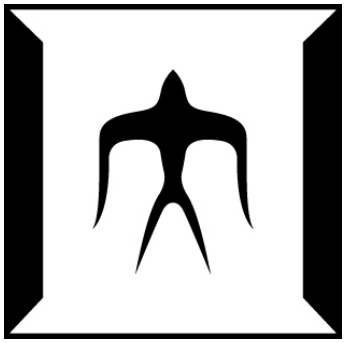
通時コーパスによる言語の歴史的変遷

山 元 啓 史

yamagen@ila.titech.ac.jp



東京工業大学

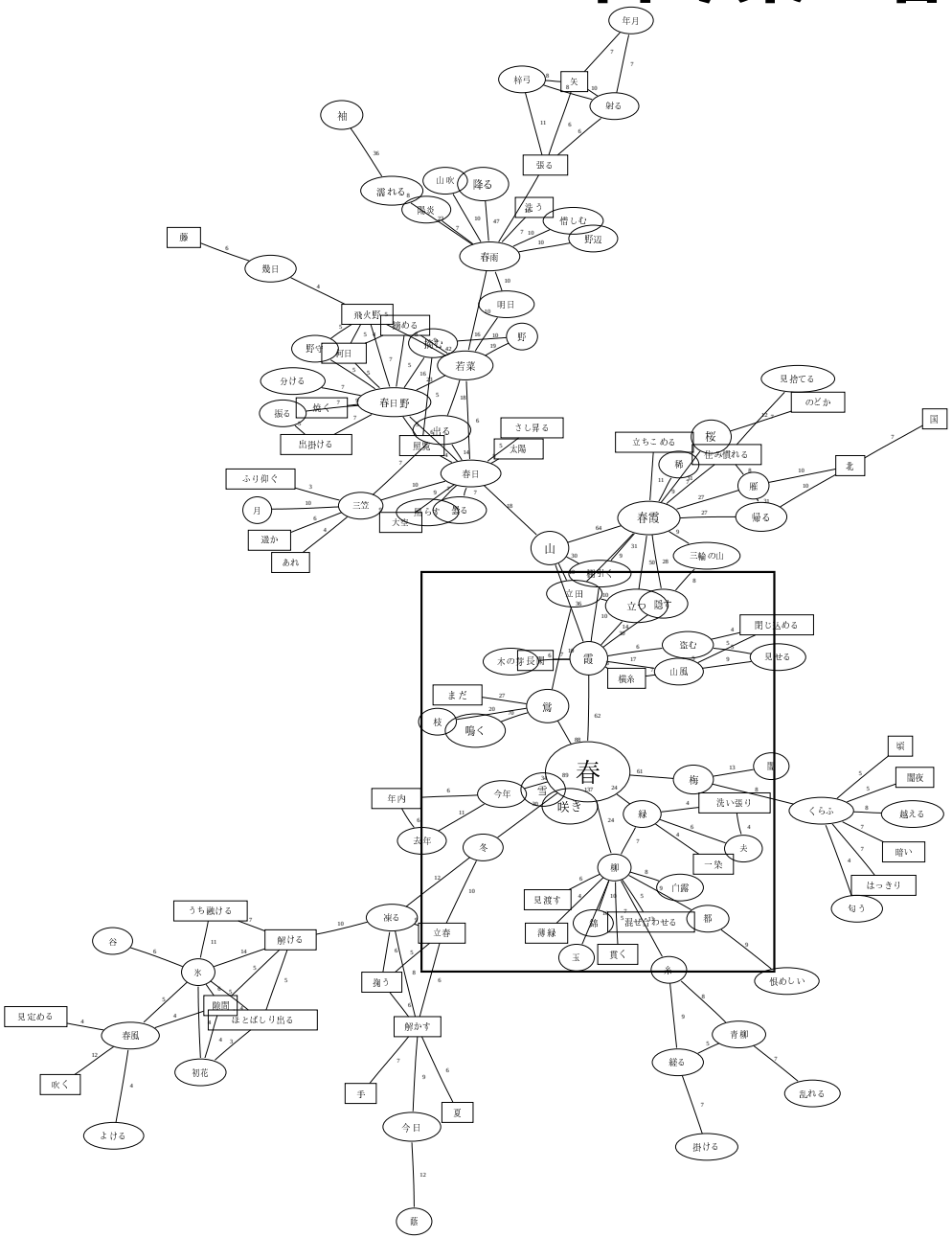
カリフォルニア大学サンディエゴ校

June 8, 2016

はじめまして

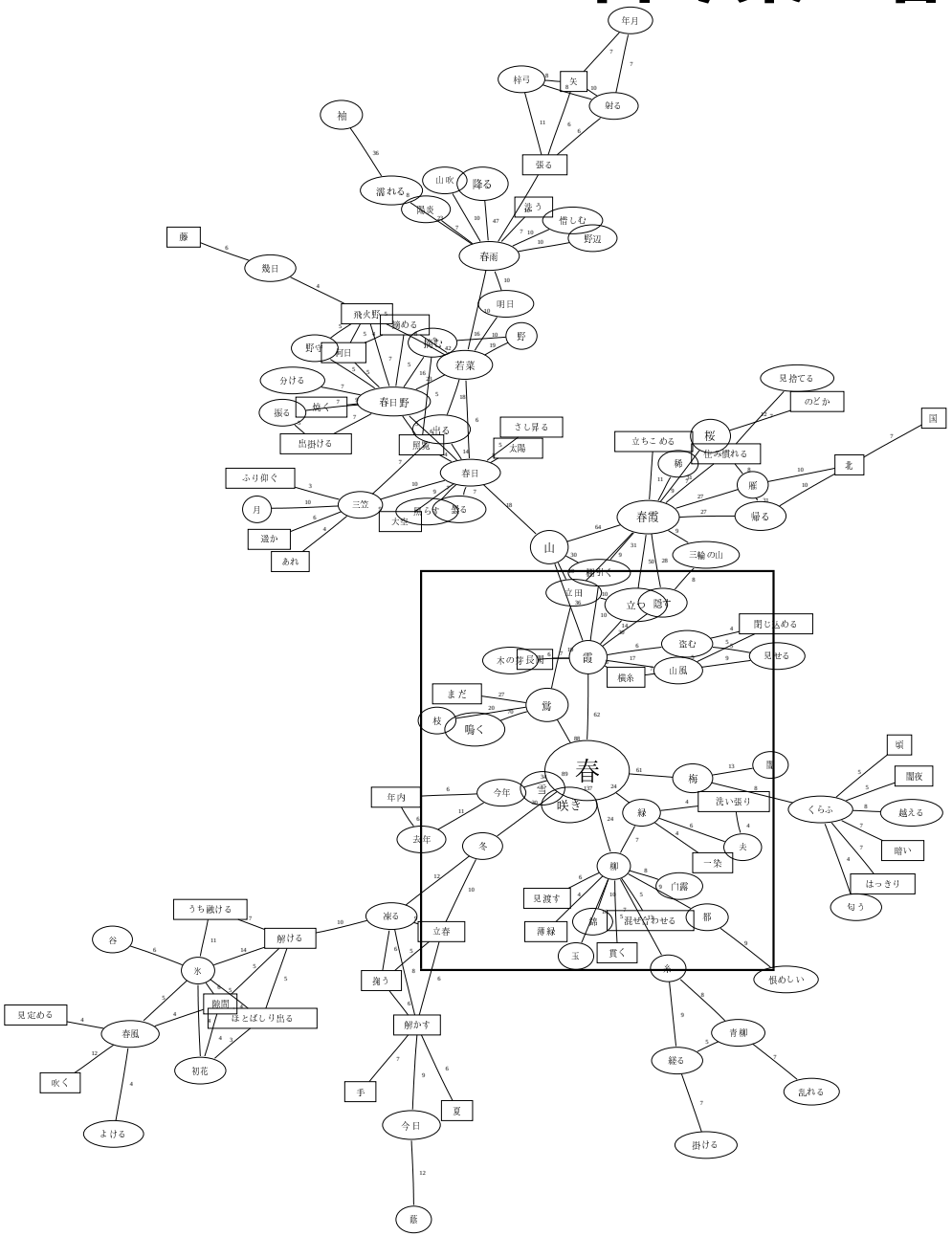
- 山元啓史（やまもとひろふみ）
- オーストラリア国立大学（言語学博士）
- 言語の変化体系を研究している。
  →具体的にどんなことかというと。

