社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
TL2010–2 (2010–05)

Webテキストにおける未知語の頻度調査

服部　　峻†　　亀田　弘之†

† 東京工科大学 コンピュータサイエンス学部
〒 192–0982 東京都八王子市片倉町 1404–1
E-mail: †{hattori,kameda}@cs.teu.ac.jp

あらまし　日々増大して行く Web という情報源から様々な知識を抽出する Web マイニングの研究が盛んに行われているが，Web テキストを形態素解析や意味解析など自然言語処理する際，システムが用いる辞書に品詞や読み，意味などが未登録である「未知語」の存在が問題になる．本稿では，Web テキストに存在する多種なメディア，多様な話題，及び，投稿日時の 3 軸に依って，どのように未知語が分布しているか頻度調査を行った結果，Web テキストを自然言語処理するシステムにおいて，どんな分野で特に未知語処理が有用（必要）かなどの知見が得られたので報告する．
キーワード　未知語，Web 文書，未登録語，新語，Web マイニング，未知語処理，自然言語処理．

Prevalence Survey of Unknown Words in Japanese Web Text

Shun HATTORI† and Hiroyuki KAMEDA†

† School of Computer Science, Tokyo University of Technology
1404–1 Katakura, Hachioji, Tokyo 192–0982, Japan
E-mail: †{hattori,kameda}@cs.teu.ac.jp

Abstract　Mining the Web to extract various knowledge from the growing source has become one of the hottest research topics. However, while such a Natural Language Processing (NLP) as morphological analysis or semantic analysis for Web text, the existence of “Unknown Words” that are not registered in a NLP system’s dictionary (lexical database) is a serious impediment. In this paper, we survey the prevalence of unknown words in various domains of Japanese Web Text, e.g., dependency on its type of Web media, topics and upload date.
Key words　Unknown Words, Web Documents, Unregistered Words, New Words, Web Mining, Unknown Word Processing, Natural Language Processing (NLP).

1. はじめに

日々情報爆発して行く Web を解析して様々な知識を抽出する Web マイニングの研究が盛んに行われている．例えば，実世界で提供されている製品やサービスなどの評判抽出 [1], [2]，実世界のある場所である期間に味わうことができる体験（イベント）の抽出およびマッピング [3], [4]，語概念の階層構造（上位下位関係や部分全体関係など）[5], [6]，実世界オブジェクトの外観などの五感情報 [7], [8]，画像メディアの外観情報として典型画像や特異画像 [9], [10]，各実空間での様々な現象（雨量や交通事故）の時間変化 [11] などを Web からマイニングする研究がある．Web 上には様々な形態のメディア（Web サイト）が存在しているが，近年とりわけ，CGM（Consumer Generated Media）の重要性が増して来ている．個人の日記や覚え書き，論評などを日時をベースに記録するブログ（Blog）に始まり，電子掲示板サイト「2 ちゃんねる（2ch）[12]」，ツイート（つぶやき）と呼ばれる短文を投稿し合う SNS サイトの一種である「Twitter [13]」などが人気である．一般ユーザの情報発信やコミュニケーション・ツールというだけでなく，製品やサービスなどの提供者側が，それらの評判や潜在的ニーズを把握するための情報源としても注目されている．しかしながら，Web マイニングのように，Web テキストを形態素解析や意味解析などによって自然言語処理する際，Web テキスト中の「未知語」の存在が大きな問題になる．本稿では，自然言語処理システムが辞書中に品詞や読み，意味などを有している語を「既知語」と呼び，システムの辞書に未登録である語を「未知語」と呼ぶ．

本稿では，Web テキストに存在する多種なメディア（Web ニュースサイト「asahi.com [14]」の記事，各種 Blog サイトのエントリ，電子掲示板「2 ちゃんねる」のレス，Twitter のつぶやき），多様な話題（政治，スポーツ，オタク文化），及び，投稿日時（2007 年，2008 年，2009 年）の 3 軸に依って，どのように未知語が分布しているか頻度調査を行う．言い換えると，

Web テキストにおける未知語の出現頻度に，メディア依存性，話題（トピック）依存性，時間依存性が見られるか否かを検証する．本調査により，Web テキストを自然言語処理するシステムにおいて，どのような分野で特に未知語処理 [15] が有用（必要）か，また，Web テキストにおける未知語処理の改善の余地についても知見が得られることが期待される．

