# ネットニュース上の日本語使用に関する一考察
## - 口語的環境におけるカタカナ使用 -

小野真嗣[*1]・小川祐紀雄[*2]・大橋智志[*3]
Email: onomasa@mmm.muroran-it.ac.jp

*1: 室蘭工業大学 国際交流センター
*2: 室蘭工業大学 情報メディア教育センター
*3: 苫小牧工業高等専門学校 創造工学科 情報科学・工学系

◎Key Words カタカナ使用，ネットニュース，テキスト処理

## 1. はじめに

本研究は、新聞や論文等のいわゆる公的文書を通じての発信を目的とする日本語ではなく、カテゴリ別にグループ化された閉鎖的環境における比較的自由な口語体での言語発信行動の日本語に注目し、その特徴について分析を試みるべく、データの選定・収集を行い、その特殊な環境下の日本語を解析するものである。本発表は、その分析に向けた過程で取り組み、カタカナ使用に焦点を当てつつ、結果考察に必要となる中間資料としての語彙リストの作成を中心に報告する。

## 2. 本研究活動の意義

### 2.1 学生の日本語運用

昨今、コミュニケーション力不足が大学構内及び企業現場においても叫ばれる中、日本語の口語的な使用実態については未解明な部分が未だ残されている部分であり、その研究が一層求められている。一方で若者言葉と一般的に解釈される表現も SNS 等の様々なネットコミュニケーションツールを介して使われている現状があり、対面における言葉遣いとは異なる表現も容易に認識できるところも見受けられる。また、海外からの留学生も各在籍先の大学でそういった表現に触れる機会もあって、計量的な語彙頻度研究により、留学生にも教員が提示可能な資料を作成する意義や可能性はあると著者らは考えている。

### 2.2 学問間の相互補完的な活動

本研究における取組では、言語学の背景を持たない情報工学分野の学生らにも分析技術の提供という形で研究協力者として参画してもらう一方、その取組自体が彼らへの教育的側面も兼ねたものとしての性格も持つことになる。言語資料の入手に始まり、perl や python によるテキスト処理を通して、その使用実態についてカテゴリ別に統計的・計量的に分析し、その考察までを一連の活動としている。発表当日は、コンピュータ利用による言語解明に関する学際的教育実践研究として位置づけ、参画した協力学生の情報技術に関する応用的側面の理解についても触れながら報告したい。

### 2.3 新たな教育研究環境の整備

筆頭著者は、小野(2014)、三河・小野ら(2015, 2017)を通じ、文系教員としての立場で情報技術の応用的側面から学際的工学教育についての経験を有するが、異動に伴いその環境を改めて一から構築することとなり、本取組はその学際的研究環境の整備としての側面をもち、パイロットスタディとしての位置付けである。

## 3. 取組の内容

### 3.1 言語データの入手

前節までに述べた日本語分析の対象となるデータとして、研究開始当初は従業員数 1 万人規模の IT 企業が技術コミュニケーションにおいて利用しているインスタントメッセージソフトの会話データなど技術実践性のあるデータなどにも関心があった。しかしながら、入手の困難さや利用の制限等もあり、口語的表現という共通点を見出し、かつてのネットニュースの記事を再検討し利用することとした。ネットニュースは電子メールと並び、コンピュータネットワークの初期に作られた情報交換システムの一つであるが、話題によって異なるニュースグループが作られておりカテゴリ別コーパスとも見なすことができる。一例を以下に示す。

(1) a. japan.town.sapporo (札幌地方の都市に関する話題)
b. japan.tv.drama (テレビ番組の内でドラマ)
c. japan.sports.baseball (スポーツの内で野球)
d. japan.sci.space (科学の内で宇宙科学)
c. japan.soc.crime (社会の内で犯罪)
c. japan.videogames.nintendo (テレビゲームの内で任天堂に関するもの)

(1) a. を例にとると、点で区切られた階層により、そのカテゴリ内で収録された情報がわかりやすくなっている。これらそれぞれを一つの言語使用域と見なして、計量的な分析を試みるため、自前のサーバより自動的にデータをダウンロードし格納の上、コーパス化を行った。

### 3.2 言語分析の方法

分析においては、柔軟な形で分析を行うことができるよう、既存の分析ソフトウェアを用いず、フレキシブルなデータ処理を可能とするべく、プログラミング言語による強力なパターンマッチで分析を行い、カテゴリ毎に語の頻度を求める手法を取った。分析環境として、Unix 環境下での python プログラミングを施し、

