

同音異義語の解析 - 専門分野における -

田中康仁

岡坂良雄

長田孝治

日本ユニパック(株)

日本ユニパック(株)

日本総合技術研究所

1 はじめに

仮名漢字変換システムは日本語文の入力方法として多くのワード・プロセッサで使われている。この仮名漢字変換システムの大きな問題は、分かち書きの問題と同音異義語の問題である。ここでは同音異義語の問題を取り扱う。同音異義を減らす方法としては色々な方法が考えられている。特に専門用語辞書(個別辞書)を持てばよいと言われている。そして、このような仕組みはシステムとして提供されている。しかし、この辞書の中味をどのようにして作るかについてはあまり良い方法が提供されていない。筆者らはこの方法について日本科学技術情報センターのファイル进行分析するなかで一つの方法を作り出した。これは漢字列を文字別に分析し、漢字列が2~3文字列の部分抽出し、整理すれば、専門用語辞書として使えることがわかった。専門用語辞書の充実によりよいワード・プロセッサが作られることを期待する。これは日本語文入力の進歩にも大きな役割をはたすであろう。

2 専門用語辞書による同音異義語の解析

2-1 専門用語辞書の分野

専門用語辞書を作るには仮名漢字変換システムを稼働させながら学習させるという方法もある。しかし、これでは学習させるまでに時間がかかるので実際上無理である。また、専門用語辞書を作るには専門用語を仮名漢字変換の規則に従って手作業で集め、入力する方法もある。これは着実な方法であるが、初期作業の費用がかかる。そこで、ここでは日本科学技術情報センター(JICST)の抄録ファイル进行分析し、これを利用することにする。JICSTの抄録ファイルは次のような10の分野を含んでいる。

(1) 管理・システム技術 (2) 電気工学 (3) 化学・化学工学 (4) 環境公害 (5) 機械工学 (6) 原子力工学 (7) 物理・応用物理 (8) 金属・鉱山・地球科学 (9) 土木・建築工学 (10) 生命科学

これら以外の分野(例えば金融、保険……)の専門用語は集めることはできない。しかし、JICSTの抄録ファイルで占めている分野ですら仮名漢字変換用の専門用語辞書はできない。そこでこの分野の仮名漢字変換用辞書を作る。

2-2 漢字列の抽出

一つの分野で使用される用語の数はどの程度であるか、その件数がどのように増加するかを調べてみる。JICST抄録テープを利用し、抄録の中から漢字列を機械的に抽出し、分析した。調査の内容は次の通りである。

- (1) 調査データ: JICST抄録テープ(管理・システム技術VOL11巻1~12号, 12冊分)
- (2) 抽出方法: 文字種による機械的分かち書き
- (3) 抽出データ: 異なり漢字列の件数は第一表参照
- (4) 作成資料: 頭文字順KWIC, 末尾文字順KWIC

ただし、同一漢字列は1件にまとめて表示し、その漢字列が何件あったかを示した。このように同一漢字列をまとめると、次のような問題が発生する。

例1 作業する サ変動詞の語幹 作業を終る 名詞

例2 店番(テンバン), 店番(ミセバン)

“作業”という漢字列になると品詞情報がつかめない。このような場合は多い。

“店番”は前後の情報がないためテンバンかミセバンかわからないということが起る。

しかし、例2のような事例は発生件数がまれである。用語の機械的抽出、その問題点を分析するためには、圧縮した資料が役に立つ。漢字列の左右に文章を付けているため特殊な漢字列が発生しても、その原因を容易に分析することができる。また膨大な漢字列を分析するためには、おおまかな分析から、より詳細な分析へと移行すべきである。

2-3 漢字列の数量的分析(1)種類

第一表を調べると対象漢字列が増加するに従って漢字列の種類も増加している。しかし、5万件毎の種類増加は減少している。調査対象漢字列の件数を100万件、200万件と増加すれば5万件ごとの異なり漢字列の増加は、さらに減少すると予測できる。この漢字列をうまくファイルすることにより仮名変換用ファイルとなる。これは複合語処理に役立つ。約9万件程度の長単位漢字列辞書を分かち書きに利用すれば延べ5万の漢字列の11%程度の新しい漢字列が発生する。しかし、この中には次のようなものも含まれている。それゆえ、新しい漢字列の発生はもっと減るはずである。