2. Web テキスト中の未知語の語彙調査

Web テキストを形態素解析や意味解析など自然言語処理する際，システムが用いる辞書に品詞や読み，意味などが未登録である語「未知語」の存在が問題になる．Web テキストを自然言語処理するシステムにおいて，どのような分野で特に未知語処理が有用（必要）かなどの知見を得るため，以下の 3 軸のクロス分析によって未知語の頻度調査を行う．

（1） メディア依存性：

Web テキストの形態に依って未知語の頻度に差が生じるか．本調査では，Web ニュースサイト「asahi.com」の記事，及び，3 種類の CGM，各種 Blog サイトのエントリ，電子掲示板「2 ちゃんねる」のレス，Twitter のつぶやきを収集した．

（2） 話題（トピック）依存性：

Web テキストで述べられている話題に依って未知語の頻度に差が生じるか．本調査では「政治」「スポーツ」「オタク（文化）」の 3 種類の話題を収集した．

（3） 時間依存性：

Web テキストが作成され，Web に投稿された日時に依って未知語の頻度に差が生じるか．本調査では「2007 年」「2008 年」「2009 年」の 3 年間を収集した．

まず，Google ウェブ検索 [16]（ドメイン検索オプション「site:」および期間指定）と Google ブログ検索 [17] を活用し，3 軸のクロス毎，最大 50 件の Web テキストを収集する．但し，メニューや広告などを手作業で除去し，見出しや本文，投稿日時，投稿者の ID や HN などの部分だけを調査対象にしている．

次に，オープンソース形態素解析エンジン「MeCab [18]」をデフォルト設定のまま用いて，全ての標本に対して形態素解析を行う．表 1 および表 2 は，Web テキストにおける未知語の頻度（異なり数および延べ数）と占有率に関して，メディア依存性，話題依存性，時間依存性があるかを検証している．唯一 CGM ではない Web ニュース，特に「政治」記事中の未知語の比率が小さい．また，Web ニュース記事よりも Blog エントリの方がテキストに含まれる形態素数が多い．Web ニュース記事には理想的な文量があるものと推測されるが，Blog エントリでは持論を熱弁しているものがあるためである．標準偏差も調べると Web ニュース記事の方が小さく，Blog エントリは短文から長文まで様々である．表 3 は，Web テキスト中の未知語の頻度（異なり数）の上位 10 件をクロス毎に示している．

ここで，未知語の多くを半角英数字記号が占めていたため，半角英数字記号から成る未知語を除くと，表 4 および表 5 のようになる．半角英数字記号から成る語を除いた未知語の頻度（異なり数）の上位 10 件をクロス毎に示している表 6 を眺めると，話題や時期に依存した未知語が見て取れる．

3. おわりに

本稿では，Web テキスト中の未知語の頻度に関して，メディア依存性，話題（トピック）依存性，及び，時間依存性の 3 軸のクロス分析によって調査を行った．メディア依存性については，基本的に個人が作成する他の 3 種類の CGM のテキストと比べて，朝日新聞社が運営する Web ニュースサイト「asahi.com」の記事中の未知語の比率が小さい事が分かった．IPA 辞書のように既に体系付けられた語彙を想定して新聞記事が書かれる傾向があるためであると考えられる．次に，話題依存性については，「政治」に関する Web テキストよりも，「スポーツ」や「オタク（文化）」などのエンタメ関連の方が未知語の比率がやや大きい事が分かった．近年，盛り上がっている話題の方が新語が生まれ易く，システム（の辞書）にとっては未知語となり易いためであると考えられる．

今後は，各クロスに対する Web テキストの標本数を増やし，話題や期間をより細かく分けるなどして，未知語の頻度だけでなく既知語の品詞なども含め，より詳細な語彙調査を行う．

文　献

[1] 鈴木泰裕, 高村大也, 奥村学: “WebLog を対象とした評価表現抽出,” 第 6 回セマンティックウェブとオントロジー研究会, SIG-SWO-A401-02 (2004).

[2] 藤村滋, 豊田正史, 喜連川優: “文の構造を考慮した評判抽出手法,” Proc. DEWS’05, 6C-i8 (2005).

[3] Taro Tezuka, Takeshi Kurashima, Katsumi Tanaka: “Toward Tighter Integration of Web Search with a Geographic Information System,” Proc. WWW’06, pp.277–286 (2006).

