認知研究のための漢字頻度基準表の作成

野崎浩成（名市大）・横山詔一（国語研）・米田純子（国語研）

１．はじめに

文字言語の認知実験を行う場合、文字の構造・複雑さ、単語の意味・親密度など、刺激材料の属性を統制する必要がある。たとえば、記憶実験では、刺激材料の使用頻度が成績に多大な影響を及ぼすことがよく知られており、使用頻度に基づいて刺激材料を選定するのが普通である。

文字、特に漢字の使用頻度については、国立国語研究所が公刊した『現代新聞の漢字』（1976）が我が国の標準的資料として諸学界に貢献してきたことは周知の事実と言えよう。この報告書はサンプリング調査の手法を用いて漢字の使用率を推定しているため、統計的推定の精度を考慮して頻度９以下の漢字は示されていない。しかし、失語症研究者を含む幅広い領域の認知研究者からは、頻度１までのデータも参照したいとの声が聞かれる。また、この調査の対象となった新聞が昭和 41 年発行のものであるため、最近の新聞との異同を明らかにする必要もある。

そこで、1993 年版の新聞記事全文データベースを用いて文字の使用頻度を計数し、認知研究一般の便に供するデータベースの作成を試みる。

２．文字使用頻度基準表の作成

方　法

材　料

分析に用いたテキストデータは、1993 年の１月１日から同年 12 月 31 日までに発行された朝日新聞（朝・夕刊）の電子化された記事であった。このテキストデータの大部分は、CD-HIASK'93（朝日新聞社・紀伊国屋書店・日外アソシエーツ、1994）という CD-ROM フルテキストデータベースからダウンロードした約 11 万件の記事から成る（ただし、CD-ROM に格納されていない記事も 114 件手で入力してデータに加えた）。

このテキストデータから、記事見出し部分を削除して入力データとした。その理由は、「実際の新聞紙面」と「電子化されたテキストデータ」の内容を比較照合すると、見出し部に不一致が多く見られたため、見出し部分を分析の対象とするのは妥当でないと考えたからである。

最終的にテキストデータとして入力した文字数の総計は約 5,500 万文字に達し、これまでに日本でなされた文字使用頻度調査のサンプルとしては最大級の規模となった。

手続き

上記のテキストデータに対して、漢字、平仮名および片仮名について文字使用頻度を計数した。漢字は、JIS X0208 に基づいて、区点コード表の第 16～47 区分に属する第１水準漢字集合、第 48～84 区分に属する第 2 水準漢字集合、両者合わせて 6353 文字とした。同様に、平仮名は区点コード 0401－0483 に属する 83 文字、片仮名は区点コード 0501－0586 に属する 86 文字とした。

次に、異体字のチェックを行い、使用頻度が上位 1,000 位までの漢字について国語研調査（1966）の結果と比較・検討した。

結　果

①漢字の延べ字数は 2340 万 8,236 字、異なりで 4476 字である（表１）。また、平仮名は延べ 2071 万 1,361 字、異なりで 83 字であり、片仮名は延べ 360 万 8,288 字で異なりで 86 字であった。漢字使用率はおよそ 42%である。

②漢字は使用頻度の上位 1,000 字で累積使用率が約 95%に達する。さらに、上位 1,600 字で全体のほぼ 99%に達し、残りの約 3,000 字は 1%程度にすぎない（図１）。

表1．漢字の使用頻度表（高使用頻度上位１０字）

順位	漢字	区点コード	使用頻度	使用率 (‰)	累積使用率 (‰)
1	日	3892	336465	14.374	14.374
2	一	1676	285089	12.179	26.553
3	十	2929	254534	10.874	37.427
4	二	3883	223075	9.530	46.957
5	人	3145	218967	9.354	56.311
6	大	3471	218693	9.343	65.654
7	年	3915	216931	9.267	74.921
8	会	1881	214989	9.184	84.105
9	国	2581	199502	8.523	92.628
10	三	2716	173495	7.412	100.040

