

# 「現代語の意味の変化に対する 計算的・統計力学的アプローチ」 シンポジウム

持橋大地

統計数理研究所 数理・推論研究系

daichi@ism.ac.jp

(daichi@boz.c.u-tokyo.ac.jp ← はるか昔)

TKP東京駅セントラルカンファレンスセンター/  
オンライン  
2022-3-9 (水)

# はじめに

- 本シンポジウムは、国立国語研究所の共同研究プロジェクトに2019年から採択された同名のプロジェクトの報告会でもあります(2019-2022年度)

The screenshot shows a web browser displaying the website of the National Institute for Japanese Language and Linguistics (NINJAL). The URL in the address bar is [www.ninjal.ac.jp/research/project-3/institute](http://www.ninjal.ac.jp/research/project-3/institute). The page title is "現代語の意味の変化に対する計算的・統計力学的アプローチ". The main content area includes sections for "Project Leader" (持橋 大地), "Implementation Period" (2019年4月～), and "Summary". The summary text discusses the study of language meaning change over time using computational and statistical methods. At the bottom right, there is a logo for "Research Organization of Information and Systems" and "The Institute of Statistical Mathematics".

大学共同利用機関法人 人間文化研究機構  
国立国語研究所  
National Institute for Japanese Language and Linguistics NINJAL

トップ > 研究活動 > 共同研究プロジェクト > 機関拠点型基幹研究プロジェクト > 現代語の意味の変化に対する計算的・統計力学的アプローチ

## 現代語の意味の変化に対する計算的・統計力学的アプローチ

▶ プロジェクトリーダー  
持橋 大地 (統計数理研究所)

▶ 実施期間  
2019年4月～

## 概要

現代および近代日本語において、あるいは現代語一般に、単語の意味は時間を通じて一様ではなく、常に変化し続けている。こうした中、(1) どのような単語が、(2) どのように意味変化を起こしたのか、またそれはどのようなメカニズムで起きるのかを理論的に調べることは、国語学および言語学において最も重要な現代的課題の一つであると考えられる。特に、国立国語研究所において提供されている通時コーパスなどを用いれば、デジタル化されたテキストデータが大量に得られるため、上記の(1)においては

Research Organization of Information and Systems  
The Institute of Statistical Mathematics

# プロジェクトの構成員

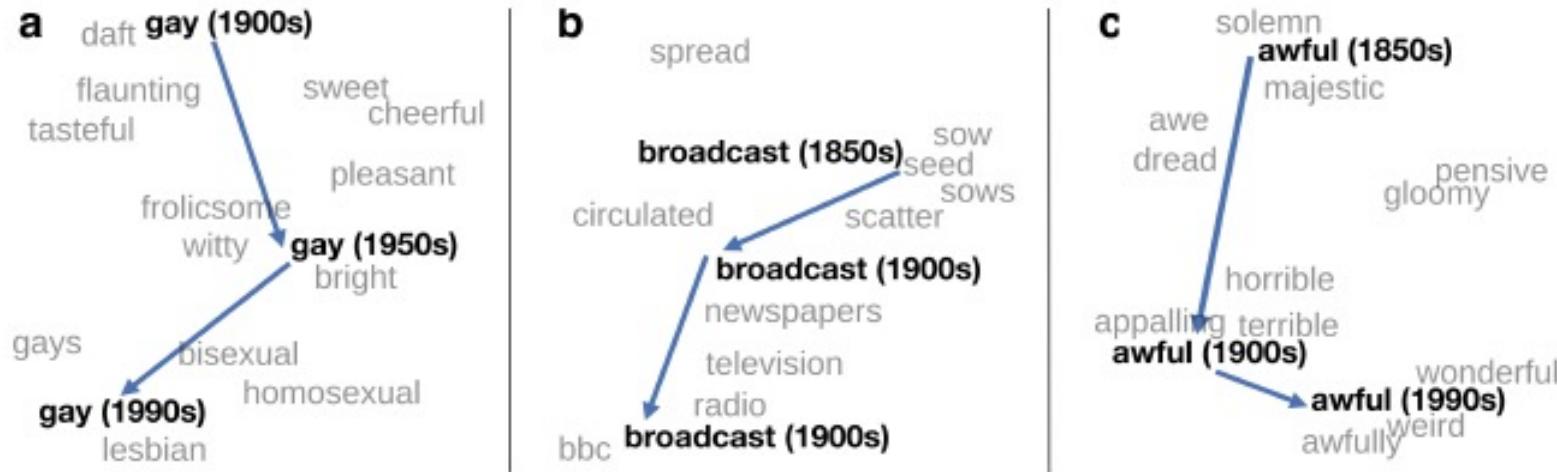
- 持橋大地 (統数研 数理・推論研究系)
- 小木曾智信 (国語研 言語変化研究領域)
- 小町守 (東京都立大 システムデザイン研究科)
- 高村大也 (産総研 人工知能研究センター)
- 坂田綾香 (統数研 数理・推論研究系)
- 小山慎介 (統数研 モデリング研究系)
- 相田太一 (東京都立大 M2)
- 井上誠一 (東京都立大 M1)
- 竹中誠 (東京都立大 D3)

# 本日の参加登録者

- 現在までに、184人の参加登録をいただきました。  
誠にありがとうございました。
- アンケート結果を、今日最後のパネルセッションで  
議論させていただきます

# 「言葉の変化」の研究

- 最近、自然言語処理でも言語学でも、多数行われている
- 英語学：「コーパスと言語変異研究会」(2020～)  
<https://corp-lang-var.blogspot.com>
- 自然言語処理：多数の研究が積み重ねられている



(Hamilton+, ACL 2016)

