Characterizing Humour: An Exploration of Features in Humorous Texts

Rada Mihalcea[1,2], Stephen Pulman[2]

[1] Computer Science Department, University of North Texas
rada@cs.unt.edu
[2] Computational Linguistics Group, Oxford University
sgp@clg.ox.ac.uk

Abstract. This paper investigates the problem of automatic humour recognition, and provides and in-depth analysis of two of the most frequently observed features of humorous text: human-centeredness and negative polarity. Through experiments performed on two collections of humorous texts, we show that these properties of verbal humour are consistent across different data sets.

1 Introduction

This paper addresses two research questions concerned with the characteristics of textual humour. First, are humorous and serious texts separable, and does this property hold for different datasets? To answer this question, we use two different data sets of verbal humour – a collection of short one-liners and a set of humorous news articles – and attempt to automatically separate them from their non-humorous counterparts.

Second, if humorous and serious texts are separable, what are the distinctive features of humour, and do they hold across datasets? In answer to this second question, we attempt to identify some of the most salient features of verbal humour, and analyse their occurrence in the two data sets.

While these are interesting issues in themselves, there is also a medium-term practical application for 'humour' recognition in the design of conversational agents of various types: detecting and responding appropriately to humour is a characteristic of natural human interaction that is conspicuously lacking in implemented systems. In the longer term, by gaining insight into the mechanisms underlying humour, we hope to increase our understanding of aspects of the creative use of language, i.e. uses of language which go beyond 'banal humorless prose' and display some reflective and self-aware properties. While these are pre-eminently displayed in creative works like novels or poetry, they are also present in more everyday phenomena like humour.

The paper is organized as follows. We first review related work in computational humour, and briefly cover some of the most recent methods for humour generation and recognition. We then describe the two data sets used in this paper, and briefly overview two machine learning techniques for text classification. Next, we address the first question, and present the results obtained in the automatic classification of humorous and non-humorous data sets. We then present some of the characteristics of verbal humour as observed in an analysis of humorous texts, and provide a detailed analysis of two of the most dominant features: human-centeredness and negative polarity. Finally, we conclude with a discussion.

1.1 Related Work

While humor is relatively well studied in scientific fields such as linguistics [1] and psychology [4, 15], to date there is only a limited number of research contributions made toward the construction of computational humour prototypes. Most of the computational approaches to date on style classification have focused on the categorization of more traditional literature genres, such as fiction, scitech, legal, and others [7], and much less on creative writings such as humor.

One of the first attempts in computational humor is perhaps the work described in [2], where a formal model of semantic and syntactic regularities was devised, underlying some of the simplest types of puns (*punning riddles*). The model was then exploited in a system called JAPE that was able to automatically generate amusing puns.

Another humor-generation project was the HAHAcronym project [16], whose goal was to develop a system able to automatically generate humorous versions of existing acronyms, or to produce a new amusing acronym constrained to be a valid vocabulary word, starting with concepts provided by the user. The comic effect was achieved mainly by exploiting incongruity theories (e.g. finding a religious variation for a technical acronym).

Another related work, devoted this time to the problem of humor comprehension, is the study reported in [17], focused on a very restricted type of wordplays, namely the "Knock-Knock" jokes. The goal of the study was to evaluate to what extent wordplay can be automatically identified in "Knock-Knock" jokes, and if such jokes can be reliably recognized from other non-humorous text. In our own previous work, we have studied the problem of automatic humour recognition using content and stylistic features [9], and have evaluated the use of large collections of humorous texts for improving widely used computer applications such as email [11].

2 Datasets for Computational Humour

There have been only a relatively small number of previous attempts targeting the computational modeling of humour. Among these, most of the studies have relied on small datasets, e.g. 195 jokes used for the recognition of knock-knock jokes [18], or 200 humorous headlines analysed in [3], and such small collections may not suffice for the robust learning of features of humorous text.

More recently, we proposed a Web-based bootstrapping method that automatically collects humorous sentences starting with a handful of manually selected seeds, which allowed us to collect a large dataset of 16,000 one-liners [9]. In this paper, we use the corpus of one-liners, as well as a new dataset that we introduce in this paper consisting of humorous news articles. By considering two different datasets, we hope to be able to derive more definite and robust conclusions about the characteristic features of humorous texts.