また、時代的変化をとらえるために、本調査と国語研究所（1966）の調査結果を比較したところ、次の③～⑤が明らかになった。

③両者の文字使用頻度について、積率相関を算出したところ、 .97（素データ）および .92（対数変換後）であった（図２に散布図）。

④両者の累積使用頻度分布を適合度の χ^2 検定によって検討した。その結果、度数の偏りは有意ではなく($\chi^2(9) = 7.45, .70 > \rho > .50$)、両調査の分布型には差がないことが明らかになった。

⑤ただし、使用頻度の時代的変動が大きい漢字も存在する。国語研の調査では使用率が 1,000 位以下であったが今回は 1,000 位以内に入った漢字は 78 字存在し、使用率が約 20 倍に増大した漢字（狙）さえある（表2、3）。

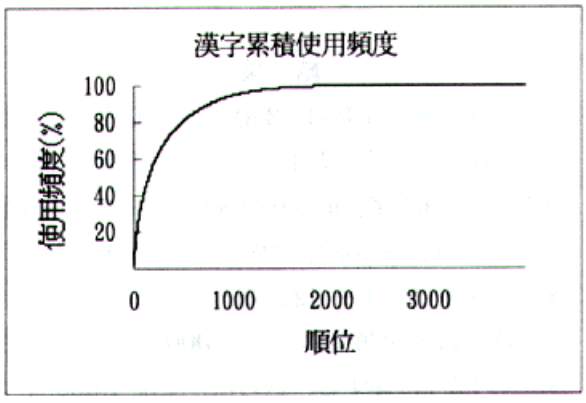

図１．漢字の累積使用頻度分布

表２．漢字使用頻度が著しく増加した漢字

’66年「順位」≦1000, かつ,’93年「順位」≦1000 そのうち、順位変動が大きい上位6字

		国語研の調査('66)		本研究の調査('93)	
漢字	区点コード	順位	使用率 (‰)	順位	使用率 (‰)
狙	3332	1983	0.010	854	0.205
崩	4288	1722	0.024	771	0.245
訟	3057	1687	0.027	967	0.161
阜	4176	1663	0.029	968	0.160
削	2679	1460	0.052	855	0.204
葬	3382	1448	0.053	892	0.189

表３．漢字使用頻度が著しく減少した漢字

’66年「順位」≦1000, かつ,’93年「順位」≧1000 そのうち、順位変動が大きい上位6字

		国語研の調査('66)		本研究の調査('93)	
漢字	区点コード	順位	使用率 (‰)	順位	使用率 (‰)
鍵	3091	734	0.291	1907	0.013
才	2645	625	0.356	1565	0.033
曇	3862	948	0.169	1871	0.014
胃	1663	893	0.190	1739	0.021
綿	4442	850	0.216	1582	0.033
糸	2769	687	0.314	1340	0.063

今後の進め方

本研究で作成された漢字頻度基準表の公開をインターネット上で進め、刺激材料の選定に有用となるデータを提供したい。

＜謝辞＞　本研究の遂行において、Eric Long 氏から多大なご協力とご助言をいただきました。

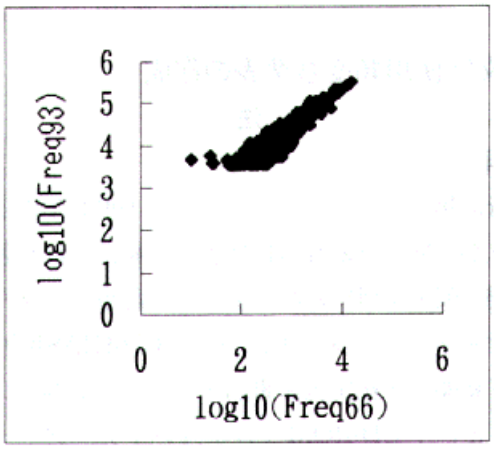

図２．対数変換後の散布図