

言語処理的アプローチによる better off 構文の定着過程の説明

永田亮¹ 大谷直輝² 高村大也³ 川崎義史⁴

¹ 甲南大学 ² 東京外国語大学 ³ 産業技術総合研究所 ⁴ 東京大学

nagata-better_off @ hyogo-u.ac.jp.

概要

本稿では、言語処理的アプローチにより、**better off** の新しい用法が定着した経緯に関する仮説を検証する。具体的には、コーパス中の用例をクラスタリングすることで用法変化を定量化し、言語学的な観点から提示された仮説に一致するかどうかを検証する。その結果は、従来の仮説を大枠で支持するものの、派生するいくつかの可能性を示唆する。また、分類結果に基づき、関連した新たなリサーチクエッションを提示する。更に、本アプローチの一般的な分析への拡張可能性についても示す。

1 はじめに

現代英語の better off は意味的にも統語的にも不規則性がある [1]。語源的には、better off は well off (裕福な) の比較級とされ、主に形容詞の叙述用法で使われる (以降、この用法を**字義的用法**と表記)。次のような例が典型的である：

(1) I am better off than before. (以前より裕福だ)

一方、better off の新しい用法は、

(2) a. You're better off dead. (死んだ方がましだ)

b. You're better off buying it. (買った方がましだ)

のように、意味的にも統語的にも字義的用法とは異なる。例文 (2) では、「裕福な」という意味は失われ、代わりに「～した方がましだ」という意味が表される。更に、統語の観点からは、better off は be 動詞の補語にもかかわらず、更にそれ自体の補語と解釈できる形容詞 (dead) や分詞 (buying it) をとるという特殊な構造を示す。以降では、補語と解釈できる句を従えるこの用法を**新用法**と呼ぶことにする。

本稿の主題は、新用法がどのように定着したかを説明することにあるが、その仮説として次の四段階を経る大谷の仮説 [2] がある：

(Stage 1) 字義的用法のみが存在する初期段階

(Stage 2) 字義的意味が一般化する段階

(Stage 3) 仮想状況を想定した用例が増加する段階

(Stage 4) 新用法が定着する段階

(Stage 1) では、例文 (1) のような経済的な状況のみを表す。(Stage 2) では、適用範囲が拡張され、経済的な状況以外に対しても「より良い」という意味を表すようになる。(Stage 3) では、現状とは異なる仮想的な状況を想定した用例が増える。例えば、You would be better off without her. (彼女がいないほうがよいだろう) のような例である。この例は、「彼女がいない」という仮想的な状況を想定している。前段階で適用範囲が拡張されたことで、仮想的な状況との比較が増えたと考えられている。そのため、非現実的な状況を表す would などの助動詞や if 節との共起が多くなる。また、仮想的な状況や条件を表す語句 (上の例文では without her) が better off に後続することも増える。(Stage 4) では、better off に後続する仮想的な状況や条件を表す語句が、その出現位置のため、主語の仮想的な状態を表す補語として再解釈される。その結果、文法的に不適格な例文 (2) のような構造が許されるようになる。それに伴い、仮想的な状況が強調され、(仮想的な状況より)「～の方がましだ」という意味を表すようになる。以上のように、この仮説では、字義的用法から出発し、適用範囲が拡張され、更に仮想的な状況を想定した事例が増え、新用法が定着すると予想する。

本稿では、言語処理的なアプローチで、この仮説を検証する。具体的には、コーパスから得られる情報のみに基づいて better off の事例をクラスタリングし、用法変化 (年代と各用法の頻度) の図として表す。その図において仮説の各段階が観測されるかを検証する。本稿の成果は次の3点に要約される。第一に、検証結果は大谷の仮説に大枠従うものの、細かい点で別の可能性があることを示す。第二に、結果の詳細な分析により、新たな仮説とリサーチクエッションを提示する。第三に、以上の結果を踏まえつつ、今回のアプローチに基いた、より一般的な言語分析の可能性を示す。