# 本研究プロジェクトでのアプローチ

- 国語研の通時コーパス(CHJ)も利用して、近代において
  - どの言葉の意味が
  - どんな風に変わったのかを統計的・網羅的に発見する試み
- データ解析チームと、統計力学チームの2チームに分けて研究を進めた
- 本シンポジウムの発表では、データ解析チームの成果および、関連した研究をされているNLP・言語学の皆様に講演をいただきます

# これまでの言語学での試みとの違い

- 対象となる言葉を限定しない
  - どの言葉が、どれくらい変わったのかを発見する
  - 自然言語処理の技術を最大限使用
  - 統計力学に基づく理論アプローチを目指す
- 言語学では、「～の変化」という形で、最初から対象が限定されていることが多い
- そもそも、言葉のどんな特徴が意味変化に結び付くのか？

# どんなデータがあるか?

- 自然言語処理で多く使われているデータ

Name	Language	Description	Tokens	Years	POS Source
ENGALL	English	Google books (all genres)	$8.5 \times 10^{11}$	1800-1999	(Davies, 2010)
ENGFIC	English	Fiction from Google books	$7.5 \times 10^{10}$	1800-1999	(Davies, 2010)
COHA	English	Genre-balanced sample	$4.1 \times 10^8$	1810-2009	(Davies, 2010)
FREALL	French	Google books (all genres)	$1.9 \times 10^{11}$	1800-1999	(Sagot et al., 2006)
GERALL	German	Google books (all genres)	$4.3 \times 10^{10}$	1800-1999	(Schneider and Volk, 1998)
CHIALL	Chinese	Google books (all genres)	$6.0 \times 10^{10}$	1950-1999	(Xue et al., 2005)

Table 1: Six large historical datasets from various languages and sources are used.

- 国語研 日本語歴史コーパス (今回は明治・大正以降を使用)

The screenshot shows the homepage of the Japanese Historical Corpus (CHJ). The top navigation bar includes links for 'Corpus Development NINJAL' (NINJAL), 'Corpus', 'ツール' (Tools), '申込方法' (Application Method), 'KOTONOHA計画' (KOTONOHA Project), '語彙調査データ' (Vocabulary Survey Data), '報告書' (Report), and 'イベント' (Events). A blue button for '国立国語研究所' (National Institute of Japanese Language and Linguistics) is also present. The main menu on the left has sections for '歴史コーパス' (Historical Corpus), 'MENU メニュー' (Menu), '概要' (Overview), and '利用・申込方法' (Usage/Application Method). The central content area displays the '概要 INTRODUCTION to CHJ' section, which includes a brief description of the corpus and its purpose. The footer features the text 'and Systems 研究' (and Systems Research) and 'The Institute of Statistical Mathematics' along with the logo of the National Institute of Japanese Language and Linguistics.

# 意味変化と統計力学

- 元々のアイデアは、個人が「周りの言語使用に合わせて意味解釈を変える」というもの  
→ イジングモデル、Pottsモデル
- 例：「乖離」は本来、納豆と議会のようにまったく違う概念であることを表していたが、単に「違うこと」を意味するように周囲が使うと、自分もそれに合わせることになり、意味解釈が変わってくる

# 持橋のアイデア : Coupled DP

Coupled Dirichlet process (持橋さん提案)

- あるネットワーク上に $N$ 人の人がいるとする。
- 時刻 $t$ における $i$ 番目の人の状態(=この人が使っている単語)を $z_i^{(t)}$ とする。
- 時刻 $t$ における状態数(=系内で使われている単語総数)を $K_t$ とする。
- 時刻 $t + 1$ における $i$ 番目の人の状態は次のように決まる。

$$p(z_i^{(t+1)} \mid z_1^{(t)}, z_2^{(t)}, \dots, z_N^{(t)}) = \begin{cases} \frac{n_{k,i}^{(t)}}{N_i + \alpha} & \text{for } z_i^{(t+1)} = k (\leq K_t) \quad N_i = \sum_{j=1}^t c_{ij}: \text{次数} \\ \frac{\alpha}{N_i + \alpha} & \text{for } z_i^{(t+1)} = K_t + 1 \quad \leftarrow \text{新しい単語の開発} \end{cases}$$

$n_{k,i}^{(t)} = \sum_{j=1}^t \textcolor{red}{c_{ij}} \delta_{z_j^{(t)}, k}$   $\cdots$   $i$ 番目の人とつながっている人のうち、 $k$ 番目の単語を使っている人の割合

- $c_{ij}$ :  $i$ 番目と $j$ 番目が繋がっていれば1, そうでなければ0。また $c_{ii} = 1$ 。
- つまり、周りの人が使っている頻度の高い単語を採用する。(平均場的)

