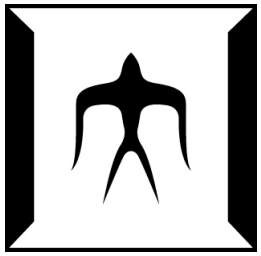
やまとことば記述の言語学

Linguistics for the description of Yamato-Japanese



山 元 啓 史

Hilofumi Yamamoto *Ph. D*

東京工業大学 / カリフォルニア大学サンディエゴ校

Tokyo Institute of Technology / University of California, San Diego

5 December 2015

自己紹介

- 山元啓史 (やまもとひろふみ) という人。
  1. 「語彙のモデリングと言語の歴史的変化」(2013 シンポジウム)
  2. 「和歌から紐解く日本語と留学生」(2014 シンポジウム)
  3. 「やまとことば記述の言語学」(2015 シンポジウム)

タイトルについて

- やまとことば記述の言語学
  1. 「やまとことば」とは
     日本で生まれたとされる日本語の語句
  2. 「記述」とは
     ほんとうの言語のありさま、ふるまいを記述する。
  3. 「記述文法」の対立語とは
     「規範文法」という。別名「学校文法」ともいう。

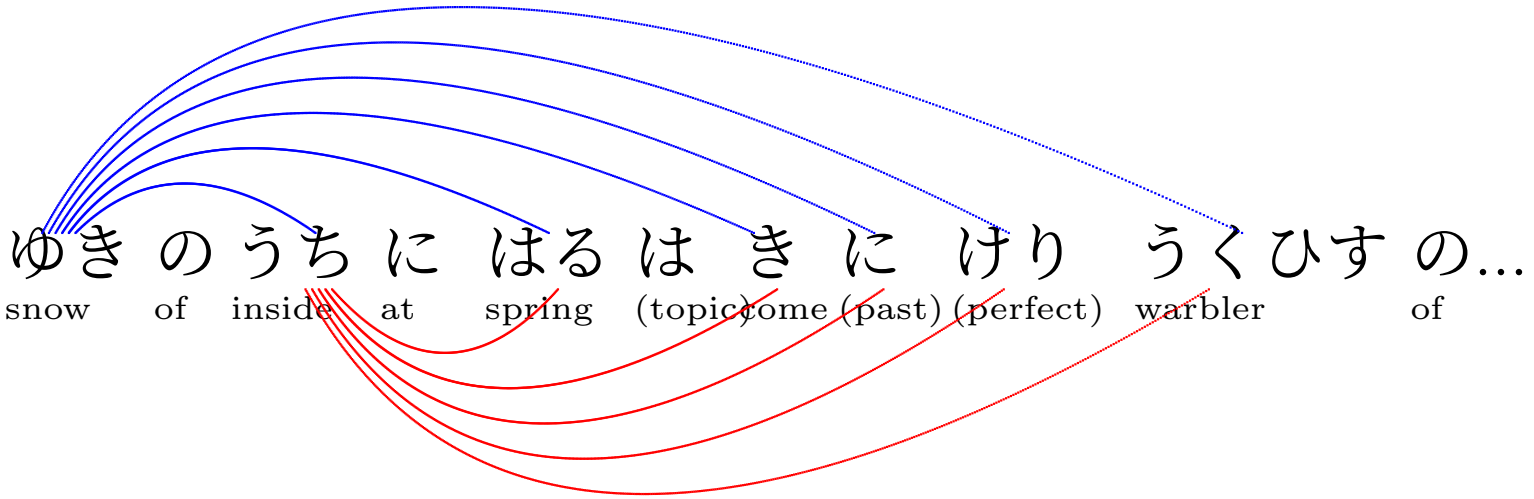
数理モデル

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

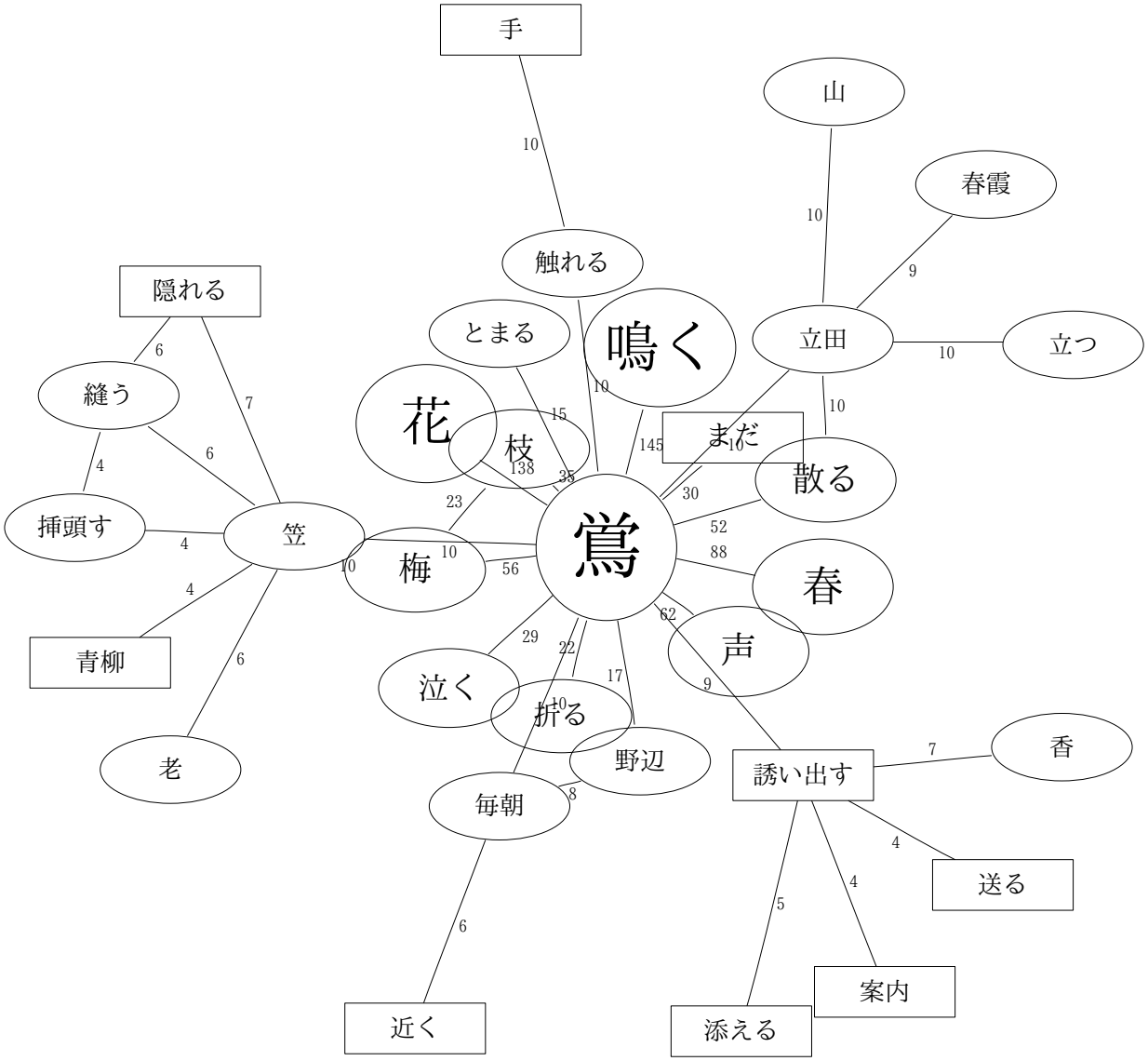
$$idf(t) = \log \frac{N}{df(t)} \qquad (2)$$

