漢字に関するデータ分析

目次

１はじめに

このレポートに取り組むにあたり、「文化庁　漢字頻度分布調査」という興味深い調査結果を見つけた。

この調査を見て、「小学校１、２年で習う漢字や画数の少ない漢字といった、いわゆる『簡単な漢字』は使用頻度が高いのか」や「辞書に載っている熟語の数が多ければ、使用頻度も高いのか」「文章の種類によって頻度分布にどれだけ差があるのか」などの疑問が生じた。

そこで、頻度分布の他にも漢検の級、画数、辞書の熟語数など計24のパラメータを用意し、hiddenを用いて可視化を行った。

２使用したデータ

・文化庁　漢字出現頻度調査

https://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/nihongokyoiku_hyojun_wg/04/pdf/91934501_08.pdf

・漢字辞典ONLINE

https://kanji.jitenon.jp/

・地名辞典ONLINE

https://chimei.jitenon.jp/

・国立国会図書館サーチ

https://ndlsearch.ndl.go.jp/

解析対象

- 漢字出現頻度調査（３）、凸版（書籍）の上位100個の漢字とした。以下、1位から順番に列挙する。

人、一、日、大、年、出、本、中、子、見、国、言、上、分、生
手、自、行、者、二、間、事、思、時、気、会、十、家、女、三
前、的、方、入、小、地、合、後、目、長、代、私、下、立、部
学、物、月、田、何、来、彼、話、体、動、社、知、理、山、内
同、心、発、高、実、作、当、新、世、今、書、度、明、五、戦
力、名、金、性、対、意、用、男、主、通、関、文、屋、感、郎
業、定、政、持、道、外、取、所

出現順位：解析対象漢字100個について、３データ内容（3/4,4/4）に記載の各パラメータに出現する順位

3 データ内容（1/4）

パラメタ名	内容	ソース
漢検の級		漢字辞典ONLINE
画数		漢字辞典ONLINE
熟語数	「（解析対象の漢字）を含む熟語の一覧」の要素数	漢字辞典ONLINE
熟語数（語頭）	「（解析対象の漢字）から始まる熟語の一覧」の要素数	漢字辞典ONLINE
熟語数（語尾）	「（解析対象の漢字）で終わる熟語の一覧」の要素数	漢字辞典ONLINE
四字熟語	「（解析対象の漢字）を含む四字熟語の一覧」の要素数	漢字辞典ONLINE
四字熟語（語頭）	「（解析対象の漢字）から始まる四時熟語の一覧」の要素数	漢字辞典ONLINE
四字熟語（語尾）	「（解析対象の漢字）で終わる四字熟語の一覧」の要素数	漢字辞典ONLINE

3 データ内容（2/4）

パラメタ名	内容	ソース
ことわざ	「（解析対象の漢字）を含むことわざの一覧」の要素数	漢字辞典ONLINE
ことわざ（語頭）	「（解析対象の漢字）から始まることわざの一覧」の要素数	漢字辞典ONLINE
ことわざ（語尾）	「（解析対象の漢字）で終わることわざの一覧」の要素数	漢字辞典ONLINE
市区町村名	「（解析対象の漢字）を含む市区町村名」のヒット数	地名辞典ONLINE
地名	「（解析対象の漢字）を含む地名」のヒット数	地名辞典ONLINE
名字	「（解析対象の漢字）を含む名字」のヒット数	漢字辞典ONLINE
名前（全体）	「（解析対象の漢字）が付く名前一覧（赤ちゃんの命名・名付け）」の総ヒット数	漢字辞典ONLINE
名前（男性）	「（解析対象の漢字）が付く名前一覧（赤ちゃんの命名・名付け）・男性」のヒット数	漢字辞典ONLINE
名前（女性）	「（解析対象）が付く名前一覧（赤ちゃんの命名・名付け）・女性」のヒット数	漢字辞典ONLINE

３データ内容（3/4）

パラメタ名	内容	ソース
国会図書館	国立国会図書館サーチで、（解析対象の漢字）で検索した時のヒット数	国立国会図書館サーチ
書籍タイトル	国立国会図書館サーチで、タイトルの欄に（解析対象の漢字）を入れて検索した時のヒット数	国立国会図書館サーチ
凸版（書籍）	凸版印刷が平成１６年，１７年，１８年に作成した組版データをもとに、「教科書」以外の ４分野の調査対象漢字数の比がおおよそ，「辞典・辞典類」「単行本」「週刊誌」「月刊誌」の順に，「１：３：１：１」の程度になるように振り分けたものが対象	漢字出現頻度数調査（３）（平成19年3月、文化庁）
朝日	朝日、読売の朝刊及び夕刊紙面 平成18年10月,11月の約367万字が対象	漢字出現頻度数調査（新聞）（平成19年、文化庁）
読売	上に同じ	上に同じ

3 データ内容（4/4）

パラメタ名	内容	ソース
ウェブ	ネット上の約13億9100万文字が対象	漢字出現頻度数調査（ウェブサイト）（平成19年、文化庁）
凸版（教科書）	*小学校用・中学校用・高等学校用の教科書が対象	漢字出現頻度数調査第2部（平成19年、文化庁）

*“情報化時代”に追いつけるか？　審議が進む「新常用漢字表（仮）」を参照した（2024年2月13日閲覧）。https://internet.watch.impress.co.jp/cda/jouyou/2008/06/20/20005.html

