パターンの生産性から見る言語の定型性

吉川 正人 (慶應義塾大学大学院)

1 はじめに

Wray & Perkins (2000)は Wray (1999)の調査に基づき，言語の本質が「定型性」であると結論付け，定型性に基づく言語モデルを提案した．言語が定型的であるということは，1) 言語処理の容易さ，2) 伝達の確実性を高めることにつながり，その点で，特に話し言葉において有益である (Wray & Perkins 2000: 17-19)．このような言語観の下では，従来言語の本質として主張されてきた言語の「生産性」及び「創造性」はあくまで二次的なものに過ぎず，日常の会話において活躍することはほとんどないとみなされる(Wray & Perkins 2000: 13)．

本稿では，このような議論を受け，1) 話し言葉，2) 書き言葉(ノンフィクション)，3) 書き言葉(フィクション)の定型性を比較し，3 < 2 < 1 の順に定型性が高まっていくことをコーパスの定量的な調査によって示す．

2 言語の定型性

Wray (1999)および Wray & Perkins (2000)は，「定型表現(formulaic sequence)」を以下のように定義している:

(1) 連続/不連続の，語その他の意味要素の配列で，使用時に丸ごと記憶から取り出されるもの

この定義には，不規則なイディオム(e.g., *by and large*)から透明性の高い表現(e.g., *NP be-TENSE sorry to keep-TENSE you waiting*)まで含まれている(Wray & Perkins 2000: 1)．共通項は，「記憶から丸ごと取りだされる」ということであり，逆に言えば，オンラインで生成されるものは定型的とは言えないことになる．

また，このような定義により，以下のことが必然化される: オンラインで生成される度合によって，定型性には段階性が存在する．100%丸ごと記憶から取り出されるものもあれば，部分的にはオンラインで生成されるスロットを持つパターンのようなものあり，それぞれ定型性の度合いは異なっていると考える．

このような連続性/段階性を含んだ定型表現の定義は以前にも試みられた(e.g., Bolinger 1976; Howarth 1986)が，これまでの定義は全て「記述的(descriptive)」なものであり，「説明的(explanatory)」なものではなかった(Wray & Perkins 2000: 6)．その点で，Wray らの定義はヒトの言語産出プロセス(および記憶のメカニズム)を組み込んだ定義であり，定型性とは何か，それは何の尺度か，といった問題に答え得るものである．

2.1. 定型性の効用

言語が定型的であるということは，以下二つの利点を生む: 1) 表現が定型であることで処理不可量が低下し，オンラインの処理が容易に行える; 2) 表現が定型であることは，解釈の可能性を狭めるか，多くの場合一様なものとし，結果伝達の確実性を高めることにつながる(Wray & Perkins 2000: 13-17)．

しかしこの二つの利点は，実際は同一の要因を別の側面から見たものに過ぎず，両者はコインの裏表の関係にあるという(Wray & Perkins 2000: 17-18)．というのも，処理負荷の軽減というのは結局対話場面における円滑なやりとりの達成のために必要とされるものであり，また，伝達の確実性というのは，聞き手にとって重要であるばかりか，聞き手に確実に意図が伝わるという点で話し手にとっても非常に重要なものである．従って，定型表現の使用はオンラインの対話において，話し手にとっても聞き手にとっても非常に有用であるということになる(Wray & Perkins 2000: 18-19)．

このような言語観の下では，従来言語の本質とされてきた生産性/創造性は，「予想外」な緊急時に対処する能力の産物として捉え直される(Wray & Perkins 2000: 13)．また，ここで「ではなぜ非定型な表現を用いるのか」という疑問が生じるが，「非定型表現使用」の動機に関しては Wray & Perkins (2000)では直接的には議論されていない．ただし，議論の後半で定型表現の発達モデルが提案されており，その中で非定型表現が優勢になる発達上の時期(2歳~思春期)が存在することが指摘されているため，その議論が間接的にこの問いに答えていると考えることはできるかもしれない．いずれにせよ，この非定型表現使用の動機に関する問題は，非常に興味深い議論ではあるが，本稿の趣旨とはやや逸れてしまうため，ここでは深く立ち入らない．

2.2. 定型性と頻度

上に見たように，Wray らは定型性を「記憶から丸ごと取りだされる」という，言語処理及び言語記憶の要因か

ら定義している．一方で，それ自体が既定要因ではないとしながらも，定型性と「頻度」の間には相関があることを述べている(Wray & Perkins 2000: 6-7)．実際，定型表現はその使用機会が多く，必然的に使用頻度は高くなる(cf. Sinclair 1991; Langacker 1991)．

