Natural Language Engineering **22** (5): 727–749. © Cambridge University Press 2015
doi:10.1017/S1351324915000145

*Computational generation and dissection of lexical replacement humor**

ALESSANDRO VALITUTTI[1**] , ANTOINE DOUCET[2],
JUKKA M. TOIVANEN[3] and HANNU TOIVONEN[3]

[1]*Department of Computer Science, University College Dublin, Belfield, Dublin, Ireland*
e-mail: alessandro.valitutti@ucd.ie
[2]*L3i Laboratory, University of La Rochelle, avenue M. Crépeau, 17000 La Rochelle, France*
e-mail: antoine.doucet@univ-lr.fr
[3]*Department of Computer Science and HIIT, PO Box 68, FI-00014 University of Helsinki, Finland*
e-mail: jmtoivan@cs.helsinki.fi, hannu.toivonen@cs.helsinki.fi

(*Received 7 January 2014; revised 26 February 2015; accepted 2 March 2015; first published online 16 April 2015*)

Abstract

We consider automated generation of humorous texts by substitution of a single word in a given short text. In this setting, several factors that potentially contribute to the funniness of texts can be integrated into a unified framework as constraints on the lexical substitution. We discuss three types of such constraints: formal constraints concerning the similarity of sounds or spellings between the original word and the substitute, semantic or connotational constraints requiring the substitute to be a taboo word, and contextual constraints concerning the position and context of the replacement. Empirical evidence from extensive user studies using real SMSs as the corpus indicates that taboo constraints are statistically very effective, and so is a constraint requiring that the substitution takes place at the end of the text even though the effect is smaller. The effects of individual constraints are largely cumulative. In addition, connotational taboo words and word position have a strong interaction.

1 Introduction

In this paper, we study automated generation of humor by lexical replacement in a given text. We consider a setting where a short text, such as an instant message, is provided to the system, and the system makes the text funny by replacing one word in it. Unintentional examples of such humor are sometimes produced by autocorrection systems e.g. on mobile phones. Intentional generation of humorous slips could be used as a component of dialogue systems to make them more human-like, or they

*We would like to thank the anonymous reviewers for their insightful comments that have greatly helped us improve the paper. This work has been supported by the Academy of Finland (decision 276897, CLiC; and the Algorithmic Data Analysis Centre of Excellence, Algodan), and by the European Commission (FET grant 611733, ConCreTe; and FET grant 611560, WHIM).
**Most of the work was done while at the University of Helsinki.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

could be used as a source of inspirational material to professionals in the creation of catchy and memorable advertizing or newspaper headlines.

The task of intentionally introducing funny changes is conceptually simple but challenging in practice. It forms a convenient setting for research into computational generation of humor since it can be used as a testbed for several elementary components of lexical humor.

We consider various ways of performing a humorous lexical substitution. We formulate them in a modular and declarative manner by considering different constraints for substitutions that a system may choose to obey in various combinations. All of these constraints potentially affect the humor response.

We identify three different types of lexical constraints as possible building blocks of humor based on lexical selection or substitution. Our hypothesis is that constraints of each of these three types make the resulting text more humorous.

(1) The *form constraint* requires the substitute to be similar to the original word either orthographically or phonetically. This effectively turns the text into a pun.
(2) The *taboo constraint* requires the substitute to be a taboo word or a word used in taboo expressions. This is a well-known feature in some jokes.
(3) Finally, the *context constraints* require that the substitution takes place at the end of the text, or that the substitution is statistically consistent with neighboring words.

Our main contribution is not in the list of humorous constraints — our list is neither exhaustive, nor novel. Rather, our contributions are in (1) formulating explicit and operationalized constraints for lexical substitution humor in Section 4, (2) using conceptual distinctions inspired by theories of humor reviewed in Section 2, (3) implementing these constraints in software in Section 5, and especially in (4) evaluating the effects of the constraints empirically in a large user study in Section 6.

As will be discussed in Section 7, the empirical evaluation shows a statistically significant increase of humor response with the use of such humorous lexical constraints.

Our preliminary work in this direction was reported in a short paper by Valitutti *et al.* (2013). In this paper, we work at a much finer conceptual level and carry out a systematic empirical analysis, allowing us to draw novel conclusions about the use and effectiveness of humorous constraints.

2 Background on humor, incongruity, and taboos

We next review and define concepts that are central to our work: funniness, incongruity, and tabooness. They are the fundamental concepts in humor studies and inspired our choice of humorous constraints.

2.1 Overview of humor theories

There is an impressive number of theories about humor, most of which are traditionally classified in three main categories, referred to as Superiority Theory, Relief

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Theory, and Incongruity Theory (Morreall 2013). The assumption of *Superiority Theory* is that we laugh at the misfortunes of others and it reflects our own feeling to be better than others. According to *Relief Theory*, laughter is caused by the release of nervous tension or psychic energy. Finally, *Incongruity Theory* claims that the humorous effect is achieved by the induction of incongruity in a playful context. Our work mostly relates to the Incongruity Theory, and we will return to it below.

In recent years, several scholars have proposed more integrative visions to the theories of humor (Martin 2007). One approach is to consider the theories as complementary ones, each emphasizing a different aspect of humor: the stimulus (incongruity), the response (relief), or the social behavior (superiority). Another viewpoint is to consider the above concepts as related to different types of humor. For example, the 3 WD taxonomy of Ruch (1992) treats incongruity resolution as a specific type of humor distinct from nonsense humor and sexual humor, the latter more directly connected to relief.

2.2 Humorousness and funniness

Humor is not a well-defined concept. An issue with the term 'humorous' is that it can be associated either to the stimulus or to the response. It can indicate an objective property of a text (i.e. a joke) or the subjective experience of the reader.

The *subjective* experience is the psychophysiological *response*, generally known as *humor appreciation* or *humor recognition* (Ruch 2008), induced by the reading or hearing of a text. It is often characterized by responses such as laughter, smiling, or mirth.

In contrast, the *objective* aspect of humorousness is the text's potential to induce humor appreciation. According to Carrell (1997), 'Humorousness is a binary category representing a stimulus' theoretical capacity to induce a humor response, while funniness is a gradable category indicating the degrees of appreciation of a humorous text, differently perceived by individuals'.

This distinction between subjective and objective aspects of humor is important not only for theoretical treatment of humor, but especially for computational generation of it. Given an output of a humor generator, its 'humorousness' could be defined in terms of the number of subjects judging it as funny, while 'funniness' could be defined by how funny they think the text is.

One of our aims in this work is to build and evaluate an actual system for producing humorous texts. In the evaluation of the system, we will focus on funniness.

We define *funniness* of a text as a numerical value (on some fixed scale) indicating the (average) degree of humor appreciation of the text, differently perceived by individuals.

