マルチエージェントシミュレーションによる不規則動詞の規則化に対する人口流入の影響

東条研究室
1310034 鈴木啓章

1　はじめに

現在の日常的に用いられている不規則動詞はおおよそ 180 語存在し、会話の中に出現する動詞の約 70% が不規則動詞である (be, have,etc...)［1］. これらの不規則動詞は Old English 時代 [AD 800 頃] に強変化動詞 (strong verb) と呼ばれ主に母音が変化することによって現在形、過去形、過去分詞などの活用形が生成されていた. Modern English では例外も含め 9 クラスに分類され、クラス内にも細かい分類がなされている［2］. しかし、上記の不規則動詞には単純に接尾辞 [-ed] をつける規則的な活用に変化しているという現象が見受けられる. 英語の歴史的な流れの中では、Old English から中期英語時代における海賊によるイングランドの侵略、ノルマン征服などの人口流入を伴った言語接触により不規則動詞の規則化の誘発、またその加速が起こっている.

本研究ではこの不規則動詞の規則化に対する人口流入の規模、頻度をシミュレーションによって検証することを目的とする. 検証のために遺伝的アルゴリズム［3］(以下 GA) をベースに、エージェントコミュニケーションと変化を進行させるような (外圧) を組み込んだモデルを作成し、複数世代を通したシミュレーションを行う.

2　研究背景

本章では、英語の時代区分と、歴史から見る人口流入と言語接触の影響について説明する.

2.1　英語の時代区分

英語の年代区分は、ノルマン征服など歴史的な事実を区切りに用いるが、3 区切りや 6 から 7 つに区切るモデルも存在する. 表 1 に 4 つの時代に区切るモデル［4］を示す.

表 1　英語の時代区分

A.D 500 -1150	Old English
A.D 1150 -1450	Middle English
A.D 1450 -1700	Early Modern English
A.D 1700 -	Modern English

各時代について簡単に説明する. Old English 時代は大ブリテン島南部でアングル、サクソン、ジュート族によって言語が確立された時期である. その後、ノルマン征服によってイタリック系言語であるノルマンフランス語との接触による影響が出始めた時代が Midle English 時代である. 活版印刷技術が西ヨーロッパに広がりはじめた時期が Early Modern English 時代、アン女王の時代以降が Modern English 時代となる.

1 で述べた区分において Old English 時代に母音交替によって活用していた動詞が不規則動詞である. またそれ以外の動詞は弱変化動詞と呼ばれ接尾辞に [-ed] をつけて活用していた.

2.2　歴史から見る人口流入と言語接触の影響

本節では Old English 時代から Middle English 時代における言語接触［5］について説明する. まず A.D800-1066 ごろに Old English に影響を与えたのは海賊 (Viking) である. 海賊はスウェーデン、ノルウェー、デンマークに居住していたデーン人と呼ばれる民族である. また海賊は古北欧語話者であった. この接触によって Old English では三人称単数の語尾に [-s] をつけるようになり、強変化動詞の規則化が始まったとされている.

次に A.D1066-1345 ごろのノルマン征服の影響について述べる. この接触によりノルマンフランス語から Old-Middle English に対して大量の語彙 (約 1 万語) の流入のが起こった. また、文法などの言語構造の簡略化や、海賊の影響で始まった規則化傾向も上昇したと考えられている.

以上より、侵略、征服など様々な言語接触を経験していることが英語の特徴である. 図 1 に海賊が居住していた地域 (青) とノルマン人によって征服された地域 (赤) のおおよその位置を示す. 図 1 より言語接触影響の規模的にもノルマン征服が大きのではないかと予想される.

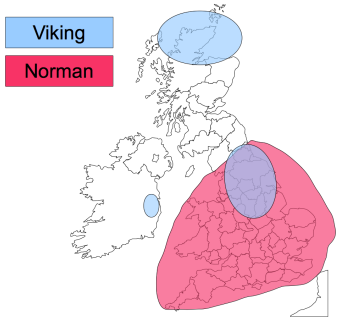


図 1　海賊とノルマン人の居住、征服地域

3　先行研究

本章では、コーパスを用いた統計的な研究について述べる.

