「言語における系統・変異・多様性とその数理」シンポジウム

異体字選好の生態学的モデル

横山詔一（国立国語研究所）
yokoyama@ninjal.ac.jp

日時：2018年2月2日 (金) 13:20〜14:00
会場：TKP東京駅大手町カンファレンスセンター ホール22E

（質問）
パソコンで
字を打つとき
どちらを
選びますか？
例：桧と檜

異体字選好実験
笹原・横山（1996）が
256ペアを用いて
調査したのが最初

01	亜	亞
	唖	啞
	壷	壺
02	媛	媛
	淫	淫
	秤	秤
03	陥	陷
	焔	焰
04	奥	奧
	襖	襖
09	葛	葛
	喝	喝
10	観	觀
	潅	灌
11	燗	爛
	澗	澗
12	徽	徽
13	侠	俠
	狭	狹
	頬	頰

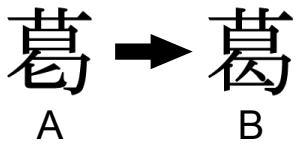
異体字ペアを見て旧字体を選んだ人の割合
（右は256ペアの一部）

同じ人に同じ調査を半年後に実施して一致度をみた再テスト法の結果によると，データの信頼性はかなり高い
（横山・笹原・當山，2006）

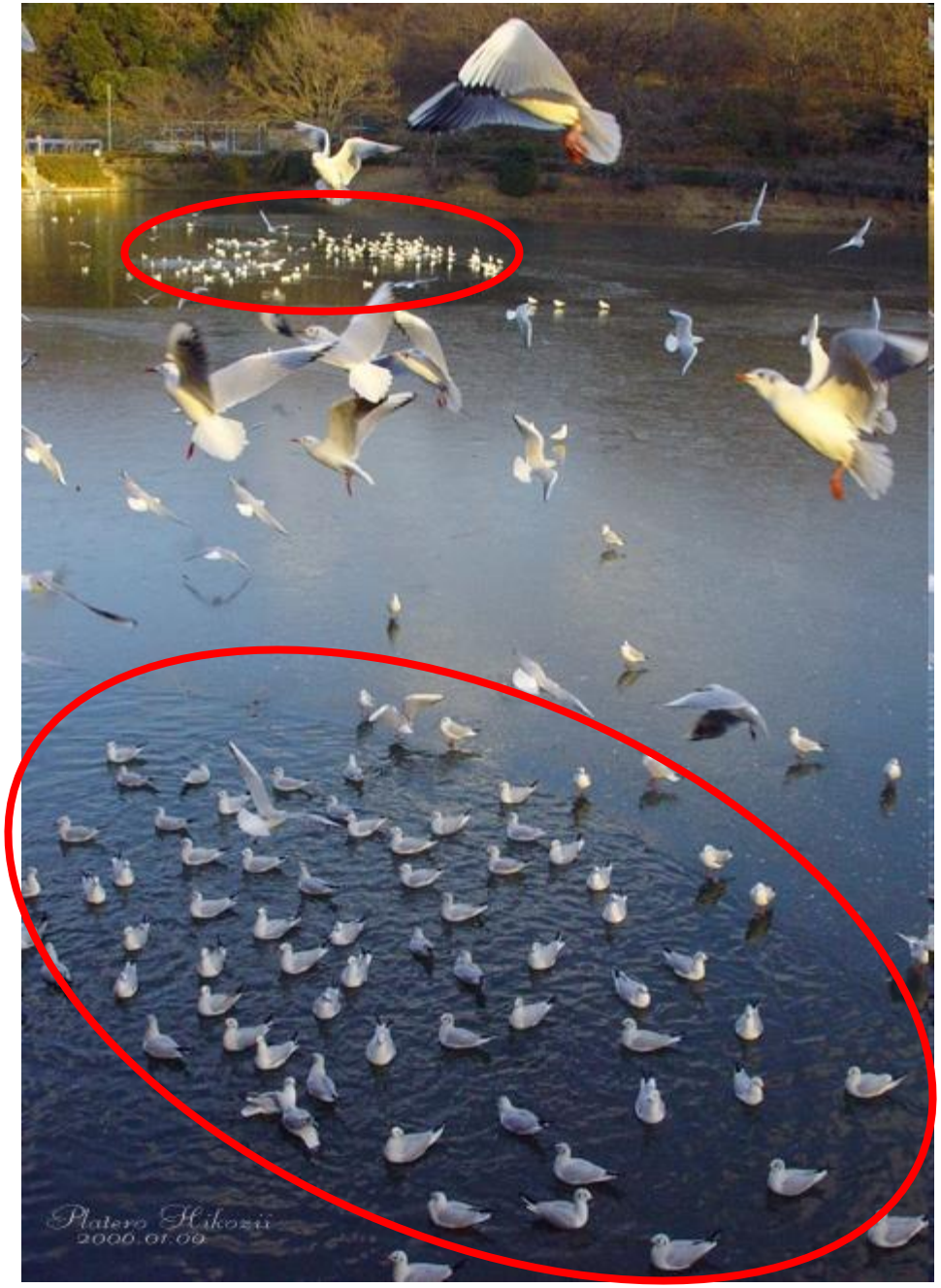
☆ゲシュタルト性
「桧」は「木」＋「会」ではない

Pair	旧字体選好 %
亜亞	2.4
壷壺	74.1
会會	4.7
桧檜	71.8
観觀	0
潅灌	84.7
尭堯	31.8
焼燒	3.5
経經	5.9
頚頸	81.2

実験心理学的な異体字選好課題を用いる理由



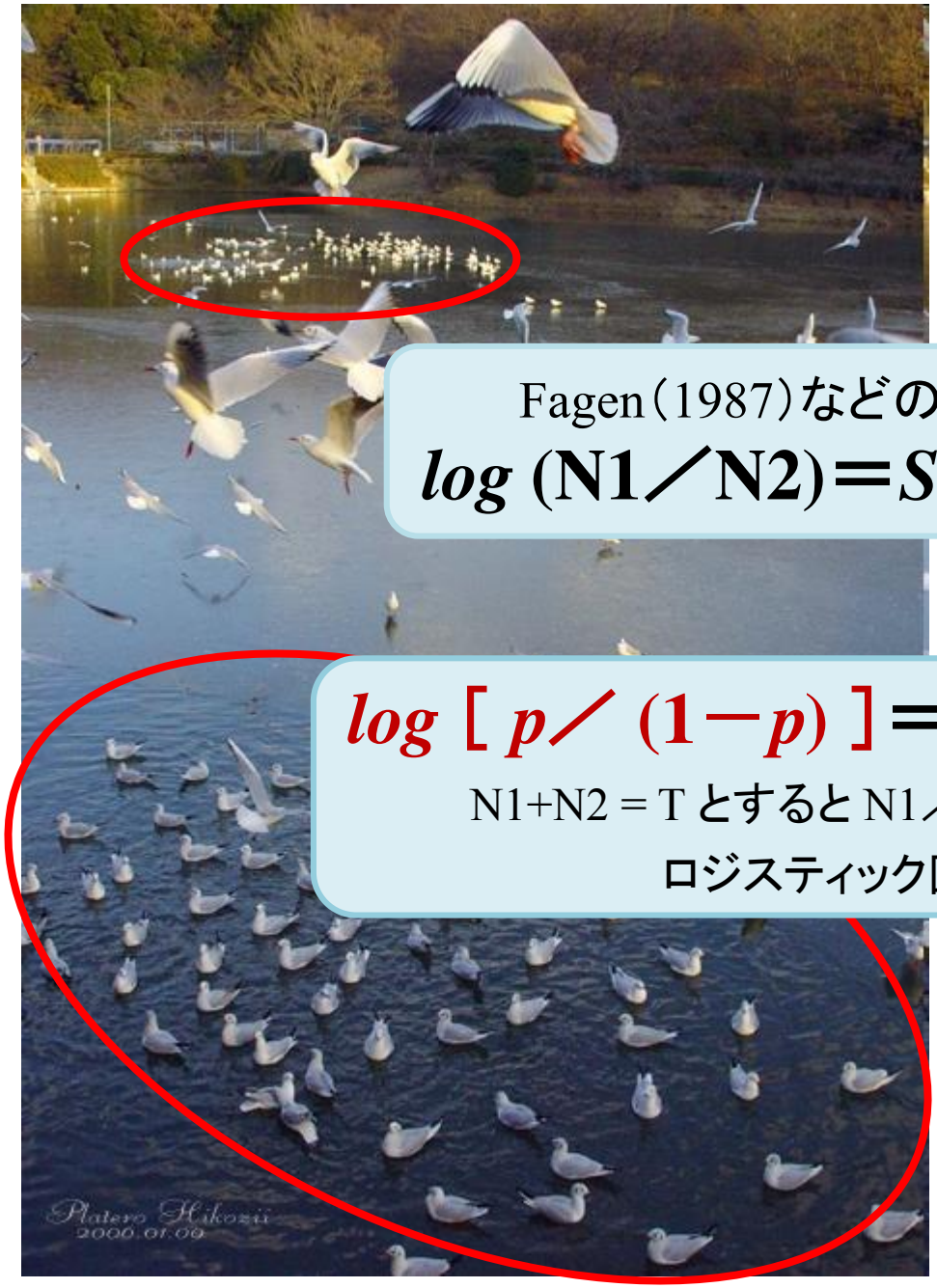
1. 文字コードなどの関係で，スマホやPCなどでは表示できない，入力できない異体字が多数ある
2. 奈良県葛城市の正式な表記はAの字体，東京都葛飾区はBの字体。Windows OS の Vista が登場するまではAの字体が表示されることが普通であった。現在はBが表示されるのが一般的
3. このような問題があるため，異体字の頻度を計数する際はテキスト入力前の原紙を目視で確認する必要が生じる場合がある
4. また，たとえば，表示や入力には問題がない「檜」を使いたいという意向を持っている場合でも，それがJIS第2水準であるために，JIS第1水準の「桧」の方が最初の変換候補として表示され，そこで妥協してしまうというケースがあるかもしれない
5. この方法は国際比較研究も容易。台湾の日本語学習者に「あなたが日本の知り合いにメールを書くとしたら，どちらの字体を選びますか」といった調査などに利用できる（横山・當山・高田・米田，2008）



理想自由分布理論
Ideal Free Distribution Theory

2つの地点に生息する捕食者数（例：トリ）と，それぞれの地点に存在するエサの量との関係を予測する理論

写真は「platero 飛孤爺 blog」2006年1月23日から引用
http://hikozii.air-nifty.com/_hikozii/2018/01/2018-df30.html
撮影地：横浜市鶴見区三ツ池公園
撮影日：2006年1月9日