[4] 倉島健, 藤村考, 奥田英範: “大規模テキストからの経験マイニング,” Proc. DEWS’08, A1-4 (2009).

[5] 服部 峻, 田中 克己: “性質継承と概念の再帰的適用に基づく Web からの概念階層抽出,” 情報処理学会論文誌（トランザクション）データベース, Vol.1, No.3, pp.60–81 (2008).

[6] Shun Hattori, Katsumi Tanaka: “Extracting Concept Hierarchy Knowledge from the Web based on Property Inheritance and Aggregation,” Proc. WI’08, pp.432–437 (2008).

[7] 服部峻, 手塚太郎, 田中克己: “文書中の地物画像を言語的記述で代替するための地物の外観情報の Web からの抽出,” 情報処理学会論文誌（トランザクション）データベース, Vol.48, No.SIG11 (TOD34), pp.69–82 (2007).

[8] Shun Hattori, Taro Tezuka, and Katsumi Tanaka: “Mining the Web for Appearance Description,” Proc. DEXA’07, LNCS Vol.4653, pp.790–800 (2007).

[9] 服部峻, 田中克己: “色名抽出と色特徴量変換に基づく典型的画像の Web 検索,” 日本データベース学会論文誌 (DBSJ Letters), Vol.6, No.4, pp.9–12 (2008).

[10] 服部 峻, 田中 克己: “Web 抽出した特異な色名と色特徴量変換に基づく特異画像の Web 検索,” 情報処理学会論文誌（トランザクション）データベース, Vol.3, No.1, pp.49–63 (2010).

[11] Shun Hattori, Katsumi Tanaka: “Mining the Web for Access Decision-Making in Secure Spaces,” Proc. SCIS&ISIS’08, TH-G3-4 (2008).

[12] 2 ちゃんねる掲示板へようこそ, http://www.2ch.net/ (2010).

[13] Twitter, http://twitter.com/ (2010).

[14] asahi.com：朝日新聞社の速報ニュースサイト, http://www.asahi.com/ (2010).

[15] 福岡知隆, 税田竜一, 久保村千明, 服部峻, 亀田弘之: “文の類似性を用いた未知語処理手法の提案とそれに基づく円滑な対話応答システムの作成,” 情報処理学会 第 72 回全国大会, 6X-2 (2010).

[16] Google ウェブ検索, http://www.google.co.jp/ (2010).

[17] Google ブログ検索, http://blogsearch.google.co.jp/.

[18] MeCab, http://mecab.sourceforge.net/ (2010).

表 1　Web テキストにおける未知語の頻度と占有率
（MeCab 解析のまま，頻度は 1 文書当たりの異なり数）
Table 1　Type Frequency of Unknown Words in Japanese Web Text with English One-byte Characters.

		話題「政治」				話題「スポーツ」				話題「オタク（文化）」				3 話題
		2007	2008	2009	3 年間	2007	2008	2009	3 年間	2007	2008	2009	3 年間	
メディア	形態素数	333.52	300.36	220.28	284.72	207.32	282.10	242.34	243.92	296.08	324.18	319.40	313.22	280.62
「asahi」	未知語数	7.66	8.54	6.62	7.61	9.22	10.86	10.32	10.13	11.96	10.26	10.64	10.95	9.56
	未知語率	0.0230	0.0284	0.0301	0.0272	0.0444	0.0385	0.0426	0.0419	0.0404	0.0316	0.0333	0.0351	0.0347
メディア	形態素数	400.96	376.00	357.52	378.16	144.04	181.66	198.84	174.85	269.42	274.80	256.52	266.91	273.31
「Blog」	未知語数	31.96	25.32	19.34	25.54	16.52	22.70	23.42	20.88	29.62	25.64	35.46	30.24	25.55
	未知語率	0.0797	0.0673	0.0541	0.0670	0.1147	0.1250	0.1178	0.1191	0.1099	0.0933	0.1382	0.1138	0.1000
メディア	形態素数	108.34	136.18	136.52	127.01	150.02	82.90	123.68	118.87	110.20	59.28	76.02	81.83	109.24
「2ch」	未知語数	21.70	27.82	33.52	27.68	35.28	24.46	35.76	31.83	18.76	14.00	24.50	19.09	26.20
	未知語率	0.2003	0.2043	0.2455	0.2167	0.2352	0.2951	0.2891	0.2731	0.1702	0.2362	0.3223	0.2429	0.2442
メディア	形態素数	45.67	44.78	56.72	49.05	38.80	43.17	45.18	42.38	31.50	45.93	46.80	41.41	44.28
「Twitter」	未知語数	18.00	16.06	17.70	17.25	13.40	16.25	20.16	16.60	13.50	16.07	16.50	15.36	16.40
	未知語率	0.3942	0.3586	0.3121	0.3549	0.3454	0.3764	0.4462	0.3893	0.4286	0.3498	0.3526	0.3770	0.3737
	形態素数				209.74				145.00				175.84	
4 メディア	未知語数				19.52				19.86				18.91	
	未知語率				0.1665				0.2059				0.1922	

