テキストにおける同音異義語の分布

著者	山崎 誠
雑誌名	言語資源活用ワークショップ発表論文集
巻	6
ページ	204-209
発行年	2021
URL	http://doi.org/10.15084/00003494

テキストにおける同音異義語の分布

山崎 誠（国立国語研究所研究系言語変化研究領域）†

Distribution of Homonyms in Japanese Text

Makoto Yamazaki (National Institute for Japanese Language and Linguistics)

要旨

日本語には漢語を中心に同音異義語が多いと言われる。国立国語研究所(1961)『同音語の研究』は同音異義語に関する総合的な研究であるが，実際の個々の文脈において同音語がどのくらい出現するかという調査は管見の限り見当たらない。本研究では『現代日本語書き言葉均衡コーパス』を利用して，1 サンプル中に漢語の同音異義語がどの程度現れるかを調査したものである。調査単位は短単位である。結果は，調査した図書館書籍（LB)の 10551 サンプルのうち，95.4%のサンプルに同音異義語の組が少なくとも 1 つ現れていた。同音異義語の組み合わせで多かったもの（頻度 10 以上）1082 組を見ると，7 割弱は「方・法」や「社・者」のような一字漢語が多く，「以上・異常」「自信・自身」のような二字漢語同士の組み合わせは約 3 割であった。またテキストに出現する同じ読みを持つ二字漢語の組み合わせを調べると，少なくとも約 6 割のサンプルに同音二字漢語の同音異義語が現れていることがわかった。

１．はじめに

日本語には同音語（同音異義語[1]）が多い，という主張は日本語の特徴の一つとしてたびたび登場する。金田一（1988: 159）は日本語の単語を形態の面から眺めたときの特色の一つとして「日本語の単語は、形の上から見て同音語の多いことが有名である。」と述べている。衣畑（編）（2019: 126）でも「現代日本語の語彙」（執筆，金愛蘭）の章で「日本語には同音語が多いといわれるが、それは、音節の種類が少なく、その組み合わせにも制限があるという音節構造上の理由に加えて、字音という限られた音節から成る漢語や、その略語が多いという語種・語構成上の理由があるからだと考えられる。」のように同音語の多いこととその理由が示されている。同音語に関する最初の総合的研究である，国立国語研究所（1961）でも刊行のことばの冒頭が「日本語には同音語が多いと書われる。」で始まる。日本語学だけでなく，徳広（2017: 11）や秋元他（2019: 46）のような日本語教育や梅澤・大澤（2017）のような情報処理の分野の文献にも同音語が多いという指摘が登場する。

このように日本語に同音語が多いということが繰り返し主張されているが，それについて具体的な証拠を挙げている文献は少ない。この主張は他の言語と比較しないと言えないものであるが，他の言語に触れているのは管見の限りでは，前述の金田一（1988: 160）と望月（1974）である。金田一（1988: 160）では，「同音語が多いことでは、上には上があって、中国語やタイ語は日本語以上である。」と述べているが具体的な典拠や数字などは挙げられていない。望月（1974）では，日中の辞書の見出し語をデータとして同音語の割合を比較している。日本語のデータとして『新明解国語辞典』（初版，金田一京助他編，1972 年），中国語のデータとして『漢語併音詞彙（増訂稿)』（中国文字改革委員会訶彙小組編，1963 年，文字改革出版社）を用い，「同音語の割合は，どちらも 35～36%程度で，ほぼ変わりが

† yamazaki [AT] ninjal.ac.jp

1 同音語に同訓意義語（「暑い」「熱い」「厚い」「篤い」等）を含める場合もある（秋元他（2019: 46））。

ない.」と指摘している[2]。望月（1974）は続けて「ただし，中国語では，声調（アクセント）を考慮に入れると，まったくの同音語は 11.6%に減ってしまう．日本語でも，それを考慮すれば，多少は減るが，4 拍の漢語のほとんどが平板型に属するように，アクセントの差は意味の弁別にあまり有効ではないということである.」と記している。