2.3 Incongruity

Incongruity is a central notion in humor studies. The term 'incongruity' is used to indicate either a group of theories about humor (collectively referred as *Incongruity*

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Theory) or a specific type of humor (called *incongruity humor*). While there is no general agreement on how to define incongruity, most definitions are related to the perception of incoherence, semantic contrast, or inappropriateness occurring while interpreting a situation or reading a text.

One of the first descriptions of the concept of incongruity was provided in the eighteenth century by the Scottish poet Beattie (1971), according to whom laughter 'always proceeds from a sentiment or emotion, excited in the mind, in consequence of certain objects or ideas being presented to it'. Koestler (1964) defined 'bisociation' as 'the perceiving of a situation or idea in two self-consistent but habitually incompatible frames of reference'. According to contemporary psychologists such as Suls (1972), the humorous effect is not due to the perception of incongruity, but to the resolution of incongruity.

In the influential Semantic Script Theory of Humor (SSTH), Raskin (1985) hypothesizes that a humorous text 'is compatible, fully or in part, with two different scripts. The two scripts with which the text is compatible are opposite'. Here, 'script' is a structured configuration of knowledge about some stereotyped or familiar situation or activity (Schank and Abelson 1977). Although Semantic Script Theory of Humor is focused on semantic (and not perceptual) aspects of humor, it indirectly shows interesting connections between semantic properties of humorous texts and their possible use to induce and resolve incongruity. In particular, the perception of semantic opposition could be considered as a kind of bisociation.

Humorous texts, and particularly jokes, are created to induce incongruity and resolution intentionally. The first part of the joke (the *setup*) is generally used to induce some form of expectations, and the second part (the *punchline*) is used to violate them, producing incongruity and then resolution. In a specific type of jokes, a mechanism of 'forced reinterpretation' is used. The setup is ambiguous, but the non-humorous interpretation is so obvious that the ambiguity is often missed during the setup. Then, however, the punchline is only consistent with the other (funny) interpretation, forcing the audience to discover it. An example of this form of jokes is: *'Alcohol isn't a problem, it's a solution... Just ask any chemist'*.

In his analysis of forced reinterpretation jokes, Ritchie (2002) emphasizes a distinction between three different elements of joke processing. (1) *Conflict* is the initial perception of incompatibility between punchline and setup according to the initial obvious interpretation. (2) *Contrast* is perceived between the initial and the funny interpretation. (3) *Inappropriateness* or intrinsic oddness occurs in the funny interpretation. All three concepts are often connected to the notion of incongruity.

Each element is associated to a different linguistic device and, correspondingly, to a different type of incongruity. Linguistic incoherence is a generator of *conflict* and ambiguity is a generator of *contrast*. Finally, inappropriateness introduces a form of 'residual incongruity' that can act as a humorous cue and make the second interpretation be perceived as funny. Humor cues indicate that the appropriate frame is one of play rather than seriousness (Mulkay 1988). In our method for humor generation, all these linguistic devices are implemented as lexical constraints.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

2.4 Taboos

Taboos are behaviors or expressions 'to be avoided' because of social disapproval. In verbal expressions, typical taboo topics include racism, Holocaust (Cory 1995), sexuality, death, sickness and disability, blasphemy, deformity, and bodily functions.

A special case of taboo behavior consists of the pronunciation of inappropriate words called *taboo words*. These are words normally avoided in everyday conversations because they are potentially offensive or embarrassing. They can be either words referring to taboo topics or strongly connotated words such as profanities or slang terms (Jay, Caldwell-Harris and King 2008).

Taboo words are often used to produce humor effects. They can be related, to different extent, to each theory type discussed at the beginning of this section. For example, they can be uttered to embarrass the hearer and, next, release tension with a proper punchline (*relief*). This is mainly the case of stand-up comedy (Seizer 2011). Another humorous effect consists of the intentional use of taboo words to simulate speech errors such as malapropisms and spoonerisms (Zwicky 1979). In this case, they are employed to emphasize ridiculous aspects of people and, thus, laugh at them (*superiority*). Finally, as suggested above, taboo words can be used as humorous cues in order to make incongruity funny (*incongruity*).

With regard to taboo words, in this research we study the extent to which their use can increase the humorous effect introduced by other constraints (e.g. phonetic similarity in lexical replacement). At this stage, we do not investigate whether this contribution is due to their role as possible humorous cues or to the superposition of taboo (either relief or superiority) humor.

3 Previous humor generation systems and evaluations

To date, there are only a small number of published methods for computational generation of humorous text, and only few have been empirically evaluated. Each work has focused on a specific type of humorous expressions. The interested reader is referred to a systematic review of the most important humor generators by Ritchie (2003).

Almost all humor generation systems are focused on the production of small texts, such as puns or simple jokes. Raskin and Attardo (1994) describe LIBJOG, a program for the generation of light bulb jokes. McKay (2002) presents WisCraic, a system capable of producing a wider range of puns, including the form of question-answering. Tom Swifty, a system by Levison and Lessard (1992), and HCPP (Homonym Common Phrase Pun) by Venour (1999), produce a quoted utterance joined to a funny remarking phrase. The JAPE program (Binsted, Pain and Ritchie 1997) produces specific types of punning riddles, and the HAHAcronym (Stock and Strapparava 2003) automatically generates humorous acronyms.

A few systems exploited the use of taboo words as a way of addressing forms of adult humor. In particular, two humor generators employing taboo words are the punning riddle generator by Sjöbergh (2006) and the FEVer program by Valitutti (2011), based on variation of a familiar expression by lexical substitution.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Recent attempts to advance humorous generation take advantage of statistical corpus-based techniques widely used in computational linguistics. For instance, Petrović and Matthews (2013) focused on a specific pattern of jokes and applied an unsupervized machine-learning method to a large-scale textual corpus. More recently, Veale (2013) provided a detailed description of corpus-based information retrieval as a way of producing ironic simili.

Two interesting systems use generation of humorous texts as a function in a larger context: Dybala *et al.* (2008) exploited the use of simple forms of pun generation in a conversational context, while Özbal and Strapparava (2012) employed humorous punning as a specific strategy for creative naming.

Empirical evaluations of the above-mentioned humor generation systems vary considerably. A major limitation for systematic and comparative studies is the lack of a generalizable definition of the performance of a humor generator. As a consequence, there is no methodology of evaluation general enough to allow us to compare the performance of different systems or of different versions of the same system.

4 A constraint-based method for humorous lexical substitution in short texts

We now propose a method for the computational generation of humorous verbal expressions as variations of existing texts by a single word substitution. The method we propose is declarative in the sense that it is essentially based on a number of constraints for the substitution. In this section we give an overview of the setting and the different types of constraints, while the next section provides details on how we implemented the constraints in an actual, running system.

