Reading Times Predict the Quality of Generated Text Above and Beyond Human Ratings

Sina Zarrieß **Sebastian Loth** **David Schlangen**
Bielefeld University
Universitätsstraße 25
33615 Bielefeld, Germany
{sina.zarriess,sebastian.loth,david.schlangen}@uni-bielefeld.de

Abstract

Typically, human evaluation of NLG output is based on user ratings. We collected ratings and reading time data in a simple, low-cost experimental paradigm for text generation. Participants were presented corpus texts, automatically linearised texts, and texts containing predicted referring expressions and automatic linearisation. We demonstrate that the reading time metrics outperform the ratings in classifying texts according to their quality. Regression analyses showed that self-reported ratings discriminated poorly between the kinds of manipulation, especially between defects in word order and text coherence. In contrast, a combination of objective measures from the low-cost mouse contingent reading paradigm provided very high classification accuracy and thus, greater insight into the actual quality of an automatically generated text.

1 Introduction

Evaluating and comparing systems that produce natural language text as output, such as natural language generation (NLG) systems, is notoriously difficult. Many aspects of linguistic well-formedness and naturalness play a role for assessing the quality of an automatically generated text. On the sentence-level, this includes grammatical and morpho-syntactic correctness, lexical meaning, fluency, and stylistic appropriateness. On the text-level, further criteria related to coherence, text structure, and content should be considered. One of the most widely applied and least controversial NLG evaluation methods is to collect human ratings. Human ratings have been used for system comparison in a number of NLG shared tasks (Gatt and Belz, 2010; Belz et al., 2011), for validating other automatic evaluation methods in NLG (Reiter and Belz, 2009; Cahill, 2009; Elliott and Keller, 2014), and for training statistical components of NLG systems (Stent et al., 2004; Mairesse and Walker, 2011; Howcroft et al., 2013).

When no extrinsic tasks or factors for evaluating an NLG system are available, human judges are typically asked to rate the quality of texts or sentences according to several linguistic criteria, such as 'A: how fluent is the text?' and 'B: how clear and understandable is the text?' (e.g. (Belz et al., 2011)). This is a hard and unnatural task for most naive users, and can be non-trivial even for experts: raters have to reflect on and differentiate between detailed, linguistic aspects of text quality, and assign scores precisely and systematically across a set of generated outputs that potentially contain various types of linguistic defects. The rating task turns increasingly difficult if they have to compare texts with multiple sentences and multiple types of linguistic defects, e.g. fluency on the sentence level, clarity and coherence on the text level. Consequently, low agreement between raters, and even inconsistencies between ratings of the same human judge have been found in previous studies (Belz and Kow, 2010; Cahill and Forst, 2010; Dethlefs et al., 2014). Standard evaluation methods for, e.g. text summarisation tend to avoid possible interactions between local sentence-level and global text-level defects. Instead, they focus on coherence and content (Nenkova, 2006; Owczarzak et al., 2012). In particular, this is due to the fact that independently rating coherence and clarity locally for each sentence and globally for an entire text is tedious, unnatural, tiring and hardly achievable for human judges.

In other disciplines of linguistic research, a range of experimental paradigms have been established that provide more systematic and objective means to assess human text reading. In particular, psycholinguistic approaches typically use objective measures such as reading times and eye movements to quantify how well human readers can process a sentence. The advantage of these measures is that humans typically focus on reading the text. Importantly, they do not consciously control their eye movements. Longer reading times or certain patterns of eye movements have been well associated with difficulties that humans encounter when reading text, e.g. apparent inconsistencies as garden path sentences (Christianson et al., 2001), and complex grammatical constructs (Traxler et al., 2002).

This paper investigates whether more objective reading measures can be exploited for evaluating NLG systems and systematically measuring text quality. However, using eye tracking for evaluation purposes is more costly than relying on ratings. Furthermore, most eye tracking studies used carefully designed stimuli to test a specific effect at a particular known position in a sentence. In sum, eye tracking is highly sensitive to pro-

Proceedings of the 15th European Workshop on Natural Language Generation (ENLG), pages 38–47,
Brighton, September 2015.

cessing difficulties. But due the costly devices and experiments, it was - to the best of our knowledge - not applied for evaluating comparably uncontrolled texts that are typical in NLG.

Thus, we have developed and tested mouse contingent reading (MCR) for evaluating generated texts. This method combines the sensitivity of eye tracking with the cost effectiveness of a rating study. The automatically generated texts are presented to human raters in a sentence-by-sentence, mouse-contingent way such that a number of parameters of the reading process are recorded, e.g. the time that people spent looking at single sentences and an entire text. We hypothesized that these parameters are more informative for the quality of a text than the user ratings of clarity and fluency.

As objective criteria for text quality are hardly available in NLG (Dale and Mellish, 1998; Hardcastle and Scott, 2008), we did not compare reading times and ratings on manual, potentially flawed annotations of text quality. Instead, we selected experimental material from a corpus-based generation framework that combines sentence-level linearisation and text-level referring expression generation (Zarrieß and Kuhn, 2013). We based our study on a set of texts that were available in 3 versions: (i) the "gold standard" corpus text, (ii) automatically linearised texts where word order deviated from the original corpus and contained potential fluency-related defects, (iii) texts with potential defects in referring expressions and linearisation which are likely to deteriorate clarity or coherence on the discourse level. We controlled the broad type of linguistic defects but not the details of each sentence or text. We argue that an objective evaluation method for NLG should clearly distinguish coherence and surface-related aspects of text quality.

