HOKKAIDO UNIVERSITY

Title	GOOGLE 日本語Ｎグラムの分析(1)：「イッパヒトカラゲ」の足跡
Author(s)	小野, 芳彦
Citation	北海道大学文学研究科紀要, 145, 33(左)-46(左)
Issue Date	2015-03-20
DOI	https://doi.org/10.14943/bgsl.145.l33
Doc URL	https://hdl.handle.net/2115/58315
Type	departmental bulletin paper
File Information	145_03_ono.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

北大文学研究科紀要 145 (2015)

Google 日本語Nグラムの分析(1)
――「イッパヒトカラゲ」の足跡――

小 野 芳 彦

1．Google 日本語Nグラムとは

本稿でGoogle 日本語Nグラムと呼ぶのは，Google 社がクロールしたインターネット上の日本語 Web 文章の 2007 年 7 月の時点のスナップショットをもとに，Google 社研究員の工藤拓・賀沢秀人両氏が日本語の単語 n-gram とその出現頻度を抽出してまとめた大規模言語リソースである（工藤・賀沢 2007）。総単語数：約 2550 億語，総文数：約 200 億文の超巨大日本語リソースである（諸元：**表 1** ）。

表 1．日本語Nグラムの諸元

総単語数	255,198,240,937
総文数	20,036,793,177
異なり 1-gram 数	2,565,424
異なり 2-gram 数	80,513,289
異なり 3-gram 数	394,482,216
異なり 4-gram 数	707,787,333
異なり 5-gram 数	776,378,943
異なり 6-gram 数	688,782,933
異なり 7-gram 数	570,204,252

単語Nグラムとは，単語単位でN（ここでは 1〜7）語の連続を集計したも

10.14943/bgsl.145.l33

のをいう。クロールで集めた元データを適切に正規化した後，文に分割し，さらに単語に分割して集計している。文の分割は，単純に句読記号（〝.〞, 〝!〞, 〝?〞, 〝．〞, 〝。〞, 〝！〞, 〝？〞）によっている。単語の分割には，工藤氏が作成し，無料で公開している形態素解析エンジン MeCab-0.96 とその形態素辞書 mecab-ipadic-2.7.0-20070801 を用いていると記されている。分割結果を一切加工していないので，解析エラーがそのまま最終データに反映されているとも記されている。本稿は，この残された解析エラーと思しきデータから，Web 上の日本語文の特徴の一端を見出そうとするものである。

配布データに付属している文書によると，集計は 1-gram（ユニグラム：すなわち単純な異なり語の集計）では頻度 50 で，2〜7-gram では頻度 20 でカットオフ（集計に含めないで捨てること）が施されている。形態素解析エンジン MeCab が辞書にないと判断した文字列も臨時の語と認定して集計するので，頻度が 50 以上であれば 1-gram に収録される。このため，異なり 1-gram 数は 256 万を超えている（**表 1**）。『日本国語大辞典』（小学館）が 50 万語と言われているので，その 5 倍を超えるのであるが，実体をみると，年月日時刻をあらわす 14 ケタの数や顔文字などが含まれていて，語数の水増しになっている。

Web 上の文章には，書籍・新聞・雑誌ならば訂正されるであろう異表記が多く残される。その中には，書き手の表現意図がこめられた異表記も多くある。これらを見つけ出すことができれば，日本語の変化の兆しをとらえることができ，さらには，個々の異表記の誕生の契機に触れることもできよう。Google 日本語Ｎグラムには，それらが一部の文脈込みで集約されている。おそらく 200 万を超えるであろう臨時の単語の中に埋もれて。

2．きっかけ

近年，日本語の変化にまつわる言葉を題材とするエッセイが多く出版されている。その一つに，歴史家で有職点前家元の高山宗東氏の『お言葉でございます』がある。氏が歴史文献に見つけた意外な現代語について，その意味