3.1　不規則動詞の出現頻度と規則化速度

Lieberman らの研究 [6] では、CELEX [7] を用いて不規則動詞の規則化速度が CELEX 内における相対出現頻度と時間の関数として表現できることを示した. 相対出現頻度は CELEX が持つ約 1770 万語の中に出現する動詞、約 331 万語を利用したものである. 不規則動詞の相対出現頻度の例を表 2 に示す.

表 2　不規則動詞の相対出現頻度 (一部抜粋)

出現頻度	動詞
10^{-1} – 1	be, have
10^{-2} – 10^{-1}	come, do, get
10^{-3} – 10^{-2}	begin, draw, help
10^{-6} – 10^{-5}	bide, shrive, pew

次に、Old(A.D 800)、Middle(A.D 1200)、Modern English(A.D 2000) の各年代である頻度を持つ不規則動詞数を求める. 図 2 の色付きの折れ線グラフがそれに当たる. その不規則動詞の数、頻度、時間を組み合わせ各年代のグラフに近似するような式 1 を導出する.

$$I(\omega,t) \approx \frac{0.4467 \times exp\left(\frac{-4.9045*10^{-6}\times(2000+t)}{\omega^{0.5088}}\right)}{\omega^{0.7099}} \tag{1}$$

$I(\omega,t)$ は、t 年後の頻度 ω である不規則動詞の数を表している. 式 1 により、未来の不規則動詞の数が予測可能になった. 図 2 の網目の部分が式 1 について $-2000 \leqq t \leqq 2500$ を与えた計算結果である. 図 2 より高頻度の不規則動詞は元から数が少なく、また規則化の影響も非常に受けにくい事がわかる. 逆に低頻度になるほど時間経過に従って不規則動詞の数は減っていくことが示されている.

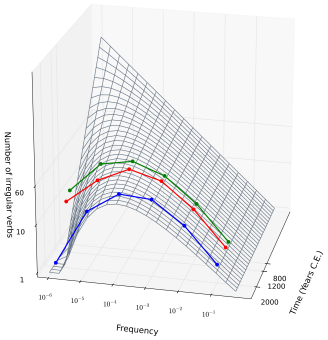


図 2　不規則動詞の分布

3.2　語形変化と外圧の影響

Ghanbarnejad らの研究 [8] は、綴り字の改正、不規則動詞の規則化などの語形変化をシグモイド $\sigma(x) = \frac{1}{1+\exp(-x)}$ として捉え、変化を引き起こす力 (外圧, 内圧) を数値的に捉えようとしたものである. 具体的には式 2 の a を外圧、b を内圧 (コミュニティ内のエージェントの接触確率) とする. 式 2 の一般解を求め、実データに近似することで a、b の値を求める. 実データは Google Ngram Corpus [10] を用いる. これは 1800 年から 2000 年に出版された電子書籍データ (3610 億語) を使って Ngram 統計を行ったものである. [8] では 1-gram を用いる.

$$\frac{d\rho(t)}{dt} = (a + b\rho(t))(1-\rho(t)) \tag{2}$$

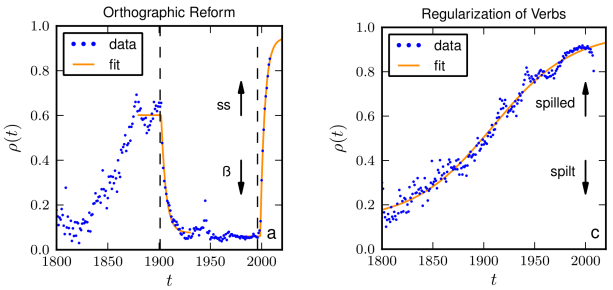
具体例としてドイツ語の綴り字の改正 [9] と、不規則動詞の規則化について述べる.