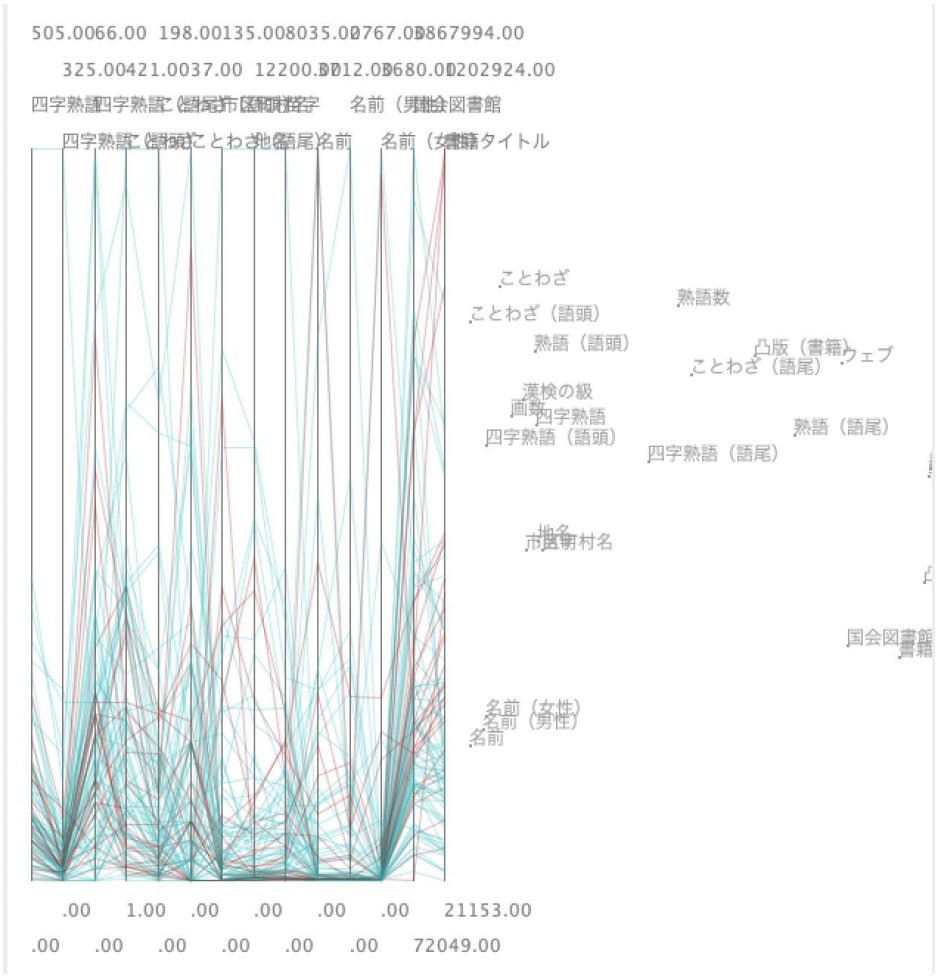
使用したcsvファイル（一部）

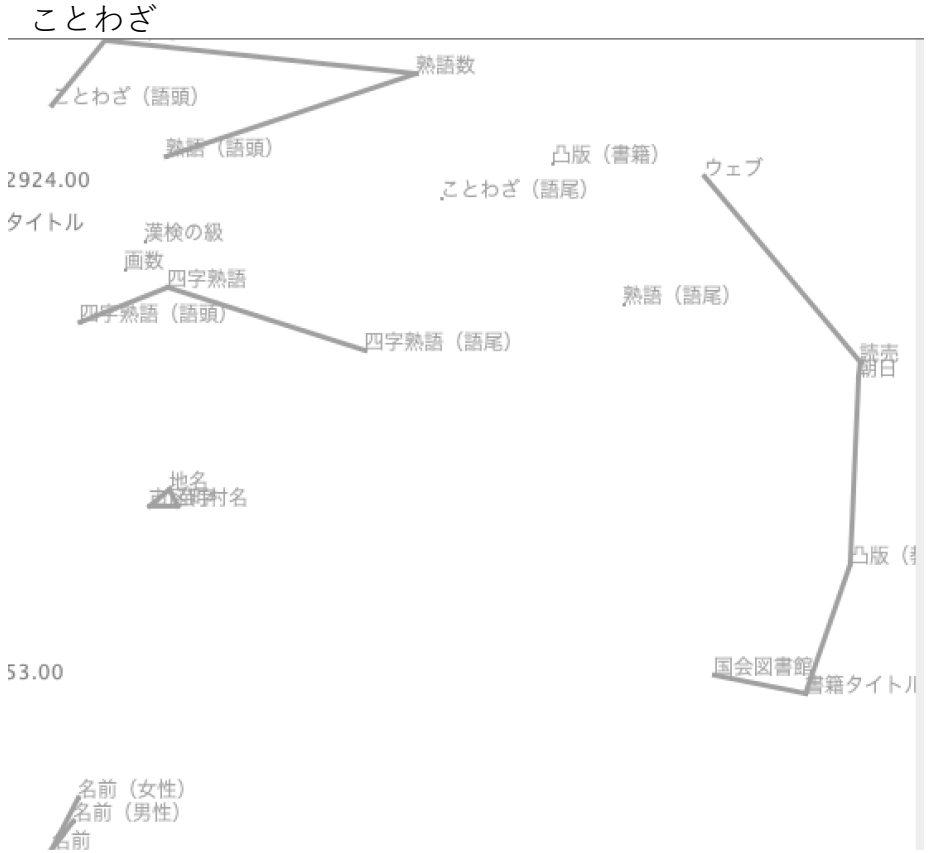
			Category	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
			数字	漢検の級	画数	熟語数	熟語（語頭）	熟語（語尾）	四字熟語	四字熟語（語	四字熟語（語尾	ことわざ	ことわざ（語	ことわざ（語	市区町村名	地名	苗字	名前
1	人	常用	0	10	2	635	207	348	188	58	51	401	129	37	1	134	81	8
2	一	常用	1	10	1	673	581	40	505	325	7	421	198	1	4	1214	704	10
3	日	常用	0	10	4	396	138	223	95	15	28	157	17	17	32	1967	817	2
4	大	常用	0	10	3	636	520	66	166	79	16	167	89	1	80	7229	2226	9
5	年	常用	0	10	6	257	104	127	47	10	15	93	25	14	0	57	71	
6	出	常用	0	10	5	682	246	94	41	16	12	172	25	1	8	905	929	
7	本	常用	0	10	5	318	169	113	28	11	4	27	7	4	18	3780	2002	
8	中	常用	0	10	4	323	173	132	117	12	19	78	5	6	32	5958	1756	
9	子	常用	0	10	3	599	51	492	77	8	27	246	28	23	17	1014	976	37
10	見	常用	0	10	7	376	236	102	27	6	8	195	37	1	15	1428	1139	
11	国	常用	0	9	8	302	165	119	46	7	15	30	3	0	16	666	510	1
12	言	常用	0	9	7	412	213	172	143	12	39	163	45	10	0	4	20	
13	上	常用	0	10	3	646	228	116	62	14	11	173	29	7	41	6049	3032	
14	分	常用	0	9	4	392	131	117	37	5	11	62	5	11	2	351	185	
15	生	常用	0	10	5	441	257	123	127	28	37	151	43	11	10	1391	1509	16
16	手	常用	0	10	4	519	120	244	46	12	7	271	139	14	8	737	519	
17	自	常用	0	9	6	167	151	4	111	56	0	41	17	0	0	41	13	
18	行	常用	0	9	6	395	170	163	127	10	65	66	7	2	3	147	322	3
19	者	常用	0	8	8	240	1	236	42	0	2	151	0	7	0	98	33	
20	二	常用	1	10	2	166	134	5	36	15	8	101	35	0	3	1095	611	7
21	間	常用	0	9	12	248	81	134	32	2	15	66	11	7	10	828	1067	

４ 予想

- 各頻度分布ランキングの間には正の相関がある
- 地名と名字には相関関係がある
- 熟語数と使用頻度には正の相関がある（ランキングとは負の相関）
- 男性の名前と女性の名前に使われる漢字には負の相関がある
- 簡単な漢字（漢検の級が下の漢字、画数の少ない漢字）は多く使われる傾向にある

