日本語文芸語彙の抽出と検証
—コーパスに基づくアプローチ—

Extraction and Validation of
Japanese Literary Vocabulary:
A Corpus-based Approach

松下 達彦

東京大学教養学部

matsushita@global.c.u-tokyo.ac.jp

本発表の概要

- 日本語の文芸語彙を抽出
- 妥当性を検討

1. 「文芸語彙」とは
2. 研究目的
3. 研究方法
4. 結果および考察
5. 今後の課題、まとめ

1. 文芸語彙とは

- 一般的テキストでの使用率に比べて、
文芸テキストでより高い使用率を占める語彙
 - * 創作に限定し、文芸評論等は含まない
 - * 本研究では、文学のジャンルは問わない
(本来は文学ジャンル(歴史、推理、児童、純文学等)による違いはあると思われるので、ジャンルに分けるほうがよいが、内容的にも労力的にも分けるのが難しい)
- 日本研究の学生、研究者や、日本語で文学を読みたい人にとっては重要な語彙であるはず

先行研究

- 経済学、化学などの専門語彙（岡1992、小宮1995・2005など）
- 専門語彙と基本語彙の中間に位置する領域特徴語
 - Academic Word List (Coxhead, 2000)
 - 日本語学術共通語彙（松下2011a）

など

学術テキストでは、学術ジャンルを問わず、一般テキストより高いテキストカバー率を示す

- 文芸語彙リストは見当たらない

2. 研究目的

学習負担軽減の重要性

- 第二言語・外国語語彙学習の負担は非常に大きい
- 中級以降, 頻出語句が減少
 - 最頻出1000語では60~70%に達するテキストカバー率, それ以降は1000語で数パーセント以下 「中級の壁」 (国立国語研究所1962など)
- 目的に即した語彙学習が重要
それなしでは学習効率が低下、挫折の原因にも

テキスト理解における語彙の重要性

- 効率性の評価: (単位語数あたりの) **テキストカバー率**
= (単位語数あたりの) 延べ語数の使用率の合計
= **そのリストの語が, 対象テキストで,**
他の語彙より高い割合で出現するか
- テキスト理解では, **カバー率を上げて未知語の密度を減らすことが最も重要な要素の一つ**

(Hu & Nation, 2000など)

(高田2006 にまとめ)

本研究の目的

- 文芸テキストで一般的テキストより高いカバー率を示す語彙リストを作成
- カバー率の検証 → 妥当性, 有用性を検証
 - 留学生などの語彙学習負担の軽減
 - より有効な語彙学習カリキュラムの開発

3. 研究方法

使用したテキストと計数単位

- 『現代日本語書き言葉均衡コーパス』(BCCWJ) モニター公開データ(2009年版)(国立国語研究所2009)
 - 書籍部分(約2800万語) + Yahoo知恵袋(約500万語)
 - 合計約3300万語
- 計数単位(語の区切り): UniDicの短単位
(ほぼ形態素レベル)
 - 形態素解析器: MeCab(工藤2006)
 - 解析用辞書: UniDic(伝ほか2009)

(UniDicの出力をAntWordProfiler で使用するため、
テキストエディタ上でマクロを作成して加工)

特徴語の抽出(1)

- AntConc (Anthony, 2007) のkeyness 機能を利用
- 特徴語の抽出には
 - 対象コーパス
 - 参照コーパスの二つが必要(対象語彙が含まれる割合を比較)
- 対数尤度比 (log-likelihood ratio) (Dunning, 1993)
 - (正規分布などの)特定の分布を要求しない
 - テキストの大きさが異なる場合にも比較可能な値を返す (Leech, Rayson, & Wilson, 2001)
 - 適度な割合で特徴語を抽出 (Chujo & Utiyama, 2006)

特徴語の抽出(2)

- 対象コーパスは分類せず、参照コーパスを四分類
- 4回特徴語抽出(Keyness:平均以上): 4種類共通の語彙
 - 旧日本語能力試験4級の高頻度語彙 → 除外
 - 上位20000語を超える低頻度語彙 → 除外
(松下(2011b)の留学生用語彙頻度リスト参照)
- **対象コーパス**
 - 文芸テキスト約830万語(創作テキストのみ)
 - 原則:NDC900番台のうち、1の位が0および4~9のもの
 - ただし、不審な場合はテキストを見て一つ一つ判定
- **参照コーパス**