In our data, there is a single human-authored version of each text which is free of errors. We do not know whether a deviation of the other versions is an error or an acceptable alternative realisation. Thus, in contrast to typical eye tracking studies we do not aim at detecting the effect of a particular type of error. Our assumption is more conservative: we expect that a set of automatically generated texts that deviates significantly from a set of corpus texts on several levels of linguistic realisation (referring expressions and linearisation) has lower quality than texts that only deviate from the corpus on a single level (linearisation). To further accommodate for the fact that we do not control the exact degree of acceptability of the potential defects, we add a set of filler texts that we manually manipulated to contain severe errors in coherence.

Based on the human ratings and MCR data collected for a set of automatically generated texts, we investigated whether a regression model can predict which types of linguistic defects were present in the text read by the participant, i.e. which generation components were used to generate it. We find that it is possible to achieve a good prediction accuracy for text quality, despite the fact that there is uncertainty with respect to the exact number and types of errors in the texts. However, the accuracy of the regression models varies considerably according to the type of predictors: Human ratings can hardly discriminate incoherent automatically generated texts from original corpus texts and texts containing defects in word order. A regression model based on reading time predictors achieved a very good fit and largely outperformed the rating model in separating different levels of quality in NLG output. This suggests that some effects were not reliably reflected in the subjective ratings that are consciously controlled and calculated by the participants. However, these effects were accounted for by the objective reading measures that are (mostly) outside of conscious control.

Section 2 provides background on research in NLG evaluation. Section 3 introduces our MCR paradigm. The generation framework we used to collect our experimental material is presented in Section 4. Section 5 describes the experimental design. The models are discussed in Section 6.

2 Background on NLG Evaluation

In recent years, the NLG community has become increasingly interested in comparative evaluation between NLG systems (Gatt and Belz, 2010; Koller et al., 2010; Belz et al., 2011; Banik et al., 2013; Hastie and Belz, 2014). Generally, evaluation methods for assessing NLG systems fall into three main categories: 1) automatic evaluation methods that compare system output against one or multiple reference texts, 2) human evaluation methods where human readers are asked to judge a text, typically with respect to several criteria. If the NLG component is embedded in an end-to-end system, such as a dialogue system, 3) extrinsic factors of task success and usefulness of the NLG output can be measured. For corpus-based NLG components such as surface realisers or referring expression generators, extrinsic factors cannot be assessed, but in this case, reference or gold text outputs are often available. Langkilde (2002) first suggested to use automatic evaluation measures inspired from methods in machine translation, such as BLEU (Papineni et al., 2002) or NIST (Doddington, 2002), that measure the n-gram overlap between the system and some reference text, sentence or phrase. The advantage of such automatic and cheap evaluation methods can be enormous. If tightly integrated in the development cycle of an NLG system, they allow fast and empirically optimised implementation decisions. In turn, a lot of research on NLG evaluation focussed on defining and validating automatic evaluation measures. Such a metric is typically considered valid if it correlates well with human judgements of text quality (Stent et al., 2005; Foster, 2008; Reiter and Belz, 2009; Cahill, 2009; Elliott and Keller, 2014). However, automatic evaluation measures in NLG still have a range of known conceptual deficits, i.e. they do not reflect appropriateness of content (Reiter and Belz,

2009), or meaning (Stent et al., 2005). Thus, many studies and evaluation challenges in NLG additionally collect human ratings to assess the quality.

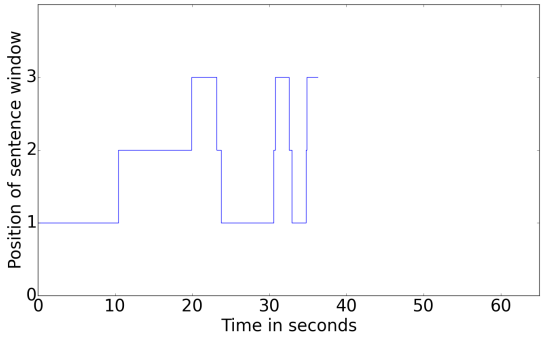
Compared to the large body of work on automatic evaluation measures, there has been little research that assessed the validity of human evaluation methods. Hardcastle and Scott (2008) provided an extensive discussion of human and automatic evaluation for text quality. They proposed a Turing-style test where participants are asked to judge whether a text was generated by a computer or written by a human. Belz and Kow (2010) showed that higher agreement between human raters can be obtained if they compare two automatically generated texts, instead of assigning scores to texts in isolation. Belz and Kow (2011) found that human judges preferred to use continuous rating scales over discrete rating scales. Siddharthan and Katsos (2012) investigated two offline measures inspired from psycholinguistic studies of sentence processing for assessing text readability, namely magnitude estimation and sentence recall. They demonstrate that the sentence recall method did not discriminate well between sentences of differing fluency if sentences were short. On the other hand, human judgements, did not discriminate well between surface level disfluencies and breakdowns in comprehension.