の変遷などを辿ったエッセイであるとの紹介がある。その中の「ヒトカラゲ」と題する一節に以下のような会話スタイルを交えた解説がある。

「そのふたりが，お前と一緒にするなという意味で，何度も『イッパヒトカラゲにするな』というんです。でもアレって『ジッパヒトカラゲ』ですよね？　イッパ云々ともいうのですか？」／「ジッパヒトカラゲ」は「十把ひと絡げ」。一緒くたにする……という意味で，数種のモノを雑多にひとまとめにしてしまうこと。一把はおよそ一掴みの分量の意で，概してそれは単一のものである。したがって通念上最初から絡げられているであろう一把を，さらに一絡げにはできない。あるいは「一把ひと絡げ」を，「一把の量に絡げる」と解釈しても，それは「纏める」という意味になるから，「一緒くた」とは真逆である。／「── 故に『イッパヒトカラゲ』という言葉は成立しません……とはいうものの，わたしもどこかで『イッパヒトカラゲ』って聞いた気がするなあ。でもその時は強い違和感を持たなかったので，そのまま聞き流したような……。」(204–205 ページ)

ちょっと面白かったのは，今はまだ正しい言葉ではないゆえにキーボードで一括変換できない「イッパヒトカラゲ」には「一把」「一波」「一派」など，さまざまな当て字があるということ。例えば「一派」と表記すると，「何らかの集団に属しているからといって，一緒くたにするな」とも，意味が通ってしまう。／「無能なものが多い常務の派閥に属しているからといって，構成員を一派ひと絡げにするな！」── といった按配。(206 ページ)

筆者は「イッパヒトカラゲ」という言葉を聞いたことがなかった。Google で検索したところ，このような多くの誤用と，それを誤用として指摘しているブログ記事[1] がヒットした。そこで，Google 日本語Nグラムに「十把一絡

[1] ブログ記事　**山岸教授の日英語サロン**（2007 年 03 月 24 日）
高山（2010）の「ヒトカラゲ」と同趣旨のブログ記事である。「いっぱひとからげ, 一派一からげ, 一派一絡げ, 一把一絡げ, 一把ひとからげ」の用例を 2 例ずつ Google から引いたもの。

げ」の誤用がどのように収録されているのかを調べてみた。さらに，そこから「イッパ」以外の誤用を発掘し，その誤用の原因を推定した。本稿は，その報告である。

3．Nグラムに見える異表記

慣用句・成句は形態素（解析用）辞書および仮名漢字変換辞書に一語の資格でしかも正規形でのみ登録されているはずである。したがって，仮名漢字変換を正常に経由すれば非正規の異表記は発生しない。書き手の誤解や漢字変換の誤操作が非正規形の原因であると考えられる。ここでは，Nグラムに見える異表記[2] を手がかりに，「一派一絡げ」などの非正規形が発生する経緯を推定する。

慣用句全体の形態素構成と頻度を見る前に，異表記の広がりを把握する目的で主要部「絡げ(からげ)」と後半部「一絡げ(ひとからげ)」の形態素構成と分布を見てみよう（**表 2**）。

1〜2 行目に示すのはカタカナ表記である。カタカナはそれが連続する限り一語とみなしていて，ユニグラムに「ヒトカラゲ（89), カラゲ（354)」が存在する。カタカナの「カラゲ」が「ひとからげ」の後半を形成していると確実にいえるのは 354 のうち 28 の「一-カラゲ[3]」(バイグラムに存在)しかない。

ひらがな列の形態素解析は前後の言葉が影響する。「ひとからげに(する)」のような言い回しが多く使われることと，感動詞「げに」がいまだ使われていると想定されていることが影響してか，MeCab は「ひとからげ／ひとからげに」を 4 通りに，また，2 つから 4 つの単語に分割するようである。この

[2] 本節では「異表記」という言葉を「語形成の誤解析（日本語Nグラムの付属文書にいう「解析エラー」)」によって過剰に分割や結合をうけた臨時の語形をも含むものとして扱う。

[3] 形態素境界あるいは語境界をハイフン (-) で示す。本稿の表にある文字列が n 個のハイフンで繋がれている場合，右にある数は（n+1）グラムの集計をあらわす。

表 2．日本語Nグラムに見える「ひと(一)-からげ(絡げ)」の形態素構成と頻度

カラゲ	354	一-カラゲ	28
		ヒトカラゲ	89
から-げ(に)		ひと-から-げ	2371
		ひと-から-げ-に	295
		ひと-から-げに	3629
		一-から-げ	23
		一-から-げに	3441
からげ	12758	ひと-からげ	1439
		一-からげ	2778
		一-と-からげ	33
		ひ-とっ-からげ	23
絡げ	5122	ひと-絡げ	191
		一-絡げ	3486
		人-絡げ	21

