# *Dajare* Generating Support Tool - Towards Applicable Linguistic Humor Processing

Pawel Dybala　　Rafal Rzepka　　Kenji Araki
Graduate School of Science and Technology
Hokkaido University
{paweldybala, kabura, araki}@media.eng.hokudai.ac.jp

**Abstract:** Humor processing is still a heavily neglected field of natural language processing. In this paper we propose a design of *dajare* (Japanese puns) generating support tool. Our system generates phonetic pun candidates and selects plausible phrases, which can be used in generating of pun-including utterances. At the current state of development the system can be used as a *dajare* generating support tool. This research is a part of PUNDA Project, aimed to create a humor-equipped module for conversational systems.
**Keywords:** humor, jokes, puns, pun generators, *dajare*

## 1. Introduction

### 1.1 Humor – the natural feature of language

The phenomenon of humor plays an important role in human language processing. It is the ability to understand and generate humor that makes our utterances more "human-like" and it is hard to imagine a person that would not be equipped with at least some of it. Therefore, it is obvious that nowadays, in the age of daily human-machine relations, the field of humor processing should not be neglected, as it holds the key to make the dialogue more natural. This is why humor generating and recognition shall be treated as an important part of natural language processing research.

### 1.2 Humor as a cure

Another, maybe even more important role of humor in our lives is its positive influence on our mental and physical health. Recent research showed that humor therapy proved successful even in such heavy diseases as cancer [1]. This makes the research on humor processing even more crucial, as humor-equipped machines may perform also medical functions, and maybe even save many people's lives.

### 1.3 Linguistic jokes – language as a material and provider of funniness

Although still no agreement between researchers on the definition of the word "humor" has been reached, we can always use the advantage of commonsensical approach, allowing us to describe humor as, for example, "something that makes things funny". In other words, we recognize humor by the funniness of its products, and these are generally known as "jokes".

The word "joke" is usually associated with written or spoken texts. However, jokes can also take other forms, such as graphical or musical. In our research we focus on the particular subtype of the first group (jokes with textual form), called "linguistic jokes" or "word plays". In this sort of jokes language is not only the material, but, in a specific way, it also becomes the provider (not necessary the only one) of funniness. Therefore, considering quite advanced level of NLP research, humor based on linguistic features seems to be the most computable and possible to process.

### 1.4 Japanese – the language of homophones

Pepicello and Green [2] stated that one of the most important humor features is its ambiguity. In linguistic jokes the ambiguity concerns different aspects of language, with phonetics as the most basic, and probably most important factor. Phonetic ambiguity often takes form of a phenomenon called "homophony", which is very common in Japanese language. Therefore Japanese provides its speakers with a vast possibility of creating puns, and, as a matter of fact, these jokes, called *dajare*, seem to be one of the most popular humor types in Japan [3] For

this reason we decided to use Japanese as a basic language for our research.

## 2. Proposal of the system

### 2.1 Pun generators – state of the art

The idea of creating humor-capable computer system is not itself a very new one. Especially pun generating engines have been the subject of some research project, with few Japanese pun generators projects among them. Although, to the author's knowledge, none of these attempts has proven fully successful, some of them are interesting and worth mentioning, such as JAPE [4] and BOKE [5] by Kim Binsted and Osamu Takizawa. The first generates a very simple sort of puns, called "punning riddles" in English, and the latter is its Japanese conversion. The evaluation experiment proved that both systems did succeed in generating texts recognized as puns, however, their level was significantly lower than of this produced by human.

The system proposed by Yokogawa was generating phonetic pun candidates with articulation similarities as basic patterns. Some of the results were evaluated as potential pun candidates, but, as the meaning was not considered, the system could not be treated as an actually working pun generator [6].

The system proposed by Tanizawa was in fact a humor-equipped Q-A system, which provided a more natural (simple dialogue) environment for the puns. However, as the evaluation experiment showed, the quality of humorous answers was relatively poor, with AH (Average of Humor) rate at the level of 10% [7]

### 2.2 PUNDA Project - outline

The research described in this paper is a part of PUNDA (PUN-*DAjare* system) – a joint research project aimed to create a fully integrated Japanese pun generating module, and implement it into a non-task oriented conversational system. Although the goal is the same as of projects described above, we would like to propose some different concepts to achieve it, such as highly-detailed "*dajare* types set", individualized sense of humor or Web-based lexical corpus (our new approach is described in details in [8] and [9]).

The development of PUNDA Project includes following steps:

**Step 1**: *Dajare* Types Set Extraction – based on phonetic classification of *dajare*, proposed in our previous research. This step is described in **section 3.1**

**Step 2**: PUNDA Module construction – using phonetic patterns extracted in **step 1**, pun generating module will be constructed. This step is currently under development, and its current results are described in **sections 3.2, 3.3 and 3.4**.