5 結果と考察

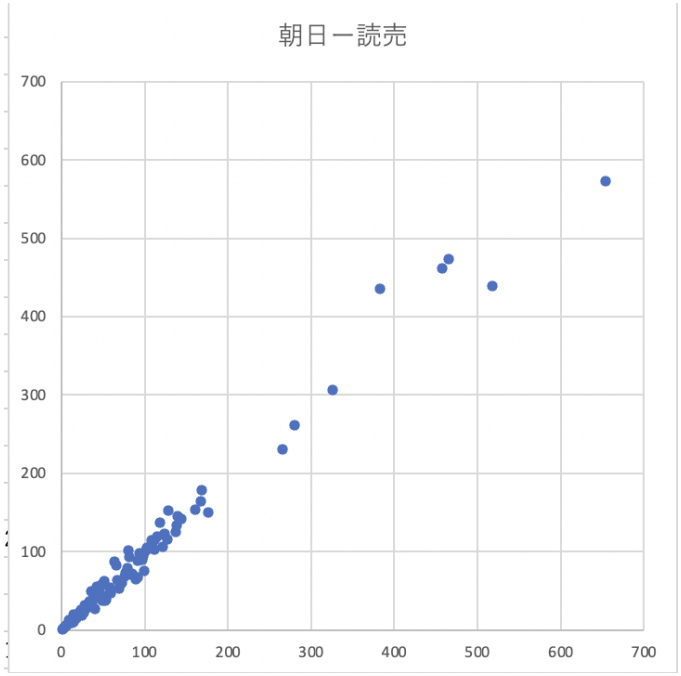




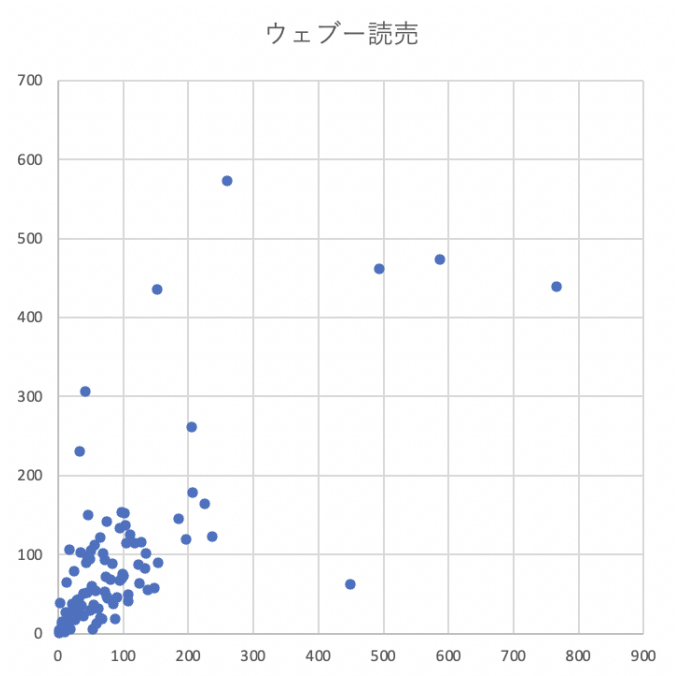
左：クラスタ数を2として、スライダーを動かさずに描画したもの。
右：スライダーを右に動かし、関係性がある要素を抽出したもの。

出現順位同士の相関

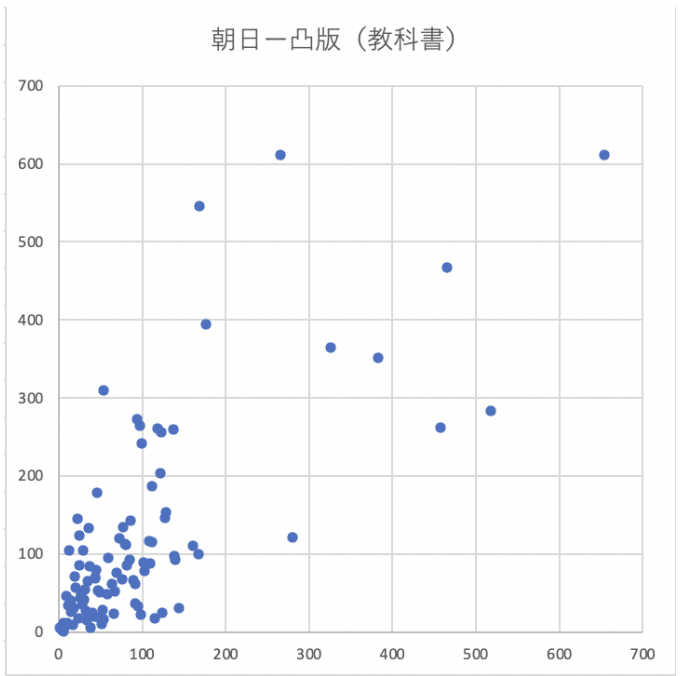
- Hiddenによる分析結果からは、「朝日-読売」間の相関が最も大きいが、「ウェブと読売」「朝日と凸版（教科書）」にも関連があることがわかる。
- 散布図と相関係数からも、「朝日-読売」間に非常に強い正の相関、「ウェブと読売」「朝日と凸版（教科書）」にも正の相関があることが見て取れる。