ように，MeCab によるひらがな列の形態素解析はわれわれの直観的なものとはかなり異なっている。ちなみに，音韻的な強勢を含んだ「ひとっからげ」は

ひ　　動詞, 自立, ＊, ＊, 一段, 連用形, ひる, ヒ, ヒ
とっ　　動詞, 非自立, ＊, ＊, 五段・ラ行, 連用タ接続, とる, トッ, トッ
からげ　動詞, 自立, ＊, ＊, 一段, 連用形, からげる, カラゲ, カラゲ

と解析されていて（mecab-0.993 win 版，辞書は IPAdic)，動詞の連結と見られているようである。また，和語「ひと」の送り仮名のゆれである「一と」も，「**一-と-からげ**（33)」と形態素に分割されてしまう[4]。

一般に，仮名漢字変換システムや形態素解析エンジンは，このような非正規形に弱く，お手上げであるといえる。

[4] このような送り仮名の余剰や過少まで含めると，完成形では以下が観測される。
「**十-把-一-と-からげ**（30)」,「**十-把-一-絡**（1129)」

前半の「じっぱ」を含めた「じっぱひとからげ」類の形態素構成を見る前に，ユニグラムに存在する「ぱひとからげ(2908)」を別に見ることにしよう。「ぱ」で始まる和語・漢語はないから，これは，頻度の高い未知語として集計されてユニグラムに存在すると推測される。前には何が来ているのか，バイグラムで確認する。

っ-ぱひとからげ（836），いっ-ぱひとからげ（139），
はっ-ぱひとからげ（115），じゅっ-ぱひとからげ（1807）

促音単独形態素「っ」の前は，トライグラムで確認するとほとんど「じ」である（**「じ-っ-ぱひとからげ（825）」**）。

この「ぱひとからげ」を後半にもつバイグラム・トライグラムの集計を**表3**にまとめる。

表3．「じっぱひとからげ」類ひらがな列の形態素構成と頻度

	ヒトカラゲ	ひとからげ	ひとからげに	ひとからげで	後3者計
ぱひとからげ		2908	208	93	3209
っ-ぱひとからげ	219	836	52		888
じ-っ-ぱひとからげ		825	49		874
じゅっ-ぱひとからげ	133	1807	70	22	1899
いっ-ぱひとからげ	51	139	79	60	278
はっ-ぱひとからげ		115			115
2-gram 計		2897	201	82	3180

これらの異形態[5]は，先頭の音節末が促音化した音読みの数であるという共通性をもっていることに注意されたい。

[5] 一列目「ヒトカラゲ」には，完成形である「ジッパヒトカラゲ, ジュッパヒトカラゲ, イッパヒトカラゲ, ハッパヒトカラゲ」のユニグラム上の集計を記した。2〜4列は完成形および完成形に助詞の「に, で」が後接した形のバイグラム・トライグラム上の集計を記した(助詞が付いたものは単独のものとは独立に集計されていることに注意)。先頭行と最終行の数値の差は，バイグラム集計の過程でカットオフされた数をあらわす。

ちなみに，MeCab 0.993 Win 版では「はっぱひとからげ」は，「はっ：感動詞, はっ」+「ぱひとからげ：名詞, 一般, ＊」「ひゃっぱひとからげ」は，「ひ：動詞, 自立, 一段, 連用形, ひる」+「ゃっぱひとからげ：名詞, 一般, ＊」と解析される。まともに数であると認識できる気配もない。このような分解では，「ひゃっぱひとからげ」は頻度 50 以上でなければ，カットオフされてしまう。それで，ユニグラムに存在しないのであろう。

4．漢字かな交り表記と仮名漢字変換

漢字表記は種々みられる。音(読み)からこの単語を想起する際，前部の「じっぱ」が「十把」であるとの認識がそもそも低いことがこの混乱の原因であろう。

漢字「十」の読み「じっ」は，現代の主要な中型（すなわち学習用）漢和辞典では慣用読みとされていて例外扱いされるが，実は「じゅう」という呉音の古い発音「じふ」が「カ・サ・タ・ハ」行に連声する際に促音化した規則的なものである[6]。このことは，高校までの国語では教えていないようなので，「十把」を「じゅっぱ」と読むことが日常化している。それにもかかわらず，国語辞典でも，また，それを引き継いでいる仮名漢字変換辞書でも「じっぱ」から引かなければ該当する漢字に到達できないのが現状である。