表 2　Web テキストにおける未知語の頻度と占有率
（MeCab 解析のまま，頻度は 1 文書当たりの延べ数）
Table 2　Token Frequency of Unknown Words in Japanese Web Text with English One-byte Characters.

		話題「政治」				話題「スポーツ」				話題「オタク（文化）」				3 話題
		2007	2008	2009	3 年間	2007	2008	2009	3 年間	2007	2008	2009	3 年間	
メディア	形態素数	917.06	788.26	583.20	762.84	503.12	732.22	631.88	622.41	741.60	789.38	823.34	784.77	723.34
「asahi」	未知語数	10.28	12.56	7.76	10.20	13.66	13.88	14.24	13.93	15.56	14.10	13.68	14.45	12.86
	未知語率	0.0112	0.0159	0.0133	0.0135	0.0272	0.0190	0.0225	0.0229	0.0210	0.0179	0.0166	0.0185	0.0183
メディア	形態素数	1399.5	1125.8	1249.2	1258.2	303.96	446.34	513.22	421.17	748.52	791.06	807.06	782.21	820.51
「Blog」	未知語数	82.96	44.44	31.76	53.05	23.94	40.28	36.40	33.54	49.30	38.78	113.30	67.13	51.24
	未知語率	0.0593	0.0395	0.0254	0.0414	0.0788	0.0902	0.0709	0.0800	0.0659	0.0490	0.1404	0.0851	0.0688
メディア	形態素数	193.22	272.32	291.06	252.20	306.62	135.52	230.92	224.35	192.62	85.60	130.86	136.36	204.30
「2ch」	未知語数	31.72	46.16	68.98	48.95	64.22	34.86	56.84	51.97	28.36	18.66	44.06	30.36	43.76
	未知語率	0.1642	0.1695	0.1902	0.2370	0.2094	0.2572	0.2461	0.2376	0.1472	0.2180	0.3367	0.2340	0.2206
メディア	形態素数	53.67	53.28	70.32	59.09	42.60	50.67	52.24	48.50	34.50	53.60	56.72	48.27	51.95
「Twitter」	未知語数	19.33	17.11	18.80	18.41	14.00	17.58	22.16	17.91	13.50	17.13	17.62	16.08	17.47
	未知語率	0.3602	0.3212	0.2673	0.3163	0.3286	0.3470	0.4242	0.3666	0.3913	0.3197	0.3106	0.3405	0.3411
	形態素数				583.07				329.11				437.91	
4 メディア	未知語数				32.66				29.34				32.00	
	未知語率				0.1403				0.1768				0.1695	

表 3　Web テキスト中の未知語の上位 10 件（MeCab 解析のまま，異なり数）

Table 3　Top 10 Unknown Words in Japanese Web Text with English 1-byte Characters.

	メディア「asahi.com」																	
	話題「政治」						話題「スポーツ」						話題「オタク（文化）」					
	2007		2008		2009		2007		2008		2009		2007		2008		2009	
1	2007	48	2008	44	2009	50	2007	45	[	38	2009	30	2007	48	2008	49	2009	39
2	[	22	[	31	12	29	02	13	]	38	/	18	[	13	1	17	[	14
3	]	22	]	31	1	14	03	9	/	38	12	13	]	13	ボーイズラブ	12	]	14
4	5	10	應大	8	11	13	01	8	2008	11	11	10	11	12	コミックス	11	09	11
5	7	10	2	7	/	12	12	6	04	10	[	9	12	12	ＢＬ	8	ボーイズラブ	10
6	應大	9	6	7	2	9	05	5	10	9	]	9	ボーイズラブ	9	4	7	10	10
7	9	8	11	7	3	9	07	5	05	6	09	9	08	9	10	7	12	10
8	10	7	オバマ	6	5	8	22	4	11	6	10	8	1	8	[	6	コミックス	8
9	11	7	9	6	22	8	)	4	12	6	04	7	09	7	]	6	11	7
10	28	6	ポピュリズム	5	20	7	(	4	29	6	1	6	ブログ	6	11	6	ＢＬ	6

