Children's evaluation of computer–generated punning riddles

Kim Binsted[*], Helen Pain, Graeme Ritchie
Department of Artificial Intelligence
University of Edinburgh

Abstract

We have developed a formal model of certain types of riddles, and implemented it in a computer program, JAPE, which generates simple punning riddles. In order to test the model, we evaluated the behaviour of the program, by having 120 children aged eight to eleven years old rate JAPE–generated texts, human–generated texts, and non–joke texts for 'jokiness' and funniness. This confirmed that JAPE's output texts are indeed jokes, and that there is no significant difference in funniness or jokiness between JAPE's most comprehensible texts and published human–generated jokes.

[*]Now at Sony Computer Science Laboratory Inc., Takanawa Muse Building., 3-14-13 Higashi-Gotanda Shinagawa-ku, Tokyo 141, Japan.

1 Introduction

There is at present no general theory of humour which is fully detailed and rigorous. Even in the slightly more restricted area of *verbal humour* — humour which is transmitted through language — there is no agreed or well developed model of the underlying mechanisms (see (Attardo, 1994) for a useful survey). We are interested in the operation of verbal humour, and have made a start by investigating one particular class of humorous item – the *punning riddle* (see (Binsted and Ritchie, 1996; Binsted and Ritchie, 1997) for discussion of our general theoretical assumptions and the validity of tackling a limited phenomenon initially). We have developed a formal model of punning riddles, of a sort which are very common, especially in children's joke books. For example:

(1) What do you get when you cross a sheep and a kangaroo?
A woolly jumper. (Webb, 1978)

Our model was then implemented in a program, JAPE (Joke Analysis and Production Engine), which generates riddles from a general purpose lexicon. JAPE's jokes seem to be very similar in quality to published, human–generated jokes. For example:

(2) What's the difference between money and a bottom?
One you spare and bank, the other you bare and spank. [JAPE]

Claims of JAPE's success, however, are empty without empirical justification. We have therefore carried out a controlled evaluation of JAPE's output. The purpose of this confirmatory evaluation was to analyse in detail the behaviour of the final system, and to examine its successes and failures. In this paper, after outlining the model, and some of the background, we will describe the procedure and results of the confirmatory evaluation in some detail.

2 The domain

Punning riddles were chosen over other forms of verbal humour for several reasons. There are thousands of them available for study, in joke books like the Crack–A–Joke Book (Webb, 1978) and Super Duper Jokes (Young, 1993). They seem to exhibit regular structures and mechanisms which could be captured in a formal model. Finally, they rely on linguistic information that is likely to be found in normal lexical resources — unlike non-linguistic verbal jokes, which use a great deal of common sense knowledge.

Groups of words, or *texts*, also have a written form, a spoken (phonological) form and an interpretation, and so can be ambiguous. For example, the texts written *Please announce her* and *Please, an ounce sir* have the same phonological form, and so that form is ambiguous. (Notice that we are adopting the term *text* as shorthand for 'group of words', not necessarily in written form.) A text containing an ambiguous part will also be ambiguous, although the context may disambiguate it. Here we will use *homonyms* to refer to two words that sound the same but are spelled differently, and *homophones* to describe two text segments that sound the same and may or may not have the same spelling.

The two kinds of (relatively low–level) ambiguities that we have concentrated on are:

spelling ambiguity: This is where one phonological form corresponds to two (or more) written forms and senses. For example, the phoneme sequence [s,ia,r,ia,l] could be written either as *cereal* or as *serial.* (For simplicity, we are using the ARPAbet system (Robinson, 1996) for expressing phonemes.) The ambiguity lies in the fact that there are several possible phoneme to text mappings. This joke uses spelling ambiguity:

(3) What do you get when you cross a rabbit with a lawn sprink-

ler?

Hare spray. (Young, 1993)

word sense ambiguity: This is when one phonological form and the associated written form correspond to two or more senses. Whether the word is spoken or read, the listener/reader cannot tell which sense of the word is meant. For example, *bank* is word sense ambiguous, in that it has two senses: the side of a river, and a financial institution. Here is a riddle that uses word sense ambiguity:

(4) Where do snowmen keep their money?

In snow banks. (Young, 1993)

In our corpus of joke books, jokes using other kinds of high–level ambiguity can be found, but these two types are very widespread, and we have chosen to model them for several reasons. Information about the pronunciation of words is available, whether in the form of phonological descriptions of words (such as one might find in a dictionary), or a list of phonologically similar words (homonyms or near–homonyms). More complex ambiguities related to the whole structure of the sentence would be more difficult to generate and to detect.

More importantly, the strategies employed by riddles containing these low–level ambiguities are simple and general (see section 3 below), whereas the strategies employed by the other kinds of riddle are often specific to the exact ambiguity used in that particular joke.

3 Mechanisms used in punning riddles

There are three main strategies used in puns to exploit spelling or word sense ambiguity: *juxtaposition*, *substitution*, and *comparison*. This is not to say that other strategies do not exist; however, none were found among the large number of punning jokes examined.

It is useful in discussing riddles to introduce the idea of *confusability.* If a text is ambiguous in one form, then its variations in other forms are *confusable.* For example, the phoneme sequence [s,ia,r,ia,l] is spelling ambiguous, and it has two written forms: *cereal* and *serial.* These two written forms are completely confusable. Two texts which contain confusable or identical segments are considered to be *partially confusable.* Confusable segments are often substituted, juxtaposed or compared in jokes, as discussed below.

Juxtaposition

Juxtaposition is the most simple mechanism, simply placing the confusable segments near each other and treating them as a normal construction. For example:

(5) What do you call a weird market?
A bizarre bazaar. [JAPE]

(6) What do you get if you cross a dog and a kangaroo?
A pooch with a pouch. (Ertner, 1993)

Substitution

This mechanism works by substituting one confusable segment for another, as part of a larger text, and using the resulting text as if it were a sensible construction. That is, the constructed text is used normally, but with a new interpretation that is a combination of the interpretations of the elements which make it up.

For example, the word *purr* is confusable with the first syllable of *purgatory.* If we substitute *purr* for *pur*, we get the constructed text *purrgatory.*

We must also construct a plausible interpretation for the construction — 'cat afterlife', for example. If the constructed text and its new interpretation are used as if they were normal, we get a joke. For example:

(7) Where do cats go when they die?

Purrgatory. (Ertner, 1993)

Comparison

This mechanism explicitly compares two confusable texts, usually by asking for similarities or differences in the question part of the riddle. Positive comparison riddles ask for similarities (e.g. "How is A like B?"), and negative comparison riddles ask for differences (e.g. "What is the difference between A and B?"). Negative comparison riddles most often contain pairs of texts constructed with metathesis. Metathesis pairs are similar in a regular way (they have exchanged sounds), and are (partially) confusable. Spoonerisms are a kind of metathesis, in which the initial sounds of a phrase are exchanged to form a pair of confusable texts. For example:

(8) What's the difference between a short witch and a deer running from the hunters?

One's a stunted hag and the other's a hunted stag. (Webb, 1978)

However, not all comparison riddles use metathesis, and not all riddles containing metathesis pairs are comparison riddles.

4 Overview of the model

The mechanisms described above — juxtaposition, substitution, and comparison — all work by constructing a segment of text not already in the lexicon. The riddle then uses the constructed word or phrase as if it were a semantically sensible construction. The effective meaning of this construction is a combination of the meanings of the pieces of text used to build it. For example:

> (9) What do you give an elephant that's exhausted?
>
> *Trunkquillizers.* (Webb, 1978)

In this joke, the word *trunk*, which is confusable with the syllable *tranq*, is substituted into the valid English word *tranquillizer*. The constructed word *trunkquillizer* is given a meaning, referred to in the question part of the riddle, which is some combination of the meanings of *trunk* and *tranquillizer*:

> *trunkquillizer*: A tranquillizer for elephants.

This is not the only meaning for *trunkquillizer* that could produce valid jokes. For example:

> (10) What kind of medicine gives you a long nose?
>
> *Trunkquillizers.*

is a joke (if not a good one) based on a different definition for *trunkquillizer* — namely, "a medicine that gives you a trunk". The constructed meaning combines notable semantic features of both of the valid words/phrases used to construct it, so that the riddle question is a reasonable description of the constructed concept. Compare the way that, in non-humorous communication, a real word/phrase can be the answer to a question:

> (11) What do you call someone who douses flames?
>
> *A firefighter.*

The *trunkquillizer* example uses the constructed meaning of a constructed word or phrase to build a question that *would* have that word or phrase as its answer, if it really existed. This question becomes the first part of the riddle, and the constructed word/phrase becomes the punchline.

At this point, it is important to distinguish between the task of building the *meaning* of the constructed segment, and the task of using that meaning to build a question with that segment as an answer. For example, the following questions use the same meaning for *trunkquillizer*, but refer to that meaning in different ways:

(12) What do you use to sedate an elephant?

(13) What do you call elephant sedatives?

(14) What kind of medicine do you give to a stressed-out elephant?

There are therefore several different jobs to be done in constructing a joke of this type: a non-lexicalized word/phrase must be constructed, a plausible description of the meaning for that segment must be built, and a question-answer pair must be constructed, using the word/phrase and the description of its meaning. All of these tasks must be supported by a suitable amount of lexical information. In this model, step one is done by the *schemata*, step two is done by the *description generator*, and step three is done by the *templates*.

Fuller details of these mechanisms can be found in (Binsted, 1996), but they can be briefly summarised as follows. A schema stipulates relations between lexical items and constructed items (and their possible descriptions). This provides, roughly speaking, the underlying semantic configuration of the riddle, and contains all the real 'humorous' information within our rules. The description generator formulates a possible description (in terms of semantic information from the lexicon) of some (imaginary) word or phrase constructed by an instantiated schema; this does no real humorous work. A template takes all the various lexical material that has been arranged by the schema and the description generator and produces actual English text, using a *sentence-form*; i.e. a mixture of fixed text (e.g. *What do you call a* ___) and slots to be filled. Each template includes constraints on what lexical material can be used to fill the slots in its sentence-forms.

As noted above, all this is supported by a lexicon, which must contain semantic, syntactic, phonological and orthographic information about a large range of words and common phrases (e.g. *serial killer*). In our implementation, the central part of JAPE's lexicon is WordNet (Miller et al., 1990) for the syntactic, semantic and orthographic information. What is particularly useful about WordNet is the fact that it classifies word-senses into a network of synonyms, hyponyms, etc., which allows more subtle processing than might otherwise be possible. For phonological information, JAPE's lexicon also includes the British English Example Pronunciation dictionary (Robinson, 1996), and a homonym list (Townsend and Antworth, 1993). JAPE also gives each output text scores from the MRC psycholinguistic database (Wilson, 1987) (see section 6 below). However, that database is not used directly in pun generation.

Although we defined a total of 13 schemata (see Appendix A), only 9 of these were used in the evaluation, as the others required lexical information that WordNet did not supply. There were 21 templates.

5 Background – JAPE-1

A preliminary version of the program, JAPE-1, was constructed in 1993, and has been described elsewhere (Binsted and Ritchie, 1994; Binsted and Ritchie, 1997). That prototype used slightly cruder notions of 'schemata' and 'template', and the role of the description generator was filled by a less elegant mechanism.

Also, it used a small, specially created lexicon, although care was taken to ensure that the lexical entries were general purpose (not aimed at producing humour).

An informal exploratory evaluation of that system was carried out (see (Binsted and Ritchie, 1997) for more details). Primarily, that evaluation was to point the way to improvements in the model behind the system. This information was then used both to improve JAPE-1's design, and to guide the path of the research. The exploratory evaluation was *not* intended to be a rigorous examination of JAPE-1's abilities.

Summarising very briefly, various riddles produced by JAPE-1 were presented to volunteer adult subjects along with a questionnaire asking both for numerical ratings of the jokes (on a scale of 1 to 5) and qualitative comments. Although the approach used had limitations, such as the failure to include human–generated jokes and nonsense texts for comparison, it did provide useful qualitative information that guided the development of JAPE-2.

The average point score for all the jokes JAPE-1 produced was 1.5 points, over a total of 188 jokes. Most of the jokes were given a score of 1 ("a joke, but a pathetic one"). Interestingly, all of the nine jokes that were given the maximum score of five by one judge were given low scores by its other judge — three got zeroes, three got ones, and three got twos.