2 分析方法

分析の基本はクラスタリング結果を用法変化のグラフとして表すことにある。横軸と縦軸をそれぞれ年代と各クラス内の事例数としたグラフである。

分析対象として、Corpus of Historical American English (COHA) [3] 中のノイズを除去した CCOHA [4] を用いる。1810～2010 年代の文書を収録する約 4.7 億語からなる歴史コーパスである。各文書には、作成年とジャンルの情報が付与されている。付録 A に示す前処理を施し、better off を含む 2,591 文を得た。

2.1 事例のクラスタリング

事例のクラスタリングには、*k*-means++ [5] を用いる。*k*-means++ を適用するためには、各事例をベクトルに変換する必要がある。本研究では、言語モデルである BERT (bert-base-uncased) を用いた。各事例を文単位で BERT に入力し、better off の出現位置に対応する最終層の隠れ状態を連結してベクトルとした。なお、bert-base-uncased を用いているため、大文字／小文字の区別はされない。

クラスタ数 *k* は、仮説が想定する四段階に合わせて 4 とした。また、次に述べる 人手の分類に合わせて、*k* = 3 としたクラスタリングも行った。なお、クラスタ番号の割り当てには任意性があるが、年代と頻度から決定した。すなわち、出現年代が古い順に 1 から番号を割り当てた。同じ年代で同時に出現するクラスタについては、頻度が高い方に若い番号を割り当てた。以降では、表記の煩雑さを避けるため、各事例に割り当てられるクラスタ番号をクラスと表記することとする（クラスタそのものに言及するときはクラスタと表記する）。これに合わせて、人手の分類結果もクラスと表記する。

2.2 人手による分類

比較のために、人手でも 2,591 事例を分類した文献 [2] を参考に、次の 3 クラスに分類した。なお、判断に迷う事例および形容詞の叙述用法以外の事例は、以降の分析では対象外とした（2,591 事例のうち 24 が該当した）。

（説明の都合上、この順で導入するが、）クラス 3 は新用法に対応する。判断基準は、better off が形容詞（例：dead）、分詞（例：learning it）、to-不定詞句（例：to stay here）に従える事例とした。この条件を満たす事例は全てクラス 3 とした。

クラス 2 は、仮想的な状況を想定した用法に対応する。文献 [1] に倣い、仮想的な状況を表す語句を伴う事例とした。具体的には、would や may などの助動詞、if 節、probably、maybe などの法副詞を含む事例をこのクラスとした。

それ以外をクラス 1 とした。仮説に対応させると、(Stage 1) と (Stage 2) に属する事例をまとめたクラスとなる。本来であれば、字義的用法と経済的意味以外に拡張された用法を区別して二つのクラスに分けるべきであるが、その判別は非常に困難であるため、文献 [2] に従い両者を一つにまとめた。

以上の基準に基づいて、2,591 事例に対して人手でクラスを割り当てた。作業は、第一著者と第二著者で独立に行った。各作業者が独立でシングルチェックを行った。この時点での両者の一致率は 95% ($\kappa = 0.92$) であった。その後、すり合わせを行い最終的なクラスを決定した。

3 結果

図 1 と図 2 にそれぞれ、クラスタリング／人手の分類により得た用法変化の図を示す。横軸が年代、縦軸が各クラスの事例数に対応する。両図は非常に類似するといえる。二つのグラフ間の平均二乗誤差（の平方根）は 8.5 であった¹⁾（人手の分類結果から得た図同士の平均誤差は 2.1）。また、クラスの一一致率は 72% ($\kappa = 0.56$) であった（人同士 95% ($\kappa = 0.92$))。

仮説の検証に先立ち、クラスタリングの結果が人手の分類に対応するかを確認した。各クラスタの重心²⁾に最も近い事例 10 件を観察した。

クラスタ 3 については、全 10 件とも補語に現在分詞に従える事例であった（一方、クラスタ 1 と 2 については、一件も該当しなかった）。クラスタ 3 の全事例を確認したところ、83% (=430/521) は、形容詞、現在分詞、to-不定詞のいずれかを従える用例であった。したがって、クラスタ 3 は新用法に（ほぼ）対応するといえる。