従って，ここから頻度に関する数的尺度で定型性を算定することができるという可能性が考えられる．勿論，Wray らが指摘するように，低頻度だが定型性を持つような表現は確かに存在している(e.g., *That's another one mess you've gotten me into*, Wray & Perkins 2000: 7).[1] しかしながら，多くの定型表現は(その使用の動機付けから考えて)高頻度であることが充分に予想される上，逆に非定型な創造的・分析的表現が(定型表現よりも)高頻度であるということはほとんど考えられない．

以上より，本稿では，頻度に基づく数的尺度によって定型性を算定し，定量調査によって言語の定型性の一面を示すという手法をとることとする．具体的には，広範囲に渡って利用されている確立した定型性の指標というものは存在していないため，逆に生産性の指標である「タイプ-トークン比」を算出し，その「低さ」を定型性の「高さ」であると看做すことで定型性の指標に代える．

3 仮説

上に述べたように，Wray & Perkins (2000)の議論では定型性の使用はオンラインの対話上の要請によるものであるところが非常に大きいとされている．ここから，逆に非定型な創造的・分析的表現はそのような要請の弱い状況下では相対的に頻度が向上することが予想される．

従って，表現の定型性は，対話場面での話し言葉が最も高く，逆に書き言葉で，さらに創造性の求められるフィクションではそこまで高くならないということが予想される．ここで，本稿では以下のような「仮説」を提案する:

(2) 表現の定型性は，オンラインでの伝達上の要請が低くなるにつれ低下する

この仮説が導く予測は，表現の定型性は[話し言葉 > 書き言葉(ノンフィクション) > 書き言葉(フィクション)]と順に低くなっていく，ということである．フィクションの定型性が低くなるという予想は，創造性が求められるということ以外にも，対話上の要請である「伝達の確実性の向上」という要因にあまりとらわれないためであるという理由にもよる．

以下では，このような仮説に基づき，前節で述べたタイプ-トークン比を利用した定型性の定量調査を報告する．

4 調査

定型性の調査で困難なのは，「何の定型性を計るか」という問題である．「定型表現」というものが，言語表現の規模に寄らず，また具体的な語句のみならずスロットを含むような抽象的パターン(e.g., *NP be-TENSE sorry to keep-TENSE you waiting*)にも及ぶことがそのような困難を生んでいる．

本稿では，定型性算定の対象を，これまで主に技術的問題から大規模な調査が行われていなかったと思われる，スロットを含む抽象的なパターン(の一部)に限定し，調査を行う．具体的には，KfNgram というフリーソフト(http://kwicfinder.com/kfNgram/kfNgramHelp.html) を用い，N-gram ベースのパターンのリストを作成し，その全パターンにおけるタイプ-トークン比を算定することで，パターンの定型性を求める．

4.1. データ

調査には，ウェブ上のコンコーダンサー等の使用により一部データの閲覧のみ可能なものではなく，全データがダウンロード等により利用可能であり，かつある程度の規模・信頼性を有するコーパスを用いる必要があったため，そのような要件を満たし，かつフリーで使用のできるコーパスとして，American National Corpus (ANC) のフリー公開分である Open ANC (OANC: http://americannationalcorpus.org/OANC/index.html) を用いた．

ANC は，現在構築途上のアメリカ英語の均衡コーパスであり，British National Corpus のアメリカ英語版として開発されている．最終的には BNC と同等の 1 億語規模になる予定だが，現時点ではセカンドリリースで 2200 万語の規模となっている．

このうち約 1500 万語分が OANC としてフリー公開されている．データはテキスト形式の文字(起こし)データ，XML 形式のメタデータ及びタグ情報の 3 種類から成っている．今回はテキストデータのみ使用し，タグ情報等は使用しなかった．

OANC には話しことばが 3,217,772 語含まれているが，そのうち大半は電話による会話(Switchboard Data)であり，対面会話は 198,295 語のみであった．本調査ではこの対面データのみを利用した．また書き言葉に関しては，

[1] Wray らは指摘していないが，諺や慣用句の類はまさに「低頻度で定型」な表現の代表ではないかと思われる．実際，諺などに対しては生起頻度に敏感な統計指標(e.g., t-score)ではその定着度が図りにくく，低頻度でも結びつきが強ければ高く出るスコア(e.g., MI-score)を用いることが有用であることがよく指摘される(e.g., 園田, 高見 2005: 132)．

計 11,406,155 語収録されている．このうち，今回使用したのはノンフィクション 330,524 語およびフィクション 61,746 語のデータのみである．