2.1 One-liners

A one-liner is a short sentence with comic effects and an interesting linguistic structure: simple syntax, deliberate use of rhetoric devices (e.g. alliteration, rhyme), and frequent

use of creative language constructions meant to attract the readers' attention. While longer jokes can have a relatively complex narrative structure, a one-liner must produce the humorous effect "in one shot", with very few words. These characteristics make this type of humor particularly suitable for use in an automatic learning setting, as the humor-producing features are guaranteed to be present in the first (and only) sentence.

Starting with a short *seed* set consisting of a few one-liners manually identified, the algorithm proposed in [9] automatically identifies a list of webpages that include at least one of the seed one-liners, via a simple search performed with a Web search engine. Next, the webpages found in this way are HTML parsed, and additional one-liners are automatically identified and added to the seed set. The process is repeated several times, until enough one-liners are collected.

Take my advice; I don't use it anyway.
I get enough exercise just pushing my luck.
I took an IQ test and the results were negative.
A clean desk is a sign of a cluttered desk drawer.
Beauty is in the eye of the beer holder.

Fig. 1. Sample examples of one-liners

Two iterations of the bootstrapping process, started with a small seed set of ten one-liners, resulted in a large set of about 24,000 one-liners. After removing the duplicates using a measure of string similarity based on the longest common subsequence, the resulting dataset contains 16,000 one-liners, which are used in the experiments reported in this paper. The one-liners humor style is illustrated in Figure 1, which shows five examples of such one-sentence jokes.

2.2 Humorous News Articles

The second dataset we consider consists of daily stories from the newspaper "The Onion" – a satiric weekly publication with ironic articles about current news, targeting in particular stories from the United States. It is known as "the best satire magazine in the U.S."[3] and "the best source of humour out there"[4].

We collected all the articles published during August 2005 – March 2006, which resulted in a dataset of approximately 2,500 news articles. We cleaned all the HTML tags, eliminated the header containing information specific to the newspaper, and finally removed all the news articles that felt outside the 1000–10,000 character length range. This process left us with a final dataset of 1,125 news stories with humorous content. Figure 2 shows a sample article from this dataset.

[3] Andrew Hammel, German Joys, http://andrewhammel.typepad.com
[4] Jeff Grienfield, CNN senior analyst, http://www.ojr.org/

Canadian Prime Minister Jean Chrétien and Indian President Abdul Kalam held a subdued press conference in the Canadian Capitol building Monday to announce that the two nations have peacefully and sheepishly resolved a dispute over their common border. Embarrassed Chrétien and Kalam restore diplomatic relations. "We are – well, I guess proud isn't the word – relieved, I suppose, to restore friendly relations with India after the regrettable dispute over the exact coordinates of our shared border," said Chrétien, who refused to meet reporters' eyes as he nervously crumpled his prepared statement. "The border that, er... Well, I guess it turns out that we don't share a border after all." Cheétien then officially withdrew his country's demand that India hand over a 20-mile-wide stretch of land that was to have served as a demilitarized buffer zone between the two nations." Really, I think the best thing for us to do is forget about the whole thing as quickly as possible," Cheétien added.

Fig. 2. Sample news article from "The Onion"

3 Automatic Humour Recognition

The first question we are concerned with is whether the humorous texts represent a distinct genre that can be easily and reliably distinguished from other non-humorous datasets. To answer this question, similar to our previous work [9], we formulate the humor-recognition problem as a traditional classification task, and feed positive (humorous) and negative (non-humorous) examples to an automatic classifier.

In particular, in this study we are concerned with the *semantic* characteristics of humour, and therefore we focus our attention on content classification, as opposed to stylistic features as used in previous work [9]. The content of humorous texts is thus "compared" against the content of serious texts using standard text classification techniques.

To perform the classification task, in addition to positive (humorous) examples, we also need a set of negative (serious) texts. For each humorous dataset, a collection of negative examples was constructed, identified as texts that are non-humorous, but similar in structure and composition to the humorous examples. We do not want the automatic classifiers to learn to distinguish between humorous and non-humorous examples based simply on text length or obvious vocabulary differences. Instead, we seek to enforce the classifiers to identify humor-specific features, by supplying them with negative examples similar in most of their aspects to the positive examples, but different in their comic effect.