クラスタ 2 についても人の分類基準に対応することが確認できた。重心最近傍 10 件中 9 件は、would などの助動詞を伴う事例であった。残り 1 つは、“He believes they are better off here.” というものであった。従属節の内容は実際に生じた事象でなく、主語が想定する事象であるので、この事例も仮想的な状

1) 平均的に各年代で 9 事例程度ずれがあることを意味する。

2) 直感的には、クラスの代表的な事例（をベクトルで表したもの）と解釈できる。

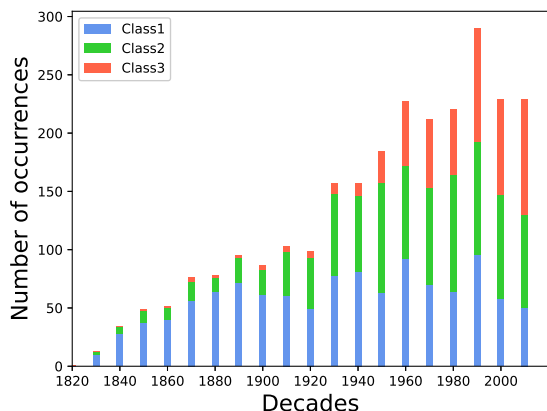


図1 クラスタリングを用いて得た用法変化の図 ($k=3$).

況を表すと解釈することもできる。その他のクラス2の事例をランダムに36件サンプリングしたところ70%は人手の分類基準を満たす事例であった。

クラス1は、完全には一致しないが、人間の分類基準に準ずることを確認した。重心最近傍の10件のうちクラス2,3の条件を満たすものはそれぞれ0件、3件であった。ただし、他の二つのクラスに比較すると多様な事例を含んでおり、別の解釈も存在する可能性もある。

コーパスから得られる情報のみに基いた教師なし分類（クラスタリング）と人手の分類がよく一致することは興味深い。単語の分布から見ても better off の三クラスは特徴的であることを示唆する。

話を本題に戻し、用法変化の図に基いて仮説の検証を行う。図1, 図2共に、1960年代頃より新用法であるクラス3が増加し始める。この時期以降が新用法が定着した（Stage 4）に対応する。その少し前の1910～1950年代には、クラス2（仮想的な状況を表す語句を伴う用例）が増加傾向にある（参考のために、比率にした図を付録Bに示す）。ただし、人手で作成した図2では、その増加はそれほど顕著ではない。クラス3の定着と並行してクラス2が増加しているようにも見える。一方で、2000年代以降は若干であるが減少している点も興味深い。

図3に、クラス数を4にした場合の結果を示す。紙面の関係から根拠は付録Bに譲るが、各クラスターの解釈は次のとおりである。クラス1から順に、(1) than を伴う典型的な比較の意味が強い用法、(2) than 以外の語句（here など）が後続する用法、(3) 仮想的な状況を表す語句を伴う用法、(4) 新用法と解釈した。図3から、1860年代ごろまではクラス1の事例が主流であることがわかる（（Stage 1）に対応）。その後、クラス1は相対的に減少し、代わりにクラ

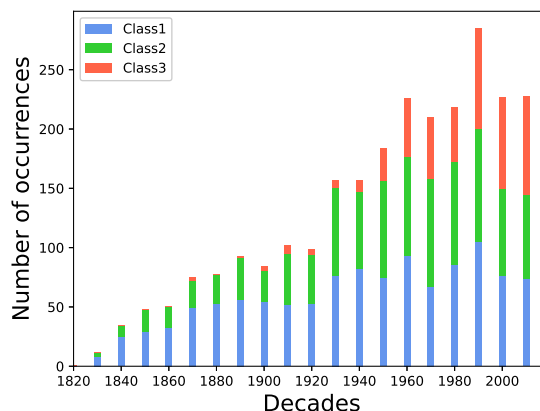


図2 人手の分類結果から得た用法変化の図。

ス2とクラス3が増加する（（Stage 2～3）に対応）。新用法であるクラス4が定着する前の1920～1950年代に、クラス3が増加しているが顕著ではない。

以上の結果は、次のようにまとめられる。用法変化の図は、大谷の仮説を大枠で支持する。特に、字義的用法から仮想的な状況を表す語句を伴う用法、新用法に移行している傾向はどの図でも確認できる。一方で、細かい点では仮説に必ずしも一致しない部分も見られる。次節で、仮説との差異を考察する。

