bakker et al., 2021).

The existing body of work therefore suggests that self-reporting, if done properly, remains one of the more feasible methods. Though it need not be the only instrument of choice. While concerns about self-reported data persist, and must be accounted for, there is enough evidence to suggest it remains a valid choice in research (Chan, 2008). However,

many self-reporting techniques previously mentioned tend to take the shape of a simple questionnaire, usually asking participants to rate a text or their emotions on a Likert scale or similar. Such work has since been expanded upon to offer newer, more intuitive self-reporting tools.

The PANAS (Positive-Negative Affect) method aims at capturing emotions on a continuous scale between two dimensions: positive vs negative affect. This is generally accomplished by choosing 10 emotional terms which varying belong to each affect, then asking participants to rate whether these terms correspond to their current emotional state, generally on a 5- or 7-point Likert scale. In doing so the PANAS method appears to draw some inspiration from Russell's circumplex model of emotion (Watson et al., 1988). This provides an easy-to-use method which generates results in a standardized format, making it easier to reproduce tests and/or compare the resulting data against other datasets. Further research proved that this method remained consistent and reliable, even outside of clinical contexts, but noted that it appears as though the two dimensions are at least partially interdependent, meaning results will trend towards similar patterns (Crawford & Henry, 2004). While this doesn't necessarily render the method unusable or unreliable it must be considered when interpreting the data. However, PANAS does offer many questions to subjects. While the choices of terms are backed by scientific literature it can arguably be somewhat overwhelming for subject, requiring them to use additional cognitive resources. This could perhaps influence their response as previously discussed. Moreover, this could potentially increase the time taken to respond, which could in turn introduce a temporal bias, affecting the reported emotions. Other methods should then, perhaps, focus on creating an even more intuitive, easy-to-use interface which allows for rapid response times without sacrificing the more grounded approach of PANAS.

3.2 Alternative Methods

In researching different means of evaluating effect on emotions several alternative methods can be proposed. For the sake of brevity, we will only elaborate on those which were potential candidates for our own research.

Emotion Wheels have been independently proposed in several instances, often based on similar ideas and prior research. One prominent example is the Geneva Emotion Wheel. A self-reporting method for emotion which lays out between 12 and 20 emotional terms in a circular wheel formation (hence the name), where each emotional term can also be rated on its perceived intensity by a series of circles moving inward towards the wheel's center, decreasing in size relative to the emotion's intensity. At the center one or two "neutral" options are also available.

The wheel can be divided into four quadrants slightly resembling the Valence-Arousal circumplex model developed by Russel. The circumplex model employs the term "Arousal" to denote emotions which are higher or lower in "activity" (for example anger vs. sadness), this can lead to complications when similar emotions may present varying degrees of arousal. Furthermore, it is not always clear whether the "arousal" or "activation" indicates a type of sympathetic response correlated with an increase in neurological and/or nervous activity, or if it indicates an "activation" in a sense of motivation or desire to act. For this reason, the GEW opts instead for the notion of "control" to replace Arousal, as doing so more closely aligns with existing literature on the appraisal of emotional responses. Control posits a difference between action and inaction, but also a sense of conduciveness or

obstruction for the subject, preserving the “activation” measure of Arousal.(Scherer et al., 2013; Siegert et al., 2011)

Reproducing the circumplex model offers several advantages, for one it fits with existing literature on the subject, thus providing some continuity between previous work and current research. Additionally this means the data produced by the study can be more easily re-used and compared to other studies with minimal changes necessary, also making reproduction of a study far easier.(Scherer et al., 2013).

The GEW has gone through several iterations over the years, mostly related to choosing its emotional labels, finding a good balance between thorough inclusion of all possible sentiments and making sure not to overcomplicate the wheel, as more complicated systems such as the Self-Assessment Manikin can sometimes present a challenge for annotators who are not instructed on its use.(Siegert et al., 2011). Prior versions included up to 20 emotional labels, largely derived from previous literature, and although these models performed very well in practice(Siegert et al., 2011) they still tended towards excessive complexity at the cost of usability.(Scherer et al., 2013) The current third version of the wheel as it is presented by the University of Geneva has enhanced its choice of labels by using results from the GRID study.

The GRID was a newly composed instrument aimed at providing a more universal means of reporting emotions which would also be easier to translate to different languages(Fontaine et al., 2007). Grid-based systems were already a mainstay within the field and largely dominant for their simplicity and effectiveness(Russell et al., 1989; Scherer, 2005), GRID specifically sought to enhance these existing models and argued that more than two-dimensions were necessary to properly encapsulate all emotions. The model uses 24 emotion terms derived from both research literature and daily use, and 144 emotional features. It is employed as a questionnaire (generally web-based), showing participants one out of the 24 emotion terms at random and asking them to assign a likelihood score to each of the 144 emotional features on a 9-point Likert scale ranging from *extremely unlikely* (1) to *extremely likely* (9). The GRID provided more evidence of the validity of Valence and Control as viable axes along which to measure emotions. It offered a more rigorously tested standard set of emotional terms which could also be readily translated, making the new version of the GEW much easier to employ in other contexts. The GEW still only uses two dimensions, instead of the proposed 4, largely because two dimensions on the wheel are intuitively easier to understand when printed on paper or displayed on a screen, which is it’s intended use.(Scherer et al., 2013)

The GEW has enjoyed a good deal of success in its application, both for the reliability of its data and its ease of use for subjects and researchers alike. It provides an adaptable, reliable framework and results which can easily be reproduced or verified.(Coyne et al., 2020, 2020; Z. Li & Mao, 2012; Sacharin et al., 2012; Siegert et al., 2014; Turumugon et al., 2019). Due to these factors, we believe the GEW to be a solid choice of instrument to measure emotional responses and perform extrinsic evaluations on their effectiveness of CA-NLG systems.

4.Research Question

As we have discussed several methods for measuring emotions exist, with methods such as PANAS having a well-established history in clinical applications. Methods such as the GEW, however, are more recent and thus questions can be raised as to whether such methods can accurately report emotions, compared to existing methods such as PANAS. If we wish to perform more rich evaluation of CA-NLG system, could we potentially look to these methods to help measure whether the system induces the desired emotional response? Suppose we can, which should we choose?

To perform valid evaluations such tools must be:

1. Consistent. Results should be reproducible.(Crawford & Henry, 2004)
2. Fast. Subjects should not have to take too much time to report their emotions, or otherwise risk introducing temporal biases into the responses.(Mauss & Robinson, 2009; Mitchell et al., 1997)

As such we can raise the following question: does the Geneva Emotion Wheel outperform PANAS in terms of consistency and/or speed of response? As the GEW was designed to be intuitive and all-encompassing we could hypothesize that it would perform better in application, an experiment could be designed to test this hypothesis. Specifically, we would need to measure the consistency of responses for the GEW and PANAS, as well as the time taken to complete the task by subjects. If the GEW returns faster, more consistent result it could prove a valuable tool for measuring emotions in the context of evaluating CA-NLG applications. If not, it would suggest using a more tried method such as PANAS could be a better option.

5. Experimental Setup

To judge the viability of the GEW as a tool for measuring emotions in testing CA-NLG a metric or test needed to be chosen. The chosen metric would be test-retest reliability, as this is a commonly accepted means of testing a method's viability. To do this we designed a questionnaire which would feature several texts with a known affective or emotional charge and ask subjects to report their own emotional states while answering it, subjects would then later receive the same questionnaire, featuring the same texts, again and be asked to perform the same task again. Results of these two rounds of questionnaires would be used as a basis to determine the test-re-test reliability. To better determine the GEW's validity we decided to design a double, between subjects, experiment featuring both the GEW and the PANAS system. As PANAS is a proven, mature, and widely used method it seemed this would be a useful comparison to make.