# 坂田さん(統数研)のアイデア

Penalized Potts model

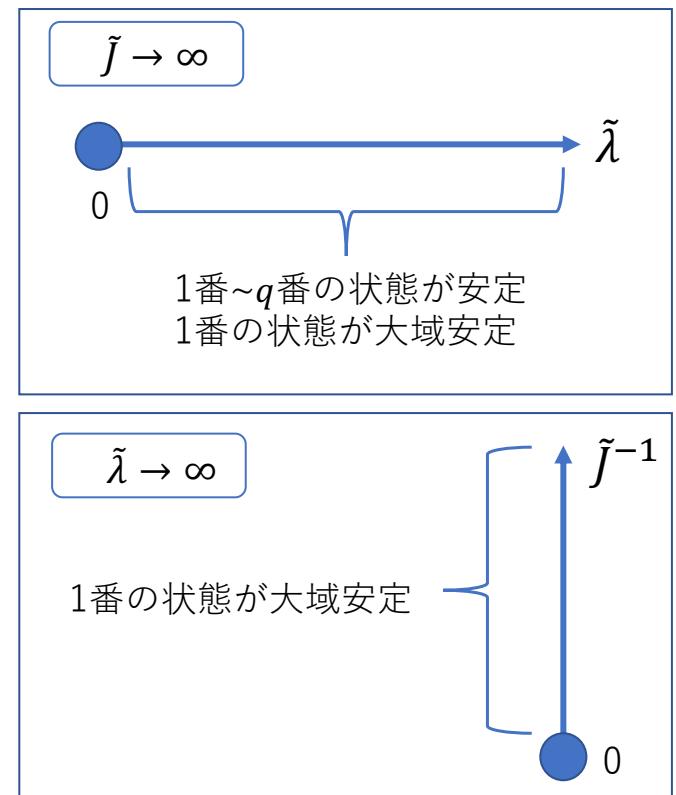
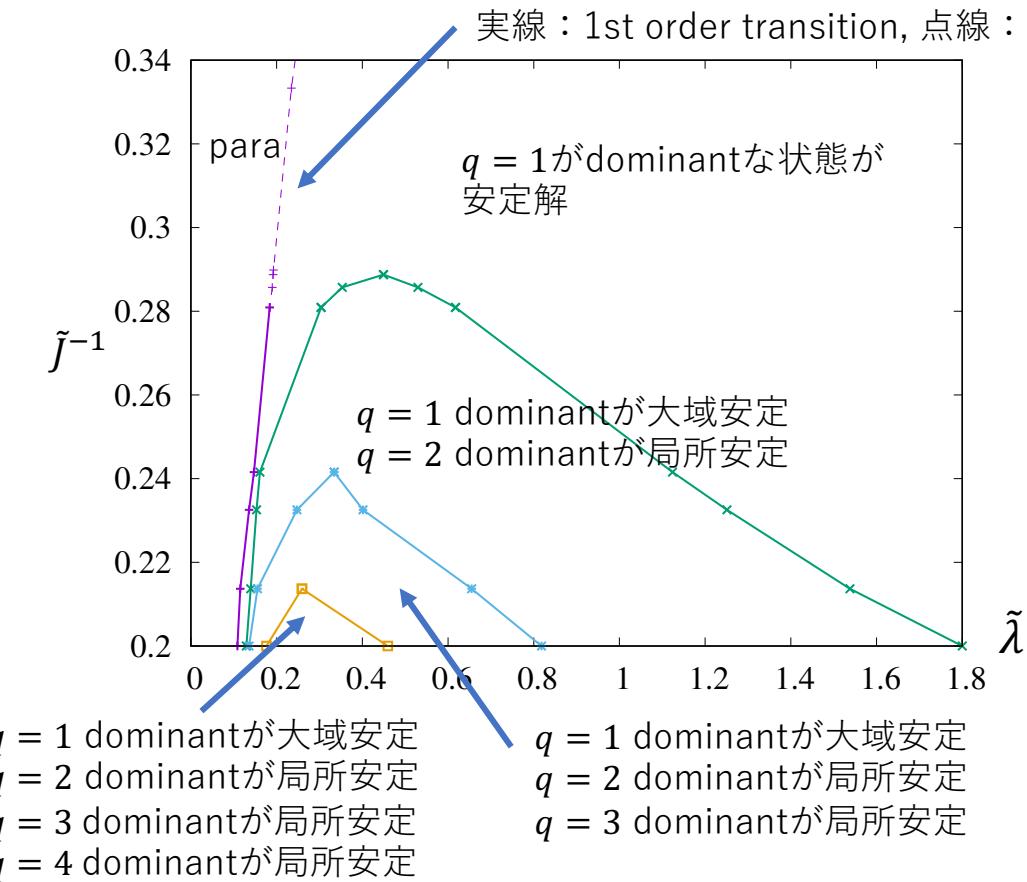
- 状態数に対してペナルティを入れる

$$H = - \sum_{i < j} J_{ij} \delta(S_i, S_j) + \lambda \sum_i S_i, \quad S_i \in \{1, 2, \dots, \infty\}$$

- 書き換えると  $\lambda \sum_i S_i = \lambda \sum_q q \sum_i \delta(S_i, q)$
- 状態数をあらかじめ決めるのではなく、  
ペナルティで状態数を決める
- ダイナミクスを考えると、  
 $\lambda$  で決まる有限の確率で新しい状態への遷移が起こる