使用テキストから、文芸テキスト以外を四分類

- 全領域の学術的専門テキスト約290万語
- 人文系一般テキスト約800万語
- 社会系/理科系一般テキスト約840万語
- 「Yahoo知恵袋」約520万語

下位コーパス分野別の異なり語数と延べ語数 BCCWJ 2009 版（書籍＋「Yahoo知恵袋」）

学術領域	学術分野	一般テキスト		専門テキスト		合計		備考
		異なり語数	延べ語数	異なり語数	延べ語数	異なり語数	延べ語数	
人文	言語・言語学	21252	403305	7831	102504	23708	505809	
	歴史	49700	2096004	11835	138139	51514	2234143	
	民族学・文化人類学	39759	1083009	3040	19666	40150	1102675	
	哲学、宗教	36253	1503013	9269	125917	38229	1628930	
	芸術	35501	967809	5042	39744	36177	1007553	
	文学	68439	8251999	5592	36852	68915	8288851	
	人文・芸術(その他)	46304	1973098	683	3414	46337	1976512	
文系	小計		16278237		466236		16744473	
	政治学	26299	920841	8814	115166	27900	1036007	
	法学	16502	511059	10074	333946	19542	845005	
	経済学	20015	684404	12534	367555	23525	1051959	
	経営学、商学	22087	846432	10788	310716	24489	1157148	
	社会学	30362	1318930	12960	333772	33008	1652702	福祉、労働、ジェンダーも含める。
	教育学	20157	621050	10417	262063	22675	883113	教科教育法もすべてここに含める。
社会	社会科学(その他)	18993	424164	4114	36168	19652	460332	運輸、通信、心理学、時事問題も含める。
	小計		5326880		1759386		7086266	
	数学	3497	40397	1959	19472	4352	59869	
	物理学	2368	25239	1280	9430	2920	34669	
	天文学・地球惑星科学	8181	101565	2583	21765	9035	123330	
	化学、金属・鉱山	4682	37469	2553	23275	6017	60744	
	工学(建築・土木)	16242	307617	7662	114099	18443	421716	
理工	工学(機械、電気、海洋・船舶)	12993	195762	5495	72049	14820	267811	
	理工系自然科学(その他)	18530	399470	8426	145175	21018	544645	情報工学、製造工業、生活科学(一部)、情報科学、図書館学など
	小計		1107519		405265		1512784	
	生物学	14680	262283	4064	41071	15672	303354	
	農学	14932	238989	3376	28584	15860	267573	林学、水産学、畜産学・獣医学を含む。
	薬学	3610	24703	1103	10197	4017	34900	
	医学	16657	485896	5955	82800	17961	568696	
生物・医学	歯学	1740	11551	874	3814	2174	15365	
	看護学	2348	19255	2491	23505	3744	42760	
	生物系自然科学(その他)	28254	943822	6749	74567	29490	1018389	スポーツ、衛生学、生活科学(一部)、環境 など
	小計		1986499		264538		2251037	
	インターネットQ&Aフォーラム(「Yahoo知恵袋」)	54215	5224852			54215	5224852	
	合計		29923987	46996	2895425	144231	32819412	

特徴語の抽出(3)

- AntConc (Anthony, 2007) のkeyness 機能を利用
 - 対数尤度比が4回とも**平均値以上**
- 「日本語を読むための語彙データベース」
(松下2011b, ダウンロード可)の留学生用語彙ランクで**初級から超上級に分類**
(ランク付けの方法については松下(2010)参照)

