拡張固有表現タグ付きコーパスの構築

橋本 泰一 † 乾 孝司 † 村上 浩司 ‡

† 東京工業大学 統合研究院
‡ 奈良先端科学技術大学院大学 情報科学研究科
{hashimoto, inui}@iri.titech.ac.jp, kmurakami@is.naist.jp

概要：従来，日本語における固有表現タグ付きコーパスは，評価型ワークショップ IREX の固有表現抽出タスクのために作成された毎日新聞をもとにしたコーパス（CRL 固有表現データ，のべ 19,254 表現，異なり 7,153 表現）が唯一であった．本論文では，新聞記事と白書の 2 種類の文書に対して，200 種類に固有表現を分類している「関根の拡張固有表現階層」の定義に基づき作成したコーパスについて報告する．．新聞記事においては，毎日新聞 31 日分の記事（8,584 記事）に対して，のべ 252,763 表現（異なり 63,545 表現）がタグ付けされた．白書においては，400 文書に対して，のべ 74,203 表現（異なり 23,857 表現）がタグ付けされた．

キーワード：拡張固有表現階層, コーパス

Constructing Extended Named Entity Annotated Corpora

Taiichi Hashimoto † Takashi Inui † Koji Murakami ‡

† Tokyo Institute of Technology
‡ Nara Institute of Science and Technology
{hashimoto, inui}@iri.titech.ac.jp, kmurakami@is.naist.jp

Abstract：In Japanese, CRL named entity data which was constructed in IREX workshop is only a corpus annotated tags of named entities. The corpus consists of about 1,100 articles of Mainichi newspaper and contains a total of 19,254 tags. We presents two new named entity annotated corpora. Our corpora consists of about 8,500 articles of Mainichi newspaper and 400 white papers, and our definition of named entity tags is Sekine's extended named entity hierarchy. The Mainichi newspaper corpus contains a total of 252,763 tags and the white paper corpus contains a total of 74,203 tags.

Keywords：extended named entity hierarchy, corpus

1 はじめに

固有表現とは，「机」「椅子」「空」「愛」といった一般的な概念を表す表現ではなく，物，イベントや考え方を表す言語表現（例：夏目漱石，東京オリンピック，日本）であり，質問応答，情報抽出，機械翻訳，テキストマイニングなどに用いられる自然言語処理における重要な基礎知識である．これまで，日本語においては，評価型ワークショップ IREX において，新聞記事に対して固有表現タグ付きコーパス（CRL 固有表現データ）が構築され，そのデータをもとに日本語における固有表現抽出に関する研究が進み，様々な抽出手法 [9, 10, 13, 14, 15] が提案されてきた．

IREX で定義された固有表現の種類は，組織名，人名，地名，日付表現，時間表現，金額表現，割合表現，固有物名の 8 種類であり，毎日新聞記事にのみタグが付与されている．しかし，このコーパスを利用して開発された固有表現抽出器を，質問応答システム，情報抽出システムやテキストマイニングに利用しようとしても実際に抽出できる固有表現の種類が少なく，新聞以外の分野の文書に対する精度も十分満足のいくレベルではない．さらなる高度な言語処理システムの発展に向けて，より詳細に定義さ

れた固有表現の定義のもと様々な分野のタグが付与された言語資源の作成が必要である．

本研究では，様々なジャンルの固有表現タグ付きコーパスの構築に向けて，200 種類にもおよぶ固有表現を定義した「関根の拡張固有表現階層」をもとに，毎日新聞約 8500 記事，白書 400 文書に対して固有表現タグ付けを行ったコーパスについて報告する．

2 関根の拡張固有表現階層

固有表現タグ付きコーパス構築に向けて，固有表現の定義として，「関根の拡張固有表現階層」(以下，拡張固有表現)[*1]を採用した．「関根の拡張固有表現階層」は，MUC(Message Understanding Conference) プロジェクトで策定された固有表現の定義 [1]，それを基に策定された日本における IREX プロジェクトの定義 [3]，ACE(Automatic Content Extraction) プロジェクト[*2]の定義をもとに，関根が拡張を行った固有表現の定義 [2, 4, 5] である．関根は，質問応答システム，情報抽出，機械翻訳，情報検索，要約などの自然言語処理技術への応用を目的として，この定義の策定を行っている．

