自由回答アンケートにおける単語の表記揺れとその解消

笠原 要 [†]　　永森 千晴 [††]　　加藤 恒昭 [†††]

[†] NTT コミュニケーション科学基礎研究所　　[††] NTT アドバンステクノロジ

[†††] 東京大学大学院総合文化研究科

1　はじめに

被験者が任意の回答を記述する自由回答アンケート調査において問題となる単語の表記揺れに関して、実際に行われたアンケートの集計作業に基づいて、その傾向を報告する。

筆者らは、人間が行う単語の類似性判別を工学的に再現する概念ベース [笠原 97] の研究を行っている。その評価基準を作成するため、刺激語に対する関連語 (類義語、対義語、連想語) を収集するアンケート調査を行なった。調査では自由回答、すなわち回答用紙に自由に単語を記入する方法を採用した。結果の集計においては、表記が全く同じ回答語をまとめて、その回答者数を単語の出現頻度とすることを考えていたが、得られた結果 (例えば、表 1) を見ると次のことが明らかとなった。つまり、単語が持つ意味やそれと関連する概念を研究するという立場に立った場合、意味論的、語用論的にほぼ同じとみなして構わないと判断される単語 (例えば、表 1 中の「月の輪熊」「月の輪グマ」「ツキノワグマ」) が異なる複数の表記を持っており、それらを別の単語と扱うよりは、同じ単語の表記揺れであるとしてまとめ上げて出現頻度を議論した方がより適切であると感じられたのである。

表 1: 刺激語「熊」に対する類義語の自由回答アンケート結果 (一部)

パンダ (18), 白熊 (12), 白態 (1), 白クマ (1), しろくま (1), ベアー (10), ベア (9), ベア (bear) (1), 月の輪熊 (1), 月の輪グマ (1), ツキノワグマ (1), テディベア (1), テディーベア (1), テディ・ベア (1), 動物 (3), ヒグマ (3)　　(括弧内は、回答した被験者数)

ここで必要となるのは、回答語のどれとどれが表記揺れの関係にあり、どのようなものが異なる単語であるかの判断である。これは、自由回答アンケートで非定型の回答を定量的に集計する場合に常に必要となるもので、内容が類似した回答を１つのカテゴリーにまとめるカテゴリー化の作業の一種であると考えられるが、作業の時間的なコストを要するために、予め選択肢を用意するアンケート調査への変更が勧められている程困難な作業である [辻 87]。また、このカテゴリー化の基準は、調査の目的や内容に応じて変化するため、実際の自由回答アンケート調査の集計作業では、調査ごとにその都度議論等を通じて ad hoc にカテゴリー化の方針を作成せざるをえないことも推測される。

これは、また、言語学研究における単語の表記揺れの問題に関連する。文献 [国研 83] では、複数の国語辞典等や新聞、雑誌に現れる単語の語形や表記の揺れを調査し、その揺れがどのような単語表記間の対立から発生しているかについての分類を試みている。その内容は、今回実施した自由回答アンケートの集計においてもしかし、国語辞典は、見出し語の表記揺れの全てを収録しているわけではないし、自由回答アンケートでは、国語辞典に収録されていない単語が現れることも多いため、そこにまとめられた内容だけで筆者らの問題が完全に解決するわけではない。

単語の収集という作業を自由回答以外の形式で行うことは困難であろうし、自由回答形式で単語の収集を行った場合、多かれ少なかれ筆者らが直面したのと同様の表記揺れの問題に出会うことになると思われる。このような背景から、筆者らの調査の集計過程で現れた表記揺れの問題がどのようなものであったかを報告すると共に、それを筆者らがどのように解決したかについて報告し考察する。前述の通り、自由回答アンケートにおけるカテゴリー化の方針、この場合は表記揺れであるか別単語であるかの判定基準は、調査の目的と内容に応じて変化せざるをえな

[†] 笠原 要 NTT コミュニケーション科学基礎研究所 619-0237 京都府相楽郡精華町光台 2-4 e-mail: kaname@cslab.kecl.ntt.co.jp

いが、単語の収集という作業は様々な場面に共通するものであるから何らかの示唆はあると思われるし、そうでなくても、１つのテストケースとして参考にしてもらえば幸いである。

以下の議論を明確化するために、本稿で取り扱う単語の表記揺れを以下のように定義する。

> **特定の質問に対して単語で回答する自由回答アンケート調査を集計する際に、調査の目的と質問の内容を鑑みた場合、語形や表記が異っていても同一の単語を指し示した回答であると判断できる一連の単語表記を“表記揺れ”と呼ぶ。**

この定義は、前述の文献[国研 83]における語表記の揺れ、語形の揺れ、文字の揺れ等を含み、更に誤用の問題も含んだ広いものとなっている。これらを“表記揺れ”と一括することには若干の問題もあるかと思うが、例えば、片仮名語において、語表記と語形の区別が困難という指摘もあり、広い意味で“表記揺れ”という語を用いることとする。