相関係数　0.991



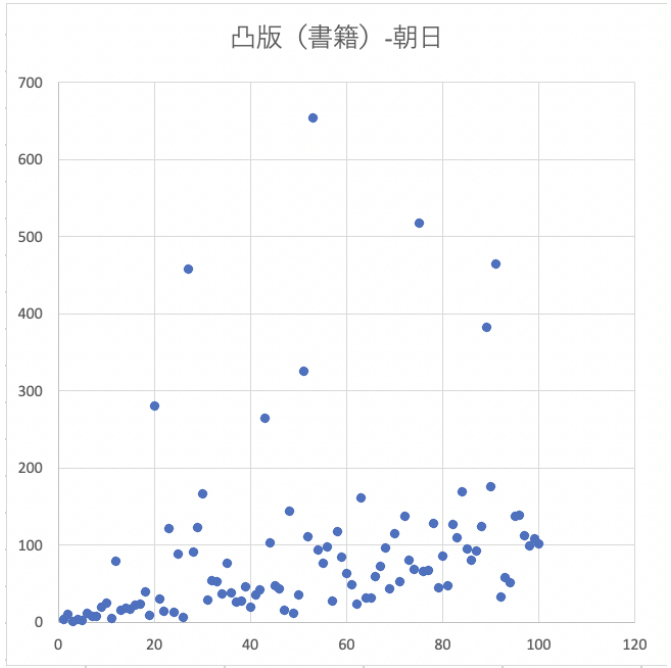
相関係数 0.715



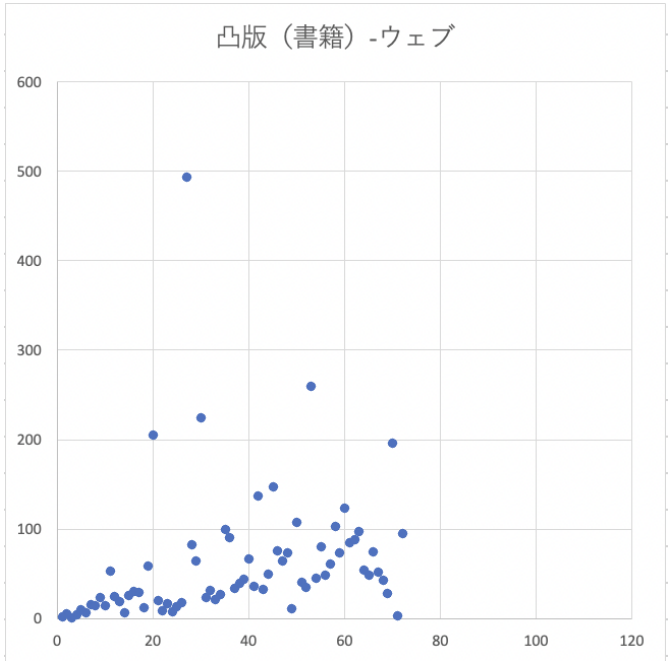
相関係数0.737

凸版（書籍）と各パラメータの関連

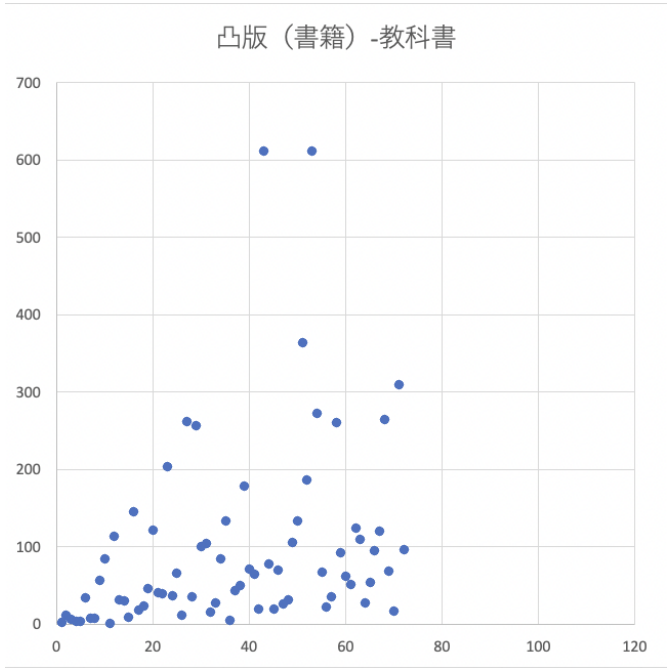
最初の予想とは異なり、「凸版（書籍）」と新聞、ウェブ、教科書に含まれる漢字の頻度分布には強い相関がみられなかった。



相関係数 0.303



相関係数 0.384



相関係数 0.304

ランキングの相関に関する考察

- **会社が異なっても、新聞に使われている漢字の頻度分布はほぼ同一である。**

⇨同一の内容を扱っていることが多いため。両者とも1位が「日」となっている等、新聞に特有の傾向が見られる。

- **新聞に登場する漢字の頻度分布は、教科書やウェブに使われている漢字の頻度分布と相関がある。**

⇨新聞と教科書に使われている文体が似ているためではないか？（どちらも堅い・正式な文体）

また、WEBもニュースを扱うものが一定割合存在するため傾向が似たのではないかと推測

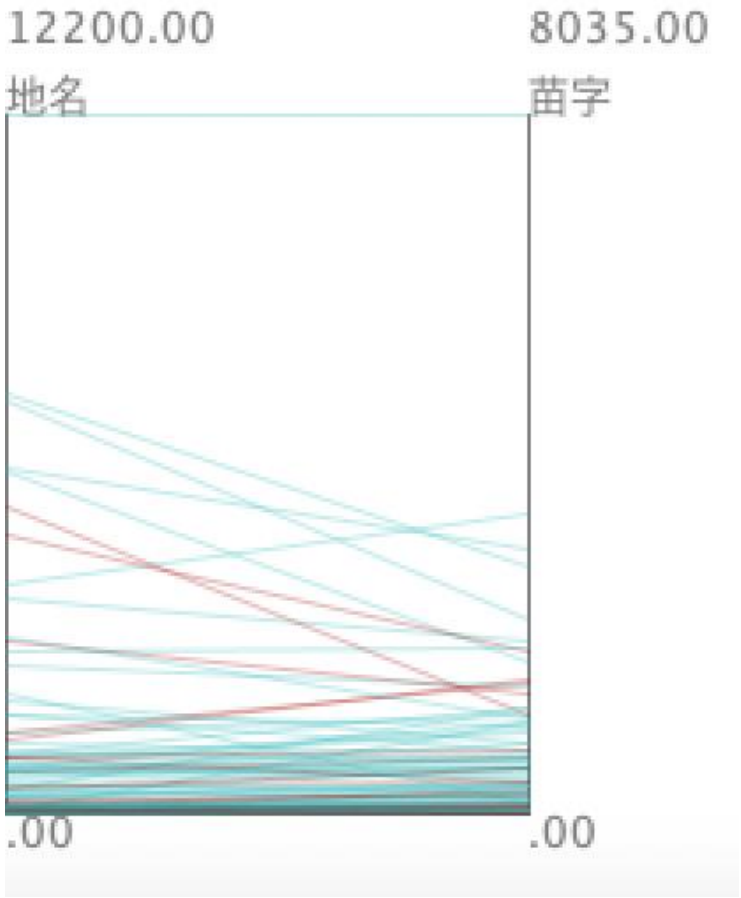
- **書籍に登場する漢字と新聞・ウェブ・教科書に使われている漢字には正の相関があるものの弱い**

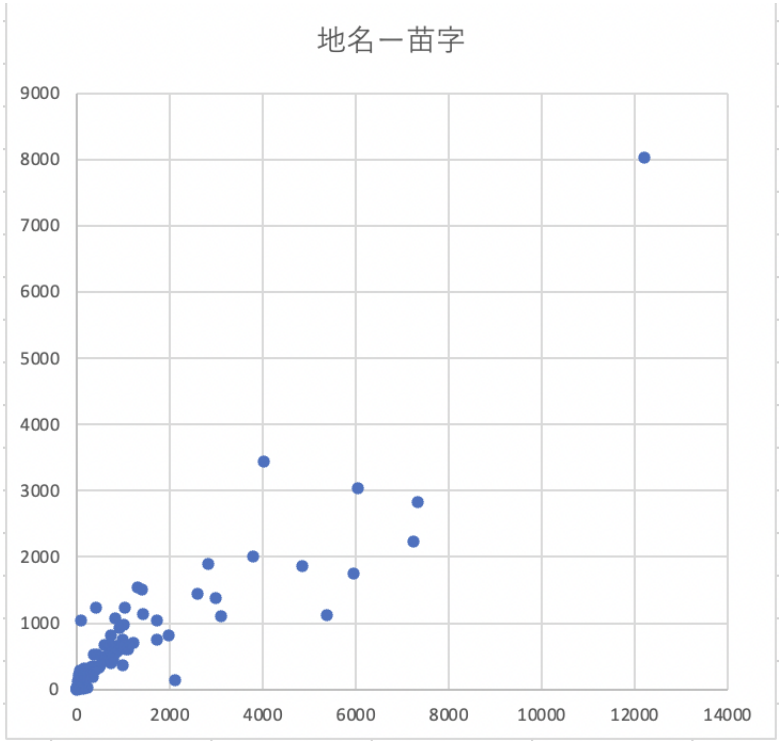
⇨書籍には物語のようなフィクションが多く含まれるため、新聞と傾向が異なったのではないか？