古今集「春」のモデル

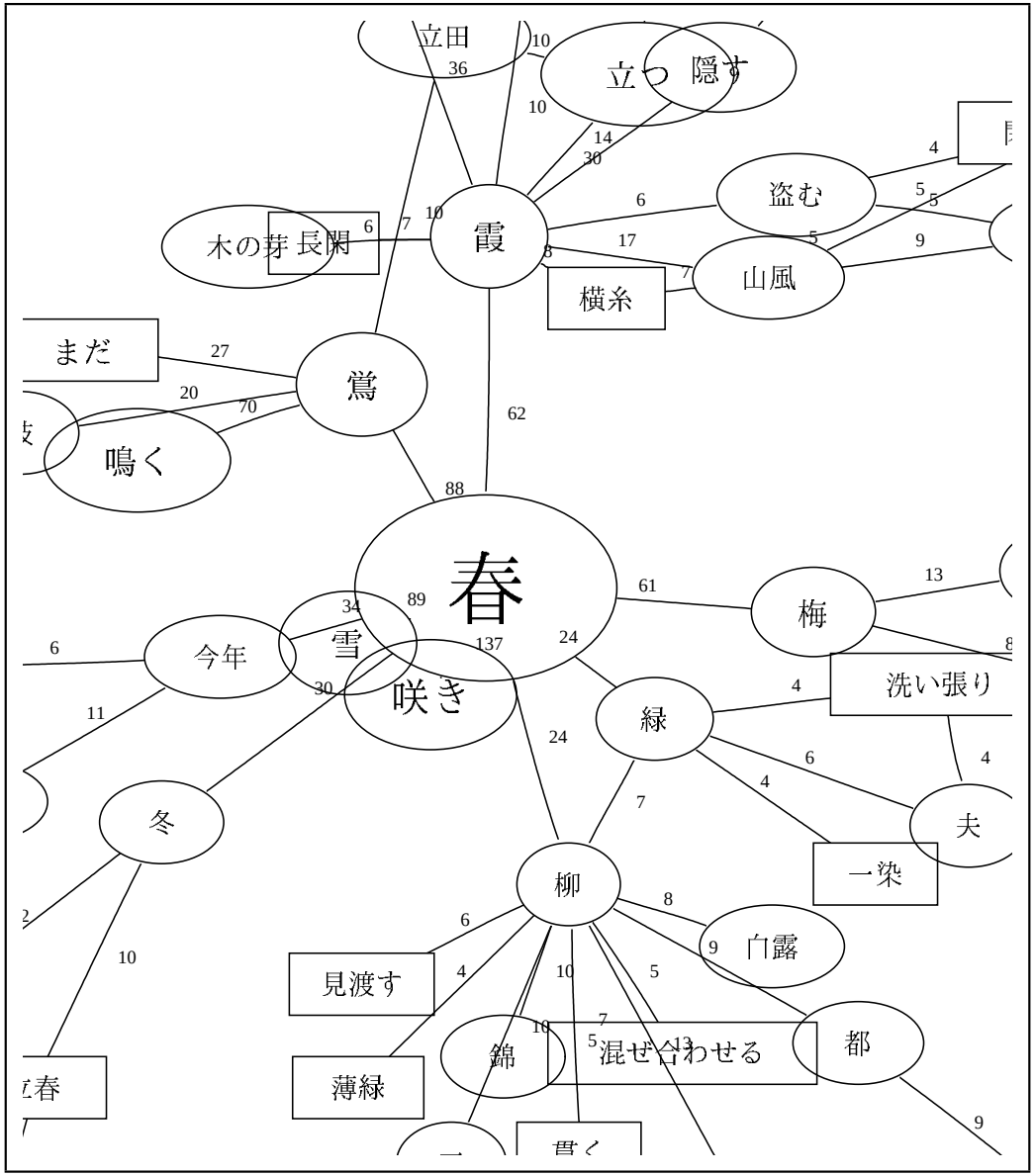


春 (90/652,2.68): CT cw.>15.6;
non-dist=off; idf=on(2)

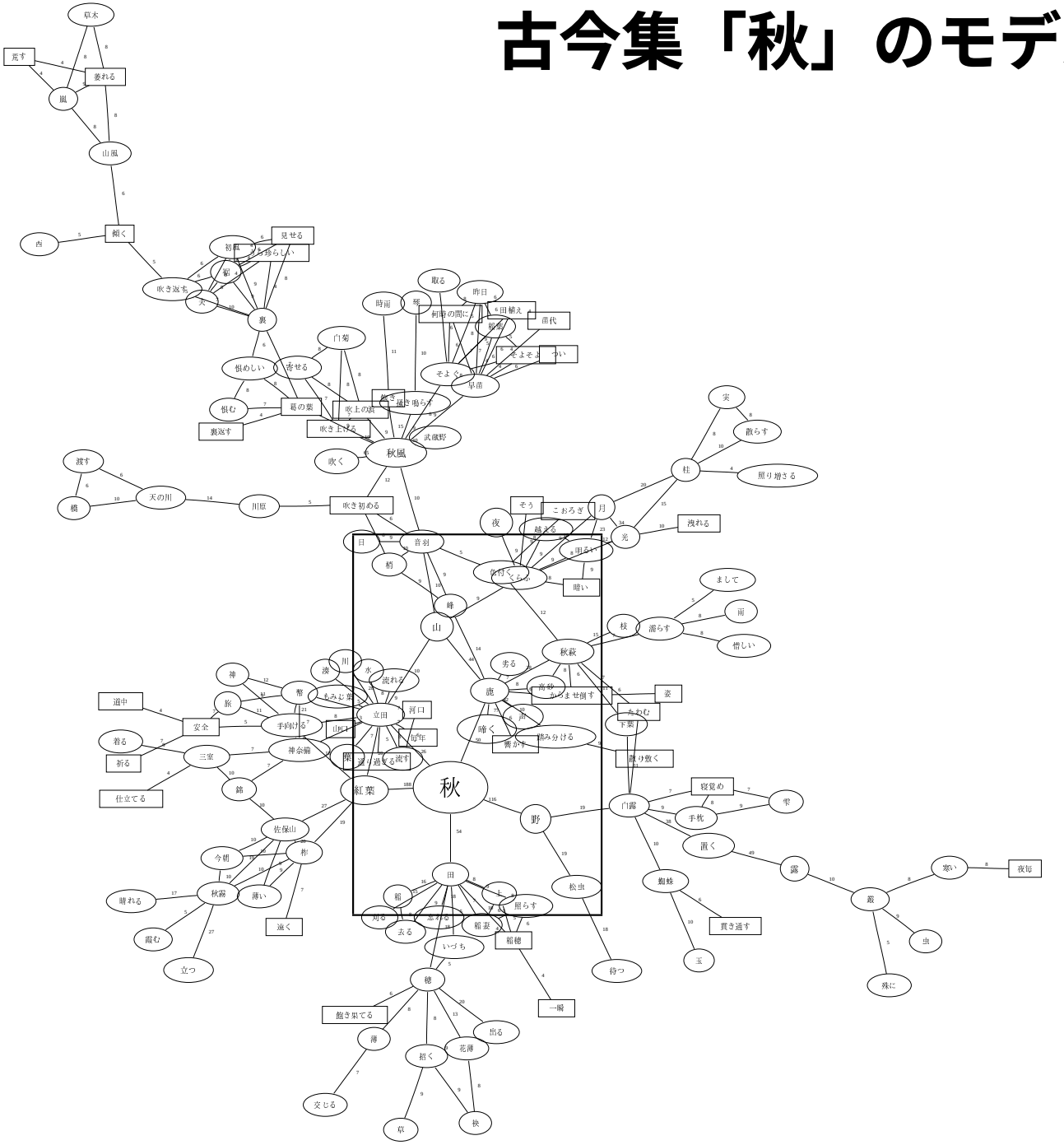
古今集「春」のモデル



春 (90/652,2.68): CT cw.>15.6;
non-dist=off; idf=on(2)