Initial designs had hoped to use texts generated by CA-NLG systems such as Affect-LM or Emotional Chatting Machine. However due to technical difficulties setting up these systems it was decided this would likely take too much time, thus a corpus of emotionally annotated text would be used instead. The chosen corpus was AffectiveText, an English-language corpus composed of several headlines gathered from online news outlets between 2011 and 2015, annotated with five emotions, their respective intensities, as well as an insensitive of valence.(Bostan & Klinger, 2018) This corpus was chosen as it likely closely mimics the desired situation of having subjects read generated texts. 30 headlines total were chosen from the corpus, three for each annotated emotion, as well as 3 "neutral" headlines which did not have any strong emotional values (below 25 out of 100). 15 would be used for each answering method, meaning the PANAS and GEW questionnaires would not feature the same headlines.

As per the designs of PANAS and GEW each question would effectively ask for 20 to 22 inputs: one Likert scale value for each emotion. These would be substantially different between the two systems though, as PANAS' scale asks the subject to fill out every emotion on a scale of 1 (does not describe my feelings at all) to 5 (perfectly describes my feelings). Whereas the GEW permits participants to not report certain emotions, leaving their value at 0. The GEW permits participants were additionally able to report "no emotion", or "other" and describe their state using a text input. In the event "no emotion" was chosen all other values would be considered as 0, and in the even all values were 0 "no emotion would be assumed.

Additionally, the first round of questionnaires contained several demographic questions including: age range, gender identity, whether the mother tongue of the respondent was English or not. All included an option not to offer this information as a matter of privacy. These questions were not included in the follow-up questionnaires.

Respondents were also asked to include a valid e-mail address to which the second round of questionnaires would be sent out, seeing as the first round was distributed via anonymous hyperlink. Responses to the first and second round from participants were identified and linked using a unique identifier assigned in the first round. This identifier would then be included in the metadata of each participant's follow-up questionnaire. Doing so allowed us to anonymize all questionnaires after the collection period, at which point participants could no longer opt-out or have their data removed.

6. Results

In the first round 90 surveys were sent out. 52 were completed, with the split slightly favoring responses using the PANAS. Three weeks later a second round was sent out, this time responses rate was about 36 people out of 52. 21 people for the Geneva Emotion Wheel, 15 for PANAS. It is therefore important to note that the overall sample sizes of the experiment ended up very small, which will heavily influence statistical analysis and our ability to draw any definitive conclusions. It may already be wise to warn that any conclusions or inferences draw from these results cannot be argued to be generalizable to a larger population without further research.

Demographics of respondents were somewhat skewed towards the 25-34 age range, which represented over half of all respondents, and gathering more male than female participants. Non-native English speakers represented the majority of respondents, with more than half indicating English not being their mother tongue

English as Mother Tongue	**GEW**	**PANAS**	**Combined**
Yes	5	5	10
No	10	16	26
Prefer not to say	0	0	0

Table 6-1 Demographic data concern English skills

While the PANAS questionnaire required all participants to fill out all emotions, the GEW allowed participants to leave as many unanswered as desired. Because of this the GEW returned many empty values which were replaced with values of 0. In the case where GEW participants had entered text into the “other” field the text was preserved extracted and a value of 1 was given for the “other” field in the results table. Most entered text mostly mentioned not understanding the headline or declared a lack of interest in the subject. Several participants contacted us privately afterwards to ask about the lack of certain emotion terms in the Wheel but did not indicate this absence in the actual questionnaire.

Data from participants who filled out the first questionnaire but not the second was not used, as this would be not allow us to assess test-retest reliability and would further complicate statistical analysis. A full table of results and anonymized demographic data is available in the appendix.

Gender Identity	**GEW**	**PANAS**	**Combined**
Female	7	7	14
Male	7	12	19
Non-binary/third gender	1	2	3
Total	15	21	36

Table 6-2 Demographic data concerning gender identity

7. Statistical Analysis

Due to the small sample size the statistical analysis encountered several complications, primarily a lack of normal distribution in the data and an outsized effect of outliers, as we’ll demonstrate. For this reason, it does not seem prudent to draw definitive conclusions and

any statistical significance is presented mostly as a potential avenue of interest for future research.

First, we used Cronbach's Alpha to measure the internal consistency of the questionnaire items.(Bujang et al., 2018; Cronbach, 1951). This alpha was designed to test reliability of surveys using Likert scales by analyzing the variance and covariance of responses and find if the questions correlate, though it should be noted that this says nothing about the validity of the results. Results are visible in table 3.

Since most of the resulting values are within perfectly acceptable ranges (as a score higher than 0.7 is generally agreed to represent good consistency(Bujang et al., 2018; Cronbach, 1951)) we decided to compare them using the **cocron** package(Diedenhofen & Musch, 2016). This method determines if the difference between two alphas is significant or simply random. Interestingly, while no significant differences were found between the first and second GEW questionnaire (P = 0.3669) nor the first GEW and PANAS questionnaire (P = 0.1095), a significant difference was found for the second PANAS questionnaire compared to both the preceding PANAS questionnaire and the second GEW questionnaire. Since the content of both rounds of questionnaires was the same this would suggest participants in the second reported slightly varied the scores for their previously reported emotions. It is unknown, however, what would cause such variance.

Survey Round	GEW	PANAS	Cocron P-Value
First	0.9277	0.9682	0.1095
Second	0.8964	0.9878	0.0002
Cocron P-Value	0.3669	0.0392	

Table 7-1 Cronbach's Alpha scores and Cocron calculated P-values

7.1 Completion Times

One question of interest was the time it took the average respondent to complete the questionnaire, seeing as time and accompanying cognition are liable to introduce biases into the self-evaluation of a respondent's emotional state. Ideally, the self-reporting tool should be used as quickly as possible to avoid biasing responses as discussed before. However, the current design of the surveys did not provide a means to measure the time taken on each individual question, but only of the overall time spent on the entire survey, which we will present here. As can be seen on tables X and graph X the average response time for the GEW was much higher, roughly between 28 and 33 minutes (1700 and 2000 seconds), with an extreme degree of variability, with one outlier up to 57 hours (207610 seconds).

Filtering out outliers of over 2.7 hours (10000 seconds) yielded much more comparable results. While the unfiltered results and tests are still presented in the appendix we decided to continue testing on the filtered results. While we believe the outliers are liable to skew the data too much to obtain reliable results, this does unfortunately cut the sample size further down to 31 subjects in total.

As can be seen in Figure 1 and Table 4, results appear to slightly favor PANAS in terms of completion speed, at an approximate mean of 16 minutes (978,6842 seconds) against an

approximate mean of 18 minutes (1121,25 seconds) for the GEW. Notable, though not unexpected, is the difference in mean completion times between the first and second

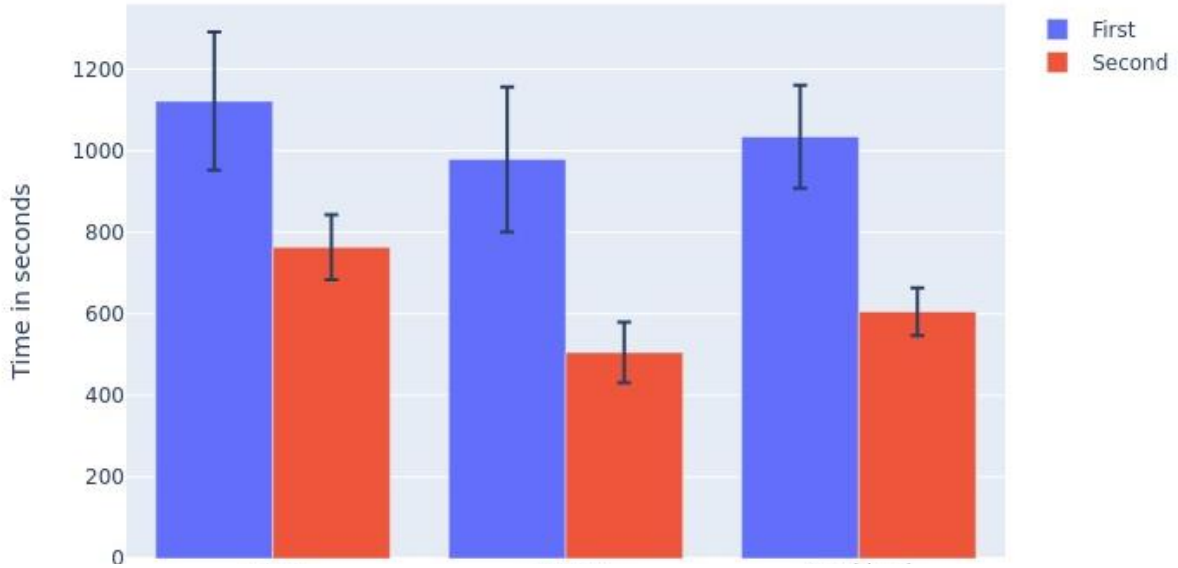

Figure 7-1 Means of time taken to complete the questionnaire. Outliers filtered out

round of surveys. As such we decided to test whether these differences were significant.