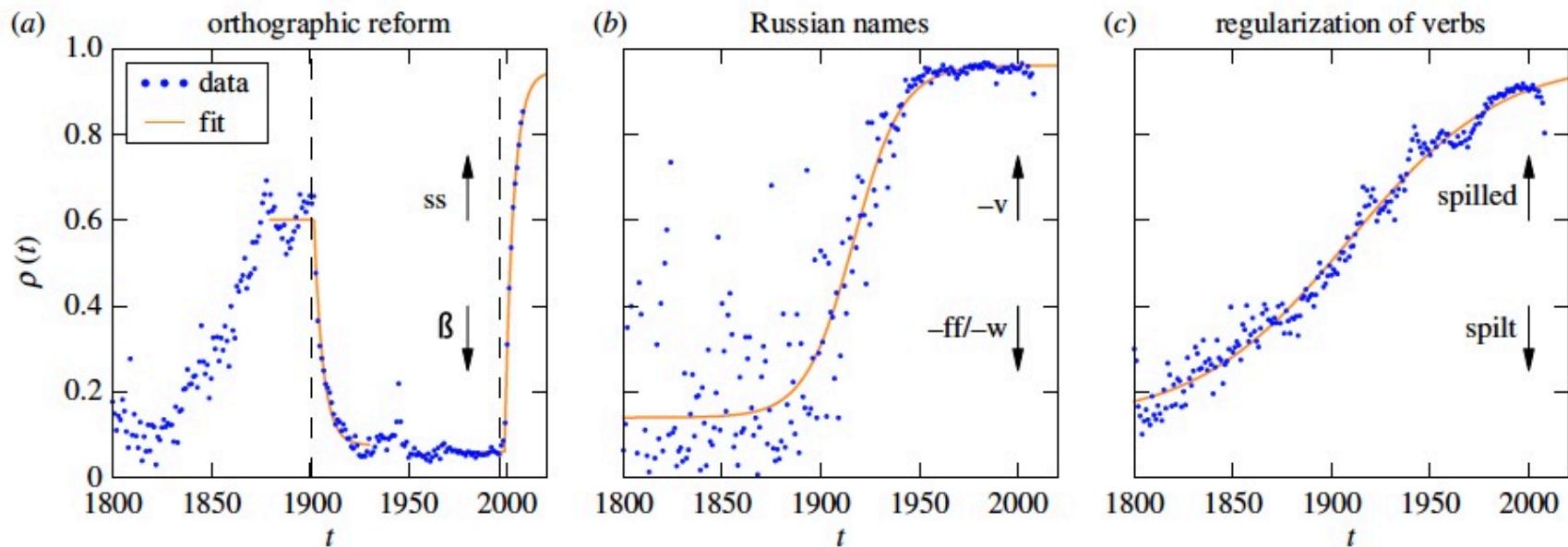
# 坂田さん(統数研)の解析

## 極限(1)



# 言語変化とS字カーブ

- 言語変化に関する理論的な研究は、そのほとんどが物理の分野で行われている (Physical ReviewやPRL等)
- “Extracting information from S-curves of language change” (J. of Royal Society Interface, 2014) を紹介
- 意味の変化は分からないので、綴りの変化の時系列



# S字カーブの導出

- 時刻tで感染している個体の確率  $\rho(t)$  は、感染していない個体の確率  $1-\rho(t)$  に「 $\rho(t)$  の定数倍+ベースライン」を加えたファクターで変化するため (Bassモデル)、

$$\frac{d\rho(t)}{dt} = (a + b\rho(t))(1 - \rho(t)).$$

- これを解くと、

$$\rho(t) = \frac{a(1 - \rho_0) - (a + b\rho_0)e^{(a+b)(t-t_0)}}{-b(1 - \rho_0) - (a + b\rho_0)e^{(a+b)(t-t_0)}}$$

- $a$ が $>0$ ならば、完全に内生的 (S字カーブ)
- $b$ が $>0$ ならば、完全に外生的 (指数分布)

# 推定方法

1.  $p(\text{データ}|a=0)$  と  $p(\text{データ}|b=0)$  の確率を比べて、どちらが大きいか見る
2. 系数aとbを推定する
  - 外生的である確率は、

$$G \equiv G^{\text{exo}} = \frac{a}{b} \log_e \left( \frac{a+b}{a} \right)$$

3. Bassモデルを一般化した

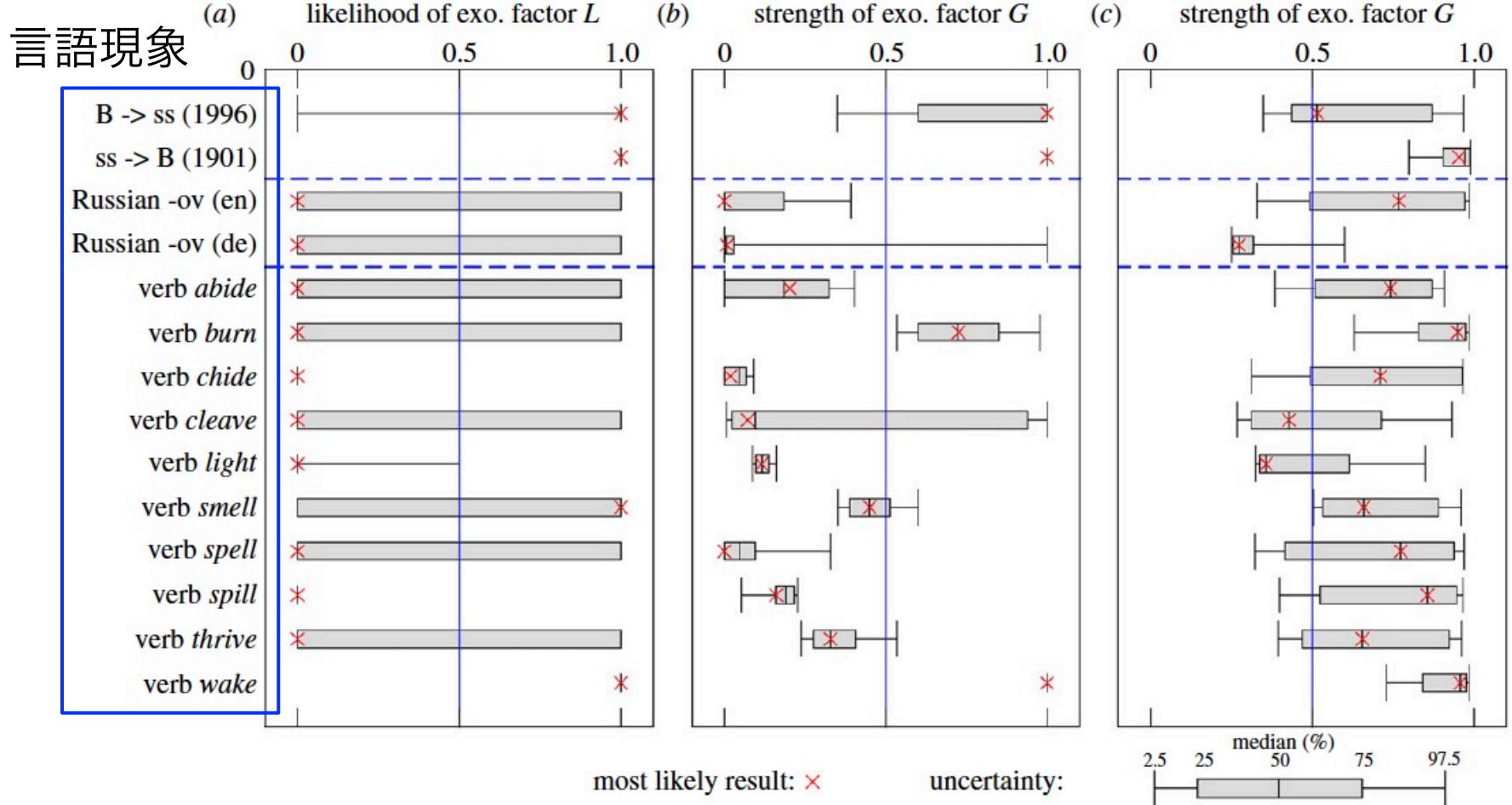
$$\dot{\rho}(t) \equiv \frac{d\rho(t)}{dt} = g(\rho(t))(1 - \rho(t))$$