2　関連語の自由アンケート調査

2.1　調査内容

既に述べた通り、実施した調査の目的は、刺激語に対する関連語を収集することである。関連語としては、「ある単語と似たような意味を持った単語」（類義語）、「ある単語と反対の意味や対となるような意味を持った単語」（対義語）、「ある単語から連想される単語」（連想語）を個別に調査した。

刺激語としては、誰でも知っているような基本的な単語200 語を文献[天野 00]に基づいて選択した。これは、6 万語に1 から7 までの値を取る単語親密度が付与されたデータベースであり、親密度の値が高いほどその単語はなじみがあると感じられる。ここでは親密度5 以上の200 語をランダムに選出し、刺激語とした。刺激語の一部を下記に挙げる。

> 踊る, 洋食, 馬, ハムスター, 春夏秋冬, タイガー, 検定, レベル, バーボン, ダイナミック, 引っ越し, ブラウス, 祭り, ボディーガード, 食卓, アルプス, 口座, 無神経, 着く, きつね, 爆笑, 関係, シーツ

アンケート調査は1 つの関連性について100 名の大学生に対して行い、200 の刺激語に対して、指定された種類の関連語を調査用紙に記入させる方法で行った。回答の目安として、1 語の刺激語に対して、10 秒以内に少なくとも3 語の回答語を記入することを指示した。

なお、この調査の詳細とそこで得られた知見については、稿を改めて報告する予定である。

2.2　調査結果

調査用紙を回収し、各々について被験者が記入した関連語をコンピュータに投入した。その際に、誤字と思われるものもそのまま投入した。例えば表１にみられる刺激語「熊」に対する回答語「白態」のように、多分誤字であり、文字の類似性より被験者が意図したであろう回答（この場合は「白熊」）を推定できる場合もあったが、投入作業者の主観や推測をできるだけ排除することを意図し、修正はしないであるがままの文字列をデータとして投入した。ただし、正しい文字として存在しないような誤字については、投入者が正しい文字を推定できる場合には修正を行った。

3 種類の関連語のアンケートの回答について、同じ刺激語に対して同じ表記の単語を何名の被験者が記入したかを表わすデータが、本稿の報告のもととなるものである。100 名の被験者による回答語数は、1 つの刺激語あたり延べ語数の平均で282 語(類義語)、220 語(対義語)、644 語(連想語)、異なり語数の平均で90 語(類義)、72 語(対義)、278 語(連想) となった。

3　回答語の表記揺れとその解消

上記の自由回答アンケート調査の回答データに対して、どのような単語の表記揺れが現れたかを報告する。そして表記揺れの解消を行う際に定めた基本方針について説明し、それに基づいて作成した44 の規則と、その効果に

ついて述べる。

3.1 問題事例とその分類

ここでは、表記揺れ解消の検討を行った際に、表記揺れか否かを判断しなければならないと考えられた問題の主なものについて、例を挙げ、分類を試みる。表 2 に問題となった回答語の集まりを刺激語と共に示す。

表 2: 回答語の表記揺れに関わる問題例

例	刺激語	関連性	回答語
1.1	爆笑	連想	面白い, おもしろい
1.2	コマーシャル	連想	じゃま, 邪魔
1.3	熊	連想	さけ, 鮭
1.4	四季	連想	紅葉, もみじ
2.1	スペイン	連想	マタドール, マタドル
2.2	スーツケース	連想	ジュラルミン, ジェラルミン
2.3	スーツケース	連想	アタックケース, アタッシュケース
2.4	考える	連想	うでくみ, うでぐみ, 腕組み
3.1	やかましい	類義	みみざわりだ, 耳ざわりな
3.2	体調	連想	崩す, 崩れる
3.3	考える	連想	集中, 集中する
3.4	ダイナミック	連想	大胆, 大胆だ
4.1	スーツケース	連想	つめこむ, つめる
4.2	倒す	連想	戦い, 闘い
4.3	コマーシャル	連想	コマーシャルソング, CMソング
5.1	熊	連想	冬眼, 冬眠
5.2	熊	連想	三ヶ月, 三日月
5.3	住所	連想	年齢, 年令
5.4	倒す	連想	なぎ, なぎ倒す

(1) 表記に用いられる文字種の多様性

回答語として、漢字を含む表記を持つ単語とその読みに対応する平仮名表記や片仮名表記を持った語が現れることがある。これらの間の関係が一対一であれば、両者を、同じ単語の表記揺れであると判断することは妥当であると考えられる。例 1.1、例 1.2 はそのような場合である。

しかし、平仮名表記や片仮名表記と複数の漢字表記が対応づけられる場合、つまり、同音異義語が存在する場合には、問題が複雑になる。例 1.3 の「さけ」に対する漢字表記としては、刺激語の存在を無視した場合、「鮭」以外にも「酒」、「咲け」等が考えられるため、「さけ」と「鮭」を同じ単語の表記揺れであると即断できない。

同様に、漢字表記を持つ単語が複数の異なった読みを持つ場合も考えられる。例 1.4 の場合、「紅葉」は「こうよう」であるかもしれず、「紅葉」と「もみじ」を同じ単語の表記揺れであると即断できない。