The method takes as input a short English text (e.g. *'I've done my part'*). Examples of such short texts include status updates in social media, Twitter tweets, and mobile phone (SMS) messages. The method then performs a single word substitution (e.g. *'part'* → *'fart'*), and returns the resulting text (*'I've done my fart'*). Just a single substitution is carried out since the text is assumed to be short and the intended effect is to turn the text into a funny one-liner.

In the rest of this paper, we propose lexical constraints that contribute to the funniness of such substitution humor and study their effects.

We consider three different types of lexical constraints:

- *Form constraints* are related to the morphological, phonetic, and syntactic forms of the original and the substitute word.
- *Semantic/connotational constraints* are related to the meanings or emotional connotations of the words. For instance, the words could be selected according to some prefixed topic or linguistic usage.
- *Contextual constraints* concern the position and textual context of the substitution.

In the next paragraphs, we outline the more specific constraints that we use within the three categories.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

4.1 Punning using form constraints

We use a form constraint, denoted in the sequel by FORM, to turn the given text into a pun. The similarity can be either sound or spelling-based and does not require exact identity. Hempelmann (2003) calls these *paronomasic puns*.

- *Constraint* FORM *requires that the original word and its substitute are similar in form, either phonetically or orthographically.*

The use of a phonetically or orthographically similar word as a replacement makes the resulting text pseudo-ambiguous, since the original intended meaning can often also be recovered. There are then two 'conflicting' and 'contrasting' interpretations — the literal one and the original one — increasing the likelihood of humorous incongruity.

4.2 Taboo humor using semantic/connotational constraints

A classic example of a semantic/connotational constraint is taboo humor. In our setting, it can be implemented by requiring that the substitute word is a taboo word. For additional granularity, we use two types of taboo words, connotational and taboo-inducing.

Connotational taboo words are unspeakable words where the taboo is in the utterance itself, not necessarily in the topic. These are the typical taboo words; according to Jay *et al.* (2008), 'Taboo words have uniquely strong connotative meanings'. Typical examples include profanities such as 'shit' and 'faggot'. The connotation can be a strong negative opinion toward some people, making the term politically incorrect.

We define *taboo-inducing words* as words capable to induce taboo meanings when used as a substitute in a short text, while not necessarily being taboos in themselves. We identified and collected three types of words that potentially are taboo-inducing:

(1) Words that denote taboo topics (e.g. sex, diseases, bodily functions, physical defects, etc.), even though they are not emotionally connotated and are not considered to be forbidden words in themselves. For instance, 'vagina' and 'excrement' refer to topics that are taboos. Depending on the context, the use of these words can be either appropriate or not.
(2) Words denoting animals, used as insults when directed to other people (e.g. 'pig' or 'whale' when used to insult an oversize person) (Leach 1964).
(3) Neutral words capable of implying intimacy or allusion when used in an expression (e.g. 'wife' in the expression *Let everything turn well in your wife!*). Examples of allusive meaning of words and phrases according to the context are discussed elsewhere (Sherzer 2002).

Accordingly, we have two constraints for taboo words:

- *Constraint* CONNOT *requires that the substitute word is a connotational taboo word.*
- *Constraint* TABIND *requires that the substitute word is a taboo-inducing word.*

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

We say that CONNOT and TABIND are both TABOO constraints, and that the TABOO constraint is satisfied if either specific condition is satisfied. In terms of humor theory, taboo words or expressions directly introduce 'inappropriateness' to the text.

4.3 Local coherence and forced reinterpretation using contextual constraints

The context and position of the substitution can have subtle effects. If the substitute word does not form a coherent or likely compound with its immediate predecessor or successor, then the text is likely to make little sense. On the other hand, if the modified text does make sense, then the changed interpretation, possibly a taboo one, is potentially expanded to the phrase level. We hypothesize that this introduces a stronger semantic or connotational 'contrast' and thus probably contributes to making the text funnier.

We measure such local coherence using *n*-grams:

- *Constraint* NGRAM *requires that the substitute and its neighboring word(s) form a coherent compound (see Section 5.3 for details).*

The semantic/connotational contrast is potentially even stronger if the changed word comes as a surprise at the end of a seemingly innocent text. The humorous effect then is similar to the one of the forced reinterpretation jokes. We formulate this as a constraint regarding the position of the substitution:

- *Constraint* WDPOS *requires that the substitution takes place among the last words of the text.*

We say that NGRAM and WDPOS are both CONTEXT constraints.

4.4 Comparison to previous work

Similar constraints are implemented in various ways in the systems reviewed above in Section 3. In particular, all of them perform a selection according to part of speech, in order to produce well-formed expressions. JAPE (Binsted *et al.* 1997), HAHAcronym (Stock and Strapparava 2003) and FEVer (Valitutti 2011) implement phonetic controls either in terms of phonetic distance or rhyme. Semantic/connotational constraints are expressed as antonymy or domain opposition (JAPE, HAHAcronym) or taboos (FEVer). Contextual constraints are implemented in JAPE and HAHAcronym in terms of template and grammatical rules, respectively. Finally, while we are not aware of *n*-grams being used before for humor generation, they have been discussed as a feature for humor recognition (Taylor and Mazlack 2005).

5 Implementation

The goal of our implementation is to allow testing the effects of the humorous constraints in a systematic manner, not to maximize funniness. We will return to the latter topic in Discussion.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

We next describe our exact computational implementation of the proposed constraints for the English language.

5.1 Form constraints

The form constraints aim to make the word substitution recognizable as a pun. We adopt a definition of punning including both phonetic and orthographical similarity.

More exactly: Two words are considered *orthographically similar* if one word is obtained from the other one with a single character deletion, addition, or replacement. We call two words *phonetically similar* if their phonetic transcriptions (in CMU pronunciation dictionary) are orthographically similar according to the above definition. The definition of both orthographic and phonetic similarity can be considered a simple implementation of Levenshtein distance (Levenshtein 1966).

The FORM constraint is satisfied if :

(1) the substitute word is either a noun, an adjective, a verb, or an adverb (analyzed using TreeTagger[1]), and
(2) the words are orthographically or phonetically similar, as defined above.

In each replacement, both the original word and its substitute are constrained to be English words. As a collection of English words, we used WordNet lexical database (Fellbaum 1998) and the CMU pronunciation dictionary[2]. The latter also provides a collection of words not normally contained in standard English dictionaries while commonly used in informal language. This increases the space of potential replacements.

5.2 Taboo constraints

For the experiments of this paper, we collected a sample of 698 different taboo words using three partially overlapping sources and manually annotating the words.

To obtain a representative sample of different taboo words, we used sources that presumably have different ratios of taboo-inducing and connotations taboo words. The first of them is a website that lists examples of funny auto-correction mistakes[3]. In many cases, the word substitution generates a taboo meaning. The second source was the domain SEXUALITY of WordNet-Domains (Magnini and Cavaglià 2000), presumably including taboo words of both classes. Finally, as a source of words commonly used as connotational insults, we used two online dictionaries of slang terms[4]. In total, we obtained 768 words at this phase.