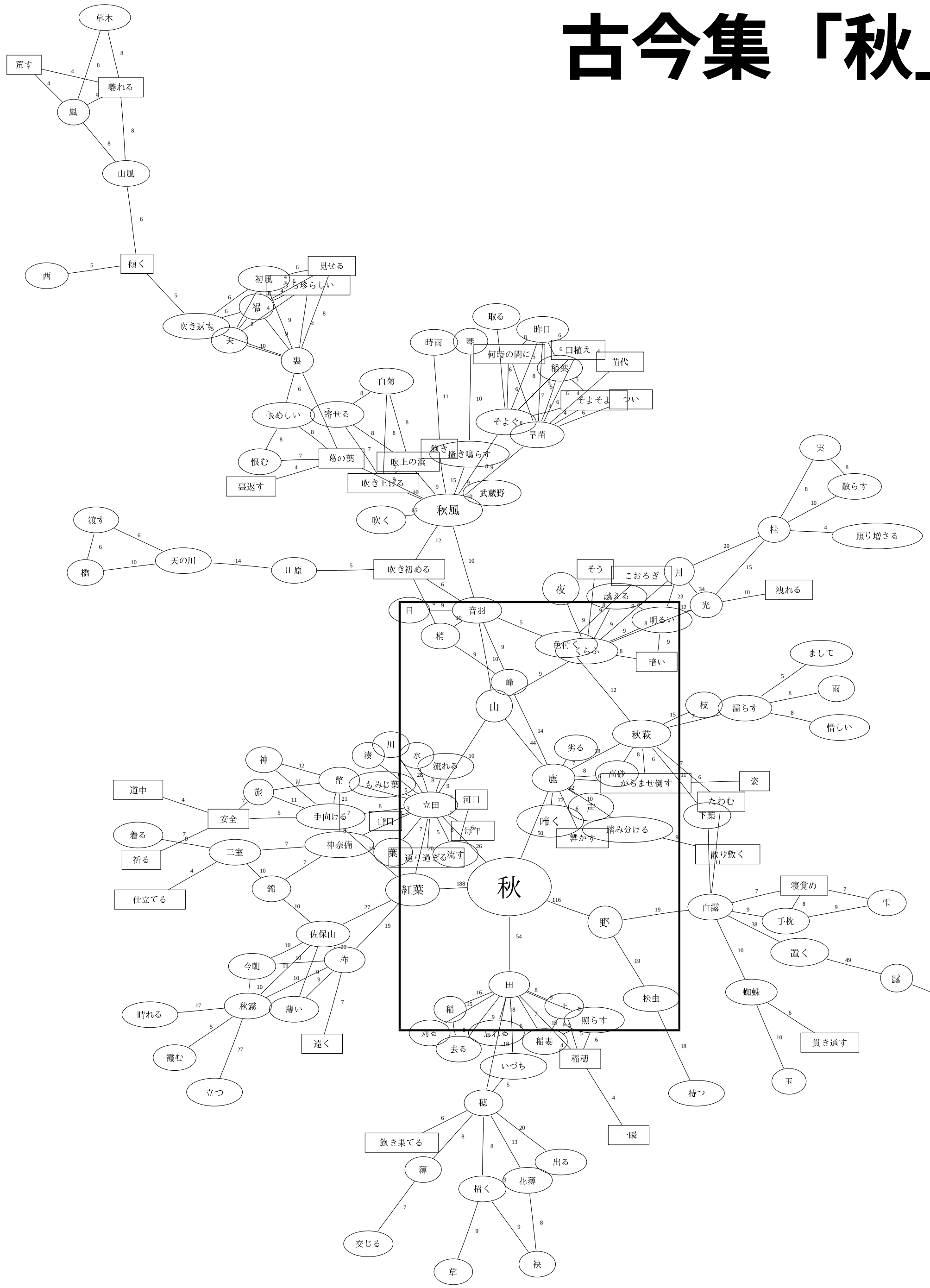


古今集「秋」のモデル

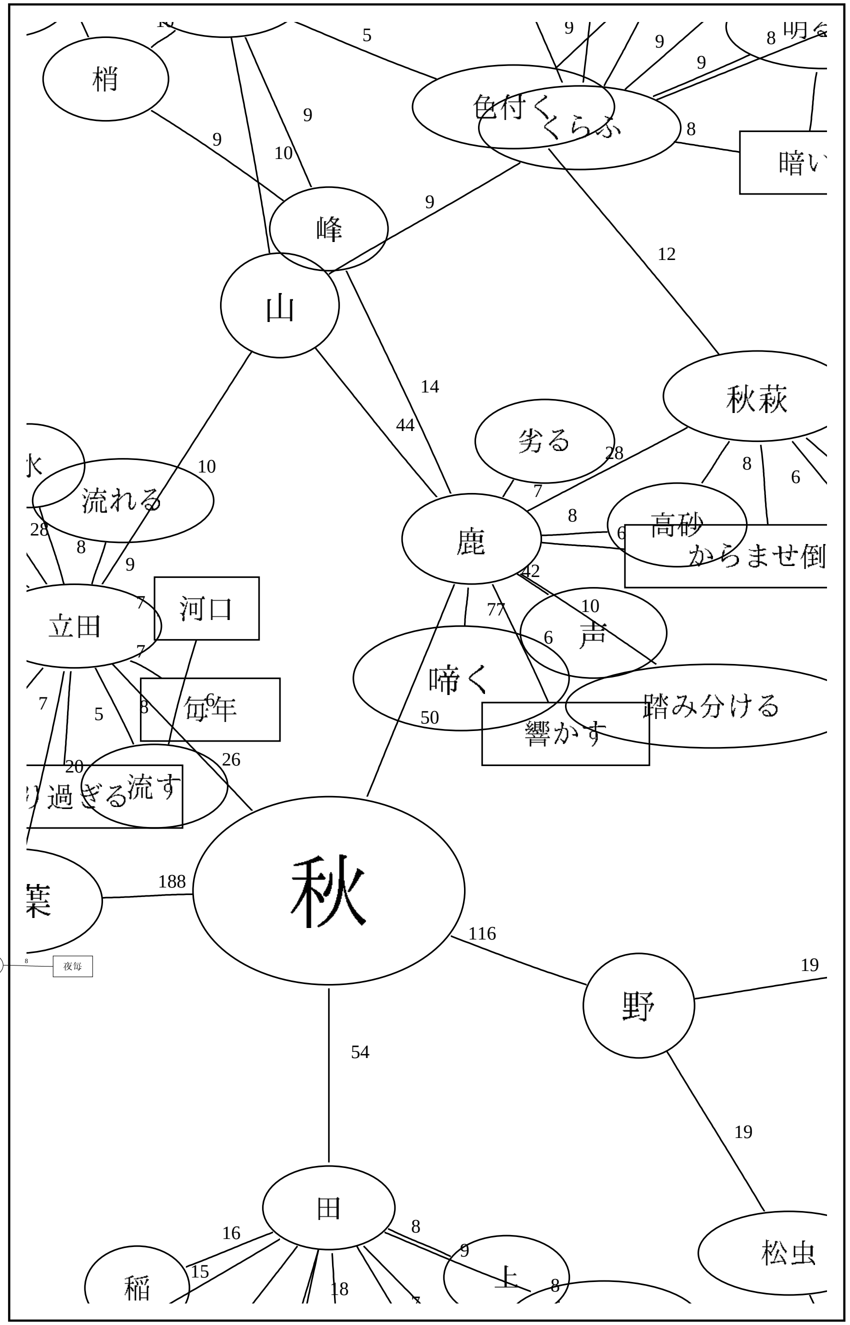


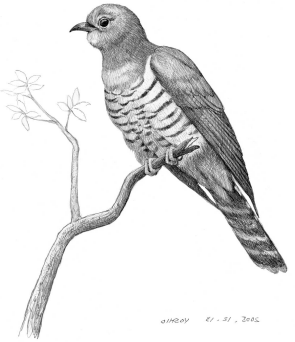
秋 (133/988,2.26): CT cw.>16;
non-dist=off; idf=on(2)

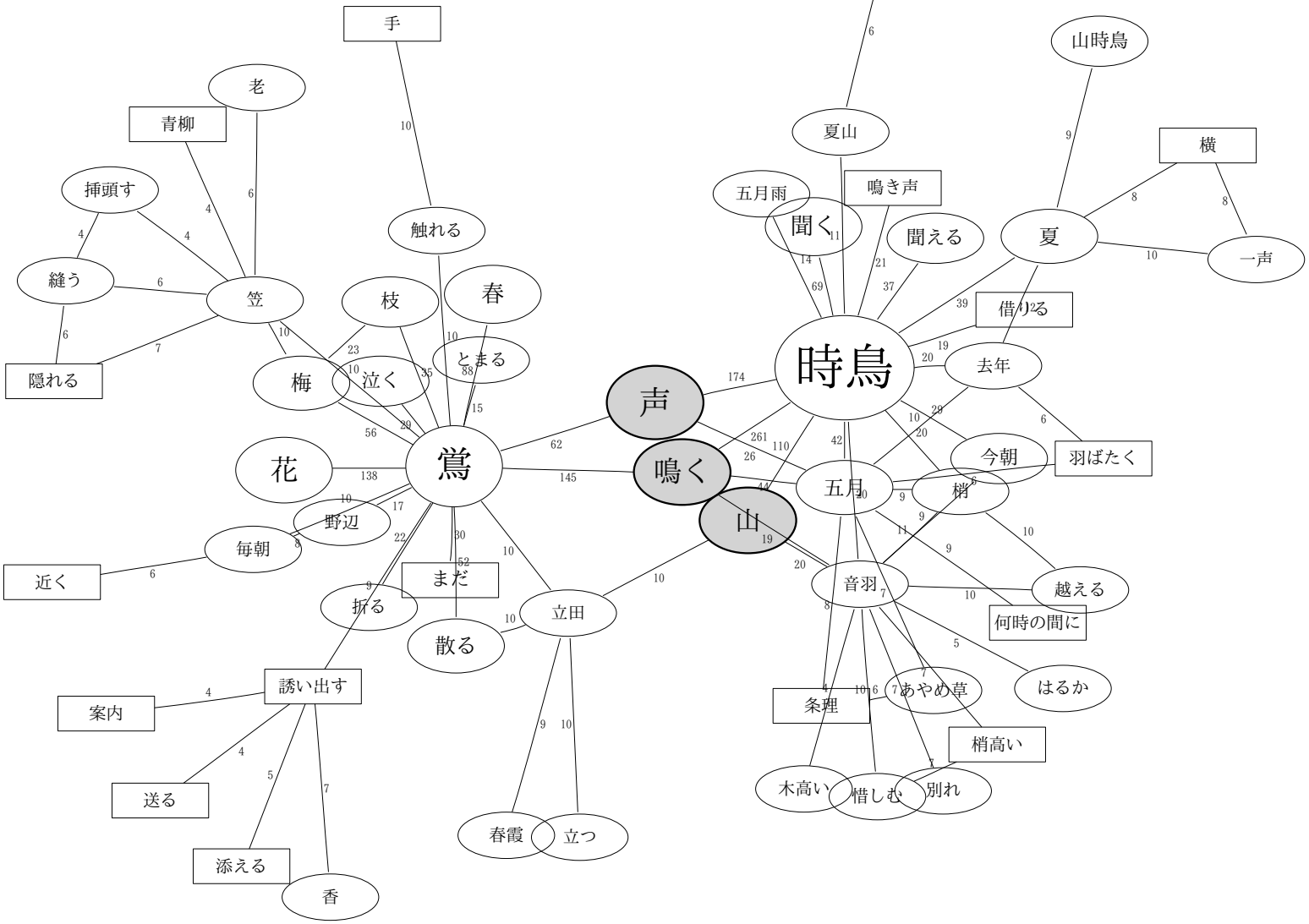
古今集「秋」のモデル



秋 (133/988,2.26): CT cw.>16;
non-dist=off; idf=on(2)







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

言語はどんな形をしているのだろう

- モデル作りは言語の普遍性研究として始められた。
- ことばの意味のむずかしさ
  - 本箱→下駄箱／ふでばこ／あみだな
  - 「そこがみそだ」
  - 「指折り数える」・「足を折る」・「筆を折る」
  - 「コツを呑み込む」→ learn the ropes

ことばの意味のむずかしさ

- 意味の変化
  - 頭が切れる（天才）／頭が切れる（怒り）
- 発生した語の形は変わることはあまりないが、意味は変化している (Goodenough, 1981)。
- 語は語そのもので独立して意味を持たない

  (Lyons 1981)