共出現パターンを作る

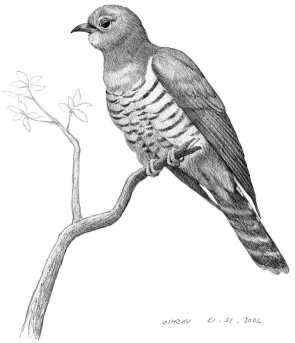
ひとつのテキストに出てくる任意の２つの単語の組み合わせパターン

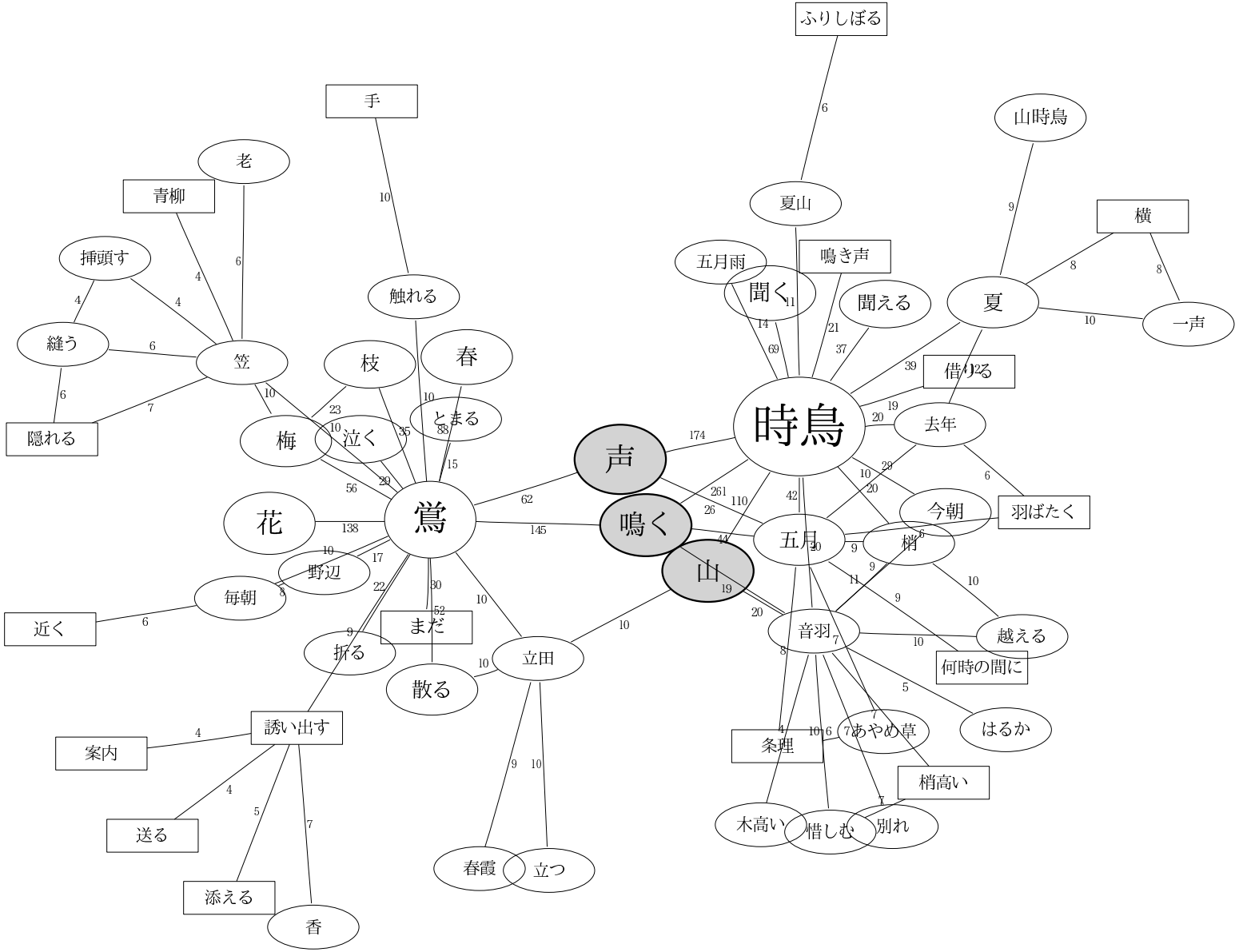


すべての歌毎にこれを繰り返すとパターンの種類は 5,000 以上



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





鴬-CT-23-229-3.73-15 時鳥-CT-40-370-3.27-16

Table 1: タグづけ済みの八代集シソーラス

01:000002:0001	A00	BG-01-4240-01-0100	袖	そで	袖 02
01:000002:0002	A00	BG-02-5130-01-2100	漬つ	ひつ	漬つ 47
01:000002:0003	A00	BG-08-0064-16-0100	て	て	て 64
01:000002:0004	A00	BG-02-1515-08-0105	掬ぶ	むすぶ	掬ぶ 47
01:000002:0005	A00	BG-09-0010-04-0200	き	き	き 74
01:000002:0006	A00	BG-01-5130-03-0201	水	みづ	水 02
01:000002:0007	A00	BG-08-0061-07-0100	の	の	の 61
01:000002:0008	A00	BG-02-5160-01-0101	凍る	こほる	凍る 47
01:000002:0009	A00	BG-09-0010-03-0300	り	り	り 74
01:000002:0010	A00	BG-08-0061-10-0100	を	を	を 61
01:000002:0011	A00	BG-01-1624-02-0100	春	はる	春 02
01:000002:0012	A00	BG-02-1513-01-0100	立つ	たつ	立つ 47
01:000002:0012	A10	BG-02-1521-06-0200	立つ	たつ	立つ 47
01:000002:0012	A20	BG-02-3330-11-0200	立つ	たつ	立つ 47
01:000002:0012	A30	BG-02-3391-02-1100	立つ	たつ	立つ 47
01:000002:0013	A00	BG-01-1641-02-1100	今日	けふ	今日 02
01:000002:0014	A00	BG-08-0061-07-0100	の	の	の 61
01:000002:0015	A00	BG-01-5151-01-0100	風	かぜ	風 02
01:000002:0016	A00	BG-08-0065-14-0100	や	や	や 65