4 考察

仮説では、仮想的な状況を想定する用法が増えたことにより、新しい意味と構造が生じたと予想する。図2では、新用法の better off が生じたのは1870年代である。その後、新用法が少数使われる期間がしばらく続く（以降で何度か取り扱うため、この現象をトッピング現象³⁾と呼ぶことにする）。図では、クラス2が増加した1910～1950年代以前にトッピング現象として新用法は生じている。このことを考慮すると、1860年代以前にクラス2が増加した（すなわち（Stage 3）が起こった）可能性も否定できない。残念ながら、今回用いた CCOHA では1860年代以前の better off の事例が少なく判断がつかない。

また、大谷の仮説とは（Stage 3）と（Stage 4）の順序が逆になる説も考えられる。すなわち、何らかの別の要因で新用法が生じ、その結果、仮想的な状況を想定する用法が増えたという説である。実際、図1～3はそうのように解釈もできることを前節で見た。新用法が生じた別の要因として、ある個人または一部のグループが新用法を使い始めたということが

3) グラフ上の小さなバーが、スイーツの上の粉砂糖やピザのトッピングのように見えるためこのように名付けた。

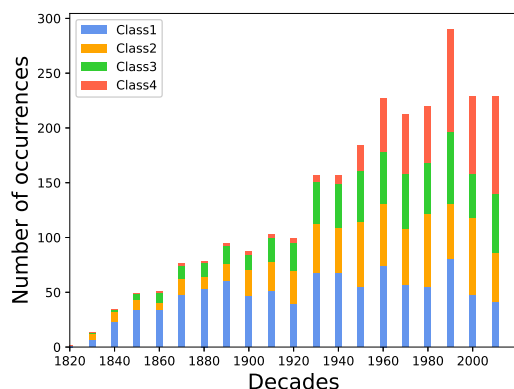


図3 クラスタリングを用いて得た用法変化の図 ($k=4$).

考えられる。各年代／ジャンルにおける新用法の出現頻度を求めてみると、1900年代以前は全て文学作品で出現することが確認できた（付録Bの図5に頻度情報を示した）。1910年代には、雑誌、ノンフィクションにも出現し、その後、全てのジャンルに広がる。事例数が少ないため、推測の域を出ないが作家のような創造的な文章を書く一部の人により新用法が作り出され、他者に広まった可能性もある。

更に、前節でみたように、2000年代以降になると仮想的な状況を想定する用法が減少傾向になる。クラス数4の場合、クラス3では62%の事例が助動詞を伴うが、クラス4になると39%まで減少する（具体的な頻度は付録Bの表1に示した）。この現象は、助動詞が担っていた仮想的な状況が *better off* に移行した語彙化 (lexicalization) の例として説明できる。すなわち、わざわざ助動詞でマーキングせずとも、*better off* だけで仮想的な状況を想起させられるようになったということである。この現象は、少なくとも、仮想的な意味を想定する用法が新用法に強い関わりをもつことを示唆する。

以上のように、今回得られた結果は、仮説の精緻化や修正を促す多くの示唆を与える。現段階では、大谷の仮説から派生した複数の説が考えられ、更なる調査が必要であると結論付けられる。特に、(1) 仮想的な状況を想定する用法が増えて新用法が増えたのか、その逆であるのか、(2) そもそも新用法がいつどのように生じたのか、(3) なぜ新用法になると助動詞との共起が減るのか、というのは新たな疑問であり、今後の更なる研究が待たれる。