Survey Round	**GEW**	**PANAS**	**Combined**
First	1121,25	978,6842	1033,87097
Second	763,3333	505,2105	605,129032

Table 7-2 Mean time taken to complete the questionnaire. Outliers filtered out

We tested the combined duration datasets using the Shapiro Wilk test with the null Hypothesis that both the collection completion times followed a normal distribution. In both cases the test returned a $P > 0.05$, rejecting the null hypothesis and allowing us to infer that the data, even when filtered for outliers, is not normally distributed. (A table containing the full results of both tests can be found in the appendix)

Because of this we chose the two-tailed Mann Whitney U test to compare the two groups of completion times for the first and second rounds of questionnaires. This is because the Mann Whitney U does not assume normal distribution and is robust even with small and unequal sample sizes(Fay & Proschan, 2010). The test returned a statistic of 669.0, and a P-value=0.00812 meaning we can reject the null hypothesis that the distribution of completion times for the first and second questionnaires is the same.

Returning to our initial question concerning the completion times of the GEW against those of PANAS we decided to perform the Mann Whitney U test again, this time however using a one-tailed variant, inferring from the differences in mean completion time between answering we posit the hypothesis that the distribution underlying the GEW completion times is greater than the one underlying PANAS. Additionally, since the Mann Whitney U test cannot indicate the strength of a possible correlation, we also calculated a Point Biserial correlation coefficient to determine how strongly completion times correlate to a difference in answering method. Where the answering method (GEW or PANAS) is used as a Boolean independent variable and the completion time as a continuous, dependent variable.

Survey Round	MW Statistic	MW P-value	PB Correlation	PB P-value
First	141	0,141226482	0,100660474	0,590028944
Second	177	0,005622924	0,390613391	0,029809345

Table 7-3 Results from the calculation of a one-tailed Mann Whitney U test and the Point Biserial correlation on the relation between completion times and answering methods, for both rounds of questionnaires. (Outlying completion time values filtered out)

As can be seen in table 5, in the first round of questionnaires no significant differences or correlations were found. Both the Mann Whitney U and Point-Biserial tests returned P > 0.05 and retain the null hypothesis positing no significant difference or correlation. However, in the second round of questionnaires a difference can be found: with a Point Biserial correlation coefficient of 0,39 it appears there is a weakly positive correlation between the choice of answering method and completion times, with the Mann Whitney U test supporting the hypothesis that it the distribution underlying the completion times for PANAS is smaller than those of the GEW.

We further tested other demographic categories for possible correlations with completion times, using the Kruskal Wallis H test. This is since these remaining categories comprised 3 or more labels, making the Mann Whitney test unsuitable. Furthermore, unlike other tests such ANOVA, the Kruskal Wallis test does not assume normal distribution. However, as all P-values exceeded the Alpha of 0.05 no significant results were obtained.

Survey Round	Age Range	English as Mother Tongue	Gender Identity
First	0,0711497	0,66338967	0,153998967
Second	0,23331782	0,930633674	0,533728892

Table 7-4 P-values of the Kruskal Wallis H test comparing completion times against several other demographic categories for each round of questionnaires.

7.2 Test re-test correlations

Determining a suitable method to measure the test-retest reliability proved somewhat challenging. Pearson correlation coefficients could be calculated between responses from the first and second questionnaire, under the assumption that if the responses were consistent across the test and re-test the scores would show a positive correlation. However, two approaches were possible.

The first approach would compare the answers of the first questionnaire to those of the second on a per-participant basis, arguably measuring the consistency of the participants. The second approach would be to compare on a per-question basis, or to be exact to compare the values of each reported emotion for each question, which would arguably measure the consistency of answers. The first method would retain the original sample sizes of 21 and 15 participants with 396 and 360 datapoints respectively, whereas the second would provide a sample size of 396 and 360 items with 21 and 15 datapoints each. The second method was chosen, in the hopes that this would alleviate some of the problems inherent to the small participant pool but generated a different complication altogether: since the Likert scale answers were now pooled per emotion, per question there would be cases in which all participants provided the same answer for one emotion.

This vector of answers would then effectively have a standard deviation of zero, meaning the Pearson correlation formula would then divide over zero, making it impossible to return a result. To address this a slightly modified calculation was used, under which any vector of answers which had a standard deviation of zero was rejected unless it was a perfect match with the second vector. In the latter case a perfect correlation was assumed and would be returned as 1.0, as would be the case with a standard Pearson correlation calculation of identical vectors. Using this variant 90 out of the 396 (approx. 23%) entries for the GEW sample had to be discarded against 63 out 360 (approx. 18%) for PANAS. No entries were discarded when computing correlations on a per-participant basis.

Resulting from these two operations two samples of Pearson correlation coefficients were obtained. In order to compare the two samples for significant differences their values were transformed using the Fisher-Z transformation(Diedenhofen & Musch, 2015; Seifert, 2020; Silver & Dunlap, 1987). Two Shapiro Wilk tests were performed to determine if both samples' data was normally distributed, both revealed this data to be extremely non-normal, obtaining P-values of 8.88189835619759e-19 for the GEW sample and 1.9861372807517664e-16 for the PANAS sample. Figure 7-2 represents a visualization of this non-normality: on the x-axis are 'bins' of Fisher-Z transformed Pearson correlation coefficients, separated in intervals of 0.1. The y axis represents the number times a score fits inside a particular `bin`.