さて，「じゅっぱ」から「十把」に到達するためには，「十」が「じゅっ」という読みを持つことを仮名漢字変換辞書が知っていなくてはいけないが，近年はちゃんとそのように設定されているようである。その点は問題ないが，「把」が助数詞で元の「は」の転呼音である「わ」という読みを持ち，促音に続く場合に異音「ぱ」となることを更に知っていなければならない。このような音便・異音を個別に変換アルゴリズムで処理するのは変換の負荷になるので，たいていの仮名漢字変換システムでは「わ」・「ぱ」という読みを該当する漢字に与えているようである。MS-IME2007 では「羽」「派」にのみ「ぱ」

[6] 沖森 2010 などを参照。

の読みを与えていて，「把」には「わ」と「は」の読みしか与えられていない。したがって，「じゅっぱ」は「十把」には変換できず，「十羽」や「十派」になるのである。

ちなみに，「じゅうわ」からなら「十把」に到達する。「ホウレンソウ十把」という表記であれば，通常「じゅうわ」と読むのだから当然である。それでも，「じゅうわ一絡げ」というような表記はNグラムに現れていないところを見ると，慣用句の面目は保たれているといえる。つまり，だれも「じゅうわひとからげ」と入力したりしないということである。

表 4．漢字かな交り「十把ひと絡げ」類の助数詞別集計

	-ぱ	-把	-派	-羽
ひと-から-げ		1495	26	246
ひと-から-げに	(3209)	1992	30	277
ひと-からげ		958	20	105
一-からげ	207	1554	275	354
一-から-げに	283	1885	263	308
ひと-絡げ		90		
一-絡げ		169	479	934

表 5．漢字かな交り「十把ひと絡げ」類の完成形[7]の集計

	○-ぱ	○-把	○-派	○-羽	10-把	10-羽	○-束	
ひと-から-げ		1368	24	202	43	24	30	
ひと-から-げに		1914	28	261			48	
ひと-からげ		936		90				
一-からげ	158	1490	268	325			29	十
一-から-げに	252	1839	260	302				
ひと-絡げ		86						
一-絡げ		17557	458	871			90	

（つづく）

表 5．（つづき）

	○-ぱ	○-把	○-派	○-羽	10-把	10-羽	○-束	
ひと-から-げ		61	106				21	一
ひと-から-げに		71	41					
ひと-からげ		49	86					
一-から-げに		43	77					
一-絡げ		106	102					
一-絡げ		24		20				百

この言葉を音で聞いて知っているだけで，そもそも「じっ」が「十」の読みだと見当がつけられない場合，入声音である「一」「六」「七」[8]「八」「百」と混同する可能性は高い。「じっ」と「いっ」は先頭の子音の有無しか違いがないため，最も混同されやすいと考えられ，事実そのとおりである。子音も母音も異なる「ひゃっ」「はっ」はそれよりはるかに少ない。そして，漢字に変換されないで「イッパヒトカラゲ」や「はっぱひとからげ」などのカタカナ表記・ひらがな表記で残されているものもある[9]。

---

[7] 白抜きの欄は，「十把一絡げ」「一派」が単独の形態素として集計されたことをあらわす。したがって，それらは表 4 の助数詞ごとの集計には含まれていないことに注意されたい。また，十把一絡げの頻度が 5 桁に達するのは，それが一括で変換できるからで，慣用句・成句の取り扱いが基本的に正規形の単語登録で扱われることの正当性・妥当性を示す数字だと言えよう。逆に，「十把一絡げ」がかくも多様な異表記をもつのは，「じっぱ」の低認知度がもたらす，特異現象である惧れも大きい。

[8] 原理的には，「七」も促音化してよさそうだが，どんな助数詞にでも「しち」のまま，あるいは，和語「なな」に続けるというように促音化しないので，Nグラムには現れない。「六」も現れないがその理由には見当がつかない。

[9] 「はっぱひとからげ」の使用例は以下のとおり。

2013/07/25 – 応援してたのに当選できなかった候補者の人はとても残念だけど，ボクたちも「政治家は信用できない」「あの政党はだめだ」って *はっぱひとからげ* に言うんじゃなくて，ひとりひとりをきちんと見て，応援できる人を見つけることがとても大切だと思っ……（zeronomics.seesaa.net/article/370208686.html）