各カテゴリのカタカナ語の生起分布及び頻度を求める手法をとることとした。

## 4. 取組の評価

### 4.1 語彙リストの作成

ローカルサーバにデータを格納した点を上述したが、python のプログラミングによりコーパス中に生起するカタカナ語を抽出し、カテゴリ毎・年代毎に生起分布を調査した。図１はあるカテゴリ内における語彙の頻度をまとめたリストの例である。

| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | text | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | total |
| 2 | スキー | 2070 | 3528 | 955 | 288 | 269 | 376 | 46 | 2 | 7534 |
| 3 | ニュースグループ | 263 | 585 | 477 | 651 | 534 | 624 | 598 | 26 | 3758 |
| 4 | フォロー | 66 | 240 | 212 | 265 | 173 | 194 | 184 | 8 | 1342 |
| 5 | シーズン | 338 | 592 | 192 | 65 | 80 | 41 | 0 | 0 | 1308 |
| 6 | ネットニュース | 45 | 173 | 144 | 200 | 164 | 207 | 230 | 10 | 1173 |
| 7 | コブ | 474 | 214 | 78 | 39 | 203 | 5 | 0 | 0 | 1013 |
| 8 | ターン | 263 | 494 | 116 | 9 | 58 | 17 | 0 | 0 | 957 |
| 9 | ツアー | 290 | 476 | 92 | 32 | 54 | 9 | 0 | 0 | 953 |
| 10 | ザウス | 318 | 411 | 105 | 35 | 56 | 8 | 0 | 0 | 933 |
| 11 | テスト | 72 | 185 | 109 | 153 | 123 | 144 | 138 | 6 | 930 |
| 12 | ネット | 165 | 551 | 166 | 25 | 7 | 9 | 0 | 0 | 923 |
| 13 | メール | 109 | 197 | 136 | 112 | 97 | 96 | 92 | 4 | 843 |
| 14 | ゲレンデ | 284 | 367 | 95 | 23 | 19 | 31 | 0 | 0 | 819 |
| 15 | オフミ | 0 | 233 | 347 | 53 | 15 | 1 | 0 | 0 | 649 |
| 16 | リフト | 206 | 325 | 73 | 18 | 7 | 15 | 0 | 0 | 644 |
| 17 | ピロ | 242 | 184 | 92 | 62 | 25 | 0 | 0 | 0 | 605 |

図１　語彙リストの例(一部)

### 4.2 予備的分析による結果と考察

人の趣味が多種多様であるのと同様、ネットニュースにおいても書き込みが多い分野があれば少ない分野もある。表１は各カテゴリにおいて生起した異なり語数(type)と述べ語数(token)を示しているが、スポーツを例にとると、巨人・阪神に関する野球内容の書き込みの方が、日本ハムの記述に比べて、非常に多い傾向が読み取れる。ただ、巨人と阪神の各カテゴリでは、異なり語数に違いはほぼないものの、述べ語数に差が表れていることがわかる。また、頻度の高い語の例も合わせて表１で例示するものの、ジャンルによって違いや傾向があることが読み取れる。

表１　カテゴリ別の使用語彙の傾向(一部)