(2) 外来語等の表記

外国語の単語が日本語に持ち込まれる場合には、できるだけ音が近い片仮名を充てて外来語とすることが一般的であろう。この場合の外来語の片仮名表記法については、[官報 91] のように許容する表記揺れの規則が制定されている。しかし、個人のレベルでの揺れは、そこに示されたものにとどまらない。同一の外国語と思われるものについても、様々な表記揺れが発生している。

例 2.1 の「マタドール」と「マタドル」は、[官報 91] に記載されている範囲内であるために、それを根拠として表記揺れとみなすことができる。例 2.2 の「ジュラルミン」と「ジェラルミン」は同一の外国語「duralumin」から発生していると推測されるが、[官報 91] にはそれを同一視する変換規則は存在しない。「ジュ」-「デュ」が述べら

れるにとどまっている。この問題は、被験者の誤解や誤用の問題とも関わってくる。例2.3 では、外国語「atachécase」に対してかなり例外的と考えられる「アタックケース」が回答されている。正しい表記と誤った表記の境界を客観的に定めることは難しいと考えられる。この問題を音の類似性から来るものととらえると、外来語にとどまらず、例2.4 に示した清濁の揺れもここに含まれる。

(3) 品詞や活用の違い

例3.1 のように、同じ刺激語に対する回答語として、異なった活用形の単語が現れることがある。活用形の違いは同一の単語の揺れであるとするなら、同一の語幹に対する活用形であるかどうかを判断することはそれほど難しくはないので、それを利用できる。例3.2 は、対となる自動詞と他動詞で、当然、音や意味は類似している。これを品詞の情報を基準として別の単語とするかが問題となる。関連した事例としては、サ変動詞化する名詞とそれに「する」がついた動詞の終止形という例3.3 や、形容動詞の語幹と終止形という例3.4 等があげられる。

(4) 単語の同義性・類義性

表記や音が類似しており、さらに意味が近いという場合、これが別の単語であるかどうかの判断は簡単でない。例4.1 は、表記と活用が異なるが音の一部と意味が近いと思われる回答語の例である。例4.2 は、音が同一で意味も近い場合であり、国語辞典[金田一 88] では、見出し表記の揺れとして扱われている。他にも「暑い」「熱い」や「交代」「交替」など、漢字表記の揺れなのか別の単語なのかの判断は必ずしも自明ではない。類義性という面では、例4.3 のように略語とその元となった単語も同様の問題がある。

(5) 誤りなど

自由回答形式のアンケート結果であることに起因する事例も見られる。そのひとつは被験者の誤りに起因すると推察されるものである。その中には、例5.1 のように存在しないと思われる単語が現れることあるし、例5.2 のように、存在はしているがおそらく被験者の意図とは異なるであろうと思われる単語も現れている。また、例5.3 に示した「年令」は、文献[国研 83] でも指摘されている通り、ある辞典では誤用（正しくは「年齢」）とされているが、別の辞典では慣用として見出しに「年齢」と併記されている。このように誤用と慣用の間は連続的であり、これも簡単に判断できるものではない。また、アンケート内容に強く関連するものであるが、例5.4 のように刺激語を含んだ単語から刺激語を取り除いた単語の部分が回答語となっている場合もある。特にこの例では、「なぎ」が「凪」である可能性も捨てきれないことが問題となる。これは一般的には揺れと呼べるものではないが、その扱いについての指針が必要になる。

3.2 表記揺れ解消の試み

前節より明らかな通り、自由回答アンケートの回答語の表記揺れには様々な要因が関連しており、ある回答語と別の回答語が同じ単語の表記揺れなのか、それとも別の単語なのかの判断は、多くの場合自明ではない。この判断基準、表記揺れ解消規則を作成するにあたって、筆者らは、その規則に基づいて行う作業が機械的なものとなること、つまり作業者の主観が入り込むことなく、誰が作業しても同じ結果になることを最重視した。そして、そのような判断基準が作成できないものについては、原則として、回答語どうしを別の単語とすることとした。

ここで、作業者がその判断に用いる情報としては、まず第一に、**特定の国語辞典の見出し表記**を用いた。つまり、国語辞典中の見出しの表記揺れとして扱われている表記が回答語中に表れている場合、同一の単語と判断するものである。国語辞典は表記に関して長年検討が加えられた結果として得られたものであり、基本的な判断基準となりうると考えた。しかし、複数の国語辞典ではその判断が異なることがあるため、適当な一冊[金田一 88] を利用した。この基準は、前節の(1) - (5) を全般的に支えることになる。特に同義性の問題などについては、アンケートの本来の目的と関わる重要な判断となるが、国語辞典だけを判断の根拠として採用することにより、明確な作業基準が確立できることになる。

それ以外に用いたのは、完全に並べ上げることのできる、もしくは品詞種別などの客観的な用語で記述できる条件のみとした。例えば、長音記号の有無やバビブベボ／ヴァヴィヴヴェヴォの違いによる表記の対立は、その判定に主観が入る混む余地はないとして採用し、この場合は、その対立を表記揺れとした。

