学習者に特化した単語難易度

◎江原　遥, 佐藤一誠, 大岩秀和, 中川裕志
東京大学

Psychological
and neurological modeling

研究課題

第二言語（外国語）学習者は,

実際にどのような単語を知っているか？

仮説：

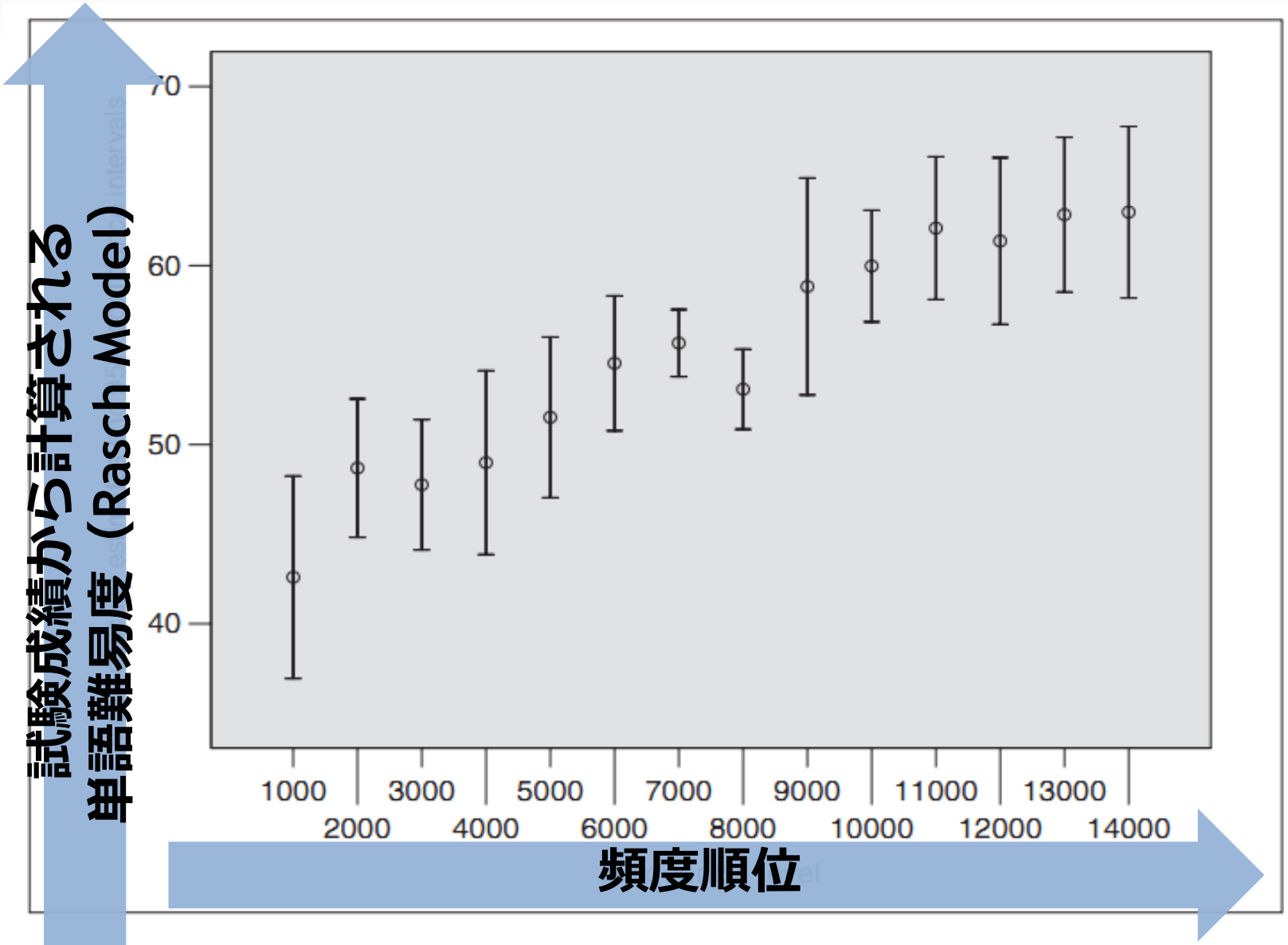
第二言語学習者は大きなコーパス中での頻度の高い順に単語を知っているのではないか？