対面会話およびノンフィクションに関してはある程度の量が確保できたため，対面会話のデータ量に合わせ，ノンフィクションのデータはランダムに半分を抽出し，この 2 者に関してはおおよそのデータ量を揃えた．ただしフィクションに関しては少量のデータしか存在していないため，このような不均衡はあまり好ましいことではないが，やむを得ず全データをそのまま使用した．以下に全データのステータスを表(Table. 1)にして示す．Table. 1 および以降，話しことば(対面会話)を SP，書き言葉(ノンフィクション)を NF，書き言葉(フィクション)を FC と表記する．

Table. 1: 各データ情報

	SP	NF	FC
File size (KB)	983	1010	366
Token	200928	169476	62904
Type	8844	15813	12536
Type/Token	0.044	0.093	0.199

“Token”は総語数，”Type”は異なり語数となっている．集計はパターンを生成した際に KfNgram が算出したものを参考にしているため元データの情報とは若干異なっている．

4.2. 方法

パターン生成および頻度のカウントは，全て先述の KfNgram を利用した．KfNgram は以下のような処理を行う: 1) 入力したテキストデータに対し n-gram を作成し，2) その n-gram の構成要素のうち一つを変項としたパターンを(pattern)生成し，3) 各パターン毎にその事例を頻度付きで提示する．n-gram および n-gram ベースのパターンは一覧が全てテキストファイルとして保存されるため，データを直接利用することができる．n-gram 生成にあたり，n は任意の数値をユーザー側で指定することができる．今回は，有意義なパターンが生成されうる範囲として，n = 2 ~ 5 まで n-gram およびパターンを生成した．

n-gram の生成は，入力として与えられたテキストデータの改行を全て削除し 1 行データとしたのち，隣接語彙をグルーピングし最後に集計するという形で行われるようである．

その後，こうして生成された n-gram に基づき，共通語彙を持つものを束ね，非共通語彙を変項，共通語彙を定項とするパターンが生成される．つまり，“XY” および “XZ”という 2-gram が存在する場合 “X*” (“*”は変項)というパターンが生成される．

変項化される語彙は一つのみであり，従って 3-gram からは 2 語が定項で 1 語が変項の，4-gram からは 3 語が定項で 1 語が変項の，5-gram からは 4 語が定項で 1 語が変項のパターンがそれぞれ生成されることになる.[2] また，パターン生成は複数の n-gram(のタイプ)における共通語彙を発見する形で行われるため，各パターンにつき少なくとも二つの事例(のタイプ)が存在することになる.[3]

5 結果・考察

以上の方法を用いて，n-gram 及び pattern を生成した．以下に，n = 2 ~ 5 の全てのパターンに関して，その pattern 数(P)と変項(variant: V)数を提示する．尚，変項数とは変項の異なり，つまりタイプ数であって，トークン数ではない．従って，変項として同一の語彙を持つ事例が複数あっても，variant としては変項の語彙が単一である限り 1 としてしかカウントされていない. この点に関しては以下で詳述する．

Table. 2: n = 2~5 のパターン数及び変項数

N	SP		NF		FC	
	Ps	Vs	Ps	Vs	Ps	Vs
2	9054	129818	16089	175269	9298	74143
3	52718	263572	46109	187120	13781	49887
4	51453	158727	23609	66348	3894	9808
5	19069	46798	7246	17306	770	1685

5.1. タイプ-トークン比とパターン-変項比

今回調査対象がスロットを含む抽象的なパターンであったため，タイプ-トークン比の産出は厳密な意味でのタイプ数とトークン数の比の計算にはなっていない．というのも，パターンにはそれぞれパターンの事例としてのタイプ(= n-gram のタイプ)とその事例の実例としてのトークン(= n-gram のトークン)が存在する．例えば SP の 3-gram ベースのパターンで “at the * ”というものが 233 回観察されているが，この中には “at the age”, “at the

[2] 即ち，複数語彙が変項となるパターンは全てもれることになる．これは KfNgram の明らかな，そして重大な欠点であるが，このような組み合わせ論的なアルゴリズムで複数語彙が変項となるパターンを生成すると膨大な量が生成されてしまうため，大規模コーパスを用いた調査には使用が困難であるという欠点がある．

[3] 上の脚注 3 で述べた点に加え，これも実際は少し問題を孕む処理である．変項が一つのパターンというのも理論的には実在してはならない理由はなく，実際アナロジーを用いて新奇表現を生成するには変項を一つしかもたないようなパターンも利用されていると考えることは充分に可能である．この点と脚注 3 で指摘した点を克服する，本当の意味で網羅的なパターン生成を実装しているプログラムとしては Pattern Lattice Builder (黒田, 長谷部 2009)が存在するが，現時点では大規模データの処理には対応できない．また，非組み合わせ論的な統計ベースのパターン生成アルゴリズムとしては Adios (Solan et al. 2005)が存在するが，現時点では Adios Lite というデモ版の使用のみが可能で，パターンを 100 生成するとプログラムがストップしてしまうため，実用的ではない．

beach", "at the beginning" 等の具体的な事例が含まれ，その事例の異なり数は 96 である．これが " at the * "というパターンに対する変項の数ということになる．今回の計算には，全体のパターン数(e.g., SP の 3-gram ベースでは 52718)を全体の変項数(e.g., SP の 3-gram ベースでは 263572)で割ったもの，言ってみれば「パターン-変項比(pattern-variable ratio)」を利用した．次節でこれを提示する．