ひとつ特徴的なのは、得られた回答語の集合を参照した判断を加えたことである。前節の(1) で述べた平仮名表記と同じ読みの漢字表記の問題では、「平仮名表記を持った回答語は、それと同じ読みを持つ漢字表記の単語が同じ刺激語の回答語として複数現れる場合は、別の単語として扱い、そのような漢字表記の単語がただ一つしかない場

合は、その単語の表記揺れとする」という基準を用いた。例えば、先の例で、「熊」の連想語として、「酒」や「咲け」が回答に含まれていない場合は、「さけ」は「鮭」の表記揺れとされる。一方、例えば、刺激語「街」に対する連想語として「がいとう」、「街灯」、「外灯」、「街頭」が現れたが、この場合、「がいとう」はいずれかの単語の表記揺れではなく、それ一つで単語とされる。このような方法によって、作業者の主観を交えることなく、しかも常識的な線で回答語を表記揺れとしてまとめることができる。

ただし、残念ながら、幾つかの事例では、「慣用」という判断基準を入れざるを得なかった。例えば、「マドリード」「マドリッド」がそのような基準で表記揺れと判断される。この慣用の判断については、複数の作業者の合意が必要であるとした。

これらの方針に基づいて、調査結果に対して回答語の表記揺れ解消作業を進めながら、最終的には 44 の表記揺れ解消規則を作成した。詳細は別紙に示す通りであるが、これを分類したものが表 3 である。

前章で眺めた現象のうち、(1) 表記に用いられる文字種の多様性は、分類 1 の規則で、(2) 外来語等の表記は分類 3 の規則で、(3) 品詞や活用の違いは分類 7,8 の規則で、(4) 単語の同義性・類義性にかかわる問題は分類 2,10 の規則で主に対処されている。これらは、国語辞典の情報に基づく判断と客観的に記述された条件に基づいて処理される。ただし、一部に慣用という判断基準が含まれるのは前述の通りである。

(5) 誤りなどのうち、誤りに関するものは分類 11 で扱われているが、その方針は、いくら誤りと推測されても別の単語とするというものである。同様に、刺激語に関連する問題は、分類 9 で扱われているが、刺激語が除かれた単語の部分である可能性が高いものも、「なぎ」のように他の単語である可能性があれば、別の単語としている。これらは、いずれも、機械的作業を可能とする判断基準が作成できないものについては、原則として、回答語どうしを別の単語とするという原則を反映したものである。

以上のような表記揺れ解消規則に加えて、カテゴリー化においては、複数の回答語が同じ単語の表記揺れと判断された時、それらの代表となる表記、代表表記を決定する必要がある。つまり、「熊」「くま」「クマ」が同じ単語の表記揺れと判断された場合、その後、どの表記でこの単語を代表させるかを決定するのである。この基準としても、筆者らは、回答語の集合を参照した判断と、国語辞典の情報を利用した。つまり、同じ語の表記揺れとされた表記の中で、最も多くの被験者が回答に用いた表記を代表表記とする方針とした。そして、2 つの表記の回答者数が同一の場合、それらが国語辞典でも表記揺れとされている場合は、より一般的なものが先になっているので、その内容を利用した。このいずれの方法でも決定でできないものについては、作業者の判断に任せた。以上が原則であり、個々の表記揺れ解消規則における代表表記の決定については、別紙に示した通りである。

表 3: 表記揺れ解消の規則の分類

No	解消規則の分類名	規則数	規則番号 (別紙)	適用回数
1	文字種の違いに関する規則	4	1,2,3,4	125
2	漢字表記に由来する規則	4	5,6,7,8	18
3	外来語の片仮名表記に関する規則	3	9,10,11	17
4	記号に関する規則	3	12,13,14	6
5	文字・単語の新旧に関する規則	4	15,16,17,18	4
6	接辞に関する規則	5	19,20,21,22,23	8
7	品詞の違いに関する規則	6	24,25,26,27,28,29	-
8	語尾の形に関する規則	3	30,31,32	2
9	刺激語の有無に関する規則	3	33,34,36	-
10	意味が近い単語に関する規則	7	37,38,39,40,41,42,43	-
11	被験者の誤りに関する規則	1	44	-
12	その他	1	35	-
	計	44		180

3.3 表記揺れ解消の結果

前節で説明した回答語の表記揺れ解消のための 44 規則について、適用される傾向を調査した。刺激語 100 語より 10 語を選出し、連想語・類義語・対義語 160 語 (回答語の表記揺れ解消後) について検証した。表 3 には、表記揺れと判定する際に用いた規則の適用回数が示されている。分類 1(文字種に関する規則) の規則が、回答語の表記揺れを決定する際に数多く適用されている。規則の具体性の違いもあるが、表記揺れの大きな原因が文字種の多様性にあることがわかる。一方、後半の分類は、表記揺れと判定する際の適用回数は少ない。これらの規則は主として、音や表記、意味の近い回答語どうしは表記揺れではないと判定するための規則であり、実際に作業者は、44 規則を全般的に考慮して表記揺れ解消の作業を行っていたのである。