２．先行研究

同音語に関する本格的な研究は国立国語研究所（1961）である。同書では，国語辞典を初め各種用語集等を元にした総合的な分析を行っている。辞書や用語集を用いているため，具体的な文脈における分析はほとんど行われていない[3]。巻末には 154 ページにおよぶ同音語集[4]が付けられているが使用頻度の情報は付けられていない。田中（1971）は新聞の語彙調査のデータに基づくものである。出現頻度の付いた同音短単位表が付いている。中野（1989）は教科書の語彙調査の結果に基づいたものである。意味分野と同音語の関係の分析が行われたほか，調査で用いられた M 単位・W 単位による同音語のリストが掲載されている。山崎（2004）も雑誌の語彙調査のデータに基づくものであり，「意味分野と使用頻度から約 8 割の同音語が識別できる可能性がある」と述べている。

これらの研究は大規模なデータ全体を対象として同音語をマクロな観点から分析するものである。中野（1989）の同音語のリストからは高校教科書で「化学」が 293 回，「科学」が 129 回出現していることが分かるが，それらが同じ文脈で現れているのかどうかは分からない。同様のことは山崎（2004）にも言える。

そこで本研究では，同音語が実際に一定の長さの文脈にどのくらいの割合で現れるのかを明らかにすることを目的とする。もし，同じ文脈に現れないのであれば，同音語の取り違えのような心配は比較的少ないのではないかと推測される。

３．同音語の認定基準

同音語の認定基準は複数のものがある。まず，話し言葉（音声）を基準とするか，書き言葉（表記）を基準とするかの違いがある。前者は日本語の場合アクセントの違いが同音語の認定に関わってくる。例えば，「公開」と「航海」とではアクセントが異なるため同音語にはならない。書き言葉（表記）を基準とした場合も，「州都（しゅうと）」と「シュート」のように標準的な表記で書き分けがあると同音語にはならない[5]。また，別の観点として，語彙素レベルと出現形レベルでの違いがある。前者では活用が捨象されるが，後者は活用形も考慮される。したがって，「来る」の命令形「来い」と「故意」「鯉」などが同音語となる。更に言語単位の違いも同音語の認定に関わってくる。動詞のテ形を活用形として認める，あるいは，短単位の連続も含めれば「切手」と「切って」が同音語になる。国立国語研究所（1961: 40）で例が挙げられている，「指揮権」と「識見」，「歯科医」と「斯界」などの同音複合語なども言語単位の違いの例と見なすことができる。

４．データ

本稿で用いるデータは，『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）の図書館サブコーパス（LB）である。データは BCCWJ の DVD 版（ver. 1.1）を利用した。LB には 10551 のサンプルが含まれており，固定長と可変長[6]を合わせた語数の平均値は 3886.6 語，中央値は 2924 語である。

---

[2] 引用は林（1982: 132）より。

[3] 事業所におけるテレタイプ，カナタイプのデータを用いた分析がいくつか報告されている。

[4] 田中（1971: 132）によれば，この同音語集には 7803 セット（約 25000 語）の同音語が収められているとのことである。

[5] ここでの書き分けは，平仮名と片仮名の違いではなく，長音表記の違いのことである。

[6] 固定長，可変長については，https://ccd.ninjal.ac.jp/bccwj/sampling.html　を参照のこと。

また，この調査では，同音語の範囲は語彙素読みが同じ短単位の漢語に限った。

５．結果

５.１　同音語が出現したサンプル

10551 サンプルのうち，約 95.4%である 10062 サンプルに少なくとも 1 つ以上の同音語の組が現れていた。図 1 は，1 サンプルあたりの同音語の組数（延べ）の分布を示したものである。平均値は 9.81，中央値は 7 である。1 サンプルあたり，7 つ程度の同音語の組が現れていたことになる。