5.2. 結果

以下に n = 2~5 全てのパターン-変項比(以下 P-V 比)を提示する:

Table. 3: n = 2~5 の P-V 比

n	SP	NF	FC
2	0.069743795	0.091796039	0.125406309
3	0.200013659	0.246414066	0.276244312
4	0.324160351	0.355835896	0.397022838
5	0.407474678	0.418698717	0.456973294

各行において最大値を網かけした．すぐにわかるように，すべての行においてフィクションが最大値を示している．尚，前述の通り，変項数は各パターンに対して最低 2 つは存在するため，P-V 比が 1 になることはなく，最大値は 0.5 になる．直感に沿うよう，この点を反映し全値に 2 を掛けたものを以下に提示する．

Table. 4: P-V 比×2

n	SP	NF	FC
2	0.13948759	0.183592078	0.250812618
3	0.400027318	0.492828132	0.552488624
4	0.648320702	0.711671792	0.794045676
5	0.814949356	0.837397434	0.913946588

統計的にも，SP-NF 間の t 検定，NF-FC 間の t 検定は共に有意差を示した．p 値はそれぞれ $p = 0.03$, $p < 0.001$ である．

5.3. 考察

以上から，仮説の予測通り，話し言葉で最も定型性が高く，フィクションで最も定型性が低いという結果が示されたと言える．ここで，この結果が何を意味しているのか，若干の考察を加える．

考えねばならないのは，「P-V 比(の低さ)が一体何を表わしているのか」という点である．P-V 比は，先に述べたように，パターンの総数を変項の総数で割った数値である．この値が高くなればなるほど，パターン一つ一つの使用の幅が狭まり，言ってみればその「利用価値」が軽減される．逆に言えば，同一パターンが多数の変項を持つ場合，そのようなパターンは非常に有用であり，同じパターンを用いて多数の表現を算出することができるようになる．つまり，パターン一つ一つを見ればむしろ生産性が高いが，パターン全体をみると，あるきまりきったパターンしか使用せずに，多数の異なった表現を産出できるということになり，結果定型性が高いということになる．従って、P-V 比(の低さ)はパターンの定型性算定には適した指標であると言えよう。

6 結語

本稿では，定型性を言語の本質とする Wray らの議論(Wray 1999; Wray & Perkins 2000)に対し，定量的な検証を試みた．その際，「オンラインでの伝達上の要請が低くなれば定型性は減少する」という仮説を立て，そこから[話言葉 > 書き言葉(ノンフィクション) > 書き言葉(フィクション)]の順に定型性が低下するという予測をした．調査の結果，n-gram ベースのパターンに対してはこの予測は当てはまり，仮説を支持する結果となった．

課題としては，調査対象のパターンを，複数語彙がスロットになるようなものも含める形に拡大し，より網羅的な調査を行うということが挙げられる．脚注3であげたツール等を有効活用し、今後実施していく必要があろう。

参考文献

Bolinger, D. 1976. Meaning and memory. *Forum Linguisticum, 1*, 1-14.

Howarth, P. 1998. Phraseology and second language proficiency. *Applied Linguistics, 19*, 24-44.

黒田航, 長谷部陽一郎. 2009. Pattern Lattice を使った(ヒトの) 言語知識と処理のモデル化. 『言語処理学会第 15 回大会発表論文集』670-673.

Langacker, R. 1991. *Foundations of Cognitive Grammar, Vol. 2*. Stanford: Stanford University Press.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Solan, Z., Horn, D., Ruppin, E., & Edelman, S. 2005. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences, 102 (33)*: 11629–11634.

園田勝英, 高見敏子. 2005. コーパスに基づく語彙研究. 斉藤俊雄, 赤野一郎, 中村純作 (編)『英語コーパス言語学』(第二版)(pp. 121-143) 東京: 研究社.

Wray, A. 1999. Formulaic language in learners and native speakers. *Language Teaching, 32*, 213-231.

Wray, A. & Perkins, M. 2000. The functions of formulaic language: An integrated model. *Language & Communication, 20*, 1-28.

連絡先　吉川 正人　machayosihkawa@dream.com