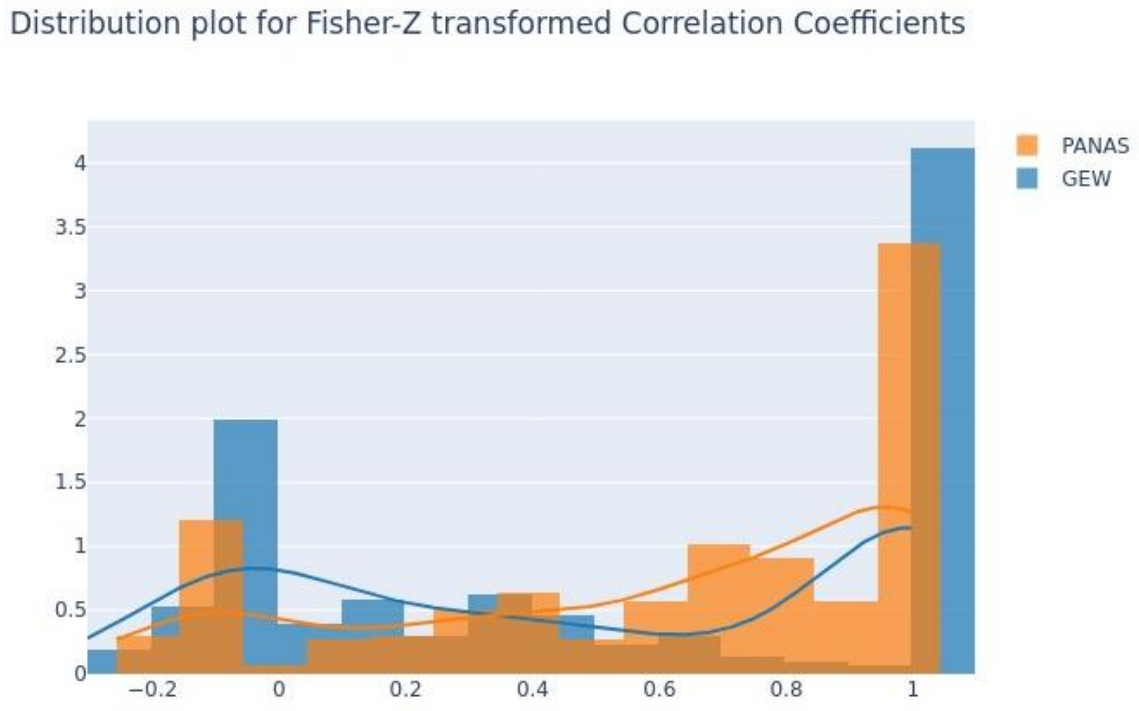

Figure 7-2 Histogram and Curve of distributions of Fisher-Z transformed Correlation scores. Range of scores on the x-axis, number of instances of that score on the y-axis

Notably the scores ranged broadly from -0.2 to 1. Such negative scores indicate a very weak negative correlation, whereas 1 indicates a perfect correlation. The high number of perfect correlations is likely a by-product of our earlier filtering method, which assigns perfect correlation scores to comparisons between two sets of answers, both the same singular value (for example all zeroes).

The original intent had been to perform a two-tailed Welch T-test to compare the two samples, however such a test assumes a normal distribution of the mean of the two samples. Accordingly, a Mann Whitney U test was once again used, since it has no such assumptions, which returned a statistic of 43591.0 and a P-value of 0.3778, retaining the null hypothesis that there is no significant difference in the underlying distributions. However, since this does not tell us anything about the means of the samples, another test was performed using a method to compare correlations(Sheskin, 2011) similar to the Cocor package(Diedenhofen & Musch, 2015) which calculates the probability of a significant difference between two correlation coefficients using the following formula.

$$z = \frac{z_{r_1} - z_{r_2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Where Z_{r1} and Z_{r2} represent the Fisher-Z transformed correlation coefficients of the first and second set of answers, and n_1 and n_2 their respective sample sizes.

To perform this calculation two weighted averages of correlations were calculated, along with their standard errors and deviations. Provided in table 7-5 and visualized in figure 7-3.

Measure	**GEW**	**PANAS**
Weighted Average of Correlations (Fisher-Z transformed)	0,503266	0,607635
Standard Deviation	0,4707	0,40464
Standard Error	0,026908	0,02348
Weighted Average (back transformed)	0,464682	0,54246

Table 7-5 Data concerning averaged Pearsson correlation coefficients

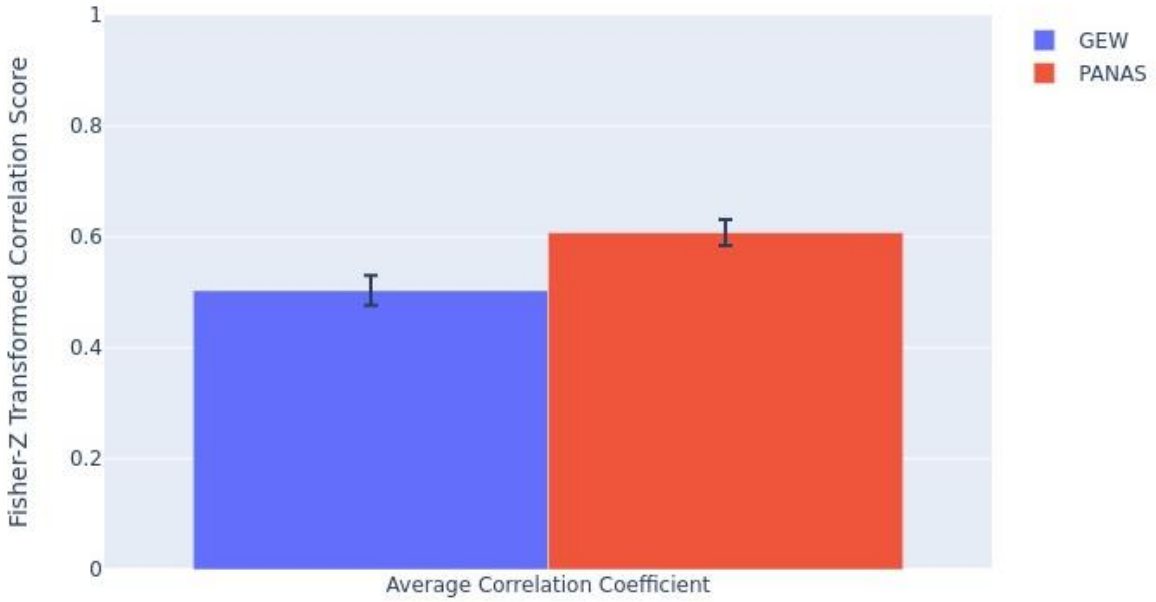

Figure 7-3 Averages of Fisher-Z transformed correlation scores for each method

Testing the Fisher-Z transformed averages with a two-tailed test yielded a t-statistic of -1.2749 and a p-value of 0.2023, retaining the null hypothesis that there is no significant difference between the two averaged correlations.

The same methods were applied, also, to the data using per-participant correlations, although results were different, they provided similar observations. These results can be found in the annex for comparison.

8. Discussion

While all values for Cronbach's Alpha are well within satisfying bounds, it is nonetheless interesting to see the second PANAS questionnaire perform significantly different, even outperforming the first round. Since no changes were implemented in the interim, as per design, it is difficult to pin down a likely cause. One possibility is that increased familiarity with the task led to a better performance for the second round, but we do not have a means to verify this as of now. Nonetheless it is advisable to include comparisons of alphas in future research as it may yet reveal interesting effects.

Considering the difference in mean completion time and the results of the Mann Whitney U test, it is possible respondents finished the second questionnaires significantly faster. This may be due to familiarity with the provided headlines, or perhaps recalling previous answers. This could suggest a longer waiting period is desirable in re-testing.

While we hypothesized that the Geneva Emotion Wheel would register faster completion times, due to its layout and lack of requirement in filling out every emotion, it appears the opposite occurred. We cannot conclude that this is inherent to the usage of the GEW itself, as it might be the case that the selected headlines presented in the GEW questionnaires required more time to read and understand. Unfortunately, we lack a control group to confirm this. Future research may perhaps consider using the same texts for all answering methods to prevent such possible confounds.

Unfortunately, the current results did not permit us to establish a significant difference in consistency between the GEW and PANAS. Both provided averaged correlation coefficients suggesting reasonable consistency, being near 0.5, suggesting that both methods at least can perform satisfactorily. The provided framework of statistical tests may be useful in future research to establish a similar comparison for other methods of self-reporting emotions. For now, more research is perhaps needed to further establish the usefulness of the GEW and determine whether it can find suitable application in measuring the effects of Affective NLG systems.

9. Conclusion

In this work we explored the field of Affective Natural Language Generation (CA-NLG), and the research done into creating systems which could control the affect of the generated output. We discussed methods of evaluating such systems and argued there is too much focus on evaluating systems intrinsically, using methods such as BLUE or expert human judging. We proposed that this bias towards intrinsic evaluation misses critical information provided by extrinsic evaluation, such as the effectiveness of a system in accomplishing it's intended task.

This situation is not a product of coincidence, as extrinsic evaluations are complicated and expensive to perform(Gkatzia & Mahamood, 2015; van der Sluis et al., 2011). Extrinsic evaluations require precise designs to effectively measure what a researcher intends to measure and are very prone to confounds and biases on the part of subjects, evaluations of subjects' emotions even more so.