01:000002:0017	A00	BG-02-1550-05-0200	解く	とく	解く 47
01:000002:0017	A10	BG-02-3060-09-0400	解く	とく	解く 47
01:000002:0018	A00	BG-09-0010-02-0100	らむ	らむ	らむ 74

さて！

研究の目的

- 年代的な言語構造の比較を行い、言語変化を明らかにすること。
- 語の対応を記載したシソーラスを作ること。これによって、...
- 語の相互関係を計算し、体系的な言語の変化の仕組みを記述すること。

問題と解決

- 「現在の意味はわかるが、元からそうであったのか？」
  →自分で求めようとしない限り、なかなかわからない。
  「筆入れ」「下駄箱」「あみだな」（現在では荷物棚）
- 「語の形」と「語の意味」は同じであるとは限らない。
  →語の形だけに頼って探してはいけない。
- 対応を計算によって、割り出す。
  →これには意味的な対応が含まれる。
  ここが重要！

研究

- 漸近的語彙対応システムの開発
- 「山吹」の常識とは？

漸近的語彙対応システムの開発

- 対応とは

  あるＡ語とＢ語は語形では違っていても、意味的に近似であるならば、同じ表現域に出現する確率は高い。

- 和歌とその現代語訳のパラレルコーパスを作る。
- それを対象に、語彙対応計算を行う。

漸近的語彙対応システムの開発

- 目的: 二十一代集 (古今和歌集（905 年頃）に始まり、新続古今和歌集（1439 年）までの 534 年間で 21 の勅撰和歌集。およそ、34,000 首のシソーラスを作ること。
- 材料: 古今和歌集とその現代語訳 10 種類

開発に関わる問題点


- 実体: 卯の花は豆腐？花？
- 単位: 卯/の/花 or 卯の花
- 意味: 頭の雪＝白髪？ 掛詞: 海松藻と見る目？
- 正書: 契りけん, 契けむ, 契けん, 契剣／思ふてふ, 思てふ, 思ふ蝶, 思蝶

ひとつずつコツコツやるか？やってもよいか？

- たくさんあるし、その時代の人じゃないし。
- 人間の判断ではゆらゆら揺れて決められない！

  →そこで、パラレルコーパス（和歌と現代語訳 10 種）を使って計算で、古語と現代語の単語対を自動的に決める！

計算方法

- Mutual Co-occurrence Rate: 村井 (2010)

$$mcr(o,t) \quad = \quad p(o\,|\,t)\; p(t\,|\,o)$$

ただし、o はオリジナル（和歌）の単語。t は現代語訳の単語。$mcr(o,t)$ は相互共出現率。$p(o|t)$ は、オリジナルと現代語訳の対応する 2 つの文に注目した時、単語 o と 単語 t が同時に出現した割合。