高山 2010 にいうように「一把一絡げ」では意味が通らない[10] ので，「一派，一波，一羽」がその意味の解釈結果の表記となるのであろう。Nグラムに「一派」があって「十派」がないのは，「派」が助数詞として変換辞書に登録されていないからであり，「10 把」「10 羽」があるのは，「把」「羽」が助数詞として（ただし「把」は「わ」という読みで）登録されているからであることがわかる。

以上のような規則的な誤りのほかに，「とっぱひとからげ」という不規則なものも，現在の Google 検索では見つかる。漢字の「十」を訓読み「と（お）」からの類推で「とっぱひとからげ」と誤って読んでいたと告白するブログ記事であったが，告白で読み方を仮名で示してくれないかぎり，そのことは表記に現れず，Nグラムに残りようがない。

それに対して，同じように誤って読んでいたという告白で取り上げられていたものであるが，Nグラムに残っているものがある。「**十-把-ひと-か-けら**（40）」と「**十-把-ひと-唐-揚げ**（27）」である。前者は，ひらがな列の読みとりがワードチャンク単位で行われるため，なじみのない「からげ」が「かけら」に再構成された[11] 読み誤りが原因の誤りであり，後者は「からげ」と「からあげ」の音の類似が原因である。

カットオフの関係で完成形は観測されていないが，「**杷一-絡げ**（20）」「**杷-**

また，「ひゃっぱひとからげ」のひらがな表記の例はカットオフされていてNグラムには見えないが，Web には現時点で以下のような使用例が見られる。

OKWave 2012-05-18 20：23：26 回答 No.2

ですので「ひゃっぱひとからげ」として処理します。

つまり「聞こえていようがいまいが，危険人物」というレッテルを貼ります。

[10]「把」にも「絡げ」にも「一」が付いているからである。後の「ひと（一）」を除いた「○把からげ」という異表記も見られるが，「**十-把-からげ（30）**」に対して「**一-っ-ぱからげで（149）**」「**一-っ-ぱからげに（51）**」と，「一」の優勢が見える。このことは，「いっぱからげ」という，意味も慣用も引き継いだ新しい（省略変異型）慣用句の誕生を意味するのかもしれない。変化の経過の一端として記録しておきたい。

[11] 文士・事物探究家松永英明の絵文録ことのは「読めてしまう」文章ネタの起源と歴史（http://www.kotonoha8.com/2009/05/10yometeshimau.html）参照。

ひと-から-げ（47)」「**十-杷-ひと-から**（25)」がNグラム上に残されている。「把」を「杷」と誤る原因は字形の類似であることは間違いないが，仮名漢字変換システムの影響も考えられる。われわれは，全体の読み仮名文字列から希望の漢字表記が得られない場合，後編集で当該漢字のみの修正を行うことが多い。その際，「把」を手書き認識や単漢字変換で入力すると，多くの候補の中から類字に見誤ることは大いに考えられる。ただ，実際は漢字の一字の読みの重複の多さのために，単漢字変換ではなく熟語から目的の文字を探し出すことを好む傾向がある。しかし，その場合でも，「把」の代表的な熟語が「把握」で，読みが「は」であるのに対して，「杷」の代表的な熟語が「枇杷」で「わ」である。この言葉の漢字表記があやふやな人にとって，「把」の助数詞としての読み「わ」から「びわ」のほうを引き出す傾向が高いのではないかと推測する[12]。

「じっぱ」と音が近く意味も通るものとして「大雑把」の「大」を除いた「雑把(ざっぱ)」からきたものであろう「雑杷ひとからげ」も見られる（木扁であることに注意)。Nグラムに残されているものと同じかどうかわからないが，2014 年 11 月現在 Google 検索で見つかるのは，あるブログサイトのタイトルである。書き手が木工家であることから名付けられた意趣的な用字（つまり一種の掛け詞）であるのかもしれない。

「把」を「束」と誤るのは，助数詞としての「わ」と「たば」の同義性からであると考えられる。このこと自体は仮名漢字変換システムと無関係であるが，このような誤った表記をする人が「束」を「わ」とも読むと信じている可能性があること，そして，その読みを仮名漢字変換システムが実装していないことをその人は容認している可能性も同時にあることは指摘できるであろう。つまり，「じゅっぱひとからげ」が「十束一絡げ」に変換されなかったのは，「束」の読みに「ぱ」が登録されていなかったせいと納得しているので