最終的に集計された回答の表記揺れ数は、最大 7 のものから表記揺れ数が 1、すなわち、揺れが全くないものまで存在した。表 4 は、ひとつの単語と判断されたものが回答時にどれだけの表記揺れを持っていたかに関する情報を 3 種類の関連語について合計したものである。異なり数で見ると、表記揺れ数の多い代表語が、全体に占める割合は必ずしも高くない。例えば、表記揺れ数 6 の代表語は、わずかに 9 語しか存在しないが、表記揺れ数 1 の代表語は約 37,000 語存在し、代表語数の 9 割以上を占める。しかし、表記揺れを考慮して実際に現れた頻度を代表語数で平均した場合、表記揺れ数 1 の回答は平均 2 回程度しか現れていないのに対して、表記揺れ数 6 の回答は 30 回程度現れており、表記揺れの解消は、単語の頻度を議論する場合に重要な問題となることがわかる。

表 5 に、特に表記揺れが多かった単語の一部を示す。主として分類 1(文字種の違いに由来する規則) の規則が関わっているが、さらに慣用や様々な規則とも関わっていることがわかる。

表 4: 回答の集計結果

回答の表記揺れ数	代表語数 (異なり)	頻度	頻度／代表語数
7	1	36	36.0
6	9	286	31.8
5	22	359	16.3
4	82	1,821	22.2
3	480	7,321	15.3
2	2,525	21,081	8.3
1	37,016	83,684	2.3

表 5: 表記揺れが多い回答語の例 (連想語)

刺激語	回答語の表記揺れ (回答者数)
バケツ	ぞうきん(25), 雑巾 (5), ぞーきん (1), ぞう巾 (1), ゾウキン (1), 雑きん (1)
カボチャ	ハロウィン(27), ハローウィン (6), ハロウィーン (5),Halloween(1), ハローウィーン (1), ハロウイン (1)
遅い	待ち合わせ(14), 待ちあわせ (2), 待ち合せ (2), まちあわせ (1), まち合わせ (1), 待合わせ (1)
金庫	泥棒(13), どろぼう (10), ドロボウ (2), どろ棒 (1), ドロボー (1), 泥ぼう (1)
マッシュルーム	缶詰(5), 缶づめ (4), 缶詰め (4), かんづめ (3), かん詰め (1), カン詰 (1)
口座	引き落とし(4), ひき落とし (2), 引きおとし (2), 引落し (2), ひきおとし (1), 引落とし (1)

(下線の表記は代表語)

4　考察

自由回答アンケート形式で単語を収集する際に問題となる単語の表記揺れの問題について述べ、可能な限り機械的に作業するという方針でそれを解消する規則群を提案した。更に、その処理によってどのようなものが表記揺れと判断されるかを示した。

表記揺れ解消規則の作成において残された問題のうち、最も重要なものは、幾つかの規則で用いられている「慣用」の判定である。しかし、例えば、「年齢」と「年令」、別紙の[規則8]に例示された「たちぐいそば」の様々な異表記と思われるもの、「マドリッド」等の幾つかの外来語等々、国語辞典等には明記されていないが、10 人に尋ねて10 人が表記揺れだと判断するものがやはり存在する。これらの判断を可能とする明文化された情報が望まれる。

一方、被験者の誤りに起因すると思われるものについては、慎重な姿勢をとったために、その多くが異なる単語として残されることとなった。これらは頻度が少ないとはいえ、収集結果を用いた今後の研究で用いられる材料として残されるので、その悪影響が心配される。アンケート時に直接コンピュータに回答を投入させるなどして、誤字や誤用の割合を減らすというのも、異なった方向からの有効なアプローチであろう。更にこのような方法では、手書きで回答させた場合に比べ、平仮名表記が少なくなるのではないかと期待される。現時点では、複数のコンピュータを用意する必要があるために、郵送による方法に比べて回答者数を多くすることができないという問題はあるが、WWW の利用などにより今後はますます現実的なものとなろう。もちろん、コンピュータへの入力においては、手書きの場合とは異なった誤りが生じることが知られており、今回とは異なる問題が新たに引き起こされる可能性も無視できない。

また、今回、表記揺れ解消の作業は作成した規則を元に人手で行ったが、作成した規則は機械的作業を目標とするものであるから、それを自動化することも不可能ではないだろう。例えば、品詞の判定などは簡単な形態素解析器と電子化辞書により実現できる。もちろん、上であげた問題は残るから、完全な自動化は困難であろうが、一部の自動化は充分に可能である。そのようなシステムはある意味での校正システムと考えられないこともない。例えば、片仮名語の表記揺れについての研究[久保田 93, 飯田 94]や人名の表記揺れ解消[高橋 92]等の成果と組み合わせることも考えられる。