$\rightarrow$ mcr(o, t) の値が、十分に大きい時、単語 o と単語 t は、文脈的に一致していると推定。

結果: 古語、現代語の一致率

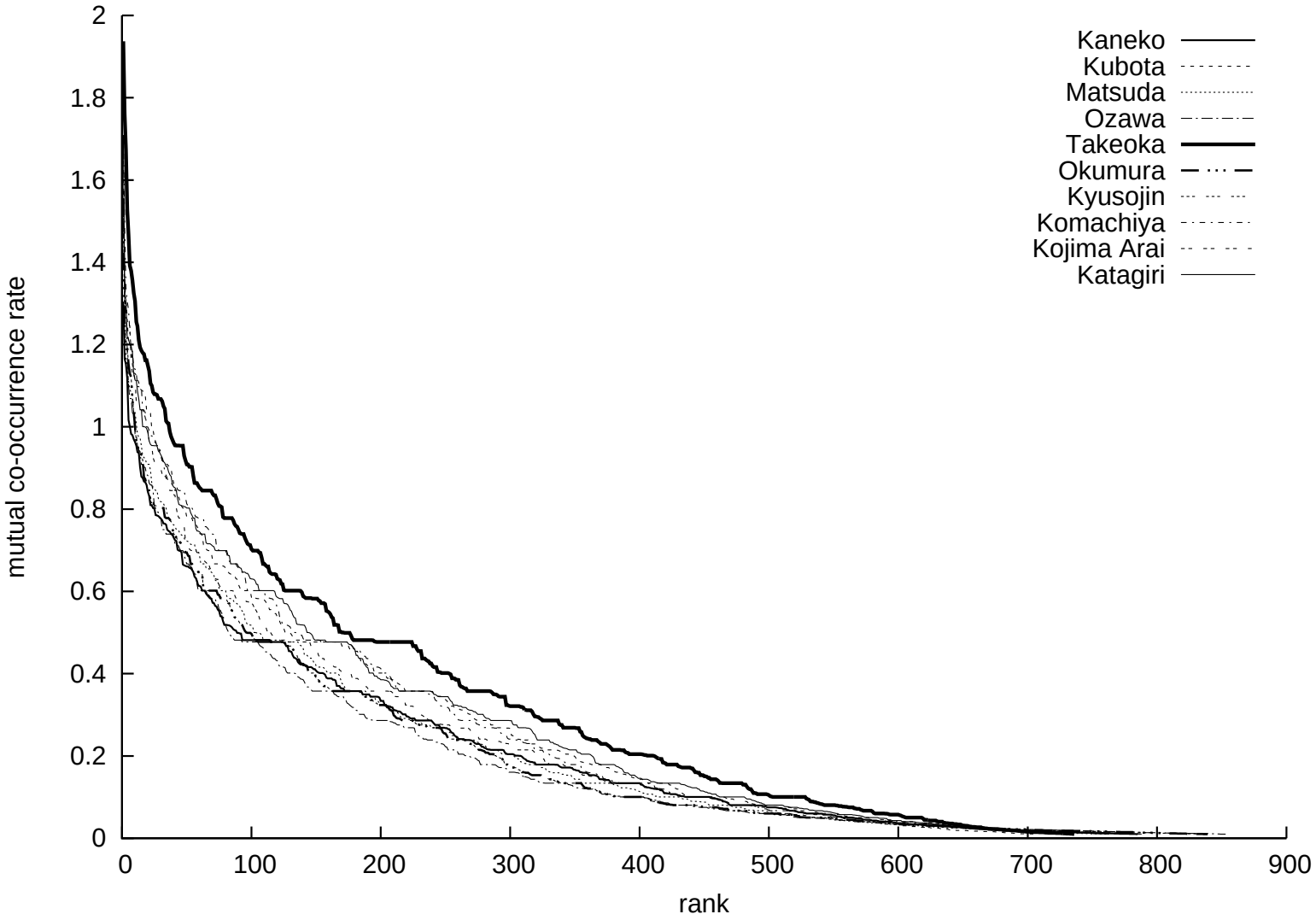


図 2:古今集と現代語訳 10 種中の任意の単語対の mcr 値の分布

結果: 推定値の上位、中位、下位ごとの単語の対応例

no.	上位対 (1.3 以上)		中位対 (0.16 より)		下位対 (0.01 以下)	
1	鳴く	鳴く	老ゆ	年老いる	異なり	あの
2	風	風	乱る	乱れる	雫	どうして
3	世の中	世の中	来	いらっしゃる	此の	この
4	人	人	問ふ	問う	随に	まま
5	春	春	問ふ	訪ねる	匂ふ	美しい
6	秋	秋	名	噂	見る	せい
7	時鳥	時鳥	変はる	変る	連れ	つく