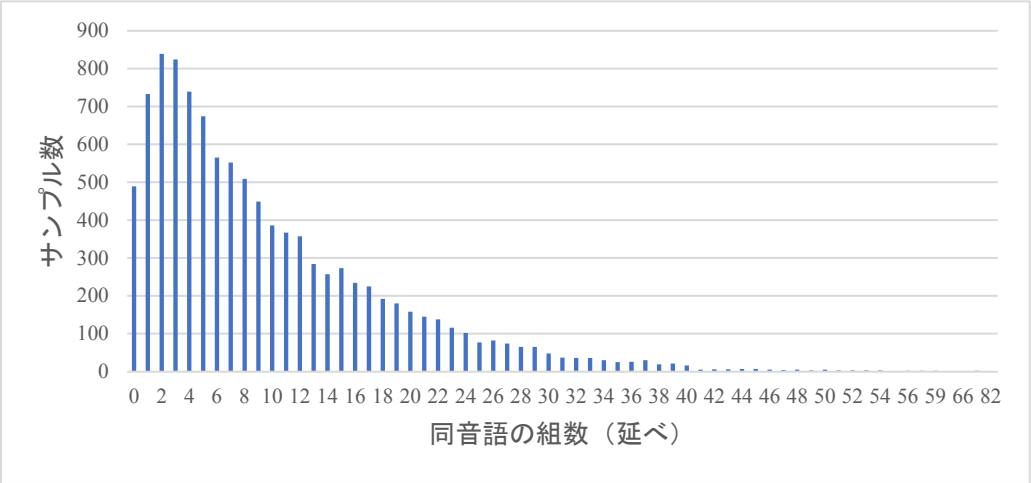

図１　1 サンプルあたりの同音語（漢語）の組数の分布

５.２　出現頻度の多い同音語

表 1 は，出現頻度の多かった同音語の組の上位 50 を示したものである。表 1 を見ると圧倒的に一字漢語が多いことが分かる。出現した同音語の組の数は 7540 組，そのうち頻度 10 以上は 1082 組であった。その 1082 組を文字数により分けると，一字漢語が 721 組，一字漢語と二字漢語の組み合わせ[7]が 21 語，二字漢語同士が 340 組であった。

表１　出現頻度の多い同音語（上位 50）

組み合わせ	頻度
様-用	2424
一-位置	1898
五-後	1867
方-法	1384
大-第	1327
社-者	1143
五-御	1004
者-車	884
敵-的	879

化-家	874
千-線	802
九-急	732
十-重	705
機-気	662
年-念	631
会-回	625
所-書	620
以上-異常	536
代-第	529

自信-自身	512
期-気	477
中-十	475
九-旧	462
性-生	453
五-語	430
店-点	427
歳-際	419
例-零	390
様-洋	389

7 「一」と「位置」，「差異」と「際」など。

代-大	364
役-約	360
以外-意外	359
三-山	349
上-場	345
器-気	345
実-日	324
時-次	320

性-所為	318
下-化	301
分-文	300
三-産	296
期間-機関	292
様-要	278
中-注	277
事態-自体	271

字-時	268
制-性	263
天-点	262
上-条	260
十-銃	257

表 2 は同音語でよく引き合いに出される二字漢語で出現頻度が多い例（頻度 50 以上）である。入力ミスでよく見るような例が多く見られる。

表 2　出現頻度の多い同音語（二字漢語，頻度 50 以上）

順位	組合せ	頻度
18	以上-異常	536
20	自信-自身	512
32	以外-意外	359
42	期間-機関	292
45	事態-自体	271
54	性格-正確	243
79	以前-依然	187
84	以来-依頼	178
92	機会-機械	164
93	家庭-過程	163
95	化学-科学	160
97	容易-用意	154
100	創造-想像	148
101	習慣-週間	147
105	時分-自分	140
112	事故-自己	130
119	減少-現象	124
120	夫人-婦人	122
121	対照-対象	120
133	当事-当時	109
138	協力-強力	108
139	以降-移行	107
142	最後-最期	105

148	指示-支持	95
150	向上-工場	95
156	一体-一帯	90
160	感心-関心	86
162	動揺-同様	83
163	生涯-障害	83
175	地震-自身	77
187	時間-次官	73
198	好意-行為	70
201	指摘-私的	68
209	一所-一緒	65
213	医師-意志	64
215	上体-状態	63
216	週刊-週間	63
217	人口-人工	63
219	上京-状況	62
231	効果-高価	59
237	史料-資料	57
243	先頭-戦闘	54
256	保証-保障	51
258	解放-開放	51
262	最近-細菌	50
264	意志-意思	50