[12] もっとも，逆の混同である「**枇杷-島（588)**」や「**枇杷-葉（68)**」という異表記もNグラムにある。江戸時代以前のように木扁と手扁の区別をあまりつけない風潮が残っているだけのことなのかもしれない。

はないかと推測されるのである。

以上のような漢字の表記誤りが書き手のあやふやな記憶に起因することは間違いないが，さらにいえば，この誤りは漢字変換が書き手のあやふやな記憶を補正しえなかった結果であることに注目したい。コンピュータやワープロの漢字変換能力は紙の国語辞典に到底及ばないと考えられていたが，最近は表記の正確性に対する信頼を勝ち得ているように思える。われわれは，携帯電話やスマートフォンの仮名漢字変換システムを国語辞書がわりに活用している。

それでは，「じっぱひとからげ」で起きている現象はどういうことであろうか。われわれは，仮名漢字変換辞書や変換システムに対して，漢字表記の正確性に対して上記のような信頼感をもっているのは確実であるが，それは現代語さらに正規の用字法に関してであり，「じっぱひとからげ」のような慣用句は古語に準ずるものとして，いまだ仮名漢字変換が捕捉していなくても容認するという意識をもっているのではないだろうか。そのため，実際は慣用句としては認識されない「じゅっぱひとからげ」という読みを入力していることに気づかず，一知半解のままあやふやな漢字表記を誤って固定化させてしまったものであろう。本稿で示したＮグラムのデータはこのような経過を強く示唆している。

5．今後の課題

最初に引用した高山 2007 の前書きに，金田一晴彦氏の言葉として以下の引用が記載されている。

> 「（前略）これこれ，しかじかで変った……という経緯を押さえておくことが，とても大切なこと。過程も含めて，全部日本語ですからね。」（前掲書 6 ページ）

これは，「言語学者（あるいは国語学者）たる者，言葉の変化を押しとどめようとしたり正そうとしたりするものではない」という趣旨，いわば学会の

コンセンサスを述べたあとの発言であり，

> そもそもの成り立ちも解ったうえで，変化の過程に通暁すること ── それが，言葉を司る者の仕事であるらしい。（前掲書 6 ページ）

と高山氏は結論付けている。本稿もこの考えに同意する。

Web 検索サービスを提供する Google などの企業は，営利企業であるとはいえ〝言葉を司る者〟として社会に大きな影響力を持つ責任ある存在である。検索ユーザーの使う言葉に対して「それは，○○ではありませんか」と訂正を控えめに提示するサービスはすでに始まっている。

Web 検索における形態素解析の導入は，検索のヒット率を挙げるためだけではなく，実は，Web 文書の整合性のとれた理解を目指しているのではないかと思われる。そのためには，言葉の使用法のゆれを検出することも重要であり，そのゆれに適宜対応できるように，形態素解析の辞書を整備するといったことも必要であると考える。本稿の場合なら，「ぱひとからげ」などの（臨時）形態素ではなく，「ひとからげ」という形態素を設けるなどがそれにあたる。このような候補をどのように抽出するかは今後の課題と考えている。

ただし，言葉のゆれ自身が仮名漢字変換システムやその辞書によって拡大されている本稿のケースでは，仮名漢字変換システム側の柔軟な対応を求めることの益が大きいであろう。つまり，「じゅっぱひとからげ」が「十把一絡げ」に変換されるようにしてあれば，あるいは，「把」のよみに「ぱ」があれば，「10 羽ひとからげ」や「十束一絡げ」のような表記はもっと少なかったであろう。これは，上にふれた「変化を押しとどめよう」とすることではない。この異表記が自立的なものでなく，システムのいわば不具合から派生したものであるからである。

このような，多少ゆるい仮名漢字変換が果たして有効か，この大規模Nグラムに基づいて検証（あるいは模擬検証）することができよう。そして，そのようなゆるめの候補にどんな類型がみられるか調べることは，言語情報学の課題足り得よう。これまた今後の課題としたい。

参考文献

沖森卓也 『はじめて読む日本語の歴史』，文昇堂，2010.
工藤拓，賀沢秀人 「Web 日本語Nグラム第 1 版」，言語資源協会（発行），2007.
高山宗東 『お言葉でございます』，東京：総和社，2010.

謝辞 日本語学と言語情報学にとって宝の山とも思える Web 言語資料を作成し公開していただいた工藤・賀沢両氏に感謝します。また，そのような研究者の自由な業務外活動を公認・奨励している Google 社に賛辞を表します。