1996 年にドイツ語の [ß] を [ss] と綴るとする改正が行われた. すべての [ß] が改正されたのではなく長母音の後ろに [ß] が来る場合に限りそのまま使用されている. この改正の浸透の仕方を [10] を用いて年代ごとに計算する. 計算式は式 3 である. 計算結果は、ある年代における [ss] の使用割合を示している. そのデータを式 2 に近似することで a、b の値を得る. 同様に不規則動詞も年代ごとに式 4 によって、動詞ごとの規則化割合を求める.

$$\rho(t) = \frac{freq(ss)}{freq(ss) + freq(ß)} \tag{3}$$

$$\rho(t) = \frac{freq(regular)}{freq(regular) + freq(irregular) + freq(pastparticiple)} \tag{4}$$

以上の結果を図 3 に示す. 青のドットが実データ、オレンジの曲線が近似曲線である. a, b の値からドイツ語の綴り字改正においては、大きな外圧 (国規模で改正を進める) によって急激な変化がもたらされている. 逆に内部で変化を進めようとする働きは小さい. 不規則動詞の規則化においては、外圧は非常に小さく、コミュニティ内部で変化を進めようとする力が働いていると言える.



ドイツ語の綴り字の変化
$a = 0.229, b = 0.0$

不規則動詞の規則化
$a = 0.001, b = 0.030$

図 3　1800 から 2000 年の各語形の変化 [8] より引用

4　研究内容

先行研究 [6, 8] で、不規則動詞の規則化は頻度と時間の関数によって表現できること、また語形変化における外圧と内圧の関係を述べた. これらは実データに基づいており、現実世界における言語変化を表している. よって本研究では不規則動詞の規則化は [6] に従うとし、シミュレーションによって図 2 を再現する. シミュレーションの中では、人工的な言語を持ったマルチエージェント環境 (コミュニティ) を仮定する. これは [8] にあるようなコミュニティ内部で起

こる力も表現可能にするためである. また、外圧として人口流入を扱う. これは言語コミュニティが様々な大きさ、頻度で外圧 (人口流入) に晒されながら複数世代を通して不規則動詞を規則化させていく状況を再現している. そして図 2 を再現することによって、人口流入の大きさ頻度も明らかにできる. また単純に人口流入が起きれば規則化が進むのではなく、エージェント同士がコミュニケーションを行うことで「より規則化が進むのか、またはそれを阻止しようとするのか」現実に近い状況でシミュレーションを行うことができる. これは本研究の大きな特徴である.

4.1 シミュレーションベースの構築

本研究で行うシミュレーションのベースとして GA を用いたモデルを構築した. 表 3 に GA の概要を示す. なお、実験は抽象的な設定で行い、実際には動的に変化すると考えられる関数等もすべて固定してある. また人口流入に関しても単純に現在は外圧とだけ考えている.

表 3　GA の概要

個体数	250 個体 (各々が 1 つのコミュニティ)
PTYPE	コミュニティ内の規則化率 (20% を 200 と表現)
GTYPE	PTYPE の 2 進数 (10bit)
交叉方法	一様交叉 (60% で交叉)
選択方法	トーナメント方式 (サイズ 2)
評価関数	ある不規則動詞の使い易さ
突然変異	外圧と仮定
バイアス	正規分布 (変化を加速させる働き)

表 3 における評価関数、突然変異、バイアスについて以下で説明を行う.

4.2 評価関数

ある不規則動詞について、それが不規則変化、規則変化どちらが使い易いのか (適応度) を評価する関数である. この関数は環境によって動的に変化することが予想される. 関数の形を決める要因は、使用頻度やエージェント同士の会話の伝わりやすさなどが考えられる.

4.3 突然変異とバイアス

本研究では突然変異は、コミュニティにかかる外圧と考えている. つまり、外圧がかかれば変化が加速されるバイアスがかかる. 例えばコミュニティ内において外圧によって規則化が進んでいるとする. コミュニティ全体が不規則変化、規則変化を使用する人間が混在する不安定な状況化にとどまることは考えにくい. よって、抑える力が働かない限り規則化は推し進められる. この力をバイアスとして表現する. また、外圧の影響が徐々に浸透していき、浸透しきれば影響は少なくなると言い換えることもできる. バイアスは図 4 の正規分布で与えられる. モデル上では外圧がかかった世代から何世代かに渡り、各個体の PTYPE を変化の方向に押し上げる働きをする. 上記の例に従えば各個体に値を加算し規則化率を押し上げていく. 加算する値は正規分布に従い、小さい値から徐々に大きな値になりまた小さくなっていく.