3 Mouse Contingent Reading

Figure 1: Screenshots of the mouse-contingent reading GUI. Top panel: at the start of each trial, all sentences are masked and the mouse cursor is positioned above them. Bottom panel: the participant has moved the mouse to the first sentence and unmasked it.

In mouse contingent reading (MCR), the reader is presented a text on a computer screen. The entire text is covered by a mask or masking pattern. Only if the reader moves the mouse cursor over a particular section of text, the mask is removed and the text is shown in clear font (see Figure 1). This paradigm is equivalent to gaze contingent reading (McConkie and Rayner, 1975; Reder, 1973) and self-guided reading (Hatfield, 2014) but it does not require an eye tracking device or touch sensitive device. However, the same metrics can be collected, i.e. the time spent on each area of interest and the scan path. Figure 2 shows an example of how the reader transits forth and back between areas of interest and how much time is spent on each area.

MCR log for original corpus text:



MCR log for generated text
(linearisation and referring expressions):

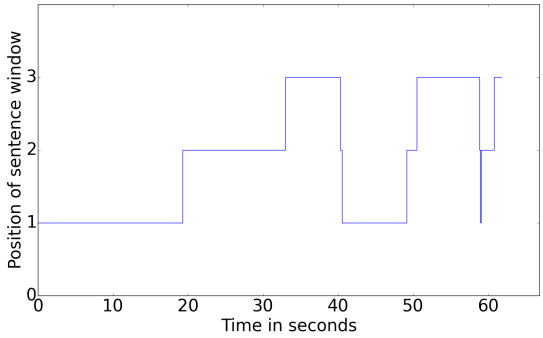


Figure 2: Visualisation of reading times recorded with MCR for a text in two different quality conditions

In reading studies, eye tracking and gaze contingent designs are the most popular paradigms. The words or phrases that a reader is currently processing and attending are indicated by fixations on them (Rayner, 1998; Rayner, 2009). However, hand motions are also highly informative to cognitive processes in general (Freeman et al., 2011). Importantly, a hand oriented paradigm requires much less technical efforts and allows a precise data acquisition. In case of MCR, the collected data approximate the comparable eye tracking data as they indicate which part of the text was attended.

4 Generation Framework

Zarrieß and Kuhn (2013) present a combined, corpus-based generation framework for two well-studied NLG

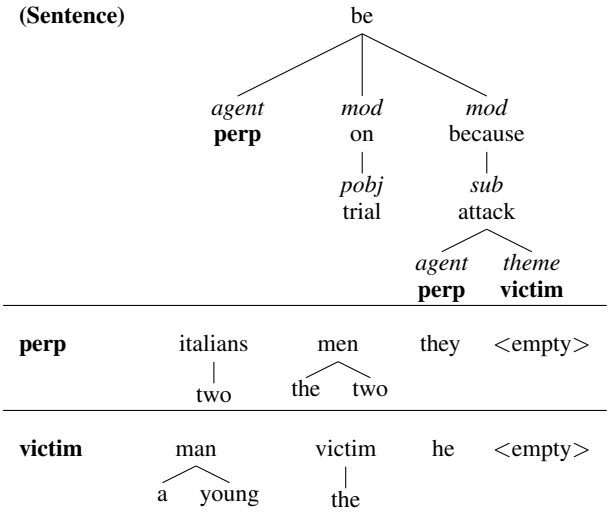


Figure 3: Example NLG input from (Zarrieß and Kuhn, 2013): a non-linearised dependency structure with slots for REs and lists for candidate RE realisations

sub-tasks: referring expression generation in context (REG) and surface realisation (linearisation). Their system generates texts from dependency-based inputs that can be more or less specified. Figure 3 illustrates an example for a dependency-based input with underspecified referring expressions. The linearisation component of the system predicts the order of words, i.e. nodes in the dependency tree. The REG component computes a ranking over candidate realisations for each RE slot and inserts the top-ranked expression into the tree. Additionally, these NLG inputs are available in a more specified version, i.e. as non-linearised dependency trees containing the referring expressions from the original corpus text. In this case, the NLG task consists of linearisation only.

Compared to other text generation tasks, such as e.g. text summarisation, this NLG framework is more restricted. The order of sentences, lexical choice and basic sentence structure are defined by the corpus-based input annotations. Our setting has two NLG components that can be switched on and off on demand. We exploit this for obtaining automatically generated text that differ in their quality. Thus, we use texts that have been generated from different components of the system. This approach is very similar to the idea of evaluating an NLG system in a architectural ablation mode, demonstrated in Callaway and Lester (2001).

The NLG inputs in Zarrieß and Kuhn (2013) were obtained from manual RE annotations and automatic dependency annotations for a corpus of 200 German newspaper articles. The texts are short reports on robberies as they can be found in the local newspaper. Thus, they all describe similar events between two referents (a victim and a perpetrator). The RE annotations also contain implicit mentions of victim and perpetrator referents in particular syntactic contexts (such as passives, or coordinations). Therefore, the RE component can delete REs that were realised in the original text, or introduce REs that were originally implicit.