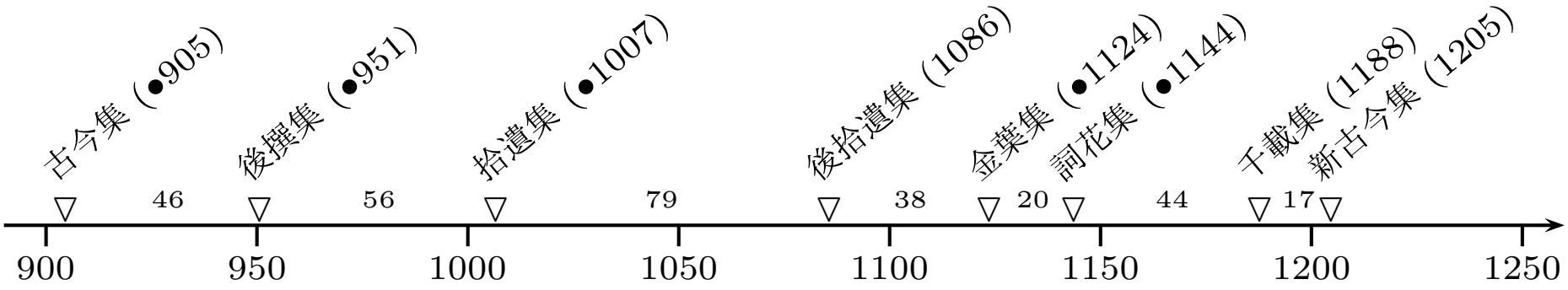
  →語彙の構造は、巨大な多次元の蜘蛛の巣の意味ネット

和歌を材料に→材料の単位が一応明確？

歌ことばのモデリング

- 体系と詳細の可視化 (山元, 2005, 2006, 2007)
- 上記の研究データは古今集に限られていた。
  →八代集に拡張して研究を続けてみたい。
- 「花といえば桜」「桜の吉野」はいつごろか。
  →古今集ではまだ成立していない。
- 新古今集では西行の時代から (片桐, 1983; 小林, 1989)。

八代集の成立



八代集の語彙の転換期

拾遺集説、◎後拾遺集説、千載集説

八代集の語彙の転換期

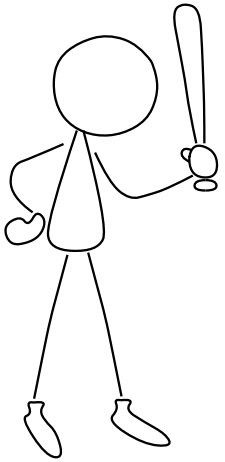
- 一般的には古今撰者の歌の排除された後拾遺集。
- 上野 (1976) 古今・後撰→〈褻〉／拾遺集以降→〈晴〉
- 川村 (1991) すでに拾遺集に見られる。
- 辻 (1998) 語彙的には千載集。
- すべての語彙が急に変わるのではない。
- 当時の流行や文化、撰者によるか。
- 転換期は語によって違うのではないか。
- → 八代集を通して「吉野」のモデルを作ってみる。

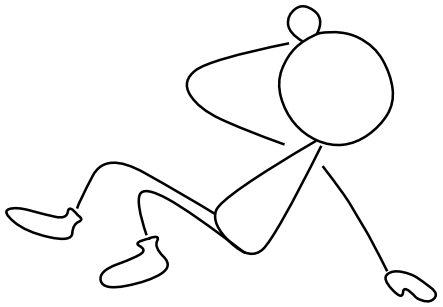
方法：材料

- 国文学研究資料館開発正保版本「二十一代集」
- 長歌を除く 9484 首の和歌テキスト
  (シソーラスの作成はすべての和歌に対して)
- kh で単位分割 (短単位) し、
- 異形同語 (立田／竜田／龍田) の問題→ t2c でシソーラスコードをつける。
- 八代集シソーラスの開発→分類語彙表を利用
- 一般語 (48732)、地名 (1408)、人名 (49)

