微小時間における日本語の変化とその法則

荒牧英治* **　　増川佐知子*

*東京大学 知の構造化センター
**科学技術振興機構 さきがけ

eiji.aramaki@gmail.com　　sachiko.maskawa@gmail.com

1　はじめに

言語は変化する．古くは平安時代の「枕草子」にも，当時の若者言葉を嘆く場面がある[1]．では，言語の変化とはどのようなものであろうか？ この問題は「*言語学における未解決のミステリ*」[1]として，これまで国語学，言語学などの各分野で音韻，語法，語彙，統語など様々な観点から取り組まれてきた．それらは大きく分けて次の２つの観点に分けられる．

第１はある言語(language)全体からみた語彙の変化である[2]．例えば，現代の日本語で頻出する 1000 語のうち万葉集においても見られるものは 326 語のみであり[2]，この間，多くの語が入れ替わったことが分かる．他にも外来語の増加[3]や漢字使用頻度の減少[4]などが指摘されている．では，その変化をどのように解釈すればよいのだろうか？ 例えば，外来語を取り入れるということは語彙数が増えることを意味する．しかし，仮に限りなく語彙が増え続ければ，いつかは人間には使いこなせない膨大な語彙数の言語が出来上がってしまうだろう．では，語彙が増える一方で，死滅する語彙もあり，なんらかの釣り合いを保っているのであろうか? 本研究では，統計力学的観点から語の頻度変化における平衡性を検証する（リサーチクエスチョン１）．

第 2 の観点は個々の語(word)からみた変化である．先の「枕草子」には「里」「言う」「文字」など現代でもほぼ通じる語が使用されている．千年以上も前から使用されているのだから，これらは日本語に定着した基本的な語といってよいと考えられる．一方，近年流行した「～なう」といった表現は，どこか正式な日本語でないという印象を受ける．だからこそ，「乱れる日本語」といった危惧が社会問題として取りざたされるのであろう．このように，少なくとも主観的には，正式な語とそうでない語の切り分けが可能であるかのように思える．では，この境界は語の頻度変化上に反映されるのであろうか？ 例えば，定着した語は高頻度で安定して使用されており，そうでない語は低頻度領域を激しく移動していくといったような違いはあるのだろうか？（リサーチクエスチョン 2）

本研究では，ツィッター[3]での発言を統計処理し，これら 2 つのリサーチクエスチョンに解答を与えることを試みる．ツィッターデータは，書き言葉であるものの話し言葉と近い性質を持つと考えられ，言語の変化に鋭敏である．また，時間情報を伴っており，本調査に適している．

本研究は次の２つの特徴をもつ．

【語全体の調査】国語学分野の先行研究は，あらかじめ注目していた語について，その振る舞いを調査する場合がある[2,3]．一方，本研究は，すべての語での調査を行うため，バイアスがかからず，全体的な挙動を知ることができる．

【微小時間の調査】いくつかの大規模調査（大西調査[5]や凸版調査[6]など）では網羅的な統計調査を行った．しかし，10 年またはそれ以上の粗い粒度の時間を対象としている．

[1]「なに事を言ひても..（中略）『と』文字を失ひて，ただいはむずる『里へいでんずる』など言へば，やがていとわろし」（枕草子 第一九五段）

[2]「言語」といった場合，日本語や英語といった言語全体(language)を指すのか，個々の語(word)を指すのか曖昧である．本論文では，言語は language，語は word として用いる．

[3] http://twitter.com/

本研究は日単位という微小時間での使用頻度の連続的変化を調査する．

2　材料／コーパス

2008 年 11 月から 2009 年 9 月までの twitter の日本語発言クロールデータ（1.77 億発言）を用いた[4]．平均の回収量は 70 万/日程度である．ただし，2009 年 3 月から 4 月にかけては twitter の仕様変更で回収率が 1/100 に大きく下がっており，十分なサンプル数が得られなかった（以降，この期間を**半脱落期間**と呼ぶ）．

語（以降，本稿では形態素を語とみなす）を抽出するために全データを形態素解析器[5]にて解析し，1 日毎に過去３０日の使用頻度を集計した（スライド単位=1 日，ウィンドウ幅=30 日）．また，集計にあたっては，日によってクロール稼働率が異なるため相対頻度（対象となる形態素頻度／すべての形態素頻度）で正規化した．