調査漢字列延べ件数	漢字列の異なり種類	5万ごとの種類の増加件数
0 ~ 5万	14,202	14,202
~ 10万	24,280	10,078
~ 15万	32,929	8,649
~ 20万	40,518	7,589
~ 25万	47,785	7,267
~ 30万	54,988	7,203
~ 35万	62,045	7,057
~ 40万	69,084	7,039
~ 45万	75,527	6,443
~ 50万	81,594	6,067
~ 55万	87,652	5,958
~ 60万	93,216	5,564

例 解法及び、上記物質

2-4 漢字列の数量的分析(2) 漢字列の長さ

漢字列の長さ(文字数)を調べた。この結果は第2表、第1図を参照していただきたい。この表を分析すると次のことがわかる。

- (1) 種類では4文字のものが最も多い。次に3文字、5文字の順になっている。
- (2) 延べ件数では2文字のものが最も多く、1文字、4文字、3文字の順になっている。
- (3) 4文字以上の種類は多いが延べ件数に占める割合は少ない。そこで1、2、3文字の漢字列を重点的に分析する必要がある。1文字のものは動詞、形容詞が多いので今後の分析にまかせる。また、4文字以上の漢字列の仮名文字は長くなるので、それにともない同音語は急速に減る。
- (4) 第三表と第四表のKWICを分析すると漢字列の造語成分が2文字、3文字漢字列の中にあることがわかる。“作業予定”を分析すると“作業”と“予定”の2つの概念から出来ていることもわかる。ま

第1表 漢字列調査件数と種類

文字数	種類	延べ件数
1文字	1,230	122,114
2文字	10,095	295,301
3文字	15,690	69,066
4文字	31,932	78,304
5文字	15,681	26,356
合計	74,628	591,141

第2表 漢字長と漢字列の発生件数



第1図 漢字長と漢字列の発生件数

た“作業”や“予定”の使用頻度が高いことがKWIC よりわかる。これは文章の中で“作業予定”という用語を使い別のところでは“作業の予定を”という分ち書きされた書き方や、個別に作業か予定を使うこのため、これら用語の使用頻度が増る。このことからある専門分野の大量の漢字列を収集し、漢字列の中から2文字、3文字の語を集めれば専門分野の造語成分が集められる。

リンク問題は	作業	をノード、作業間の順序関係をアークと	345	るため最高の
タの複雑な	作業上	に効果のあるクリティカルパスなる図型	3	協定のほか、
衛生管理者、	作業主任者	または就業制限の免許および技能講習に	2	以来大規模な
門から日々の	作業予定	が示されて、現場は、この予定の達成を	1	漸情報、特に
行なう場合の	作業予測	を明示して作業手順を示し、ついてに定	1	を考える際、
による代替、	作業交代制	などが主要対策である。	1	、在来技術と
職と労働者の	作業位置	の新しい型の業について記述。労働組織	2	目標志向的、
べ、各機間の	作業体験	に基づく展示品、規格備品目録具などを	2	力の提供する
て、これら	作業例	について説明	1	ては自動車の
ちがいによる	作業評価	、高度の技術と新機械の導入によるコス	1	識別、(2)
技術の進歩が	作業内容	に質的变化を引起している、適当な改良	15	は試作業務、
入れ時および	作業内容変更時	の教育、技能を必要とする作業に従事す	1	

第3表 頭文字順のKWIC

第4表 末尾文字順のKWIC

技術	を研究すべきこと、(3)現場の実情を明	461
自主技術	の開発に努めている	1
被創的自主技術	の開発を図るべく9テーマを取上げ研究	1
軍事技術	の情報の流れを単純なツテルによって固	2
現代技術	の矛盾の特色を明らかにすることは興味	6
近代技術	との中間技術の開発にその焦点を求むべ	3
社会技術	システムで物的生産プロセス、日常業務	3
保安技術	サービスの利用を懸念すべきか、社内に	1
安全技術	とSSVおよびその開発過程にあるES	3
類似技術	の開発、(3)段階的な判別分析技術の	1
製作技術	の調査研究、機械技術に関する技術協力	1