グラフで可視化

現実





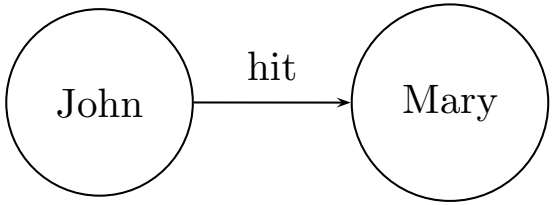
↓

抽象化

グラフで可視化

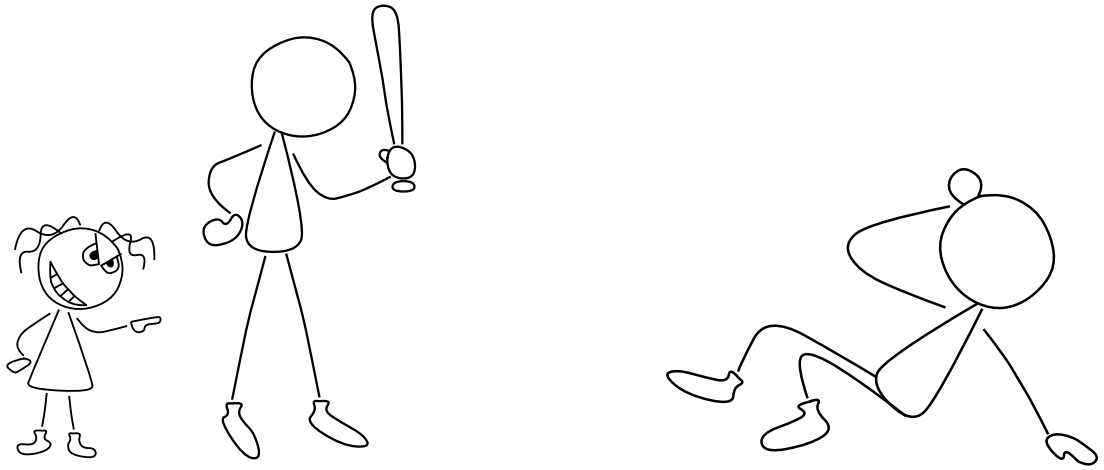
現実

抽象化

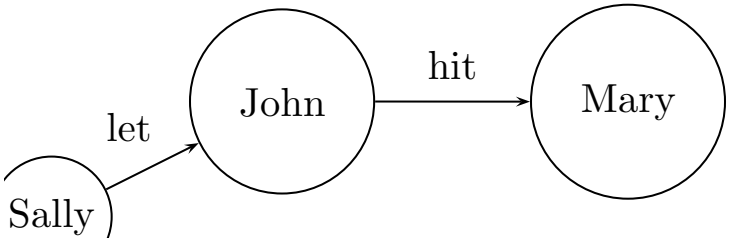


グラフで可視化

現実

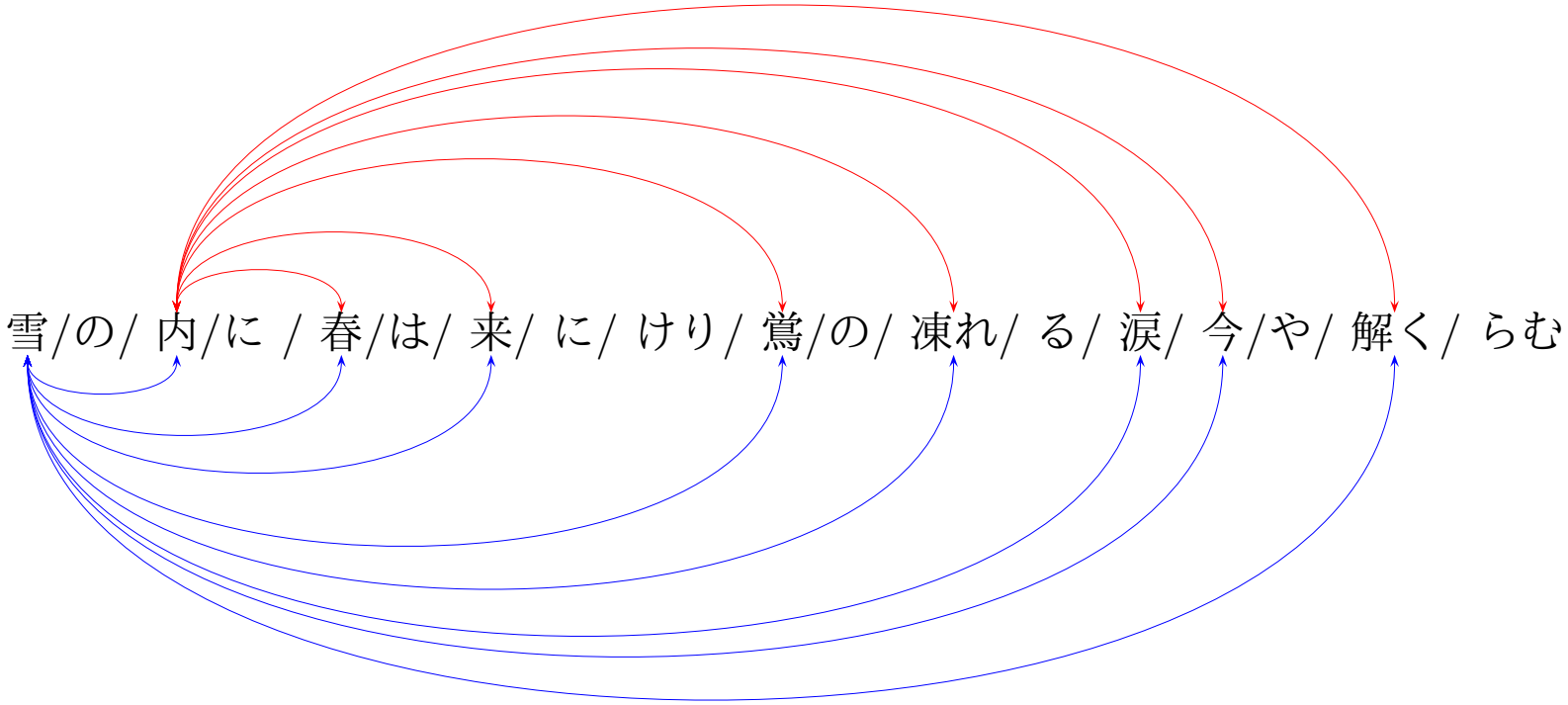


抽象化



精細化

方法：共出現パターン



- 雪–内、雪–春、雪–来、…

用語の採り方

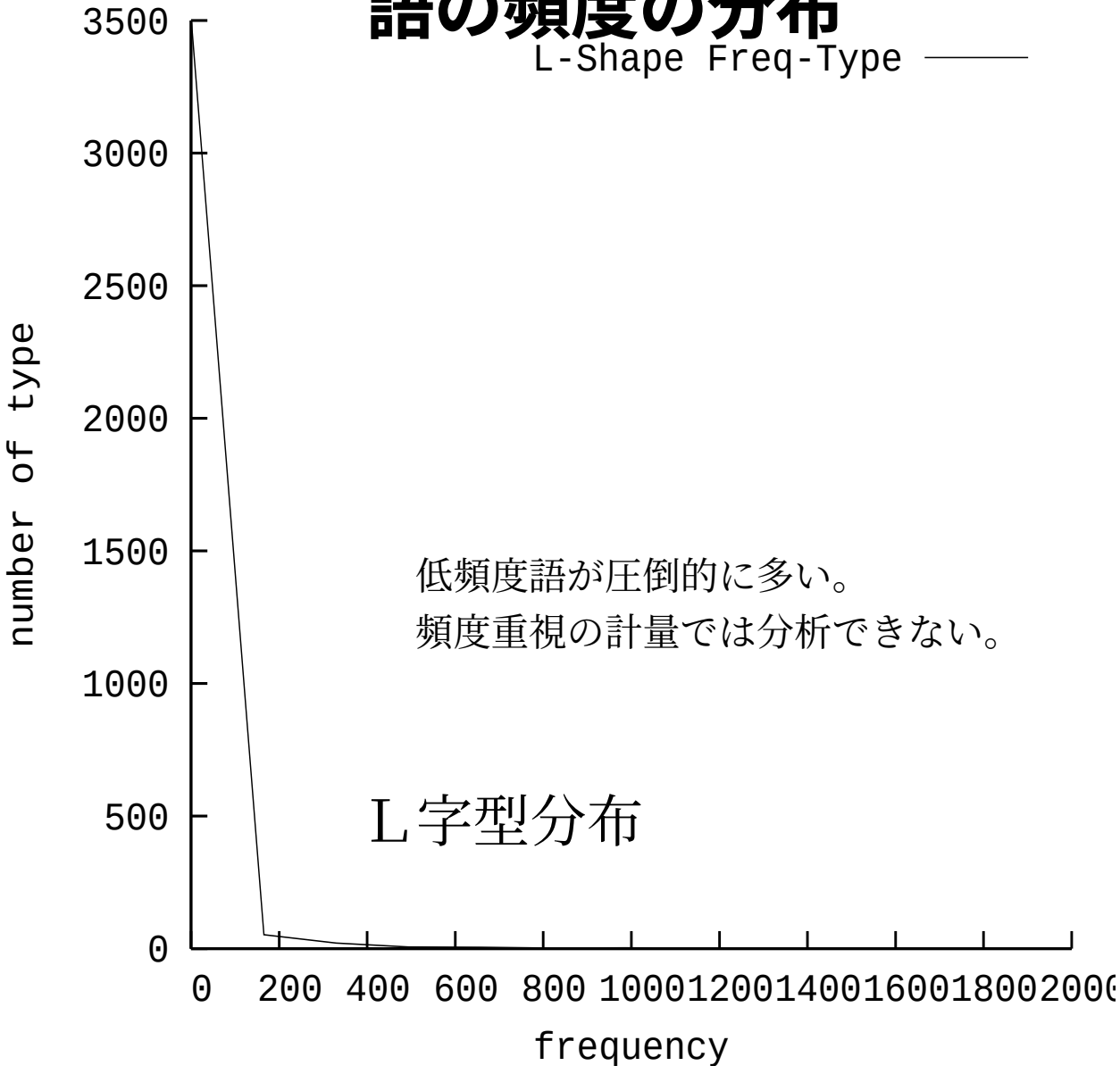
- 一般的に計量研究は、低頻度語が無視される。
- 高頻度語→キーワード性乏しい (水谷, 1983)
- 低頻度語→文章の性格に規定される (石井, 1996)
- idf (Spärck Jones, 1972) を利用する
  → 文脈を有する最小かつ代表的な単位

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

idf: inverse document frequency

$$
\begin{aligned}
idf(ari, N) &= \log \frac{N}{df(ari)} && (1)\\
&= \log \frac{9484}{1201} && (2)\\
&= \log 7.89.. = 2.07.. && (3)\\
idf(uguisu, N) &= \log \frac{N}{df(uguisu)} && (4)\\
&= \log \frac{9484}{101} && (5)\\
&= \log 93.90.. = 4.54.. && (6)
\end{aligned}
$$

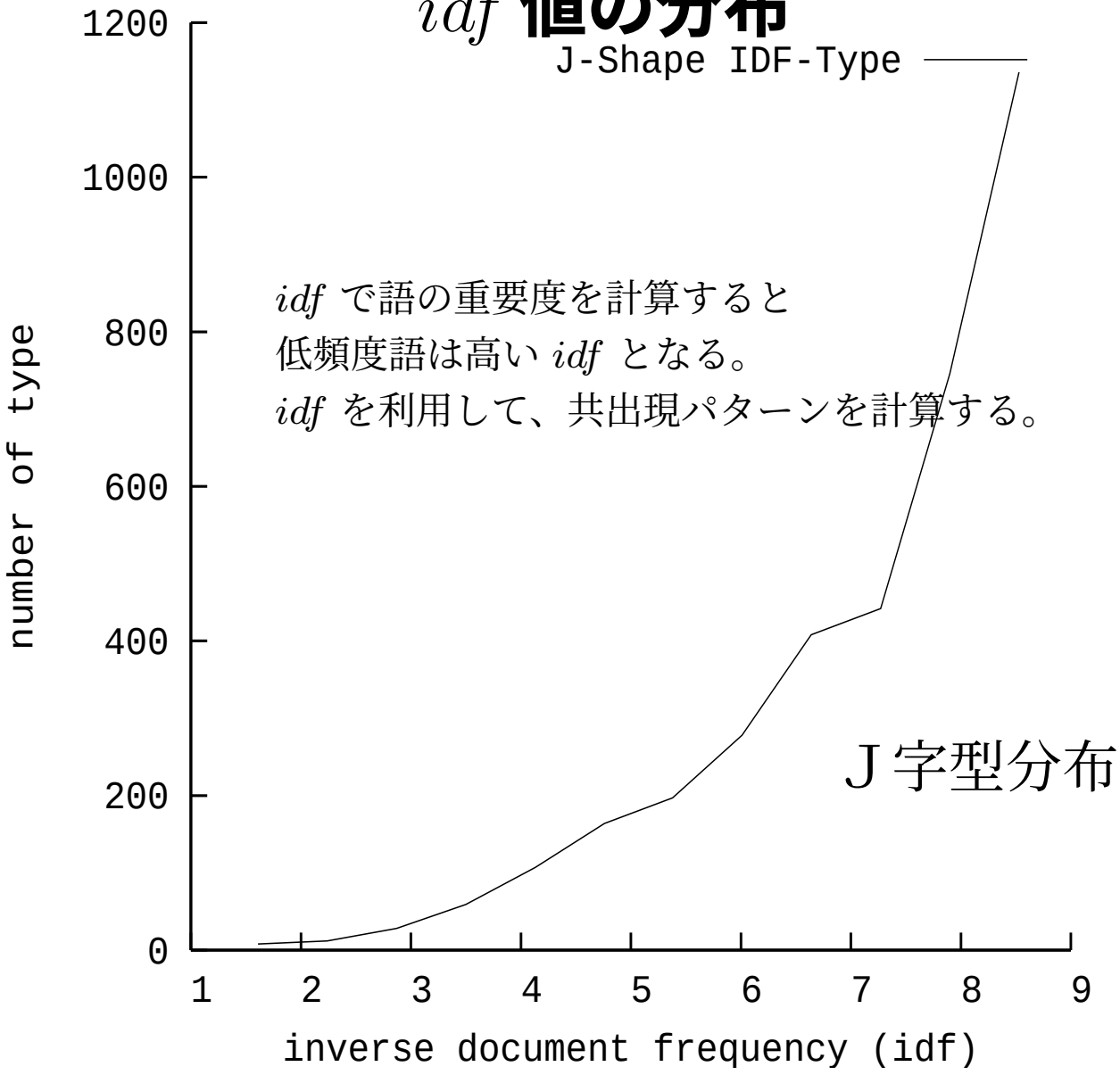
語の頻度の分布



低頻度語が圧倒的に多い。
頻度重視の計量では分析できない。

L字型分布

idf 値の分布



idf で語の重要度を計算すると
低頻度語は高い idf となる。
idf を利用して、共出現パターンを計算する。

J 字型分布

共出現パターンの特徴と計算方法

- 共出現パターンには最小単位 (2 語) で文脈が含まれる。
  月＋宿 → 月が宿る（池などに月影が映る）
  頭＋雪 → 頭の雪（白髪）
- すべてのパターンを描画すると「真っ黒の塊」になる。
  →重要なパターンのみを選び出す必要がある。
- *tfidf* を 2 語の重要度へ拡張する。