Table 1 shows an example from the original corpus text and an automatically generated version by the system. Please note that neither gold nor generated texts contain punctuation. Since the system does not predict punctuation, this was removed from the original texts. Furthermore, the automatically generated text deviates from the original corpus text in a range of linearisation and REG decisions. These deviations are not necessarily ungrammatical or incoherent (as e.g. the predicted RE in the second sentence which is still understandable and does not degrade coherence). On the other hand, there can be sentences that are clearly misconstrued such as the third sentence where ungrammatical word order and incoherent or superfluous REs result in an unclear meaning of the sentence.

Thus, we controlled for the broad, expected level of text quality, rather than applying a costly manual annotation of error types present in the generated texts. As the focus of our study is on predicting defects in text quality that are due to clarity and fluency, we selected texts from Zarrieß and Kuhn (2013)'s data set where the linearisation and REs deviated from the original corpus texts. As a sanity check for our evaluation metrics, we included the original corpus texts and further manipulated some of the generated texts such that their referring expressions would be very hard to resolve and obscure the relation between two sentences. These texts were treated as *fillers*. Each generated text is available in two versions: a) generated by the linearisation and referring expression component containing defects in the realisation of reference and word order, and b) generated by the linearisation component containing perfect referring expressions and potential defects in word order. This provided us a hierarchy of levels of text quality. Linearisation mostly affects the fluency (and sometimes the grammaticality) on the sentence level, whereas wrong predictions of referring expressions can result in incoherent transitions between sentences which affects clarity on the text level.

5 Experiment

This study tested human evaluation methods for text generation. We focussed on the problem of evaluating NLG output formed of multiple sentences and detecting whether the user experienced difficulties in reading and understanding the text.

5.1 Hypothesis

In evaluations of text generation, the quality of a text has to be assessed on different levels of linguistic well-formedness including grammatical correctness, fluency, and intelligibility or clarity. Cases of misconstrued texts are not just right or wrong but they vary on a scale from well-formed through understandable but yet difficult to read to unintelligible. Often, it is difficult to pinpoint which components and decisions of

Original corpus text	Automatically generated text
Auf dem Weg von der U-Bahn-Haltestelle Dornbusch zu seiner Wohnung in Ginnheim ist ein 27jähriger in der Nacht zum Samstag überfallen und ausgeraubt worden	Auf dem Weg von der U-Bahn-Haltestelle Dornbusch zu seiner Wohnung in Ginnheim in der Nacht zu Samstag überfallen und ausgeraubt worden ist ein 27jähriger
On the way from the metro station Dornbusch to his apartment in Ginnheim, a 27-year-old has been attacked and robbed Saturday's night	
Der Täter hatte sein Opfer gegen 1.30 Uhr zunächst scheinbar harmlos nach der nächsten Telefonzelle gefragt	Der Täter hatte **den 27jährigen** gegen 1.30 Uhr zunächst scheinbar harmlos nach der nächsten Telefonzelle gefragt
*Around 1:30 o'clock, the perpetrator had asked his victim (***the 27-year-old***) for the next phone box in a seemingly harmless way*	
Nachdem ihm diese an der Ecke Ernst-Schwendler-Straße Platenstraße gezeigt worden war machte er kehrt und verfolgte den 27jährigen	Nachdem Platenstraße diese an der Ecke Ernst-Schwendler-Straße gezeigt ihm worden war er machte kehrt und verfolgte **der Täter sein Opfer**
*After it had been shown to him on the corner Ernst-Schwendler street Platen street, he returned and (***the perpetrator***) followed the 27-year-old (***the victim***)*	

Table 1: Example corpus text and corresponding NLG output. Word order defects are underlined, generated REs that differ from corpus REs are in bold face. The English translations do not show word order problems, but predicted REs are given in brackets and bold face.

the NLG contributed to the well-formedness. In particular, a single component can affect all levels of well-formedness, e.g. the realisation of word order can impair the readability and intelligibility of a text.

We expected that naïve participants would have difficulties in independently rating different aspects of text quality, e.g. clarity and fluency. We assumed that the rating task would be even more tedious on the sentence-level such that we collected global user ratings for fluency and clarity. We hypothesised that the parameters of the reading process such as the time spent on individual sentences, and the transitions between sentences would be more objective, local measures and can at least complement ratings of perceived quality. Thus, we recorded the reading parameters in our study and aimed to identify the links between ratings, reading parameters and levels of text quality.

Suboptimal NLG decisions affect an entire sentence or text. Thus, the MCR study was designed to assess the well-formedness of larger units. We used three comparably large areas of interest formed by each sentence of the texts. In contrast to typical reading studies at the level of single sentence processing, the cursor motions were selectively recorded for transitions between sentences. These transitions are most likely related to measures at the text level that we were interested in, i.e. clarity and fluency. Furthermore, the ratings of clarity and fluency were collected with regards to the entire text.

5.2 Experimental Setting

Participants Thirty-three participants were recruited from the department's participant pool (including students and staff). They received €5 as well as candy sweets in exchange for their time and effort.

Apparatus The participants were seated in front of a typical office computer screen. A dedicated Python programme controlled the presentation of the stimuli, recorded the reading times[1] and mouse transitions, and collected the ratings. The participants interacted with the programme through a standard mouse and keyboard.