For the purpose of informing the further development of our model (and JAPE-2), the comments of the volunteer joke judges were more useful than the numerical scores they gave the texts. It was clear from the comments that the quality of the lexicon greatly influenced the quality of the resulting jokes. We learned that:

Semantic information should be included in the lexicon only if it is typical of the word being entered, so that the associations necessary for understanding the joke can be made by the audience.

The information in the lexicon should be common knowledge, again, so that the audience is able to understand the references in the joke.

Jokes should avoid using very general words (e.g. *object* or *person*), and instead use specific words (e.g. *hammer* or *carpenter*) more closely associated with the key concepts in the joke.

The templates varied a great deal in the quality of jokes they produced. A general comment that was made several times was that some of the questions were not "logically coherent" in some sense. This incoherence was often the result of using one of the templates with inappropriate word ordering. Scrutiny of JAPE-1's mechanisms suggested that this was a symptom of an overly simple interface between the schemata and the templates. This led to the inclusion of the description generator in our model.

6 Measuring readability

It was important in our evaluation to consider the influence of the 'readability' of the texts being judged by the children. It is hard to devise a realistic readability measure which can be applied mechanically to large quantities of riddles, but we decided to base our measure on the MRC psycholinguistic database (Wilson, 1987; Coltheart, 1981). This is a large lexical database compiled from a wide variety of sources. There are a total of 150837 words, and each entry has associated with it various linguistic and psychological data, ranging from the phonetic transcription to the frequency of occurrence in various corpora (e.g. (Francis and Kučera, 1982)). Not every entry has a value for each of the 26 possible headings; for example, "age of acquisition" is listed for only 3503 entries, whereas the part of speech is entered for 150769 items. The properties we are interested in here are a word's *familiarity, concreteness*, *imageability* and *age of* *acquisition*, all of which are already normalised to lie in the range 100 to 700 (Coltheart, 1981). We made the working assumption that these measures would contribute to ease of reading.

Our metric works as follows. Each text (riddle or control item) is given four scores from the psycholinguistic database: familiarity, concreteness, imageability and age of acquisition. To do this, each key word (i.e. word not supplied by the fixed text of the sentence-form of the template) in the text is given its scores for these four measures from the database. We assume that the readability of the whole text is limited by the *least* readable of the words that make it up. That is, a text is only as familiar (or imageable, or concrete) as the least familiar word it contains. For this reason, the scores for the whole text are taken to be the *worst* (i.e. lowest for familiarity, concreteness and imageability, and highest for age of acquisition) of the scores for the key words in that text. If a key word has no score for a particular measure, then no score for that measure for that text is recorded. Because age of acquisition data is quite sparse in the database, this procedure leaves very few texts with age of acquisition scores, and so that measure was abandoned for the remainder of the analysis.

For example, to calculate the psycholinguistic scores for the text:

(15) How is a shark like a bass?

They are both fish.

we take the set of key words (*shark*, *bass* and *fish*), and find their scores in the MRC database. They are:

shark: Familiarity: 516

Concreteness: 611

Imageability: 602

Age of acquisition: No score

bass: **Familiarity:** 540

Concreteness: 547

Imageability: 544

Age of acquisition: No score

fish: **Familiarity:** 548

Concreteness: 597

Imageability: 615

Age of acquisition: No score

The worst scores for each measure are:

Familiarity: 516

Concreteness: 547

Imageability: 544

Age of acquisition: No score

These are taken as the scores for this text.

`***** Figure 1 about here *****`

7 Confirmatory evaluation

7.1 Introduction

Once the development of the JAPE-2 model was completed, and its implementation finished, it was necessary to evaluate rigorously its performance. The purpose of this confirmatory evaluation was to determine whether or not the JAPE-2 program was able to generate punning riddles of a similar quality to those generated by human joke experts. A secondary goal was to discover which of JAPE-2's output texts were of the highest quality, and why. These goals can be reformulated in terms of the following research questions:

1. Are JAPE-2's output texts jokes?
2. If so, how funny are they? Are they as funny as human-generated jokes of the same type?
3. What in JAPE-2's jokes contribute to their quality? Is it the way in which the pun is constructed, the subject matter of the joke, or some other factor?
4. Has JAPE-2 replicated any human-generated jokes?

In order to answer these questions, JAPE-2 output texts, human-generated joke texts, and non-joke texts were evaluated by a large number of children.

7.2 Hypotheses

The purpose of this study was to summatively evaluate (Mark and Greer, 1993) the behaviour of a pun-generating system, JAPE-2. Since the behaviour being evaluated was pun generation, the experiment compared JAPE-2's output with human-generated puns, and with a control group of non-puns. We hoped to show that:

1. JAPE-2's output texts are more joke-like than non-jokes of a similar form. That is, joke 'experts' judge JAPE-2's output texts to be jokes significantly more often than they judge non-jokes (in question-answer form) to be jokes.

2. JAPE-2's output texts are jokes. That is, on average, joke 'experts' do not judge JAPE-2's output texts to be jokes significantly less often than they judge human-generated punning riddles to be jokes.

3. JAPE-2's output texts are not less funny than puns of the same type generated by humans. That is, joke 'experts' do not give JAPE-2's output texts significantly lower funniness scores than they give to human-generated punning riddles.

Before the above hypotheses could be evaluated, it was necessary to establish that our joke judges were, in fact, experts. That is:

4. The joke 'experts', on average, judge non-jokes to be jokes significantly less often than they judge human-generated punning riddles to be jokes.

There were also some secondary questions that this experiment aimed to answer. These were exploratory questions that could inform further development of the model, or aid other humour research.

1. Are there any correlations between the age, year in school, or reading ability of the joke judges, and the judged quality of the human-generated jokes?

2. Is there any relationship between the form of the texts and their perceived funniness or joke/non-joke status?

3. Is there any relationship between the schemata used to generate the JAPE-2 jokes and the perceived funniness or joke/non-joke status of the joke?

4. Have any of JAPE-2's jokes been heard before?

The experiment was designed so that the main hypotheses could be addressed, and the data gathered could inform the secondary questions.

7.3 Design

7.3.1 Subjects

Other experiments (e.g. (Yuill and Easton, 1993)) have suggested that native English speaking children aged 8-11 years old are able to judge whether or not a spoken text is a pun. This age group is also able to judge the funniness of a spoken text. That is, children in this age group, having heard a text, can say whether or not that text is a pun, and can say how funny they thought it was, at least on a scale of 0-2. They can also be expected to consistently judge at least twenty texts without losing concentration. Moreover, children of that age group are likely to both appreciate and understand jokes of the punning riddle genre (Ruch et al., 1990). For these reasons, 8-11 year old children were the best choice for experts on punning riddles for the purposes of this experiment.

However, there were also some problems with this group of judges. We did not know whether a child's judgements would be internally consistent, or consistent with those of other children. Also, we could expect the reading ability in this age group to vary considerably. Finally, we could expect children in this age group to be influenced by the subject matter and reading level of the texts as well as their pun nature (or lack thereof).

Therefore, the texts being judged had to be carefully selected to control for subject matter and reading level. Moreover, each child was asked to judge whether or not non-puns (i.e. sensible or nonsense questions and answers, with no punning element) were jokes, so that their ability to distinguish jokes from non-jokes could be established. Also, each child's year at school and age were recorded, so that variations in these could be taken into consideration.

Since the children's reading abilities were uncertain, they were exposed to the texts in both written and recorded form. In order to avoid bias in the experimenter's reading of the texts, an actor was asked to read all the texts onto tape with the same voice and general intonation pattern.

7.3.2 Materials

The initial materials were the set of texts that the judges were to evaluate. Because the purpose of this experiment was to compare JAPE-2-generated texts with human-generated punning riddles, representative examples of each were included. The judges' ability to distinguish jokes from non-jokes also needed to be checked, and so non-jokes were also included; the non-jokes also acted as a control set for the evaluation of the actual jokes.

There are two relevant kinds of non-jokes — sensible question-answer pairs, and nonsense question-answer pairs. A sensible question-answer pair is a two sentence text in which the second sentence is a truthful, expected answer to the question in the first sentence. For example:

(16) What do you get when you cross a horse and a donkey?
A mule.

A nonsense question-answer pair is a two sentence text with the syntactic form of a question and answer, but without any connection, sensible or punning, between the topic(s) of the question and that of the answer. For example:

(17) What do you get when you cross a murderer and a ferry?
A citrus fruit.

Thus, four sets of materials had to be prepared: the JAPE-2 generated texts, the human-generated jokes, the sensible non-jokes, and the nonsense non-jokes.

Since we are only interested in humour caused by the pun nature of a text, it was also necessary to control for subject matter and form of the joke. For this reason, all the texts in the experiment had subjects selected from the same set of possible subjects, where the *subject* of a text is the set of nouns, adjectives and verbs used in that text (see appendix B for a list of the permitted vocabulary). Also, all the texts in the experiment had forms selected from the same set of possible forms, where the *form* of a text is one of JAPE-2's twelve sentence-forms (see appendix C).

It was important that each text be judged by several different children, so that their evaluations could be compared. However, children in this age group are easily bored, and their ability to judge texts can be expected to deteriorate if they are shown large numbers of texts. We estimated, after (Yuill and Easton, 1993), that children in this age group could judge twenty texts without becoming too distracted from their task. For this reason, the texts were divided into sets of twenty. Each set had approximately the same number of each type of text. In order to eliminate ordering effects such as boredom, each set of texts was also randomised into several different sequences of texts.

A set of JAPE-2 texts was selected to be representative of JAPE-2's output. In order to obtain data for this experiment's secondary question on the relationship between readability and funniness (see subsection 7.2), it was important that all of the vocabulary used in the texts also be in the MRC psycholinguistic database (Wilson, 1987), which contains data relevant to measuring readability.

The following procedure was used to generate the texts to be used as the materials in this experiment.

1. For each of JAPE-2's nine schemata used in the evaluation, JAPE-2 generated as many output texts as could be generated using words from the MRC database.

2. For each schema, if the schema generated *more* than one–ninth (nine schemata were tested) of the required number of texts, we chose the texts

with the highest familiarity, concreteness, and imageability scores (see section 7.5.3), yet which had different subjects (i.e. did not share any nouns, verbs or adjectives).

3. Some schemata generated *fewer* than one–ninth of the required number of texts. The shortfall in the total number of texts was made up by adding the excess texts (from the over–generating schemata) with the highest familiarity, concreteness and imageability scores.

To find the set of human-generated texts:

1. We went through the selected JAPE-2 texts, and determined which of JAPE-2's sentence-forms were used in the generation of these texts. These, stripped of the semantic constraints on lexical items, became the set of allowable forms (see appendix C).

2. We went through the selected JAPE-2 texts, and made a list of all the nouns, verbs and adjectives used in these texts. This set of words, and all of their sister and daughter nodes in WordNet's hyponym hierarchy, became the set of allowable vocabulary items (see appendix B).

3. From published books of jokes, not examined during JAPE's development ((Anderson, 1987), (Byrne, 1995), (Abbott, 1993), (Fremont, 1993), (Churchill, 1976), (Phillips, 1991), (Jam, 1991), (Rayner, 1991), (Girling, 1988), (Alec, 1987), (Young, 1993), (Brandreth, 1990), (Hegarty, 1992), and (Forrester, 1994)), we selected all the jokes which use only the allowable forms and subjects. Minor adjustments of sentence-forms and subjects were done by an impartial adult, in order to fit the human-generated jokes to the experimental criteria. In other words, if one of the jokes was almost in an allowable form, minor syntactic changes could be made;

likewise, minor changes in a joke's vocabulary could be made to give it an allowable subject. What constituted a 'minor' change was left to the impartial adult's discretion, as long as the resulting text was, to their judgement, a joke.

For example, the joke:

> (18) What do sea-monsters eat?
>
> *Fish and ships.* (Young, 1993)

was adjusted in both vocabulary and form to read:

> (19) What kind of food does an octopus eat?
>
> *Fish and ships.*

so that both its form and its subject were allowable.