[Beglar2010, Nation and Beglar 2007]

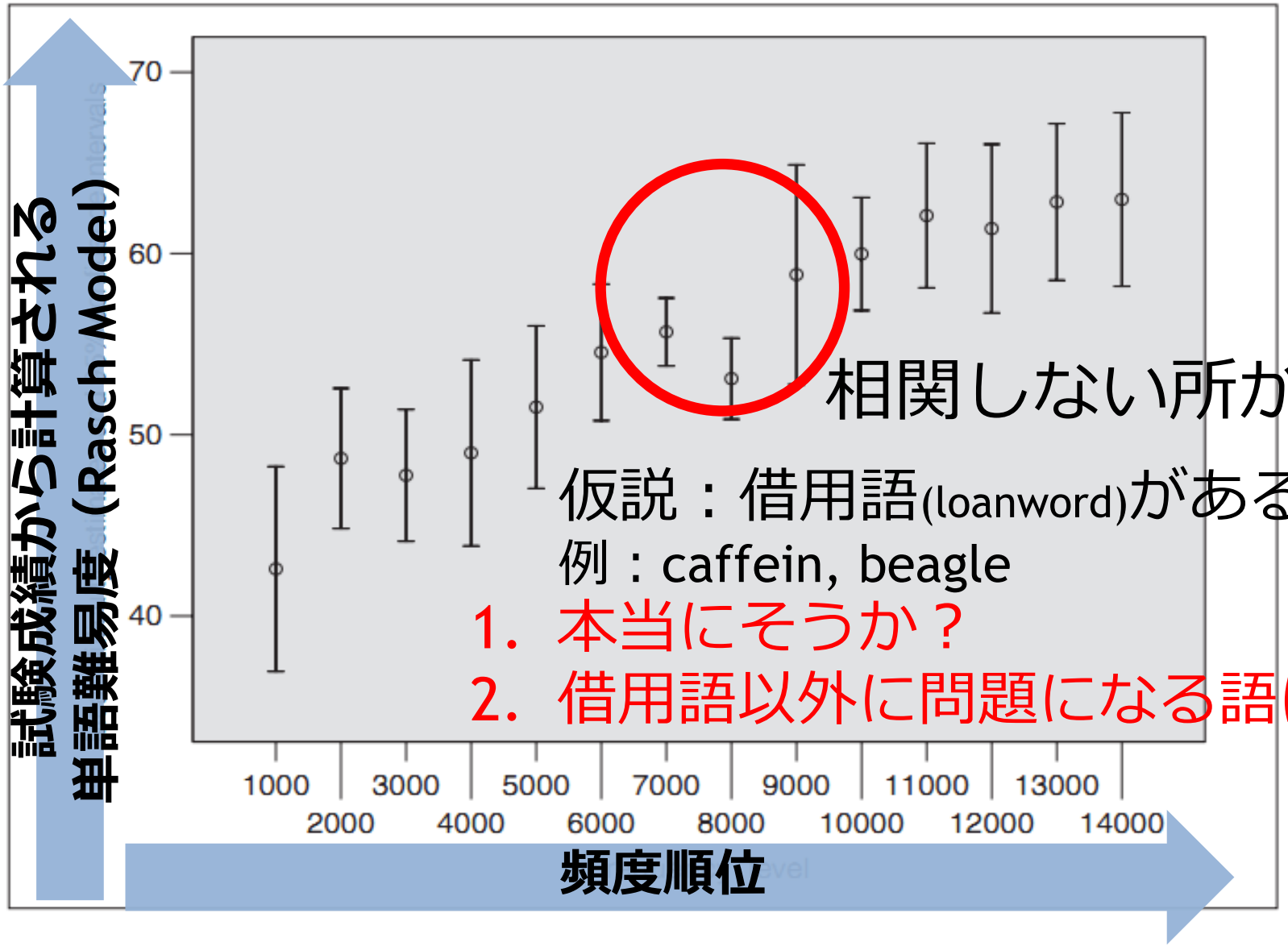
応用：

- 言語テストに使う語の選択

既存研究 [Beglar 2010] 1/2



既存研究 [Beglar 2010] 2/2



相関しない所がある

仮説：借用語(loanword)があるから

例：caffein, beagle

1. 本当にそうか？
2. 借用語以外に問題になる語は？

研究の比較

	[Beglar 2010]	本研究
目的	Vocabulary Size Testの評価	語彙予測タスクの評価（個々の単語を当てる）
語彙数	140語	11,999語
手法	多選択式	自己申告式
被験者数	197人	15人

crab: Do you like <crabs>?

a very thin small cakes
b tight, hard collars
c sea creatures that always walk to one side
d large black insects that sing at night

crab
1. never seen the word before
2. probably seen the word before
3. absolutely seen the word before but does not know its meaning
4. probably know the word's meaning
5. absolutely know the word's meaning

語彙予測タスク

学習者の語彙の一部から:

(既知, Satoshi, “*dog*”), (未知, Koji, “*dwindle*”),
(既知, Koji, “*cat*”), (既知, Satoshi, “*swap*”)

} Raschモデルはここの分析

残りの語彙知識を予測:

Satoshiは“*worship*”を知っているか?

Kojiは“*swap*”を知っているか?