本研究では，形態素単位での集計を行うので形態素と形態素の組み合わせの変化や，同じ形態素が異なる意味で使われるという変化は捉えられない．

3　RQ1：語彙の使用頻度は安定状態にあるか？

3.1 調査手法

2008 年 11 月 9 日から 30 日間を基準期間として語の使用頻度とその順位を保持しておき，Δt 時間経過後の使用頻度順位がどう変化したかを調査した（例を表 1 に示す）．変化は以下の尺度を用いて調査した．

- **N 位保存率**：基準期間で上位 N 位以内の語群が Δt 時間経過後にどれだけ N 位以内に残っているかの比率を算出した(図 1)．この指標では N 位以内での変化は追えない．
- **順位相関係数**：基準期間で上位 N 位以内の語の順位（の系列）が Δ t 時間経過後の順位とどれくらい類似しているかをスピアマン順位相関係数にて算出した(図 2)．この指標では N 位内部の順位の入れ替わりが考慮される．
- **頻度遷移分布**：語の相対使用頻度が基準期間から Δt 時間経過後にどのように変化したかを調査した(図 3)．

[4] http://d.hatena.ne.jp/code46/20100919/p1
[5] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

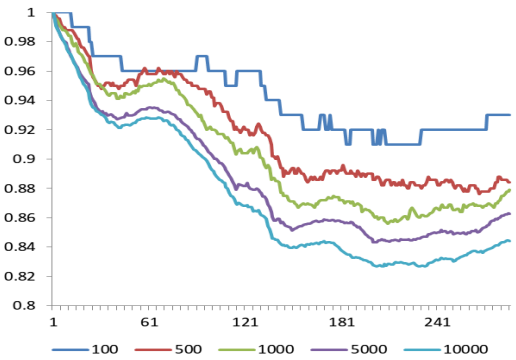

図 1: N 位保存率（Y 軸）と Δt (X 軸；単位は日).
N は 100, 500,1000, 5000 と 10000 の 4 つを調査した．

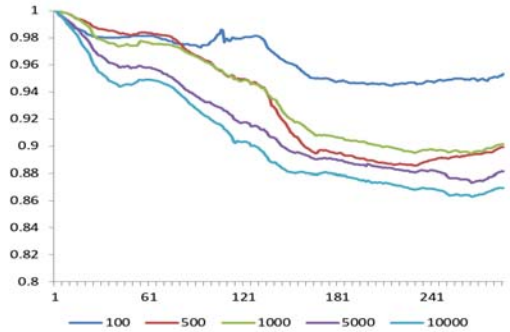

図 2: 順位相関係数と Δt.
表記は図 1 と同じ．

基準期間			2009/01/01		
順位	形態素	頻度	順位	形態素	頻度
1	。	2242380	1	。	52381
2	の	2124802	2	の	45417
3	が	1612961	3	が	33817
4	は	1512619	4	は	33271
28	です	360138	28	です	8729
29	「	358632	29	おめでとう	8102
30	？	356408	30	ございます	7850
31	」	353828	31	まして	7780
100	つ	87344	100	今	2154
1000	いえ	9128	1000	検索	213
10000	為替	630	10000	マズイ	13

表 1: 語の使用頻度は日毎に異なる．

2000 位以上→1000 位以内	1000 位以内→2000 位以上
2009, 野球, オリジナル, 暑い, インフルエンザ, ハルヒ, おん, マク, 新型, 染, けい, ドロリッチ	つかれ, 冬, おか, クリスマス, 秋, 雪, 鍋, なるほど, 2008, おめでとう, 寒, ござ, ただい, こちら

表 2: Δt 経過後(Δt=180 日)に成長した語と衰退した語．

3.2 微小時間で変化する語の使用頻度

図 1 と図 2 の両方においてグラフが右下がりになっており，基準期間から徐々に語の使用頻度が変化していることが分かる．N=100 のグラフにおいてもこの傾向が見られ，高頻度語でさえも 1 年に満たない短期間で入れ替わることが分かる．表 2 に急速に成長または衰退した語を載せる．多くの語は名詞であり，季節に依存する「雪」「鍋」「マスク」など

の一般名詞や「ドロリッチ」や「けい|おん」などの一時的に流行した固有名詞で占められている．

変化の速度については，基準期間で上位 10000 語が 1 日の発言に占める割合は約 88%であるが，10000 語のうち異なりで 15%(図 1 より)が 1 年に満たない間に入れ替わっている．