4. There were more suitable human-generated jokes than required, so we randomly chose the required number.

To find the set of sensible non-jokes:

1. The set of permissible subjects and the set of permissible sentence-forms were given to an impartial adult. She was asked to fill in the blanks of the forms with the subjects in as many ways as she could to produce 'true' questions and answers.
2. The resulting set of texts were given to a second impartial adult. He was asked to eliminate any which did not 'make sense'.
3. The resulting set was larger than required, so a suitable number of texts were randomly chosen from the set.

In order to determine the set of nonsense non-jokes:

1. Allowable subjects were inserted randomly into the permissible sentence-forms. (This was done using the part of the JAPE-2 program which carries out the final stages of constructing surface texts).

2. If any of the resulting texts accidentally happened to be either sensible question-answer texts or punning riddles, as judged by an impartial adult, then they could be eliminated from the set. This situation, however, did not arise.

3. The resulting set was larger than required, so we randomly chose a suitable number of texts from the set to use in the experiment.

In this way, all the texts had subjects selected from the same set of subjects (i.e. the subjects used in the JAPE-2 texts), and forms selected from the same set of forms (i.e. JAPE-2's sentence-forms). Moreover, they were all rateable using the MRC psycholinguistic database (see section 6). The initial set of texts was then divided into test sets of twenty texts each. Texts of each type (i.e. JAPE-2 generated, human generated, sensible and nonsense) were spread evenly across the sets. Each test set was then randomised and recorded. Since we could not be sure that each judge would finish judging their set of texts, and to eliminate ordering effects, each test set was randomised into several different sequences, so that each text was likely to be judged the same number of times. We recorded each of these sequences on a separate tape, marking it carefully (e.g. Test Set 3, Ordering 2). All the texts were recorded with the same voice, using the same intonation patterns.

In addition, one human-generated joke and one sensible non-joke was also recorded, to be used as examples for the judges. The example texts were not from the test material set.

7.3.3 Equipment and setting

The response sheets used in this experiment contained: some simple printed instructions to supplement those given by the experimenter; an area for the the child to fill in their name, age, year or form at school, and whether or not they like jokes; and twenty numbered response areas, one for each text. In each response area, there were three questions relating to the text heard on the tape:

- Was that a joke? In response, they were to circle either a "YES" or a "NO".
- How funny was it? In response, they were to circle one of five simple faces: frowning mouth open, frowning mouth closed, flat mouthed, smiling mouth closed or smiling mouth open. Under the faces there was text saying "not funny at all" "not very funny" "not sure" "funny" and "very funny".
- Have you heard it before? In response, they were to circle either a "YES" or a "NO".

There was also some space for comments at the end of the response sheet, which both the experimenters and the judges could fill out if necessary. The comments were not taken as part of the formal evaluation. Each response 'sheet' was made up of several pages, stapled together.

The example response sheets had two numbered response areas like those on the main response sheets. No other information was recorded on these; they were used only to familiarise the children with the procedure.

7.3.4 Procedure

The tape (i.e. sequence of texts) to go in each machine was chosen randomly. Children carried out the evaluation in groups of no more than five, and they were asked to confirm that they were between eight and eleven years old and that English was their native language. They were also asked if they could read, but this was not the sole check on reading ability, as the initial practice at using the response sheets (see below) acted as a further check on their literacy.

Before the experiment started, the experimenter explained that we needed their help in deciding whether some 'things' are jokes or not. The experimenter told them they were to listen to the tape in their machine, and tell us on the response sheet if what they heard was a joke, how funny it was and whether they had heard it before. These instructions were repeated briefly at the beginning of each tape, and at the top of each response sheet. The experimenter also explained that, should they wish to stop at any point, they should raise their hand and they would be allowed to go. Finally, the experimenter told the children not to tell any jokes heard during the experiment to other children, because the other children might want to participate in the experiment too. No mention was made of the fact that some of the texts were computer-generated.

The experimenter then asked the children to listen to the example tape, and fill in the example response sheet. Any obvious misunderstandings about the procedure (as opposed to the nature of puns, the meanings of words etc.) were corrected at this point.

The experimenters then helped the children fill in the first part of the response sheet, which asked for the age and year or form at school of the child. It also asked if they like jokes or not. The experimenters then started each tape recorder, ensuring that the children could hear the tapes clearly.

As the tapes finished, the experimenters asked if the children had any comments about what they heard. If the children could not write the comments themselves, the experimenters made brief notes for them on the sheet. The children were then allowed to go.

Including instructions and the writing of comments, the experiment took no more than 20 minutes of each child's time. Including turn around, each cycle took no more than 30 minutes. Each tape (i.e. sequence of texts) was judged at least once.

7.4 The experiment

The pilot for this evaluation took place at a primary school, where twenty children evaluated 100 texts (50 JAPE-2 texts, 30 human jokes, 10 sensible non-jokes and 10 nonsense non-jokes). The experiment went smoothly, and there were some significant results: the 'jokiness' of both the human–generated jokes and the JAPE-2 texts was higher than that of the non-jokes ($p < 0.005$); and the 'jokiness' of the human–generated jokes was significantly higher than that of the JAPE-2 texts ($p < 0.005$).

The main experiment took place over two days at the 1996 Edinburgh International Science Festival on an April weekend. 122 children took part in the experiment, most aged between eight and eleven years old, although a few slightly older or younger siblings were permitted to participate at the request of their parents. Two hundred texts were judged in the experiment. There were one hundred JAPE-2 generated texts, sixty human-generated texts, twenty sensible non-jokes, and twenty nonsense non-jokes. These were evenly divided into ten sets of twenty texts, which were then randomised into forty sequences.

There were no technical hitches, although two minor errors in the materials were detected too late to be fixed: one JAPE-2 text was included in two sets, and one questionnaire (seen by three children) contained one incorrect text (although the correct text was on the tape).

Almost all of the children were able to follow the instructions without any problems. One child did not fill in any funniness data, while another missed a page in his questionnaire. One child seemed to have significant difficulties reading the texts, and this was noted on his questionnaire. The remainder of the questionnaires were correctly filled out. All of the children behaved well, and all completed the full experiment. Our 'room' was a corner of a large hall, separated from the rest of the space with room dividers, and was quiet enough for our purposes.

7.5 Results

This confirmatory evaluation provided adequate data to assess the hypotheses described in section 7.2. It also gave some significant answers to some of the secondary research questions.

Of the forty sequences of texts, thirty-eight were evaluated three times, and two were evaluated four times. This means that each of the ten sets of texts was evaluated at least twelve times.

Some of the 122 questionnaires returned contained data that was flawed in some way:

- Two questionnaires were filled in by seven year olds (both siblings of other subjects). Both said they were almost eight.
- Four questionnaires were filled in by twelve year olds (again, siblings of other subjects). Three said they had just turned twelve.
- One text sequence was marred by a mismatch between the tape and the questionnaire. One text was read correctly on the tape, but the questionnaire contained a different text. This flawed sequence was evaluated by one child before it could be corrected.
- One child had obvious difficulty reading the questionnaire.

- One of the seven year olds did not fill in any funniness data.
- One child missed a page in the questionnaire. We were unable to tell whether the remainder of his responses corresponded to the appropriate texts.

Of these, only the last two were discarded completely. The three containing the mismatch were assumed to be correct otherwise, and only the data on the mismatched text was discarded. The rest have been included in the data, but the problems with them have been noted. Even after these deletions, most of the two hundred texts have been evaluated by twelve children, and all have been evaluated by at least nine.

For each text, three types of evaluation were given by the children.

Jokiness: Each text is given a zero score if evaluated as a non-joke, and a score of one if evaluated as a joke. If not evaluated (i.e. that part of the questionnaire was not filled in), no score is given. The average of all the scores for that text is taken to be the 'jokiness' of the text (i.e. the proportion of the children who judged it to be a joke).

Funniness: Each child gave each text a score from 1 ("not funny at all") to 5 ("very funny"). If a text was not evaluated for funniness, it is not given a score at all. The children were not given instructions on how to rate the funniness of a non-joke. For this reason, if a child rated a text as a non-joke, the funniness score that child gave that text was discarded. The average of all the "How funny is it?" scores for a text is taken to be the 'funniness' of the text.

Heard before: Each text is given a zero score if not heard before, and a score of one if heard before. The average of all the "Have you heard it before?" scores for a text is that text's 'heard before' score.

The 'jokiness', 'funniness', and 'heard before' scores for each text have been given in appendix D. The texts are ordered first by 'jokiness', then by 'funniness'.

7.5.1 Jokiness

The average 'jokiness' of each type of text was calculated, and is shown in figure 2. Then the significance of the differences in 'jokiness' was calculated, using the Wilcoxon Signed Rank Test (Greene and D'Oliveira, 1992). It was found that:

- The children found sensible non-jokes and nonsense non-jokes equally 'joke–like'. That is, there is no significant difference ($p > 0.05$) between the 'jokiness' scores of the two types of non-jokes. For this reason, we do not distinguish between the two types of non-jokes for the rest of the 'jokiness' analysis.

- The children could distinguish human jokes from non-jokes. That is the 'jokiness' of the human-generated texts is significantly ($p < 0.01$) higher than that of the non-joke texts. This confirms hypothesis 4 in section 7.2.

- The 'jokiness' of the JAPE-2 generated texts is significantly ($p < 0.01$) higher than that of the non-joke texts. This confirms hypothesis 1 in section 7.2.

- The 'jokiness' of the human-generated jokes is significantly ($p < 0.01$) higher than that of the JAPE-2 generated texts. This fails to confirm hypothesis 2 in section 7.2.

`***** Figure 2 about here *****`

A secondary research goal (see section 7.2) was to compare the success of JAPE-2's various schemata at generating jokes. To do this, the JAPE-2 generated texts have been categorised according to the schema that generated them. The 'jokiness' scores for each type have then been compared (figure 3). Because not all schemata were able to generate a large number of texts, most of the differences are not significant. The exceptions are that the **phonsub** schema generated texts with significantly higher 'jokiness' scores than both the **lotus** schema and the **rhyming** schema.

`***** Figure 3 about here *****`

7.5.2 Funniness

`***** Figure 4 about here *****`

Similar calculations were done for the 'funniness' scores of the types of text (see figure 4). The results are:

- There is no significant difference in funniness between the two types of non-joke ($p > 0.05$).
- Human-generated jokes are significantly funnier than non-jokes ($p < 0.05$).
- JAPE-2 generated jokes are significantly funnier than non-jokes ($p < 0.05$).
- Human-generated jokes are significantly funnier than JAPE-2 jokes ($p < 0.05$).

Recall that a particular funniness score is only used if the child who gave it also judged that text to be a joke. This is because children were not given any instructions on how to judge the funniness of a non-joke text, and they adopted several different, and inconsistent, strategies. When evaluating the funniness of texts that they judged not to be jokes, some gave only the lowest score, some gave a range of scores, and some gave no funniness score at all.

7.5.3 Interactions with readability

Several of the secondary research questions (see section 7.2) relate to the 'readability' of the texts (see section 6). The first step in the readability analysis was to compare the average readability scores for human-generated texts and for JAPE-2 texts (figure 1). All of the differences are significant; that is, human-generated jokes are significantly more familiar ($p < 0.01$), concrete ($p < 0.005$), and imageable ($p < 0.001$) than JAPE-2–generated jokes.

The next step was to find out if there was a correlation between any of the three psycholinguistic measures and the 'jokiness' scores for the human-generated jokes. No significant correlation was found for any of the measures. The same test was then performed for JAPE-2 jokes alone. Again, none of the correlations were significant.

The data for the human-generated texts and the JAPE-2 generated texts were then grouped together, to see if this larger set of texts would show a significant correlation between the psycholinguistic scores and the 'jokiness' of the texts. In fact, three correlations were found. A significant ($p < 0.01$) correlation between the familiarity score for a text and its jokiness was found, with a Spearman coefficient of .23. A correlation between the concreteness of a text and its 'jokiness' was also significant ($p < 0.002$), with a Spearman coefficient of .2878, as was the correlation between imageability and 'jokiness' ($p < 0.001$), with a coefficient of .2818. This would seem to indicate that there is a small but significant correlation between readability, in the form of the three psycholinguistic measures, and the jokiness of a text (i.e. the fraction of children rating it as a joke).