生息地2
餌の量 A2
個体数 N2

Fagen（1987）などの一般理想自由分布モデル

$$log\,(N1／N2)＝S\,log\,(A1／A2)＋\,log\,b$$

この b は選好バイアス

$$log\,[\,p／\,(1－p)\,]＝S\,log\,(A1／A2)\,＋\,log\,b$$

N1+N2 = T とすると N1／T = p また (T－N1)／T = $1－p$

ロジスティック回帰の形になっている

生息地1
餌の量 A1
個体数 N1

ロジスティック回帰の形とは

$$log\,[\,p\,/\,(1-p)\,]=Z' \qquad (1)$$

たとえば2変数の場合は

$$Z' = a_1 \times \text{変数}1 + a_2 \times \text{変数}2 + b$$

ただし，log は自然対数，つまり底は e

$p\,/\,(\,1-p\,)$ はオッズ，$log\,[\,p\,/\,(1-p)\,]$ はロジットという

これは以下のように変形できる（**逆ロジット変換**）

$$p = exp\,(\,Z'\,)\,/\,[exp\,(\,Z'\,)+1\,]$$

$$= 1\,/\,[\,1+exp\,(-Z'\,)\,] \qquad (2)$$

Yokoyama & Sanada (2009) によれば，「偏差値」から50を引いて10で割ったものを Z とすると

$$Z' = 1.7 \times Z$$

アナロジーで説明してみよう

四国の風景から

愛媛県の加茂川（西条市）

水流は石鎚山から。青石の産地。鮎が釣れる

もちろん，フィクションですが・・・

愛媛県の加茂川： 体表に「桧」や「檜」と似た模様の鮎が生息

- それぞれが群れをなす。「桧」群と「檜」群はナワバリが別
- 群間で味や大きさの平均値に差はない


「桧」模様の鮎が A2 匹
餌，報酬（鮎）→ 新旧字体刺激
捕食者（鳥）→ 私たち
☆「言語接触は報酬」仮説（横山，2018）
無意識のごちそう
「檜」模様の鮎が A1 匹

次は高知県の四万十川を訪ねてみよう

四万十川にある鮎のナワバリ（フィクションです！）

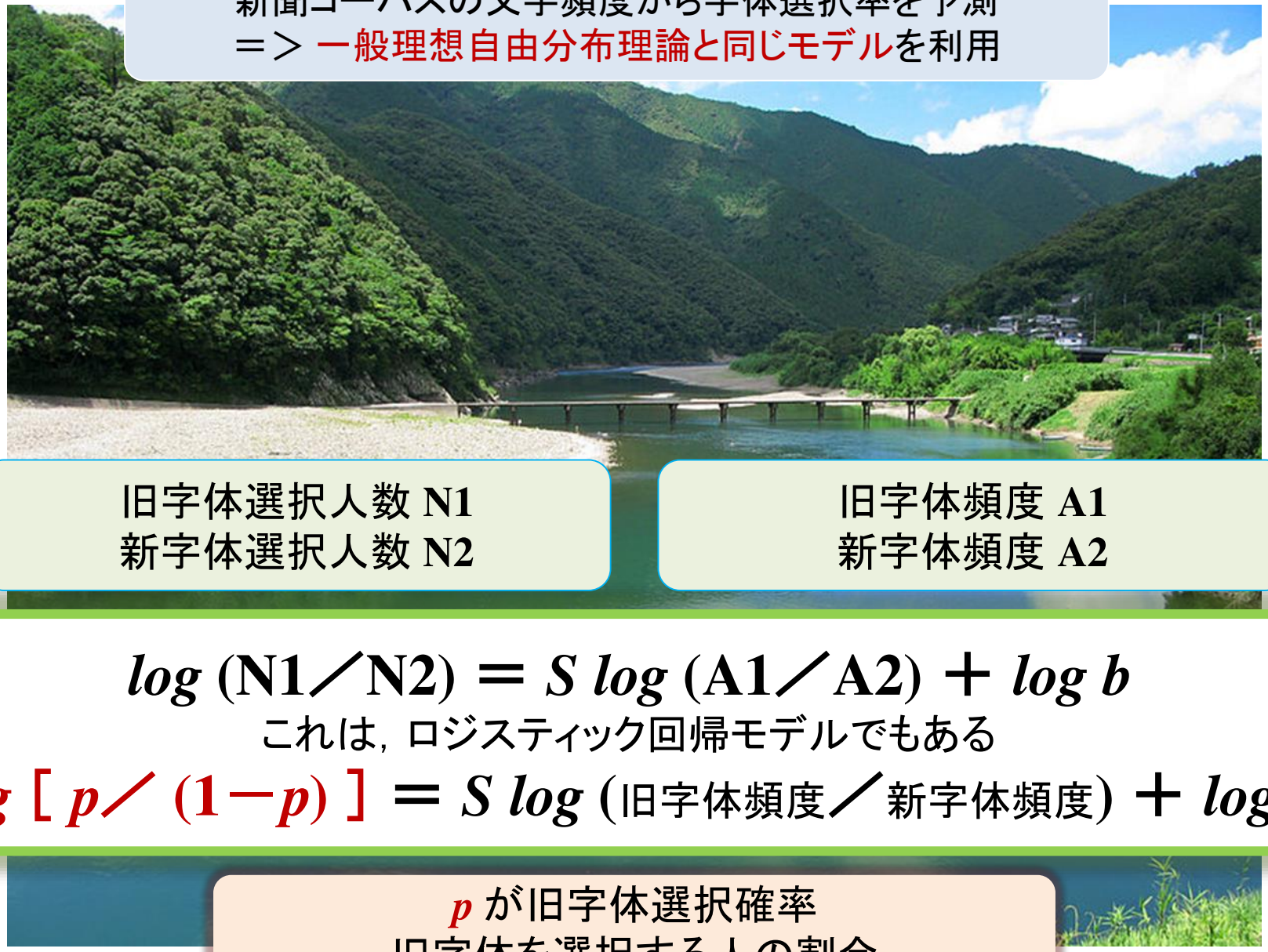
もちろん，フィクションですが・・・

高知県の四万十川： 体表に「壷」や「壺」と似た模様の鮎が生息

- それぞれが群れをなす。「壷」群と「壺」群はナワバリが別
- 群間で味や大きさの平均値に差はない

四万十川には「壷」群と「壺」群のナワバリがあるという（作り話！）

新聞コーパスの文字頻度から字体選択率を予測
＝＞ 一般理想自由分布理論と同じモデルを利用



旧字体選択人数 N1
新字体選択人数 N2

旧字体頻度 A1
新字体頻度 A2

$$log\ (N1／N2) = S\ log\ (A1／A2) + log\ b$$

これは，ロジスティック回帰モデルでもある

$$log\ [\ p／\ (1－p)\ ] = S\ log\ (\text{旧字体頻度}／\text{新字体頻度}) + log\ b$$

p が旧字体選択確率
旧字体を選択する人の割合

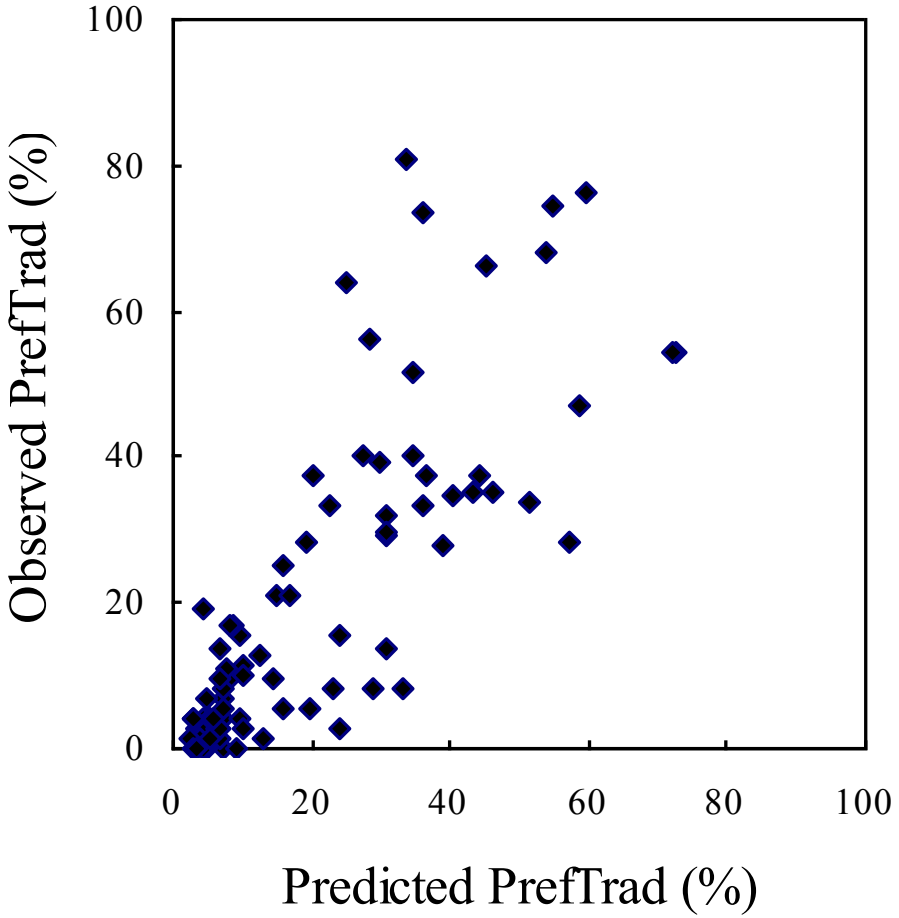
目視も併用して確認した頻度データ
（256ペアの一部）

Pair	頻度（新聞）	
	新字体	旧字体
亜亞	1035	5
壷壺	59	20
蛍螢	98	2
学學	54725	7
誉譽	2198	0
鴬鶯	16	4
鴬鸞	16	0
会會	161051	7
桧檜	230	15
覚覺	4990	3
撹攪	12	2
観觀	7794	0
潅灌	11	2
尭堯	72	45
焼燒	2553	1
区區	28396	0
欧歐	8001	0
経經	38698	9
頚頸	5	40