	メディア「ブログ（Blog）」																	
	話題「政治」						話題「スポーツ」						話題「オタク（文化）」					
	2007		2008		2009		2007		2008		2009		2007		2008		2009	
1	2007	49	2008	47	2009	49	2007	48	2008	50	2009	50	2007	50	2008	48	2009	48
2	:	19	:	21	12	34	:	22	:	18	:	19	-	23	:	21	:	27
3	-	19	.	21	:	20	-	19	(	17	3	17	1	19	2	19	12	22
4	.	16	2	20	)	20	ブログ	15	2	15	12	17	)	19	-	19	2	18
5	/	15	/	20	1	16	)	15	-	14	11	16	:	18	1	18	)	17
6	1	14	-	18	.	16	10	13	10	14	1	15	2	18	/	16	(	16
7	5	14	10	18	-	16	(	12	1	13	)	15	/	18	)	14	1	15
8	)	14	3	16	2	13	2	10	3	12	2	14	(	17	3	13	/	15
9	ブログ	13	12	16	3	13	04	10	/	12	.	14	.	16	.	13	4	13
10	23	13	)	15	11	12	1	9	)	12	(	13	10	16	(	13	10	12

	メディア「2 ちゃんねる（2ch）」																	
	話題「政治」						話題「スポーツ」						話題「オタク（文化）」					
	2007		2008		2009		2007		2008		2009		2007		2008		2009	
1	2007	50	2008	50	:	50	:	50	2008	50	ID	50	2007	50	2008	50	2009	50
2	:	50	:	50	/	50	/	50	:	50	:	50	:	50	:	50	:	50
3	/	50	/	50	(	50	(	50	/	50	/	50	/	50	/	50	/	50
4	(	50	(	50	)	50	)	50	(	50	(	50	(	50	(	50	(	50
5	)	50	)	50	2009	49	2007	49	)	50	)	50	)	50	)	50	)	50
6	ID	44	ID	42	ID	48	ID	47	ID	50	2009	49	>>	18	05	16	ID	46
7	11	19	0	23	0	30	1	29	:???	27	2	30	09	16	08	15	12	19
8	>>	15	1	21	2	28	.	29	0	22	06	30	07	15	.	13	03	15
9	01	15	http	20	://	27	://	28	02	21	01	30	08	14	10	12	2	12
10	0	13	://	20	.	27	http	26	01	18	9	26	1	13	11	12	23	11

	メディア「Twitter」																	
	話題「政治」						話題「スポーツ」						話題「オタク（文化）」					
	2007		2008		2009		2007		2008		2009		2007		2008		2009	
1	2007	3	2008	18	2009	50	2007	5	2008	12	2009	50	2007	2	2008	15	2009	50
2	:	3	:	18	:	50	:	5	:	12	:	50	:	2	:	15	:	50
3	,	3	,	18	,	50	,	5	,	12	,	50	,	2	,	15	,	50
4	th	2	th	15	th	43	web	4	th	12	th	45	th	2	AM	12	th	45
5	PM	2	PM	10	AM	32	th	3	AM	6	web	31	Sword	1	th	11	AM	29
6	Jul	2	AM	8	web	31	AM	3	PM	6	PM	30	May	1	web	9	web	25
7	4	2	web	7	Dec	24	PM	2	web	6	http	23	web	1	1	5	PM	21
8	.	2	6	7	_	20	3	2	5	4	://	23	nii	1	2	4	Dec	17
9	nd	1	7	7	PM	18	28	2	8	4	/	23	ヤキモノ	1	6	4	Nov	11
10	28	1	11	7	7	17	ガイジンスポーツ	1	11	4	.	23	Twit	1	st	4	2	11

表 4　Web テキストにおける未知語の頻度と占有率
（半角英数字記号を無視，頻度は 1 文書当たりの異なり数）
Table 4　Type Frequency of Unknown Words in Japanese Web Text without English One-byte Characters.