We then manually checked and annotated the obtained words according to the extended definition provided in Section 4.2, and removed the words not recognizable as either connotational or taboo-inducing words. While such classification may vary across age groups and cultural backgrounds, we believe this division is useful in

[1] http:www.ims.unistuttgart.de/projekte/corplex/TreeTagger
[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[3] http://www.damnyouautocorrect.com
[4] http://www.urbandictionary.com and http://onlineslangdictionary.com

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

classifying taboo words (and experiments in Section 6 will empirically confirm a difference between them). For our tests, one annotator performed the manual check, and two other annotators made a double check, even though no measurement of inter-annotator agreement was performed. Finally, we obtained a list of 184 connotational words and a list of 514 taboo-inducing words, i.e., 698 words in total.[5]

5.3 Context constraints

The context constraints consider the context or the position of the substitution.

The NGRAM constraint is aimed to increase local lexical coherence by requiring that the substitute word and one to two words preceding it (i.e., its left-context) have a statistically relevant joint probability. To obtain such information, we use a vast collection of *n*-grams, extracted from the 2012 Google Books *n*-grams collection[6] (Michel *et al.* 2011) . In the experiments our focus is on contemporary English, so we only take into account the statistics of Google *n*-grams stemming from books published after 1990. For further consistency, we here only use the subset of American English. These choices give forty-one million distinct 3-grams. While this number may seem enormous, it still produced very few replacement candidates for most two-word left-contexts in the SMS corpus we used, see below. This made the use of 2-grams essential with this dataset, as with them the number of possible replacements was much larger, with the downside of a shortened, less restrictive, left-context.

We computed the cohesion of each *n*-gram by comparing its expected frequency, assuming word independence, to its observed number of occurrences. A subsequent Student t-test allowed us to assign a measure of statistical cohesion to each *n*-gram (Doucet and Ahonen-Myka 2006). The NGRAM constraint is satisfied in a given text if the cohesion of the *n*-grams composed by the substitute word and the previous words is positive.

The word position (WDPOS) constraint is the requirement that only one of the last two words in the text may be substituted.

6 Empirical evaluation

We now move on to describe the empirical evaluation of the proposed lexical constraints. We aim to address two research questions.

(1) Primarily, we aim to dissect the effects of the proposed lexical constraints. Our hypothesis is that they increase the funniness of texts.
(2) A secondary goal is, keeping practical applications in mind, to see how well a fully automated system could produce humorous texts using a single lexical replacement without practically any analysis of the original text. An answer to this question will be extrapolated from the results for the first question.

[5] The taboo word lists are publicly available at http://www.cs.helsinki.fi/research/discovery/data/taboo/
[6] http://books.google.com/ngrams

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

First, however, we will empirically validate our lists of connotational and denotational taboo words.

6.1 Validation of taboo word classification

Because our division of taboo words into connotational ones and taboo-inducing ones is unconventional, we validated this split with a crowdsourcing experiment using CrowdFlower[7], a crowdsourcing service.

Asking subjects directly about the taboo word class is not feasible in practice due to the abstract nature of the concepts. Instead, the validation was designed so that subjects simply need to indicate if various uses of a given word are acceptable or not. The questions were designed to correspond to the hypothesis that taboo-inducing words induce a taboo only in specific contexts. This is why we provided context to each example sentence of the experiment. Furthermore, one question directly asks if the word is a profanity. Table 1 shows the questions and the included patterns for the word 'banging'. Question 1 (*Likability*) checks if it is ok to like the referent of the word. If the denotation of the word is taboo, or if liking it is taboo, e.g. because of intimacy, then the answer is negative. Question 2 (*Profanity*) directly checks if the word is a connotational taboo word. Question 3 (*Discussibility*) asks if it is ok to discuss the topic officially in public, and Question 4 (*Advisability*) if it is ok in a special context such as a medical consultation (even if the word may have a taboo meaning).

The dataset for this experiment was built as a randomized list of connotational words, taboo-inducing words, and non-taboo words such as 'book' or 'table'. The latter group was not the topic of interest here, but a small sample of non-taboo words was included to provide the subjects a wider range of words to assess.

The evaluation task was proposed to subjects as a list of four yes-no questions as described above. Each word (and corresponding list of questions) was evaluated by 40 different subjects. However, since the crowdsourcing platform performed random assignment of words and subjects, each word was rated by a different selection of subjects. For this reason, we did not perform a measurement of inter-rater agreement with multiple raters. Instead we cumulated, for each word, the answers of all subjects and selected the most frequent answer to each question (called the *crowd (majority) judgement*). For example, Question 2 for the word 'smell' received 2 'Yes' votes and 38 'No' votes, so the crowd judgement is 'No, this word is not profanity'.

We then defined words with a crowd judgement 'Yes' to Question 2 as connotational taboo words, and words with at least one crowd judgement 'No' to any of Question 1, 3, or 4 as taboo-inducing words. For instance, the word 'fart' received the crowd judgement 'Yes' to Question 2 and, thus, it was classified as a connotational word. On the other hand, 'condom' received the crowd judgement 'No' to Question 2 (i.e., it is not a profanity), 'Yes' to Questions 3 and 4 (i.e., it can be appropriate

[7] http://www.crowdflower.com

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Table 1. *Example questions used to assess the tabooness of words*

Variable	Question in crowdsourcing (example word: 'banging')
Likability	Would it be ok to say something like 'I love banging' or 'Do you like banging?' to a stranger?
Profanity	Is the word 'banging' a bad language/vulgar/swearing/ profanity word?
Discussibility	Imagine a politician saying this in public: 'We need some serious discussion about banging'. Would it be acceptable?
Advisability	Imagine a patient saying to his/her physician or psychiatrist: 'I have problems with banging. Could you advise me?'. Would it be acceptable?

in some contexts), but 'No' to Question 1 (i.e., it can be allusively inappropriate in some contexts). It is thus classified as a taboo-inducing word.

Since, the crowd judgment on taboo word classification is based on heuristic rules about the acceptability of their uses, the crowd judgement is not the ground truth. We believe our manual classification may actually be more reliable and use it in the experiments of this paper.

However, the crowd judgement is useful as a proxy to validate our manual classification. To compare our manual annotation and the crowd judgements, we measured Cohen's Kappa coefficient of inter-rater agreement between them. The coefficient is 0.445 ($p < 0.001$), showing a moderate positive correlation between the two ratings. This indicates that our annotation of taboo word classification is supported by the empirical, indirect validation. The main experiments below also show different behaviors for the two classes of taboo words.

6.2 Test setting

When designing the test environment for lexical replacement humor, we made two central decisions. The first one was to use real SMS messages as the texts to be modified using the proposed constraints. This choice reflects our desire to make the test setting as realistic as possible. Second, we crowdsourced the assessment of humor responses to a large number of independent subjects, for a large and unbiased evaluation.