応用：

分析：頻度と難易度が相関しにくい語を探す

→そのような語は言語テストで使わない

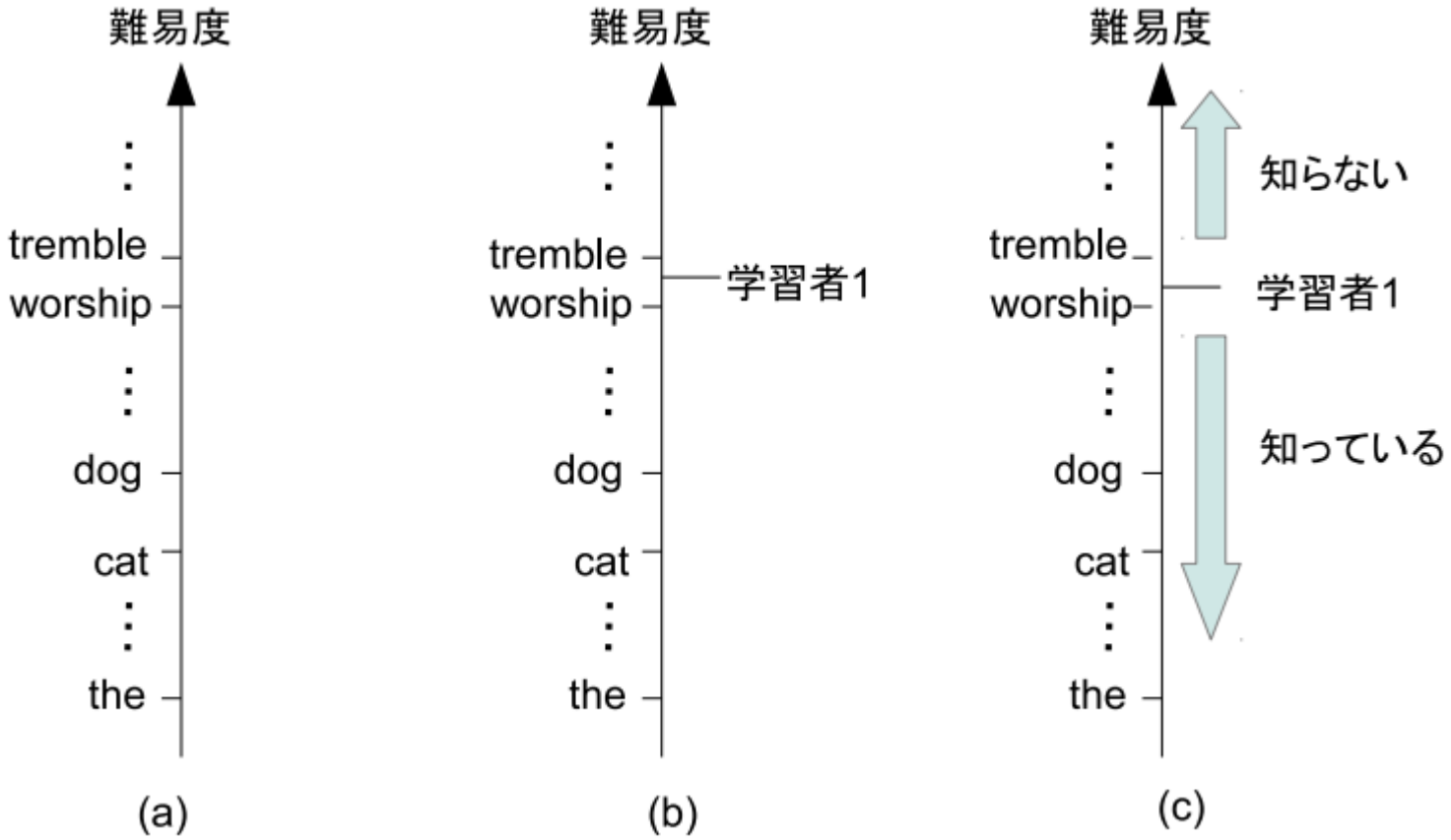
予測：読解支援システム [Ehara+, 2013, Ehara+, 2010]

