# IT による日本語分析「目で見てわかる日本の歌ことば」

山元啓史 (東京工業大学) yamagen@ryu.titech.ac.jp

## コノテーションって定義できるの？

Octopus

weird
spooky ... VS

たこやき
(fried octopus!)
= おいしい！

‘コノテーション’ は受け手に依存する意味である。

## 比較対象: 立田 と 吉野

「うたまくら」 とは:
1. 地名
2. 歌にちなんだ特別な意味

→ どうちがう？！

## 材料

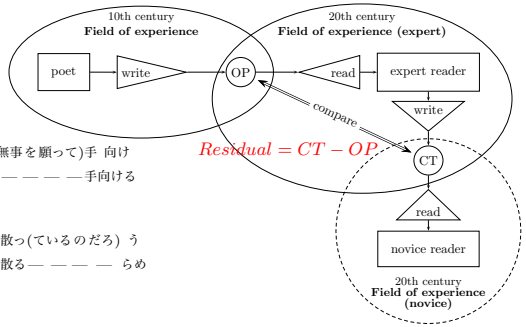
### 古今和歌集 (905 年頃)

はじめての勅撰和歌集（ちょくせんわかしゅう：「勅撰」とは「天皇の命令で選ばれた」という意味）

## OP(元歌) と CT(現代語訳) との関係

シュラムのコミュニケーション理論による。

[発信元]→ [エンコード]→ (信号)→ [デコード]→ [受信先]




## OP と CT を並べて比べると？

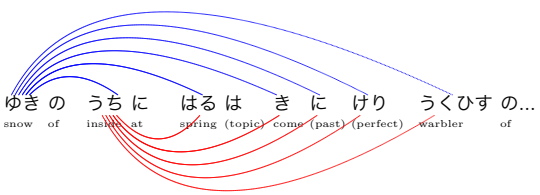
| | |
|---|---|
| CT | (秋の 末近くなって帰り道についた)龍田姫(が道中の無事を願って)手 向け |
| OP | ———— — — — — — —立田姫— — — — — — —手向ける |

| | |
|---|---|
| CT | (をする)神があるからこそ秋の木の葉(が)幣(となって)散っ(ているのだろ) う |
| OP | — — 神のあれ ば こそ秋の木の葉[の]幣 と — —散る— — — — らめ |

小町谷照彦による 298 番歌の現代語訳 (小町谷 1982)

## 共出現パターンを作る

ひとつのテキストに出てくる任意の２つの単語の組み合わせパターン




すべての歌毎にこれを繰り返すとパターンの種類は 5,000 以上

## 共出現ウエイト

Rocchio (1971)

$$w(t,d) = (1 + \log\ tf(t,d)) \cdot idf(t)$$

$w$ を $cw$ に変形する。

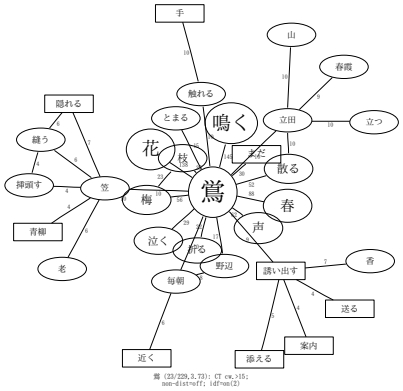
$$cw(t_1,t_2,d) = (1 + \log\ ctf(t_1,t_2,d)) \cdot \sqrt{idf(t_1) \cdot idf(t_2)}$$

document

frequency of co-occurrence

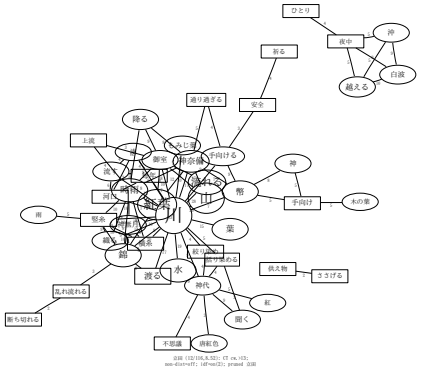
geometric mean of two weights

## グラフの例「うぐいす」




鶯 (23/229,3.73); CT cw.>15;
non-dist=off; idf=on(2)

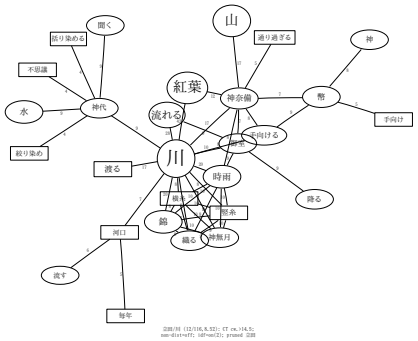
## 「立田」のモデル（刈り込み後）




立田 (12/116,8.52); CT cw.>13;
non-dist=off; idf=on(2); pruned 立田

KKS 994
風吹けば　沖つ白波　立田山
夜半にや君が　一人越ゆらむ

## 「立田川」のモデル

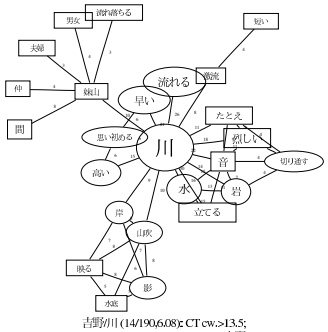



立田川 (12/116,8.52); CT cw.>14.5;
non-dist=off; idf=on(2); pruned 立田

KKS 294
たつたひめ
たむくるかみの
あればこそ
あきのこのはの
ぬさとちるらめ

KKS 300
かむなびの
やまをすぎゆく
あきなれば
たつたがはにぞ
ぬさはたむくる

## 「吉野川」のモデル

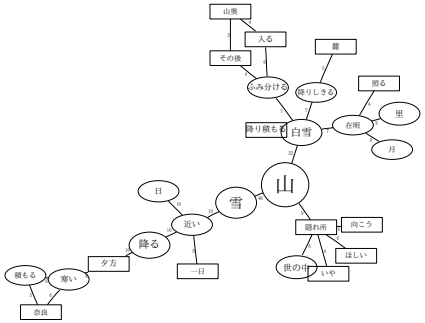



吉野川 (14/190,6.08); CT cw.>13.5;
non-dist=off; idf=on(2); pruned 吉野

KKS 828
ながれては
いもせのやまの
なかにおつる
よしののかはの
よしやよのなか

KKS 471
よしのがは
いはなみたかく
ゆくみづの
はやくぞひとを
おもひそめてし

## 「吉野山」のモデル




吉野山 (23/193,6.68); CT cw.>13.5;
non-dist=off; idf=on(2); pruned 吉野, 麓吉野

KKS 950
みよしのの
やまのあなたに
やどもがな
よのうきときの
かくれがにせむ

「かくれが」は世の中を厭い隠れるための場所

## おわりに

- 共出現パターンによる可視化をおこなった。
- ことばの使われ方によって、クラスターが明確に見られた。例）白波。
- 数学的操作が和歌のモデルにも適用できた。e.g. 枝の刈り込み, 集合演算（差分）
- 元歌と現代語訳を使ってコノテーションが示せた。例） 神の世 → 立田、世俗 → 吉野
- 相対的な卓立を見せることができた。
- 同じ語の異なる時代の意味の違いがわかるだろう。例）10 世紀と 13 世紀の「吉野桜」
- **Email**: yamagen@ryu.titech.ac.jp