$$w(t,K,N) = (1 + \log\ tf(t,K))\ idf(t,N)$$

共出現ウエイト (cw) の計算

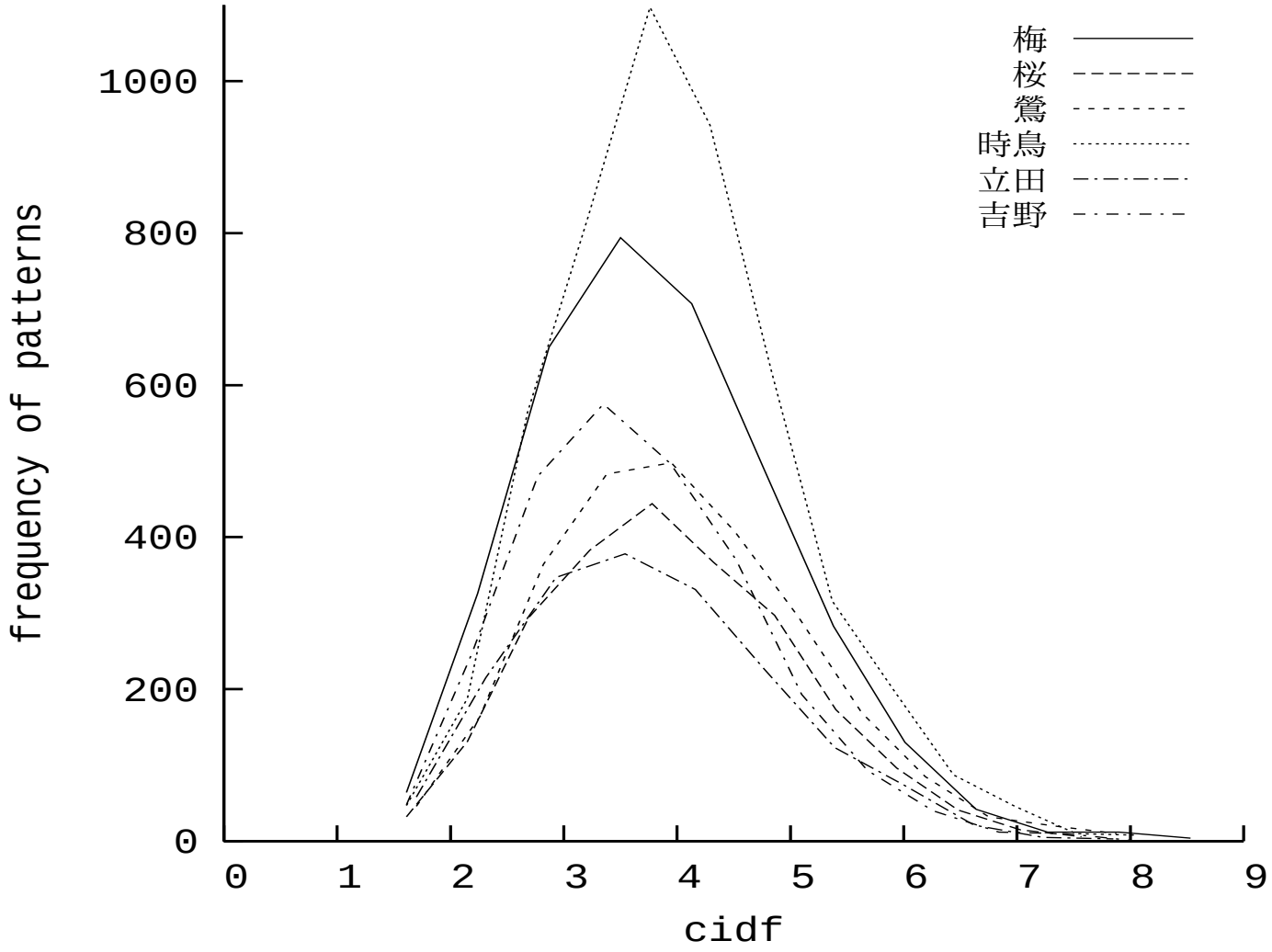
$$w(t, K, N) = (1 + \log\ tf(t, K))\ idf(t, N) \quad (7)$$

$$cidf(t_1, t_2, N) = \sqrt{idf(t_1, N)\ idf(t_2, N)} \quad (8)$$

$$ctf(t_1, t_2, K) = 1 + \log\ |\{k : t_1, t_2 \in k\}| \quad (9)$$

- K は条件により抽出されたテキスト。
- (8) は 2 語の重要度の幾何平均【単語重要度】
- (9) は K に出現したパターンの頻度【実出現頻度】
- ◎パターンの「珍しさ」の情報がない！

cidf 値の分布


梅
桜
鶯
時鳥
立田
吉野
frequency of patterns
0
200
400
600
800
1000
0
1
2
3
4
5
6
7
8
9
cidf

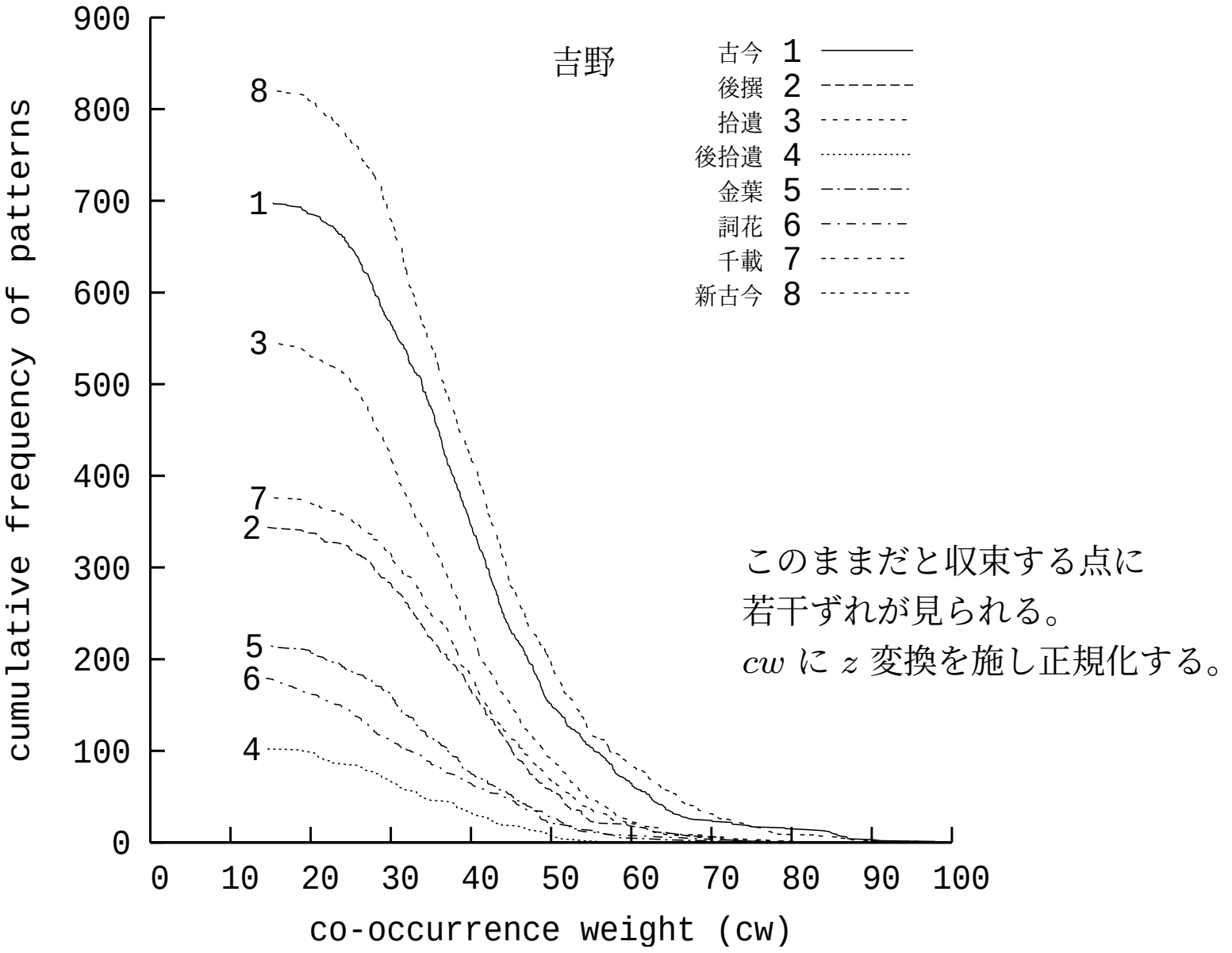
共出現ウエイト (cw) の計算

$$ictf(t_1, t_2, N) = 1 + \log \frac{|N|}{|\{n : t_1, t_2 \in n\}|} \tag{10}$$

$$cw(t_1, t_2) = ctf(t_1, t_2, K)\ ictf(t_1, t_2, N)\ cidf(t_1, t_2, N) \tag{11}$$