を使い、 $\dot{\rho}(t)$  を数値微分で求めて内生性  $g(\rho)-g(0)$  を計算する

# 実験

- Gleeson (2014)のネットワークモデルから、シミュレーションデータを生成
  - この場合、外生性因子は  $G = \sum_k P_k \sum_{m=0}^k \int_0^\infty s_{k,m} F_{k,0} dt,$
  - 小山さんが去年解説してくれた近似マスター方程式 (AME)で計算できる
  - Bassモデルの形を仮定しない3.が最も良い
- 実際の言語データで計算
  - 数値微分は不安定なので、3.は使えない
  - 1.は結果がどちらかに寄って極端
  - S字カーブを直接推定する2.がよい
  - Bootstrap法を使って、推定値の分散を求めて信頼性を考慮している

# 実験結果



- 左からそれぞれ推定法1,2,3

# 議論

- ここでは、言語の形式的な変化を扱っている
- 変化がドイツ語の  $ss <-> \beta$  のように二値かつ、競争的な場合の話
  - どちらかの形を必ず使わなければならぬ
- 意味の場合は、競争的とは限らない
  - ある意味を表すのに、どの語を使ってもよい
  - 「語」ではなく「意味」について競争がある?  
(例: “とっぽい”, “ととっぽい”, “へっちゃり” のどれを使って表現するか?)
- 二値とは限らないダイナミクスの場合どうするのか?  
(cf. grassroot changes in linguistic systems, 2014)

## 議論 (2)

- この論文では、感染を生み出すネットワークのモデルと近似マスター方程式は、データ生成にのみ用いている
  - 結果として出てきたS字カーブのみを評価
- 背後に隠れたネットワークモデルのパラメータを推定できないか?
- Hamilton (2016) でも、頻度や意味の数に関する「結果」だけを議論 → なぜそうなるか、は不明
  - 意味の数が多いと変化が速くなる、頻度が高いと変化が遅くなる、ことの理論的なモデルを作れるか?

# Feltgen+ (2017)

- Feltgen, Fagard, Nadal, “Frequency patterns of semantic change: corpus-based evidence of a near-critical dynamics in language change”, Royal Society Open Science, 2017.
- 1300年代-2000年までのフランス語のテキストを使ったデータの観察+モデル化
- S字カーブだけでなく、その前駆体の存在を示す  
→ 同じ統計的法則に従う (逆ガンマ分布)

