# 語彙体系における巨視的変数について

高村大也 takamura@pi.titech.ac.jp

## 言語の成立条件

**効率性**

「あぶない！」が
「あぶらかだぶらいらどああびららいすえがりもい！」だったら大変なことになる.

**誤解しにくさ**

「冷酒」が「あつかむ」だったら,
熱燗を頼んだのに冷酒が来るという居酒屋トラブルが大量発生するだろう.

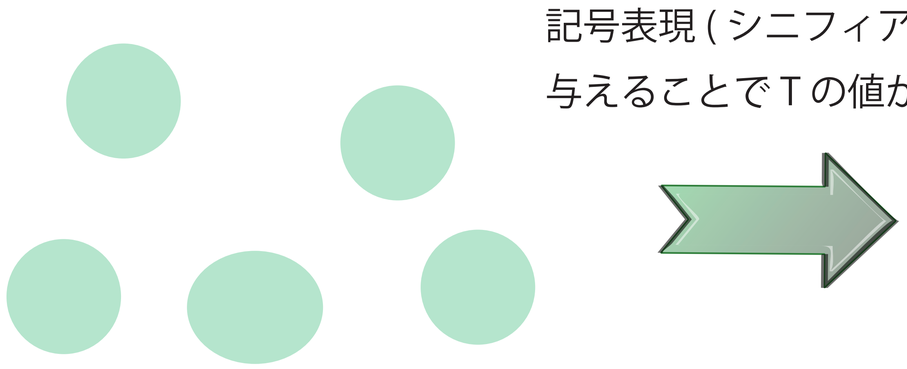
## 基本モデル

これを定量化してみよう

小さい方が良い

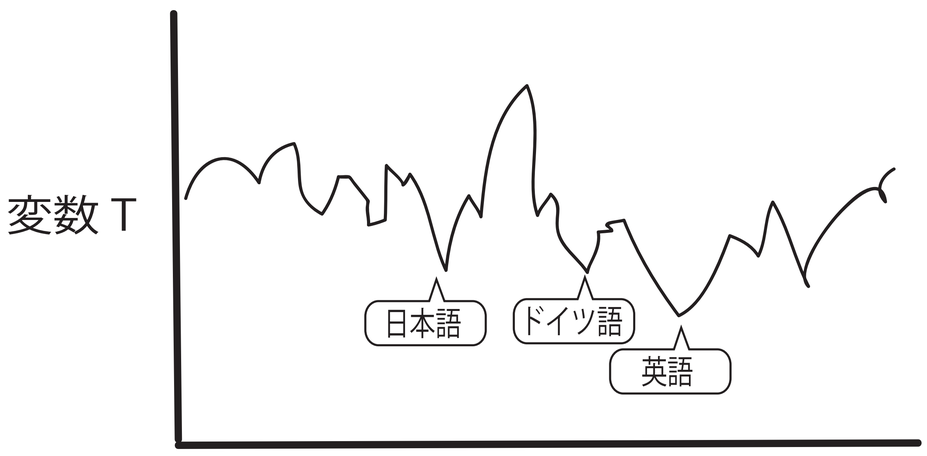
変数 T = (単語の平均長さ) + (誤解しやすさ)

記号表現 ( シニフィアン ) を
与えることで T の値が決まる




モノや概念 ( シニフィエ )

モノや概念 ( シニフィエ )




モノや概念集合への
記号表現の与え方

逆に言うと，T が小さくなるように全体の記号表現を
決めれば，言語らしきものが生成できるのでは？

## 具体化

(単語の平均長さ ) = Σ 頻度 (s)× 長さ (s)

(誤解しやすさ ) = Σ 頻度 (s)×(Σ 共起 (s, s′ )× 類似度 (w, w′ ))

s, s′ はシニフィエ,
w, w′ は対応する記号表現

## 語彙体系生成実験

### 実験設定

頻度 (s)：s のコーパス内での出現頻度 ( 最尤推定 )
共起 (s, s′ )：s と s′ がコーパス内で同じ文に出現する確率 ( 最尤推定 )
類似度 (w,w′ )：w と w′ の音素列としての近さ.
音素間の近さを与えて，DP マッチングで計算
コーパス：WSJ corpus. ただし，頻度 1000 以上の単語のみ.

### アルゴリズム

初期化：各シニフィエにランダムに母音を与える.
以下を繰り返す.
各シニフィエ s に対して
挿入：w の各位置に各音素を挿入したときの T の値を計算.
その最小値が古い T より小さければ採用.
置換：w の各音素を異なる音素に置換したときの T の値を計算.
その最小値が古い T より小さければ採用.

※ ただし，母音の連続，同音の連続は禁止する.

### 結果

元の単語と生成された記号表現の一部

| | | |
|---|---|---|
| the u | for ekekce | million hfaha |
| of e | have hfaf | its nlayaya |
| be u | it e | but aymabm |
| to or | on caymamb | he e |
| a rolo | at ayamabm | year ha |
| in akh | as mjay | company fhahajy |
| and afp | by lmam | an avmam |
| say alanj | from ayayg | will hmarnar |
| that ykehk | with eyeyg | market ycaya |

### 関連研究

Joshua Plotkin and Martin Nowak, “Language Evolution and Information Theory” , Journal of Theoretical Biology, vol. 205, pp. 147-159, (2000).

### 研究の意義

言語処理は言語学へこういう形で貢献できないか？
「計算言語学」の進むべき方向性？？