We then tried to correlate readability and 'funniness' scores for human-generated jokes. None of the correlations were significant. The same test was then performed for JAPE-2 jokes alone. Again, none of the correlations were significant. We then grouped together the data for the human-generated texts and for the JAPE-2-generated texts. We found that there was a correlation between familiarity and funniness ($p < 0.02$, coefficient .2121), concreteness and funniness ($p < 0.001$, coefficient .3152), and imageability and funniness ($p < 0.001$, coefficient .3697). This indicates that there is a small but significant correlation between the readability and the funniness of a text.

7.5.4 Other Results

The 'heard before' data for human-generated jokes and JAPE-2 jokes was also compared. For those texts judged to be jokes, more children claimed to have heard the human-generated before than claimed to have heard the JAPE-2 jokes before. This result is both statistically significant and completely unsurprising.

It was also expected (see section 7.2) that there would be some correlation between the age of the children and their ability to identify jokes. The ages of the participants in the experiment were compared with their ability to identify jokes. A child's *joke recognising ability* is taken to be the proportion of human-generated jokes that they successfully recognised as jokes. Using the Spearman test (Siegel and Castellan Jr, 1988), it was found that, although the correlation coefficient between the age of the participant and the ability to identify jokes was not large (.2596), it was significant ($p < 0.002$).

7.5.5 Improvement by elimination

One of the purposes of this evaluation was to test systematic ways of eliminating poor output texts automatically. This could be done by removing parts of

JAPE-2 which do not work as hoped, or by constructing an output filter, that eliminates texts that do not meet some criteria.

The analysis of the results shows that some schemata are significantly better than others. In particular, the schemata which require information not readily available in WordNet — **lotus, rhyming, poscomp** and **elan** — performed badly. It was possible that these schemata, hindered by lack of information, were producing poor outputs texts which, in turn, were bringing down JAPE-2's average performance.

To check this, we eliminated all the texts produced by the underinformed schemata, then recalculated the averages and significance. Although this increased the average jokiness of the JAPE-2 output texts to 0.68, the difference between this and the average jokiness of the human generated texts was still significant ($p < 0.05$). However, if the **bazaar** schema, which uses the weak juxtaposition mechanism (see section 3), is also removed, difference in jokiness between human generated texts and the remaining 28 JAPE-2 generated texts is no longer significant ($p = 0.2$). This suggests that eliminating poor or underinformed schemata would improve the overall quality of JAPE-2's output.

`***** Figure 5 about here *****`

Another approach would be to filter JAPE-2's output after generation, according to the psycholinguistic scores for the texts. To check this, we eliminated all the JAPE-2 texts with any psycholinguistic score below the average for that score for human-generated jokes. This did not significantly improve the quality of JAPE-2's output texts. Then, we removed the texts scoring (on any measure) below 350, 375, 400 and 450, and charted the results (figure 5). At the 400 threshold, the difference between the jokiness of the human-generated texts (0.80) and that of the remaining 20 JAPE-2 texts (0.72) was no longer significant ($p = 0.12$). At the 450 threshold, the difference between the jokiness of the human-generated texts and the remaining 9 JAPE-2 generated texts is not significant at all ($p = 1$). This suggests that eliminating those texts with low psycholinguistic scores would improve the overall quality of JAPE-2's output.

8 Discussion

8.1 Subjects

As noted earlier, the choice of 8-11 year olds seemed reasonable, based on relevant psychological research (e.g. (Yuill and Easton, 1993)(Ruch et al., 1990)). However, we found a significant correlation between the age of the judge and her or his ability to recognise a human-generated punning riddle as a joke (see Section 7.5.4 above).

Also, grouping the human-generated and computer-generated jokes together, we found a small correlation between the various 'readability' measures for a joke and the average 'jokiness' rating it received: for familiarity, $p < 0.01$, coefficient .23; for concreteness, $p < 0.002$, coefficient .2878; for imageability, $p < 0.001$, coefficient .2818.

It is likely that these two results are related. If joke recognition goes up with the readability of the text, and if older children tend to be better readers, then older children should be better able to understand the joke texts, and therefore be better able to recognise them as jokes. This would suggest that a slightly older age group might be appropriate for this kind of study, despite evidence that 8-11 year olds *appreciate* puns most. Our interpretation of these results could be confirmed with further experimentation.

8.2 Materials

There were some minor errors in the materials, but it is unlikely that they had a significant effect on the results (see section 7.4).

More importantly, the questionnaires did not include a “not sure” option in their yes/no questions. This forced the children to judge whether or not a given text was a joke, even if they were not sure. This may have had a positive effect, however. We suspect that many children, faced with an unfunny joke, would have marked the “not sure” box had it been available — rather than the “yes” option for ‘jokiness’ and the “not funny at all” option for ‘funniness’, as we would have wanted them to. Since many of the texts were not very funny jokes, this might have been a serious problem. Not having a “not sure” box forced the children to give a definite answer to “Was that a joke?”, and may well have biased the results (for all texts) towards “yes”.

Also, no clear instructions were given to the children on how to mark the funniness of texts that were judged to be non-jokes. As a result, a variety of strategies were followed, such as giving the non-jokes the lowest funniness score, or not giving non-jokes a funniness score at all. This inconsistency meant that we were unable to decide whether or not the children thought non-jokes could be funny, which would have been an interesting secondary result.

8.3 Filtering

One of the main purposes of the evaluation was to compare jokes generated by JAPE with those generated by humans. However, the human–generated jokes had one significant advantage: they were also *filtered* by humans.

For a joke to be remembered and retold, it must be quite good; for it to be included in an edited collection of jokes, it must be very good (compared to the range of possible jokes). All of the jokes used in this study came from published books of jokes, so must have gone through some sort of filtering process. JAPE's jokes, on the other hand, were minimally filtered (see section 7.3.2) before the evaluation (although the evaluation did suggest some ways in which they might be filtered in the future).

Although we would claim that JAPE's output texts are all well-formed *puns*, they are not all good *jokes*. Unfortunately, an automatic filtering process, to parallel the human filtering described above, would require a system that could *appreciate* humour — a much more difficult task than humour *generation*.

Another approach would be to collect a set of heuristics that, given well-formed puns, could order them according to expected funniness. Such heuristics could include preferring short jokes to long ones, preferring jokes which use slang or 'rude' words, and preferring jokes which contain accidental (i.e. not required by a schema) alliteration or rhyme (see section 5). However, such simplistic heuristics are bound to be quite crude, and would probably be theoretically uninteresting.

9 Some consequences for our model

9.1 Adjustments to the knowledge

If the output texts and their various successes and failures (see appendix D) are qualitatively compared, two main kinds of failure stand out. Some contain words that the average 8-11 year old is unlikely to know. For example:

(20) What do you call a lenient shelter?

A lax deduction. [JAPE]

Not only are *lenient*, *lax*, *tax shelter* and *deduction* quite difficult vocabulary, the joke is based on the compound nominal *tax deduction* — a phrase with which most children are unfamiliar. (For those who don't 'get it', JAPE figures that a tax deduction is a kind of shelter, so a lax deduction would be a lenient shelter.)

In other jokes, the words themselves are familiar, but the *sense* of at least one of the words used in the joke is not:

(21) What kind of curve has cheek?

A nerve ball. [JAPE]

In this joke, all of the words themselves should be familiar. However, understanding the joke requires that the listener be familiar with the term *curve ball* (a baseball term), and also know that *cheek* can mean *nerve* (as in *She has a lot of cheek saying that!*). Most British children would not have this knowledge, which includes both American and British slang.

Some jokes apparently failed because they simply do not make sense, linguistically. For example, WordNet contains the information that *running away* is a compound nominal, and is a kind of *feat*. This led to the 'joke':

(22) What do you call a clever feat?

Cunning away. [JAPE]

In building its descriptions, JAPE assumes that the last word in a compound nominal is the noun being modified, and the other words are the modifiers. In *running away*, this is not the case, resulting in both the constructed phrase (*cunning away*) and its description (*a clever feat*) not being well-formed linguistically.

Finally, a few poor jokes resulted from schemata not performing as expected. For example:

(23) How is an ugly insect like a deep kinswoman?

They're both bass aunts. [JAPE]

This is a positive comparison riddle which uses the phrases *base ant* (ugly insect) and *bass aunt* (deep kinswoman). Unfortunately, the schema that generated this text only gives *one* of the constructed phrases, assuming its homophonous pair will also be brought to mind. A better example of a riddle generated by this schema is:

(24) How is a nice girl like a chocolate birdie?

They are both sweet chicks. [JAPE]

This particular problem can be corrected by constraining the **poscomp** schema to use words with two senses, rather than homonyms (which are spelled differently), to construct its jokes.

We would draw three conclusions from the above.

- If a joke relies on a word or association which is not familiar to the listener, or which is too weak to bring the target concept to mind, then the joke will probably fail. Unfortunately, no currently available linguistic resource gives psycholinguistic information about the *sense* of words; that is, none could distinguish between the familiarity of *cheek* the body part and *cheek* the attitude.

- Joke texts must follow (most of) the linguistic principles governing the syntax and semantics of grammatical texts, *even though the resulting texts may be nonsensical.* That is, semantic constraints derived from the syntax of the text must be satisfied, even though the text may not have a semantic interpretation which makes sense in the 'real world'. Although our model was designed with this in mind, some resulting texts were still ill-formed (e.g. *cunning away*).

- Jokes are similar to puzzles, in that the mental effort required to 'solve' them is part of their pleasure. However, if a joke is too complex, it may not be understood at all. In several of JAPE's output texts, the combination of unfamiliar words and a complex schema led to puns that were too convoluted to understand.

9.2 Implementation considerations

As a result of the semantic requirements set out in section 9.1 above, any computational testing of the model would need more sophisticated linguistic resources if it was to generate riddles of the same quality as those generated by humans. are required. In particular, the system would need:

- A wider range of semantic links between words, especially between words of different syntactic categories. For example, a link between a object and an action that object is likely to perform (e.g. between *bomb* and *explode*) would greatly enrich the associations which could be used in jokes. (JAPE-1 contained a mechanism of this sort, which seemed to be useful.)

- Psycholinguistic information relating to the familiarity of particular senses of words, so that comprehensible jokes can be constructed based on associations that the listener is likely to know.

We believe that large knowledge bases containing such information would be useful for other applications in natural language research as well.

Alternatively, this problem could be avoided by building a system for *human-assisted* pun generation. Such a system would work much the same way JAPE does, but it would not rely on its lexical resources for good word-word associations. Instead, it would prompt the user for typical associations; for example, it could ask the user what a *bomb* typically does (hopefully getting the answer *explode*), rather than relying on its lexicon to provide this information, which it may not be able to do.

10 Conclusion

We have shown that children in the age range 8 to 11 years old can make several useful distinctions in the various sets of texts. We found that JAPE-2's jokes were significantly more joke-like than non-jokes, but that they were significantly less joke-like than human-generated jokes. Similarly, JAPE-2's jokes were funnier than those non-joke texts that were judged to be jokes, but less funny than the human-generated jokes. We also found a correlation between the 'readability' of the joke and its judged 'jokiness', and between the age of the judge and their ability to recognise human-generated jokes.

We conclude that JAPE can successfully generate punning riddles. Most of the generated riddles are good examples of the genre, although some are not very funny, and a few fail altogether. Even these failures are interesting, as they indicate some fixable weaknesses in our model and its implementation. Most of the less successful jokes were due to weaknesses in the knowledge bases available to JAPE, resulting in jokes which may be incomprehensible to their intended audience. There is no reason in principle why this weakness could not be remedied.

Overall, JAPE's successful generation of punning riddles is evidence that our model captures the essential features of the genre.

Acknowledgements

The first author was supported by a studentship from the National Science and Engineering Research Council, Canada.

References

Abbott, S. (1993). *Miaow! The Cat Joke Book.* London, Random House Children's Books.

Alec, S. (1987). *Smart Alec's Spooky Jokes for Kids.* Alderley Edge, UK, Beaver Publishing.

Anderson, S. (1987). *A-Z of Animal Jokes.* London, Young Corgi Books.

Attardo, S. (1994). *Linguistic Theories of Humour.* Berlin, Mouton de Gruyter.

Binsted, K. (1996). *Machine humour : an implemented model of puns.* PhD thesis, Department of Artificial Intelligence, University of Edinburgh.

Binsted, K. and Ritchie, G. (1994). An implemented model of punning riddles. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, pages 633–638, Seattle, USA.