# データ量

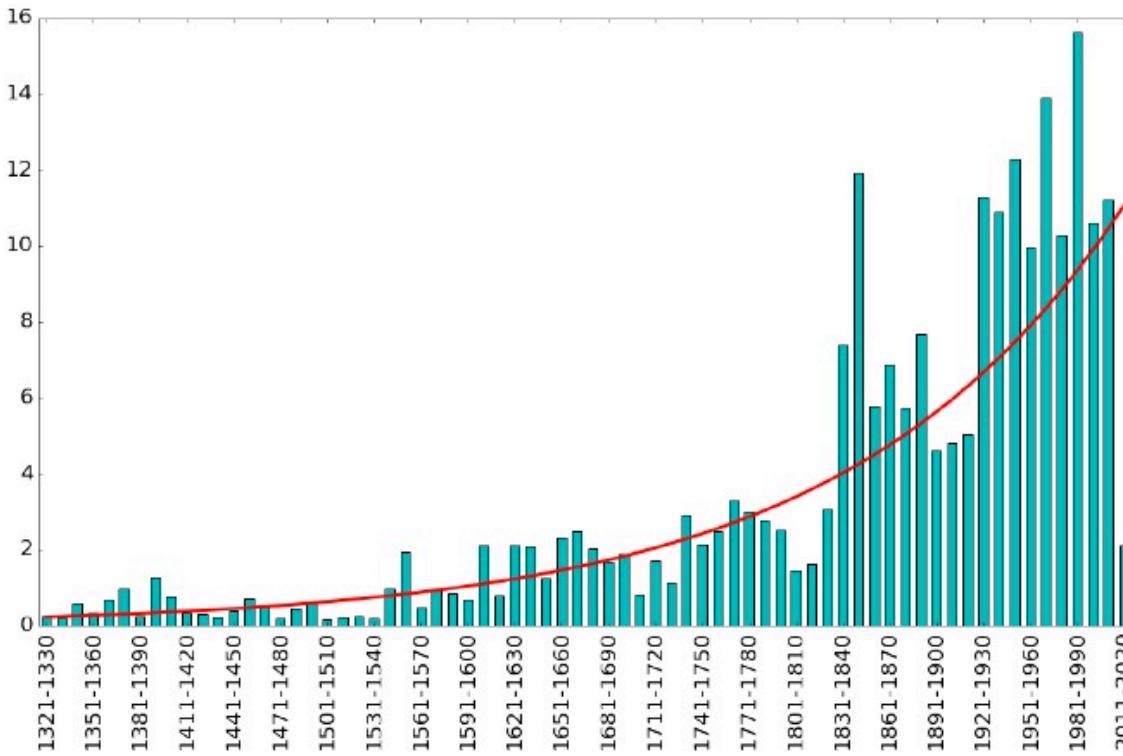
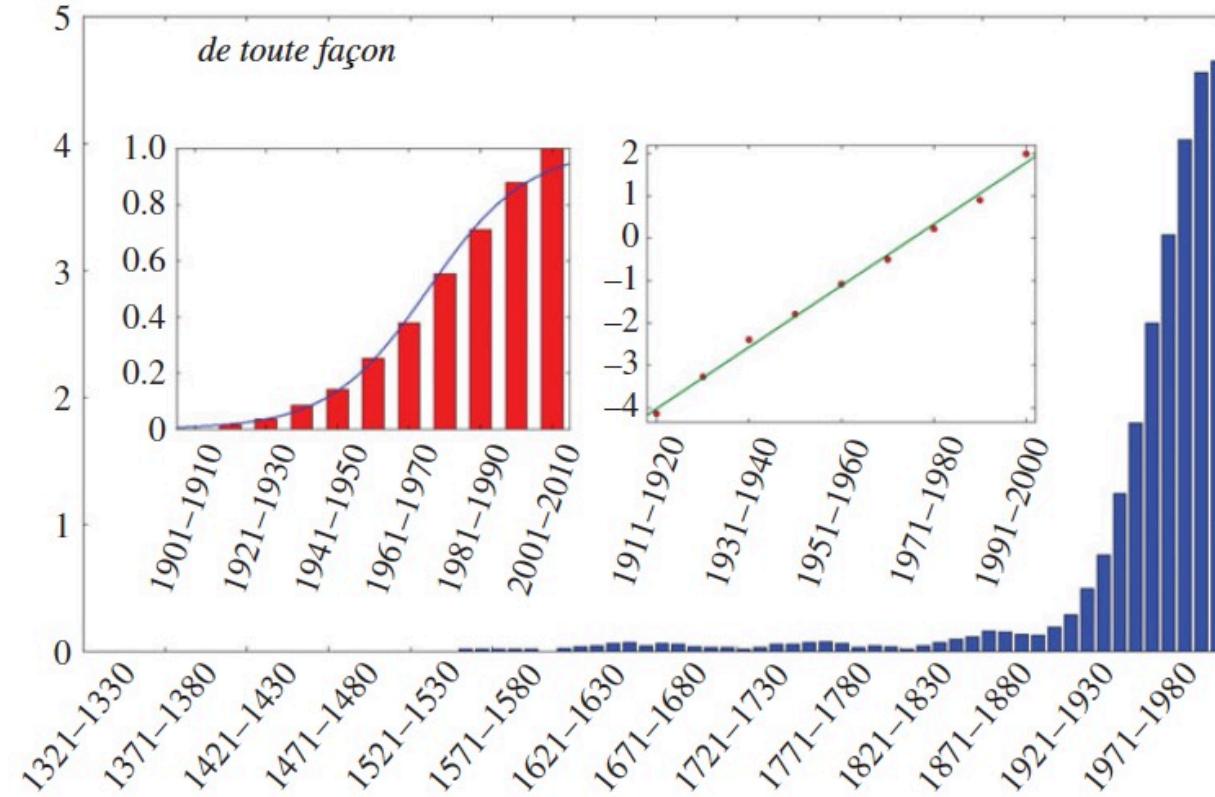


FIG. 10. Number of millions of occurrences per decade in the Frantext database. Exponential fit is shown by a red line.