2-5 KWICの分析(同音異義語について)

漢字列の2文字、3文字を分析すれば、ある専門分野の基本的用語が集められることがわかった。しかし、この中にも同音異義語が発生する。この同音異義語については同音異義語の発生したものについてのみ、それらを含んでいる漢字列を抽出、整理し、専門用語ファイルとする。

例えば“イチジ”という読みには“一次”“一時”という用語がある。また“イチジテキ”という読みには“一次的”、“一時的”という用語がある。しかし、これらに接続する用語、発生頻度には異りがある。これを利用すれば同音異義語の判別がつく。これでもなお同音異義語が発生する割合はごくまれである。第5表、第6表からわかるように“一次”、“一時”に接続する漢字列は異っている。

“イチジ”に対しては(一次、一時)、イチジテキに対しては(一時的、一次的)というように専門用語辞書を持ち、複合語を持てば同音異義語の割合は減る。この例でも60件の“イチジ”“イチジテキ”の中で判別を必要とするのは5ヶである。同音異義語が発生してもこのような工夫で同音異義語の選択を減らすことができる。

一時	4
一時点	1
一時的	13
一時的支払不能	1
一時的段階	1
一時的気まぐれ	1
一時的減退	1
一時的翻訳	1
一時的要因	1
一時解雇時訓練	1
一時貯水池	1
一時間	1
計	27

第5表 “一時”の漢字列

一次	10
一次不等式	1
一次以上	1
一次以下	1
一次元座標	1
一次元放射伝熱	1
一次元配列	1
一次処理	1
一次刊行物	1
一次加工	1
一次原料	1
一次外注	1
一次大戦	1
一次式	2
一次微分方程式	1
一次情報	3
一次情報発生源	1
一次投入要素形態別	1
一次文書	1
一次文書中央研究所	1
一次文献	2
一次汚染物	1
一次火災	1
一次的	1
一次的流れ	1
一次結合	1
一次資料	2
一次遅れ	1
一次集計	1
計	33

第6表 “一次”の漢字列

3 専門用語ファイルの作成

3-1 専門用語ファイルの作成プロセス

2の部分で述べた内容にもとづき専門用語ファイルの作成を行ってみる。データは2-2で述べたデータを用いる。データ件数は約60万件の漢字列の中から抽出した。2文字、3文字の漢字列は約2万6千種類ほどである。この中から適切でないものを削除すると約2万3千件ほどの漢字列が得られる。

2文字、3文字の漢字の中には適切でないものが含まれているのでこれらを調査し取り除く。例えば、次のようなものである。戦略毎に、一件当り、一番多い、世界第1位、100万円、か働中、その上多く、2乗和

これらのものは2文字、3文字漢字列の種類のうち1割程度である。さらに、これに読みを付けアイウエオ順に整理する。同音語を含む漢字列を抽出し、整理し読みを付ける。アイウエオ順に分類し、整理する。また2文字、3文字に含まれていなくて長い漢字列を分割することにより得られる、2文字、3文字の漢字列の中にも基本的な語がある。これについても整理しなければならない。このプロセスを図にまとめると第2図のようになる。

3-2 専門用語ファイルの同音異義語の分析

2文字、3文字の漢字列25,785種類の中から不適切なもの3,043種類を取り除き22,742種類の漢字列を得た。

この中で同音異義語が発生している組合せの数を調べてみると次のようになる。(同音語に対して何種類の漢字列があるか)

	2	3	4	5	6	7	8	9	10	計
種類	1,420	277	89	22	17	1	1	0	1	1,828
総件数	2,840	831	356	110	102	7	8	0	10	4,264