Measuring emotions presents several complications which need to be accounted for. In the case of self-reporting methods these primarily revolve around biases introduced in subjects' evaluation of their own emotional states by the physical environment, physiology, and temporal displacement from the moment of experience of an emotion (Mauss & Robinson, 2009; "Temporal Adjustments in the Evaluation of Events," 1997). In terms of physical, or physiological, measurement it is difficult to point to discrete, separate emotions (Fontaine et al., 2007), as subjects may experience several at a time. Furthermore, the underlying physical processes are not always fully understood, and in some cases previous assumptions about things like activation of certain brain regions prove to not be exclusive to one emotion (Mauss & Robinson, 2009).

We touched on the issues present in measuring emotions for the purposes of evaluating the effectives of NLG systems intended to provoke, or avoid provoking, a certain emotional reaction with their output. We discussed the limitations of existing systems such as PANAS and proposed further research into new alternatives such as the Geneva Emotion Wheel. To this end we designed an experiment aimed at validating the GEW and comparing it to PANAS based on their respective test-retest reliability. We encountered difficulties in obtaining a suitable sample size and this, unfortunately, limits our ability to make conclusive arguments. We found that both GEW and PANAS performed similarly in terms of test-retest reliability, with seemingly no significant difference in how answers form the first round correlated to answers in the second round. We found that subjects using PANAS completed their questionnaires significantly faster than those using the GEW, however we could not determine this on a per-question basis, nor could we exclude that this was not a by-product of choosing different texts for each method. A such we concluded more research including a larger sample size, and using the same texts for both methods, would be needed to further validate these results.

As the GEW comes very close to PANAS in terms of overall performance, we could suppose it is therefore a viable tool to be used in measuring emotions as a part of CA-NLG evaluation. While some technical difficulties were encountered in implementation of a digital version the GEW these could be more readily solved in future research, for example by reusing code or by using a physical, paper version instead. Additionally, since we can

provide access to the Python code used for the statistical analysis the experiment could be easily recreated.

More research is required to further test the validity of the GEW, as we believe our work here provides a good template to work from. Mostly we believe further research will require a larger sample size, should measure the response time for each question, and should not use different emotional utterances between the two methods.

Furthermore, we believe it still stands to reason that more attention should be given to extrinsic evaluation and ecological validity. Not just in the field of CA-NLG, but perhaps in the field of NLG as a whole and perhaps even the broader field of Artificial Intelligence. Such evaluations can still offer us key insights in how systems perform in real-world situations, allowing us to better understand the strengths and shortcoming of different approaches. Further validation of methods such as the GEW can provide us with better understanding of how to accomplish this, as we can design evaluations to overcome the inherent complications of self-reporting tools, such as temporal biases. We believe that we have successfully demonstrated a viable method of testing these methods in future research and point towards several problems that would need to be considered. Such as time taken on questions and internal reliability of questionnaires.

Future research could also focus on incorporating a working CA-NLG system to more closely mimic the actual intended real-world situations in which such systems would find application. While such an application was outside of the scope of this research for the time being, it would likely provide deeper insight into people's reactions to texts generated on the fly instead of prepared annotated texts.

Bibliography

Amidei, J., Piwek, P., & Willis, A. (2019). The use of rating and Likert scales in Natural Language Generation human evaluation tasks: A review and some recommendations. *Proceedings of the 12th International Conference on Natural Language Generation*, 397–402. https://doi.org/10.18653/v1/W19-8648

Bakker, B. N., Schumacher, G., & Rooduijn, M. (2021). Hot Politics? Affective Responses to Political Rhetoric. *American Political Science Review*, *115*(1), 150–164. https://doi.org/10.1017/S0003055420000519

Barrett, L. F., Robin, L., Pietromonaco, P. R., & Eyssell, K. M. (1998). Are Women the "More Emotional" Sex? Evidence From Emotional Experiences in Social Context. *Cognition and Emotion*, *12*(4), 555–578. https://doi.org/10.1080/026999398379565

Belz, A., & Reiter, E. (2006, April). Comparing Automatic and Human Evaluation of NLG Systems. *11th Conference of the European Chapter of the Association for Computational Linguistics*. EACL 2006, Trento, Italy. https://www.aclweb.org/anthology/E06-1040

Bostan, L.-A.-M., & Klinger, R. (2018). An Analysis of Annotated Corpora for Emotion Classification in Text. *Proceedings of the 27th International Conference on Computational Linguistics*, 2104–2119. https://www.aclweb.org/anthology/C18-1179

Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A Review on Sample Size Determination for Cronbach's Alpha Test: A Simple Guide for Researchers. *The Malaysian Journal of Medical Sciences: MJMS*, *25*(6), 85–99. https://doi.org/10.21315/mjms2018.25.6.9

Celikyilmaz, A., Clark, E., & Gao, J. (2021). Evaluation of Text Generation: A Survey. *ArXiv:2006.14799 [Cs]*. http://arxiv.org/abs/2006.14799

Chan, D. (2008). So Why Ask Me? Are Self-Report Data Really That Bad? In *Statistical and Methodological Myths and Urban Legends*. Routledge.

Coyne, A. K., Murtagh, A., & McGinn, C. (2020). Using the Geneva Emotion Wheel to Measure Perceived Affect in Human-Robot Interaction. *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 491–498. https://doi.org/10.1145/3319502.3374834

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *43*(3), 245–265. https://doi.org/10.1348/0144665031752934

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Diedenhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Diedenhofen, B., & Musch, J. (2016). cocron: A Web Interface and R Package for the Statistical Comparison of Cronbach's Alpha Coefficients. *International Journal of Internet Science*, *11*, 51–60.

Dörr, K. N., & Hollnbuchner, K. (2017). Ethical Challenges of Algorithmic Journalism. *Digital Journalism*, *5*(4), 404–419. https://doi.org/10.1080/21670811.2016.1167612

Fay, M. P., & Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys*, *4*, 1–39. https://doi.org/10.1214/09-SS051

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The World of Emotions is not Two-Dimensional. *Psychological Science*, *18*(12), 1050–1057. https://doi.org/10.1111/j.1467-9280.2007.02024.x

Gatt, A., & Krahmer, E. (2018). Survey of the State of the Art in Natural Language Generation: Core tasks, applications, and evaluation. *ArXiv:1703.09902 [Cs]*. http://arxiv.org/abs/1703.09902

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., & Scherer, S. (2017). Affect-LM: A Neural Language Model for Customizable Affective Text Generation. *ACL*. https://doi.org/10.18653/v1/P17-1059

Gkatzia, D., & Mahamood, S. (2015). A Snapshot of NLG Evaluation Practices 2005—2014. *ENLG*. https://doi.org/10.18653/v1/W15-4708

Haim, M., & Graefe, A. (2017). Automated News. *Digital Journalism*, *5*(8), 1044–1059. https://doi.org/10.1080/21670811.2017.1345643

Hinojosa, J. A., Moreno, E. M., & Ferré, P. (2020). Affective neurolinguistics: Towards a framework for reconciling language and emotion. *Language, Cognition and Neuroscience*, *35*(7), 813–839. https://doi.org/10.1080/23273798.2019.1620957

Huber, B., McDuff, D., Brockett, C., Galley, M., & Dolan, B. (2018). Emotional Dialogue Generation using Image-Grounded Language Models. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3173574.3173851

Kezar, L. (2018). Mixed Feelings: Natural Text Generation with Variable, Coexistent Affective Categories. *Proceedings of ACL 2018, Student Research Workshop*, 141–145. https://doi.org/10.18653/v1/P18-3020

Lester, J. C. (2012). Expressive NLG for Next-Generation Learning Environments: Language, Affect, and Narrative. *Proceedings of the Seventh International Natural Language Generation Conference*, 2–2.