3.3 語の頻度変化における釣り合い状態

頻度遷移分布を図 3 に示す．横軸は基準期間での使用頻度(x_1)，縦軸は Δt=30 日での使用頻度(x_2)を表す．$x_2=x_1$ より上にある点は頻度を上げた語，$x_2=x_1$ より下にある点は頻度を下げた語を示している．図 3 では，$x_2=x_1$ を中心軸にほぼ対称な分布となっている．仮に対称であるとすると，基準期間で頻度 $x_1=a$ であった語が，Δt 経過後にどのような頻度に変化しているかの確率分布と，Δt 経過後に頻度 $x_2=a$ となった語が，もともと基準期間にどのような頻度であったかの確率分布が等しいということを示す．

このような制約を伴う状態遷移は自然界にはしばしば見られ，熱力学における気体分子運動や，企業の成長曲線[7]では**詳細釣り合い（detailed balance）**と呼ばれる．コルモゴロフ・スミルノフ検定によって詳細釣り合い状態であるかどうか確かめた結果，相対頻度 2.5e-5 より大きい語（使用頻度上位 4000 語に相当する）については，詳細釣り合い状態であることが示された[6]．詳細釣り合い状態にあると，語の頻度分布は時間の経過とともに変化しないことになる．つまり，現時点での言語が Zipf 則にしたがっているならば，これまでも，また，これからも常にその分布を満たす．このような遷移に関する強い制約を言語変化が持っていると言える．

4 RQ2:高頻度語と低頻度語に境界はあるか?

4.1 語の成長率

高頻度語と低頻度語の間に何らかの統計的境界が存在するのかを調査した．この際，個々の語の頻度変化の尺度として，語の成長率を用いた． ここでいう語の成長率とは，語がどのように頻度を伸ばした（または減らしたか）の尺度であり，以下に定義したものを用いた：

$$\text{語}w\text{の成長率} = \frac{\Delta t \text{経過後の語}w\text{の相対使用頻度}}{\text{基準期間での語}w\text{の相対使用頻度}}$$

例えば，語の成長率が 2 であるとは，Δt 期間中に使用頻度が 2 倍になったことを意味する．

4.2 語の成長率の分布

どのような成長率の語がどれくらいあるのかという分布を調査した．これは，成長率 0.1～10.0 を 0.005 単位で区切り，その区分に入る語がいくつあるかを集計して行った．集計は 語の使用頻度 1 位から 10000 位 を 10 等分し，その頻度区分ごとに行った（図 4）．

結果，どの頻度区分においても，成長率 1（頻度変化なし）を中心にした成長率分布が得られ，その左右の裾野が対称であった．

これは，使用頻度の上昇と下降の両方の変動が，どの順位でも等確率で起こっていることを示している．すなわち，ある区分で頻度が 2 倍に急上昇した語が n 個あれば，同数の逆の変化，すなわち 1/2 倍に急低下した語が n 個 あるということを示す．

分布の幅に注目すると，順位が大きくなるほど裾野の幅は広くなっている．例えば，頻度変化が 2/3 倍から 1.5 倍にとどまっている割合は，上位 1000 語では約 98%であるのに対し，9001 位から 10000 位の語では約 92%である．これは，低頻度であるほど，その成長率の分散が大きいことを示している．

4.3 語の成長率のばらつきと順位

語の成長率のばらつき（4 分位偏差）と順位との関係を示す（図 5）．4 分位偏差とは上位 1/4 位の値から下位 1/4 の値を引いて 2 で割ったもので，データに外れ値がある場合のばらつきを評価するために用いられる指標である．高頻度（上位）の語が頻度変化は安定しているならば，その成長率のばらつきは小さくなっているはずである．さらに，低頻度（下位）との境界があるならば，どこかでグラフが変化していると予想される．図 5 では，前者は成り立つものの，後者は直線が示すように成り立たず，成長率のばらつきから見て境界がないことが分かる．

以上をまとめると，語の頻度変化は高頻度領域ではおだやかに（小さな分散で），低頻度領域では激しく（大きな分散で）起こっている．また，その分散は頻度とよく相関しており，特に境界はないことが分かった．

[6]検定方法の詳細は文献[7]を参照のこと．

5　まとめ

本研究では微小時間での語の使用頻度の変化を調査した．結果，語の使用頻度は言語全体で常時変化しており，高頻度語においてさえも入れ替わりがあることが分かった．さらに，その変化は平衡状態を保っている．これらを総合すると，我々が日本語とみなしているものは，毎日いたるところで順位の交代がありつつも，常に同じ頻度分布を保った系であることが分かる（3 章）．

また，語の成長率に関しては，日本語の中に境界は見られず，我々が日本語と呼んでいる語に，頻度変化の観点からは区別はないことが示された(4 章).