（学習者の分からなさそうな単語に自動的に訳を付ける）

Raschモデル

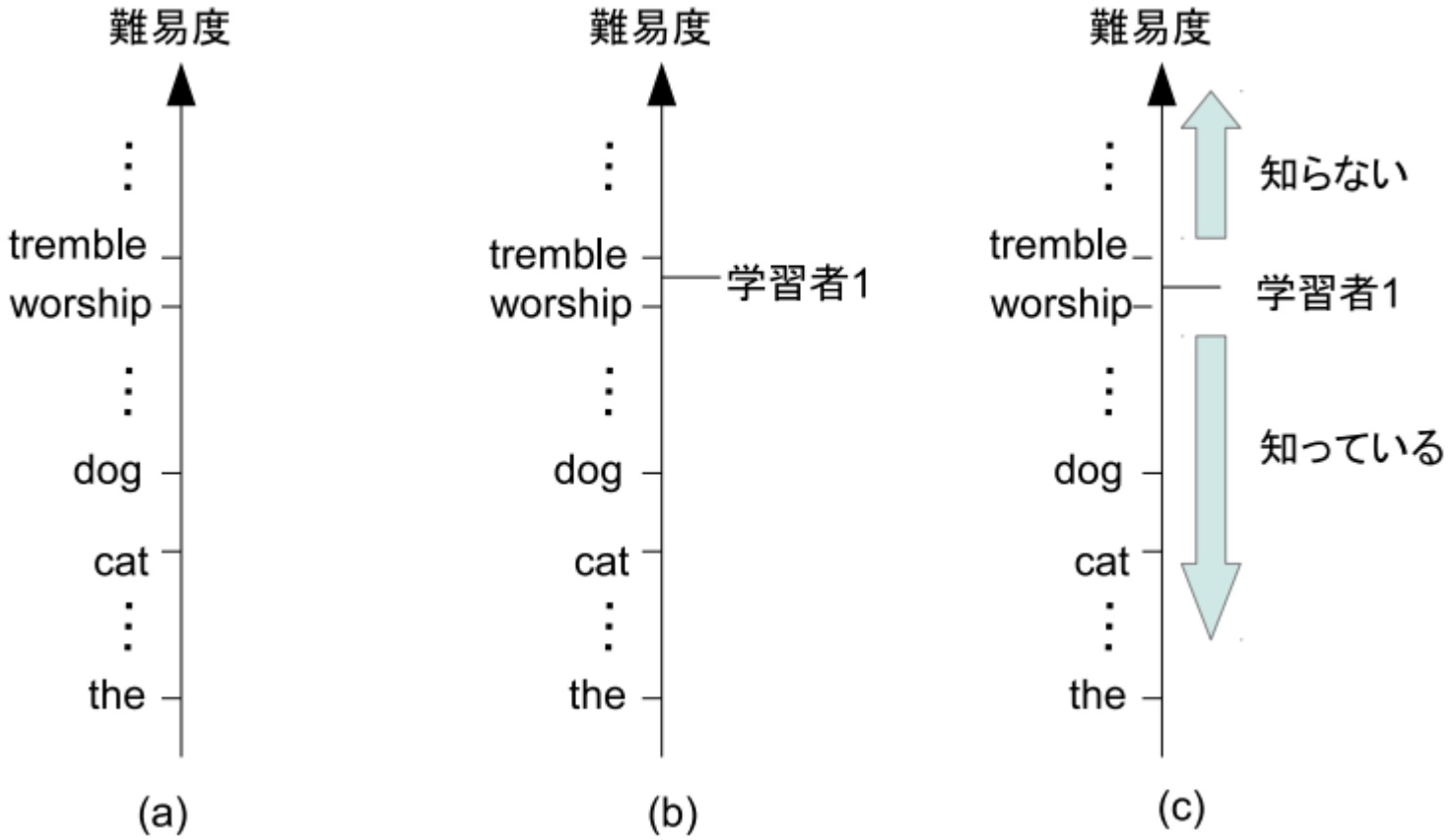
$$P(y=1|u,v) = \sigma(a_u - d_v)$$

a_u: 学習者uの語彙力
d_v: 単語vの難易度
σ: シグモイド関数



Raschモデルの問題点

“worship”は知っているが “tremble”を
知らない学習者を表現出来ない！
全学習者に取って単語難易度は同じなため.