また、書籍（特に物語文）にはいわゆる「話し言葉」が多く登場するが、新聞にはほとんど登場しないなどの文体の違いが原因とも推測できる。

地名と姓名の関連

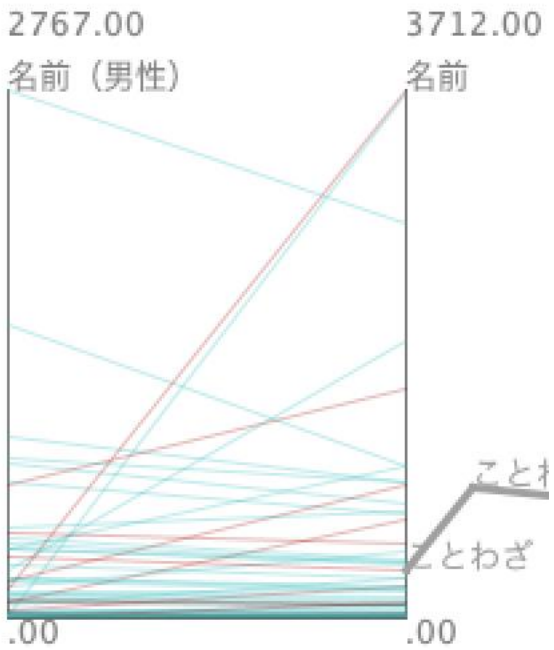
画面左のスライダーを右に動かすと、「地名-名字」のグラフが表示された。
Hiddenのグラフと散布図、相関係数より、この2要素には強い正の相関があることがわかった。

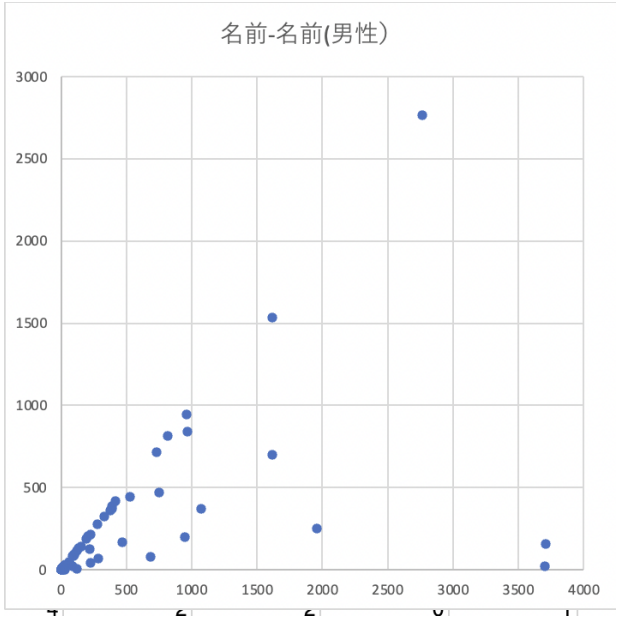




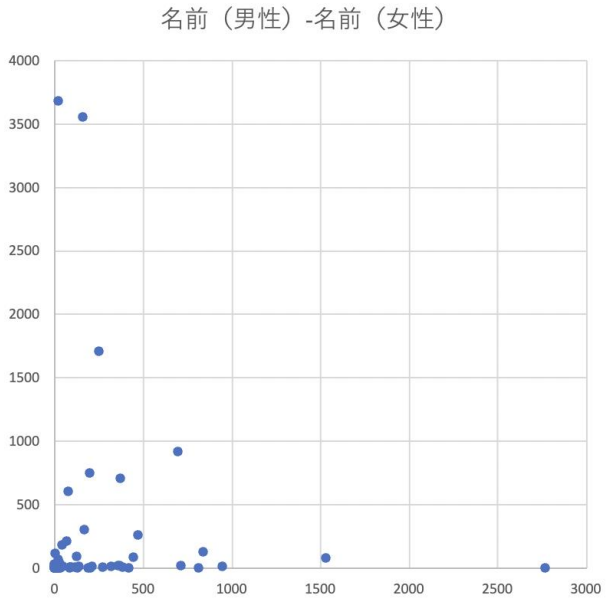
相関係数 0.901

名前の関連性

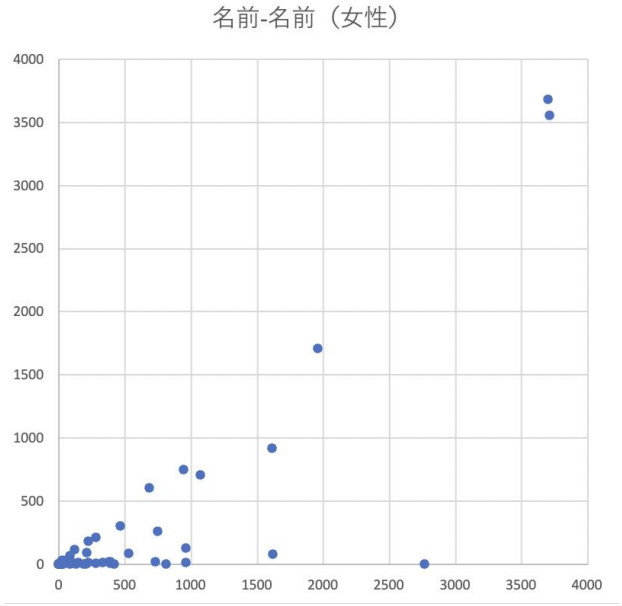




相関係数　0.545



相関係数 0.035



相関係数　0.853

さらに左のスライダーを右に動かすと、「名前（全体）」と「名前（男性）」の相関関係が示された。
「名前（全体）」と「名前-（男性）」は従属関係にあるため正の相関が見られるが、外れ値の影響で相関係数は低く出ている。
「名前（男性）」と「名前（女性）」の間には相関関係が見られなかった。

地名と姓名についての考察

・地名に多く使われている漢字は、名字にも多く使われている。

⇨地名・名字共に地理的な特徴（山、田、川など）からつけることが多いこと、明治時代に一般市民も名字を持てるようになった際、地名をそのまま名字にした人が多かったことが原因と推測される。

・「名前（全体）」は、他のどの要素とも明確な関連性がない。

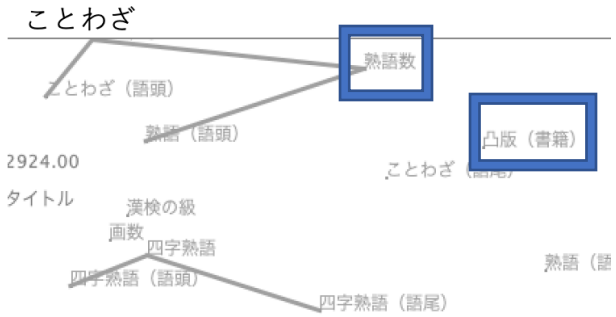
⇨名前に使われる漢字は意味や響きが重視されるため、通常の文章に登場するからといって名前に多く登場するとは限らない。