5 一般の場合への拡張可能性

本節では、今回のアプローチがより一般的な言語の分析に拡張可能かを考察する。大前提として、言語モデルの内部を直接観測することは困難であるの

で、結果の解釈には常に注意が必要である。また、今回は解釈可能なクラスが得られたが、他の現象に対しても同様のことが成り立つとは限らない。

このように注意すべき点はあるものの、本アプローチは様々な言語分析に応用できる可能性がある。本稿のように、仮説の検証と修正に用いることができる。また、何らかのアイデアを思いついた際に、用法変化を可視化し、研究をさらに進めるかどうかの意思決定にも利用できる。人手の分析は有効な手段ではあるが、時間とコストがかかる。例えば、今回の場合、第一著者は、2,519事例の分類に20時間ほど費やした。クラスタリングの場合は十数秒である。この簡便さと即時性は大きな利点である。

別の有効な応用例として、3節で導入したトッピング現象の発見を取り上げる。トッピング現象は、特異点である可能性が高いが、(ある期間内で)低頻度である上に、いつ始まるかは未知であるため、人手で発見することは困難である。一方、今回のアプローチでは視覚的に効率よく発見できる。少々作例的な例ではあるが、事例として、単語 *apple* を対象にしたところ 1930～1970年代にトッピング現象が発見された（図は付録Cに示した）。1970年代は Apple Computer Company が創業された時期であり、正に *apple* が新しい意味を獲得した瞬間である。

以上のように、言語処理的アプローチは、言語学における質的分析と組み合わせることで、実証性の高い方法論を提供することができる可能性がある。特に、表層情報のみに基づいた教師なし学習というアプローチは、実際の言語使用を基盤として言語体系を捉えようとする用法基盤モデル [6, 7] の考え方と親和性が高く、今後更なる発展が、言語学、言語処理の両面から期待される。

6 おわりに

本稿では、言語処理的なアプローチを用いて *better off* の新用法に関する仮説の検証を行った。その結果、従来の仮説を概ね支持するが、派生するいくつかの可能性あることを示した。また、(1) 仮想的な状況を想定する用法が増えて新用法が増えたのか、その逆であるのか、(2) そもそも新用法がいつどのように生じたのか、(3) なぜ新用法になると助動詞との共起が減るのか、という新たなリサーチクエッションも示した。更に、本アプローチの一般的な分析への拡張可能性も示した。

謝辞

本研究は JSPS 科研費 18K11456, 19K00657 の助成を受けたものである。

参考文献

- [1] 大谷直輝. 英語の better off 構文について. 日本言語学会第 157 回大会予稿集, pp. 246–251, 2018.
- [2] 大谷直輝. better off 構文の定着に関する認知言語学的考察. 言語処理学会第 28 回年次大会発表論文集, 2022.
- [3] Mark Davies. Expanding horizons in historical linguistics with the 400-million word corpus of historical American English. **Corpora**, Vol. 7, No. 2, pp. 121–157, 2012.
- [4] Reem Alatrash, Dominik Schlechtweg, Jonas Kuhn, and Sabine Schulte im Walde. CCOHA: Clean corpus of historical American English. In **Proc. of the 12th Language Resources and Evaluation Conference**, pp. 6958–6966, 2020.
- [5] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In **Proc. of 18th Annual ACM-SIAM Symposium on Discrete Algorithms**, pp. 1027–1035, 2007.
- [6] Ronald W. Langacker. **A Usage-based Model**. Stanford University Press, Stanford, 1988.
- [7] Michael Barlow and Suzanne Kemmer. **Usage-Based Models of Language**. CSLI Publications, Stanford, 2000.