3.1 Negative Datasets

For each humorous dataset, we collected an equal number of non-humorous examples, by mixing texts from three or four different sources. The purpose of seeking different sources for the construction of the negative non-humorous dataset is to avoid the bias that could be introduced by a specific source or genre.

For the one-liners, we created a negative dataset consisting of a mix of sentences following the same length restrictions (10–15 words). We combined: (1) *Reuters* titles, extracted from news articles published in the Reuters newswire over a period of one

year (8/20/1996 – 8/19/1997); (2) *Proverbs* extracted from an online proverb collection; (3) *British National Corpus (BNC)* sentences; and (4) sentences from the *Open Mind Common Sense* collection of commonsense statements.

For the news articles, the negative examples were collected from three different sources: (1) articles drawn from *Los Angeles Times*; (2) newstories from the *Foreign Broadcast Information Service*; and finally (3) texts extracted from the *British National Corpus*. All the non-humorous examples were constrained to have a similar structure to "The Onion" articles – stories with a length of 1,000–10,000 characters.

3.2 Text Classification

We ran classification experiments using two frequently used text classifiers, Naïve Bayes and Support Vector Machines, selected based on their performance in previously reported work, and for their diversity of learning methodologies.

Naïve Bayes. The main idea in a Naïve Bayes text classifier is to estimate the probability of a category given a document using joint probabilities of words and documents. Naïve Bayes classifiers assume word independence, but despite this simplification, they perform well on text classification. While there are several versions of Naïve Bayes classifiers (variations of multinomial and multivariate Bernoulli), we use the multinomial model, previously shown to be more effective [8].

Support Vector Machines. Support Vector Machines (SVM) are binary classifiers that seek to find the hyperplane that best separates a set of positive examples from a set of negative examples, with maximum margin. Applications of SVM classifiers to text categorization led to some of the best results reported in the literature [6].

3.3 Classification Results

For each humorous dataset, we ran classification experiments with respect to their "negative" non-humorous counterpart. The documents were tokenized and stemmed prior to classification; no other pre-processing was applied.

All the evaluations are performed using stratified ten-fold cross validations, for accurate estimates. The baseline for all the experiments is 50%, which represents the classification accuracy obtained if a label of "humorous" (or "non-humorous") would be assigned by default to all the examples in the data set. Table 1 shows the classification accuracies obtained with each of the classifiers.

Classifier	One-liners	News articles
Naive Bayes	79.69%	88.00%
SVM	79.23%	96.80%

Table 1. Classification accuracy for the two humorous datasets.

The results indicate that humorous and non-humorous data are clearly separable, using exclusively linguistic features. Not surprisingly, the classification accuracy for

the news articles is higher than for the one-liners, most likely due to the larger size of the documents in the newstories' collection. The different gap between the SVM and the Naive Bayes classification accuracies can be probably attributed to the same reason, with the SVM classifier leading to results close to 100% in the case of the newstories, but to results slightly worse than those obtained with the Naive Bayes classifier in the case of the one-liners.

Perhaps even more importantly than the classification results are the features that can be learned from the classifiers' output, which can help us characterize the linguistic properties of humour. In the following, we describe the features identified in a previous examination of linguistic properties of verbal humour, and provide an in-depth, larger-scale evaluation of the two main characteristics of humour: human-centeredness and negative polarity.

4 Characteristics of Verbal Humour

In a previous analysis of the features of verbal humour [10], we tried to identify and classify the content-based humor-specific features characteristic to the one-liner data set. By examining by hand the most discriminative content-based features learned during the text classification process, we tried to classify them into semantic classes. The following frequently occurring word classes emerged:

Human-centric vocabulary. Jokes seem to constantly make reference to human-related scenarios, through the frequent use of words such as *you*, *I*, *man*, *woman*, *guy*, etc. For instance, the word *you* alone occurs in more than 25% of the one-liners (*"You can always find what you are not looking for"*), while the word *I* occurs in about 15% of the one-liners (*"Of all the things I lost, I miss my mind the most"*). This supports earlier suggestions made by Freud [5], and later on by Minsky [12], that laughter is often provoked by feelings of frustration caused by our own, sometime awkward, behaviour.