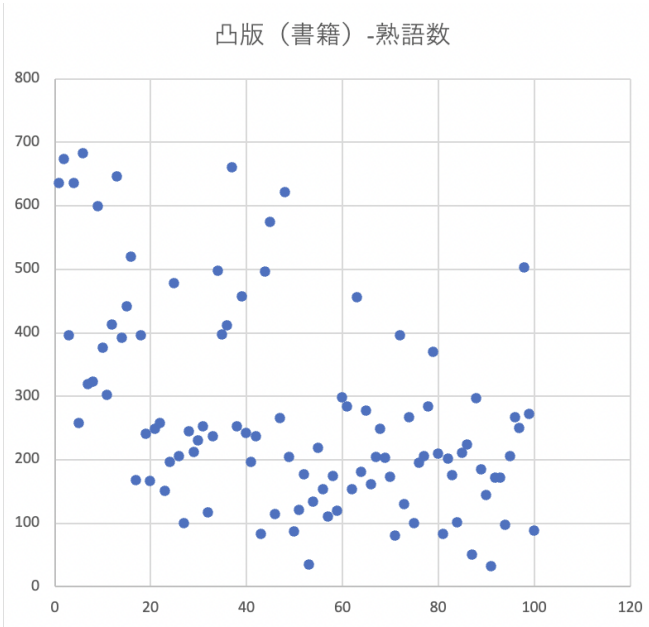
・「名前（男性）」と「名前（女性）」には明確な相関が得られなかった。

⇨分析前は負の相関があると予想していたが、明確な相関なし

値が極端に小さい要素（男性・女性を問わずほとんど使われていない漢字）が多かったことが原因？

熟語数と凸版（書籍）の関連性



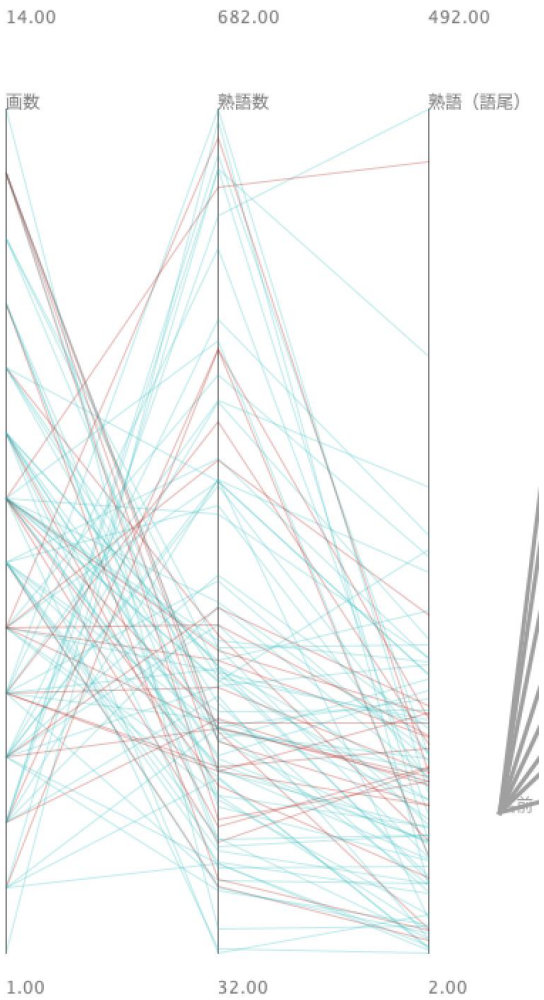


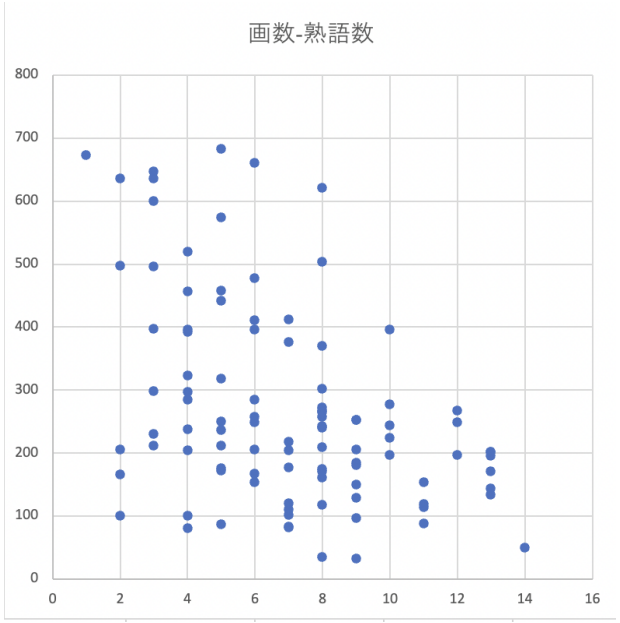
相関係数 -0.48

辞書に載っている熟語の数と書籍の登場頻度は、弱い負の相関があるものの、予想していたほど強い相関は見られなかった。

⇨**漢字の文章への登場頻度は、単に「辞書に載っている言葉の数」では決まらない。**
動詞として多く使われる（「見る」「行く」など）、熟語を形成しないが文章の中でよく使われる（◯月◯日の「月」「日」など）など、複合的な要素の影響を受けていると推測される。