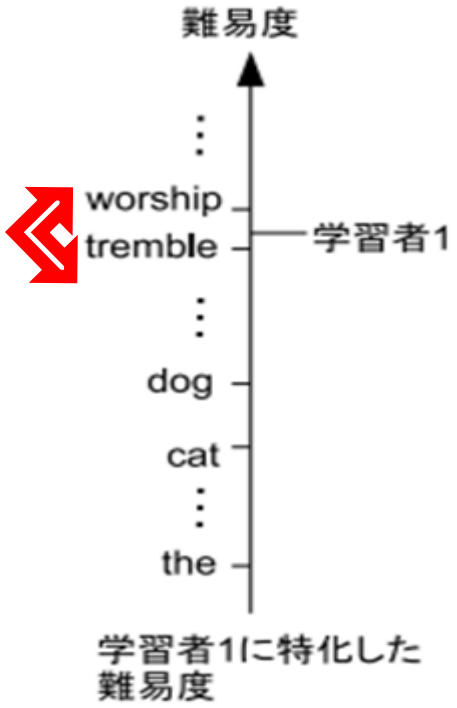
提案モデル：学習者に適応する難易度

$$P(y=1|u,v) = \sigma(a_u - f(u,v))$$

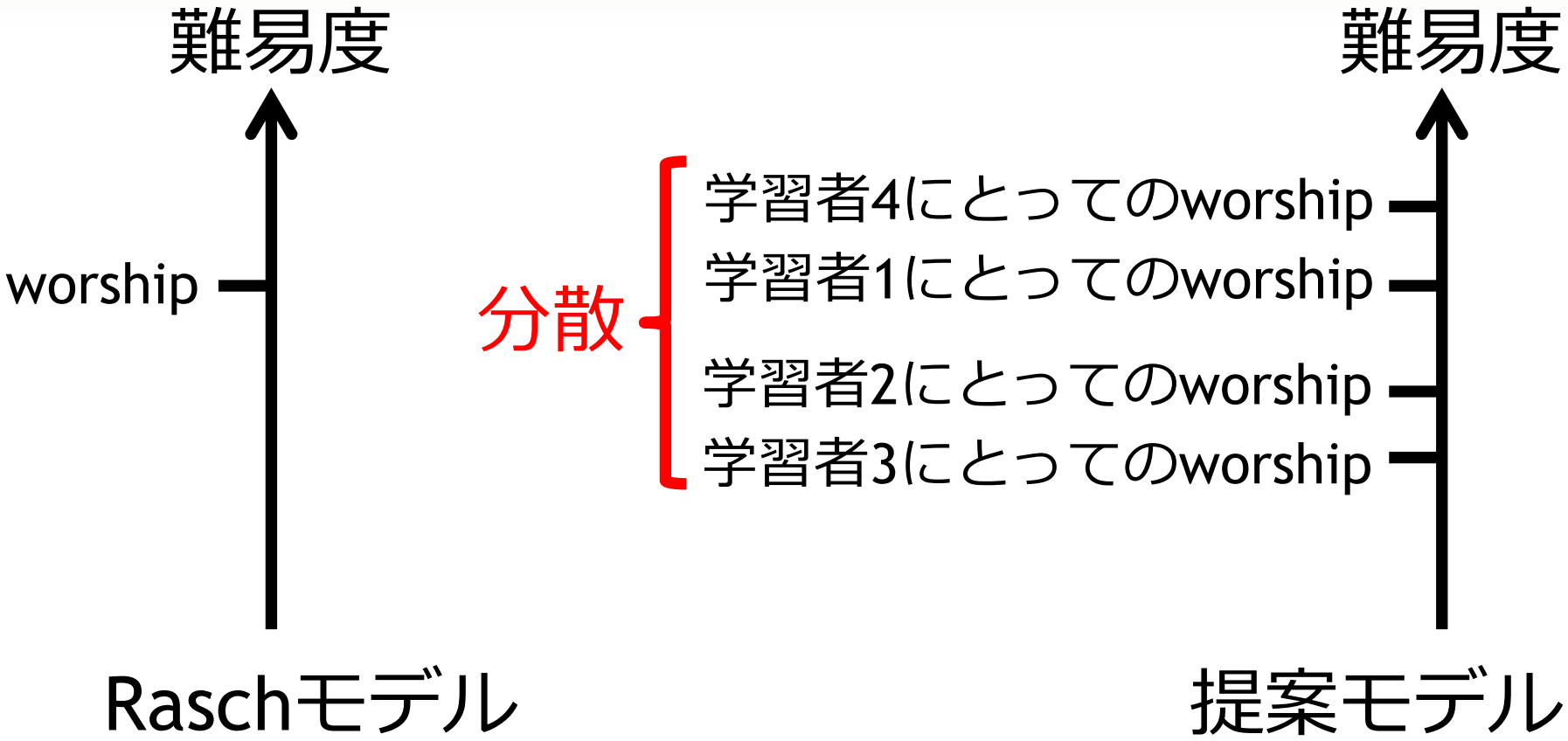
$$f(u,v) = \mathbf{w}_u^\top \phi(v)$$

学習者uにとっての単語vの難易度

BNCでのvの頻度，COCAでのvの頻度



学習者ごとに適応させると，単語ごとに分散が分かる！



分散の値：頻度と相関しない度合いの量的な指標

全体的には難しい語なのに低い語彙力の
学習者の一部が知っていたりすると分散が大きくなる

結果：難易度の分散の大きい順

$Var(v)$	語	学習者特異性が高い理由
0.993	twitter	商品名
0.886	waltz	ドメイン依存: 音楽，母語で借用語
0.849	kindle	商品名
0.833	rink	母語で “link” と同音語
0.827	launder	母語で借用語
0.825	bass	ドメイン依存: 音楽
0.823	ultraviolet	ドメイン依存：化粧品
0.818	chime	ドメイン依存: 音楽
0.804	asphalt	母語で借用語
0.802	harry	母語で “hurry” と同音語

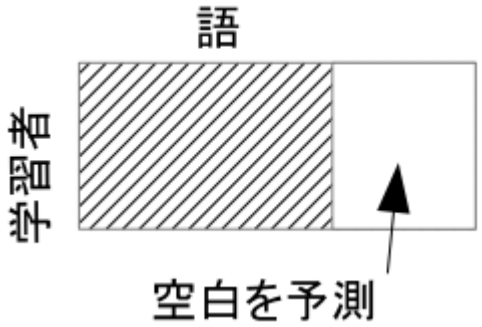
難易度の分散が大きい理由

- 商品名
- 母語で借用語
- 母語で同音語 ⇨ 自己申告式で誤って知っていると答えてしまう場合の一部が取れている.
  - rinkに対するlinkなど，より簡単な語が母語で同音の場合には，単語を「知っている」と答えてしまう.
- ドメイン依存
  - 学習者の平均的な能力が低くとも，学習者がある特定のドメインの知識が豊富な場合（例：音楽），そのドメインの語だけは知っている事がありうる.

予測精度：

訓練データに「ない」単語に対して，ある学習者がその単語を知っているか？（２値判別）

Raschモデル　66.32%
提案モデル　77.81%



- Raschモデルは分析には使えるが，予測能力は低い.
- 提案モデルは，分析の幅も広がる上，予測能力も高い.

まとめ：

- 既存：