カバー率の検証方法

- AntWordProfiler (Anthony, 2009)を利用

- カバー率の検証

文芸テキスト > 参照テキスト

となるかどうか

- 抽出時に使用したコーパスとテストコーパス

- **テストコーパス**

= **文芸語彙の抽出の際に使用していないコーパス**

テストコーパス

- (MC) **会話**: 名大会話コーパス(日本語母語話者同士の雑談), 約113万語
- (BS) **一般書**: 『現代日本語書き言葉均衡コーパス』2009年モニター版(国立国語研究所2009)「ベストセラー」部分(文芸テキスト数が53%, 専門テキストなし), 約230万語
- (PC) **一般書**: 「日英対訳文対応付けデータ」(内山・高橋2003)の日本語部分(文芸書, エッセイなど), 210万語
- (JN) **新聞**: 日英新聞記事対応付けデータ(JENAAD) (Utiyama & Isahara, 2003)の日本語部分(1989-2001の「読売新聞」記事)約568万語
- (IS) **人文・社会系専門テキスト**: 新屋・松下編(未公刊)『日本語上級読解演習 国際学アラカルト』本文部分, 約4万語
- (TB) **社会系専門テキスト**: 「中・上級社会科学系読解教材テキストバンク」(東京外国語大学留学生日本語教育センター1998)本文部分, 約19万語
- (SS) **社会系専門テキスト**: 『留学生のための専門講義の日本語』(名古屋大学 国際化拠点整備事業2010)全9冊中, 社会系の3冊分の講義テキスト部分, 約5万語
- (TN) **理工系専門テキスト**: 『留学生のための専門講義の日本語』(同上)全9冊中, 理工系の5冊分の講義テキスト部分, 約7万語
- (BM) **生物・医学系専門テキスト**: 『留学生のための専門講義の日本語』(同上)全9冊中, 生物・医学系の1冊分の講義テキスト部分, 約1万語

4. 結果および考察

抽出結果

- 文芸語彙リスト (JLW = Japanese Literary Words)
5レベル, 計1616語 (表1)
 - 中～上級に多く分布
 - 名詞(39.9%)、動詞(34.0%)
 - 学術共通語彙と比べ動詞、副詞、間投詞が多い
 - 7割以上は和語
 - 漢語が4分の3を占める学術共通語彙と対照的

日本語文芸語彙のレベル別語数および語例

レベル	日能試	留学生のための語彙 ランク(松下2011b)	語数	各レベル 最下位の2語
初級	旧3級	682-1,291	142	ちっとも 引き出し
中級		1,292-5,000	446	戸惑う 吐き出す
上級1	旧2級	5,001-10,000	483	不吉 銀色
上級2	旧1級 旧級外	10,001-15,000	345	敵機 口笛
超級		15,000-20,000	200	香菜 樹海
Total			1,616	

日本語文芸語彙の特徴

- 1) **身体部位**を表す語(「首」「指先」「まぶた」)
- 2) **身体動作**を表す語(「立ち上がる」「飛び出す」「引っ張る」)
- 3) **陳述や様態の副詞**(「ちっとも」「たちまち」「きらきら」)
- 4) **間投詞**(「おや」「へー」「ほら」)
- 5) **縮約や地域語**などの**口語的表現**(「こりゃ」「-ちまう」「-はる」「-さかい」)
- 6) **比喩に用いられる語**(「振り向く」「流れる」「かみしめる」)
- 7) 作品の背景(**舞台装置・小道具**)に用いられやすい物体を表す語(「ビール」「岩」「ソファ」)
- 総じて言えば「**人間の感情や動作を生き生きと描写する語彙**」
- 性的描写や殺人、戦争を連想させる語も目立つ。コーパスは厳密なサンプリングによるものなので、冒険、推理、サスペンス、歴史などが多く題材になっているためだと考えられる。

日本語学術共通語彙の妥当性(1)

- **テキストカバー率**

=対象語彙によってカバーされるテキストの延べ語数の割合

- **テキストカバー効率** (本研究のために考案)

=あるテキストにおける異なり語1語あたりの平均カバー率

=テキストカバー率を, そのカテゴリーの異なり語数で割り,
1000000をかけたもの

=そのカテゴリーの語を1語学習することで

そのジャンルでどの程度効率よくカバー率が上がるか

- この数値が大きいほど, そのカテゴリーの語彙の学習で
効率よくそのテキストを理解できるようになることを予測

カバー率の検証(1)(テストコーパス)

- テストコーパス: 特徴語抽出時に使用していないコーパス
- 中級の文芸語彙446語
文芸テキストにおいて約3%のカバー率
＝学術テキストにおけるカバー率の約4.5倍
- しかし、中級の学術共通語彙559語が学術テキストで示す10-15%のカバー率と比べると低い
- 文芸語彙にはある程度の領域特徴性(専門性)があるが、そのレベルは学術共通語彙ほどではない

カバー率の検証(2)(テストコーパス)