Negation. Humorous texts seem to often include negative word forms, such as *doesn't*, *isn't*, *don't*. A large number of the jokes in our collection contain some form of negation, e.g. *"Money can't buy you friends, but you do get a better class of enemy"*, or *"If at first you don't succeed, skydiving is not for you."*

Negative orientation. In addition to negative verb forms, jokes seem to also contain a large number of words with a negative polarity, such as adjectives with negative connotations like *bad*, *illegal*, *wrong* (*"When everything comes your way, you are in the wrong lane"*), or nouns with a negative load, e.g. *error*, *mistake*, *failure* (*"User error: replace user and press any key to continue"*). Both the negative verb forms and the words with negative orientations are potential reflections of the incongruity-based theories of humor.

Professional communities. Many jokes seem to target professional communities that are often associated with amusing situations, such as lawyers, programmers, policemen. For instance, about 100 one-liners in our collection fall under this category, e.g. *"It was so cold last winter that I saw a lawyer with his hands in his own pockets."*

Human "weakness". Finally, the last significantly large semantic category that we identified refers to events or entities that are often associated with "weak" human moments, including nouns such as *ignorance*, *stupidity*, *trouble* (*"Only adults have trouble with child-proof bottles"*), *beer*, *alcohol* (*"Everybody should believe in something, I believe I'll have another beer"*), or verbs such as *quit*, *steal*, *lie*, *drink* (*"If you can't drink and drive, then why do bars have parking lots?"*). As mentioned before, this kind of vocabulary seems to relate to theories of humor that explain laughter as an effect of frustration or awkward feelings, when we end up laughing "at ourselves" [12].

On a higher level, these characteristics can be classified into two main classes. First, *human-centric vocabulary*, *professional communities*, and *human "weakness"* can be grouped into the larger category of **human centeredness**. Second, *negation*, *negative orientation*, and *human "weakness"* all have to do with the broader category of **polarity orientation**. In the following, we analyse each of these categories in turn, and bring evidence of a high correlation between humorous text and each of these two features.

5 Human Centeredness

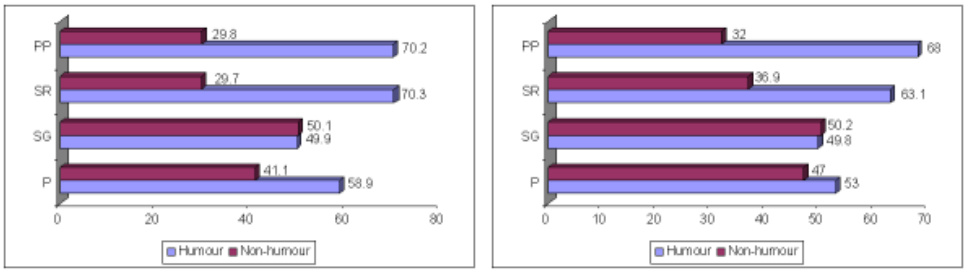
For a more robust evaluation of the human-centeredness property of the humorous texts, we implemented a system that measures the weight of the most discriminatory features learned from the text classification process with respect to given semantic classes considered relevant for human-centeredness.

Specifically, we begin by creating a list of salient features for the humorous dataset. Starting with the features identified as important by the Naive Bayes classifier (a threshold of 0.3 was used in the feature selection process), we select all those features that have a total weight exceeding a given threshold T, where a feature weight is calculated for each category (humorous/non-humorous) and is determined as the probability of seeing the feature in a given category. We then calculate the *humorous score* of a feature as the ratio between the weight in the humorous corpus and the total weight in the entire mixed corpus. This results in a score within the [0–1] interval, with a value closer to 1 indicating a feature representative for the humorous texts, and a value close to 0 corresponding to high saliency features for the non-humorous dataset. In the evaluations reported below, we use a threshold T of 100, which allows us to extract the top 1,500 most discriminatory features for each dataset.

Next, given a certain semantic class, we measure the *weight* of that semantic class with respect to the most discriminatory features by adding up the corresponding weights, and normalizing with respect to the size of the semantic class. For instance, assuming a semantic class that includes the words *I*, *me*, *myself*, with the *humorous scores* of 0.88, 0.65, and 0.55 respectively measured on the humorous dataset, the weight of the given semantic class is then measured as $(0.88 + 0.65 + 0.55)/3 = 0.69$[5].

By using semantic classes, we can generalize over the individual word features learned from the classifiers' output, and derive *categories* of words representative for the humorous data. Note that a semantic class that has no correlation with the humorous

[5] Correspondingly, the weight of the semantic class in the non-humorous texts is measured as $1 - 0.69 = 0.31$.