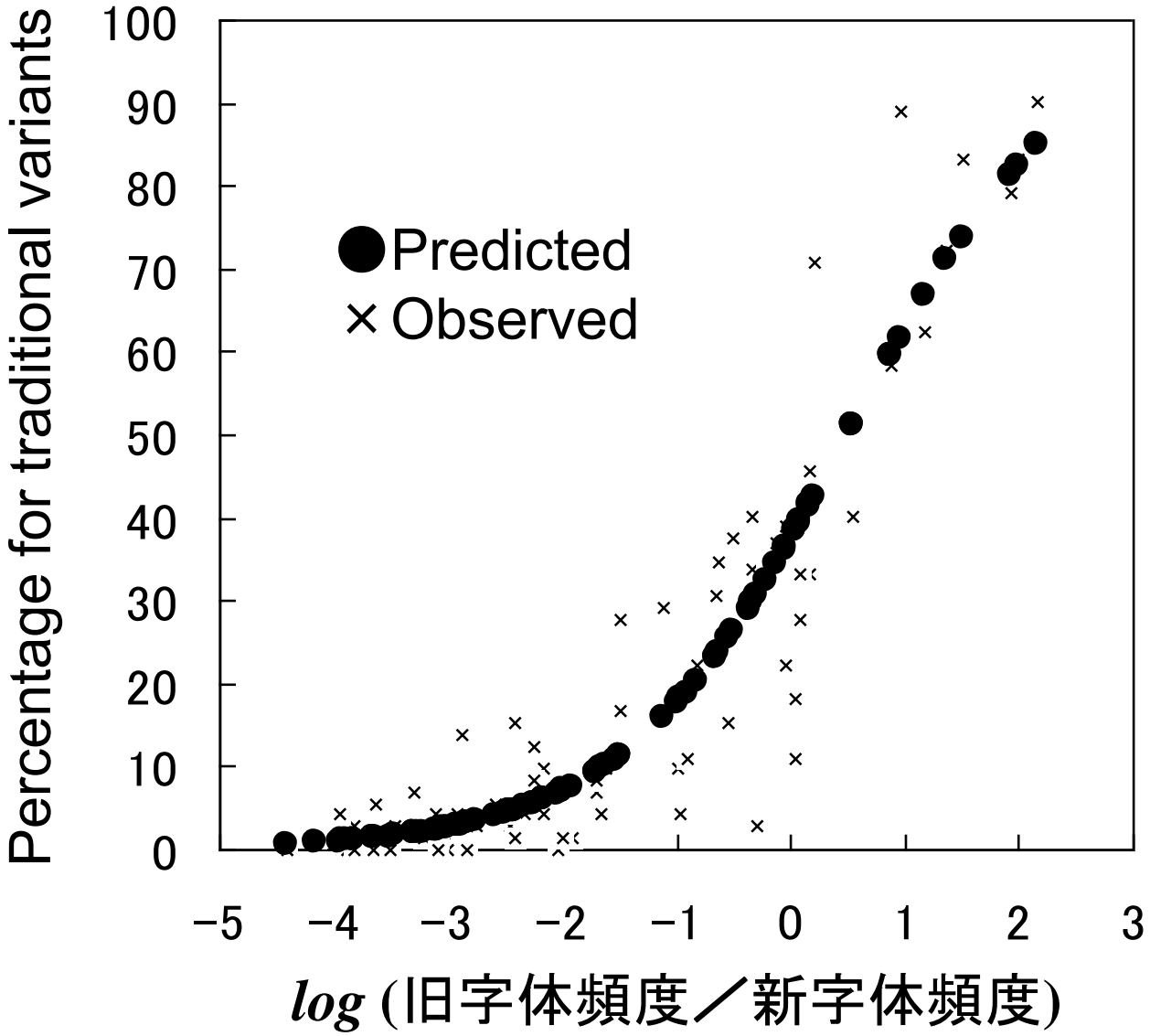
朝日新聞コーパスにおける文字頻度から字体選好確率を予測
下の表にはデータの一部を示す（256の異体字ペアから4ペアを例示）

ペア	新字体頻度	旧字体頻度	旧字体選好 観測値(%)	旧字体選好 予測値(%)
亜亞	1,035	5	1.4	13.1
壷壺	59	20	73.6	36.1
会會	161,051	7	2.8	3.1
桧檜	230	15	63.9	25.1

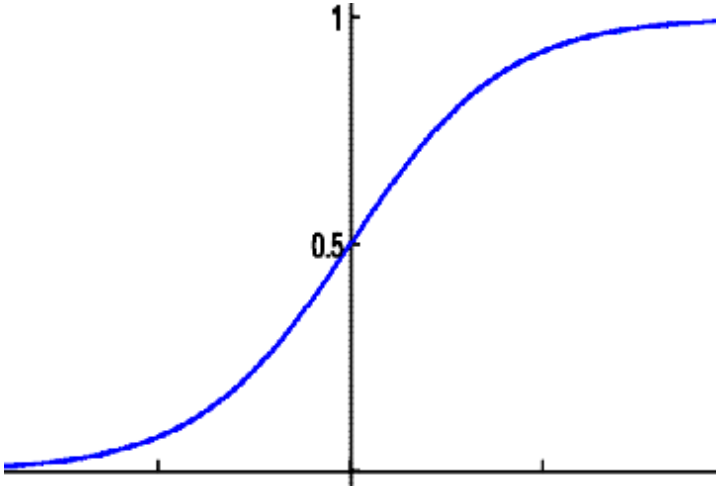
新聞コーパス頻度による旧字体選択率の予測
予測値と実測値の相関は r =.86



予測モデルの基本形はS字カーブ（ロジスティック関数）

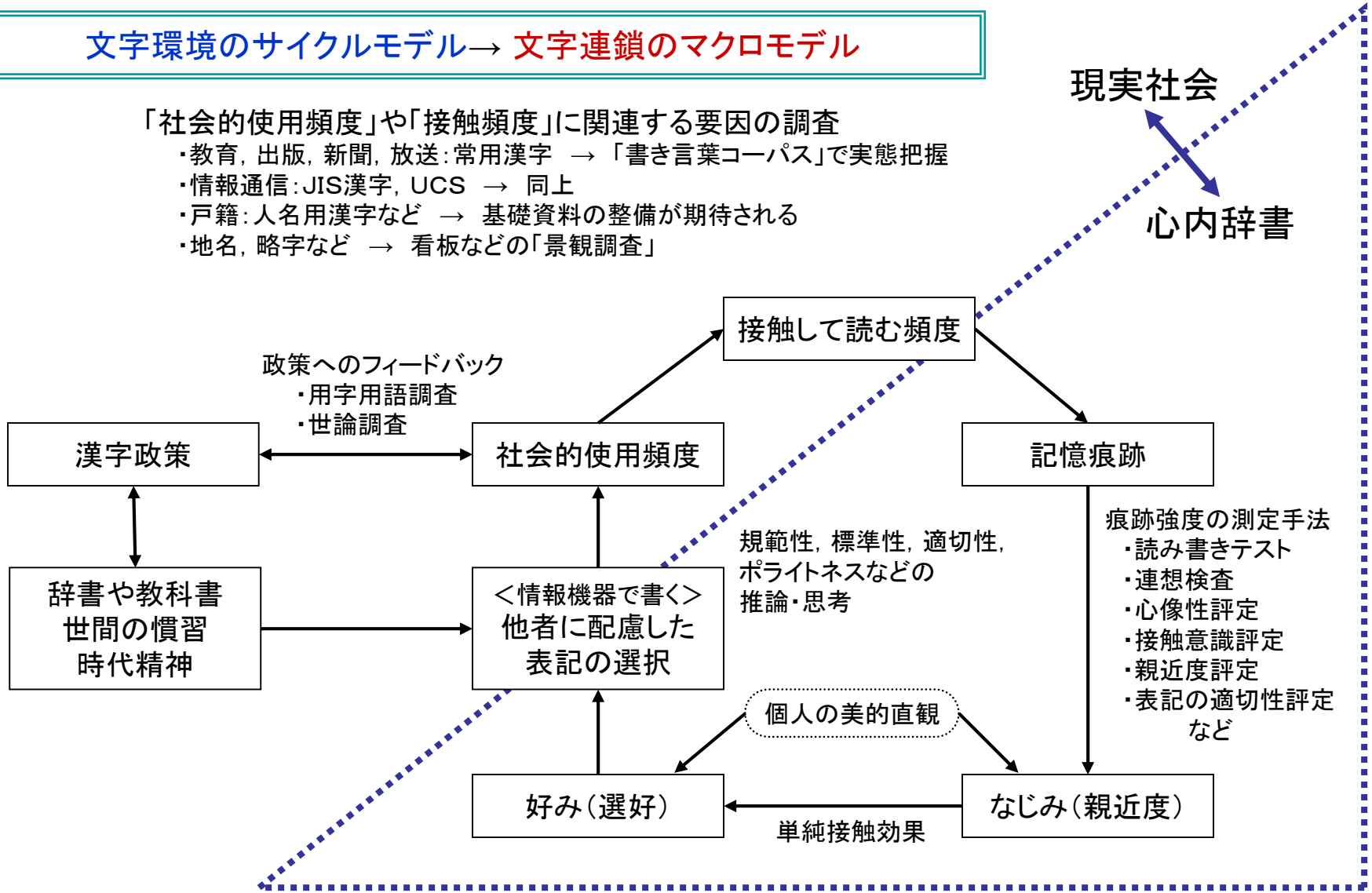


言語変化のS字カーブ
　計量言語学の分野でAltmannら（1983）などがロジスティック関数を提唱
細菌学，人口学，動物生態学などのS字カーブ
　昔から詳しく研究されている
　ヘール＝マルサスのモデルなどが有名：基本はロジスティック関数
新製品の普及過程や流行現象のS字カーブ
　ロジスティック関数ではないがきわめてよく似ている

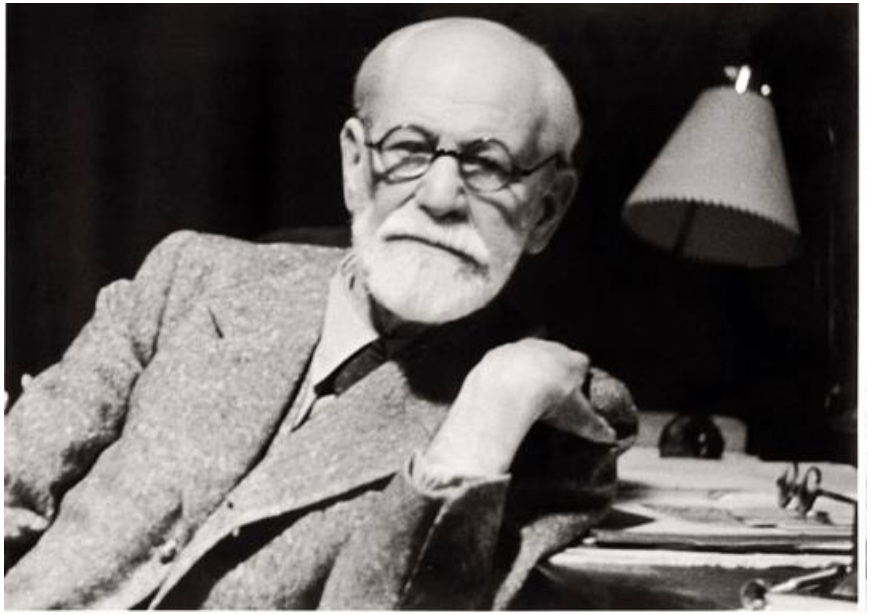


文字環境のサイクルモデル→ 文字連鎖のマクロモデル

「社会的使用頻度」や「接触頻度」に関連する要因の調査
- ・教育，出版，新聞，放送：常用漢字 → 「書き言葉コーパス」で実態把握
- ・情報通信：JIS漢字，UCS → 同上
- ・戸籍：人名用漢字など → 基礎資料の整備が期待される
- ・地名，略字など → 看板などの「景観調査」