- 今回は、“tout à l'heure”のような熟語表現に注目
  - 内容語と異なり、外的な要因に影響されない

# 最初の発見

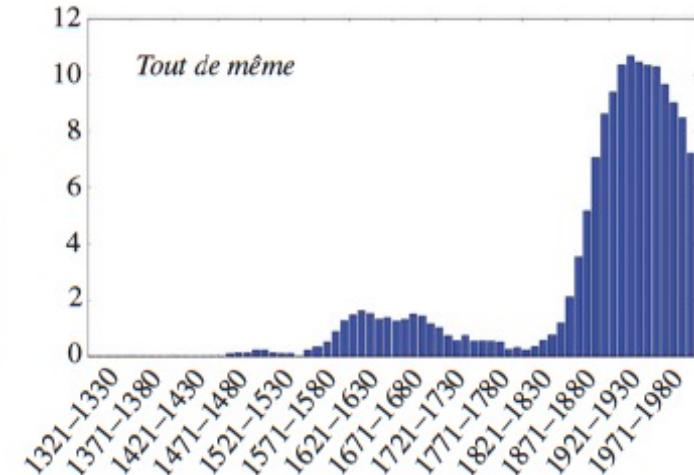
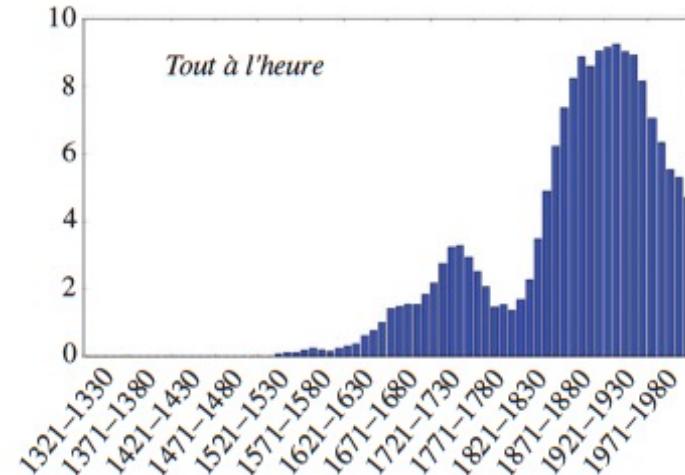
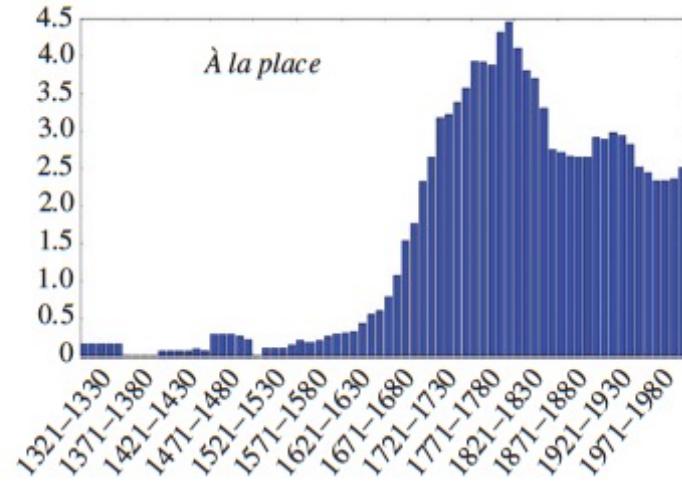
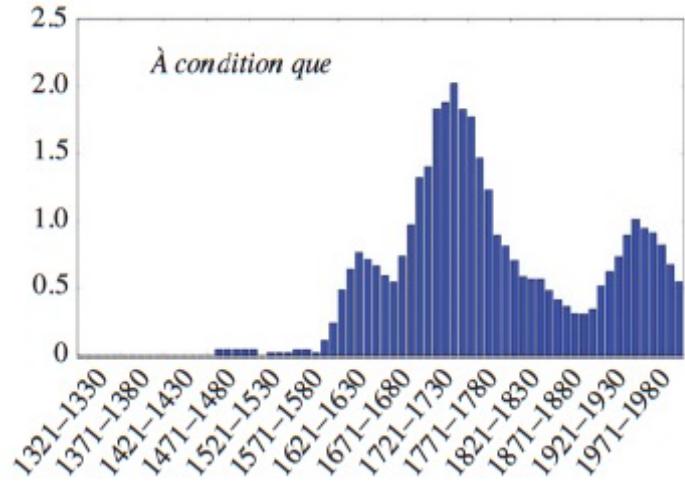
- S字でピークに達する前に、何世紀にもわたって一定の出現がある



“de toute facon”的出現／S字バーストの前に350年も一定の出現

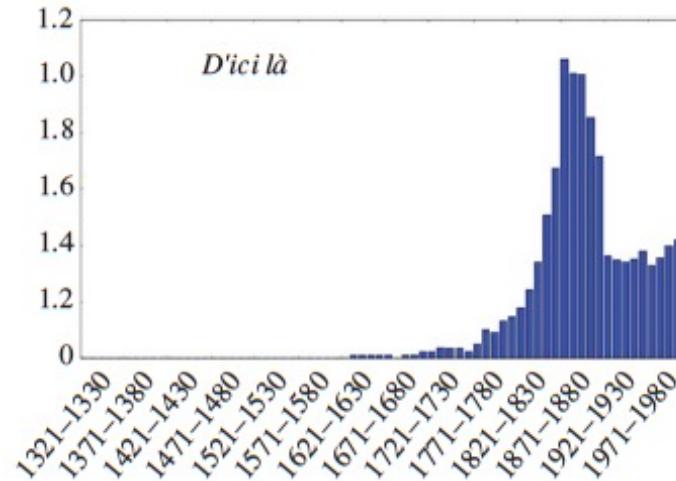
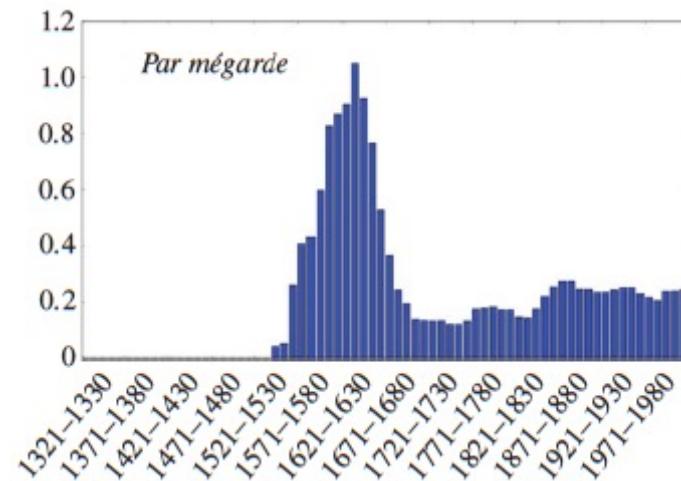
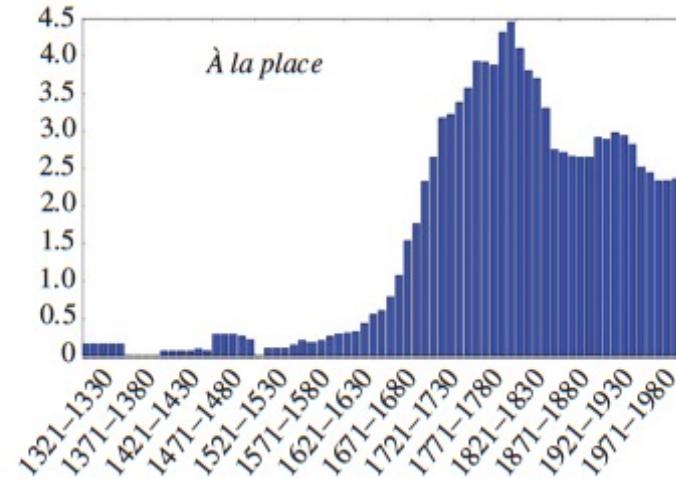
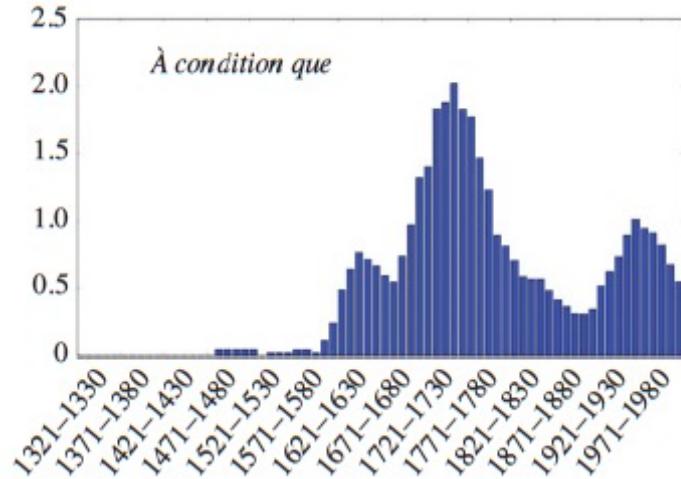
# 最初の発見