Li, J., & Sun, X. (2018). A Syntactically Constrained Bidirectional-Asynchronous Approach for Emotional Conversation Generation. *ArXiv:1806.07000 [Cs]*. http://arxiv.org/abs/1806.07000

Li, Z., & Mao, X. (2012). Emotional eye movement generation based on Geneva Emotion Wheel for virtual agents. *J. Vis. Lang. Comput.* https://doi.org/10.1016/j.jvlc.2012.06.001

Mahamood, S., & Reiter, E. (2011). Generating Affective Natural Language for Parents of Neonatal Infants. *Proceedings of the 13th European Workshop on Natural Language Generation*, 12–21. https://aclanthology.org/W11-2803

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion*, *23*(2), 209–237. https://doi.org/10.1080/02699930802204677

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J. H., & Khudanpur, S. (2010). Recurrent neural network-based language model. *INTERSPEECH*.

Mitchell, T. R., Thompson, L., Peterson, E., & Cronk, R. (1997). Temporal Adjustments in the Evaluation of Events: The "Rosy View." *Journal of Experimental Social Psychology*, *33*(4), 421–448. https://doi.org/10.1006/jesp.1997.1333

Otter, D. W., Medina, J. R., & Kalita, J. K. (2019). A Survey of the Usages of Deep Learning in Natural Language Processing. *ArXiv:1807.10854 [Cs]*. http://arxiv.org/abs/1807.10854

Portet, F., Reiter, E., Gatt, A., Hunter, J., Sripada, S., Freer, Y., & Sykes, C. (2009). Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, *173*(7), 789–816. https://doi.org/10.1016/j.artint.2008.12.002

Reiter, E. (2001). Building Natural-Language Generation Systems. *Computational Linguistics*. https://doi.org/10.1162/coli.2000.27.2.298

Reiter, E. (2018). A Structured Review of the Validity of BLEU. *Computational Linguistics*, *44*(3), 393–401. https://doi.org/10.1162/coli_a_00322

Reiter, E. (2007). An Architecture for Data-to-Text Systems. *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)*, 97–104. https://doi.org/10.3115/1610163.1610180

Reiter, E. (2011). Task-Based Evaluation of NLG Systems: Control vs Real-World Context. *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop*, 28–32. https://www.aclweb.org/anthology/W11-2704

Reiter, E., & Belz, A. (2009). An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, *35*(4), 529–558. https://doi.org/10.1162/coli.2009.35.4.35405

Reiter, E., & Dale, R. (1997). Building applied natural language generation systems. *Natural Language Engineering*, *3*(1), 57–87. https://doi.org/10.1017/S1351324997001502

Reiter, E., Gatt, A., Portet, F., & Meulen, M. van der. (2008). The Importance of Narrative and Other Lessons from an Evaluation of an NLG System that Summarises Clinical Data. *INLG*. https://doi.org/10.3115/1708322.1708349

Reiter, E., Robertson, R., & Osman, L. M. (2003). Lessons from a failure: Generating tailored smoking cessation letters. *Artificial Intelligence*, *144*(1), 41–58. https://doi.org/10.1016/S0004-3702(02)00370-3

Rieser, V., & Lemon, O. (2010). Natural Language Generation as Planning under Uncertainty for Spoken Dialogue Systems. In E. Krahmer & M. Theune (Eds.), *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation* (pp. 105–120). Springer. https://doi.org/10.1007/978-3-642-15573-4_6

Robinson, M. D., & Clore, G. L. (2002). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology*, *83*(1), 198–215. https://doi.org/10.1037/0022-3514.83.1.198

Russell, J. A., Weiss, A., & Mendelsohn, G. A. (1989). Affect Grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, *57*(3), 493–502. https://doi.org/10.1037/0022-3514.57.3.493

Sacharin, V., Schlegel, K., & Scherer, K. R. (2012). *Geneva Emotion Wheel Rating Study*. https://archive-ouverte.unige.ch/unige:97849

Santhanam, S., & Shaikh, S. (2019). Emotional Neural Language Generation Grounded in Situational Contexts. *ArXiv:1911.11161 [Cs]*. http://arxiv.org/abs/1911.11161

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, *44*(4), 695–729. https://doi.org/10.1177/0539018405058216

Scherer, K. R., Shuman, V., Fontaine, J., & Soriano Salinas, C. (2013). The GRID meets the wheel: Assessing emotional feeling via self-report. In *Components of emotional meaning: A sourcebook*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199592746.003.0019

Seifert, J. (2020, April 5). Averaging Correlations—Part I. *Medium*. https://medium.com/@jan.seifert/averaging-correlations-part-i-3adab6995042

Sheskin, D. J. (2011). Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition. In *Handbook of Parametric and Nonparametric Statistical Procedures, Fifth Edition* (5th ed., pp. 1250–1364). Chapman and Hall/CRC.

Siddharthan, A., Ponnamperuma, K., Mellish, C., Zeng, C., Heptinstall, D., Robinson, A., Benn, S., & Van Der Wal, R. (2019). Blogging birds: Telling informative stories about the lives of birds from telemetric data. *Communications of the ACM*, *62*(3), 68–77. https://doi.org/10.1145/3231588

Siegert, I., Bock, R., Vlasenko, B., Philippou-Hubner, D., & Wendemuth, A. (2011). Appropriate emotional labelling of non-acted speech using basic emotions, geneva emotion wheel and self-assessment manikins. *2011 IEEE International Conference on Multimedia and Expo*, 1–6. https://doi.org/10.1109/ICME.2011.6011929

Siegert, I., Böck, R., & Wendemuth, A. (2014). Inter-rater reliability for emotion annotation in human–computer interaction: Comparison and methodological improvements. *Journal on Multimodal User Interfaces*, *8*(1), 17–28. https://doi.org/10.1007/s12193-013-0129-9

Silver, N. C., & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, *72*(1), 146–148. https://doi.org/10.1037/0021-9010.72.1.146

Smiley, C., Schilder, F., Plachouras, V., & Leidner, J. L. (2017). Say the Right Thing Right: Ethics Issues in Natural Language Generation Systems. *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 103–108. https://doi.org/10.18653/v1/W17-1613

Strapparava, C., & Mihalcea, R. (2007). SemEval-2007 Task 14: Affective Text. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, 70–74. https://aclanthology.org/S07-1013

Temporal Adjustments in the Evaluation of Events: The "Rosy View." (1997). *Journal of Experimental Social Psychology*, *33*(4), 421–448. https://doi.org/10.1006/jesp.1997.1333

Turumugon, P., Baharum, A., Nazlan, N. H., Noh, N. A. M., Noor, N. A. M., & Rahim, E. A. (2019). Users' emotional evaluation towards kansei-based higher learning institution website using geneva emotion wheel. *Indonesian Journal of Electrical Engineering and Computer Science*, *16*(3), 1547–1554. https://doi.org/10.11591/ijeecs.v16.i3.pp1547-1554

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. https://doi.org/10.1126/science.7455683

van Deemter, K. J. (2016). *Computational Models of Referring: A Study in Cognitive Science*. The MIT Press. https://aura.abdn.ac.uk/handle/2164/8956

van der Lee, C., Gatt, A., van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. *Proceedings of*

the 12th International Conference on Natural Language Generation, 355–368. https://doi.org/10.18653/v1/W19-8643

van der Sluis, I., & Mellish, C. (2010). Towards Empirical Evaluation of Affective Tactical NLG. In E. Krahmer & M. Theune (Eds.), *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation* (pp. 242–263). Springer. https://doi.org/10.1007/978-3-642-15573-4_13

van der Sluis, I., Mellish, C., & Doherty, G. J. (2011). Affective Text: Generation Strategies and Emotion Measurement Issues. *FLAIRS Conference*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037//0022-3514.54.6.1063

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., & Eger, S. (2019). MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. *ArXiv:1909.02622 [Cs]*. http://arxiv.org/abs/1909.02622

Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. *ArXiv:1704.01074 [Cs]*. http://arxiv.org/abs/1704.01074

Appendix

Age Range	GEW	PANAS	Combined
18-24	1	2	3
25-34	5	11	16
35-44	1	1	2
55-64	1	2	3
65-74	3	3	6
Prefer not to say	1	0	1
Total	12	19	31

