5E-03

日本語単語の数値化による駄洒落現象（偶然の一致）拡張の試み

金久保　正明†

静岡理工科大学情報学部†

１．はじめに

近年、いわゆる「ことば工学」の一環として、ユーモラスな駄洒落や、駄洒落に基づくなぞなぞなどの言葉遊びをコンピュータに自動生成させる研究が行われて来た[(1)～(3)]。駄洒落は、発音が一致し、意味が異なる二つの単語や言い回しの検出により発生し、偶然の一致（符合）の検出の一種と言える。一方、数字には「ひ・ふ・み」などの読みがあり、これを用いて一部の単語は数値化する事が出来る。或る単語の数字を二倍したり、二つの単語の数値を足すと別の単語になったりという偶然の一致も駄洒落と同様な面白さを生む可能性があるかもしれない。国語辞典を用いた検出結果と、様々な偶然の一致（符合）を検出する試みの今後の展開について述べる。

２．日本語単語の数字化

数字とその読みの対応としては、表 1 のようなものを用意した。数字の読みは「いち、に、さん・・」の系統と、「ひ・ふ・み」の系統がある。これらは混ぜても良い事とした。また、0 は「おー」や「わ」などのアルファベットや形状から派生する読みも可能であり、これも取り入れた。この対応からは、同じ読み（平仮名）から、複数の数字が該当する事はない。

小学生用国語辞典[(4)]に収録されている 31,461 単語に対し、同じ読みの重複を排除した 26,680 個の読みを数字化の対象にした。「う」「ん」「っ」「ー（音引き）」は、一文字のみの省略ならば、読み全体の印象に大差ないと考えられる事から、「う」が先頭であるものを除き、省略した残りの読みを数字化した。その代わりに、これらの文字が読みの中に二つ以上登場する単語は除外する事とした。

また、二文字で数字になる場合はそれを優先し、三文字目から新たな数字化が可能かを検討した。さらに、数字の先頭が 0 になるものは除外する事とした。

表１　読みと数字の対応

0	1	2
れい・れ・わ・お	いち・ひ	にい・に・ふ
3	4	5
さん・さ・み	し・よ・よん	ご・いつ
6	7	8
ろく・ろ・む	しち・なな・な	はち・は・ば・ぱ・や
9	10	
く・きゅ・こ	じゅ・とう・と	

その結果、完全に数字に変換出来た読みは、920 語で、読みの総数に対する比率は、約 2.45%だった。また数字の桁数別には、表 2 のような結果になった。

表２　数字化後の桁数別読みの数

桁数	1桁	2桁	3桁	4桁	5桁以上
読み数	54	346	317	147	56

桁数では 8 桁の事例が 1 例で最長だった。7 桁以上の変換事例を表 3 に示す。

表３ 7桁以上の変換事例

いつとはなしに	5108742	としおとこ	1040109
ことごとく	9105109	なにくれとなく	72901079
ことごとに	9105102	はなしことば	8749108
こみみにはさむ	9332836	ひとしごと	1104510
しろうとばなれ	4610870		

３．偶然の一致（符合）を求める計算結果

ここでは、あまり複雑な計算を行うと、無理にこじつけて偶然の一致を求めた形となり、意外性が損なわれるとの予測から、簡単なものを選ぶ事とした。具体的には、或る読みの持つ数字を二倍して別の読みの数字になるかを検討した。もう一つは、或る読みと別の読みの数字を足したら、第三の読みの数字になるかを検証する事とした。他にも引き算なども考えられるが、今回はこの二つにした。

その結果、或る読みの数字を二倍にして別の読みの数字になったのは　1,255 例だった。元の

An attempt to extension the pun(accidental coincidence) based on quantifying the Japanese words.
†Masaaki Kanakubo
Shizuoka Institute of Science and Technology
Faculty of Informatics

読みの数より多くなった理由は、二倍して出来た数字に複数の読みが対応する事が多かったためである。

また、或る読みと別の読みの数字を足して、第三の読みの数字になったのは、26 万 9,300 例に上った。桁数の大きい数字同士の和の方が意外性は高いと考えたため、双方 500 以上で成立した場合に絞ると、566 例になった（双方 1,000 以上では 100 例程度しか該当しなかった）。