Binsted, K. and Ritchie, G. (1996). Speculations on story puns. In *Proceedings of the International Workshop on Computational Humour*, pages 151–159, Enschede, Netherlands.

Binsted, K. and Ritchie, G. (1997). Computational rules for generating punning riddles. *Humor*, 10(1):25–76.

Brandreth, G. (1990). *The Teddy Bear Joke Book.* London, Armada.

Byrne, J. (1995). *Mirthful Kombat - Jokes with BYTE!* London, Random House Children's Books.

Churchill, E. R. (1976). *The Six-Million-Dollar Cucumber.* London, Piccolo Pan Books.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, 33(A):497–505.

Ertner, J. D. (1993). *Super Silly Animal Riddles.* New York, Sterling Publishing Co. Inc.

Forrester, M. (1994). *The Vampire Joke Book.* London, Puffin Books.

Francis, W. N. and Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston, Houghton Mifflin.

Fremont, E. (1993). *Cackles from the Crypt.* London, Random House Children's Books.

Girling, B. (1988). *The Schoolkids' Joke Book.* London, Armada.

Greene, J. and D'Oliveira, M. (1992). *Learning to use statistical tests in psychology.* Milton Keynes, UK, Open University Press.

Hegarty, J. (1992). *The Upside Down Joke Book.* London, Random Century Children's Books.

Jam, E. (1991). *The Raspberry Joke Book.* London, Knight Books/Hodder and Stoughton.

Mark, M. A. and Greer, J. E. (1993). Evaluating methodologies for intelligent tutoring systems. *AI and Education: Special Issue on Evaluation*, 4(2/3):129–153.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., and Tengi, R. (1990). Five papers on WordNet. *International Journal of Lexicography*, 3(4). Revised March 1993.

Phillips, L. (1991). *Wackysaurus Dinosaur Joke Book.* London, Puffin Books.

Rayner, S. (1991). *Ready Teddy Go Joke Book.* London, Puffin Books.

Robinson, T. (1996). The British English example pronounciation dictionary. Online version.

Ruch, W., McGhee, P., and Hehl, F.-J. (1990). Age differences in the enjoyment of incongruity-resolution and nonsense humour during adulthood. *Psychology and Aging*, 5:348–355.

Siegel, S. and Castellan Jr, N. J. (1988). *Nonparametric statistics for the behavioral sciences.* New York, McGraw-Hill Book Company.

Townsend, W. and Antworth, E. (1993). *Handbook of Homophones.* Online version.

Webb, K., editor (1978). *The Crack-a-Joke Book.* London, Puffin.

Wilson, M. (1987). The MRC psycholinguistic database. Technical report, Informatics Division, Science and Engineering Research Council, Rutherford Appleton Laboratory, Oxford.

Young, F. (1993). *Super-Duper Jokes.* USA, Farrar, Straus and Giroux.

Yuill, N. and Easton, K. (1993). The role of linguistic ambiguity in understanding and improving children's text comprehension. CSRP 296, University of Sussex.

A Summaries of schemata

In the list below 'SAD' denotes the artificially constructed description of the punchline phrase. The schemata listed in Figure 3 are:

phonsub: This schema substitutes a word for a confusable segment of another word. The SAD is constructed from the entries for both words. For example:

> (25) What kind of ears do engines have?
> *Engine-ears.* (Girling, 1988)

hopchew: Similar to **coatshed**, this schema negatively compares two confusable verb phrases, constructed with metathesis. For example:

> (26) What's the difference between a hungry kangaroo and a lumberjack?
> *One hops and chews, the other chops and hews.* (Ertner, 1993)

negcomp: This schema negatively compares two confusable compound nominals, constructed with metathesis. For example:

> (27) What's the difference between a pretty glove and a silent cat?
> *One's a cute mitten, the other's a mute kitten.* (Ertner, 1993)

poscomp: This schema constructs a word-sense ambiguous phrase, and positively compares its two senses. For example:

> (28) How's a red–haired loonie like a biscuit?
> *They're both ginger nuts.* (Webb, 1978)

lotus: This schema substitutes a homophone for the first word in a lexicalized compound nominal. The SAD is constructed from the entries for the homophone and the noun phrase. For example:

> (29) What do you call a hairy beast in a river?
> *A weir–wolf.* (Forrester, 1994)

rhyming_lotus: Similar to **lotus**, this schema substitutes a rhyming word for the first word in a lexicalized compound nominal. The SAD is constructed from the entries for the rhyming word and the compound nominal. For example:

> (30) What do you call a police dog?
> *A copper spaniel.* (Young, 1993)

elan: This schema substitutes a homophone for the last word in a lexicalized compound nominal. The SAD is constructed from the entries for the homophone and the compound nominal. For example:

(31) How wide is every cemetery?
A grave yard. (Young, 1993)

bazaar: This schema juxtaposes two homophones in a compound nominal. The SAD is generated from the entries for the homophones. For example:

(32) How does a whale cry?
Blubber blubber. (Young, 1993)

vn: This schema negatively compares two confusable verb-object phrases, constructed by reversing the order of two pairs of homophones, and then compares what one is able to do with what the other is *unable* to do.

(33) What's the difference between an elephant and a flea?
An elephant can have fleas, but a flea can't have elephants. (Young, 1993)

B Allowable vocabulary items

The following is a list of allowable vocabulary items for the texts used in the confirmatory evaluation. WordNet sister and daughter nodes (in the inheritance hierarchy) of these words were also allowable vocabulary items.

adult, age, ale, alien, animal, ant, ape, apricot, aunt, away, back, bad, bag, bail, bake, bale, ball, bank, bare, bargain, base, basement, bass, bath, beach, beak, bear, beast, beat, beech, beet, beloved, better, bill, bird, bitter, blade, blunder, boil, bolt, bond, bottom, bow, boy, brake, bread, break, bright, broil, brush, bump, burn, buy, by, call, car, cast, caste, cede, cellar, cent, cheek, child, clever, close, clothes, clothing, clown, coarse, colour, corn, course, crude, crush, cunning, cure, curiosity, curve, dancing, dark, dear, deduction, deep, deer, depressed, desolate, dirty, dolt, door, doorway, dormitory, dough, draw, dye, earth, education, egg, end, engine, entrance, error, fail, failure, fair, fare, fast, feat, final, fir, fish, flair, flare, flash, fleece, fool, foul, fowl, frail, frank, full, fun, fur, garbage, genuine, gilt, girl, golden, groan, grown, guilt, hail, hall, hare, haul, hobby, hoe, horse, hour, house, human, ice, idea, in, inn, insect, iron, jam, jolt, jump, just, kinswoman, knight, labyrinth, lament, last, lax, leave, leg, lenient, level, light, line, link, lobby, locomotive, lodge, low, mail, maize, male, mammal, manner, manor, maze, melt, menu, mighty, miss, mistake, mite, moan, money, monkey, nerve, nice, night, noise, note, nude, odour, old, one, opinion, out, pain, pane, pause, peach, pelt, penny, period, person, personality, place, plain, plane, play, pleasant, poetry, position, post, potato, pouch, power, pupil, purse, rake, rancid, rarity, real, rear, reasonable, reel, regret, remedy, rhyme, right, rite, ritual, road, root, rotation, route, running, rush, sack, sad, sail, sale, sand, scent, sea, see, seed, sentiment, servant, shelter, shoot, shower, simple, smart, so, sole, son, sore, soul, sound, sour, sow, spank, spare, speck, spot, square, squeak, squeak, stale, stew, story, straight, stranger, student, stupid, style, sunburn, sunshine, sweat, sword, tail, tale, tax, tender, term, terrible, thaw, thought, tie, time, tractor, tree, trick, true, tush, ugly, up, vegetable, vulgar, wagon, wail, wares, waste, water, weak, wear, weather, well, whale, whirl, wolf, wool, world.

C Allowable sentence structures

The following are the allowable sentence structures for all types of text used in the confirmatory evaluation. Some were not used at all. Here, *a* could be replaced by an appropriate determiner.

- What is the difference between a ___ and a ___? You ___ a ___ , but you ___ a ___.
- What is the difference between a ___ and a ___? A ___ ___, wh ile a ___ ___.
- What is the difference between a ___ and a ___? One ___ and __ _, but the other ___ and ___.
- How is a ___ like a ___? They are both ___.
- What is the difference between a ___ and a ___? One is a ___, th e other is a ___.
- What do you get when you cross a ___ and a ___? A ___.
- What do you call a ___? A ___.
- What do you call ___ you can ___? A ___.
- What do you call a ___ that can ___? A ___.
- What kind of ___ can ___? A ___.
- What kind of ___ can you ___? A ___.
- What kind of ___ can ___ a ___? A ___.
- What kind of ___ can you ___ at/in [location]? A ___.
- What do you call a ___ that can ___ a ___? A ___.
- What do you call a ___ that you can ___ at/in [location]? A ___.
- What do you use to ___ a ___? A ___.

D Scores for each text

These are the average 'jokiness', 'funniness' and 'heard before' scores for each text, with their set number and source (H = human, J = JAPE, N = nonsensical, S = sensible), ordered by 'jokiness'. Scores for 'jokiness' range from 0 (none of the children who were asked to rate the text thought it was a joke) to 1 (all of the children who were asked to rate the text thought it was a joke). Scores for 'funniness' range from 1 to 5, with 1 meaning "not funny at all" and 5 meaning "very funny". Scores for 'heard before' range from 0 (none of the children who were asked to rate the text had heard it before) to 1 (all of the children who were asked to rate the text had heard it before).

Jokiness	Funniness	Heard	Source	Set	Text
1	4.33	0	J	4a	20. What's the difference between leaves and a car? One you brush and rake, the other you rush and brake .
1	4.08	0.33	H	4a	16. What do you get when you cross cars and sandwiches? Traffic jam.
1	3.92	0.54	H	2a	8. What kind of vegetable can jump? A spring onion.
1	3.92	0.46	H	2a	6. What do you call a cat with eight legs? An octopus.
1	3.92	0.08	H	2a	10. What do you get when you cross a house with a pancake? A flat.
1	3.83	0.18	H	6a	16. What do you call a bad dream with teeth? A bite-mare.
1	3.77	0.62	H	2a	9. What kind of food do octopuses eat? Fish and ships.
1	3.73	0.17	H	10a	15. What nuts can you use to build a house? Walnuts.
1	3.73	0.08	H	7a	3. What do you call a fun cow? A-moo-sement.
1	3.67	0.25	H	4a	18. What kind of clothing can a spider wear? A coat of arms.
1	3.67	0.18	H	4a	2. What kinds of babies do winds have? Chill-dren.
1	3.67	0.08	H	6a	8. What do you call a boy who eats six bowls of raspberries? Berry greedy.

Jokiness	Funniness	Heard	Source	Set	Text
1	3.5	0.17	H	8a	19. What do you use to talk to a skunk? A smelly-phone.
1	3.36	0.08	J	10a	14. What kind of boy burns? A son-burn.
1	3.31	0.23	J	2a	13. What do you call a beloved mammal? A dear deer.
0.92	3.62	0.31	H	1a	12. What do you call a deer with no eyes? No eye-deer.
0.92	3.92	0.33	H	9a	7. What kind of fruit fixes taps? A plum-ber.
0.92	3.91	0.18	H	3a	14. How is a window like a headache? They are both panes.
0.92	3.75	0.25	H	5a	18. What kind of tent has hair? A wig-wam.
0.92	3.75	0.08	H	6a	20. What do you call a cold aunt? Aunty-freeze.
0.92	3.73	0.18	H	10a	17. What do you use to talk to an elephant? An elly-phone.
0.92	3.64	0.10	J	3a	6. What do you get when you cross a bag and a human? A purse-on.
0.92	3.58	0.17	H	8a	4. What do ghosts eat for pudding? Scream cakes.
0.92	3.58	0.17	H	6a	10. What do you call a lizard on the wall? A rep-tile.
0.92	3.58	0.08	J	5a	20. What's the difference between money and and a bottom? One you spare and bank, the other you bare and spank.
0.92	3.55	0.10	H	10a	20. What kind of Christmas tree does a hedgehog have? A porcu-pine.
0.92	3.5	0.5	H	5a	1. What do you get if you cross a zebra with a kangaroo? A striped jumper.
0.92	3.5	0.36	H	4a	19. How is a car like an elephant? They both have trunks.
0.92	3.33	0.17	H	5a	19. What's the difference between a piece of cotton and a tattered toy? One is a bare thread and the other is threadbare.
0.92	3.33	0	J	4a	5. What do you call true dancing? A real reel.
0.92	3.18	0.33	H	7a	7. What do you call a dead author? A ghost writer.
0.92	3.10	0	J	10a	18. What do you call a bright night? Light time.
0.92	3.08	0.45	H	3a	9. What's the difference between a pony and a sore throat? One is a horse, and the other is hoarse.
0.92	3.08	0.17	H	8a	11. What does a vegetable earn? Celery.
0.92	2.83	0.17	J	6a	11. What kind of beast has a fleece? A wool-f.
0.92	2.82	0	J	3a	19. What's the difference between a straight bill and a nude noise? One's a square beak and the other's a bare squeak.
0.85	3.38	0.33	J	2a	1. What do you get when you cross a monkey a nd a peach? An ape-ricot.