		話題「政治」				話題「スポーツ」				話題「オタク（文化）」				3 話題
		2007	2008	2009	3 年間	2007	2008	2009	3 年間	2007	2008	2009	3 年間	
メディア	形態素数	329.24	296.08	214.80	280.04	203.32	277.24	237.1	239.22	290.94	319.72	315.5	308.72	275.99
「asahi」	未知語数	3.38	4.26	1.14	2.93	5.22	6.00	5.08	5.43	6.82	5.80	6.74	6.45	4.94
	未知語率	0.0103	0.0144	0.0053	0.0100	0.0257	0.0216	0.0214	0.0229	0.0234	0.0181	0.0214	0.0210	0.0180
メディア	形態素数	377.46	356.42	343.12	359.00	132.64	167.22	183.44	161.10	250.48	258.68	232.66	247.27	255.79
「Blog」	未知語数	8.46	5.74	4.94	6.38	5.12	8.26	8.02	7.13	10.68	9.52	11.60	10.60	8.04
	未知語率	0.0224	0.0161	0.0144	0.0176	0.0386	0.0494	0.0437	0.0439	0.0426	0.0368	0.0499	0.0431	0.0349
メディア	形態素数	88.08	110.04	105.70	101.27	119.18	60.50	91.04	90.24	93.46	46.10	53.52	64.36	85.29
「2ch」	未知語数	1.44	1.68	2.70	1.94	4.44	2.06	3.12	3.21	2.02	0.82	2.00	1.61	2.25
	未知語率	0.0163	0.0153	0.0255	0.0191	0.0373	0.0340	0.0343	0.0352	0.0216	0.0178	0.0374	0.0256	0.0266
メディア	形態素数	28.33	30.44	40.08	32.95	26.20	27.92	26.30	26.81	18.50	31.33	31.78	27.20	28.99
「Twitter」	未知語数	0.67	1.72	1.06	1.15	0.80	1.00	1.28	1.03	0.50	1.47	1.48	1.15	1.11
	未知語率	0.0235	0.0566	0.0264	0.0355	0.0305	0.0358	0.0487	0.0383	0.0270	0.0468	0.0466	0.0401	0.0380
	形態素数				193.32				129.34				161.89	
4 メディア	未知語数				3.10				4.20				4.95	
	未知語率				0.0205				0.0351				0.0324	

表 5　Web テキストにおける未知語の頻度と占有率
（半角英数字記号を無視，頻度は 1 文書当たりの延べ数）
Table 5　Token Frequency of Unknown Words in Japanese Web Text without English One-byte Characters.

		話題「政治」				話題「スポーツ」				話題「オタク（文化）」				3 話題
		2007	2008	2009	3 年間	2007	2008	2009	3 年間	2007	2008	2009	3 年間	
メディア	形態素数	911.46	781.92	577.28	756.89	497.88	726.40	625.20	616.49	735.24	784.52	819.00	779.59	717.66
「asahi」	未知語数	4.68	6.22	1.84	4.25	8.42	8.06	7.56	8.01	9.20	9.24	9.34	9.26	7.17
	未知語率	0.0051	0.0080	0.0032	0.0054	0.0169	0.0111	0.0121	0.0134	0.0125	0.0118	0.0114	0.0119	0.0102
メディア	形態素数	1332.1	1090.6	1224.1	1215.6	287.34	419.54	488.98	398.62	715.32	766.20	722.86	734.79	783.01
「Blog」	未知語数	15.58	9.26	6.74	10.53	7.32	13.48	12.16	10.99	16.10	13.92	29.10	19.71	13.74
	未知語率	0.0117	0.0085	0.0055	0.0086	0.0255	0.0321	0.0249	0.0275	0.0225	0.0182	0.0403	0.0270	0.0210
メディア	形態素数	163.36	228.26	226.44	206.02	248.60	103.12	177.56	176.43	166.78	68.10	89.04	107.97	163.47
「2ch」	未知語数	1.86	2.10	4.36	2.77	6.20	2.46	3.48	4.05	2.52	1.16	2.24	1.97	2.93
	未知語率	0.0114	0.0092	0.0193	0.0133	0.0249	0.0239	0.0196	0.0228	0.0151	0.0170	0.0252	0.0191	0.0184
メディア	形態素数	35.00	37.94	52.90	41.95	29.40	34.17	31.44	31.67	21.50	38.20	40.76	33.49	35.70
「Twitter」	未知語数	0.67	1.78	1.38	1.27	0.80	1.08	1.36	1.08	0.50	1.73	1.66	1.30	1.22
	未知語率	0.0190	0.0469	0.0261	0.0307	0.0272	0.0317	0.0433	0.0341	0.0233	0.0454	0.0407	0.0365	0.0337
	形態素数				555.12				305.80				413.96	
4 メディア	未知語数				4.71				6.03				8.06	
	未知語率				0.0145				0.0244				0.0236	