最後に、今回、異なった文字種の対立については、回答語の集合を参照して表記揺れとすることにしたが、「クマ」「熊」が単なる表記揺れかという点については疑問も残る。例えば、表記の親近性と単語の熟知性や出現頻度との関連性の研究[広瀬 84]や、状況によって表記に用いる文字種が異なるという研究[八田 99, 岩原 00]もある。この点は、単語収集の目的なども関連して、より細かい判断が必要となるだろう。

5　おわりに

本稿では、単語で回答する自由回答アンケートに共通する問題である、広い意味での単語の表記揺れの問題に対して、ケーススタディとして我々の調査の集計過程で現れた表記の問題がどのようなものであったかを報告し、それを我々がどのように解決したかについて説明した。作業者が機械的に適用できることを重視して作成した回答語の表記揺れ解消のための規則には、慣用や誤用等の問題が課題として残されているが、類似したアンケート調査における表記揺れ解消規則作成の参考となれば幸いである。

参 考 文 献

[天野 00]　天野, 近藤：日本語の語彙特性, 三省堂 (2000).

[飯田 94]　飯田, 中村：変形ルールと禁則ルールを用いた片仮名の表記揺らぎの解消法, 情報処理学会論文誌, Vol. 35, No. 11, pp. 2276 –2281(1994).

[岩原 00]　岩原, 八田：日本語書字における情動情報の伝達メカニズム, AI 学会研究会資料, SIG-J-A002-9, pp. 39 – 44 (2000).

[笠原 97]　笠原, 松澤, 石川：国語辞書を利用した日常語の類似性判別, 情報処理学会論文誌, Vol. 38, No. 7, pp. 1272–1284 (1997).

[官報 91]　内閣告示第2 号 外来語の表記, 官報 (1991).

[金田一 88]　金田一, 池田（編）：学研 国語大事典 第二版, 学習研究社 (1988).

[久保田 93]　久保田, 庄田, 河合, 玉川, 杉村：片仮名表記の統一方式, 情処研報, NLP-97-16, pp. 111 – 117 (1993).

[国研 83] 国立国語研究所：現代表記のゆれ, 秀英出版 (1983).

[高橋 92] 高橋, 岩瀬：人名の読みからの検索法, 情処研報, NLP-91-4, pp. 25 – 31 (1992).

[辻 87] 辻, 有馬：アンケート調査の方法, 朝倉書店 (1987).

[八田 99] 八田, 岩原：日本語書字における表記選択メカニズムについて, AI 学会研究会資料, SIG-LSE-9902-3, pp. 16 – 21 (1999).

[広瀬 84] 広瀬雄彦：仮名単語の認知における全体処理の検討, 心理学研究, Vol. 56, No. 1, pp. 44 –47 (1984).

(別紙) 表記揺れ解消規則一覧

注: 例は、その規則が適用された例。ただし、他の規則を併用している場合もある。

[**規則 1**] 異なる文字体系間の対立 (平仮名／片仮名／漢字／アラビア数字／アルファベット) は規則 2 に該当する場合を除き、表記揺れとする。代表表記は回答者数が最も多い表記とし、回答者数が同じ場合の優先順は、漢字、平仮名、片仮名、アルファベット、アラビア数字とする。(ただし、回答語が外来語で、かつ回答者数が同じ場合は、片仮名表記を優先する)

例: 風邪 (8), かぜ (6), カゼ (6) → 風邪 (20)　　死んだふり (4), 死んだフリ (1) → 死んだふり (5)

[**規則 2**] どの表記揺れグループに属するか特定できない場合 (例えば、平仮名の回答語が複数の漢字の回答語の読みと一致する場合) は、別の単語とする。

例: 街灯 (14), がいとう (1), 外灯 (1), 街頭 (1) → 街灯 (14), がいとう (1), 外灯 (1), 街頭 (1)

[**規則 3**] アルファベットの大文字／小文字の対立は、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は大文字を含む表記を優先する。

例: ドリーム (41), Ｄｒｅａｍ (1), ｄｒｅａｍ (1) → ドリーム (44)

[**規則 4**] 記号とその読みは、別の単語とする。

例: 丸 (5), ○ (2), まる (1) → 丸 (6), ○ (2)　　! (1), ビックリマーク (1) → ！, ビックリマーク (2)

[**規則 5**] 国語辞典 [金田一 88] で、同じ見出し語の表記として記載されている表記の対立は、表記揺れとする。代表表記は回答者数が最も多い表記とし、回答者数が同じ場合の優先順位は、国語辞典の掲載順とする。

例: 整える (7), ととのえる (4), 調える (1) → 整える (12)　　戦い (3), たたかい (1), 闘い (1) → 戦い (5)

[**規則 6**] 国語辞典 [金田一 88] で同じ見出し語の送りがなとして扱われている表記の対立は、表記揺れとする。代表表記は回答者数が最も多い表記とし、回答者数が同じ場合の優先順位は、国語辞典における優先順とする。

例: 売り上げ (1), 売上げ (1) → 売り上げ (2)

[**規則 7**] 国語辞典 [金田一 88] で、語義文の記載がなく、他の見出しを指定し、参照せよとしている見出し語の表記と、参照先の見出し語の表記との対立は、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は参照先の見出し語の表記を優先する。