以下に具体例を示す（人手で抜粋した）。

（1） 或る単語を 2 倍して他の単語になる例

・イチゴ(15)二つに未練(30)がある
・二つの文(23)を読む(46)
・二つの逸話(50)は永遠（100）に伝わる
・その広場(168)には、二本のヤシ(84)がある
・号令(50)を二回掛けると答礼(100)が返る
・二回困惑(909)して、ひやひや(1818)した
・魅惑(309)、魅惑(309)、無批判(618)
・シート(410)二枚にくるまれた埴輪(820)
・黒板(998)に二つの予告(499)が書かれた
・不統一(2105)な点が二つあるシフト(4210)

（2） 二つの単語を足して第三の単語になる例

・いち早く(1889)コミック(939)読んでニヤニヤ(2828)
・やにわに(8202)クレーム(906)の付いた言葉(9108)
・得用(1094)という広告(999)の触れ込み(2093)
・婚約(989)でパニック(829)冷や冷や(1818)
・三国志(3594)は複合語(2955)も多い労作(639)
・素人(4610)を蝕む(6486)一苦労(11096)
・母親(8808)がやにわに(8202)火縄銃(17010)
・ハンサム(836)と触れ込み(2093)の男がニコニコ(2929)

４．自動生成への展開と課題

多数生成された単語の組合せから、自動的にフィルタリングするには、まず意味が成立するものに絞る必要がある。語彙体系のデータベースを用いて、動詞カテゴリーとその目的カテゴリーを用意すれば、例えば「二つの目的語を動詞する」といった文章テンプレートに当てはめる事が出来る。場所のカテゴリーとそこに在り得るもののカテゴリーを定義すれば、「場所には二つの（在り得る）物がある」といったテンプレートに当てはめる事が出来る。また、上記の例にあるように、関連するカテゴリーの単語であれば助詞を用いずに「○○、○○、●●」（黒丸が計算結果の方）と並べる方法もある。

意味の整合性があるものに絞った後は、これまでの駄洒落研究で行われた様々な評価の方法が援用出来るだろう。例えば、ユーモラスな単語群を予め定義して該当度を見るとか、単語親密度のデータを用いる事も考えられる。

偶然の一致は意外性を生み、意外性はこれまでの多くの研究でも絞り込みの重要な指標とされている。意外性を数値化して評価に持ち込む事が特に重要と考えらえる。意外性は主観的なものなので、プログラムによる定量的な評価は難しい。客観的な指標の一つとしては、その偶然の一致の発生確率が挙げられ、低いほど意外性は高いと予想できる。例えば駄洒落であれば、一致する文字数が多いほど発生確率は低いと予想出来る。

５．他の偶然の一致（符合）の検出課題

しかし、大量データから意外性のある偶然の一致を検出する方法としては、駄洒落以外にまだ多くの形態がある。例えば、或る山の高さとその近くを走る国道の数字が一致しているといった、数字データの一致である。他にも、「品川駅が港区にあり、目黒駅が品川区にある」といった矛盾する関係が複数並んでいるといった現象、「大阪府に福島という地名があり、福島県に大坂という地名がある」といった相互関係など、様々な事例を挙げる事が出来るが、これらは意外性のあるトリビアとして、しばしば話題に上るため、自動生成も無意味とは言えない。

一致の起こる確率の低いものほど、トリビアとしての価値が高まる事は実に容易に予想されるが、それぞれの一致の在り方に於いて、どのように確率を計算するかは今後の課題である。

６．おわりに

単語の読みを数値化して、計算により駄洒落と同様の偶然の一致を得る事を検討、実験で或る程度の事例を得た。今後の課題としては、予め定義した単語カテゴリーと対応する文章テンプレートを用いた自動抽出、さらに面白いものを絞り込む工夫、特に発生確率に基づく意外性の検出、他の様々な形態の偶然の一致への拡張などが挙げられる。

参考文献

(1) 松澤和光, 堀浩一, 金杉友子, 阿部明典: ことば工学入門, 人工知能学会誌(2000)
(2) 滝澤修, 柳田益造, 伊藤昭, 井佐原均: 日本語修辞表現の工学的解析－駄洒落・アイロニー・トートロジー, 信学技報(1997)
(3) ヨーナス・シューベルグ, 荒木健治: 日本語を対象とした謎掛けの自動生成, 情報処理学会研究報告(2007)
(4) 小学新国語辞典, 光村教育図書, 2002