- 1万語レベルまでの文芸語彙

会話でも文芸テキストと同等の高いカバー率

→ 日常会話の語彙との共通性が高い

・・・参照コーパスに会話テキストがない

会話コーパスを参照コーパスに入れれば・・・

- ・会話＋文芸

- ・文芸のみ の2種類の語彙に分けることが可能

=今後の課題

- 10000語を超える低頻度レベル

日常会話にはない、文芸書に特徴的な語彙が多い

カバー率の検証(3)(テストコーパス)

- 文芸テキストと学術テキストの**カバー率**の比較

文芸テキスト > 学術テキスト

初中級文芸語彙 3～5倍

6000～10000語レベル 8倍

11000語～20000語レベル 15倍～25倍

カバー率は高くないが、他の語彙よりは遥かに効率的

ジャンル別、レベル別の日本語文芸語彙のテキストカバー効率

Corpus Code				MC	BSB	UPC	BCCWJ	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	BCCWJ-T	JS-Bn	JS-Tn
ジャンル				会話	小説・エッセイ等	エッセイ・小説等	全体	新聞	学術	学術(入門)	学術	学術(入門)	学術(入門)	学術	学術(研究)	学術(研究)
領域					人文	人文		社会	社会	社会	社会	生物・医学	理工		生物・医学	理工
コーパス・サイズ(延べ語数)				1,129,538	2,298,828	2,102,178	32,819,424	5,675,357	186,768	50,601	42,152	13,904	74,645	2,895,425	719,802	2,705,026
レベル	日能試	留学生のための語彙リンク (松下2011b)	語数	テキストカバー効率 = あるテキストにおける異なり語1語あたりの平均カバー率												
			142	474	201	248	149	55	74	80	88	68	93	72	44	46
中級		1,292-5,000	446	72	62	63	42	17	12	9	15	7	18	14	9	9
上級 1	旧 2 級	5,001-10,000	483	15	13	15	9	3	1	1	2	3	3	2	1	2
上級 2	旧 1 級	10,001-15,000	345	2	6	7	4	1	0	0	0	0	0	0	1	0
超級	旧 級外	15,000-20,000	200	1	4	4	3	0	0	0	1	0	0	0	1	0
Total			1,616	66.8	41	46	28	11	10	10	12	9	14	11	7	7

ジャンル別、レベル別の日本語文芸語彙のテキストカバー率

Corpus Code		MC	BSB	UPC	BCCWJ	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	BCCWJ-T	JS-Bn	JS-Tn		
ジャンル		会話	小説・エッセイ等	エッセイ・小説等	全体	新聞	学術	学術(入門)	学術	学術(入門)	学術(入門)	学術	学術(研究)	学術(研究)		
領域		人文		人文	社会		社会	社会	社会	生物・医学	理工	生物・医学		理工		
コーパス・サイズ(延べ語数)		1,129,538	2,298,828	2,102,178	32,819,424	5,675,357	186,768	50,601	42,152	13,904	74,645	2,895,425	719,802	2,705,026		
レベル	日能試	留学生のための語彙ランク(松下2011b)	語数	テキストカバー率												
			142	6.73	2.86	3.52	2.12	0.78	1.05	1.13	1.25	0.97	1.33	1.02	0.63	0.66
中級		1,292-5,000	446	3.23	2.76	2.83	1.87	0.74	0.52	0.39	0.66	0.29	0.81	0.62	0.42	0.40
上級 1	旧 2 級	5,001-10,000	483	0.75	0.64	0.73	0.41	0.15	0.06	0.03	0.08	0.16	0.12	0.08	0.05	0.07
上級 2	旧 1 級	10,001-15,000	345	0.07	0.21	0.23	0.14	0.02	0.01	0.00	0.01	0.01	0.02	0.02	0.02	0.01
超級	旧 級外	15,000-20,000	200	0.01	0.08	0.09	0.05	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.02	0.00
Total			1,616	10.79	6.55	7.40	4.59	1.70	1.64	1.55	2.01	1.43	2.27	1.75	1.14	1.14

日本語学術共通語彙の妥当性(2)

- JLW は文芸語彙の効率的な学習に適している
おそらくは抽出方法も妥当、ただし検討の余地あり
 - JLWのカバー率は上級や超上級では極めて小さいが
 - 他の語彙より、効率よく
文芸テキストのカバー率を上げられる
 - 上級以降ではカバー率を上げるために数千語も必要
- たとえ0.1%でも効率よく学べることは重要