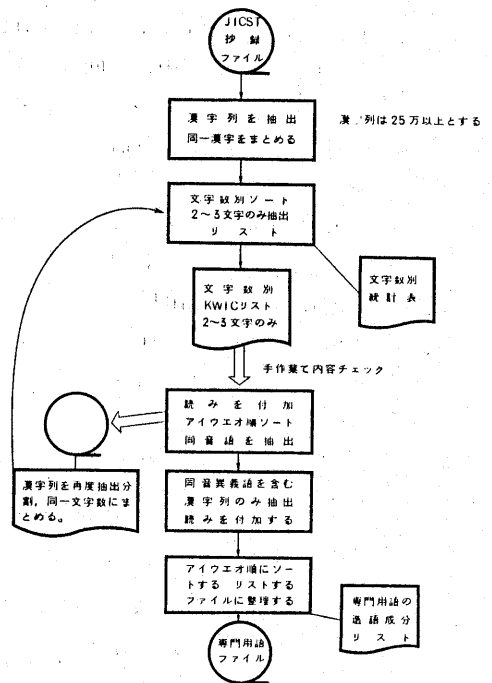
第7表 同音異義語の発生割合

22,742件の中で4,264種類の同音異義語があることがわかる。同音異義語の発生は18.7%である。さらに1つの同音語に対する漢字列の種類は2.33件である。国語辞典の中の同音異義語の発生は(18)

87,216件中29,346件である。この%は33.6%である。

分野を限定することにより同音異義語の発生割合が少ないことがわかる。

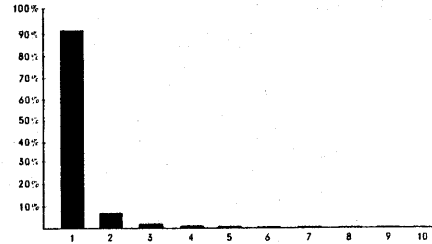
次に同音語の中で頻度の多い語順にならべた時各順位の累積件数と%について調べてみた。これは延漢字列で調べてみた。



第2図 漢字列の処理プロセス

順位	1	2	3	4	5	6	7	8	9	10	合計
件数	122,709	9,774	1,095	248	90	28	5	4	1	1	133,955
%	91.60	7.30	0.82	0.19	0.07	0.02	0.00	0.00	0.00	0.00	100.0

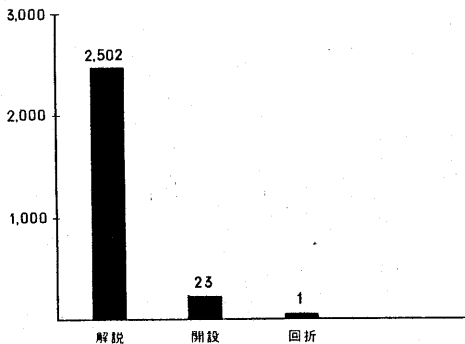
第8表 同音語の頻度の多い語順にならべた各順位の累積件数と%



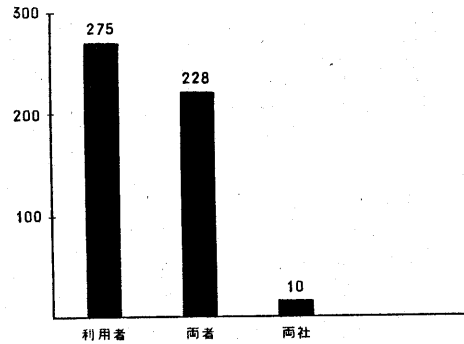
第8表、第3図からある限られた分野をみると総数では、同音異義語の発生には第一位と第二位では大きな差があることがわかる。しかし個別に内容を分析してみると次のような二つの場合にわけることができる。同音語の第1位と第2位頻度に大きな差のあるもの、例えば「カイセツ」のようなものである。(第4図参照)

第3図 同音語の頻度の多い語順にならべた時の各順位の累積グラフ

また、同音語の第1位と第2位頻度に大きな差のないもの、例えば「リョウシヤ」のようなものである。(第5図参照)これらについては今後、十分研究しなければならない。同音異