Table 0-1 Age ranges of participants

English as Mother Tongue	GEW	PANAS	Combined
Yes	4	5	9
No	8	14	22
Prefer not to say	0	0	0

Table 0-2 DEMOGRAPHIC DATA CONCERN ENGLISH SKILLS

Gender Identity	GEW	PANAS	Combined
Female	6	6	12
Male	5	11	16
Non-binary/third gender	1	2	3
Total	12	19	31

Table 0-3 DEMOGRAPHIC DATA CONCERNING GENDER IDENTIY

Survey Round	GEW	PANAS	Cocron P-Value
First	0.9277	0.9682	0.1095
Second	0.8964	0.9878	0.0002
Cocron P-Value	0.3669	0.0392	

Table 0-4 Cronbach Alphas for each Questionnaire, and P-values of Cocron comparisons

Survey Round	GEW	PANAS	Combined
First	1121,25	978,6842	1033,87097
Second	763,3333	505,2105	605,129032

Table 0-5 Mean time taken to complete the questionnaire. Outliers filtered out

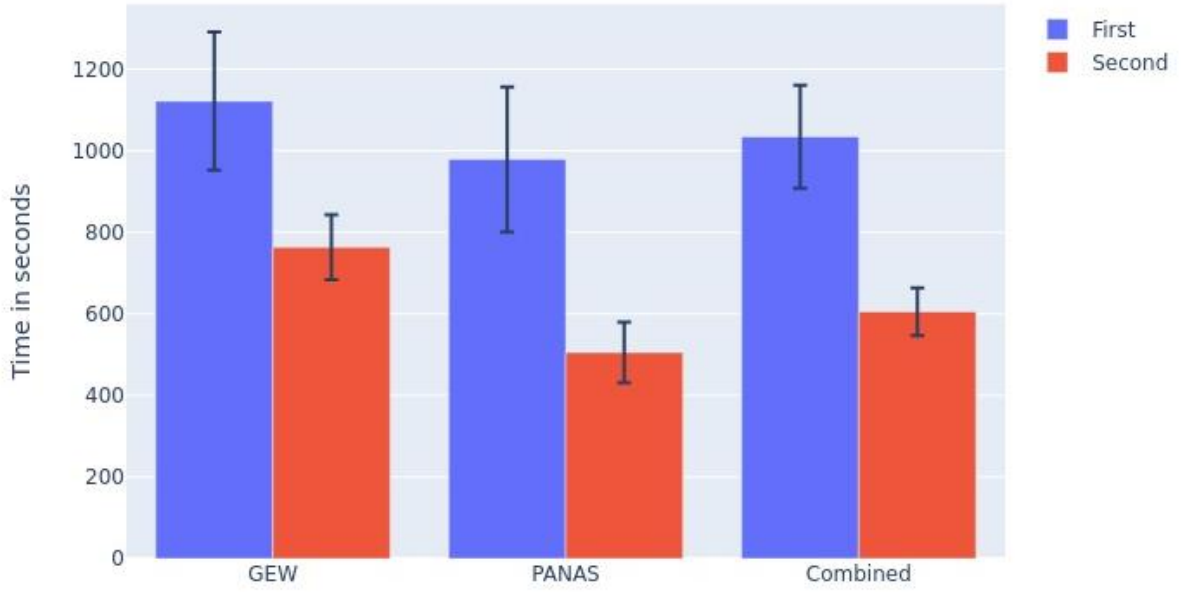

Figure 0-1 MEANS OF TIME TAKEN TO COMPLETE THE QUESTIONNAIRE. OUTLIERS FILTERED OUT

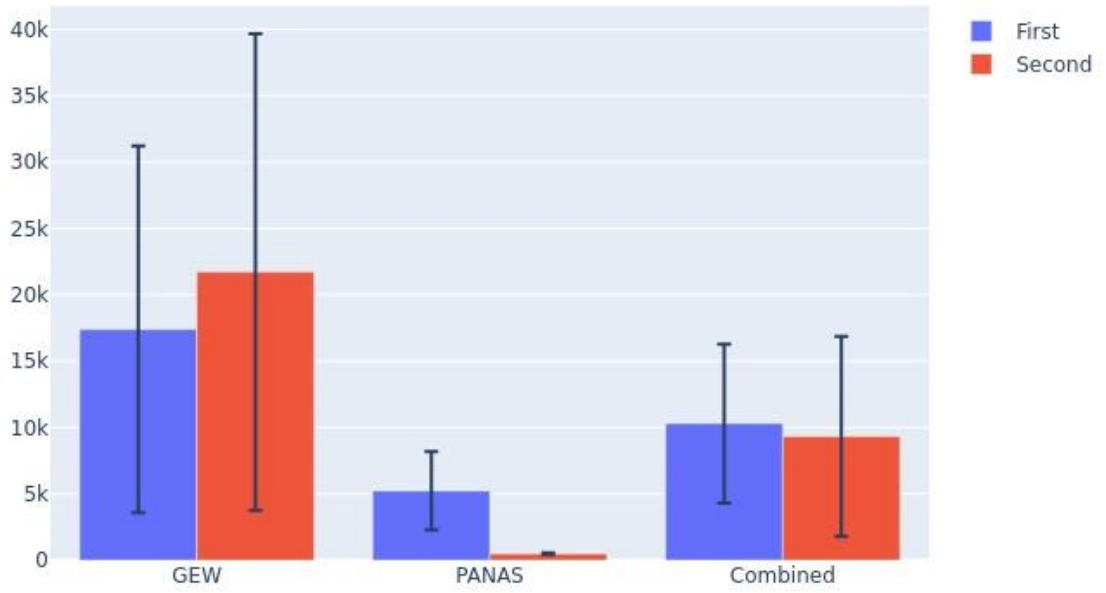

Figure 0-2 MEANS OF TIME TAKEN TO COMPLETE THE QUESTIONNAIRE. OUTLIERS RETAINED

Survey Round	MW Statistic	MW P-value	PB Correlation	PB P-value
First	141	0,141226482	0,100660474	0,590028944
Second	177	0,005622924	0,390613391	0,029809345

Table 0-6 Results from the calculation of a one-tailed Mann Whitney U test and the Point Biserial correlation on the relation between completion times and answering methods, for both rounds of questionnaires. (Outlying completion time values filtered out)

Survey Round	Age Range	English as Mother Tongue	Gender Identity
First	0,0711497	0,66338967	0,153998967
Second	0,23331782	0,930633674	0,533728892

Table 0-7 P-values of the Kruskal Wallis H test comparing completion times against several other demographic categories for each round of questionnaires. (Transposed Data)

Survey Round	Age Range	English as Mother Tongue	Gender Identity
First	0,204217	0,697624	0,35475
Second	0,355821	0,804718	0,549014

Table 0-8 P-VALUES OF THE KRUSKAL WALLIS H TEST COMPARING COMPLETION TIMES AGAINST SEVERAL OTHER DEMOGRAPHIC CATEGORIES FOR EACH ROUND OF QUESTIONNAIRES. (NON-TRANSPOSED DATA)

T statistic	df	p value 2 sided	Difference in mean	lb	ub
-2,9225329	592,3985	0,003604436	-0,104369118	-0,17451	-0,03423

Table 0-9 Values of a 2-sided T-test comparing correlations (transposed data)

T statistic	df	p value 2 sided	Difference in mean	lb	ub
-4,77616	28,59602	4,88E-05	-0,24241	-0,34628	-0,13854

Table 0-10 VALUES OF A 2-SIDED T-TEST COMPARING CORRELATIONS (Non-TRANSPOSED DATA)