Materials and Conditions From the set of NLG outputs with potential defects in clarity and fluency, as described in Section 4, we randomly selected 16 texts. A subset of twelve texts were presented in three conditions: a) the original corpus text without any defects (*gold*), b) automatically linearised texts that could include defects in word order (*lin*), and c) automatically linearised texts with automatically generated referring expressions, i.e. these texts could include defects in word order and referring expressions (*sem*). The remaining four sentences were hand manipulated such they would be clearly distorted in terms of syntax, reference and intelligibility (*filler*).

The critical texts were assigned to one of three lists such that all lists contained four texts per condition and each text occurred once per list. Additionally, all lists included the four filler items.

Procedure The participants were welcomed to the lab and handed a written consent form. If they agreed to take part in the study, the participants were handed written instructions asking them to read the texts displayed on screen using the mouse and to rate them for clarity and fluency afterwards.

The session started with an additionally selected practice item to familiarise the participants with the design of the study. In the experimental session, the 16 items were presented in random order. Each trial started as soon as the participant confirmed the ratings to the previous item. The mouse cursor was positioned

[1]The reading times were approximated by measuring the dwell time of the mouse cursor on a sentence. This is equivalent to measuring the dwell time of the point of gaze in gaze contingent reading.

above the three sentences such that the entire text was masked. The participants initiated the clock by moving the cursor to the first sentence. The sentences unmasked as soon as the mouse cursor entered the white space surrounding the script and was masked again as soon as the mouse left this area (see Figure 1). Thus, at any point in time either one or no sentence was presented without the masking pattern. Once the participant had completed reading the text, they clicked a confirmation button below all sentences.

This button click triggered the display of two rating questions. First, a fluency rating was elicited by asking "How well does the text read? Is it formulated in a linguistically correct way?" Secondly, "How clear and understandable is the meaning and content of the text?" asked for a clarity rating. Instead of a discrete Likert-scale, we adopted the magnitude estimation paradigm, i.e. the participants were instructed to score sentences relative to each other by assigning them a number (Siddharthan and Katsos, 2012). The entire session including instruction and debriefing lasted about half an hour.

6 Results

In total, we collected reading and rating data of 528 experimental trials from thirty-three participants. In the following, we analyse the subjective ratings and the objective reading parameters with respect to our experimental conditions and investigate whether they can separate texts with different types of linguistic defects. Table 2 provides an overview of the measures we calculated and used as predictors for regression models.

6.1 Ratings

In the magnitude estimation paradigm, each participant uses their own numerical scale for assigning fluency and clarity scores. The raw scores were standardised with a z-transformation such that 0 is the mean rating for each participant. The variables **fluency-*z*** and **clarity-*z*** indicate to which extent a participant's rating of a text is better or worse than their mean rating.

As shown in Table 2, the overall tendency for fluency and clarity z-scores was as expected: on average, participants assigned the highest scores to *gold* texts, followed by *lin*, *sem* and *filler* texts. This suggests that on average the participants rated the evaluation criteria as intended and that the hierarchy of perceived text quality corresponds to our assumptions. Furthermore, the relatively low standard deviation between the means of the participants' ratings indicates that z-scores obtained from magnitude estimation ratings are relatively consistent.

6.2 Reading Measures

Using the MCR design, we recorded the time spent for reading single sentences and the text also the scan-path, i.e. the number and order of transitions and regressions between sentences. For identifying the most informative predictors, we derived a number of measures from the raw data that are described in the following.

Reading Time Using the dwell times (the time span that a particular sentence was not masked) and number of words per sentence, we computed the **speed**[2] and the **pace**[3] as first order derivatives. Nine predictors at sentence level and three at text level were computed. In addition, we computed the time required to read the entire text for the first time. Normalising this time span by the total reading time of the text provides a measure for how much time was spent on regressions within the text compared to the first pass.

Scan-path Next to dwell times, the scan-path can inform about the quality of a text. This could be reflected in how often particular sentences were visited and how often the participant transited between sentences. However, our regression analyses showed that reading time measures are generally more effective than scan-path predictors (see Section 6.3 below). In Table 2, we show the means for **path-log** as a log-normalised measure for the total number of transitions between sentences.

Standardising The individual differences in reading times and scan-paths between individual participants were pronounced. Additionally, they were also differences between texts, e.g. their content and lengths. As with the ratings, we added a standardised (z-score) measure of the reading parameters to the list of predictors (e.g., pace-total-z). This z-score is based on the mean and standard deviation of one parameter of one participant. For accommodating the variance in between the texts, we computed a second z-score (termed 'z2' in the following) based on the mean and SD of an items' reading times for all sentence-level predictors. This score reflects how a text or sentence compares to the other items.

As shown in Table 2, the means of the reading time measures do not comply with the expected quality hierarchy in the same way as ratings. Thus, it was not the case that lower quality texts are generally read more slowly than more coherent or gold texts. For instance, *sem* texts were read slowest (total time) whereas fillers can be identified by a high pace and large number of transitions, i.e. a long scan-path. *Sem* and *lin* texts can be distinguished in terms of the local, sentence-based reading times, e.g. 'speed_sent_z2' or 'time_sent3_z2'. Thus, the ratings and the MCR data appear to provide disjoint information such that one cannot substitute the other, e.g. a low clarity rating does not imply a prolonged reading time and vice-versa.