- S字でピークに達する前に、何世紀にもわたって一定の出現がある→緩和時間



# 次の発見

- S字でピークに達した後も、非自明な振動をする



# モデル化の方法

- ここでは、統計力学的に多数の群衆を考える代わりに「代表的なエージェント」を考えて、その振る舞いを微分方程式で記述するアプローチ
- Original formに対し、新しいformを使う確率  $p(x)$  を1次元上の微分方程式で表して、挙動を観察する
  - 複数のformが共存するようなことは、単純化のため考えていらない

# モデル化の概要

- X : new form  
(文脈 $C_0$ では元の意味)
- Y : original form  
(文脈 $C_1$ でのある概念を  
もともと表す単語)

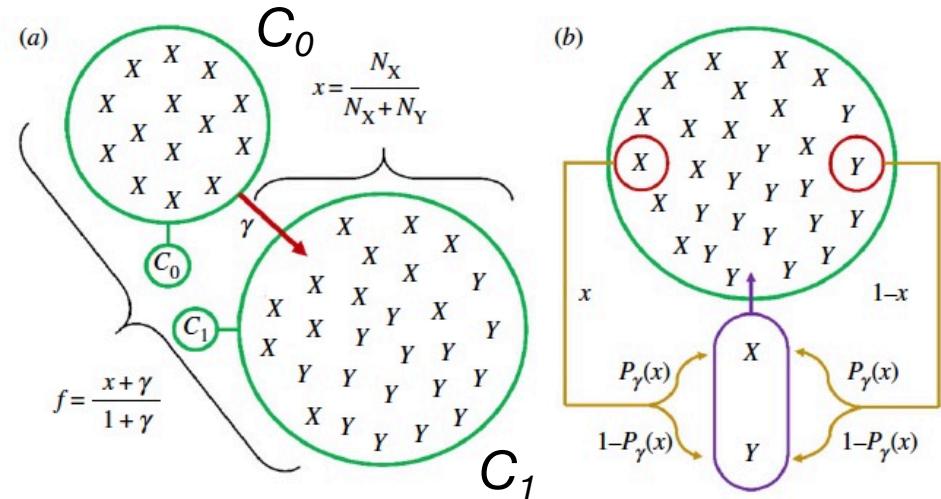
以下、基本的に $C_1$ での話

- $x(t)$  : 時刻tにおける $C_1$ でのXの出現回数  
 $f(t)$  :  $C_0$ と $C_1$ を含めたXの実効的出現回数

$$f(t) = \frac{N^1(t) + \gamma N^0(t)}{M + \gamma M} = \frac{x(t) + \gamma}{1 + \gamma}$$

- ここから、以下のような関数を仮定

$$P(f) = \frac{1}{2} \left\{ 1 + \tanh \left( \beta \frac{f - (1-f)}{\sqrt{f(1-f)}} \right) \right\}$$



# まとめ

- データ解析チームでは、統計的に透明な方法で意味変化の解析を行い、どんな言葉が意味変化するのか、そもそも語義の数は幾つなのか、などの網羅的なモデル化を行った
- 統計力学チームでは、特に物理分野で蓄積されている言語の物理モデルをもとに検討を行い、意味変化の微分方程式モデルにまで行き着いた
  - これまでの変化研究は、綴りや文法といった「見える」現象の変化が対象
  - 意味の変化は、そもそも見えないので、先に意味変化のデータをデータ解析チームが抽出する必要がある  
→ 本シンポジウムの発表は、その結果が中心