5. 今後の課題、まとめ

今後の課題(1)

- 一部、不適切に見える語あり

例) 「安」

解析ミスや採録基準のレベル設定の問題か

(Leechほか (2001)は対数尤度比3.8を5%有意の基準に)

用例を確認して、何らかの基準で除去することを検討

- 会話語彙との区別
- ジャンル別文芸語彙の探究

今後の課題(2)

- 複合辞・複合語・連語などは抽出できていない
短単位の語より頻度の高いものも多いはず
- 学術的な語が日常語の**比喩表現**などと同形の場合、**抽出できていない可能性あり**
(コーパスは多義表現に弱い)
例:「注ぐ」
「液体」-「力」「情熱」「心血」「精力」「愛情」
「(信濃)川が(日本)海に」 「視線」「火に油」「酒」

まとめ

- 日本語文芸語彙は文芸作品を読むために効率の良い語彙リストである
- が、そのレベルは日本語学術共通語彙(松下2011a)ほどではない
- 文芸語彙は10000語レベルまでは会話語彙との共通性が高い
- 会話語彙との区別や、ジャンル別の文芸語彙の抽出が今後の課題

ダウンロード

- 語彙リストおよび関連資料は、以下のサイトから

URL: www.wa.commufa.jp/~tatsum/

- このリストを活用した実践をなさった方は、ぜひ教えてください。
よろしく願い申し上げます。

引用文献(1)

- 内山将夫・高橋真弓 (2003). 日英対訳文対応付けデータ.
<http://www2.nict.go.jp/x/x161/members/mutiyama/align/index.html>
- 小木曾智信 (2007). 「茶まめ」(形態素解析ユーザーインターフェース)
https://www.tokuteicorpus.jp/dist/modules/system/modules/menu/main.php?page_id=1&op=change_page
- 工藤 拓 (2006). MeCab Ver. 0.98pre3 (形態素解析器)
<http://mecab.sourceforge.net/>
- 岡 益巳 (1992) 「非漢字圏の留学生のための日本経済基本用語表」『岡山大学経済学会雑誌』23(4), 191-229
- 国立国語研究所 (2009) 『現代日本語書き言葉均衡コーパス』2009年モニター版 (非公刊だが、当時は申し込みにより利用可能)
- 小宮千鶴子 (1995) 「専門日本語教育の専門語 — 経済の基本的な専門語の特定を目指して —」『日本語教育』86, 81-92

引用文献(2)

- 小宮千鶴子(2005)「理工系留学生のための化学の専門語-高校教科書の索引調査に基づく選定」『専門日本語教育研究』7, 29-34
- 伝 康晴・山田 篤・小椋秀樹・小磯花絵・小木曾智信 (2009). UniDic version 1.3.11 (解析辞書)
<http://www.tokuteicorpus.jp/dist/> (Ver. 1.3.0.は 2007)
- 松下達彦 (2010). 「日本語を読むために必要な語彙とは？ -書籍とインターネットの大規模コーパスに基づく語彙リストの作成-」『2010年度 日本語教育学会春季大会 予稿集』日本語教育学会, 335-336.
- 松下達彦 (2011a)「日本語の学術共通語彙(アカデミック・ワード)の抽出と妥当性の検証」『2011年度 日本語教育学会春季大会 予稿集』244-249
- 松下達彦 (2011b)「日本語を読むための語彙データベース」
<http://www.geocities.jp/tatsum2003/>

引用文献(3)

- Anthony, L. (2007). AntConc Version 3.2.1 (text analysis tool) <http://www.antlab.sci.waseda.ac.jp/software.html>
(Version 1.0 first published in 2002)
- Anthony, L. (2009). AntWordProfiler Version 1.2 w (word profiler)
<http://www.antlab.sci.waseda.ac.jp/software.html>
(Version 1.0 first published in 2008)
- Chujo, K., & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System*, 34, 255-269.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213-238.

引用文献(4)

- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19, 61–74.
- Hu, M. H. & Nation, P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403-430.
- Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. Harlow: Longman.
- Utiyama, M. and Isahara, H. (2003) Reliable Measures for Aligning Japanese-English News Articles and Sentences. *ACL-2003*, 72-79.