Given the large number of possible constraint combinations as well as the monetary costs of crowdsourcing their evaluations, we also needed to be selective about the number of tests to be carried out. We decided to take punning (i.e., constraint FORM) as the context of all the experiments and to study the effects of other constraints and combinations only in the presence of FORM. This essentially halves the number of tests required (as tests without FORM are not needed) and allows us to focus on more interesting phenomena.

We next describe the evaluation setting in detail.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

6.2.1 Preparation of humorous texts

As input texts, we employed *NUS SMS Corpus*[8], a collection of 10,116 real SMS messages (Chen and Kan 2013). The texts make heavy use of abbreviations and colloquial expressions, providing a challenging but realistic case for linguistic processing.

As pre-processing, part-of-speech tagging was applied on all sentences in the corpus. Since, we chose to only consider substitutions of nouns, adjectives, verbs, and adverbs, we discarded messages where none was recognized, leaving us with 8,602 messages for test use.

To systematically study the effects of the constraints, we applied them in all twelve possible combinations containing FORM: with no other constraint, any one of the other four constraints, any feasible pair of the other four constraints (note that the TABOO constraints CONNOT and TABIND are mutually exclusive), and the fullest possible sets of constraints, containing three of them in addition to FORM. We call these combinations of constraints the different *experimental conditions* of our study.

Based on preliminary tests, we identified a sample size of seventy-five texts per experimental condition as a suitable compromise between statistical power and crowdsourcing expense. To obtain such samples, we first carried out every possible lexical replacement under each of the experimental conditions, one at a time, so that the resulting messages have exactly one word substituted. We then randomly picked seventy-five modified messages for each of the experimental conditions. However, in the two conditions consisting of four simultaneous constraints (FORM+CONNOT+NGRAM+WDPOS and FORM+TABIND+NGRAM+WDPOS), the system was only able to generate twenty and seventeen examples, respectively, so the total number of messages tested is 787. Table 2 shows an example output of the humor generator under each of the experimental conditions.

6.2.2 Crowdsourcing

We evaluated the constraints empirically using CrowdFlower. Crowdsourcing allows performing the evaluation with a high number of subjects, also for tasks with a high cognitive cost. This has been successfully applied, for instance, to the evaluation of information retrieval over digital libraries (Kazai *et al.* 2011). We are aware of only one previous use of crowdsourcing for assessment of humorous texts (Hempelmann *et al.* 2012) . Unlike the present work, it was performed on texts collected from the web and not produced automatically.

We recruited subjects from English-speaking countries using the crowdsourcing service. We did not control the age of subjects. The majority of crowdsourcing users are known to be between eighteen and thirty-five years old (Ross *et al.* 2010)

and they thus are suitable subjects for tests concerning mobile messaging and taboo humor. Each subject was assigned a random set of twenty messages out of the total of 787 messages. Some subjects chose to evaluate several such sets; the

[8] http://wing.comp.nus.edu.sg/SMSCorpus

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Table 2. *Examples of outputs of the system, each generated under the corresponding experimental condition. Recall that the* FORM *constraint (word similarity) can be based on orthographic or phonetic similarity*

Experimental Condition	Text Generated by the System	Original Word	Substitute Word
FORM	*Harpy birthday!*	happy	harpy
FORM+CONNOT	*Hi, come back homo now*	home	homo
FORM+TABIND	*Now u makin me more curious... smell me pls...*	tell	smell
FORM+NGRAM	*Me kill in singapore.super best hor.*	still	kill
FORM+WDPOS	*Tmr u going to school? I meet u in pool?*	school	pool
FORM+CONNOT +NGRAM	*Which fart of town would you be in?*	part	fart
FORM+CONNOT +WDPOS	*Later we go where to shit?*	eat	shit
FORM+TABIND +NGRAM	*Is ur paper today in e porn or aft?*	morn	porn
FORM+TABIND +WDPOS	*Can sell e book.how much u wan smell?*	sell	smell
FORM+NGRAM +WDPOS	*Remember to get the phone book from cat person.*	that	cat
FORM+CONNOT +NGRAM+WDPOS	*Remember to get the phone book from fat person.*	that	fat
FORM+TABIND +NGRAM+WDPOS	*Ok i am leaving now pee u*	see	pee

total number of subjects evaluating the messages was 524. Then, thirty additional messages were designated as 'gold units', i.e., messages not included in the evaluation, but preliminarily used by the crowdsourcing service to check which users can be trusted (as further explained below). The crowdsourcing evaluation was running until each message was judged by at least ninety different subjects.

We asked the subjects to assess each individual message for its funniness on a scale from 0 to 4, as follows. First, they were required to judge if the text is funny or not and select 0 in the latter case. Next, if the text was considered funny, they were asked to score its funniness with a number between 1 and 4. In this way, we collected judgements that allow analysis of humor either as a Boolean or a graded variable.

In order to identify and remove potential scammers in the crowdsourcing system, we simply asked subjects to select the last word in a given message. If a subject

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

failed to answer correctly in at least three gold standard units, all her judgements were removed by the crowdsourcing system. As a result, 1.4% of judgments were discarded as untrusted. This obviously is a suboptimal method to remove scammers and may weaken the statistical power of the results.

From the experiment, we have a total of 70,848 trusted assessments of messages; 90 assessments of each of the 787 test messages plus 18 extra assessments.

6.2.3 Evaluation measure

Since, we are interested in the effects of different experimental conditions rather than individual texts, we consider statistics over all texts produced under the same condition. More specifically, we measured the *funniness*, i.e. the intensity of humor response, defined as the mean score of texts generated under a given experimental condition.

6.3 Results

6.3.1 Humor responses under different experimental conditions

Table 3 gives the average and the standard deviation of funniness under each of the twelve experimental conditions. Consider FORM as a baseline condition (first row of the table). The mean funniness is 0.55, indicating that on average the texts can hardly be called funny.

On the following rows, we additionally list the absolute and relative improvements in funniness when compared to the plain FORM constraint, to make the effects of different constraints more explicit. For instance, FORM+CONNOT together have mean funniness of 0.79 (standard deviation 0.23), which is 0.24 higher than the mean funniness of FORM. In relative terms, CONNOT increases the mean funniness by 44%. A general observation taken from the table is that the measures tend to improve as constraints are added, with CONNOT and TABIND contributing most; a statistical analysis will be given below. The best results were obtained under condition FORM+CONNOT+NGRAM+WDPOS (second row from the bottom). The mean funniness is 0.92, indicating a 67% increase in funniness, even if the absolute number still is not high. We next study the effects in more detail.

6.3.2 Regression analysis of the effects of humorous constraints

To analyse the magnitudes and statistical significances of the effects we fitted a linear regression model to the data, regressing funniness on the presence of the four constraints (CONNOT, TABIND, NGRAM, WDPOS). The analysis was performed in *R*, with dummy coding of the constraints as independent binary variables without any interaction terms.