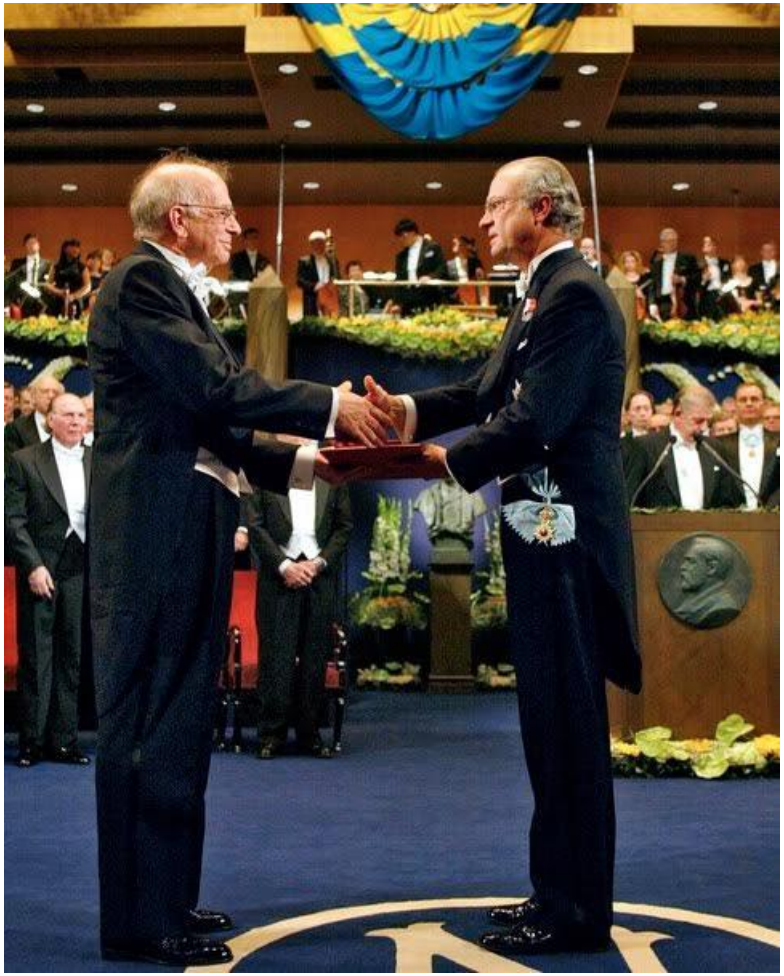


ここまでは数理生態学のお話
次は心理学の伝統的理論による説明



フロイト先生

感覚・知覚心理学と認知心理学にもとづくモデル化



- 認知心理学の分野から
  Kahneman & Tversky (1979)
  2002年ノーベル経済学賞を
  受賞した研究を参考に

  価値関数（Value function）の
  ようなものを考えてみる
- 感覚・知覚心理学の分野から
  フェヒナーの法則
- 動物心理学の分野から
  一般対応法則
- 上記すべてを理論的につなぐ

価値関数(value function)の導入

Kahneman & Tversky(1979) ノーベル経済学賞を受賞した研究を参考に

- 字体選好は,その字体を選択した場合にもたらされる心理的な利得と損失の収支を示す「価値関数」に左右されると仮定
- 「言語接触は報酬」仮説に立脚すると，価値関数に関係がありそうな要因として，字体と人間との「親疎関係」に着目するのは不自然ではないだろう
- 本研究は，親疎関係や好意度をうまく捉えるための近似的な指標として，「親近度（なじみ）」を取り上げた
- 価値関数**V**を下に示す。旧字体親近度（familiarity of traditional form）を**Fam1**，新字体親近度（familiarity of simplified form）を**Fam2**とし，両者の差が利得と損失の収支になると仮定（ここでは旧字体を選んだときの収支）

$$V = Fam1 - Fam2$$

価値関数(新旧字体間の親近度の差)から選好確率を予測できるか？

- 価値関数V，すなわち新旧字体間の親近度の差は

  V＝Fam1－Fam2

  ＝旧字体親近度－新字体親近度

- 旧字体を選好する確率 p をロジスティック回帰分析で予測
- 説明変数はV，つまり価値関数
- 親近度は天野・近藤(1998)による7段階評定データ。1文字ずつ測定したもの