表 6　Web テキスト中の未知語の上位 10 件（半角英数字記号は無視，異なり数）

Table 6　Top 10 Unknown Words in JAP Web Text without ENG 1-byte Characters.

	メディア「asahi.com」								
	話題「政治」			話題「スポーツ」			話題「オタク（文化）」		
	2007	2008	2009	2007	2008	2009	2007	2008	2009
1	應大 9	應大 8	オバマ 6	ヘス 3	Ｉ 5	Ｉ 4	ボーイズラブ 9	ボーイズラブ 12	ボーイズラブ 10
2	捉 4	オバマ 6	ＣＯＰ 6	ＢｏＡ 3	ロレーナ・オチョア 4	トップアスリート 3	ブログ 6	コミックス 11	コミックス 8
3	ナントカ 3	ポピュリズム 5	習近 5	ＭＢＣ 3	アニカ・ソレンスタム 4	モチベーション 2	ＢＬ 5	ＢＬ 8	ＢＬ 6
4	ＥＰＡ 2	マケイン 3	シー・チンピン 5	キム・ヘス 3	ディープスカイ 3	スプリンターズ 2	ＤＶＤ 4	イカサマアシスタント 3	ツッコミ 5
5	ミッドタウン 2	ロールズ 3	ラスムセン 3	巨塔 2	Ｖｏｌ 3	プエナビスタ 2	アキバ 3	ポックリ 3	キログラム 2
6	岐夫 2	ょう 3	濤 2	トレ 2	ＡＣＬ 3	マセラティ 2	ディープ 3	ツンデレ 3	アーカイブス 2
7	アルカイダ 2	リヴァイアサン 2	習氏 2	ヒカル 2	フィル・ミケルソン 3	クロフネ 2	アジコ 3	アキバ 2	アニマゲ 2
8	ＳＡＲＳ 1	ワーキングプア 2	ＮＧＯ 1	ＦＣ 2	パーオン 3	フットサル 2	コミックス 3	ツッコミ 2	ＩＴ 2
9	イジメ 1	トマス・ホップズ 2	ユーチューブ 1	アイフィルム 2	ＬＰＧＡ 3	アスリート 2	ＰＣ 3	妖人 2	ハイジ 2
10	ＮＧＯ 1	ポリティクス 2	ＣＨＡＮＧＥ 1	エイペックス 2	オフィシャルスポンサー 3	ウェブ 2	カンジ 3	ユンケル 2	ＭＡＮＧＡ 2
	メディア「ブログ（Blog）」								
	話題「政治」			話題「スポーツ」			話題「オタク（文化）」		
	2007	2008	2009	2007	2008	2009	2007	2008	2009
1	ブログ 2	ブログ 8	ブログ 10	ブログ 15	ブログ 7	ブログ 9	ブログ 14	ブログ 7	ブログ 11
2	思社 2	オバマ 2	習近 6	ジャージ 2	モチベーション 3	プレイ 3	オタ 11	オタ 6	オタ 6
3	マシ 2	フトコロ 2	オバマ 4	アマチア 1	コーチ 3	フットサル 2	コミケ 7	アキバ 5	アキバ 4
4	モチベーション 2	マシ 2	濤 3	ミラン 1	ＦＣ 3	アレ 2	アキバ 5	ツッコミ 3	ヲタ 3
5	バラエティ 2	ゴネ 2	ウェブ 3	スカブー 1	トレ 3	ヌンチャク 2	エロゲ 4	キモイ 3	コラボ 3
6	ダーティ 2	ヒドイ 2	習副 2	プレオ 1	プレイ 2	wwww 2	ツッコミ 3	ヲタ 3	サーセン 2
7	ジジェク 2	ＣＨＡＮＧＥ 2	ツイッター 2	ＵＳＣ 1	ジャージ 2	カワイイ 2	エヴァンゲリオン 3	イケメン 3	ハルヒ 2
8	ブロガー 2	アキバ 1	デタラメ 2	ボカ 1	アメフト 2	アスリート 2	ワケ 3	コレ 3	www 2