The magnitudes of effects are shown in the top part of Table 4. The column 'Absolute effect' gives the regression coefficients of constraints, the column 'Relative effect' the relative sizes of the coefficients with respect to the plain FORM constraints (cf. relative improvements in Table 3).

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Table 3. *Mean funniness values and their standard deviations over texts under different experimental conditions; the absolute and relative improvements in terms of mean funniness over the plain* FORM *constraint*

Experimental Condition	Funniness (mean and std.dev.)	Absolute improvement over FORM	Relative improvement over FORM
FORM	0.55 ± 0.11	–	–
FORM+CONNOT	0.79 ± 0.23	0.24	44%
FORM+TABIND	0.70 ± 0.19	0.15	27%
FORM+NGRAM	0.58 ± 0.10	0.03	5%
FORM+WDPOS	0.58 ± 0.14	0.03	5%
FORM+CONNOT+NGRAM	0.78 ± 0.24	0.23	42%
FORM+CONNOT+WDPOS	0.91 ± 0.32	0.36	65%
FORM+TABIND+NGRAM	0.73 ± 0.21	0.18	33%
FORM+TABIND+WDPOS	0.73 ± 0.21	0.18	33%
FORM+NGRAM+WDPOS	0.64 ± 0.23	0.09	16%
FORM+CONNOT+NGRAM+WDPOS	0.92 ± 0.29	0.37	67%
FORM+TABIND+NGRAM+WDPOS	0.79 ± 0.25	0.24	44%

The CONNOT and TABIND constraints have more substantial effects of 0.26 (47%) or 0.15 (27%), respectively, while those of WDPOS and especially NGRAM are more subtle. Statistically, constraints CONNOT, TABIND, and WDPOS are all highly significant whereas the NGRAM constraint is not (column 'Signif. level (linear regr.)').

In the linear model, constraints CONNOT and WDPOS have larger effects than they have individually in the empirical observations (Table 3). For easier comparison of the predictions made by the linear model and the empirical observations, in the bottom part of Table 4 we list the cumulative effects of different combinations of constraints as predicted by the linear regression model, i.e., by simply summing up the absolute effects from the top part of Table 4.

Given the differences between predictions and observations, related especially to CONNOT and WDPOS, we additionally fitted a linear regression model with pairwise interaction terms included (results not shown). In this case, the interaction between the constraints CONNOT and WDPOS is indeed statistically significant ($p < 0.05$), indicating that their combination (a connotational taboo word at the end of a text) is more powerful than either one alone. No other pairwise interactions are statistically significant.

6.3.3 Permutation-based analysis of effects of humorous constraints

We also carried out a permutation-based statistical analysis of the effects of the constraints on funniness of texts. We implement the null hypothesis that a constraint has no effect by randomly swapping results between two sets as follows. Results obtained with the constraint being tested and some set of other constraints are swapped with those obtained with the exact same set of other constraints but *not* the constraint being tested. We then compute the difference in mean funniness when using the constraint or when not using it while keeping the other constraints

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Table 4. *Absolute sizes of the effects of the constraints, their relative effects, and their statistical significance levels. The significance levels have been corrected for multiple testing using Bonferroni correction (linear regression) or Westfall-Young correction (permutation tests). The cumulative effects are derived from the linear regression model by summing the absolute effects*

Constraint	Absolute effect (linear regr.)	Relative effect wrt FORM	Signif. level (linear regr.)	Signif. level (perm. test)
CONNOT	0.26	47%	$p < 10^{-5}$	$p < 10^{-5}$
TABIND	0.15	27%	$p < 10^{-5}$	$p < 10^{-5}$
NGRAM	0.03	5%	–	$p < 0.05$
WDPOS	0.07	13%	$p < 10^{-5}$	$p < 10^{-5}$
	Cumulative Effects (predictions by linear regr.)			
CONNOT+NGRAM	0.29	53%		
CONNOT+WDPOS	0.33	60%		
TABIND+NGRAM	0.18	33%		
TABIND+WDPOS	0.22	40%		
NGRAM+WDPOS	0.10	18%		
CONNOT+NGRAM+WDPOS	0.36	65%		
TABIND+NGRAM+WDPOS	0.25	45%		

constant, and average the result over all combinations of the other constraints. The sizes of effects obtained this way from the original data are almost identical with the factors obtained using linear regression (cf. Table 4, leftmost column in the top part).

Comparing the sizes of effects to those obtained with randomly permuted datasets, we obtain empirical p-values for each of the constraints (Table 4, rightmost column in the top part). They are corrected for multiple testing using the Westfall-Young procedure (Westfall and Young 1993). This permutation test has more statistical power than regression analysis, and indicates that NGRAM is also statistically significant at level $p < 0.05$ (after correction for multiple testing) despite its small absolute effect.

7 Discussion

As already explained above, we chose an extremely simple task of humor generation (i.e. single word substitution). The punning function implemented for this study was even simpler than the one used in our previous work (Valitutti *et al.* 2013). In fact, as formal constraints we did not use rhymes but only orthographic and phonetic similarity, both implemented through a simple implementation of Levenshtein distance. Our experiments have been performed in an artificial test setting (crowdsourcing) but using a real corpus of SMSs — with all their issues, especially the use of colloquial language and abbreviations. In our tests, a fully automated system processed the SMSs and produced the texts used in the experiments, without any curation before submitting them to evaluation on the crowdsourcing platform. All our results should be interpreted with the challenging context of real SMSs and crowdsourcing evaluation in mind.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

7.1 Effects of constraints

The empirical results (Tables 3 and 4) give the following answers to our research question about the effects of the lexical constraints:

- Taboo constraints CONNOT and TABIND are effective in increasing the average funniness. The absolute sizes of the effects are modest, but the relative ones over plain punning (constraint FORM) are substantial: 47% increase of funniness with connotational taboo words and 27% increase with taboo-inducing words. Both constraints are statistically highly significant. According to the results, using taboo words as substitutes thus seems to successfully introduce inappropriate incongruity to the text (cf. Section 2.3). According to the theory, it may embarrass the reader, ridicule the writer, or simply characterize the context as playful.
- Context constraints NGRAM and WDPOS produce subtler effects. The word position constraint gives 13% increase and the *n*-gram constraint 5% increase on funniness. The effect of word position is statistically highly significant but the significance of the *n*-gram constraint is questionable (significant in a permutation-based test, insignificant in linear regression). In any case, the result confirms that simple constraints other than taboo words can be used to induce greater funniness.
- The effects of the constraints are largely cumulative (with one exception, see below): using more than one constraint at the same time increases the overall funniness approximately according to the sum of the individual effects.
- There is a statistically significant interaction between the CONNOT and WDPOS constraints: while connotational taboo words are funny, they are especially funny at the end of the text. Given how strong the statistical effect of CONNOT already is (44% observed improvement over plain punning), it is impressive that the addition of WDPOS can further improve the effectiveness in a statistically significant way (to 65% improvement over punning). Constraining the substitution to take place at the end of the message thus seems to successfully induce a humorous surprise effect (the contrast incongruity type) especially with connotational taboo words.
- Finally, a negative result somewhat contrary to our expectations is that no significant interaction was discovered between taboo-inducing words and contextual constraints. By definition, taboo-inducing words are sensitive to the context where they are used. Obviously, the context constraints used here (WDPOS, NGRAM) are not particularly useful in providing appropriate contexts for taboo-inducing words. As a result, the taboo-inducing words have smaller effects than connotational taboo words.