8	時鳥	ほととぎす	燃ゆ	燃える	立ち返る	言う
9	散る	散る	濡づ	濡れる	有り	つく
10	見る	見る	難し	むずかしい	有り	まさしく

まとめ

1. 動詞:

   「落つ→落ちる（自動詞）」「捨つ→捨てる（他動詞）」

2. 名詞:

   「木綿付鳥（ゆふつけどり）→鶏」「朝な朝な→毎朝」

3. 古語固有:

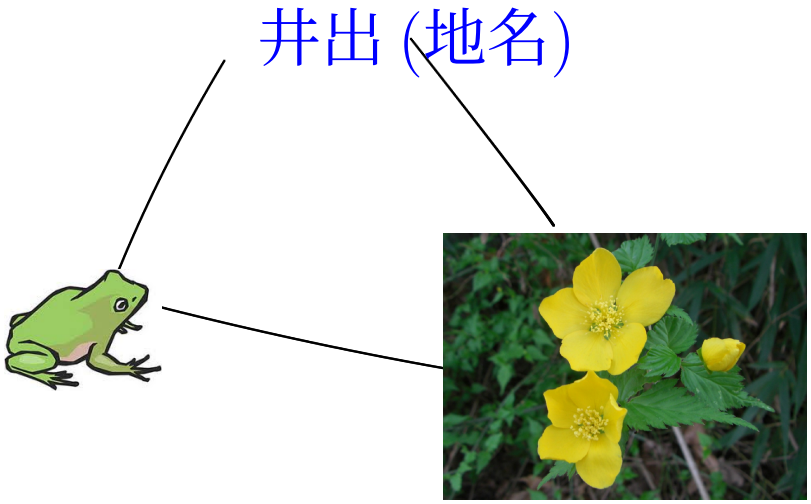
   機械的に、古語「名」に対して「噂」が推定できた。

4. 推定範囲:

   単語対の推定が可能なのは 0.2 あたりまで。

これを応用する

- 漸近的語彙対応システムで語彙対が得られる。
- 語彙対をシソーラスに登録する。
- シソーラスにより、やまとことばの体系を可視化する。
- 発見があるかどうか検討する。