9	サヨク 2	プーチン 1	ソーシャル 2	サイボウズ 1	エクササイズ 2	ヤバイ 2	コミ 3	ヲタク 2	シャア 2
10	ヘーゲル 2	ロザン 1	マシ 2	チャンバラ 1	ワケ 2	メタボリック 1	アニオタ 2	アニソン 2	ザク 2
	メディア「2 ちゃんねる（2ch）」								
	話題「政治」			話題「スポーツ」			話題「オタク（文化）」		
	2007	2008	2009	2007	2008	2009	2007	2008	2009
1	フーシェ 6	スレ 10	スレ 15	スレ 13	フーリガン 3	スレ 16	スレ 8	キモオタ 4	ガンダムオタク 7
2	タレーラン 6	コピペ 4	スレッド 8	サンスポ 8	メッシ 3	アスリート 4	キモイ 6	キモイ 4	キモイ 5
3	プーチン 3	ニューススレ 3	ぐた 8	スレッド 6	www 2	ベッカム 4	オタ 6	オタ 4	ケツ 3
4	ロベスピエール 3	ぐた 3	づる 4	コソーリ 5	ロビーニョ 2	スポーツチャンバラ 3	ブログ 3	アニオタ 4	キチガイ 3
5	メルマガ 2	スレッド 2	ニューススレ 4	圧太 5	プレイ 2	オバマ 2	アキバ 2	スレ 2	アキバ 2
6	ＩＴ 1	モテ 1	ウホ 3	スポーツパークテ 5	wwwww 2	ニュートラル 2	キモヲタ 2	キチガイ 1	スレ 2
7	ノムヒョン 1	ブログ 1	習副 3	ザバス 2	ヤキブタ 2	アメフト 2	ストーカー 2	ガンオタ 1	ガノタ 2
8	ネオコン 1	ネチズン 1	ネトウヨ 3	ぐた 2	チビッ 2	マイク・ロイコ 2	エヴァ 2	ＳＮＳ 1	ハルヒ 2
9	アカポス 1	サヨク 1	ＰＣ 2	レクサス 2	チビデブ 2	サッカーヲタク 2	ツンデレ 2	ゲームオタ 1	オタ 2
10	メルケル 1	ググ 1	プーチン 1	コーチ 2	レズ 2	バラク・オバマ 2	アスペ 1	エヴァ 1	パクリ 1
	メディア「Twitter」								
	話題「政治」			話題「スポーツ」			話題「オタク（文化）」		
	2007	2008	2009	2007	2008	2009	2007	2008	2009
1	バーテンダー・スタン 1	マ 5	ツイッター 4	ガイジンスポーツ 1	コンマイスポーツクラブ 1	リーボック 2	ヤキモノ 1	マクロス 1	モテ 1
2		儿 5	Ｔｗｉｔｔｅｒ 2	コンシューマー 1	キウチミツアキ 1	ミシマガ 2		モエキャラ 1	ブログ 1
3		コ 5	アクター 1	デュクシング 1	ょもす 1	コンピ 2		ラジルギノア 1	ジャニコン 1
4		アミン・ダダ 1	オカルト 1		アキバ 1	ベロン 1		ミズモ 1	ステイタス 1
5		ブログ 1	ハイポリティックス 1			ヤバイ 1		コンバース 1	ハルヒ 1
6		ＥＴＣ 1	ゲッペルス 1			メッシ 1		コードギアス 1	オタイベント 1
7		センセー 1	ユーチューブ 1			アベシ 1		ゲロカエルン 1	スイーツ 1
8		オバマガール 1	オバマ 1			アディダス 1		ギアス 1	フジョシ 1
9		キムタク 1	パロンシュレイヒ 1			デジカメ 1		サブカル 1	ジャニオタ 1
10			ソーシャルメディア 1			ノム 1		カワイイ 1	コミケ 1