In a nutshell, taboo words can make a text recognizable as humorous. Contextual constraints have subtler effects, but the word position constraint clearly helps induce a humorous surprise effect especially in the case of connotational taboo words.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

7.2 Application of humorous constraints

Consider a system that aims to generate humorous texts using lexical substitution. How can the ideas and results from this paper be used for practical applications? To address the question, let us first classify the contributions of this paper into two categories.

First, we have described the implementations of five specific lexical constraints (FORM, CONNOT, TABIND, NGRAM, and WDPOS) and empirically analyzed their effects on funniness when applied on SMSs (Tables 3 and 4). These specific techniques and results are immediately applicable in automated humor generators working on SMSs or other similar short texts such as social media status updates.

Second, and more importantly, (1) we have shown how different factors that potentially contribute to the funniness of a text can be integrated into a unified framework as constraints on lexical substitution, (2) we have identified different general types of constraints (those concerning the form of the substitute word, its semantics or connotations, or the context of substitution), and (3) we have given an example of how to empirically evaluate the effects of given constraints.

For a practical application of humorous text generation, we recommend — according to the three contributions above — to first consider if the task can be formulated as lexical substitution (as was done in this article). Then, a key design question is what kind of constraints to use. The different constraint types identified in this paper are intended to help in that design task, and the specific ones we have used here can be used as illustrative examples. One should then test the constraints on data that are representative of the application at hand. Empirical results such as those in Tables 3 and 4 could be collected using crowdsourcing, as we have done in this paper, or potentially as feedback from users of the system.

Despite the large scale of the empirical evaluation in this paper, the evaluation has limitations that have decreased the funniness values and increased the p-values compared to some other possible settings. (1) The texts of the SMS corpus vary a lot in quality, style, and content. This adds a large amount of noise to the experimental data, making it difficult to substitute words in a humorous way, and also making it statistically harder to separate the effects of the investigated constraints from the random effects caused by variation in the source texts. (2) Another major source of noise in the evaluation is the likely presence of crowd-working scammers who give non-informed answers, e.g. by random. Our scammer removal mechanism is simple, and a better method should help improve the accuracy of the results. If feedback can be collected directly from the users of the actual application, then this problem could disappear.

In the application scenarios that we envision, either a text is given and (almost) any word may be replaced, or a collection of texts is given and any of them can be used as the original text (and any word may be replaced). The system could now consider a larger number of possible words to be substituted, possibly even in a large number of original texts, and choose the best one. Results such as those in Tables 3 and 4 can be used to estimate the funniness of the resulting text, depending on the constraints used.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Humor generators such as the ones considered here could have applications in conversational systems or in tools for creative writing. There also are applications directly related to auto-completion and auto-correction. As is well-known, they occasionally produce incorrect output that turns out to be unintentionally funny. The methods described here could be useful in detecting such cases and preventing especially inappropriate expressions.

In addition to applications for end users, like the ones mentioned above, the proposed method may have applications in empirical humor studies. They commonly face the difficulty of producing a suitable dataset. The samples should be collected according to some prefixed conditions, typically defined by a specific set of linguistic parameters, and the sample size should be large enough to allow statistical tests. An automated generator of stimuli, i.e., of humorous texts, could offer valuable help.

8 Conclusions

We have considered the task of generating humorous texts by substitution of a single word in a given short text. The conceptual simplicity of this task allows a systematic empirical study into some central mechanisms of verbal humor.

Inspired by theories of humor, we recognized and categorized three types of properties of word substitutions, those relating to the forms of the words, to their meaning and connotations, and to the context of the substitution. These properties can be used as constraints in a humor-generation system.

We focused on five specific constraints on word substitution, chosen according to their theoretical potential to induce a humor response — FORM: using a substitute phonetically or orthographically similar to the original word; CONNOT: using a connotational taboo word as substitute; TABIND: using a taboo-inducing word as substitute; NGRAM: using a substitute that matches the preceding word; WDPOS: substituting a word close to the end of the text.

In our tests, we used a corpus of real SMSs as source texts and crowdsourced the evaluation of the automatically modified texts to independent subjects. The tests involved 12 different empirical settings, 787 texts, 524 test subjects, and over 70,000 assessments in total. To the best of our knowledge, this is the first time that an evaluation of this scale and detail has been performed on machine-generated humorous texts, and with positive results.

We used funniness as a measure of the humor response. The effects of the studied constraints were analysed in detail using statistical tests. The main empirical findings are the following:

(1) We have isolated the effect of each lexical constraint and shown that each of them contributes to the humorous effect with a different weight.
(2) The effects of the constraints are cumulative and provide first evidence of a compositional nature of the humorous effect in the studied context.
(3) There are combinations of constraints that support each other and amplify their individual contribution. In particular, connotational words and replacement at the end of the text produce the highest effect.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

The results suggest that fully automated production of humorous text can be feasible. Even if the practical value of SMS modification is questionable, we consider the empirical results significant for proving the potential of automated humor generation. We envision the approach used in this paper to have applications, e.g. in automated production of funny marketing messages, in intelligent conversational agents, in helping creative writing, as well as in empirical studies of humor.

In the future, it would also be interesting to use a similar setting to investigate more subtle ways to generate humor. Other semantic or connotational domains besides taboos could be explored. We are also interested in studying the effect of semantic opposition between the original word and the substitute, as well as the use of several semantically related substitutes.

Finally, we would like to experimentally identify and separate those humorous constraints that induce incongruity from those that make the incongruity funny. If feasible, this would shed further light onto how to produce humorous linguistic expressions.

References

Beattie, J. 1971. An essay on laughter, and ludicrous composition. In *Essays*. Reprinted by Garland (Original work published by William Creech, Edinburgh, 1776), New York: Garland Publishing.

Binsted, K., Pain, H., and Ritchie, G. 1997. Children's evaluation of computer-generated punning riddles. *Pragmatics and Cognition* **2**(5): 305–354.

Carrell, A. 1997. Joke competence and humor competence. *Humor* **10**: 173–185.