Jokiness	Funniness	Heard	Source	Set	Text
0.85	3.08	0.33	H	1a	6. What kind of animal plays cricket? A bat.
0.85	3.08	0.31	H	1a	2. What do you call a monkey bed? An ape-ricot.
0.83	3.83	0.33	J	9a	13. What kind of pupil has sweat? A stew-dent.
0.83	3.58	0.17	H	8a	9. What do you call a clever skunk? A fast stinker.
0.83	3.58	0.17	H	4a	3. What kind of food do cannibals eat? Human beans.
0.83	3.55	0.17	J	7a	18. What do you call an Earth rotation? A whirl-d.
0.83	3.45	0.10	H	9a	3. What do you call a ghost summer race? A dead heat.
0.83	3.42	0	H	3a	5. How is mathematics like a headache? They are both sum trouble.
0.83	3.36	0.83	J	7a	4. What do you call a tender blade? A sore-d.
0.83	3.33	0.17	J	4a	10. What do you get when you cross a penny and an odour? A cent scent.
0.83	3.27	0.08	J	7a	16. What do you call a pleasant period? The nice age.
0.83	3.25	0.08	J	8a	7. What do you call an adult moan? A grown groan.
0.83	3.18	0.33	H	7a	1. What do plumbers have for Christmas dinner? Plumbed pudding.
0.83	3.17	0.10	J	4a	13. What do you get when you cross a remedy and a rarity? A cure-iosity.
0.83	3.10	0.25	J	10a	5. What kind of hall has a doorway? A door-mitory.
0.83	3.10	0	J	10a	9. What do you call a corn labyrinth? A maize maze.
0.83	2.75	0	J	6a	19. What kind of term has clowns? A fool term.
0.83	2.67	0.18	H	3a	11. What do you get if you cross a car with a vile substance? Crude oil.
0.77	3.38	0.17	H	2a	15. What kind of dog has no tail? A hot dog.
0.77	3	0.15	H	1a	17. What do you get if you cross a flower with a monkey? A chim-pansy.
0.77	2.69	0.15	J	2a	4. What's the difference between a terrible pouch and a desolate rear? One's a bad sack and the other's a sad back.
0.77	2.54	0.08	J	1a	5. What's the difference between a sea and a sale? You can sail a sea, but you can't see a sale.
0.75	3.75	0.10	H	8a	13. What do you call a fight between an apple and an orange? Fruit punch.
0.75	3.58	0.25	H	9a	9. What do you call a bird that lives under ground? A miner bird.
0.75	3.58	0.08	H	5a	10. What kind of being drinks beer? An ale-ien.

Jokiness	Funniness	Heard	Source	Set	Text
0.75	3.5	0	H	4a	12. Wht do you get when you cross ghosts with trees? Cemetries.
0.75	3.45	0	J	10a	11. What's the difference between a seed and a so? You can sow a seed, but you can't cede a so.
0.75	3.42	0	H	6a	13. What do you get when you cross a chicken and a power pack? A battery hen.
0.75	3.36	0.25	H	7a	10. What kind of apple is bad-tempered? A crab-apple.
0.75	3.33	0.10	J	9a	17. What do you call an old bottom? A stale end.
0.75	3.18	0.17	H	10a	16. What kind of ghost rings the door bell? A dead ringer.
0.75	3.17	0.08	J	5a	5. What kind of boy has the post? A mail child.
0.75	3	0.25	H	7a	14. What do you call a line of ghosts? A dead line.
0.75	3	0.08	J	3a	4. What's the difference between a horse and a wagon? One bolts and jumps, the other jolts and bumps.
0.75	2.92	0.25	H	6a	7. What bird is red and steals? A robin.
0.75	2.92	0.08	J	5a	12. What do you call a rancid shower? A sour bath.
0.75	2.75	0	J	9a	14. What kind of iron has a position? Caste iron.
0.69	2.85	0.10	H	1a	8. Why is a gold coin like a criminal? They are both gilty.
0.69	2.62	0.08	J	1a	13. What do you call a depressed engine? A low-comotive.
0.67	3.33	0.18	J	5a	15. What's the difference between a pane and a brake? You can break a pane, but you can't pain a brake.
0.67	3.27	0.36	H	7a	19. What do you call a hot sheep? A woolly sweater.
0.67	3.08	0.33	J	8a	14. What do you use to colour an animal? Hare dye.
0.67	2.92	0.17	H	3a	16. What's the difference between a fish and a fly? A fish can fly but a fly cannot fish.
0.67	2.91	0.08	J	7a	20. What do you call a poetry pause? A rhyme out.
0.67	2.83	0	J	8a	5. What kind of girl has an error? A Miss take.
0.67	2.83	0	J	5a	8. What kind of penny has an opinion? A cent-iment.
0.67	2.42	0.25	J	9a	20. What kind of idea melts? A thaw-t.
0.64	3.10	0.10	J	9a	4. What do you call bare garbage? Nude waste.
0.62	3.08	0.08	J	2a	19. What's the difference between a tractor and a servant? One hoes and bales, the other bows and hails.
0.62	2.77	0.08	J	2a	14. What do you get when you cross a road and a basement? A route cellar.

Jokiness	Funniness	Heard	Source	Set	Text
0.62	2.46	0	J	1a	10. What's the difference between a potato and an egg? One you broil and bake, the other you boil and break.
0.62	2.38	0.15	J	2a	20. What do you get when you cross a bird and a blunder? A fowl up.
0.58	3.36	0.17	H	10a	1. What do you call a sick bird? An ill eagle.
0.58	3	0	J	6a	14. What kind of dark has horses? Knight time.
0.58	2.92	0.10	H	5a	6. What kind of dairy product has muscles? Hard cheese.
0.58	2.82	0	J	10a	2. What kind of leg can shoot? A bow leg.
0.58	2.73	0.10	J	10a	12. What do you get when you cross style and flash? A flair flare.
0.58	2.67	0.27	J	3a	3. What do you call a nude animal? A bare bear.
0.58	2.67	0.17	H	3a	17. What is the difference between a butcher and a fish? One has scales and sees meat and the other has scales and meets seas.
0.58	2.5	0.08	J	6a	18. What do you call a fun spot? A play-ce.
0.58	2.5	0.08	J	5a	4. What's the difference between a just bond and nude failure? One's fair bail and the other's a bare fail.
0.58	2.5	0.08	J	5a	2. What kind of fish has personality? A soul sole.
0.58	2.42	0.08	J	4a	14. How is simple vegetable like a level sou nd? They're both plain beats.
0.58	2.33	0	J	9a	2. What do you call better water? Well water.
0.58	2.25	0	J	3a	18. What kind of bird is dirty? A foul fowl.
0.55	2.64	0.10	H	9a	5. What's the difference between a game and a romance novel? One is exciting, the other is yecch writing.
0.55	2.33	0.17	J	8a	3. What do you call a genuine bill? A frank note.
0.54	3.08	0.15	H	1a	4. What is the difference between an enormous hen and a huge coward? One is a giant chick en and the other is a chicken giant.
0.54	2.54	0.17	J	1a	20. What kind of tree has sand? A beach beech.
0.54	2.08	0	H	2a	11. What's the difference between a picture and a church? One is a centre-piece and the other is a peace centre.
0.5	3.17	0	H	9a	15. How is a boy scout like a tin of raspberries ? They're both prepared.
0.5	2.73	0.08	N	10a	8. What is the difference between a ghost and a spot? You can miss hay, but you cannot help a dormitory.
0.5	2.33	0	J	3a	13. What's the difference between clothes and wa res? You can wear clothes, but you can't close wares.
0.5	2.08	0.17	J	8a	20. What do you call stupid bread? A dough-lt.

Jokiness	Funniness	Heard	Source	Set	Text
0.5	2.08	0	J	3a	10. What kind of manor has style? A manner house.
0.46	2.38	0	N	2a	12. What do you call a gilt explanation?
	¿				A true principal.
0.42	2.83	0.08	H	5a	7. What do you call an Aboriginal in the blood? A foreign body.
0.42	2.58	0.08	H	8a	1. What paper does a hang man read? A noose-paper.
0.42	2.55	0.08	H	10a	3. What ribbon do lawyers use? Red tape.
0.42	2.33	0.18	N	9a	11. How is an ocean like a saucer? They are both sensitive.
0.42	2.33	0	J	4a	1. What kind of curve has cheek? A nerve ball.
0.42	2.27	0.10	J	7a	15. What kind of tree has a pelt? A fur tree.
0.42	2.25	0.17	H	9a	19. What has four legs and one arm? A pit bull.
0.42	2.17	0.08	J	6a	4. What do you call a smart ritual? Rite smart.
0.42	2.17	0.08	J	6a	1. What kind of speck has power? A mighty mite.
0.42	2.08	0.17	J	4a	17. What kind of tush has a story? A tale end.
0.42	2.08	0	J	4a	9. What do you get when you cross a bitter and a stranger? An ale-ien.
0.42	1.91	0.27	S	3a	8. What's the difference between a lemon and an orange? One is yellow and the other is orange.
0.42	1.91	0.08	J	7a	13. What do you call an ugly instrument? A base bass.
0.42	1.91	0	J	7a	9. What do you get when you cross sunshine with a menu? Fare weather.
0.42	1.75	0.08	J	3a	12. What kind of draw has a lobby? An entrance haul.
0.42	1.73	0	J	7a	8. What do you call a mammal's lament? A whale wail.
0.38	2.31	0	N	2a	18. What do you get when you cross a remedy with a mall? A coarse line.
0.36	2.45	0.10	S	9a	6. What kind of entrance can you open? A door.
0.33	2.5	0.08	N	6a	3. What do you get when you cross an amusement and an eagle? A bad bowl.
0.33	2.45	0.10	S	10a	4. How is a robin like an eagle? They are both birds.
0.33	2.42	0.08	S	9a	8. What is the difference between a sheep and a wolf? One eats grass, the other eats meat.
0.33	2.42	0	N	5a	16. What kind of monkey can remedy? A spare peach.
0.33	2.33	0.08	N	9a	1. What kind of enigma can burn? A cold raspberry.