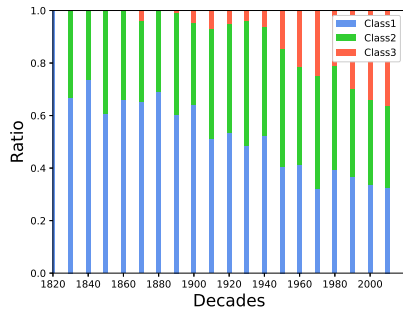


図4 人手の分類結果から得た用法変化（比率）。

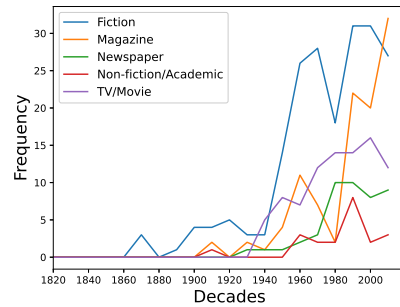


図5 年代／ジャンル別の新用法頻度。

A コーパスに対する前処理

今回の分析に合わせて、CCOHA に対して次のような前処理を行った。ノイズと思われる文書は除外した。具体的には、「@@年.txt」（例：@@1525.txt）のように年とファイル名と思われる文字列を含む文書は分析対象外とする（この処理により、全 140,743 文書から 140,665 文書へ減少）。また、文書中のタグ（<P></P>など）は除去した。

この結果から、better off を含む文を抽出した。文脈が短い事例（better off の左右の単語数が 5 未満）は対象外とした。更に、伏字が含まれている文も対象外とした⁴⁾。その結果、2,591 文を得た。人手の分類の際に 提示順の影響が出ないように、ランダムに並べ替えた。

B クラスタリング結果の詳細

図4に、人手での分類結果から得られた用法変化の図を示す。本文とは異なり、縦軸は各クラスの相対頻度である。

$k = 4$ とした場合のクラスの解釈根拠は次の通りである。表1に、各クラスにおける than, 助動詞⁵⁾、新用法が出現した割合を示す。表1によると、クラス1でのみ、than を伴う頻度が圧倒的に高い。これらの事例は、better off の典型例であるといえる。クラス2は、than の代わりに別の句（at home, without it, now など）が後続する事例が増える。このことは、経済的な意味からより一般的な意味へ拡張されたことを示唆する。ただし、クラス1でも、経済的な意味以外でも使用されている事例も多数存在するので、大谷の仮説の (Stage 1) と (Stage 2) に完全

4) COHA および CCOHA では、著作権の制限により、一定の割合で文章の一部が伏字になっている。

5) ここでは、構文解析は行わず、単純に better off の右側／左側に、than/助動詞が出現するかを数えた。これは better off 構文は特殊であり、構文解析がうまくいかないことが多いことを考慮したためである。なお、ここでは助動詞を can, could, may, might, shall, should, will, would とした。

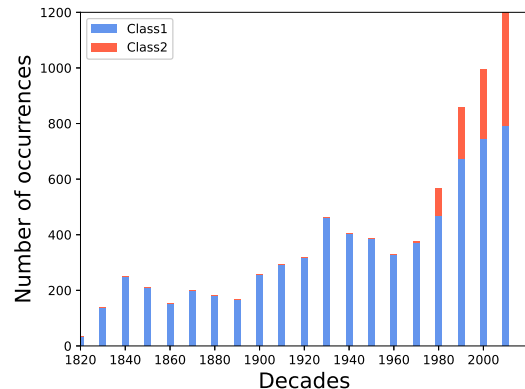


図6 単語 apple におけるトッピング現象。

に対応するわけではない。更に、表1の助動詞の列に目を向けると、クラス3での出現頻度が相対的に高いことが分かる。興味深い点として、クラス4では助動詞の頻度がクラス3に比べて低くなっている；両者の差は有意である（母比率の差の検定、 $p < 0.01$ ）。この点については、4節で考察している。

図5に、各年代／ジャンルにおける新用法の出現頻度を示す。新用法はフィクションから別のジャンルへ広がっていることが分かる。

C apple におけるトッピング現象

単語 apple に対してトッピング現象を観察してみる。今回のアプローチを用いて用法変化を図示した⁶⁾。ただし、クラス数は2とした。その結果は図6に示す通りである。図6より、1930～1990年代にトッピング現象がみられる。

表1 各クラスで than, 助動詞, 新用法が出現する割合。

クラスラベル	than	助動詞	新用法	全事例数
1	0.43	0.33	0.01	944
2	0.06	0.24	0.07	637
3	0.03	0.62	0.04	548
4	0	0.39	0.81	462

6) bert-base-uncased を用いているため、大文字／小文字の情報は利用していないことに注意されたい。