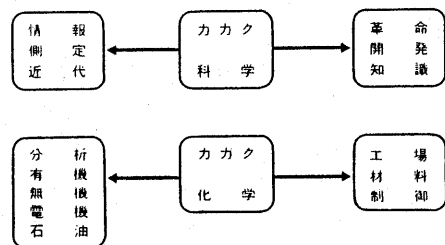
第4図 同音語の第1位と第2位頻度に大きな差のある例



第5図 同音語の第1位と第2位頻度に大きな差のない例

語の前後に接続する用語を集め、それをカテゴリー別にわけると用語接続の性質がわかる。(第6図参照)

頻度分析だけによるシステムでも、このように専門用語をうまく作れば対処できることがわかる。2～3文字の同音異義語のうち第2位以降で選択を必要とするものは約1万1千件程度で約60万漢字列の2%弱である。これから全漢字列に含まれ第2位以降で選択を必要とするものは倍の割合であるとしても約4%弱程度である。



第6図 同音語の語接続

4 この専門用語収集方法の利点

仮名漢字変換用専門用語辞書の作り方について、これまで述べた。ここでは、この方式の利点、今後の発展などについて述べる。

(1) 自動的に作成できる。

この方式の良い点は膨大な資料を分析し、データの中から専門用語ファイルを作り出すことができる。しかも、これらは同音異義語の割合が一般辞書より少い。基礎的用語をみつけたため、応用範囲が広い、自動的に作り出される。

(2) 最適化がはかられている。

専門用語の基礎的用語がデータ分析の過程で作りに出されるため準備すぎがない。システムの最適化がデータ分析過程で自動的に行われている。

(3) 専門的知識を特に必要としない。

膨大な実際のデータを分析するため、一般用語にない特別な基礎的用語をたやすく集めることができる。不適切なデータを削除するだけで、用語について、あまり考えることがない。

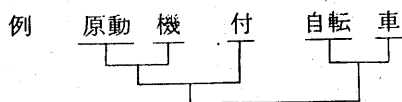
(4) 用語の使用頻度

用語の使用頻度が入っているため辞書を小さくすることも簡単にできる。このため小型機の効率よりシステム作りができ、同音異義語も少い。

これらの利点のほかはこの方法で集めた資料をもとに4文字以上の漢字列を分解したいと考え、研究中である。

(5) 漢字列の分解

長い専門用語を分割し語の結合順序を自動的に見つけ出すことにも利用できる。



また、長い専門用語を文に変化させるために語を自動的に分割するために使用することができる。

例 作業内容 作業の内容 工場用地 工場のための用地
被認定者 認定される者

漢字列を文に変換する研究は九州大学の吉田教授、九州芸術工科大学稲永講師等によって優れた研究がなされている。さらに、大量のデータを集収している。筆者等はこれらの研究を踏まえ新しい方法を見つけようとしている。

5 おわりに

専門用語辞書の充実により仮名漢字変換の同音異義語の減少を達成できた。同音異義語はこの方法で全て解決したわけではなく、さらに分析してゆきたい。同音異義語は語と語の関係を持つことにより減ると言われているがこれは今後の課題である。たとえば アメ(雨, 飴)がフル(降る)この時は雨と降るが関係ある語である。これらは今後に期待していただきたい。この研究を進めるにあたって資料を提供して下さった日本科学技術情報センターの中井浩氏、佐藤雅之氏に深く感謝する。