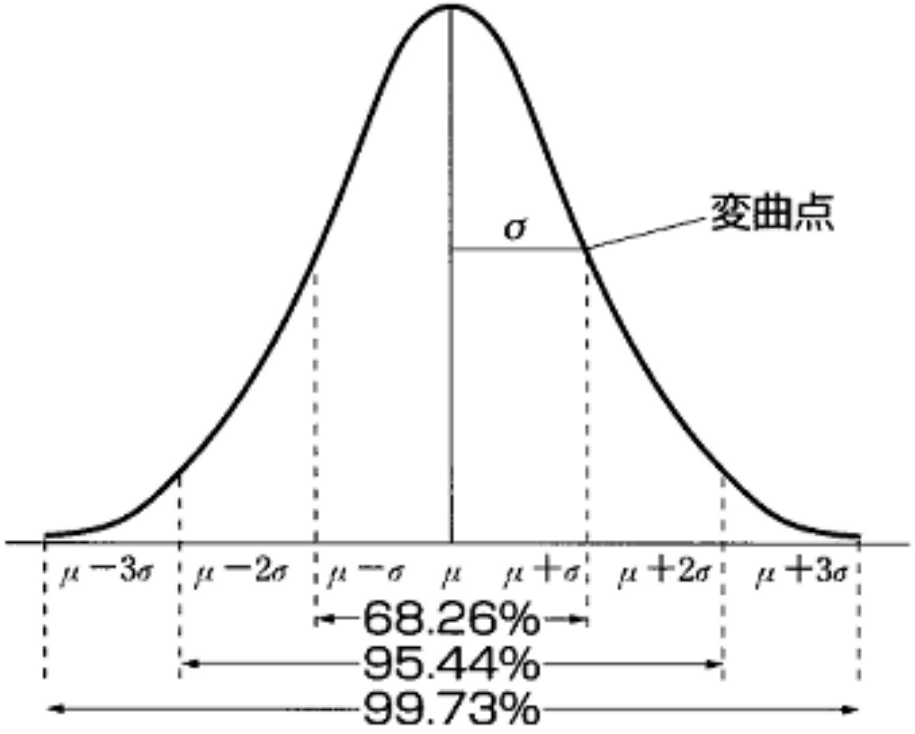
- K は条件抽出されたテキスト。N はすべてのテキスト。
- 【単語重要度】は 2 語の重要度の幾何平均
- 【実出現頻度】は K に出現したパターンの頻度
- 【組合重要度】は N に出現したパターンの頻度

cw 値の累積パターン数



このままだと収束する点に
若干ずれが見られる。
cw に z 変換を施し正規化する。

1σ と変曲点

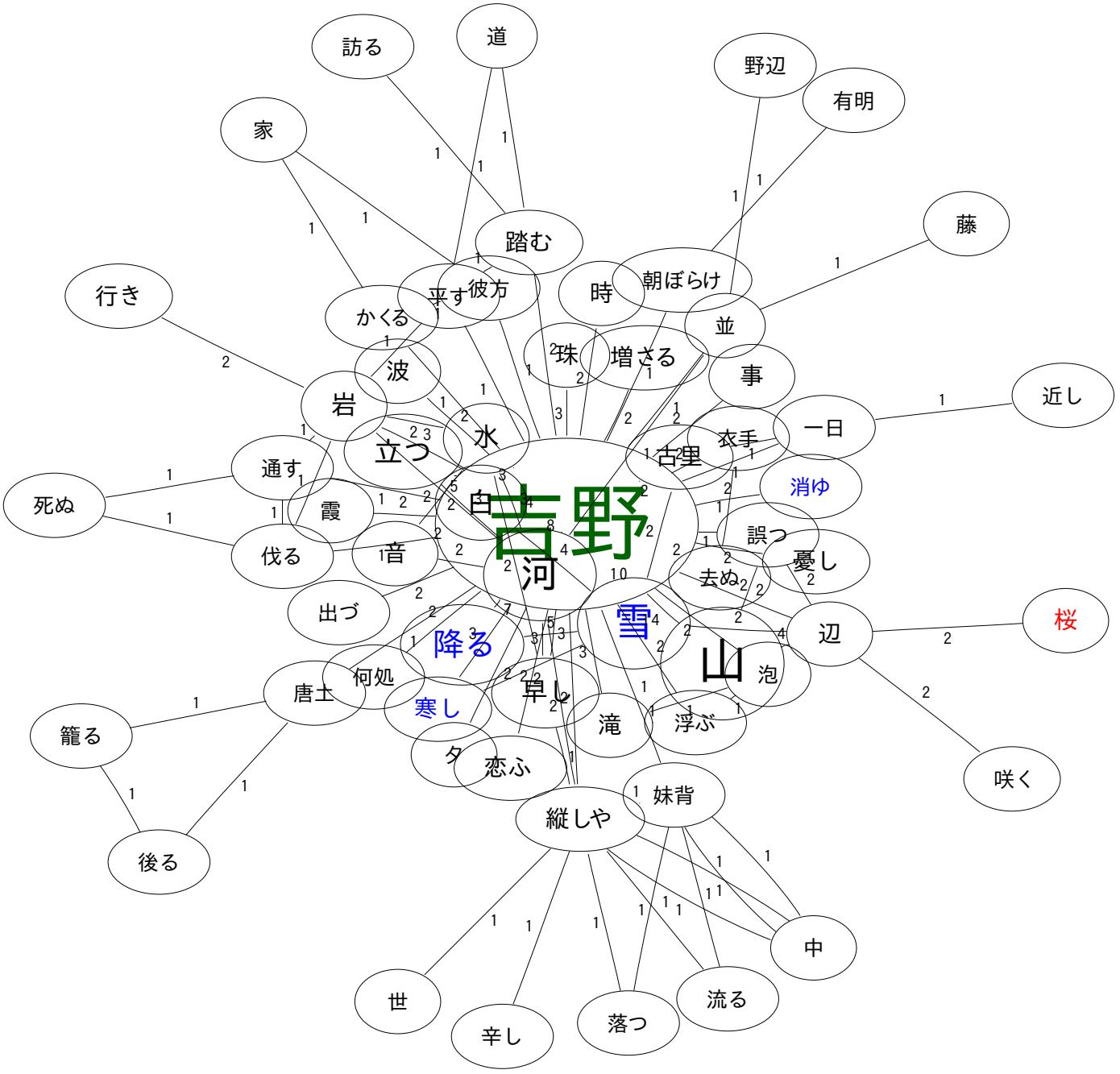


下方向カーブから

右方向カーブへ

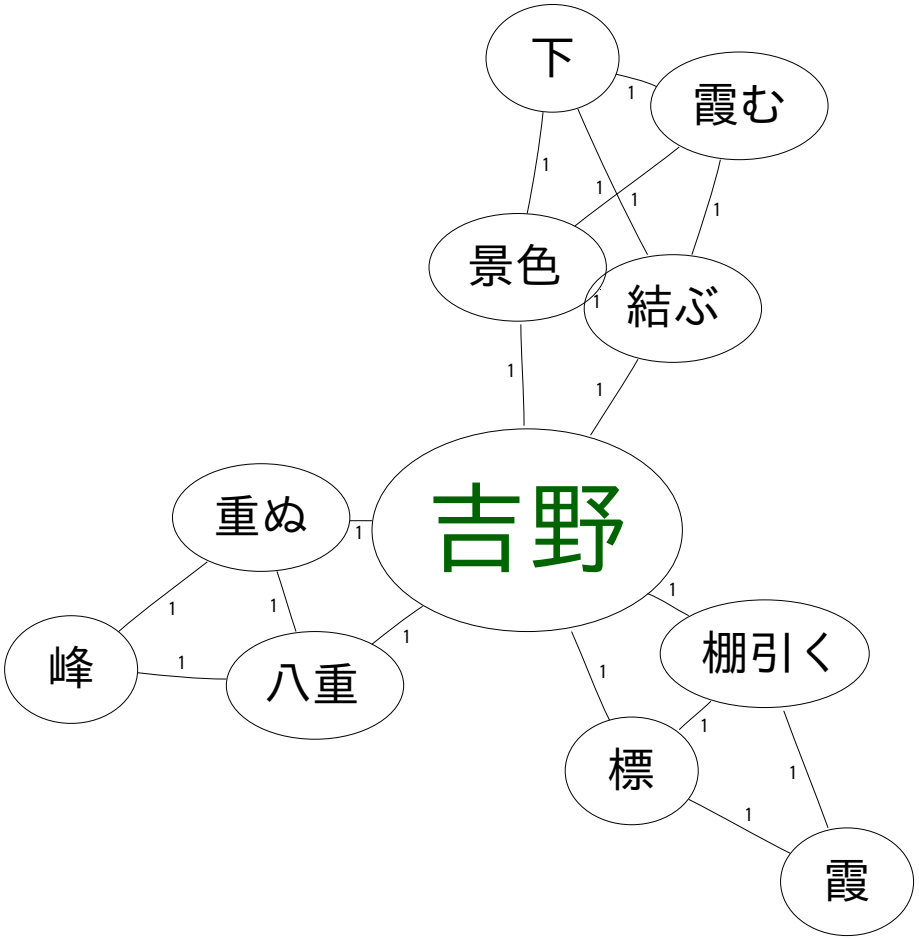
右裾野の 16 %が自動的に選択される。
(人為的に収束点を選ばなくてもよい)