  - 頻度と，試験から計算される難易度が相関しない単語が存在する.
    - 理由：借用語のせいだろう（未検証）
    - 個々の単語について，「相関しない度合い」の量的な指標は与えられていなかった.
- 本研究：
  - 個人に適応する単語難易度を提案し，単語ごとの難易度の分散を「相関しない度合い」の量的な指標として提案.
  - 借用語以上に，商品名も有力な候補であることが示唆された.
  - また，予測精度もRaschモデルより向上した.

ご清澄ありがとうございました

主要参考文献

- Beglar, D.
  A Rasch-based validation of the Vocabulary Size Test
  Language Testing, **2010**, *27*, 101-118
- Nation, I. & Beglar, D.
  A vocabulary size test
  The Language Teacher, **2007**, *31*, 9-13

データセット：

公開しています.

http://yoehara.com/esl-vocabulary-dataset/

項目反応理論の識別力パラメタとの違い

```
      [,1]  [,2]
 [1,] "twitter"    "0.377761"
 [2,] "launder"    "0.428807"
 [3,] "kindle"     "0.434475"
 [4,] "ultraviolet" "0.436943"
 [5,] "gadget"     "0.480944"
 [6,] "twitch"     "0.494876"
 [7,] "spoke"      "0.514651"
 [8,] "warring"    "0.537003"
 [9,] "cram"       "0.543278"
[10,] "ketchup"    "0.549178"
[11,] "curt"       "0.550045"
[12,] "spire"      "0.552409"
[13,] "bean"       "0.559044"
[14,] "mathematical" "0.562629"
[15,] "headmaster" "0.581117"
```