添付資料として“同音語頻度順同音異義語表”と“同音異義語の第2項以降の頻度の総和が高いものの表”の一部を付け加えた。

〔参考文献〕

- 1) 中井浩, 岡野弘行, 佐藤雅之, 中瀬純夫, 長田孝治, 古賀勝夫, 石川徹也
日本語における語基と構文についてⅠ, Ⅱ, Ⅲ 昭和53年度第19回全国大会講演
論文集 情報処理学会
- 2) 中井浩, 岡野弘行, 長田孝治 Word Base切り出しシステムとそのJ ICST キーワードへ
の適用 第14回情報科学技術研究会発表論文集1977 日本科学技術情報
センター
- 3) 中井 浩 J ICSTにおける自然言語処理(Ⅰ), (Ⅱ) 情報管理1978 VOL20 NO11,
VOL20 NO12
- 4) 野村雅昭 現代漢語の語構成について 情報管理VOL18 NO11 Feb 1976
- 5) FACOM OSIV KUIN(事務処理用語)/JEF解説書 富士通
- 6) 野村雅昭 接辞性字音語基の性格 国立国語研究所報告61「電子計算機による国語研究Ⅱ」
(1978.3)
- 7) 野村雅昭 四字漢語の構造 国立国語研究所報告54「電子計算機による国語研究Ⅶ」
(1975.3)
- 8) 野村雅昭 三字漢語の構造 国立国語研究所報告51「電子計算機による国語研究Ⅵ」
(1974.3)
- 9) 野村雅昭 同字異音 一字音形態素の造語機能の観点から一 井田祝夫博士 功績記念国語学
論集 1979.2
- 10) 野村雅昭 造語法 岩波講座 日本語9「語彙と意味」1977
- 11) 野村雅昭 漢字パターン分類 国立国語研究所報告67「電子計算機による国語研究Ⅹ」
1980.3
- 12) 田中康仁 専門用語の自動抽出 計量国語学会誌 12巻8号
- 13) 田中康仁 漢字列長単位用語の抽出 計量国語学会誌 13巻1号
- 14) 学術用語集 電気工学編 文部省編 コロナ社
- 15) 樺島忠夫 日本語はどう変わるか 一語彙と文字一 岩波新書
- 16) 田中康仁 専門用語の自動抽出 計算言語学25-1 1981.2 情報処理学会
- 17) 牧野 寛 カナ漢字変換 情報処理学会「日本文の入力方式」昭和56年7月2~3日
情報処理学会シンポジウム
- 18) 伊藤均, 亀山正俊, 坂下善彦, 渡辺治, 大川清人 国語辞書のファイル構成上の諸特性
情報処理学会第22回(昭和56年前期)全国大会
- 19) 日立 文字用語辞書ハンドブック
- 20) 田中康仁, 長田孝治 仮名漢字変換システムの専門用語辞書
第23回 プログラミング・シンポジウム報告集 情報処理学会
- 21) 稲永紘之, 吉田将 日本語処理のための機械辞書
情報処理 VOL23 NO2 1982.2
- 22) 牧野 寛 カナ漢字変換入力法 情報処理 VOL23 NO6 1982.6

1 ケイヒ	経費 5997	経皮 1				25 ジカン	時間 773	自館 2
2 ショウカイ	紹介 2973	照会 9	詳解 4	商会 2		26 トクテヨウ	特徴 554	特長 212
3 カイセツ	解説 2502	開設 23	回折 1			27 カイシャ	会社 677	下位者 1
4 ケイサンキ	計算機 2314	計算器 3				28 キカイ	機械 546	機会 106
5 ジョウホウ	情報 2307	乗法 5	上方 4	上法 1		器械 2		
6 ホンコウ	本稿 1527	本項 3	本誌 2	本校 1		29 タイショウ	対象 605	対称 26
7 ヒヨウカ	評価 1447	費用化 5	標価 1	評家 1		対照 15	大正 1	
8 キジユツ	記述 1403	既述 1				30 ユウコフ	有効 637	意向 4
9 カンケイ	関係 1377	環系 8	管型 4	管系 2		友好 3		
10 ケイカク	計画 1365	傾角 2				31 ガイヨウ	概要 638	外洋 2
11 ショウ	使用 1185	仕様 127	試用 3	私用 2		32 テイギ	定義 617	提議 9
12 コウカ	効果 1054	高価 54	硬貨 7			33 コウジョウ	向上 610	工場 1
13 キヨウイク	考課 5	高架 2	硬化 1			高上 1		
14 キノウ	機能 1025	昨日 3	掃納 1			34 ホウコク	方向 607	奉公 1
15 テキョウ	適用 1023	摘要 2				35 ホウシキ	方式 582	法式 1
16 コウセイ	構成 967	公正 20	功勢 6	校正 6		36 ソウカ	増加 580	増価 1
17 セイサン	生産 1006	精算 2	製産 1	清算 1		37 カクリツ	確立 377	確率 182
18 サクセイ	作成 951	作製 56	作制 1	削正 1		38 キジユン	基準 468	規準 83
19 ヒカク	比較 917	皮革 3				39 キョウテヨウ	強調 490	協調 51
20 セイヒン	製品 840	成品 3				40 テイジ	提示 467	呈示 68
21 ニホン	日本 828	二本 3				41 カダイ	課題 520	過大 14
22 ショウライ	将来 777	招来 12				42 ネンカン	年間 517	年刊 2
23 シテキ	指摘 776	私的 6	至適 2	時的 1	史的 1	年鑑 2		
24 サイキン	最近 778	細菌 1				43 リョウシヤ	利用者 275	両者 228
						両社 10		
						44 ジレイ	事例 499	辞令 1
						45 ケイコウ	傾向 494	形鋼 2
						係光 1		
						46 ゲンショウ	減少 364	現象 111
						原償 1		
						47 カガク	科学 252	化学 212
						価額 1		
						48 テンカイ	展開 460	点解 3
						49 カテイ	過程 347	課程 55
						家庭 41		