**Step 3**: Preliminary research – evaluation of human-made *dajare*, which will allow us to create some models of "linguistic sense of humor". The survey on the subject is now being conducted.

**Step 4**: Integration with ML-Ask system – the problem of joke timing will be solved by implementing Ptaszynski's Emotive Analysis System into PUNDA Module. [8, 9]

**Step 5**: Joking conversational system – the functional PUNDA Module will be implemented into a non-task oriented conversational agent. As the result, it will be able to generate jokes and insert them smoothly into the conversation with the user.

### 2.3 *Dajare* generating supporting tool

The system described in the next section has been developed as a part of realization of **Step 2**, which itself is based on **Step 1** of PUNDA Project. At its current shape, it can be used as a *dajare* generating supporting tool, providing the user with pun candidates for given word or phrase. It has been developed from the previous phonetic pun candidate extraction algorithm, described in [9].

## 3. System outline

### 3.1 *Dajare* phonetic classification and *Dajare* Types Set Extraction

The first phase of the system development, *dajare* types set extraction, is based on *dajare* phonetic classification, proposed in our previous research. We gathered the examples of human-generated *dajare* and divided them into 12 groups, most of which are internally divided into subgroups and sub-subgroups [3].

At this point of research we focus on groups 1-3, with their subgroups, which provides us with the amount of 6 *dajare* generation patterns:

i. **Homophony** (カエルが帰 *Kaeru ga kaeru* <The frog comes back>)

ii. **Mora addition**

ii.a **Initial mora addition** (スイカは安いか *Suika wa yasuika* <Watermelon is cheap>)

ii.b **Final mora addition** (カバのかばん *Kaba no*

*kaban* <Sea cow's bag>)
ii.c **Internal mora addition** (布団が吹っ飛んだ *Futon ga futtonda* <Futon flew away>)
iii. **Mora omission**
iii.a **Final mora omission** (スキーが好き *Sukii ga suki* <I like skiing>)
iii.b **Internal mora omission** (ステーキはすてき *Suteeki wa suteki* <Steaks are cool> [3, 8]

### 3.2 Phonetic candidate extraction

In the first step, our system proposes all possible phonetic changes using patterns described above. For example, for the word *katana* (a Japanese sword) it will be *katana* (homophone), **katana* (*akatana, ikatana, ukatana*... - initial mora addition), *katana** (*katanaa, katanai, katanau*... - final mora addition), *ka*tana, kata*na* (*kaatana, kaitana, kautana*... - internal mora addition), *kata* (final mora omission) and *Kana* (internal mora omission), with all possible Japanese sounds of one mora in place of "*". This gives the list .[1] In the next step, the plausibility of acquired phonetic candidates is checked in the Internet, with 10 000 Yahoo hits rate as the minimal point. All words with hits rate higher than this borderline are extracted and form the list of possible phonetic pun candidates.

### 3.3 Evaluation experiment

At this point of system development we conducted an evaluation experiment to make sure that we are proceeding in the right direction. From our human-generated *dajare* corpus we chose examples matching phonetic techniques covered currently by the system (6 patterns mentioned above, with the amount of 1 mora as change restriction). From every pun the base word was selected and used as an input for the system. Then it was checked if words actually used in human-generated jokes appeared among the candidates proposed by the system. We decided not to include homophones in the evaluation test, because of the obvious 100% correctness of the phonetic candidates for this group – however, it still is included in the algorithm. From other 5 patterns, we used 36 base words as an input to the system.

As the result, 78% of words used in human-generated puns appeared among the candidates proposed by the system. Other candidates were also proposed. This gave the system an acceptable and promising accuracy for continuing the research.

### 3.4 Selection of plausible candidates

At that point of development, one of the major problem with the system was the amount of extracted candidates. Of course, it depended on the word, its length and commonness, however some input phrases were given above 100 candidates, many of which included base words with added particles (for example, "*katana ga*" <Japanese sword> plus subject indicating particle <ga> for "*katana*"). As linguistic joke obviously can not have the same meaning element as the word it came from, it was necessary to throw out such phrases and select candidates that would be really plausible for generating puns.

To do this, for each phonetic candidate we extracted a proper Yahoo snippet, and analyzed it using POS and morphological analyzer MeCab[10]. If any of the lines was recognized as the base word, the candidate was deleted. Than, if any of the lines was recognized as the candidate itself, it meant that the word exists, it has a different meaning than the base word, and as such it can be used to generate a pun (we call these candidates "exact match candidates"). Finally, for all candidates that left after these two selection, nine other snippets were extracted, and again analyzed with MeCab. If any combination of joint lines included the candidate, the coocurrence of these words were checked in Yahoo, and, if there were more than one possibility of analysis, the combination with higher coocurrence was chosen ("combined match candidates").

