Evaluating humorous properties of texts

Graeme Ritchie[1], **Robyn Munro**[2], **Helen Pain**[3], **Kim Binsted**[4]

Abstract. The success of a humour-generation program is usually assessed by having human judges rate texts. However, there has been little consideration of the patterns shown by such judgements, particularly in terms of consistency. We present two small studies which attempt to gauge the consistency of human judgements about humorous aspects of texts, and discuss some of the methodological issues involved.

1 MOTIVATION

In developing affective natural language generation systems, the question arises of how best to evaluate the performance of a system. Ideally, the NLG system would function as part of some larger task, and rigorous evaluation would assess the contribution of the generated texts to some desired qualities of the overall system, such as efficacy, usability or pleasantness. However, when a language generator is being developed, there is a practical need to be able to test whether the generated text meets certain requirements (one could think of this as *formative evaluation*, by analogy with educational testing). This may have to be done without the full context of some larger task-oriented system. Also, even when evaluating a full system, the contribution of the NLG component will be clearer if we have some idea of the nature of the texts it produces. This leads to the notion of trying to evaluate the quality of text produced by an NLG system, an area which has attracted an increasing amount of reflection in recent years (e.g. [8], [9], [2]).

We focus here on one particular class of texts, the generation of which would constitute one form of affective NLG, namely *humorous* texts. In particular, we focus on jokes, as a small, manageable genre of text for controlled study (see [11, Ch.2] for methodological arguments in favour of this restricted focus).

There are a few, usually small, studies in which the quality of computer-generated humorous text is considered (e.g. [7], [15]). These have all been done by showing texts (under experimental conditions) to human judges, and asking for ratings of the texts. This method, which has also been used for evaluating non-humorous generated text, seems relatively straightforward, easy to administer, and clear in its findings. However, it has a tacit assumption: that ratings by human judges of the humorous properties of texts will be relatively systematic. If judges rate texts in a random manner, then it is not convincing to claim success in humour-generation by showing that computer output is rated as randomly as human-written output is. None of the existing studies of computer-generated humour included any check of agreement across judges, or the consistency of the rating of texts (either control or computer-generated items). It is this issue which we wish to examine here.

Away from the area of computer generation, there are findings which show *correlations* between preferences for particular jokes or types of joke, most notably Ruch's development of the 3WD test [12], but these have not explored consistency, nor compared judgements of jokes with judgements of non-jokes.

We summarise here the results of two studies which explore the extent to which judges make consistent ratings of texts in terms of humour. The studies are very preliminary, but they do raise questions about what might be a suitable methodology for assessing the success of a humour-generating program.

Both the studies investigate two possible notions of 'humorous': whether a text is a joke or not (*jokehood*) and how funny the text is (*funniness*); see [11, Ch. 2] for discussion of this distinction. Informally, the initial conjectures were that jokehood would show consistency of ratings across judges, but funniness would be very varied.

2 STUDY 1: PUNNING RIDDLES

2.1 Data collection

As part of a project to study computer-generated jokes, data were collected involving judgements, by young children, about the humorous properties of short texts (all of the same general form - question and short answer). Fuller details are given in [4] and [5], so only a brief outline of the data collection methods are given here. The analysis here, of consistency, was not part of the original project, but was carried out retrospectively on the collected data some years later.

Data items were of 4 distinct types (total quantities[5] in parentheses):

J: computer-generated texts (80). These were output items from Binsted's JAPE computer program [4], which contained rules intended to create punning riddles; e.g. : *What do you get when you cross a bird and a blunder? A fowl up*.

H: human-written jokes (60). These were punning riddles selected from published joke books, chosen as far as possible to be similar in structure and genre to the target text type of the computer program; e.g. *What kind of animal plays cricket? A bat*.

S : sensible question & answer (30). A number of non-humorous, factually correct texts were constructed in a constrained way, consisting of a question and a single-phrase answer; e.g. *What kind of yellow fruit can you eat? A banana*.

N : nonsense question & answer (30). A number of texts made up of a question and a single-phrase answer were constructed, using random content words (nouns, adjectives, etc.) from the vocabulary employed in the other items; e.g. *What do you get when you cross a remedy with a mall? A coarse line*.

[1] Computing Science, University of Aberdeen, UK. email: g.ritchie@abdn.ac.uk
[2] formerly Informatics, University of Edinburgh, UK.
[3] Informatics, University of Edinburgh, UK. email: h.pain@ed.ac.uk
[4] Information and Computer Sciences, University of Hawaii, USA. email: binsted@hawaii.edu

[5] As taken from the original data files.