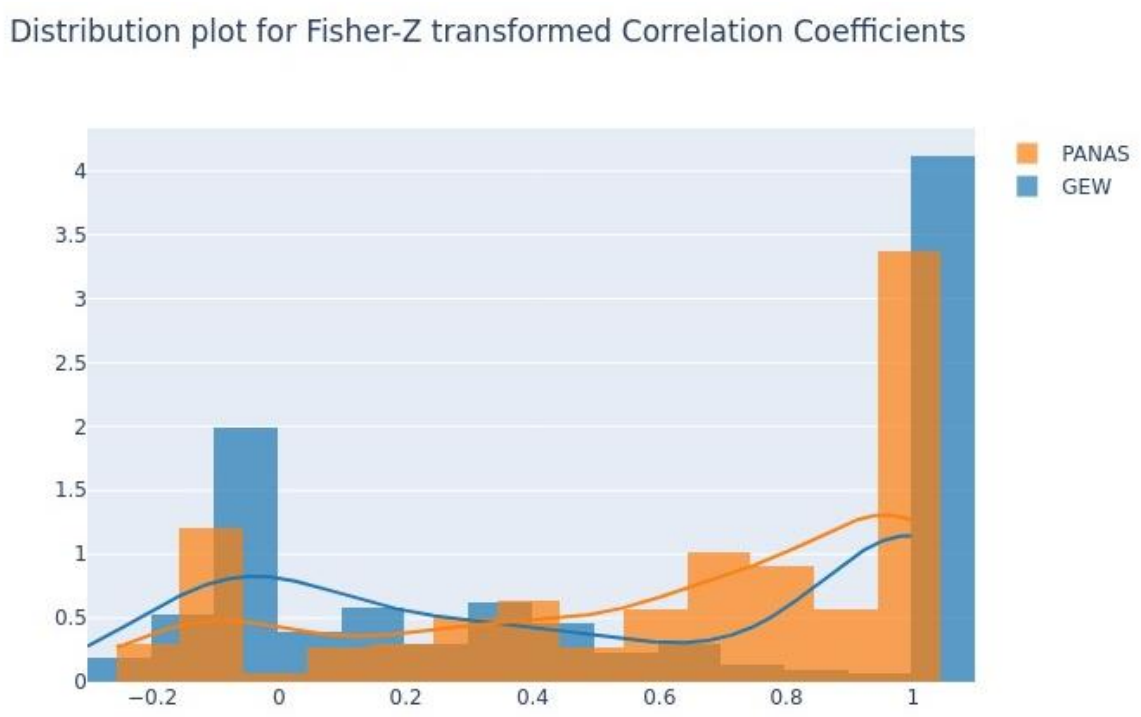

Figure 0-3 Histogram and Curve of distributions of Fisher-Z transformed Correlation scores. Range of scores on the x-axis, number of instances of that score on the y-axis

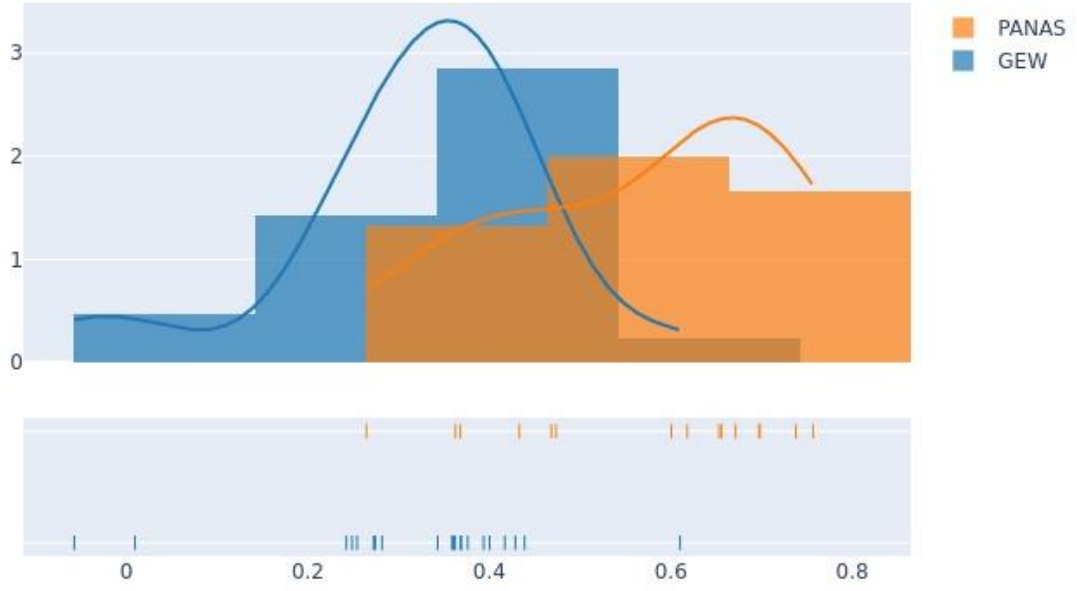

Figure 0-4 Same Histogram as Figure 12-3, but non-transposed data

Measure	**GEW**	**PANAS**
Weighted Average of Correlations (Fisher-Z transformed)	0,503266	0,607635

Standard Deviation	0,4707	0,40464
Standard Error	0,026908	0,02348
Weighted Average (back transformed)	0,464682	0,54246

Table 0-11 Data concerning averaged Pearsson correlation coefficients (Transposed data)

Measure	**GEW**	**PANAS**
Weighted Average of Correlations (Fisher-Z transformed)	0,321526	0,563934
Standard Deviation	0,142292	0,15549
Standard Error	0,031051	0,040147
Weighted Average (back transformed)	0,310886	0,510891

Table 0-12 DATA CONCERNING AVERAGED PEARSSON CORRELATION COEFFICIENTS (NON-TRANSPOSED DATA)

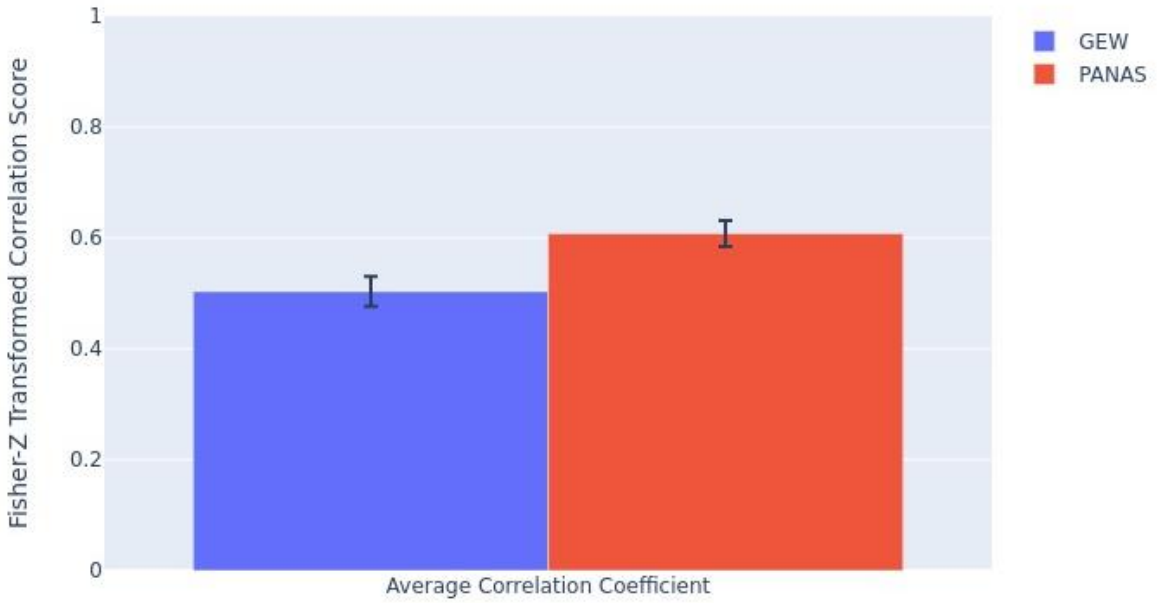

Figure 0-5 AVERAGES OF FISHER-Z TRANSFORMED CORRELATION SCORES FOR EACH METHOD (Transposed Data)