6.3 Regression Models

For testing whether and to what extent the user ratings and the reading times discriminate between the types of generated texts, the measures were used as predictors in regression models. This provides insight into

[2]Number of words in a sentence or text divided by the summed reading times

[3]Summed dwell time divided by length of text or sentence

		Filler		Sem		Lin		Gold	
		mean	sd	mean	sd	mean	sd	mean	sd
Ratings	fluency_z	-0.41	0.38	-0.32	0.50	-0.04	0.36	0.77	0.47
	clarity_z	-0.53	0.40	-0.18	0.49	0.17	0.36	0.54	0.51
Text RTs	pace_total_z	0.79	0.33	-0.22	0.64	-0.17	0.59	-0.40	0.65
	path_log	2.29	0.42	2.04	0.41	2.10	0.37	2.14	0.43
	speed_total_z	-0.44	0.12	0.22	0.79	0.07	0.81	0.15	0.87
	time_total_z	0.04	0.37	0.15	0.46	-0.09	0.49	-0.10	0.62
	time_1stpass_z	-0.29	0.36	0.30	0.49	0.09	0.48	-0.11	0.47
Sentence RTs	pace_1sent_z2	0.00	0.79	-0.04	0.83	-0.02	0.75	0.06	0.75
	pace_2sent_z2	-0.00	0.70	0.02	0.74	-0.01	0.64	-0.00	0.71
	pace_3sent_z2	0.00	0.72	0.08	0.65	0.07	0.78	-0.15	0.60
	speed_1sent_z2	-0.00	0.79	0.12	0.86	-0.10	0.71	-0.02	0.72
	speed_2sent_z2	0.00	0.70	0.21	0.79	-0.10	0.64	-0.11	0.62
	speed_3sent_z2	-0.00	0.72	0.19	0.63	0.01	0.72	-0.20	0.55
	time_1sent_z2	-0.00	0.79	0.04	0.85	-0.06	0.73	0.02	0.74
	time_2sent_z2	-0.00	0.70	0.12	0.77	-0.06	0.64	-0.06	0.67
	time_3sent_z2	-0.00	0.72	0.15	0.65	0.03	0.76	-0.18	0.58

Table 2: Means and SD for ratings, text-based and sentence-based reading times. SD is computed on mean values per participant and indicates agreement/consistency between participants.

the type of relation between the text quality conditions on the one hand, and multiple evaluation metrics on the other hand. The dependent variable of our models was the text condition, with four possible values - *filler*, *sem*, *lin* or *gold*. We used hierarchical binary regression[4] and we fitted three binary regression models that iteratively distinguish a particular text type from a set of remaining texts. The hierarchy of the models corresponds to the types of errors and to the level of text quality: First, we applied a *filler* model that should separate *filler* texts (25%) from all other texts. These items were manually distorted and thus, contained a greatest number of defects. In the next step, we excluded the fillers items and build a model that separates *sem* texts (33%) from the remaining *lin* and *gold* texts. The *sem* texts were automatically linearised and included generated referring expressions, thus the remaining items were expected to entertain less defects. Finally, we designed a model that separates *lin* (50%) from *gold* texts.

We were interested in how well different sets of predictors perform in the regression analysis. For each step (*filler, sem, lin*) of the text quality hierarchy, we evaluated the following models: a) **Ratings** based on fluency and clarity z-scores, b) **Text RTs** based on text-level time, space, speed, time-1stpass and their z-scores, c) **All RTs** based on Text RTs and sentence-level time, pace, speed and corresponding z-scores (computed over texts), d) **Combined** based on a combination of Ratings and All RTs.

We excluded non-significant predictors using stepwise backward regression. Therefore, each model in-

Predictors	% Fit	% Acc.	# Coef.	R^2
	Filler vs. other (Sem, Lin ,Gold)			
	Majority BL: 75%			
Ratings	76.33	76.14	1	0.133
Text RTs	81.06	81.06	2	0.359
All RTs	100.0	96.78	11	0.885
Combined	100.0	96.02	11	0.905
	Sem vs. other (Lin, Gold)			
	Majority BL: 66.66%			
Ratings	67.91	67.93	2	0.143
Text RTs	66.92	64.89	5	0.113
All RTs	100.0	94.19	14	0.926
Combined	100.0	94.94	15	0.943
	Lin vs. other (Gold)			
	Majority BL: 50%			
Ratings	66.66	66.29	2	0.26
Text RTs	68.18	65.15	9	0.23
All RTs	75.75	67.42	18	0.412
Combined	77.65	74.62	17	0.521

Table 3: Hierarchical binary regression for text quality conditions, using different sets of predictors ('RTs' stands for reading times, 'All RTs' include text and sentence reading measures).

[4]Ordinal or multinomial regression can handle multiple values in the dependent variable, but uses more complex statistics and the resulting models are harder to interpret.

cludes a different number of coefficients. In Table 3, we report the performance of the final models obtained from backward regression in terms of the goodness of fit (% Fit), the prediction accuracy in ten-fold cross validation (% Acc.), the number of significant predictors (# Coef.) and Nagelkerke's R^2.