Children aged 8 to 11 completed questionnaires, each with 20 items (suitably balanced and randomised) accompanied by audio versions on tape. No mention was made of computer-generated jokes. Required responses for each item were:

- Is this a joke? [YES/NO]
- How funny is it? [5 point scale]
- Have you heard it before [YES/NO]

Although each item in the total set of items was judged by more than one subject, not all items were judged the same number of times, and no items were seen by all subjects. In total, there were data sets for 120 participants.

2.2 Results

The conjectures which motivated this work were stated briefly and informally at the end of Section 1, but we have not yet presented these as precise hypotheses about variables involved in the two studies. The question of how best to quantify, statistically, the intuitive notion of 'consistency' is not totally clear.

2.2.1 Jokehood

The percentages of joke and non-joke ratings for each text type in Study 1 are shown in Table 1.

	J	H	S	N	All
J	55.90	61.18	47.93	45.63	54.70
NJ	44.10	38.82	52.07	54.64	45.30

Table 1. Study 1: % age of joke/non-joke ratings, by text type

Tests such as χ-square and Wilcoxon Signed Ranks showed various differences (or lack of differences) in the balance of joke/non-joke ratings across the four types [4, 5]. However, such tests do not address the question of consistency of the ratings. A possible measure of consistency for the jokehood judgements is to apply the Sign (binomial) test (two-tailed) to the aggregate ratings for each item, and determine what proportion of the texts show significant skew away from a chance outcome; see Table 2.

p	J	H	S	N	All
< 0.05	15.00	23.33	6.67	6.67	15.00

Table 2. Study1: % age of items showing significance for jokehood

It could be argued that, since this approach involves a number (200) of applications of the Sign Test, we are really testing that large number of hypotheses, and so a correction (e.g. Bonferroni) should be made, resulting in a lower threshold than $p < 0.05$. However, it is a rather odd perspective to treat every trial (item) as a separate hypothesis. This draws attention to a drawback of using the Sign Test in this way: it does not yield a single overall measure of the statistical significance of the outcome of the whole experiment (but see Section 4 below).

In view of the very low percentage of items showing significance at the 0.05 level, there was little point in exploring a lower threshold.

2.2.2 Funniness

For funniness, we are also interested in consistency, although our initial conjecture is that there will *not* be much consistency (owing to variations in personal taste). A number of indicators of variation in funniness ratings were considered.

On the 5-point scale, out of 200 items, 192 had (across all raters) minimum ratings of 1, and 195 had maximum ratings of either 4 or 5. The standard deviation, which gives some indication of spread of values, had – across all items – a minimum of 0.64, a mean of 1.19 and a maximum of 1.65 (where the mean across the funniness rating means for all items was 2.6). This does seem to suggest quite a wide spread of values.

The funniness ratings (on a 5-point scale) were then simplified by mapping all scores 1-2 into a rating of *low* (L), and those of 4-5 into *high* (H), with ratings of 3 omitted. The structure of the data was then similar to that for jokehood, and analogous tests could be applied. Table 3 shows the proportions of the H/L rated items for each text type (omitting judgements not rated as either H or L).

	J	H	S	N	All
H	39.73	48.43	33.57	29.37	39.87
L	60.27	51.57	66.43	70.63	60.13

Table 3. Study 1: % age of high/low funniness ratings, by text type

Out of 200 items, 21 had exactly equal numbers of H and L scores. From the remaining 179, only 20 (10% of the original total) had an imbalance between H and L scores that was significant under the Sign Test ($p < 0.05$); see Table 4.

p	H	J	S	N	All
< 0.05	3.33	7.50	6.67	23.33	10.0

Table 4. Study 1: % age of items showing significance for funniness

At $p < 0.001$, none of the items showed a significant H/L imbalance.

3 STUDY 2 : NARRATIVE JOKES

3.1 Data collection

The aim of this study [10] was to address the central question in the current paper: the consistency of judgements about the humorous qualities of short texts.

The participants were 80 undergraduate students between the ages of 18 and 24 years of age, all of whom spoke English to a native standard and had no problems with reading or writing.