画数と熟語数の関係





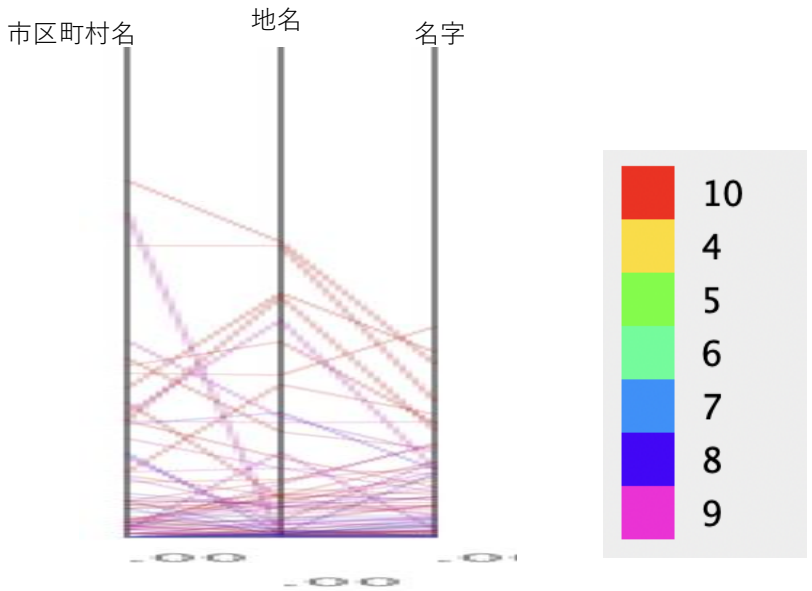
相関係数 -0.443

2つのスライダーを調整すると、「画数-熟語数」のグラフが表示された。
グラフ、散布図、相関係数より、弱い負の相関があることがわかる。

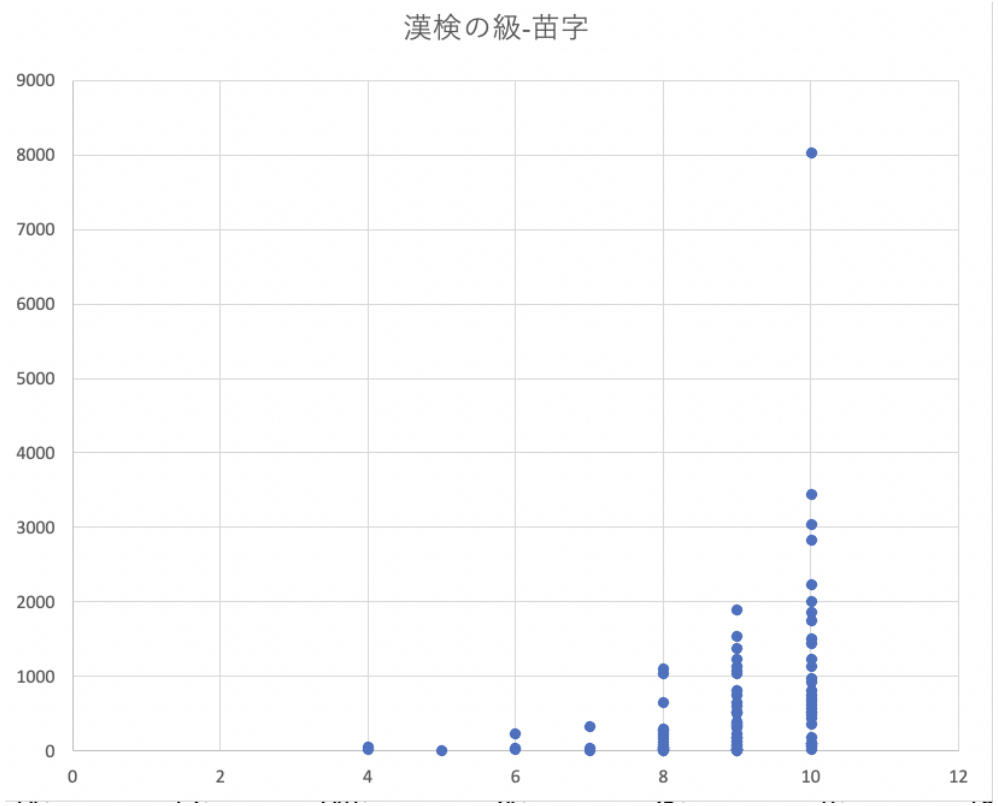
これは、「画数が少ない、書きやすい漢字ほど熟語に多く使われている」ことを示している。

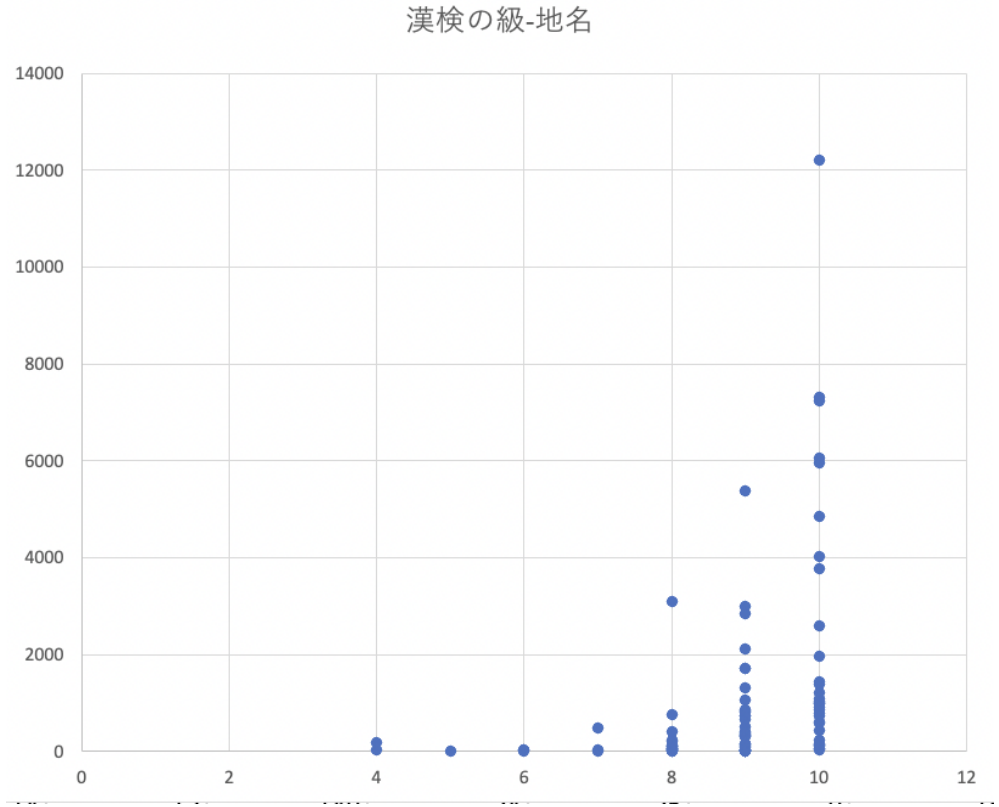
漢検の級による分類

今まで数値として処理していた漢検の級をカテゴリとして、漢検の級ごとの傾向を掴もうとした。
スライダーを左に動かすと、先ほど得られた「地名と名字の関係」が表示された。
赤が10級、ピンクが9級の漢字を表している。これら2つの級の感じが上位にきていることがわかる。
このことから、名字や地名には平易な漢字が使われることが多いと推測できる。