$$\log\left[\, p \,/\, (1-p) \,\right] = W\,(\mathrm{Fam1}-\mathrm{Fam2}) + B$$

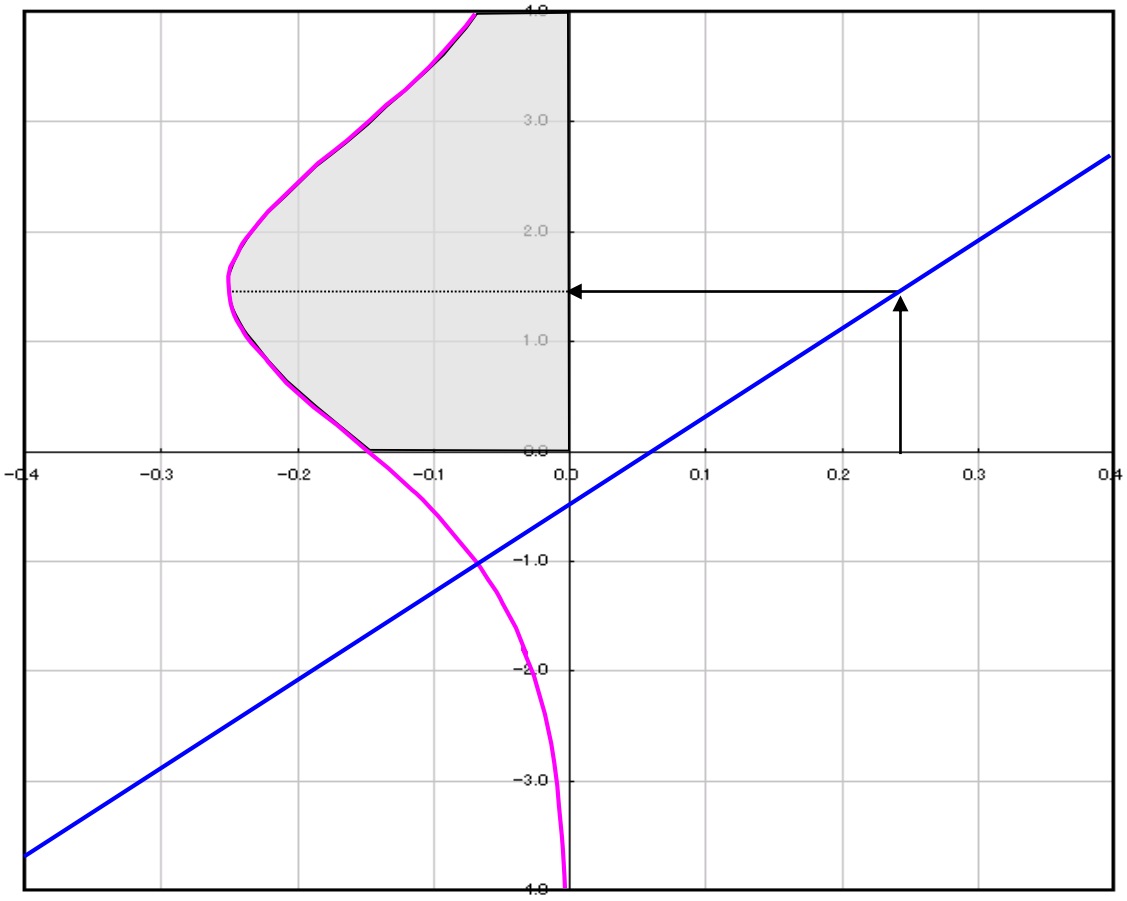
これを「親近度の比較判断モデル」と命名

「サーストンの比較判断の法則」から着想

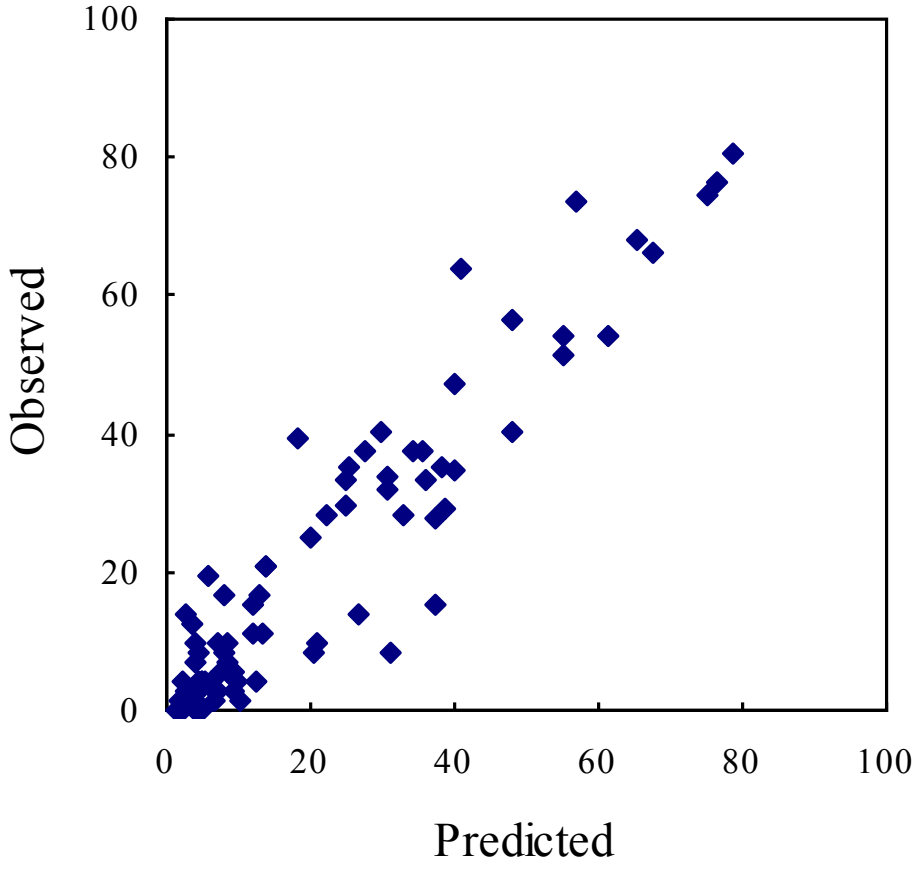
ロジスティック回帰分析の模式図（横山・真田，2007）

ヨコ軸が価値関数 V の値

正規分布で影のついた部分の面積が旧字体を選択する確率 *p*



新旧字体間の親近度差で選好確率を予測すると予測値と実測値の相関は $r = .90$前後に



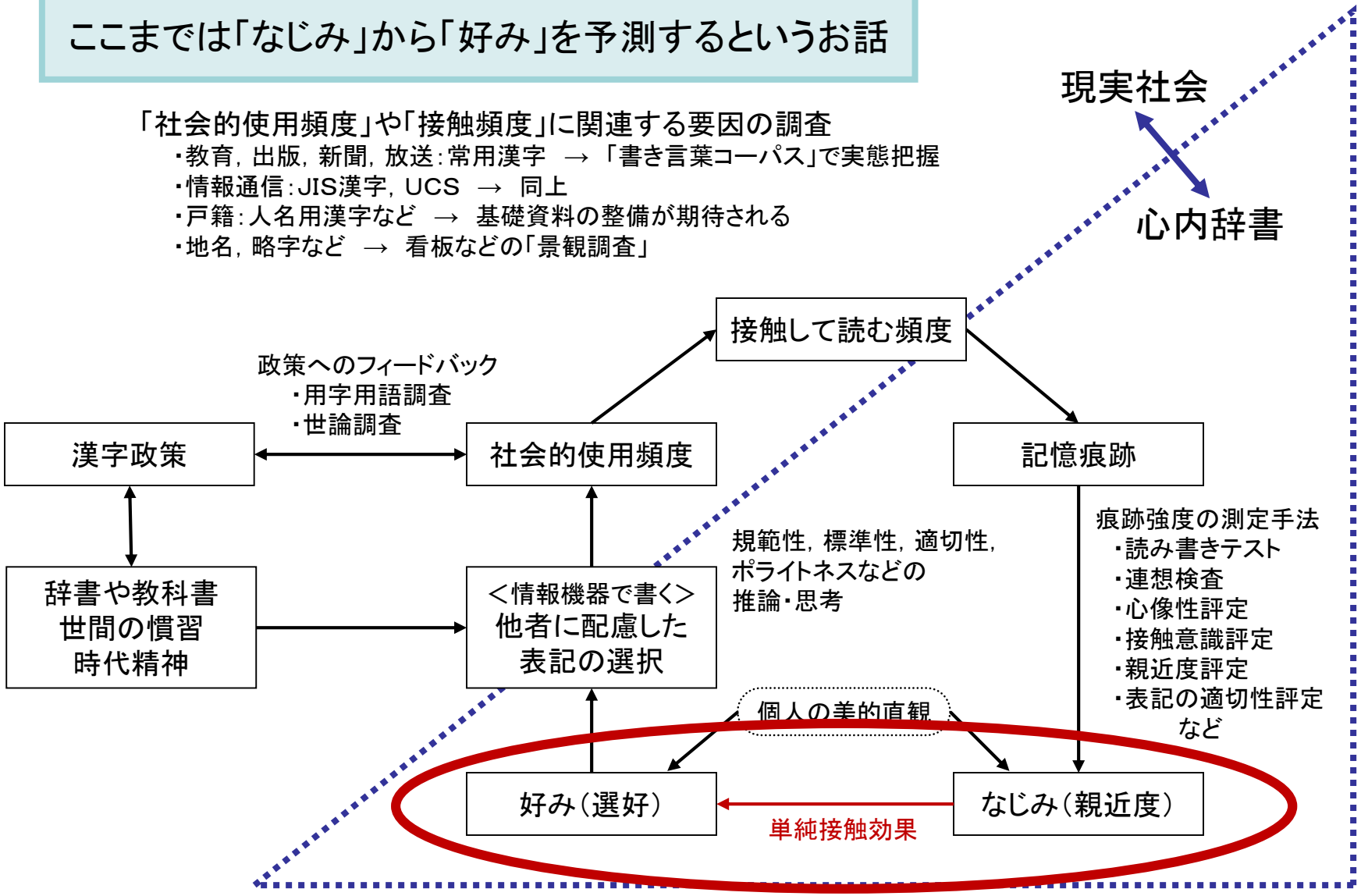
東京と京都における10年間の経年実験

$$log\,[\,p/(1-p)\,] = W\,(\mathrm{Fam1}-\mathrm{Fam2})+B$$

	r	*W*	*B*
Experiment 1			
Tokyo'96	.900	0.835	-0.547
Kyoto'98	.935	1.033	-0.517
Experiment 2			
Tokyo'05	.857	0.703	-0.752
Kyoto'06	.913	0.824	-0.274

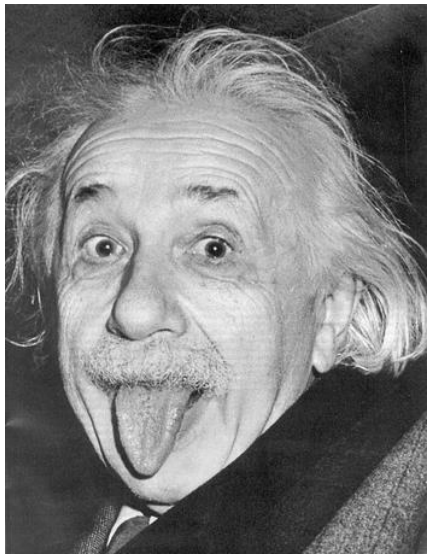
ここまでは「なじみ」から「好み」を予測するというお話

「社会的使用頻度」や「接触頻度」に関連する要因の調査
- ・教育，出版，新聞，放送：常用漢字 → 「書き言葉コーパス」で実態把握
- ・情報通信：JIS漢字，UCS → 同上
- ・戸籍：人名用漢字など → 基礎資料の整備が期待される
- ・地名，略字など → 看板などの「景観調査」

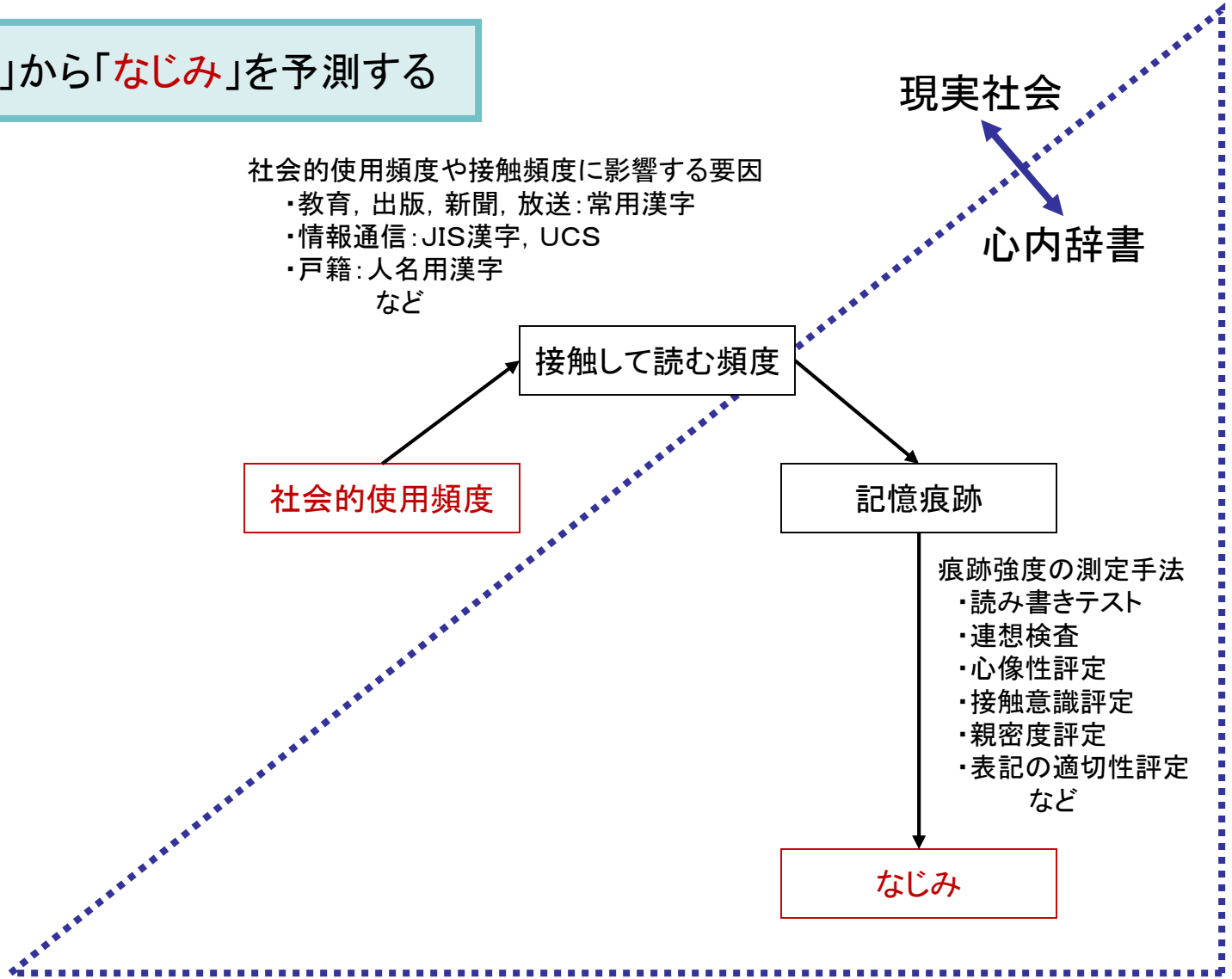


- どうやら familiarity → preference の予測精度はきわめて高い
- では frequency →（memory trace, familiarity）→ preference は？
- 検証したいモデルは

  現実社会での用例の分布状況 →（記憶痕跡，なじみ）→ 選好
- 世の中で使用されている文字の頻度：新聞のフルテキストデータ（すなわち，新聞コーパス）を用いてカウント