同音語頻度順同音異義語表

1	リョウシヤ	利用者 275,	両者 228,	両社 10					
2	トクチョウ	特徴 551	特長 215						
3	カガク	科学 252	化学 212	価額 1					
4	カクリツ	確立 377	確率 182						
5	シコウ	思考 88	指向 63	施行 47	試行 34				
		志向 32	至向 1						
6	キカン	機関 176	期間 156	気管 3	器管 3				
		基幹 3	帰還 3	器管 2	季刊 2				
7	セイカク	正確 265	性格 151	精確 6					
8	セイト	精度 151	制度 148						
9	シヨウ	使用 1185	仕様 127	試用 3	私用 2				
10	ヨウセイ	要請 140	養成 117						
11	ジキ	磁気 223	時期 96	時機 11	次期 6				
		自機 2							
12	セイサク	政策 258	製作 90	制作 25					
13	ケンシヨウ	減少 364	現象 111	原債 1					
14	キキ	機器 150	危機 105	危期 3	器機 3				
15	キカイ	機械 546	機会 106	器械 2					
16	キカク	規格 175	企画 98						
17	カテイ	過程 347	課程 55	家庭 41					
18	ジコ	事故 151	自己 96						
19	ヨウイン	要因 346	要員 90						
20	キョウリョク	協力 298	強力 89						
21	キジュン	基準 468	規準 83						
22	センモンカ	専門家 245	専門化 74						
23	テイシ	提示 467	呈示 72						
24	カイトウ	回答 125	解答 72						
25	ココ	個々 164	個個 72						
26	コウカ	効果 1054	高価 54	硬貨 7					
		考課 5	高架 2	硬化 1					
27	シジ	指示 77	支持 69						
28	シカク	視覚 63	資格 52	視角 10	四角 1				
29	ジツシヨウ	実情 124	実状 63						
30	ヨウイ	容易 261	用意 62						
31	ジュウファン	十分 293	充分 60						
32	ケイキ	景気 34	計器 27	契機 26	掲記 3				
		型機 3							
33	サクセイ	作成 951	作製 56	作制 1	削正 1				
34	セイキ	世紀 40	正規 29	生起 23	精微 4				
35	ソウイ	相違 124	相異 39	創意 11	総意 2				
36	ソクト	速度 92	測度 52						
37	イツカン	一環 72	一貫 51						
38	キョウチヨウ	強調 490	協調 51						
39	コウセイ	構成 967	公正 20	功勢 6					
		校正 6	鋼製 6	更生 2	更正 2				
		矯正 1	後正 1						
40	イシヨウ	以上 293	異常 31	委譲 13	移譲 4				
		異状 1							
41	ハンエイ	反映 105	繁栄 45						
42	ホンヨウ	保証 166	補償 24	保障 21					
43	フカ	付加 40	負荷 35	不可 4	賦課 3				
		府下 1	付価 1						
44	カイホウ	解法 259	解放 21	開放 20	会報 2				
45	ジヨウキヨウ	状況 352	情況 42						

同音異義語の第2項以降の頻度の総和が高いものの表