例: 鮭 (7), サケ (4), さけ (3), シャケ (3) → 鮭 (17) (「しゃけ【鮭】」の項目に、「さけ【鮭】」を参照せよとの指示あり)

[**規則 8**] 国語辞典 [金田一 88] に収録されている表記と、収録されていないが慣用と見なせる表記との対立は、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は国語辞典に収録されている表記を優先する。

例: 年齢 (5), 年令 (3), 年れい (1) → 年齢 (9)

立ち食いそば (4) 立ち喰いそば (4) たちぐいそば (1) 立ち食いソバ (1) 立喰いそば (1) → 立ち食いそば (11)

[**規則 9**] 長音記号の有無による表記の対立は、表記揺れとする。

例: ベアー (3), ベア (1) → ベアー (4) 代表表記は回答者数が多い表記とし、回答者数が同じ場合は長音記号を含む表記を優先する。

[**規則 10**] バビブベボ／ヴァヴィヴヴェヴォの違いによる表記の対立は、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合はバビブベボを含む表記を優先する。

例: ビバルディ(11), ヴィヴァルディ(3), ヴィバルディ(1) → ビバルディ(15)

[**規則 11**] その他の外来語の表記の対立は、慣用と見なせるもののみ、表記揺れとする。代表表記は回答者数が最も多い表記とし、回答者数が同じ場合は国語辞典 [金田一 88] に収録されている表記を優先する。回答者数が同じで、かつ国語辞典に収録されている表記が無い場合は、最も一般的であると思われる表記を優先する。

例: パエリア(8), パエリヤ(7) → パエリア(15) マドリード(1), マドリッド(1) → マドリード(2)

[**規則12**] 「」の有無による表記の対立は、表記揺れとする。代表表記は「」を含まない表記とする。

例: 考える人(17),「考える人」(2) → 考える人(19)

[**規則13**] 「・」、「＝」、「.」、全角スペース等の区切り記号の有無や違いによる表記の対立は、表記揺れとする。人名の片仮名表記は、たとえ「・」を含む表記を記入した被験者がいなくても、代表表記の姓名の間に「・」を入れる。それ以外の場合の代表表記は、回答者数が多い表記とし、回答者数が同じ場合は区切り記号を含まない表記を優先する。

例: ケビンコスナー(11), ケビン・コスナー(6), ケビン＝コスナー(2) → ケビン・コスナー(19)

[**規則14**] 「！」、「？」等の文末記号の有無による対立は、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は文末記号を含まない表記を優先する。

例: Ｓｈａｌｌ　ｗｅ　ｄａｎｃｅ(1), Ｓｈａｌｌ　ｗｅ　ｄａｎｃｅ?(1), シャルウィーダンス(1) → シャルウィーダンス(3)

[**規則15**] 旧字／新字の違いによる表記の対立は、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は新字を含む表記を優先する。

例: 慶応(1), 慶應(1) → 慶応(2)

[**規則16**] 歴史的かな遣いによる表記の対立は、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は歴史的かな遣いを含まない表記を優先する。

例: やはらか(1), やわらか(1) → やわらか(2)

[**規則17**] 単語とその文語形との対立は、別の単語とする。

例: ない(2), 無い(2), なし(1), 無し(1) → 無い(4), 無し(2)

[**規則18**] 単語とその俗語との対立は、別の単語とする。

例: やわらかい(21), 柔らかい(5), やらかい(1) → やわらかい(26), やらかい(1)　　あたたかい(2), あったかい(1) → あたたかい(2), あったかい(1)

[**規則19**] 接頭語「お」の有無による表記の対立は、「お」を含む表記が国語辞典[金田一88]の見出しに収録されている場合を除き、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は「お」を含まない表記を優先する。

例: お金(14), 金(10) → お金(24)　　すもう(5), 相撲(5), おすもう(1) → 相撲(11)

[**規則20**] 接頭語「大」の有無による表記の対立は、別の単語とする。

例: 荷物(9), にもつ(1), 大荷物(1), → 荷物, 大荷物(11)

[**規則21**] 「一」「一つ」等の数を表す語の有無による表記の対立は、別の単語とする。

例: 一年(7), １年(3), 年(1) → 一年(10), 年(1)

[**規則22**] 単語の単数形と複数形の違いによる表記の対立は、別の単語とする。

例: 山(28), 山々(1) → 山(28), 山々(1)

[**規則23**] 接尾語の有無による表記の対立は、別の単語とする。

例: ラテン(9), ラテン系(1) → ラテン(9), ラテン系(1)

[**規則24**] 自動詞と他動詞との対立は、別の単語とする。

例: くずす(11), 崩す(9), 崩れる(1) → くずす(20), 崩れる(1)

[**規則25**] 名詞と動詞との対立は、別の単語とする。　　例: 休み(1), 休む(3) → 休み(1), 休む(3)

[**規則26**] 名詞と形容詞との対立は、別の単語とする。例: 黒(7), くろ(1), 黒い(5) → 黒(8), 黒い(5)