(News articles) (Oneliners)

Fig. 3. Semantic classes reflecting human-centeredness within humorous texts. PP = personal pronouns; SG = social groups; SR = social relationships; P = persons.

features of a text will result in an approximately equal weight (0.50) measured on the humorous and non-humorous texts.

To measure the human-centeredness characteristic of humorous texts, for each dataset we extracted the top 1,500 most discriminatory features, and subsequently measured the weight of four semantic classes that we considered relevant for the property of human-centeredness: *persons*, *social groups*, *social relations*, and *personal pronouns*. The first three categories are derived automatically from WordNet, by listing all the nouns found in the synsets subsumed by the node {*person, individual, someone, somebody, mortal, human, soul* (20,676 nouns are extracted), {*relative, relation*} and {*relationship, human relationship*} (351 nouns), and {*social group*} (2,393 nouns). The fourth category is constructed by listing exhaustively all the personal pronouns in the English language.

Figure 3 shows the weight of each semantic class with respect to humorous and non-humorous data, for each of the two datasets (one-liners and news articles). Our hypothesis concerning the human-centeredness of humour seems to be confirmed, with a much higher weight measured for the semantic classes of *persons*, *social relationships*, and *personal pronouns* in humorous texts. In particular, social relationships (e.g. *wife*, *husband*, *son*) and personal pronouns (e.g. *I*, *you*) seem to have high prevalence in humorous data. Rather surprisingly, social groups do not correlate with humorous texts, having an equal weight distribution between humorous and non-humorous data. Although we initially thought that this WordNet class would help us uncover the category of *professional communities*, on closer inspection it turns out that the nouns relevant for such communities (e.g. *programmer*, *lawyer*) are represented under the semantic class of *person*. Instead, the *social group* category includes more organization-related nouns such as *church*, *university*, or *council*, which are not necessarily representative for humorous text.

6 Polarity Orientation

The second humour characteristic we are investigating is concerned with the polarity orientation of humour. In a previous manual analysis of humorous features (Section 4), we observed a frequent use of negative verbal forms in humorous texts, as well

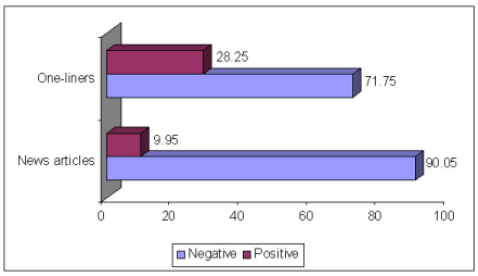


Fig. 4. Polarity orientation of humorous data.

as other words with negative orientation (e.g. negative adjectives), or denoting human "weakness." In order to take this analysis to the next step, and investigate on a larger scale the polarity orientation of humour, we have implemented a tool for automatic sentiment analysis, and used this tool to annotate the two humorous datasets used in the current study.

Starting with a dataset annotated for "positive" and "negative" orientation, we implemented a classification system that has the ability to automatically indicate the semantic orientation of a text. Specifically, we are using the dataset of 10,662 short text fragments introduced in [13], and feed the 5,331 "positive" and the 5,331 "negative" fragments into a Naive Bayes classifier. In a ten-fold cross validation experiment, the accuracy of the system was determined as 78.15%, which compares favorably with previous results reported on the same dataset [13].

Using this sentiment analysis tool, we automatically annotate the two humorous datasets, with results shown in Figure 6. These results seem to confirm our hypothesis that humour tends to have a strong negative orientation, with 71.74% of the one-liners being labeled as negative, and as many as 90.04% of the news articles from "The Onion" having a negative annotation. Interestingly, regular text also tends to have a slight tendency toward the negative, with 56.26% of the mix of "serious" sentences being determined as having a negative orientation. General "serious" news articles are even more negative, with 67.60% labeled as negative, perhaps reflecting the general negative trend of the stories typically reported in the news.

Interestingly, by analyzing the annotations, several of the examples labeled as positive seem to include words with a negative orientation, whose strength was perhaps not high enough to be selected as negative by the automatic classifier. For instance, "CURSOR: What you become when your system crashes." is labeled as an example with positive orientation, despite the word "crashes" that seems to indicate a negative outcome. Conversely, "I love deadlines, especially the whooshing sound as they fly by." is labeled as negative, perhaps because of a frequent occurrence of "deadline" in negative contexts, despite the fact that this one-liner does not have a clear negative connotation. A larger training dataset with polarity annotations, perhaps integrating manual annotations of jokes, is likely to improve the accuracy of the annotations.