まず，「頻度」から「なじみ」を予測する



頻度(frequency)で選好確率を予測できるか？
心理物理学(Psychophysics)を援用してみよう！

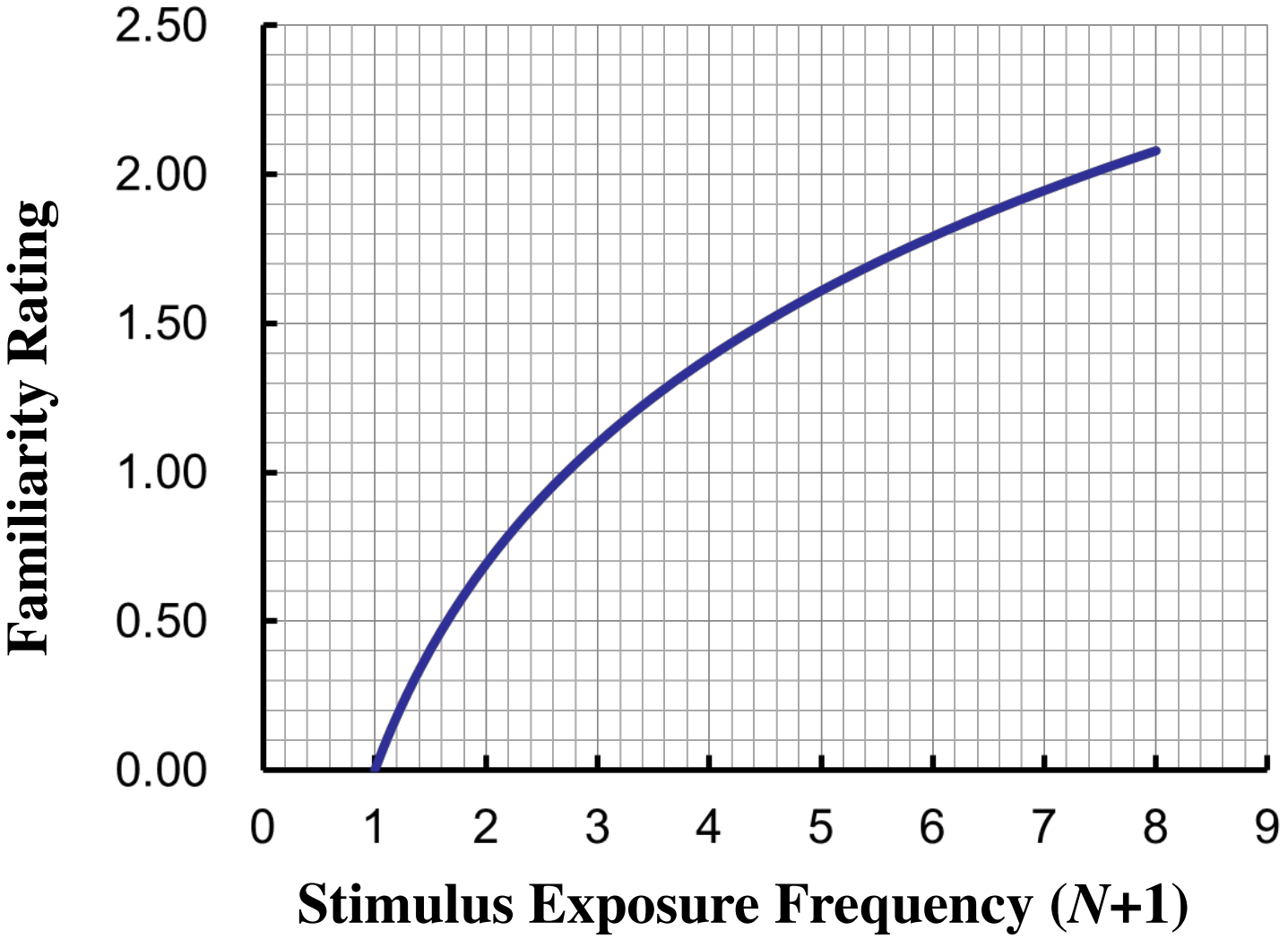
- フェヒナーの法則（Fechner's law）：　ある刺激から生じる感覚尺度を E，刺激強度を I，定数を K とすると

$$E = K \log(I) \quad (1)$$

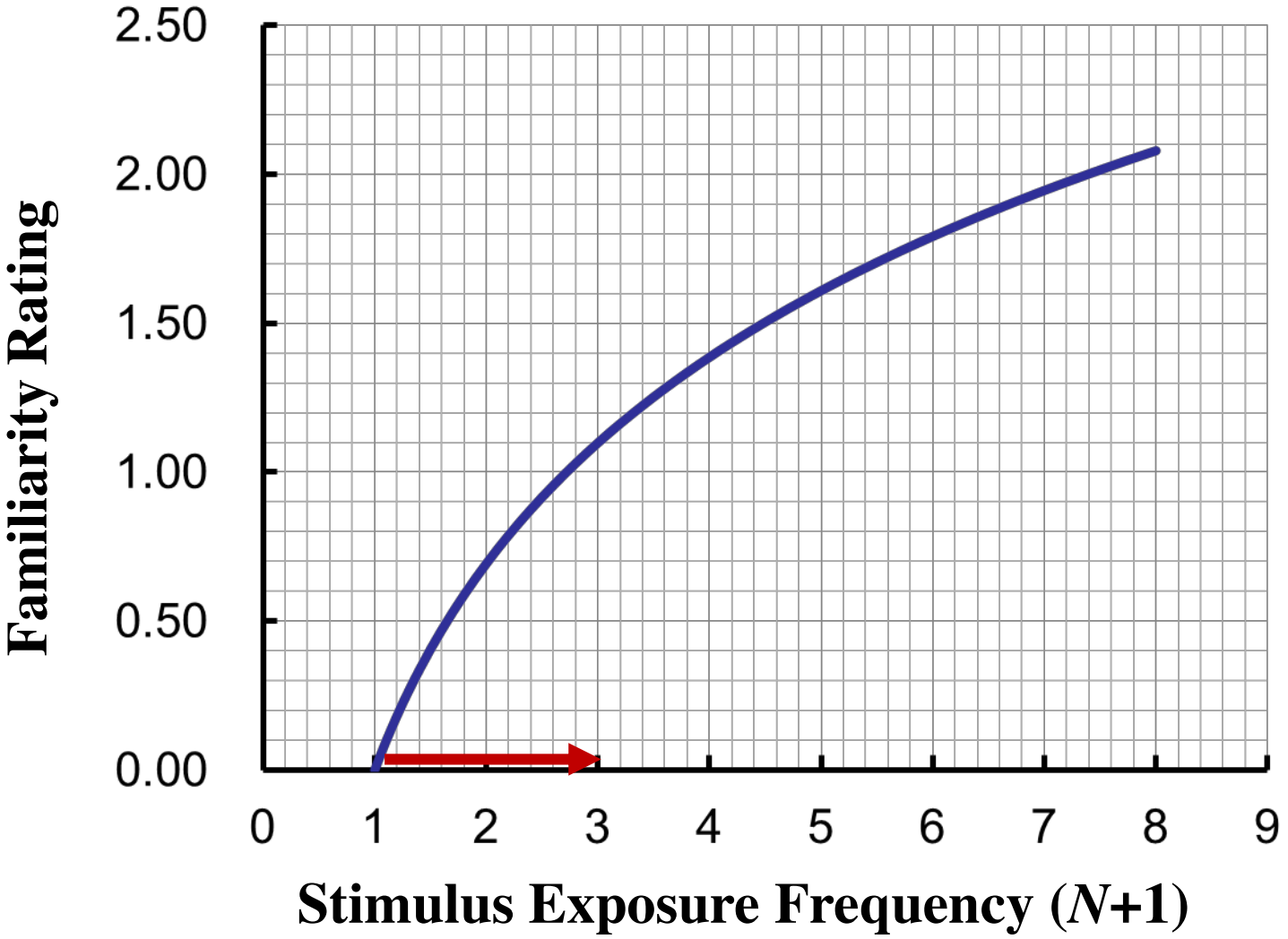
漢字親近度と漢字頻度の関係に式(1)を適用すると

- 漢字親近度＝ $K \log$ (漢字頻度) ＋ C　　(2)
- 旧字体親近度＝ $K \log$ (旧字体頻度) ＋ C　(2.1)
- 新字体親近度＝ $K \log$ (新字体頻度) ＋ C　(2.2)

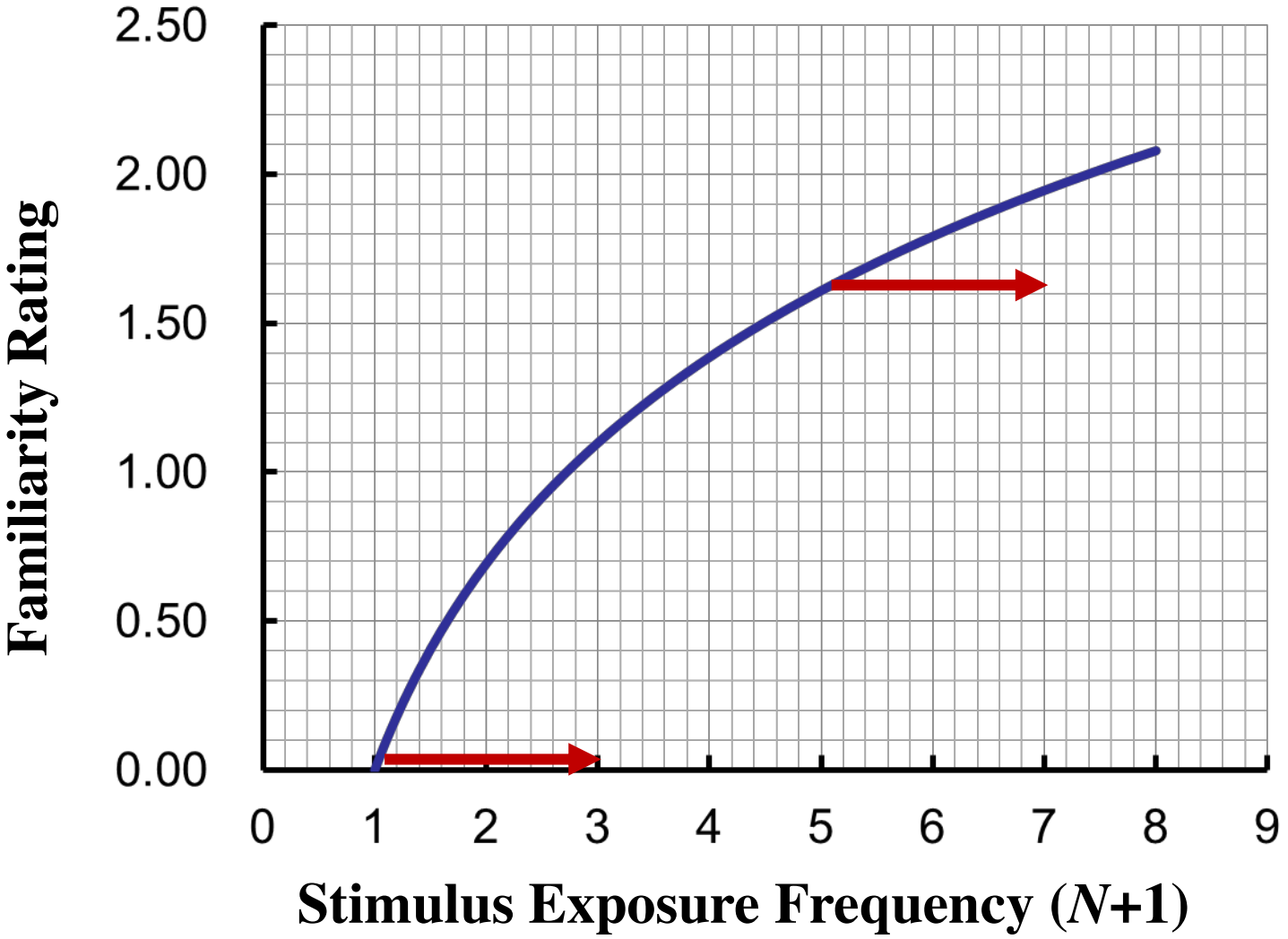
ヨコ軸：独立変数 → 接触経験（接触回数に1を加えている）
タテ軸：従属変数 → 親近度評定の値（0から4の5段階評定）