古今集



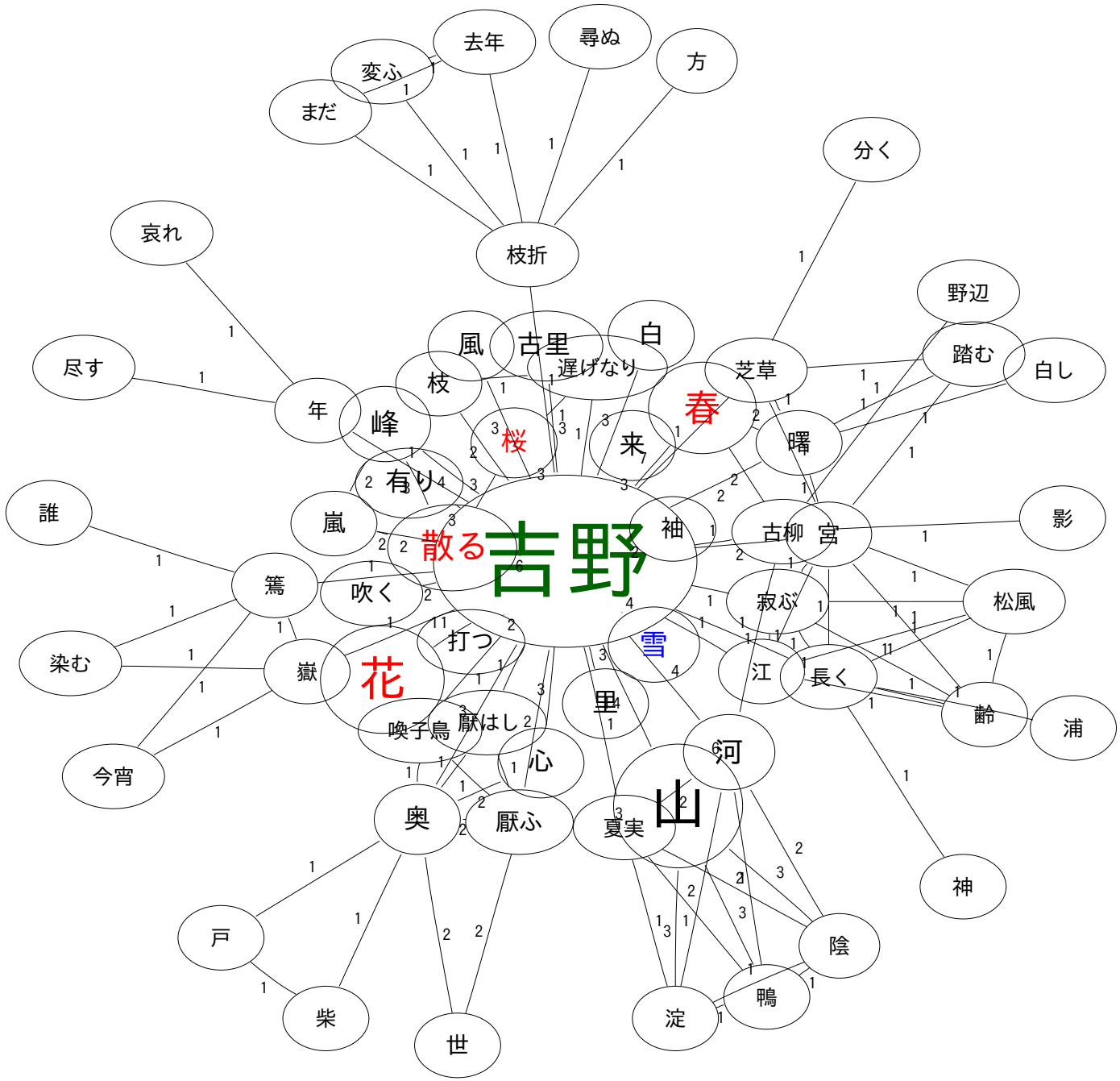
吉野 (24/92/97,4.63)cw > 0.00 K:1-1 U:2 L:0.00 M :16 Z:1.00

後拾遺集



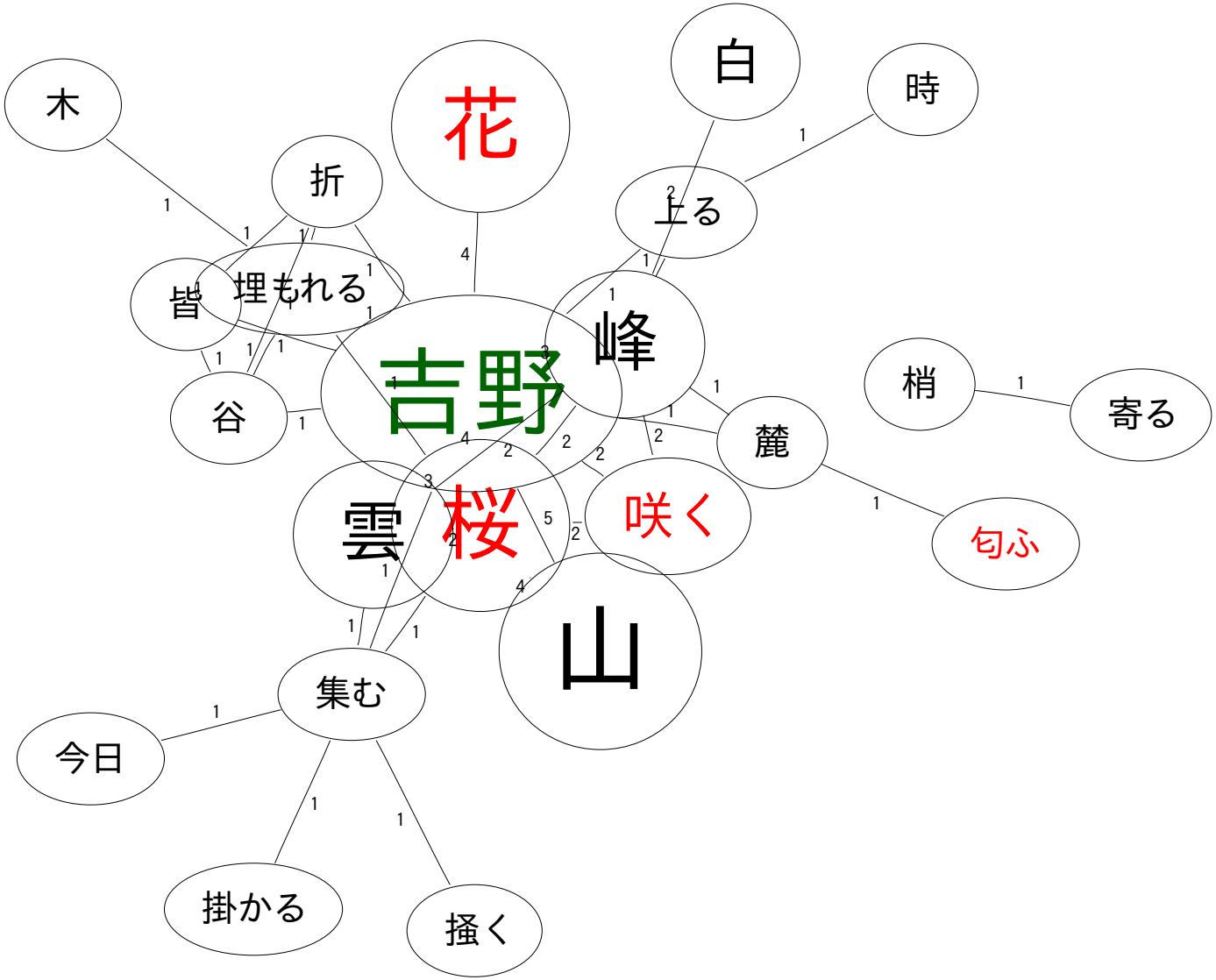
吉野 (3/92/97,4.63)cw > 0.00 K:4-4 U:2 L:0.00 M :16 Z:1.00

新古今集



吉野 (24/92/97,4.63) cw > 0.00 K :8-8 U :2 L :0.00 M :16 Z :1.00

金葉集



吉野 (5/92/97,4.63) cw > 0.00 K :5-5 U :2 L:0.00 M :16 Z :1.00

雪あるいは桜を含むパターン (1)

表２　各集の「吉野」のモデルから抽出した雪あるいは桜を含むパターン

	t_1–t_2	cw	z	ctf	$idf(t_1)$	$idf(t_2)$
古今集 (24)	雪–吉野	86.06	3.33	10	3.18	4.63
	雪–降る	65.15	1.76	5	3.18	3.26
	桜–辺	64.32	1.70	2	3.43	4.69
	雪–寒し	63.36	1.62	2	3.18	4.92
	雪–辺	61.87	1.51	2	3.18	4.69
	雪–白	60.36	1.40	4	3.18	3.18
	雪–古里	55.34	1.02	2	3.18	4.37
後撰集 (11)	雪–吉野	54.69	1.33	3	3.18	4.63
	雪–降る	52.40	1.12	3	3.18	3.26
	雪–崩る	51.40	1.03	1	3.18	8.06
	桜–吉野	51.28	1.02	2	3.43	4.63
拾遺集 (15)	雪–吉野	80.25	3.74	8	3.18	4.63
	雪–消ゆ	55.90	1.54	2	3.18	3.83
	雪–山	54.92	1.46	8	3.18	2.08
	雪–峰	54.35	1.40	2	3.18	3.95
	雪–宿	52.42	1.23	2	3.18	3.37
	雪–古道	50.48	1.05	1	3.18	7.77
後拾遺集 (3)	N/A					

雪あるいは桜を含むパターン (2)

表２　各集の「吉野」のモデルから抽出した雪あるいは桜を含むパターン

	t_1–t_2	cw	z	ctf	$idf(t_1)$	$idf(t_2)$
金葉集 (5)	桜–吉野	72.27	3.34	4	3.43	4.63
	桜–峰	52.17	1.44	2	3.43	3.95
	桜–咲く	51.68	1.40	2	3.43	3.71
	桜–雲	51.00	1.33	2	3.43	3.43
	桜–山	49.48	1.19	4	3.43	2.08
	桜–集む	48.33	1.08	1	3.43	6.59
	桜–埋もれる	47.56	1.01	1	3.43	6.38
詞花集 (6)	N/A					
千載集 (9)	N/A					
新古今集 (24)	桜–吉野	63.56	1.64	3	3.43	4.63
	桜–散る	62.38	1.55	3	3.43	3.14
	雪–吉野	62.18	1.53	4	3.18	4.63
	桜–遅げなり	56.96	1.14	1	3.43	9.16

考察

- 重み (cw) は z 変換により正規化を行った上で、1σ 以上を取り出すと一律に決められ、歌集間の比較が可能。
  →なぜ $1\sigma(16\ \%)$ で文脈がうまく見えるのか。
- 「雪の吉野」から「桜の吉野」へは金葉集が転換期である。→定着期であるかどうかは不明
- 千載集、新古今集では「桜」より「花」という言い方が多い。→定着してきた証拠か

モデリングのまとめ

- モデリングによる「体系」の構築とその可能性
- 実例を示す研究方法→「体系」が示しにくい。
- 語彙リスト、一覧表による方法→実例に戻らなくてはならない「もどかしさ」がある。
- モデルからテキストへ参照→モデリングシステム

  `http://etymology.jp/waka/poem.cgi`

  XML(SVG) フォーマットの採用。
- 八代集シソーラスの公開

質問

- 和歌の数理モデルについては

  http://warbler.js.ila.titech.ac.jp/~yamagen/

  をご覧ください。
- 東工大へもお出でください。学部向け言語学の授業あり。
- 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@ila.titech.ac.jp