In order to create texts which systematically varied their humorous properties, but which were nevertheless similar in other respects, we adapted data used by [3] and [13]. These earlier studies had created 16 items in which there was a *setup* (a short narrative of about three sentences) followed by a choice of four short (one sentence) possible endings. Subjects in these studies were asked to select the correct ending for the text. The four possible endings were always of the same four types: *correct punchline* (JK) – something which combined with the setup to form a joke; *humorous non-sequitur* (HNS) – an absurd action which did not integrate with the setup; *associated non-sequitur* (ANS) – an event which superficially connected to the situation in the setup, but which did not follow on; *straightforward* (SF) – an event which combined with the setup to form a coherent, non-humorous narrative. For example:

> A ship is cruising in the Caribbean. One day a girl falls overboard and her father screams: "I'll give half my fortune to save

her." A fellow jumps in and saves the girl. The father says, "I'll keep my promise. Here's half my fortune."
JK: The fellow answers, "I don't want money; all I want to know is who shoved me."
HNS: Then the fellow tips his hat to the girl and his toupee slips off.
ANS: The fellow says, "I usually get seasick on boats."
SF: The fellow answers, "Thank you. I need the money."

By appending each of the 16 setups to each of its 4 possible endings, we created 64 items, 16 of each of the 4 types. These were then made into suitably balanced and randomised 16-item questionnaires, where each item had 4 questions:

- Do you consider the text a joke or not a joke? [Joke/ Not a joke]
- How funny did you find the text? [7-point scale from 'not funny at all' to 'very funny'].
- How aversive, or how dislikable, did you find the text? [7-point scale from 'not aversive' to 'very aversive'].
- Have you heard this text, or one similar, before? [3 choices: 'definitely yes', 'not sure', 'definitely no']

3.2 Results

3.2.1 Jokehood

As in Section 2.1, Table 5 shows the proportion of jokehood judgements, and Table 6 shows how many items showed a significant bias in one direction. The second row of Table 6 shows the results for $p < 0.001$; see remark about p values in Section 2.2.1.

	JK	HNS	ANS	SF	All
J	95.61	21.62	13.36	20.07	37.78
NJ	4.39	78.38	86.64	62.22	79.93

Table 5. Study 2: % age of joke/non-joke ratings, by text type

p	JK	HNS	ANS	SF	All
< 0.05	100	68.75	81.25	68.75	79.69
< 0.001	100	31.25	68.75	50	62.5

Table 6. Study 2: % age of items showing significance for jokehood

3.2.2 Funniness

On the 7-point scale, out of 64 items, 63 had (across all raters) minimum ratings of 0 or 1, and 29 had maximum ratings of either 5 or 6; hence, around 44% of items had a difference of 4 points or more across their ratings. The standard deviation had – across all items – a minimum of 0.55, a mean of 1.23 and a maximum of 1.83 (where the mean funniness rating for all items was 1.44).

Next, the funniness ratings (on a 7-point scale) were simplified by mapping all scores 0-2 into a rating of *low*, and those of 4-6 into *high*, with ratings of 3 omitted (much as in Study 1).

The low and high ratings (as percentages of total low & high ratings) are shown in Table 7.

Using the Sign Test on individual items gave the results in Table 8. For *all* the items in HNS, SF, and ANS, there were majorities for low funniness, with only two failing to reach statistical significance (ratings splitting 10:4 for these). For the JK texts, only 1 joke reached significance, with a 13-to-1 majority voting it highly funny; of the other 15 JK items, 5 were voted high, 8 were voted low and 2 tied.

	JK	HNS	SF	ANS	All
H	49.77	14.5	7.39	3.17	16.81
L	50.22	85.50	92.61	96.83	83.19

Table 7. Study 2: % age of high/low funniness ratings, by text type)

p	JK	HNS	SF	ANS	All
< 0.05	6.2	87.5	100.00	100.00	73.44

Table 8. Study 2: % age of items showing significance for funniness

In Study 2, the conjecture (that there will be variation) was broadly supported *for items in the JK* category; for the other three (non-joke) types of text, there was high *agreement* (that these items were not very funny). That is, this study suggests that there is great variation of opinion about the funniness *of jokes*, but general consensus that other types of text (or at least those used in this study) are definitely not funny. (This latter trend tends to support the jokehood results for this study.)