ただし，以上の結論は，１年を超えていない観測によって得られたものである．よって，季節要因を除去できておらず，一般的な議論を行うためには，さらなる長期観察が必要である．

最後に，本研究ではツィッター上の発言を扱ったが，これは文書における話し言葉に相当する． 音声での話し言葉や，書き言葉など他の伝達形式においても本研究での知見が共通するかどうかは不明であり，今後の課題である．

謝辞:本研究は，JST 戦略的創造研究推進事業（さきがけタイプ）「情報環境と人」及び，科研費補助金(若手研究 A)による．本論文を書くにあたって有益な議論をいただい．た日本学術振興会（京都大学）遠藤智子氏，産業技術総合研究所黒嶋智美氏，金沢学院大学石川温先生，及び，貴重かつ膨大なデータを提供くださった兼山元太氏 (クックパッド) に感謝いたします．

参考文献

[1] 高田博行:歴史社会言語学の拓く地平，月刊言語，Vol38，No. 3，pp34-41，2009.

[2] 宮島達夫: 近代語彙の形成，国立国語研究所論集 3「ことばの研究 3」, 1967.

[3]飛田良文:明治以降の語彙の変遷, 言語生活 182 号, 1966.

[4] 安本美典: 漢字の将来，言語生活 137 号，1963.

[5] 大西雅雄:日本基本漢字, 三省堂, 1941.

[6] 凸版印刷:単語出現頻度調査, 1976

[7] 青山秀明，家富洋，池田裕一，相馬亘，藤原義久: パレート・ファームズ～企業の興亡とつながりの科学～，日本経済評論社，2007.

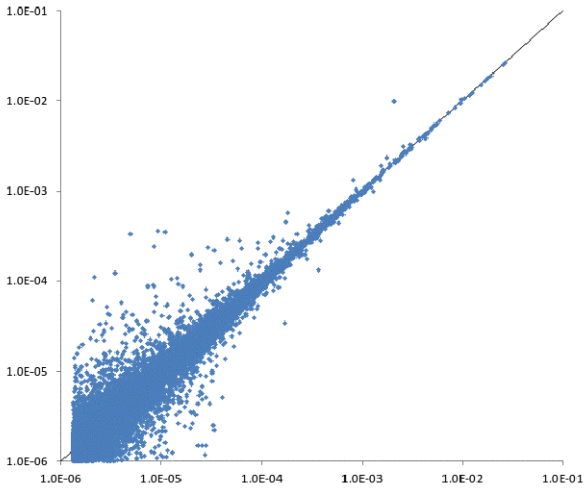

図 3：語の相対頻度の変動（Δt=30 日）
X 軸は基準期間での相対頻度 x_1, Y 軸は Δt 経過後の相対頻度 x_2.を示す.

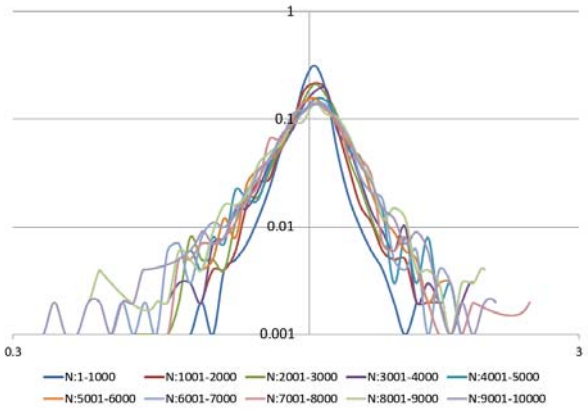

図 4: 語の成長率分布 (Δt=30 日)
X 軸は語の成長率，Y 軸, 成長率の確率分布を示す．N は順位を表し，1000 位ごとに線を変えてある

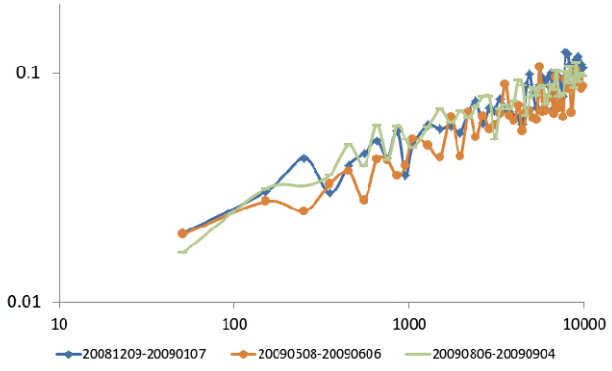

図 5:順位と成長率のばらつき
X 軸,語の順位．Y 軸, 成長率の 4 分位偏差