[**規則27**] 形容詞の終止形と連用形との対立は、連用形が副詞として国語辞典[金田一88]の見出しに収録されている場合を除き、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は終止形を優先する。

例: すごい(2), すごく(2) → すごい(4)

[**規則28**] 助詞の有無や違いによる派生語の対立は、別の単語とする。

例: ２人(6), 二人(5), ２人の(2), 二人の(1) → ２人(11), ２人の(3) ふんわり(2), ふんわりとした(1), ふんわりしている(1) → ふんわり(2), ふんわりとした(1), ふんわりしている(1)

[**規則29**]

規則24〜28 以外の派生語は、別の単語とする。例: コンディショニング (1), コンディション (2) →コンディショニング (1), コンディション (2)

[**規則30**] 形容詞と、形容詞の連用形に動詞「なる」がついた語との対立は、別の単語とする。

例: 固い (2), 固くなる (1) → 固い (2), 固くなる (1)

[**規則31**] サ変動詞の活用語尾「する」の有無による表記の対立は、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は、活用語尾を含まない表記を優先する。

例: アピール (1), アピールする (1) → アピール (2)

[**規則32**] 形容動詞の活用語尾の有無による表記の対立は、表記揺れとする。代表表記は回答者数が多い表記とし、回答者数が同じ場合は、活用語尾含まない表記を優先する。

例: 派手 (15), はで (2), ハデ (2), 派手だ (1), 派手な (1) → 派手 (21)

[**規則33**] 刺激語の有無による対立は、別の単語とする。

例: レオナルド (3), レオナルド熊 (4) → レオナルド (3), レオナルド熊 (4) (刺激語は「熊」)

[**規則34**] 刺激語＋助詞の有無による対立は、助詞や活用語尾等が先頭もしくは末尾にあるためにそれだけでは意味をなさない場合を除き、別の単語とする。助詞や活用語尾等が先頭もしくは末尾にある場合は、刺激語を補って表記揺れとする。

例: ぬいぐるみ (8), 縫いぐるみ (1), 熊のぬいぐるみ (1) → ぬいぐるみ (9), 熊のぬいぐるみ (1)
緊張の糸 (1), の糸 (1) → 緊張の糸 (2)(刺激語は「緊張」)

[**規則35**] 読みの清濁による平仮名表記の対立は、表記揺れとする。代表表記は回答者数が多い表記とする。回答者数が同じ場合の優先順は、漢字、平仮名とする。同数の平仮名表記どうしでは、国語辞典[金田一 88] に収録されている表記を優先する。

例: うでくみ (1), うでぐみ (1), 腕組み (1) → 腕組み (3)

[**規則36**] 刺激語を補わなければ単語として成り立たない濁音を持つ表記は、清音を持つ表記及び漢字表記とは別の単語とする。

例: さお (12), ざお (1), 竿 (4) → さお (16), ざお (1)(刺激語は「釣り」)
ごと (1), 事 (1) → ごと (1), 事 (1)(刺激語は「心配」)

[**規則37**] 助詞「の」の有無による表記の対立は、別の単語とする。

例: 胃のもたれ (1), 胃もたれ (1) → 胃のもたれ (1), 胃もたれ (1)

[**規則38**] 規則5 に該当しない同義語、類義語は、別の単語とする。

例: 重要 (1), 大切 (1) → 重要 (1), 大切 (1)

[**規則39**] 単語と、その単語の意味を含むような語及び句との対立は、別の単語とする。

例:考える人 (17),「考える人」(2), 考える人の銅像 (1) → 考える人 (19), 考える人の銅像 (1)　　ロダン (20), ロダン彫刻 (1) → ロダン (20), ロダン彫刻 (1)

[**規則40**] 単語とその略語との対立は、別の単語とする。

例: コマーシャルソング (2), CMソング (2) → コマーシャルソング (2), CMソング (2)

[**規則41**]

音や意味が似ている擬声語や擬態語は、別の単語とする。例: ランラン (3), ランランラン (5) → ランラン (3), ランランラン (5)　　コロコロ (1), ゴロゴロ (1) → コロコロ (1), ゴロゴロ (1)

[**規則42**] 人名とその略称との対立は、別の単語とする。

例: K．コスナー (1), ケビンコスナー (11), ケビン・コスナー (6), ケビン＝コスナー (2) → K．コスナー (1), ケビン・コスナー (19)

[**規則43**] 括弧書きによる注釈がある場合は、注釈の無い単語とは別の単語とする。例: 速い (5), はやい (1), 速い（速度が）(1) → 速い (6), 速い（速度が）(1)

[**規則44**] 誤字が含まれると思われる単語は、被験者が意図したと推定できる表記の単語があっても、それとは別の単語とする。

例: 冬眼 (1), 冬眠 (8) → 冬眼 (1), 冬眠 (8) (刺激語は「熊」)　　猟師 (1), 漁師 (1) → 猟師 (1), 漁師 (1) (刺激語は「熊」)