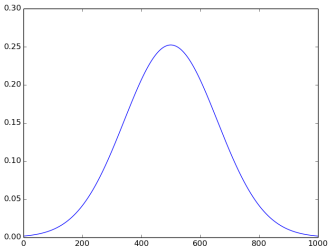


図 4　変化を加速させるバイアス

4.4 実験

4.2、4.3 の元、規則化が進行するシミュレーションを行った. GA の設定は表 3 と同様である. コミュニティの規則化率はスタート時 20% とし、評価関数は分散 30000、平均 1000 の正規分布を仮定した (固定). つまり規則化を進めたほうが適応度が上がる設定である. 外圧がかかる世代は 50 世代目とし、その後 20 世代にわたってバイアスがかかるよう設定した. 図 5 に 100 回シミュレーションを行った平均を示す.

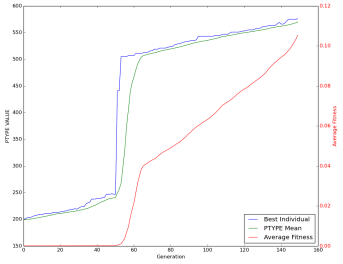


図 5　変化を加速させるバイアス

図 5 の縦軸は全コミュニティの平均規則化率を示す. 図中の緑の曲線がその移り変わりを表現している. 青線は最も適応度の高いコミュニティの規則化率、赤線は平均適応度である. 緑の曲線からバイアスがかかりはじめた 50 世代目から徐々に上昇を始め、最も強いバイアスがかかる中間付近では急激な上昇を見せている. その後はまたゆっくりとした変化に収束する様子が表現できている.

5 まとめと今後の課題

GA を元にシミュレーションモデルの作成を行った. 外圧が加われば変化が加速されるという実際の語形変化に則した変化を意図的な突然変異と数世代に渡るバイアスという形で表現した. 今後の課題は、規則化だけでなく不規則化に対応できようなモデルに拡張する必要がある. 特にバイアスを動的に設定できる仕組みを導入する. 例えば外圧の大きさだけ設定すれば、バイアスの大きさや影響する世代を自動的に判断し実行する.

またコミュニティの設定のために、人工的な動詞の作成やエージェントコミュニケーションを導入しコミュニティの状況によって変化する評価関数 (複数の要因パラメータを持ち、関数の形が変化する) を作成する. 現段階では外圧をただ PTYPE の値を変化させるだけで扱っている. 今後は人口流入を明確に定義し、扱えるようにする.

参考文献

［1］ *Pinker, S. The irregular verbs. Landfall 8385 (Autumn issue, 2000)*

［2］ *Pinker, S , Prince, A. On language and connectionism: analysis of a parallel distributed processing model of language acquisition. Cognition 28,p73-193 (1988)*

［3］ 伊庭 斉志, 遺伝的アルゴリズムの基礎-*GA* の謎を解く オーム社 *(1994)*

［4］ *Tom McArthur, THE ENGLISH LANGUAGES, Cambridge University Press (1998),* 英語系諸言語, 牧野武彦 監修, 山田 茂, 中本 恭平 訳, 三省堂 *(2009)*

［5］ *Philip Gooden, THE STORY OF ENGLISH (2009),* 物語 英語の歴史, 田口孝夫 監修, 悠書館 *(2012)*

［6］ *Erez Lieberman, Jean-Baptiste Michel, Joe Jackson, Tina Tang , Martin A. Nowak, Quantifying the evolutionary dynamics of language. Nature Vol449 (11 October 2007)*

［7］ *CELEX http://wwwlands2.let.kun.nl/members/software/celex.html*

［8］ *Ghanbarnejad, Fakhteh, et al. Extracting information from S-curves of language change. Cornell University Library arXiv preprint arXiv:1406.4498 (2014)*

［9］ *Chris Upward, Spelling Reform in German. Journal of the Simplified Spelling Society, J21, 1997-1 pp22-24,36*

［10］ *Google Ngram Viewer, https://books.google.com/ngrams*