| category | type | token | sample words |
|---|---|---|---|
| ゲーム (ﾌﾟﾚｲｽﾃｰｼｮﾝ) | 6902 | 80135 | ｹﾞｰﾑ、ｿﾌﾄ、ﾌﾟﾚｽﾃ、ﾄﾞﾗｸｴ、ｻﾀｰﾝ、ﾌﾟﾚｲ、ﾊｰﾄﾞ |
| ゲーム (任天堂) | 1577 | 26438 | ｹﾞｰﾑﾌﾞﾗﾝﾄﾞ、ﾍﾟｰｼﾞ、ｸｿｹﾞｰ、ｷﾞｬﾙｹﾞｰ |
| スポーツ (野球・巨人) | 4862 | 143341 | ﾁｰﾑ、ﾌｧﾝ、ﾎｰﾑﾗﾝ、ｼｰｽﾞﾝ、ｺｰﾁ、ｼﾘｰｽﾞ、ﾒｼﾞｬｰ、 |
| スポーツ (野球・阪神) | 4196 | 69256 | ﾀｲｶﾞｰｽ、ﾔｸﾙﾄ、ｼﾞｬｲｱﾝﾂ、ｼｰｽﾞﾝ、ﾄﾚｰﾄﾞ、ｺｰﾁ |
| スポーツ (野球・ハム) | 751 | 18740 | ﾍﾝ、ﾌﾟﾛ、ｼﾞｬｲｱﾝﾂ、ﾗｲｵﾝｽﾞ、ｶｰﾌﾟ、ｸﾞﾙｰﾌﾟ、ﾀｲｶﾞｰ |
| スポーツ (スキー) | 4966 | 70563 | ｽｷｰ、ｼｰｽﾞﾝ、ｺﾌﾞ、ﾀｰﾝ、ﾂｱｰ、ｹﾞﾚﾝﾃﾞ、ﾘﾌﾄ、ﾌﾞｰﾂ |
| 都市 (札幌) | 5408 | 47613 | ﾗｰﾒﾝ、ﾛｰｶﾙ、ﾆｭｰﾖｰｸ、ﾙｰﾄ、ﾎﾃﾙ、ｹｰｽｹ、 ｽｷｰ |
| 都市 (横浜) | 2195 | 22397 | ﾗｰﾒﾝ、ｸﾞﾙｰﾌﾟ、ｿﾌﾄ、ﾛｰｶﾙ、ﾊﾏ、ｷﾞﾘｼｬ、ｶﾚｰ、ﾌｶﾋﾚ |
| 都市 (大阪) | 2560 | 26450 | ﾗｰﾒﾝ、ﾆｭｰﾖｰｸ、ﾐﾅﾐ、ﾋﾞﾙ、ｷﾀﾞﾀﾛｰ、ｴｽｶﾚｰﾀｰ |
| テレビ (一般) | 5459 | 61352 | ﾁｬﾝﾈﾙ、ﾄﾞﾗﾏ、ｽﾎﾟｰﾂ、ﾄｲﾚ、ｻｯﾁｰ、ﾅｲﾄｽｸｰﾌﾟ |
| テレビ (CM) | 3853 | 36446 | ﾁｬﾝﾈﾙ、ﾆｭｰｽ、ｽﾎﾟｰﾂ、ﾅｲﾄｽｸｰﾌﾟ、ﾃﾚﾋﾞﾄﾞﾗﾏ、ｱｲﾄﾞﾙ |
| 旅行 (海外) | 1657 | 18744 | ﾎﾃﾙ、ｱﾒﾘｶ、ﾂﾎﾞ、ﾂｱｰ、ｵｰｽﾄﾗﾘｱ、ｱﾝｹｰﾄ、ﾁｹｯﾄ、ﾄﾞﾙ |
| 旅行 (国内) | 744 | 13851 | ｱﾝｹｰﾄ、ﾂﾎﾞ、ﾊﾞｽ、ﾌﾟﾚｲ、ｻｰﾋﾞｽ、ｽﾎﾟｯﾄ、ｶﾞｲﾄﾞ、 |
| 留学 | 358 | 12211 | ｱﾒﾘｶ、ﾛｼｱ、ｷﾞﾘｼｬ、ｶﾅﾀﾞ、ｹﾍﾞｯｸ、ﾜｰｷﾝｸﾞﾎﾘﾃﾞｰ |

### 4.3 教育的効果

研究協力者の参加学生には、分析に携わった後に、記述式回答による設問を通じた事後アンケートを実施している。その分析については、発表当日に報告する。

## 5. おわりに

実際のコミュニケーション運用を観察する目的から、カテゴリ別に整理されているネットニュースに注目し、その資料収集を経て語彙リストの作成までをコンピュータ上で行うことに成功した。現在はPerl からPython に主流が移る過渡期でもあるようだが、プログラミングを通じた言語研究はこの先もまだまだ発展する余地があり、分析技術の向上が期待される。数年後には情報教育においてプログラミング指導も必須化されるが、その目的の一つとして言語分析の要素も入ると、ますますコンピュータを利用した言語研究の裾野が広がるのではなかろうか。今後も言語研究活動を情報教育的アクティビティの一つとして再考してみたいと考えている。

**謝辞**

本研究は、平成 28 年度室蘭工業大学研究推進経費(B)の助成金により実施されたものです。また本研究に際して、快くご協力を引き受けて頂いた室蘭工業大学情報電子工学系学科に所属する開米拓実君、並びに中田涼介君に対し心より感謝の気持ちと御礼を申し上げたく、ここに謝意を示します。

**参考文献**

(1) 小野真嗣：“工業高専における文系教員による本科卒業研究指導の実践　－developer/user 二つの側面を持つ学生の語学意識から－”，高専教育，37 号，pp.501-506（2014）.
(2) 三河佳紀，小野真嗣，渡辺暁央，小藪栄太郎，三上拓哉：“安全運営のための鉄軌道事故記録DB の構築と技術教育への活用　－高専における学際的工学教育構築への取組み－”，コンピュータ＆エデュケーション，38 号，pp.98-103（2015）.
(3) 三河佳紀，小野真嗣，渡辺暁央，小藪栄太郎：“鉄道技術の関する高専教育の再考　－クラブ活動における課外実習による取組－”，苫小牧工業高等専門学校紀要，52 号，pp.9-15（2017）.
(4) 田中二郎：“オープン・ニュース”，http://open-news.com/，（2016）.
(5) 渡邊克宏：“Unified fj NetNews archive”，http://katsu.watanabe.name/unifiedfj/，（2011）.