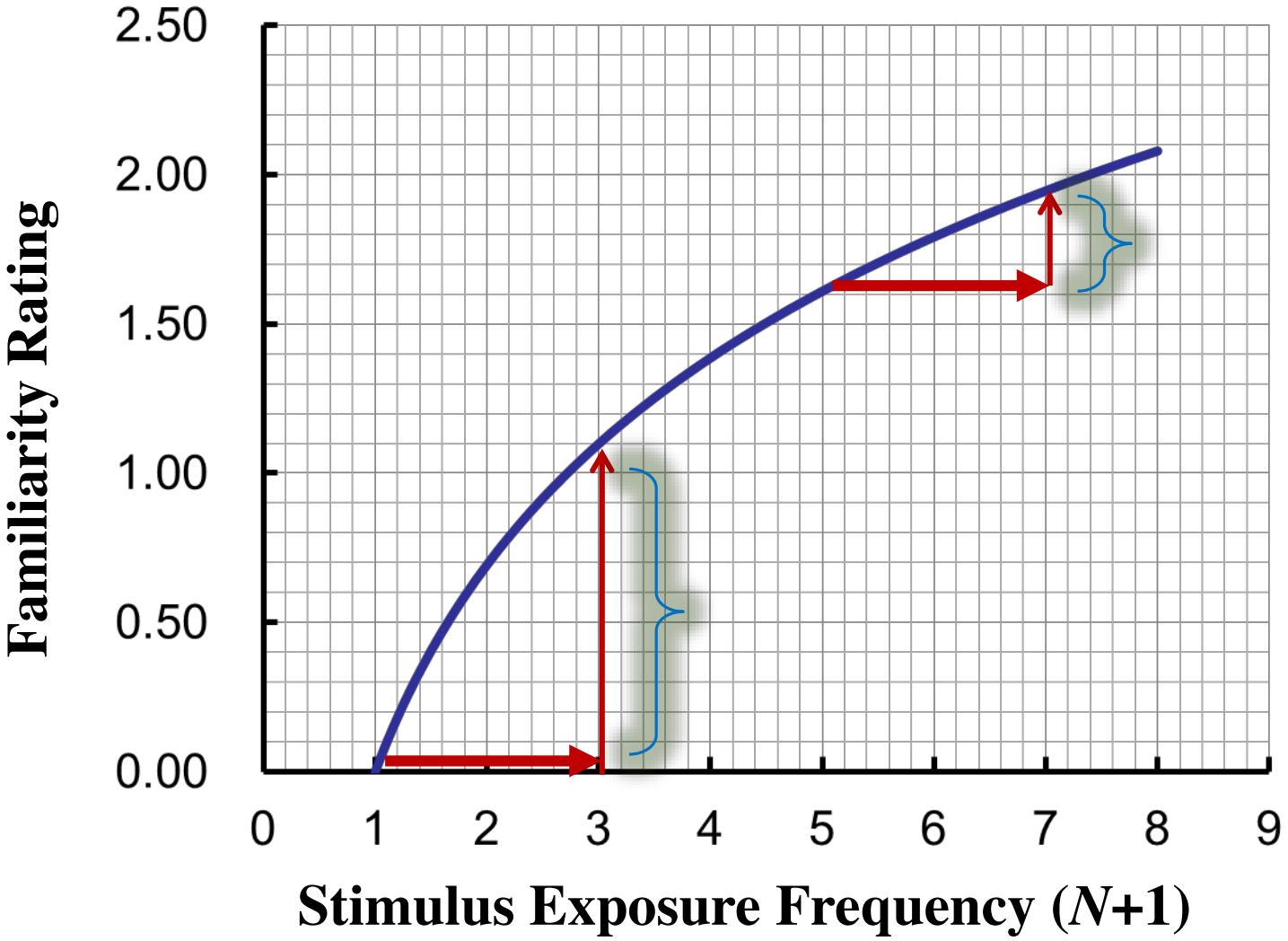
接触回数が少ない場合に接触経験が2回増えた



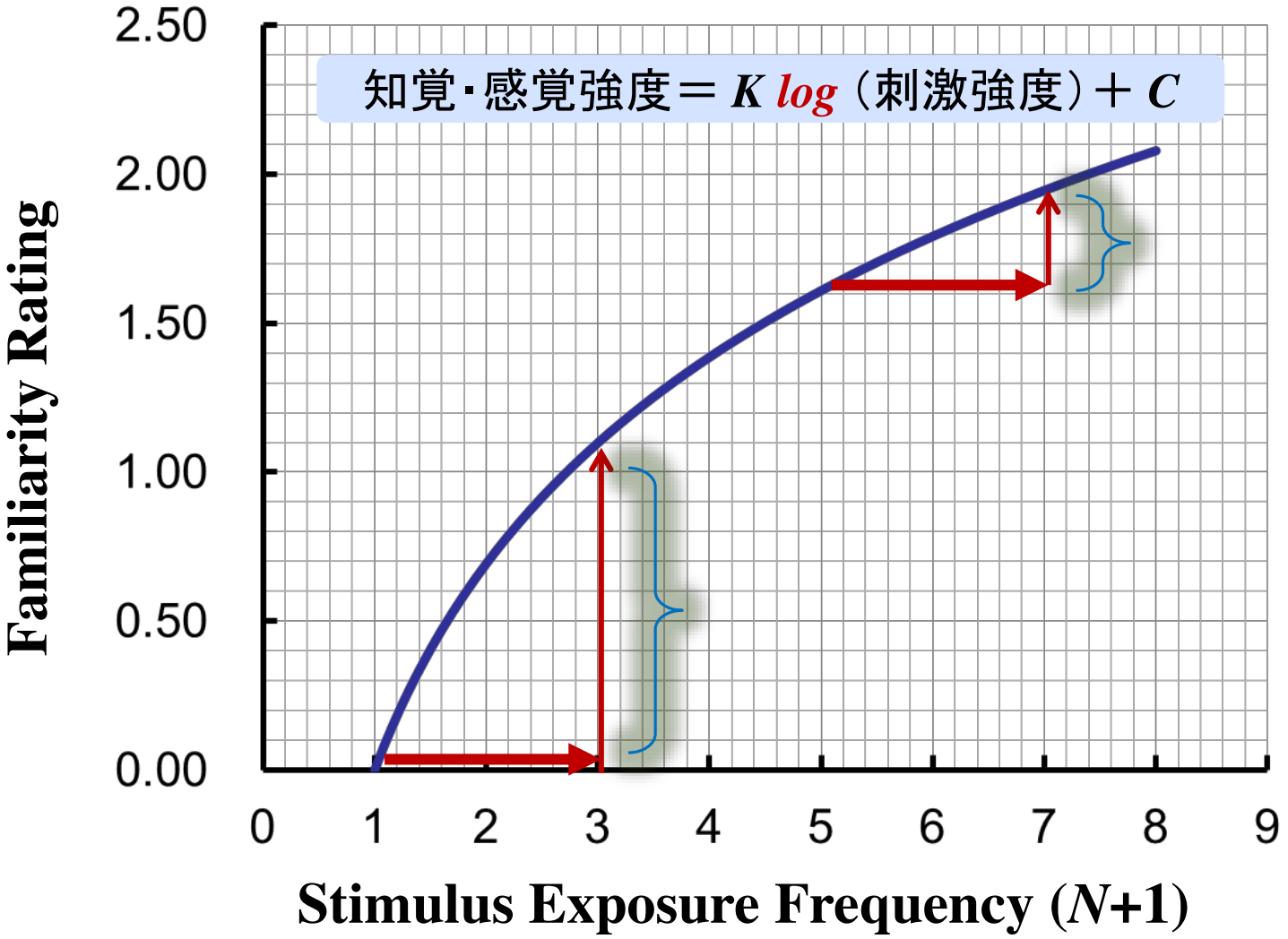
接触回数が少ない場合に接触経験が2回増えた
接触回数が（相対的に）多い場合に接触経験が2回増えた



接触経験の増加が親近度評定値の向上に及ぼす効果は，接触回数が少ない場合の方が大きい



Fechner's law（フェヒナーの法則）：経済学の限界効用逓減法則も同じ



価値関数(新旧字体間の親近度の差)を「頻度」で書き換えると

- 字体選好確率はロジスティック回帰分析でかなり精度よく予測できる場合がある
- 説明変数は価値関数 V
- 価値関数 Vは，新旧字体間の親近度の差

$$log\,[\,p/(1-p)\,] = W\,(\text{旧字体親近度} - \text{新字体親近度}) + B$$

先の式(2.1)と(2.2)を代入・整理し，

$W \cdot K$ を S とおくと

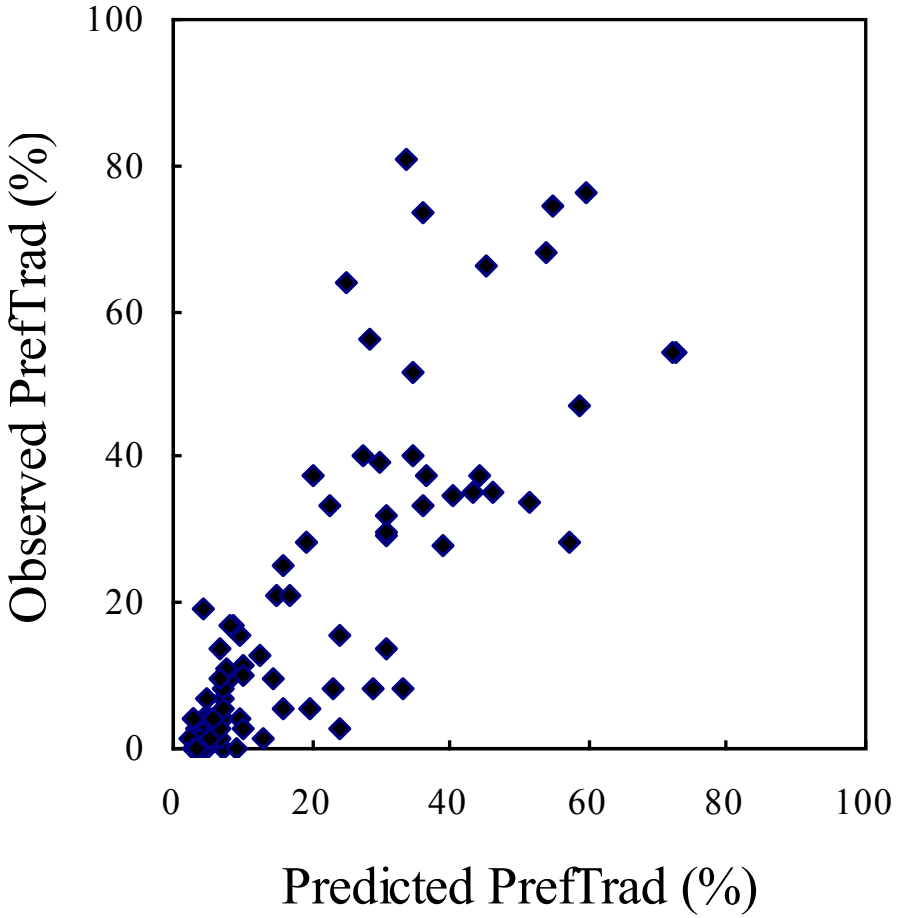
$$log\,[\,p/(1-p)\,] = S\,log\,(\text{旧字体頻度}/\text{新字体頻度}) + B$$