５.３　語の長さと同音語

表 3 は語の長さと同音語の関係を示したものである。表 3 は，横軸が語の長さ（対象が漢語なので，ここでは漢字数[8]），縦軸が同音語となる語の数を示す。語の数が 1 というのは，サンプル内で，その読みとなる語が 1 つだけであった，すなわち，サンプル内に同音語が存在しなかったことを示す。そのような語の合計が 10551 サンプルに出現したことを意味する。これを二字漢語（縦の 2 の列）で見ると，組の数が 2 で 6525 サンプル，すなわち全体（10551 サンプル）に対して少なくとも約 6 割のサンプルに同音二字漢語の同音異義語が現れていることがわかる。

表 3 サンプルにおける語の長さと同音語の組数（サンプル数）

長さ / 語の数	1	2	3	4	5	6	8	計
1	10551	10551	6087	398	61	285	21	27954
2	9882	6525	1		1			16409
3	5805	466						6271
4	2064	38						2102
5	535	1						536
6	157	1						158
7	55							55
8	15							15
9	8							8
10	1							1
計	29073	17582	6088	398	62	285	21	53509

表 4 サンプルにおける語の長さと同音語の組数（語数）

長さ / 語の数	1	2	3	4	5	6	8	計
1	413367	2133471	9947	412	64	285	21	2557567
2	64136	20297	1		1			84435
3	14516	504						15020
4	3121	39						3160
5	649	1						650
6	174	1						175
7	58							58
8	15							15
9	8							8
10	1							1
計	496045	2154313	9948	412	65	285	21	2661089

[8] ただし，文字数 6 以上は漢語だが，ひらがな表記の語彙素である。

また，表 4 はサンプルにおける語の長さと同音語の組数ベースで集計したものである。これによると，同一サンプル中に現れる同音語の二字漢語数は，20842 語（表 4 の 2 の列の 2 行目から 6 行目の和）で全体（2154313 語）の約 0.97%であることが分かる。一方一字漢語では，この値が約 16.7%になる。すなわち，サンプルにおける同音語は一字漢語の場合が圧倒的に多いということが分かる。ただし，仮名漢字変換の際に一字漢語が単独で入力される場面はそれほど多くないと想像されることから，一定の文脈ではそれほど同音語に悩まされることなく入力ができるのではないかと推測される。

6. まとめと今後の課題

今回の調査では漢語のみに限って同音語の出現状況を調べたが，書き言葉の同音語で問題になるのは仮名漢字変化における語表記の選択の場面である。それを考慮すると，出現形のレベルでの同音語も視野に入れなければならないし固有名を含むすべての語種を対象にすべきである，たとえば，動詞や形容詞の活用形と他の語との同音語がどのくらいあるのかについても調査が必要であろう。

文　献

秋元美晴，押尾和美，丸山岳彦（2019）『日本語教育よくわかる語彙』アルク.

梅澤猛, 大澤範高 (2017)「音声による属性情報付加を用いたかな漢字変換候補の選択手法」「第 79 回全国大会講演論文集」2017(1), 17-18.

衣畑智秀（編）（2019）『基礎日本語学』ひつじ書房.

金田一春彦（1988）『日本語』（上），岩波書店.

国立国語研究所（1961）『同音語の研究』（国立国語研究所報告 20）秀英出版.
DOI: https://doi.org/10.15084/00001232

田中章夫（1971）「新聞語彙調査の同音語と同形語」『電子計算機による国語研究』3（国立国語研究所報告 39），121-145. DOI: https://doi.org/10.15084/00001008

徳弘康代（2017）「漢字語彙のモーラ音素の有無による類音語の調査とその資料の作成」「JSL 漢字学習研究誌」9（11-20）. DOI: https://doi.org/10.20808/jslk.9.0_11

中野洋（1989）「高校教科書の同音語」，国立国語研究所『高校・中学校教科書の語彙調査分析編』（国立国語研究所 99）77-131. DOI: https://doi.org/10.15084/00001350

林大（1982）『図説日本語』角川書店.

望月八十吉（1974）『中国語研究学習双書 13 中国語と日本語』光生館.

山崎誠（2004）「意味分野と使用頻度からみた同音語―二字漢語の場合―」『国際シンポジウム比較語彙研究 7』139-146.

資　料

国立国語研究所（2011）『現代日本語書き言葉均衡コーパス』
https://ccd.ninjal.ac.jp/bccwj/