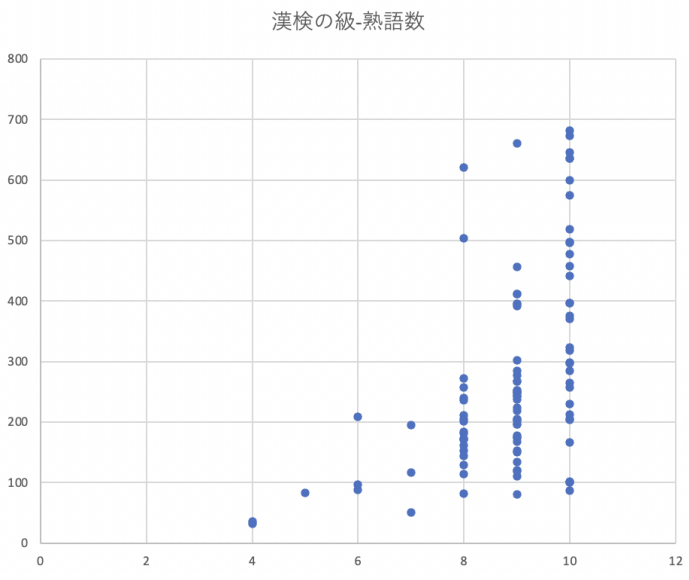
漢検の級による分類

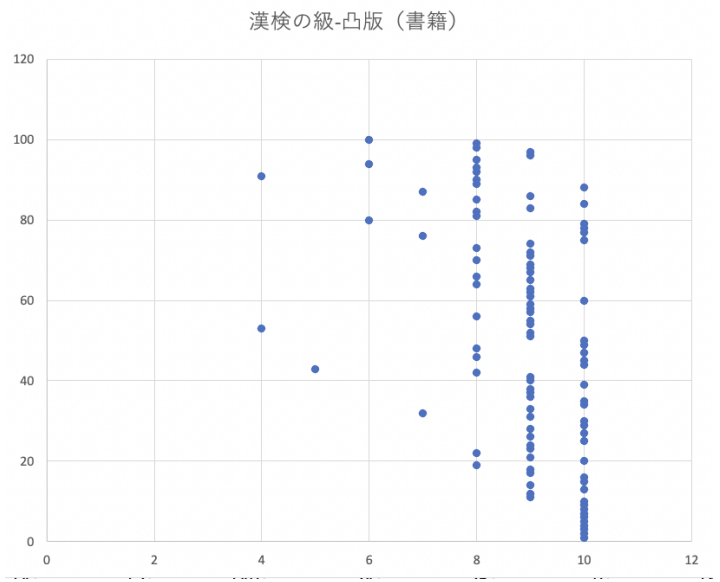


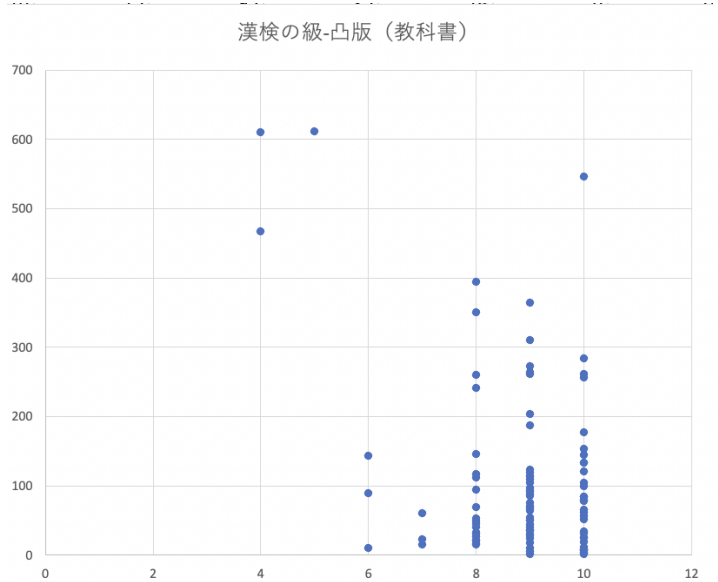


散布図からも、名字や地名に多く使われている漢字は、漢検の級が下である傾向がわかる。
⇨名字や地名に多く登場する漢字は、小学校1,2年生でも書けるものが多い。

漢検の級による分類

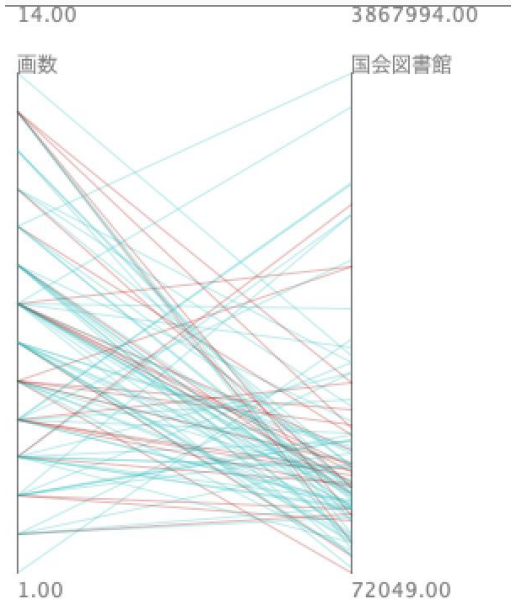


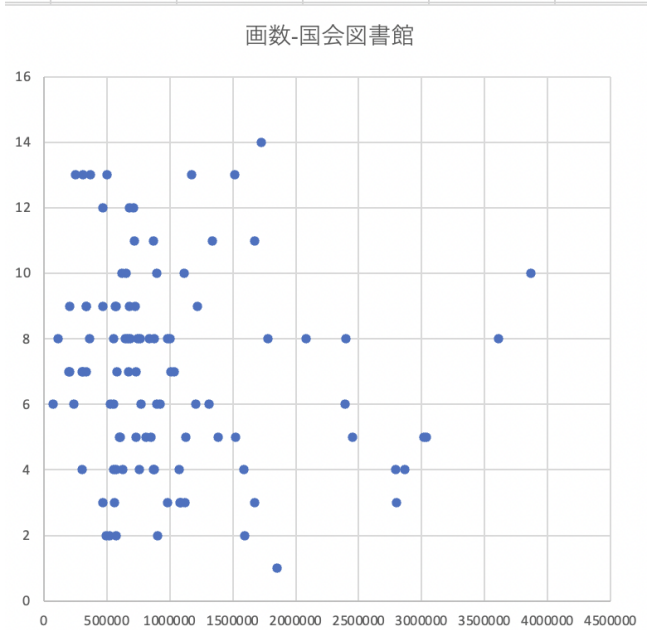




漢検の級と熟語数、書籍への登場頻度、教科書への頻度には相関があると推測していたが、最初の可視化では関連が弱いように見えた。
散布図からは、漢検の級と熟語数にやや正の相関があるとがわかる。
また、教科書は書籍と比べて、級が下の漢字がランキング上位に固まっている。これは、教科書（特に小学校のもの）は、意図的にその学年で習う漢字を取り入れるからだと推測できる。
これらの相関が可視化されなかったのは、級が離散的な値であるためであると考えられる。

画数-国会図書館検索ヒット数





画数と国会図書館の検索ヒット数の関係は、右の図では遠かったが、左のグラフでは表示された。
しかし、散布図を見ても明確な相関があるとはいえない。
⇨**スライダーの位置によっては、相関関係の弱い要素が抽出されてしまうこともある**
（画数が離散的な値であるため、誤差が大きくなった？）

6 全体の考察

・頻度分布に対する予想はおおむね正しかったが、思ったより相関関係が見られないペアもあった。

⇨文章の種類によって、漢字の頻度分布にはばらつきがある

・「簡単な漢字の方が多く使われる」という予測も、「熟語数」や「地名」に対してはおおむね正しかったが、**文章への登場頻度ととはあまり相関がなかった。**

⇨簡単な漢字だからといって、文章で多く使われているとは限らない。

ただし、今回は書籍に使われている漢字の上位100件を解析に用いたため、**ほとんどの漢字が小学校1~3年で習うものだった（漢字間の難易度の差があまりなかった）。**

サンプル数を増やした場合、頻度分布との関連性も見られるのではないかと推測