Table 2 shows that that the clarity and fluency ratings of *filler* and *sem* texts are lower on average. But the data in Table 3 indicate these ratings hardly achieved any error reduction compared to the majority baseline, i.e. these ratings were not informative with regards to identifying these texts. This is particularly remarkable in the case of *fillers* as they are clearly erroneous and should be identified by any reliable metric. The rating model performs slightly better in the last step of the hierarchical regression, i.e. for distinguishing linearised texts and original corpus texts. We attribute this effect to the pronounced difference between fluency ratings of gold texts as compared to other texts (see Table 2).

The global text-level predictors **Text RTs** perform slightly better in the case of *fillers*, and comparably worse in the *sem* and *lin* conditions. However, when we add sentence-level reading times to the set of predictors, the model achieves an accuracy of 96% for discriminating *sem* texts and 94% for fillers which is above and beyond the rating model. The high accuracy shows that mouse contingent reading data are very informative with regards to the quality of automatically generated texts.

We note that combining the reading parameters and the ratings in the *filler* and *sem* model did not improve the accuracy compared to only using the reading parameters (see Table 3). Thus, we attribute most of the predictive power of the combined model to the reading measures. In the *filler* model, the clarity rating score was statistically significant, but did not add to the prediction accuracy of the model. In the *sem* model, the fluency rating is significant. When distinguishing *gold* and *lin* texts, the reading parameters were less effective predictors compared to the *filler* and *sem* models. This affected the model's accuracy such that the fluency ratings contributed significantly to the model. However, including the reading parameters still improved the model's accuracy substantially. This suggests that the measures we recorded with sentence-by-sentence reading are especially informative for predicting quality defects on the level of text clarity and coherence.

6.4 Predictors

In Table 4, we present the plain coefficients for the final *filler*, *sem* and *lin* models that we obtained from combining sentence-level and text-level reading measures and ratings. The stepwise backward regression procedure excludes different subsets of predictors from the initial models. For instance, the text-level reading times, such as total speed and pace, are not significant for identifying *filler* and *sem* texts. On the other hand, they discriminate between *lin* and *gold* texts. The

Predictor	**Filler**	**Sem**	**Lin**
(Intercept)	-13.457	8.704	0.271
pace-1sent-z2	8.122	-52.90	-
pace-2sent	22.033	-21.38	9.188
pace-3sent	-	76.32	8.067
pace-3sent-z2	-	-66.76	-
speed-1sent-z2	-	213.2	-
speed-2sent	0.035	0.023	0.021
speed-2sent-z2	9.683	12.65	-
speed-3sent	-0.084	0.097	0.009
speed-3sent-z2	-	-	0.732
time-1sent	-1.17	-0.7345	0.607
time-1sent-z2	-	-259.4	-
time-2sent-z2	-21.5	-120.5	-1.223
time-3sent	2.047	-5.399	-
time-3sent-z2	-2.55	62.61	-
pace-total	-	-	-31.852
pace-total-z	-	-	4.298
display-3sent-log	-	-	1.331
time-1stpass-z	-	-	2.04
speed-total	-	-	-0.003
speed-total-z	-	-	3.062
time-1stpass-z	-	-	-1.23
time-total-z	-	-	-2.283
fluency-z	-	-1.629	-1.068
clarity-z	-1.352	-	-

Table 4: Sentence-level, text-level, rating-based predictors and their coefficients in final *filler*, *sem* and *lin* models from Table 3

filler and *sem* models used sentence-level predictors for time, pace and speed of particular sentences. This pattern suggests that defects in referring expression realistion, which are present in *filler* and *sem* texts but not in *lin* and *gold* texts, deteriorate the clarity and coherence of NLG output which is reflected in longer reading times on particular sentences in a text.

6.5 Discussion

Generally, our results corroborate the common claim that quality of generated text is a multi-faceted and graded phenomenon which cannot be reduced to a small number of quality criteria that can be easily assessed in a rating task. Despite the fact that averaged ratings seem to correspond roughly to the expected hierarchy of text quality, a regression analysis of individual ratings for text instances shows a more detailed picture. A combination of reading time metrics identifies generated texts that contain defects in word order and referring expressions with high accuracy, while the rating predictors could hardly discriminate between instances of different text quality conditions. We found that objective sentence-level and text-level reading time measures can complement each other and account for complex interactions between aspects of text quality. This result has implications for standard human evaluation set-ups in NLG, summarization and possibly Machine Translation which are often based on several self-reported rating criteria.

We showed that an experimental paradigm such as MCR provides low-cost and natural means for recording objective reading measures while sidestepping the technical and practical requirements of an eye-tracking study. Our MCR set-up is based on a simple GUI that presents pieces of text in a mouse-contigent way and can be deployed on crowd-sourcing platforms.

Further research is needed to understand how predictors generalise and how the metrics can be applied to a reliable comparison of NLG systems. The fact that standardising across participants and across texts was effective, implies that prior knowledge about individual reading behavior of a participant is needed to accurately identify texts where understanding and reading difficulties occurred. Such parameters could be collected by introducing additional error-free and clearly erroneous texts into the experiment. The acquired data would reflect a burn-in phase for the predictors and provide the data for standardising the metrics.