4 THE KAPPA TEST

The Kappa (κ) test [14, Sect 9.8] is used in many studies to rate overall agreement between judges, generally in situations where there is a need to establish reliable ratings of data (e.g. in marking up a language corpus for further analysis [6]). It might seem, therefore, that it would neatly fulfil the need for an overall rating of the degree of consistency in our ratings.[6]

Although we are not interested in the classification of the items, but in the actual consistency itself, it is interesting to explore the results of κ on our data. The literature suggests that 'agreement' is indicated by κ as follows: $> 0.8 =$ very good, 0.6 to 0.8 $=$ good, 0.4 to 0.6 $=$ moderate, 0.2 to 0.4 $=$ fair, $< 0.2 =$ poor.

There is already evidence (e.g. Table 2) that there was little agreement on jokehood in Study 1, and the κ value (for all the Study 1 jokehood data) is indeed extremely low: 0.053.

However, the κ figures for Study 2 demonstrate the way in which this measure can give counter-intuitive results when applied to skewed data. Applied to all the Study 2 data, $\kappa = 0.5173$, merely 'moderate'. This is slightly surprising, as inspection of the raw data shows there were clear majority verdicts for most items (as hinted at by Tables 5 and 6). The low rating is because the items had a predominance of texts which were constructed *not* to be jokes (HNS, SF, ANS), producing a skew in the judgements (overall, most items were judged as non-jokes). The effect is even more noticeable if we consider the types of text separately. The JK texts, all of which had overwhelming majority judgements, produce, when considered apart from the other three types, an abysmal κ value of 0.0150. If the JK and ANS data are combined – thereby creating a data set more balanced between 'probably J' and 'probably NJ' items – the κ score shoots up to 0.83 (very good). Thus κ appears to say that our judges agree very well on this combined set, but hardly agree at all on either half of it.

It is far from clear that the κ test is the appropriate test for our methodological question about consistency.

[6] The usual version of the κ test assumes that all judges rate all items, but it is straightforward to adjust the formulae for a situation (as here) where the set of judges rating an item varies.

5 CONCLUSIONS

It is hard to draw firm empirical conclusions from either of these studies, which are merely first attempts at probing the issues. In particular, it is unclear what is the correct methodological approach, especially regarding statistical tests. With those caveats, a few tentative observations can be made.

Study 1 does not show the expected consistency in judgements about jokehood. There could be a number of reasons for this. The most radical would be that this demonstrates a wider truth: that there is rarely agreement, even about jokehood, when people judge texts. A number of less sweeping excuses are also possible: perhaps this particular genre (punning riddles) is rather vulnerable to confusion about whether a text is a joke, or maybe young children, particularly when put in an experimental setting, find it difficult to make measured judgements about the concept of 'joke'. For funniness, the data does conform to the expectation that there is a wide variety of opinions; interestingly, the N (nonsense) items showed the greatest degree of agreement.

In Study 2, there is strongly suggestive support for the conjecture that judges are consistent in judging whether texts are jokes, particularly where the text has been constructed to be a joke. However, in view of the statistical difficulties outlined earlier, it is hard to claim that this is firmly corroborated. The funniness judgements behaved quite differently on texts constructed as jokes (where great variety did occur) from texts constructed as non-jokes (where there was much agreement).

The studies differed greatly in the type of texts and the judges used, which could contribute to the differing patterns of results.

Even if the hypotheses in both studies had been firmly established statistically, these are just two small studies, focussing on two very narrow text types and with different participant groups; this merely scratches the surface of the issue. It is also not clear whether any such results would be generalisable to further types of text. A claim that is universal across all texts cannot be proven by specific studies (although it could be refuted), but a large number of supportive studies would be highly suggestive.

If further studies supported the regularities shown in Study 2 about jokehood judgements, then it would be feasible to maintain the position outlined in Section 1 – that jokehood is a relatively stable concept amenable to testing with human judges. This would also mean that it would make sense to have an NLG system generate texts which were jokes; that is, this would be a well-defined and testable task. However, the variations in funniness judgements (for all texts in Study 1, and for joke texts in Study 2) suggest that the *effects* of supposedly humorous texts (on the user) might not be predictable. However, the analyses reported in [4] and [5] of the data items in our Study 1 did indicate that *on the whole* the set of computer-generated humorous texts were rated as more humorous than control items, even if no computer-generated text was given an overwhelming verdict of "joke" or "very funny". Similarly, analysis of the Study 2 data (in [10]) showed statistically significant differences between the ratings of the text types. Hence, the evaluation of the success of a humour-generating program could be measured in this aggregated or averaged form, rather than the ratings of individual items. Also, this pattern suggests that making a text "humorous" could be regarded not as a clear-cut attribute (as "syntactic well-formedness" might be in a model inspired by generative linguistics), but rather as a vaguer tendency. That is, a more realistic aim for an NLG system might be to take steps which will make it "more likely" that the text will be perceived as "humorous", rather than guaranteeing the humorous property – thus tackling the vaguer goal of "try to be more humorous" rather than the discrete goal of "create a joke".