Jokiness	Funniness	Heard	Source	Set	Text
0.33	2.33	0	J	9a	18. What do you call golden regret? Gilt guilt.
0.33	2.18	0.08	S	10a	13. What is the difference between a raspberry and a walnut? One is a berry, the other is a nut.
0.33	2.17	0	J	6a	9. What do you call jam time? Crush hour.
0.33	2	0.08	S	10a	6. How is celery like broccoli? They are both vegetables.
0.33	1.92	0.17	S	5a	9. What is the difference between a scent and an odour? One is a good smell and the other is a bad smell.
0.33	1.67	0	N	3a	20. What kind of dancing has a principle? A cunning personality.
0.31	2.15	0	J	2a	3. What do you get when you cross a link and a lodge? A tie inn.
0.31	2.15	0	J	1a	18. How is an ugly insect like a deep kinswoman ? They're both bass aunts.
0.31	1.92	0	J	1a	7. What do you call a clever feat? Cunning away.
0.31	1.77	0	J	1a	14. What do you get when you cross a bargain and a hobby? A buy line.
0.31	1.62	0	N	1a	11. What do you get when you cross a regret with an adult? A primary utterance.
0.27	2.18	0.10	S	3a	2. What kind of craft has sails? A ship.
0.25	2.36	0	J	10a	19. What do you call a weak tail? A frail end.
0.25	2.25	0.33	S	9a	16. What kind of reptile has legs? A lizard.
0.25	2.18	0.10	N	10a	7. What kind of load can you hate? A baron.
0.25	2.17	0	S	4a	7. How is a remedy like a cure? They both relieve pain.
0.25	2.08	0.08	N	5a	17. What's the difference between a beloved deer and a dim feint? One is a fair bird and the other is dirty gold.
0.25	1.92	0.08	N	4a	11. What kind of feat can squeak? Cunning clothes.
0.25	1.92	0	J	8a	6. What do you call an ugly fish? A base bass.
0.25	1.91	0.08	S	7a	6. What do you call a tender spot? A sore.
0.25	1.83	0.27	S	8a	8. What kind of animal can fly? A bird.
0.25	1.83	0	J	9a	12. What do you call a vulgar education? A coarse course.
0.25	1.82	0.33	S	7a	11. What kind of dish do you use to eat stew ? A bowl.
0.25	1.82	0.08	N	7a	12. What do you call a greedy tape? A fast crab.
0.25	1.75	0.25	S	8a	2. What do you call a smelly animal? A skunk.

Jokiness	Funniness	Heard	Source	Set	Text
0.25	1.75	0.17	N	3a	7. What's the difference between a bass and sand? A penny dims and faints, and an odour costs and regrets.
0.25	1.75	0	N	3a	1. What do you call a grown bargain? A gilt hobby.
0.25	1.5	0.17	S	3a	15. What do you call a green mineral? Jade.
0.25	1.45	0.08	N	7a	5. What do you get when you cross a scout with a clever celery? A man.
0.23	2	0	J	2a	7. What do you call a final trick? A last one.
0.23	1.77	0.17	S	1a	19. What do you call a stick with leaves? A tree.
0.23	1.62	0.15	J	1a	16. What kind of squeak has clothing? A clothes call.
0.23	1.62	0	S	2a	16. How is an ape like a chimpanzee? They are both monkeys.
0.23	1.54	0	N	1a	9. What kind of call can you blunder? An alien fowl.
0.18	1.67	0	N	8a	17. What is the difference between an aunt and teeth? One taps and whirls, the other writes and punches.
0.17	2	0.17	N	5a	11. What kind of sole is low? Bitter clothing.
0.17	2	0.10	S	8a	10. What is the difference between straw and ti n? You can burn straw, but you can't burn tin.
0.17	2	0.08	N	10a	10. What kind of mare has a walnut? A plumber.
0.17	1.83	0.17	S	5a	14. What's the difference between a tree and a tractor? One has leaves and the other has wheels.
0.17	1.83	0.08	N	6a	15. What is the difference between earth and a beast? One is an idea, the other is steadfast.
0.17	1.75	0.08	J	6a	2. What do you call a reasonable menu? A fair fare.
0.17	1.75	0.08	N	6a	17. How is a red thaw like a sword? They are both ribbons.
0.17	1.67	0.10	S	5a	3. What do you call a mouse noise? A squeak.
0.15	1.69	0	N	2a	17. What's the difference between a bear and a scent? An instrument can course but a tree cannot move.
0.15	1.46	0.08	S	1a	3. What's the difference between a bird and a horse? One has wings and the other has legs.
0.10	1.83	0.17	N	8a	16. What kind of berry has a bell? A sick newspaper.
0.10	1.67	0	J	8a	12. What do you call a lenient shelter? A lax deduction.
0.08	1.92	0	N	9a	10. What is the difference between an entrance and antifreeze? One is straw, the other is a sinful mammal.
0.08	1.73	0.08	N	7a	2. What is the difference between a company and an idea? A boy can melt, but a sore cannot twist.
0.08	1.64	0.08	N	4a	6. What kind of engine has a lodge? An apricot purse.

Jokiness	Funniness	Heard	Source	Set	Text
0.08	1.55	1	S	7a	17. What kind of place can you sleep in? A dormitory.
0.08	1.5	0	S	4a	8. What do you get if you cross an occupation an d pleasure? A hobby.
0.08	1.42	0	S	4a	15. What kind of animal can swing through trees ? A monkey.
0.08	1.46	0	N	1a	15. What's the difference between a nude and an animal? You can bare a fish but you cannot scent a base.
0.08	1.38	0.08	S	2a	5. How is a cod like a bass? They are both fish.
0.08	1.31	0.10	S	1a	1. What kind of yellow fruit can you eat? A banana.
0	1.64	0.10	N	4a	4. How is a cent like an education? They are both impartial mammals.
0	1.58	0	N	8a	15. What do you call a fast place? New dough.
0	1.5	0.17	S	6a	6. What kind of man fixes taps? A plumber.
0	1.5	0.08	S	6a	5. What kind of food grows on trees? Fruit.
0	1.42	0.08	S	6a	12. What do you call a person who studies? A student.
0	1.23	0.08	S	2a	2. What do you call a broken nose? Damage.

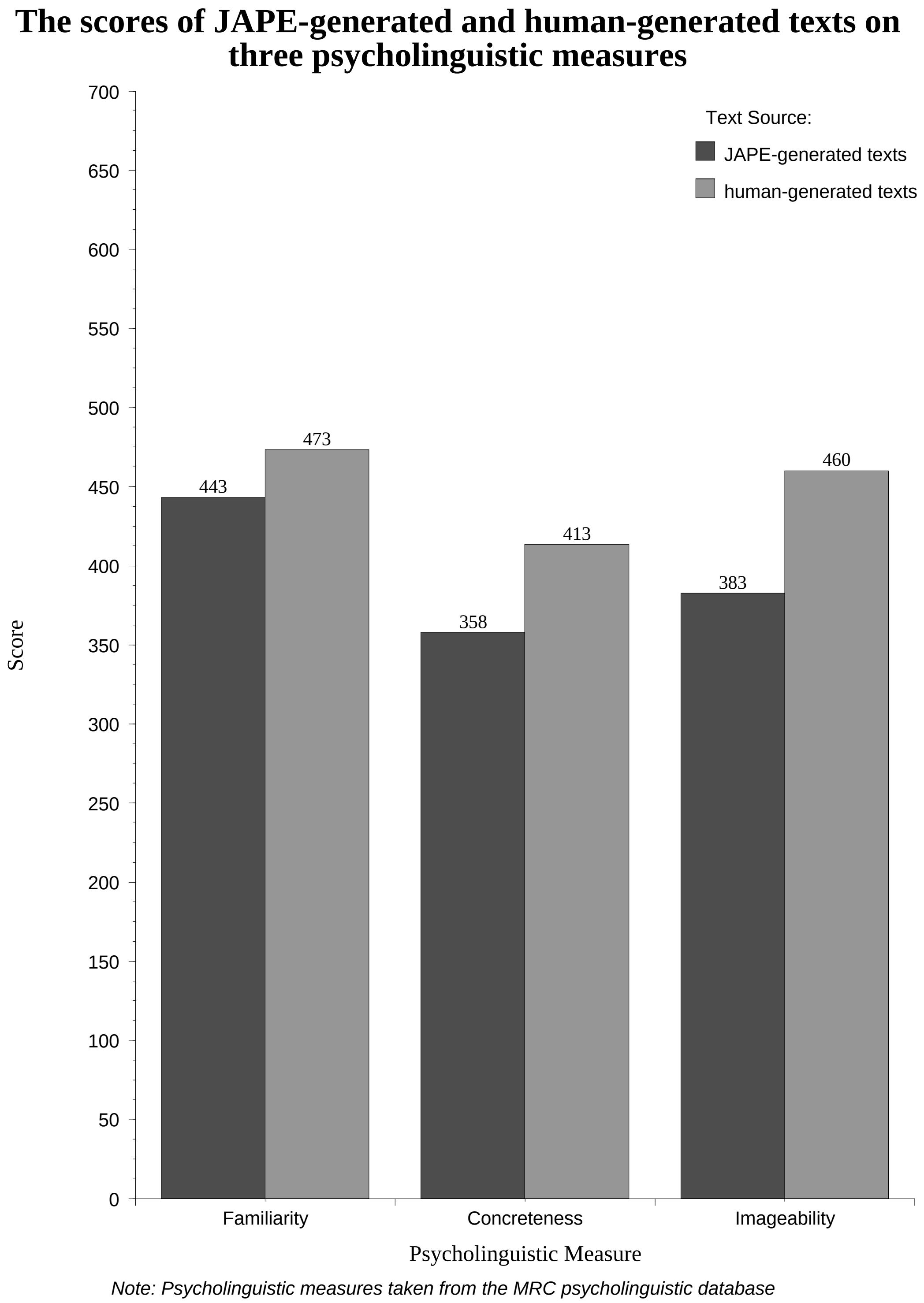

Figure 1: Psycholinguistic data for JAPE-2 texts and human texts compared.

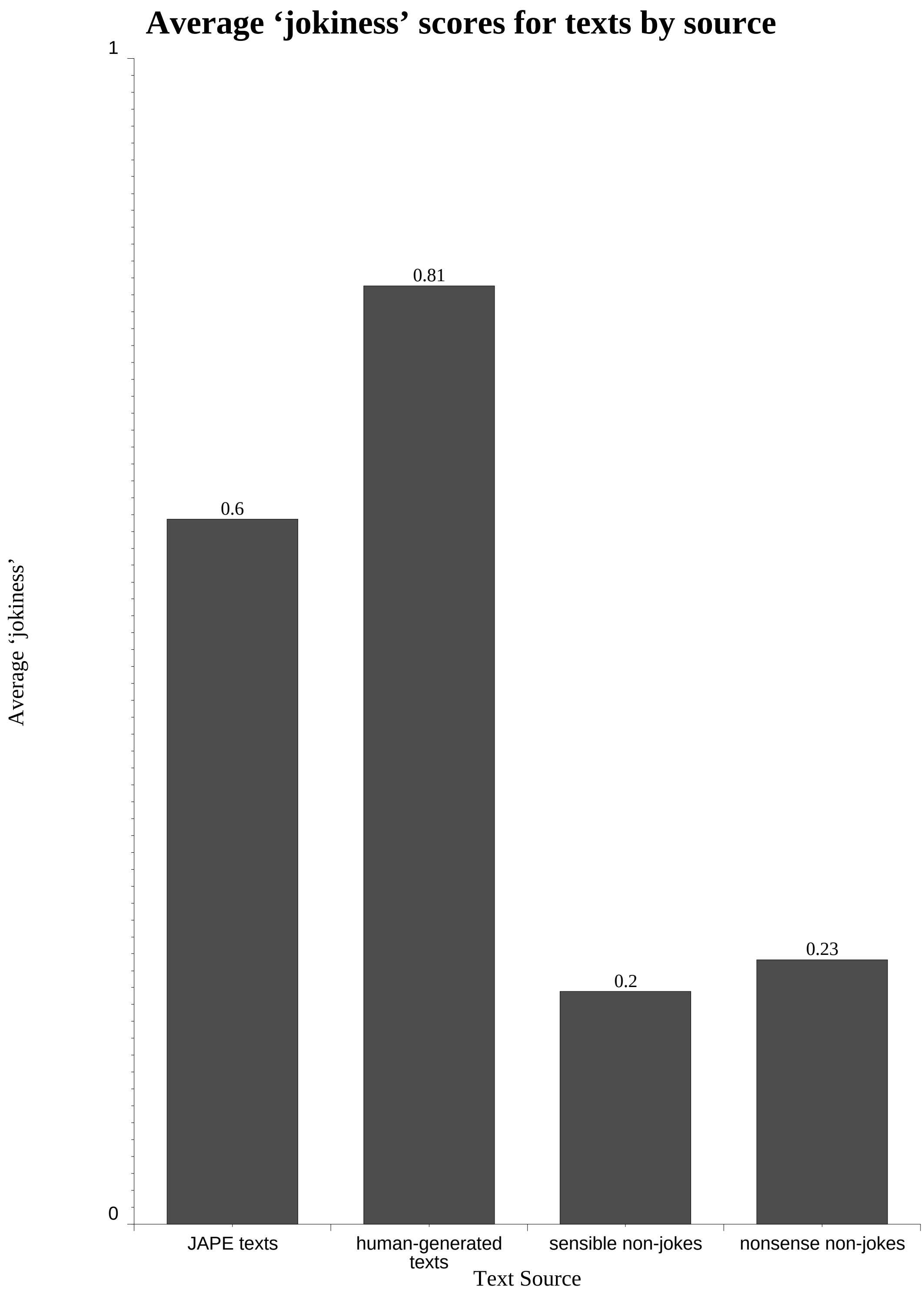

Figure 2: Average 'jokiness' scores for texts from each source.