A surprising result is that is that (error-free) gold texts were not associated with faster reading times. It is possible that the fact that users had to rate each text after reading it might have impaired their natural reading behavior. On the other hand, users might spend less time on clearly defictive texts as they were unable to integrate them syntactically and/or semantically. This effect will be investigated in future work.

7 Conclusion

Evaluating automatically generated texts is a complex task and involves dealing with a range of interacting levels of linguistic realisation. While many users can easily and naturally read texts, they cannot be expected to provide detailed, objective and systematic assessments of the linguistic quality of a text. This study suggests that there is a lot to be gained from exploring psycholinguistically plausible methods and paradigms for human evaluation in NLG. We adopted a simple and low-cost mouse contingent reading paradigm for an evaluation study in text generation. We showed that parameters of the reading process recorded with MCR, such as reading time for texts and sentences, provide very effective predictors for discriminating between generated texts of different quality levels, whereas self-reported quality ratings do not.

References

Eva Banik, Claire Gardent, and Eric Kow. 2013. The KBGen challenge. In *Proceedings of the 14th European Workshop on Natural Language Generation (ENLG)*, pages 94–97.

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *Proceedings of the 6th International Natural Language Generation Conference*, INLG '10, pages 7–15, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anja Belz and Eric Kow. 2011. Discrete vs. continuous rating scales for language 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. Association for Computational Linguistics.

Anja Belz, Michael White, Dominic Espinosa, Eric Kow, Deirdre Hogan, and Amanda Stent. 2011. The first surface realisation shared task: Overview and evaluation results. In *Proceedings of the 13th European workshop on natural language generation*, pages 217–226. Association for Computational Linguistics.

Aoife Cahill and Martin Forst. 2010. Human evaluation of a german surface realisation ranker. In *Empirical methods in natural language generation*, pages 201–221. Springer.

Aoife Cahill. 2009. Correlating human and automatic evaluation of a german surface realiser. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 97–100. Association for Computational Linguistics.

Charles Callaway and James Lester. 2001. Evaluating the effects of natural language generation techniques on reader satisfaction. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 164–169.

Kiel Christianson, Andrew Hollingworth, John F Halliwell, and Fernanda Ferreira. 2001. Thematic Roles Assigned along the Garden Path Linger. *Cognitive Psychology*, 42(4):368–407, June.

Robert Dale and Chris Mellish. 1998. Towards evaluation in natural language generation. In *Proceedings of the First International Conference on Language Resources and Evaluation*, volume 562.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. In *Proceedings of EACL 2014*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*, volume 452, page 457.

Mary Ellen Foster. 2008. Automated metrics that agree with human judgements on generated output for an embodied conversational agent. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 95–103. Association for Computational Linguistics.

Jonathan B. Freeman, Rick Dale, and Thomas A. Farmer. 2011. Hand in Motion Reveals Mind in Motion. *Frontiers in Psychology*, 2.

Albert Gatt and Anja Belz. 2010. Introducing shared tasks to nlg: The tuna shared task evaluation challenges. In *Empirical methods in natural language generation*, pages 264–293. Springer.

David Hardcastle and Donia Scott. 2008. Can we evaluate the quality of generated text? In *Proceedings of LREC*.

Helen Hastie and Anja Belz. 2014. A comparative evaluation methodology for nlg in interactive systems. In *Proceedings of LREC'14*.

Hunter Hatfield. 2014. Self-Guided Reading: Touch-Based Measures of Syntactic Processing. *Journal of Psycholinguistic Research*, October.

David Howcroft, Crystal Nakatsu, and Michael White. 2013. Enhancing the expression of contrast in the sparky restaurant corpus. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 30–39, Sofia, Bulgaria, August. Association for Computational Linguistics.

Alexander Koller, Kristina Striegnitz, Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. 2010. The first challenge on generating instructions in virtual environments. In *Empirical Methods in Natural Language Generation*, pages 328–352. Springer.

Irene Langkilde-Geary. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proceedings of the 12th International Natural Language Generation Workshop*, pages 17–24.

François Mairesse and Marilyn A Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

George W. McConkie and Keith Rayner. 1975. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17(6):578–586, November.

Ani Nenkova. 2006. Summarization evaluation for text and speech: issues and approaches. In *Proceedings of INTERSPEECH*.

Karolina Owczarzak, John M Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization*, pages 1–9. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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. Association for Computational Linguistics.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422.

Keith Rayner. 2009. Eye movements in reading: Models and data. *Journal of Eye Movement Research*, 2(5):1–10.

Stephen M. Reder. 1973. On-line monitoring of eye-position signals in contingent and noncontingent paradigms. *Behavior Research Methods & Instrumentation*, 5(2):218–228, March.

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

Advaith Siddharthan and Napoleon Katsos. 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. Association for Computational Linguistics.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *Computational Linguistics and Intelligent Text Processing*, pages 341–351. Springer.

Matthew J Traxler, Robin K Morris, and Rachel E Seely. 2002. Processing Subject and Object Relative Clauses: Evidence from Eye Movements. *Journal of Memory and Language*, 47(1):69–90, July.

Sina Zarrieß and Jonas Kuhn. 2013. Combining referring expression generation and surface realization: A corpus-based investigation of architectures. In *ACL (1)*, pages 1547–1557.