As a final result, the system provides the user with candidates for the input phrase, including the exact match phrases and combination match phrases, with the proposition of how they can be combined.

For example, for the base word テント (*tento* <tent>), the system proposed following exact match candidates: **base word**: テント(*tento* <tent>)
**exact match candidate:** てんとう (*tentoo* <to collapse>), ステント (*sutento* <stent>), パテント (*patento* <patent>), テントリ (*tentori* <competition for score>)
**combined match candidates**: してんと <視点+と> (*sitento*, divided into *siten* <point of view> and particle *to* <with> or quotation), んてんと <寒天 + と> (*ntento*, divided into *kanten* <freezing weather>

[1] However, further research showed that for the group "internal mora addition" the list of inserted sounds may be limited to sound prolongations and voiceless sounds, as only such examples were found in our human-created *dajare* corpus.

and particle *to*),てんとん <店 + とん> (*tenton*, divided into *ten* <shop> and *ton*)

Of course, not all words do have exact match candidates. For the base word スイカ (*suika* <watermelon>), system proposed combined candidates only:

**base word**: スイカ (*suika* <watermelon>)

**combined match candidates**: すいから <すい + から> (*suikara*, divided into *sui* (may mean <water>) and particle *kara* <from>), うすいか <うすい + か> (*usuika*, divided into *usui* <thin> and question particle *ka*), くすいか <くすい + かん> (*kusuika*, divided into *kusui* <name> and question particle *ka*), すいかい <さんすい + かい> (*suikai*, divided into *sansui* <landscape> and question particle *kai*), んすいか <炭水化物> (*nsuika,* included in the word *tansuikabutsu* <carbohydrate>), ますいか <麻酔 + 科> (*masuika*, divided into *masui* < anaesthesia> and *ka* < department>), スイカン <送水 + 管> (*suikan*, divided into *sousui* < supply of water> and *kan* <pipe>),キスイカ <キス + イカ> (*kisuika*, divided into *kisu* <kiss> and *ika* <squid>)

## 4. Conclusion

In this paper we presented our idea of PUNDA – Japanese Pun Generating System. We briefly described its outline, originality and current state of development. Evaluation experiments of phonetic candidate extraction module and emotive analysis module proved both of them to be successful and can be used in further research. We also introduced an algorithm for selecting plausible pun candidates. At its current development state, the system can be used as a *dajare* generating support tool, providing the user with plausible pun candidates for the given input word.

## 5. Future work

We are going to continue our research according to the development plan explained in **section 2.2**. The next step of pun candidates extraction module will begin with covering other phonetic patterns, such as mora substitution or metathesis. Also, one of the tasks we will have to deal with in the near future is the conversion of candidates from *Kana* characters into *Kanji*. Then we will create a association algorithm for binding the candidates with the base words. We are also planning to combine our system with Ptaszynski's ML-Ask Emotive Analysis System (currently under development) [9], which will allow us to solve the *dajare* timing problem (basing on utterances emotive analysis).

This will be a solid basis for the pun including sentence generation module, which will be implemented into a non-task oriented conversational system.

## References:


[1] Kazue Takayanagi "The Laughter Therapy" Japanese Journal of Complementary and Alternative Medicine, Vol. 4, 2007, No. 2 pp.51-57

[2] William Pepicello and Thomas A. Green "The Language of Riddles", Ohio State University, 1984

[3] Pawel Dybala, "*Dajare - Nihongo ni okeru dōon'igi ni motozuku gengo yūgi*" (*Dajare* – Japanese puns based on homophony), Jagiellonian University, Kraków, 2006

[4] Kim Binsted "Machine humour: An implemented model of puns", University of Edinburgh, 1996

[5] Kim Binsted, Osamu Takizawa "Computer generation of puns in Japanese", Sony Computer Science Lab, Communications Research Laboratory, September 1997

[6] Toshihiko Yokogawa "Generation of Japanese puns based on similarity of articulation", in Proceedings of IFSA/NAFIPS 2001, Vancouver, Canada.

[7] Toshifumi Tanizawa, „*Shitsumon-outougijutsu wo mochiita yuumoa outou*" (Humorous Answers in Question Answering Systems), HIroshima City University, 2007

[8] Pawel Dybala, Rafal Rzepka, Kenji Araki, "PUNDA Project – a Design For A Japanese Puns Generating system" (in: Language Acquisition and Understanding (LAU) Technical Report, Hokkaido University, Sapporo, 2007, pp. 6-11)

[9] Michal Ptaszynski, Pawel Dybala, Wen Han Shi, Rafal Rzepka, Kenji Araki, "Lexical Analysis of Emotiveness in Utterances for Automatic Joke Generation", ITE Technical Report, Vol. 31, No.47, pp.39-42 ME2007-204, 2007

[10] MeCab: Yet another part-of-speech and morphological analyzer, T.Kudo, http://mecab.sourceforge.net/