Chen, T., and Kan, M.-Y. 2013. Creating a live, public short message service corpus: the NUS SMS Corpus. *Language Resources and Evaluation* **74**(2): 299–335.

Cory, M. 1995. Comedic distance in holocaust literature. *Journal of American Culture* **18**(1): 35–40.

Doucet, A., and Ahonen-Myka, H. 2006. Probability and expected document frequency of discontinued word sequences, an efficient method for their exact computation. *Traitement Automatique des Langues (TAL)* **46**(2): 13–37.

Dybala, P., Ptaskynsky, M., Higuchi, S., Rzepka, R., and Araki, K. 2008. Humor Prevails! - Implementing a joke generator into a conversational system. In *Proceedings of the 21st Australian Joint Conference on AI (AI-08)*, vol. 5360, pp. 214–225. Berlin: Springer Verlag.

Fellbaum, C. 1998. *WordNet. An Electronic Lexical Database*. Cambridge, Massachusetts: The MIT Press.

Hempelmann, C. F. 2003. *Paronomasic Puns: Target Recoverability Towards Automatic Generation*. Ph.D. thesis, West Lafayette, IN: Purdue University.

Hempelmann, C., Taylor, J., and Raskin, V. 2012. Tightening up joke structure: not by length alone. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society 2012 (CogSci 2012)*, Sapporo, Japan.

Jay, T., Caldwell-Harris, C., and King, K. 2008. Recalling taboo and nontaboo words. *American Journal of Psychology* **121**(1): 83–103.

Kazai, G., Kamps, J., Koolen, M., and Milic-Frayling, N. 2011. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011*, ACM, pp. 205–214, Beijing, China.

Koestler, A. 1964. *The Act of Creation*. London: Hutchinson.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Leach, E. 1964. Antropological aspects of language: animal categories and verbal abuse. In E. H. Lenneberg (ed.), *New Directions in the Study of Language*, pp. 23–63. Cambridge, Massachusetts: The MIT Press.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* **10**(8): 707–710.

Levison, M., and Lessard, G. 1992. A system for natural language generation. *Computers and the Humanities* **26**: 43–58.

Magnini, B., and Cavaglià, G. 2000. Integrating subject field codes into WordNet. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, pp. 1413–1418, Athens, Greece.

Martin, R. A. 2007. *The Psychology of Humor: An Integrative Approach*. Elsevier, Elsevier: San Diego, California.

McKay, J. 2002. Generation of idiom-based witticisms to aid second language learning. In O. Stock, C. Strapparava, and A. Nijholt, (eds.), *Proceedings of the The April Fools Day Workshop on Computational Humour (TWLT20)*, pp. 77–87, Trento, Italy.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A. and Aiden, E. L. 2011. Quantitative analysis of culture using millions of digitized books. *Science* **331**(6014): 176–182.

Morreall, J. 2013. Philosophy of Humor. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab Publisher, Stanford, California.

Mulkay, M. 1988. *On Humour: Its Nature and its Place in Modern Society*. Cambridge, UK: Polity Press.

Özbal, G., and Strapparava, C. 2012. A computational approach to the automation of creative naming. In *Proceedings of the 50th annual meeting of the Association of Computational Linguistics (ACL-2012)*, pp. 703–711, Jeju Island, South Korea.

Petrović, S., and Matthews, D. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 228–232, Sofia, Bulgaria.

Raskin, V. 1985. *Semantic Mechanisms of Humor*. Netherlands: Dordrecht-Boston-Lancaster.

Raskin, V., and Attardo, S. 1994. Non-literalness and non-bona-fide in language: approaches to formal and computational treatments of humor. *Pragmatics and Cognition* **2**(1): 31–69.

Ritchie, G. 2002. The structure of forced reinterpretation jokes. In *Proceedings of the The April Fools Day Workshop on Computational Humour (TWLT20)*, pp. 47–56, Trento, Italy.

Ritchie, G. 2003. *The Linguistic Analysis of Jokes*. London: Routledge.

Ross, J., Irani, I., Silberman, M. S., Zaldivar, A., and Tomlinson, B. 2010. Who are the crowdworkers?: shifting demographics in Amazon Mechanical Turk. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 2863–2872, Atlanta, Georgia.

Ruch, W. 1992. Assessment of appreciation of humor: studies with the 3 WD Humor Test. In C. D. Spielberger and J. N. Butcher (eds.), *Advances in Personality Assessment*, vol. 9, pp. 27–75. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ruch, W. 2008. Psychology of humor. In V. Raskin (ed.), *The Primer of Humor Research*, pp. 17–100. De Gruyter Mouton, Hillsdale, New Jersey.

Schank, R., and Abelson, R. 1977. *Scripts, Plans Goals and Understanding: An Inquiry into Human Knowledge Structures*. Erlbaum, Hillsdale, NJ.

Seizer, S. 2011. On the uses of obscenity in live stand-up comedy. *Anthropological Quarterly* **84**(1): 209–234.

Sherzer, J. 2002. *Speech Play and Verbal Art*. University of Texas Press, Austin, Texas.

Sjöbergh, J. 2006. Vulgarities are fucking funny, or at least make things a little bit funnier. Technical Report TRITA-CSC-TCS 2006: 4, School of Computer Science and Communication, the Royal Institute of Technology, Stockholm.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press

Stock, O., and Strapparava, C. 2003. HAHAcronym: humorous agents for humorous acronyms. *Humor: International Journal of Humor Research* **16**(3), pp. 297–314.

Suls, J. 1972. A two-stage model for the appreciation of jokes and cartoons: an information-processing analysis. In J. Goldstein and P. McGhee (ed.), *The Psychology of Humor*, pp. 81–100. New York: Academic Press.

Taylor, J., and Mazlack, L. 2005. Toward computational recognition of humorous intent. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society (COGSCI 05)*, pp. 2166–2171, Stresa, Italy.

Valitutti, A. 2011. How many jokes are really funny? Towards a new approach to the evaluation of computational humour generators. In *Proceedings of 8th International Workshop on Natural Language Processing and Cognitive Science*, pp. 189–200, Copenhagen, Denmark.

Valitutti, A., Toivonen, H., Doucet, A., and Toivanen, J. M. 2013. 'Let everything turn well in your wife': generation of adult humor using lexical constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 243–248, Sofia, Bulgaria.

Veale, T. 2013. Humorous similes. *HUMOR: The International Journal of Humor Research* **21**(1): 3–22.

Venour, C. 1999. The computational generation of a class of puns. Master's thesis, Kingston, Ontario: Queen's University.

Westfall, P. H., and Young, S. S. 1993. *Resampling-Based Multiple Testing*. New York: John Wiley & Sons.

Zwicky, A. M. 1979. Classical malapropisms. *Language Sciences* **1**(2): 339–348.

https://doi.org/10.1017/S1351324915000145 Published online by Cambridge University Press