これは「言語接触は報酬」仮説にもとづくモデル

先に示した
目視も併用して確認した
頻度データ
（256ペアの一部）

Pair	旧字体選好 %	頻度（新聞） 新字体	旧字体
亜亞	2.4	1035	5
壷壺	74.1	59	20
蛍螢	54.1	98	2
学學	7.1	54725	7
誉譽	3.5	2198	0
鴬鶯	65.9	16	4
鴬鸞	25.9	16	0
会會	4.7	161051	7
桧檜	71.8	230	15
覚覺	1.2	4990	3
撹攪	10.7	12	2
観觀	0	7794	0
潅灌	84.7	11	2
尭堯	31.8	72	45
焼燒	3.5	2553	1
区區	1.2	28396	0
欧歐	24.7	8001	0
経經	5.9	38698	9
頚頸	81.2	5	40

繰り返しになりますが，新聞コーパス頻度による旧字体選択率の予測
予測値と実測値の相関は r =.86

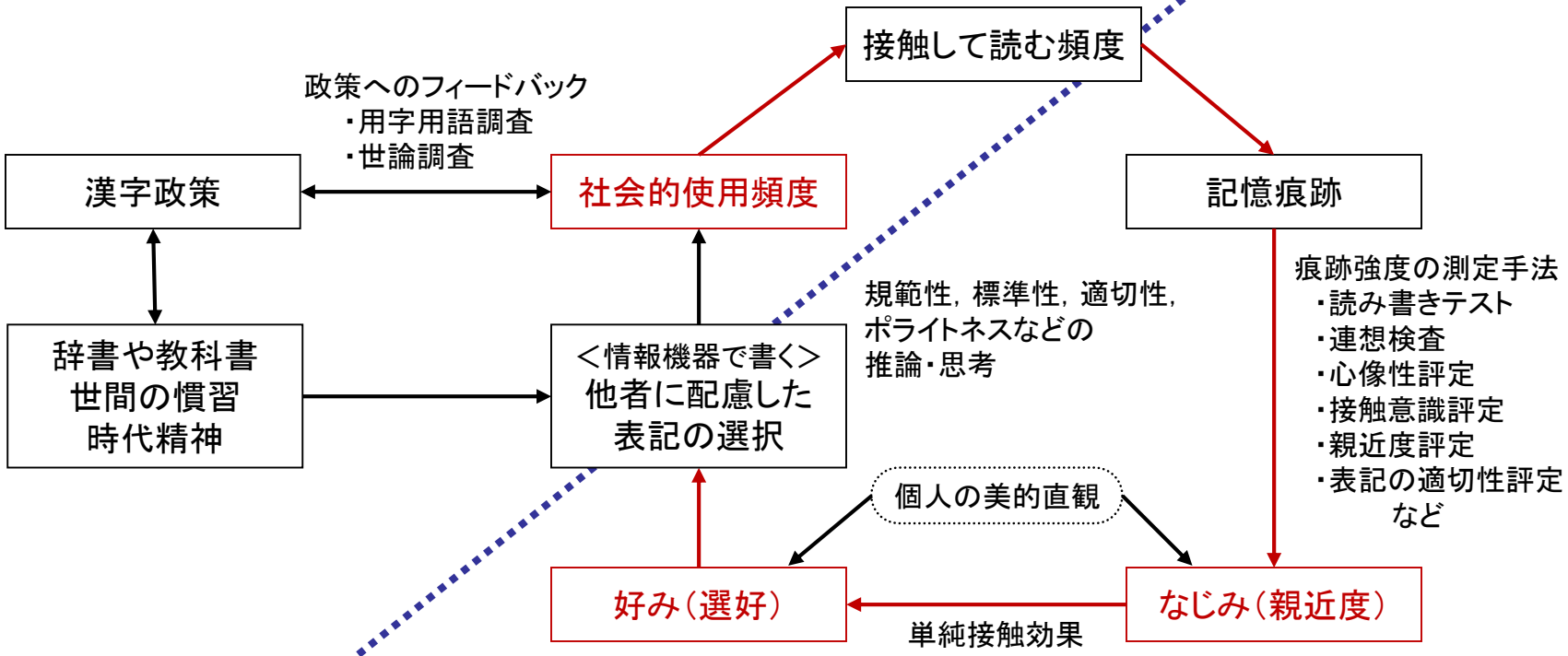


「社会的使用頻度」→（なじみ）→「好み」の連鎖モデル
人間は，言語刺激に接触する頻度の「分布」を学習する

現実社会

心内辞書

「社会的使用頻度」や「接触頻度」に関連する要因の調査
- ・教育，出版，新聞，放送：常用漢字 → 「書き言葉コーパス」で実態把握
- ・情報通信：JIS漢字，UCS → 同上
- ・戸籍：人名用漢字など → 基礎資料の整備が期待される
- ・地名，略字など → 看板などの「景観調査」



一般理想自由分布理論（Generalized Ideal Free Distribution Theory）と一般対応法則（Generalized Matching Law）の関係

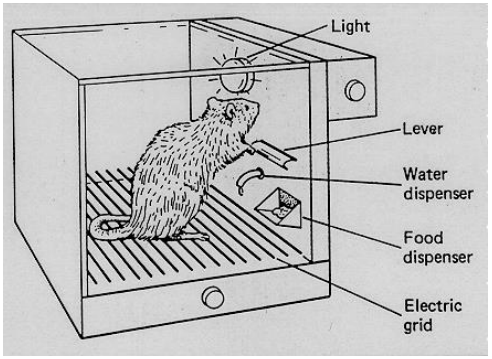
- 「一般理想自由分布理論」は生態学のFagan(1987)などが提唱
- 「一般対応法則」は動物行動学の分野でBaum(1976)などが発見

  反応比(R1／R2)は報酬比(r1／r2)と次のように対応

  $$\mathbf{R1／R2 = \mathit{b}\ (r1／r2)^{\mathit{S}}}$$

- 両辺の自然対数をとって展開すると

  $$\mathit{log}\ (\mathbf{R1／R2}) = \mathit{S}\ \mathit{log}\ (\mathbf{r1／r2}) + \mathit{log}\ \mathbf{\mathit{b}}$$



☆ これは一般理想自由分布理論のモデル式と同じ形

＊なお，*b* は反応バイアス，*S* は感度。また *b* と *S* はテスト理論の困難度と識別力に関係あり

- *p* を旧字体選択確率，r1 をコーパスで計数した旧字体頻度，r2 を新字体頻度とすると

  $$\mathit{log}\ [\ \mathit{p}／\ (1－\mathit{p})\ ] = \mathit{S}\ \mathit{log}\ (\text{旧字体頻度／新字体頻度}) + \mathit{log}\ \mathbf{\mathit{b}}$$

  ここで $\mathit{log}\ b$ を B とおくと，先の「言語接触は報酬」仮説モデルと同じになる

まとめにかえて

1. 人間は，言語刺激に接触する頻度の「分布」を学習する。また，人間はその「分布」を変化させる原動力の一つである
2. 言語生活研究でインフォーマント（調査対象者）から収集できるのはスモールデータであることが多い。スモールデータを扱う手法も重要
3. 文字コードなどの関係で，コーパスを使って異体字の頻度を正確に計数することはできない場合があることに留意が必要。文字研究者による目視も併用すべき
4. その原因の一つとして「JIS漢字のネジレ問題」が有名
5. たとえば，現在は「檜」がJIS第2水準で，「桧」はJIS第1水準。しかし，1995年ごろまでのPCでは「檜」がJIS第1水準で，「桧」はJIS第2水準であることがしばしばあった
6. ちなみに，コンピュータによる日本語研究を世界で初めておこなったのは国立国語研究所（国語研報告56『現代新聞の漢字』，1976）
7. 最初のJIS漢字規格（通称78JIS）を作ったのは国立国語研究所第3代所長の林 大（はやし・おおき）

＊本日の参考文献

Fagen, R. (1987). A generalized habitat matching rule. *Evolutionary Ecology*, **1**, 5-10.

Kunst-Wilson, W R., & Zajonc, R. B. (1980). Affective discrimination of stimuli that cannot be recognized. ***Science***, **207**, 557-558.

笹原宏之・横山詔一・エリク=ロング著 (2003).『現代日本の異体字―漢字環境学序説』(国立国語研究所プロジェクト選書 №2) 三省堂

豊島正之 (1999).「書評　横山詔一、笹原宏之、野崎浩成、エリク・ロング編『新聞電子メディアの漢字---朝日新聞CD-ROMによる漢字頻度表---』」,『日本語科学』, **6**,

高田智和・横山詔一編 (2014).『日本語文字・表記の難しさとおもしろさ』彩流社

高田智和・馬場基・横山詔一編, 石塚晴通(監修) (2016).『漢字字体史研究 二: 字体と漢字情報』勉誠出版

横山詔一 (2006).「異体字選好における単純接触効果と一般対応法則の関係」,『計量国語学』, **25**(5), 199-214.

Yokoyama Shoichi & Wada Yukiko (2006). A logistic regression model of variant preference in Japanese kanji: an integration of mere exposure effect and the generalized matching law. *Glottometrics*,**12**, 63-74. RAM-Verlag, Germany

Yokoyama Shoichi & Sanada Haruko (2009). Logistic Regression Model for Predicting Language Change. Reinhard Koehler(Ed.). *Studies in Quantitative Linguistics* **5**, 176-192, RAM-Verlag, Germany

横山詔一・笹原宏之・野崎浩成・エリク＝ロング編著 (1998).『新聞電子メディアの漢字―朝日新聞CD-ROMによる漢字頻度表』(国立国語研究所プロジェクト選書 №1) 三省堂

横山詔一・笹原宏之・當山日出夫(2006).「文字コミュニケーションにおける異体字の選好と親近度:再検査法による信頼性の検討」,『社会言語科学』, **9**(1), 16-26

横山詔一・當山日出夫・高田智和・米田純子 (2008).「台湾日本語学習者は日本人の字体選好をいかに推論するのか」『情報処理学会研究報告: 人文科学とコンピュータ研究会報告』, 2008(8), 43-50

☆本研究は以下の成果物でもある

JSPS科研費17K18501, ならびに国立国語研究所基幹研究「通時コーパスの構築と日本語史研究の新展開」