- 殊な場合
- ァというパラメタがあり，
- うな傾向を示す．

- **しかし識別力パラメタを求める問題は非凸であり，最適解を求めることが数値計算的に難しい．また，識別力パラメタを含めても，予測能力はRaschモデルと同程度である．**

計算上の優位性：最適解

- Rasch Modelは，項目反応理論の1母数モデル
- 項目反応理論には，2母数モデルもある．

識別力パラメータ：

「能力の高い学習者は軒並み高得点，能力の低い学習者は軒並み低得点」のように

設問が学習者の能力を分ける度合いを示すパラメタ．

しかし，この識別力パラメータを入れた２母数モデルは，非凸なので，通常，最適解は求まらない．

- しかし，提案モデルは凸なので，最適なパラメタを求められる．

Parameter Estimation

$$\mathbf{W} = \{\mathbf{w}_u | u \in U\}$$

We want to minimize:

$$l(\mathbf{W}, \mathbf{a}, \mathbf{w_0}) = \sum_{i=1}^{N} nll(y_i, u_i, v_i; \mathbf{w}_{u_i}) + \frac{\lambda}{2} \sum_{u \in U} \|\mathbf{w}_u - \mathbf{w}_0\|^2 + \frac{\eta_w}{2} \|\mathbf{w}_0\|^2 + \frac{\eta_a}{2} \sum_{u \in U} a_u^2$$

Convex for all variables (W, a, w_0)

Global optimal can be obtained.

Optimization details basically follow (Kajino+, AAAI2012)

語の定義

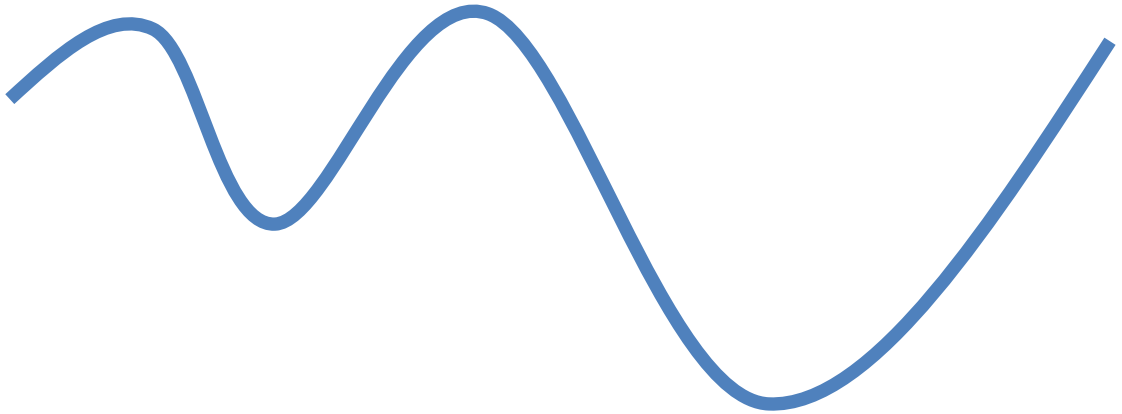
- Word family [Nation+, 2007]
  活用形のまとめ：
    study：study, studying, studied
  意味のまとめ：
    bank： 銀行，土手．区別せずにカウント．どちらか１つの意味を知っていれば良い.
  SVL12000も，word familyより少し活用形について細かいものの，同様なまとめ方をしている.

項目反応理論の識別力パラメタとの違い

- Raschモデルは項目反応理論の特殊な場合
- 項目反応理論には識別力パラメタというパラメタがあり，このパラメタは提案法と似たような傾向を示す．
- **しかし識別力パラメタを求める問題は非凸であり，最適解を求めることが数値計算的に難しい．**
  - **実際に，Rのirtoysを用いた所，11,999語については一度に求められず，1,000語までしか計算できなかった．**
  - **初期値依存，近似解などの問題**
- **また，識別力パラメタを含めても，予測能力はRaschモデルと同程度である．**

凸性

非凸：



凸：