We have focussed here on the possible difficulties of using conscious judgements to compare the humorous aspects of human and computer-generated texts (as was the main reason for the JAPE evaluation described in Section 2.1). However, there might be other methodologies which could be helpful. For example, some way of measuring genuine amusement (e.g. via facial expression [1]), or subsequent changes in mood (e.g. [16]), might be more reliable.

In spite of the inconclusive results, we believe that the methodological questions addressed here are worthy of consideration, and that we have at least made a start on investigating these questions.

ACKNOWLEDGEMENTS

The data collection, and most of the analysis of Study 2, took place while the authors were at the University of Edinburgh. KB's role in Study 1 was supported by a grant from the Natural Science and Engineering Council of Canada. The writing of this paper was partly supported by EPSRC grant EP/E011764/01. We are grateful to Hiram Brownell and Prathiba Shammi for the data used in Study 2.

REFERENCES

[1] Marian Stewart Bartlett, Joseph C. Hager, Paul Ekman, and Terrence J. Sejnowski, 'Measuring facial expressions by computer image analysis', *Psychophysiology*, **36**, 253–263, (1999).

[2] Anja Belz and Ehud Reiter, 'Comparing automatic and human evaluation of nlg systems', in *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, (2006).

[3] A.M. Bihrle, H.H. Brownell, J.A. Powelson, and H. Gardner, 'Comprehension of humorous and nonhumorous materials by left and right brain-damaged patients', *Brain and Cognition*, **5**, 399–411, (1986).

[4] Kim Binsted, *Machine humour: An implemented model of puns*, Ph.D. dissertation, University of Edinburgh, Edinburgh, Scotland, 1996.

[5] Kim Binsted, Helen Pain, and Graeme Ritchie, 'Children's evaluation of computer-generated punning riddles', *Pragmatics and Cognition*, **5**(2), 305–354, (1997).

[6] Jean Carletta, 'Assessing agreement on classification tasks: the kappa statistic', *Computational Linguistics*, **22**(2), 249–254, (1996).

[7] Justin McKay, 'Generation of idiom-based witticisms to aid second language learning', in *Proceedings of the April Fools' Day Workshop on Computational Humor*, eds., Oliviero Stock, Carlo Strapparava, and Anton Nijholt, number 20 in Twente Workshops on Language Technology, pp. 77–87, Enschede, Netherlands, (2002). University of Twente.

[8] Chris Mellish and Robert Dale, 'Evaluation in the context of natural language generation', *Computer Speech and Language*, **12**, 349–372, (1998).

[9] Nestor Miliaev, Alison Cawsey, and Greg Michaelson, 'Applied NLG system evaluation: FlexyCAT', in *Proceedings of 9th European Workshop on Natural Language Generation*. ACL, (2003).

[10] Robyn Munro, 'Empirical measurement of humorous effects'. 4th Year Project Report, School of Informatics, University of Edinburgh, 2004.

[11] Graeme Ritchie, *The Linguistic Analysis of Jokes*, Routledge, London, 2004.

[12] Willibald Ruch, 'Assessment of appreciation of humor: Studies with the 3WD humor test', in *Advances in personality assessment: Volume 9*, eds., Charles D. Spielberger and James N. Butcher, chapter 2, Lawrence Erlbaum, Hillsdale, NJ, (1992).

[13] P. Shammi and D.T. Stuss, 'Humor appreciation: a role of the right frontal lobe', *Brain*, **122**, 657–666, (1999).

[14] Sidney Siegel and N. J. Castellan, Jr., *Nonparametric Statistics for the Behavioural Sciences*, McGraw-Hill, 2nd edn., 1988.

[15] Oliviero Stock and Carlo Strapparava, 'The act of creating humorous acronyms', *Applied Artificial Intelligence*, **19**(2), 137–151, (2005).

[16] David Watson, Lee Anna Clark, and Auke Tellegen, 'Development and validation of brief measures of positive and negative affect: The PANAS scales', *Journal of Personality and Social Psychology*, **54**(6), 1063–1070, (June 1988).