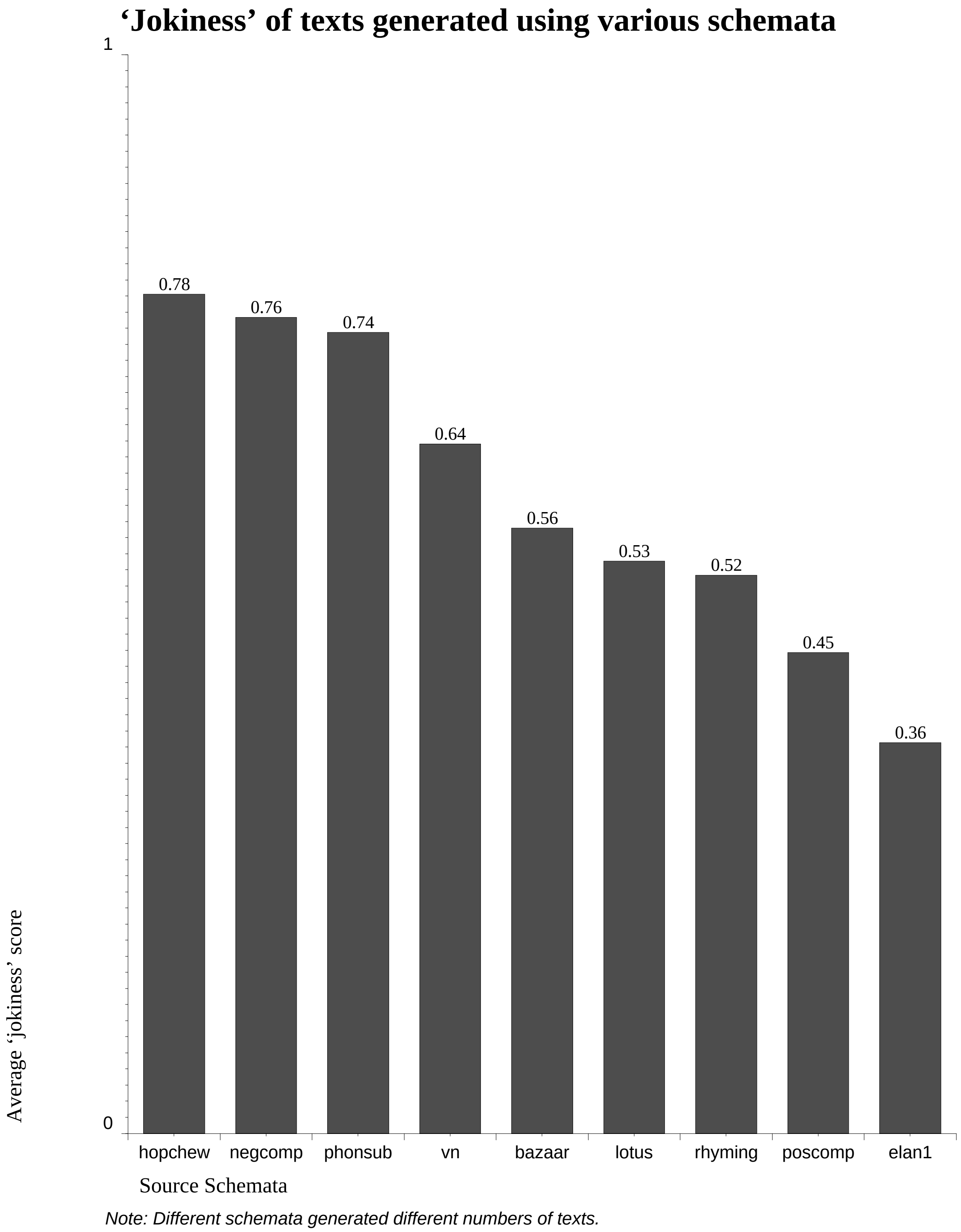

Figure 3: Average ‘jokiness’ scores for texts generated by various schemata.

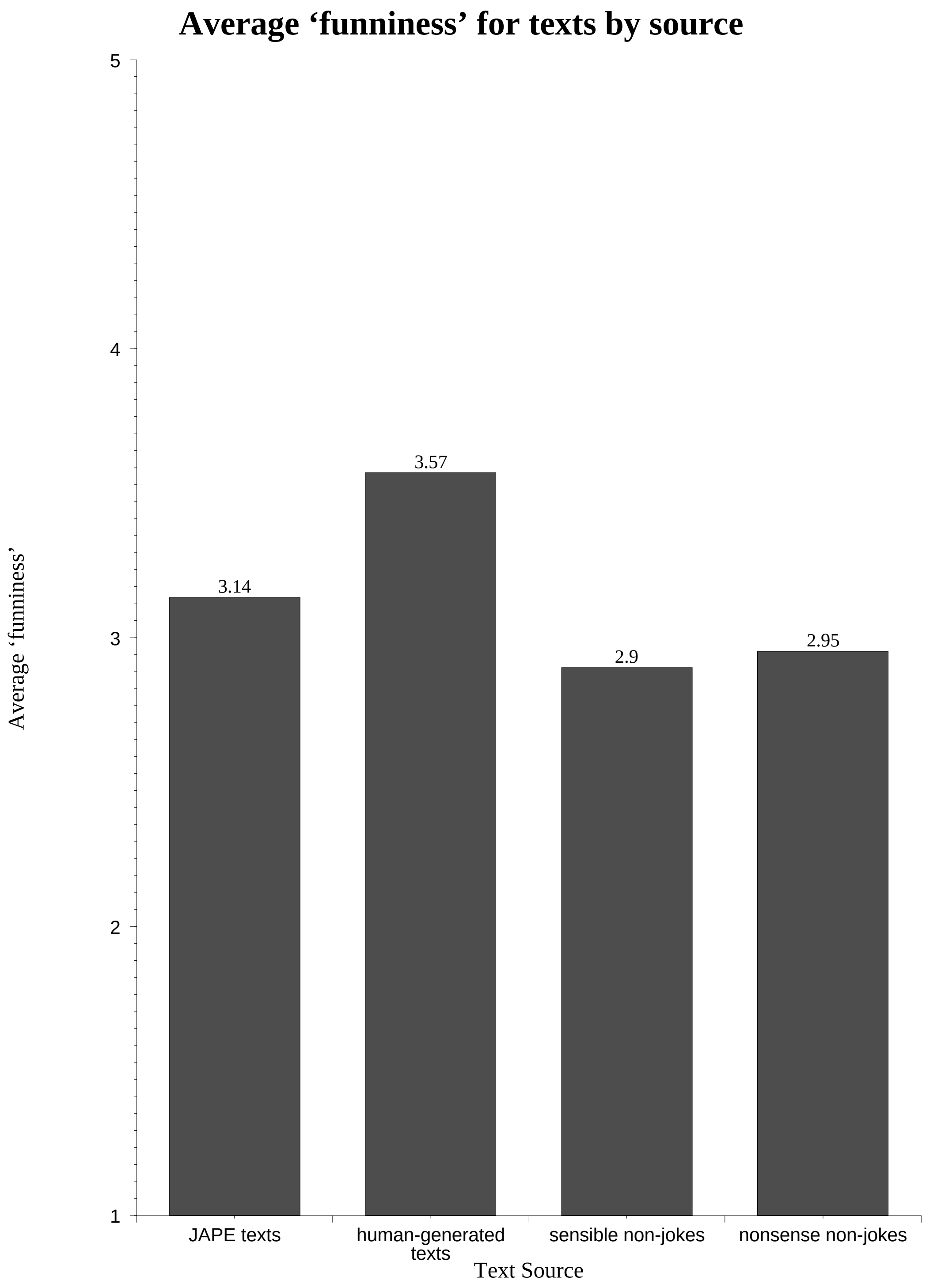

Figure 4: Average 'funniness' scores for texts from each source.

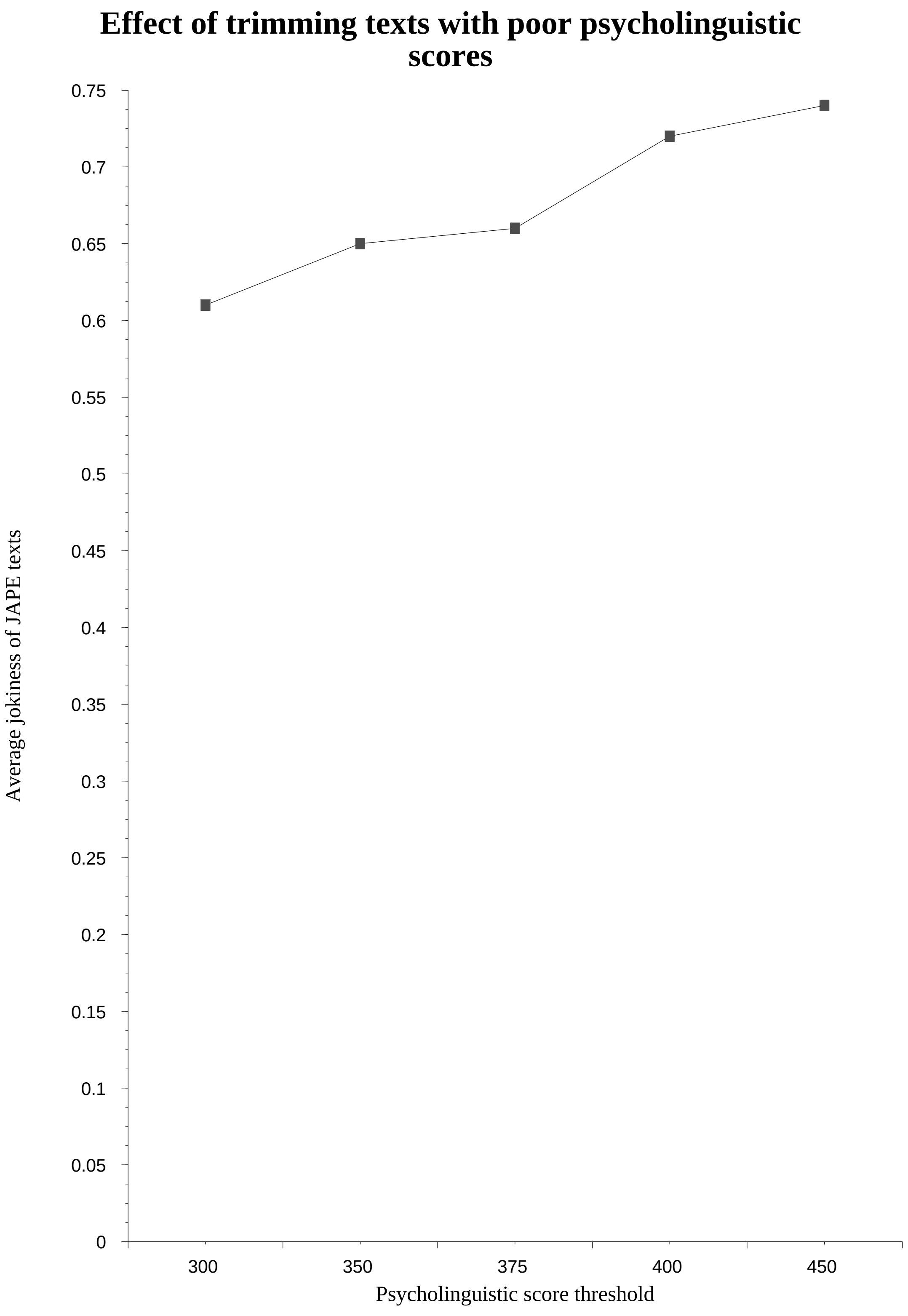

Figure 5: The effect of trimming JAPE output texts with (any) psycholinguistic score beneath a given threshold.