「山吹」の常識とは？

- 井出は京都の地名で和歌では有名。
- 井出-蛙-山吹は和歌の世界の 黄金率 として知られる。

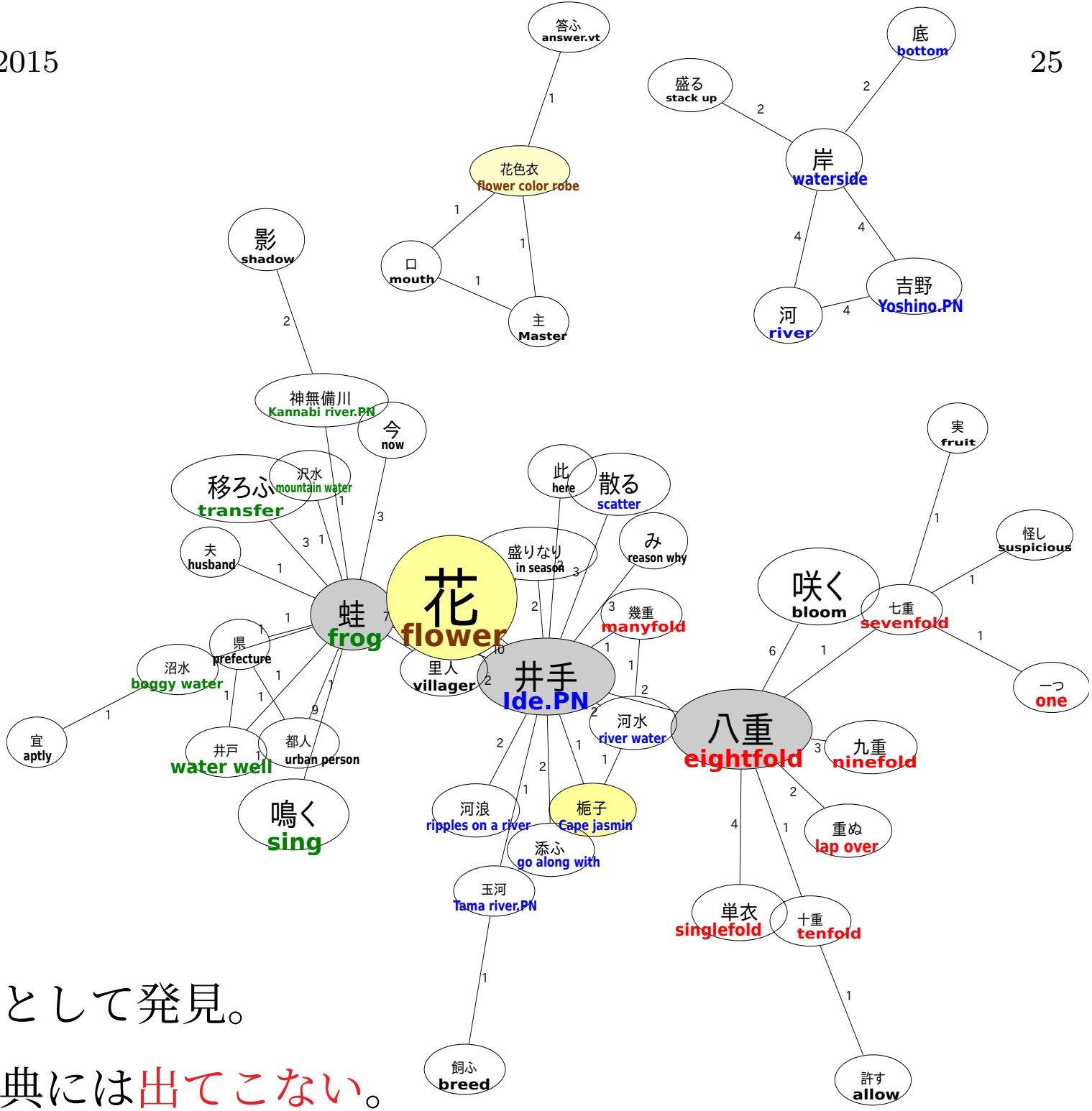
「蛙と山吹」歌川広重

「山吹」の常識は本当か？

和歌、俳諧、絵画など

かなり強い関係のようだ。

これが本当ならば、
可視化できるはず！



- 八重をハブノードとして発見。
- これは歌ことば事典には出てこない。

まとめ

1. 項目を執筆する研究者が意識しないと辞書に記載されない。
2. コンピュータでは、規則にしたがう限り、関係を網羅できる。
3. この方法は「山吹」に限られた方法論ではない。

おわりに

1. 日本語史上の発見は多いであろう。
2. 日本語の歴史を計算手続きでたどるのは難しくない。
3. この記述方法は日本語だけの方法論ではない。

質問

- 和歌の数理モデル・東工大への留学

  http://warbler.ryu.titech.ac.jp/~yamagen/

  をご覧ください。

- その他ご質問については:

  山元啓史 Hilofumi Yamamoto までお気軽にどうぞ。

  yamagen@ryu.titech.ac.jp