7 Discussion and Conclusions

The questions with which we began were: (1) Are humorous and serious texts separable, and does this property hold for different datasets? and (2) If so, what are the distinctive features of humour, and do they hold across datasets?

In answer to the first of these questions, we have shown that humorous and serious texts can be separated at the linguistic level, and also that this holds for at least two different datasets: short one-liners, and longer news articles. Of course, there are many other types of humorous and non-humorous prose and it may be that some of these are more difficult to separate.

In trying to address the second question, by analysis of the linguistic features that emerged as important for the classifiers, we hypothesized two main characteristics of humour: human-centeredness and negative orientation, which were validated through larger scale experiments of annotations on the two datasets. In a sense, one might have predicted the human centeredness *a priori*, given that humour seems to be a specifically human property, but the negative orientation we found is less obvious: indeed, from the generally positive effects associated with humour, one might have expected the opposite.

As Ritchie [14] suggests, it is probably misguided to look for **the** defining property of humour, but we may make some speculations on the basis of our findings as to one of its possible functions. It does not seem completely implausible that some varieties of humour act as a kind of "natural therapy" whereby tensions related to **negative** scenarios concerning **humans** (us) are relieved, by emphasizing them in a context which leads to them being exorcised through laughter.

References

1. ATTARDO, S. *Linguistic Theory of Humor*. Mouton de Gruyter, Berlin, 1994.
2. BINSTED, K., AND RITCHIE, G. Computational rules for punning riddles. *Humor 10*, 1 (1997).
3. BUCARIA, C. Lexical and syntactic ambiguity as a source of humor. *Humor 17*, 3 (2004).
4. FREUD, S. *Der Witz und Seine Beziehung zum Unbewussten*. Deutike, Vienna, 1905.
5. FREUD, S. *Der Witz und Seine Beziehung zum Unbewussten*. Deutike, Vienna, 1905.
6. JOACHIMS, T. Text categorization with Support Vector Machines: learning with mny relevant features. In *Proceedings of the European Conference on Machine Learning* (1998), pp. 137–142.
7. KESSLER, B., NUNBERG, G., AND SCHUETZE, H. Automatic detection of text genre. In *Proceedings of the 35^{th} Annual Meeting of the Association for Computational Linguistics (ACL97)* (Madrid, July 1997).
8. MCCALLUM, A., AND NIGAM, K. A comparison of event models for Naive Bayes text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization* (1998).
9. MIHALCEA, R., AND STRAPPARAVA, C. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Human Language Technology / Empirical Methods in Natural Language Processing conference* (Vancouver, 2005).
10. MIHALCEA, R., AND STRAPPARAVA, C. Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence 22*, 2 (2006), 126–142.

11. MIHALCEA, R., AND STRAPPARAVA, C. Technologies that make you smile: Adding humor to text-based applications. *IEEE Intelligent Systems 21*, 5 (2006).
12. MINSKY, M. Jokes and the logic of the cognitive unconscious. Tech. rep., MIT Artificial Intelligence Laboratory, 1980.
13. PANG, B., AND LEE, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics* (Barcelona, Spain, July 2004).
14. RITCHIE, G. *The Linguistic Analysis of Jokes*. Routledge, London.
15. RUCH, W. Computers with a personality? lessons to be learned from studies of the psychology of humor. In *Proceedings of the The April Fools Day Workshop on Computational Humour* (2002).
16. STOCK, O., AND STRAPPARAVA, C. Getting serious about the development of computational humour. In *Proceedings of the 8^{th} International Joint Conference on Artificial Intelligence (IJCAI-03)* (Acapulco, Mexico, August 2003).
17. TAYLOR, J., AND MAZLACK, L. Computationally recognizing wordplay in jokes. In *Proceedings of CogSci 2004* (Chicago, August 2004).
18. TAYLOR, J., AND MAZLACK, L. Computationally recognizing wordplay in jokes. In *Proceedings of CogSci 2004* (Chicago, August 2004).

This article was processed using the LaTeX macro package with LLNCS style