拡張固有表現の大きな特徴は，固有表現の種類の豊富さである．MUC では，組織名，人名，地名，日付表現，時間表現，金額表現，割合表現の 7 種類，IREX では，MUC の 7 種類に固有物名を加えた 8 種類を固有表現として定義している．一方，拡張固有表現（バージョン 7.1.0）では 200 種類のタグの定義を行っている．これは様々な自然言語処理技術への応用を考慮し，新聞記事や百科事典などに見られる概念や単語を考慮していることに起因する．

拡張固有表現に関する従来の研究では，主に固有表現辞書を獲得する研究 [6, 7, 8] が行われており，固有表現解析手法（タガー）に関する研究は少ない [11]．この主な原因は，大規模なタグ付きコーパスが存在しないことが原因であると思われる．

3 拡張固有表現タグ付きコーパス

3.1 拡張固有表現タグ付きコーパスの概要

本研究では，2 種類の拡張固有表現タグ付きコーパスを構築した．一つは，毎日新聞をもとにしたコーパス，もう一つは，白書をもとにしたコーパスである．

毎日新聞の新聞記事（31 日分，8,584 記事）に対して，最新の拡張固有表現 (Version 7.1.0) の定義に従い人手によりタグ付けを行った．白書コーパスは，文部科学省科学研究費補助金特定研究「日本語コーパス」プロジェクトにより現在モニター公開されている白書（約 500 万語，1,500 サンプル）からランダムに 400 サンプルを選び，拡張固有表現 (Version 7.1.0) の定義にそって，人手によりタグ付けを行った．タグ付け結果の概要を表 1 に示す．

3.2 拡張固有表現タグの頻度分布による比較

白書は，毎日新聞に比べ，タグ付けされた文書数が約 20 分の 1 と少ない．しかし，白書の文書は，毎日新聞の文書に対し 1 文書当りの平均文字数では約 13 倍多く，そのため総文字数では約 3 分の 2 の違いしかない．

毎日新聞に付与された拡張固有表現のタグ数は，のべ 252,763 個，異なり 79,632 個であった．白書に付与された拡張固有表現のタグ数は，のべ 74,203 個，異なり 23,857 個であった．付与された固有表現タグの数や種類は，白書に比べ毎日新聞の方が比較的多い．これは，白書に比べ，毎日新聞の方が文書の内容が多種多様であるため，様々なタグが付与されたためである．さらに，プロスポーツの試合結果に関する記事，株価の変動に関する記事や選挙結果に関する記事など，組織名や人名が列挙されている記事があり，特に多くの固有表現を含む記事が存在することもタグの数の増加に起因している．

拡張固有表現の各タグにおける表現の頻度分布を表 2，表 3，表 4 に示す．毎日新聞では「人名」，「地位・職業名」，「日付表現」，「国名」，「市区町村名」が多く，白書では「日付表現」，「割合表現」，「国名」，「政府組織名」，「金額表現」が多かった．毎日新聞は事件事故に関する内容の文書，白書は政治的活動に関する内容の文書を多く含む傾向にあるためである．

3.3 拡張固有表現タグの曖昧性に関する比較

同一の表現に付与されたタグの種類の数の分布とその表現例について表 5 に示す．タグの曖昧性が多い表現の多くは数値と地名であった．この傾向は毎日新聞と白書でほとんど変わらない．拡張固有表現の定義において，数値表現は数字が表す対象によって付与されるタグが異なるため，様々なタグが付与される傾向にある．以下に，表現「**1**」の例を示す．

- **施設数**
  ... ５４年度末現在，全国の主要港等に病院３，診療所２，保養所６８，海外福祉施設 **1**（ラスパルマス），船員保険総合福祉センター２，休養所７か所が設けられている。...
- **組織数**
  ... ３０，０７８企業によって，合併８４，企業組合 **1**，協業組合２９９，協同組合２１２及び

[*1] http://nlp.cs.nyu.edu/ene/
[*2] http://www.nist.gov/speech/tests/ace/

表 1　拡張固有表現タグ付きコーパスの概要

	文書数	総文字数	1 文書当りの平均文字数	表現数		1 文書当りの平均表現数
				のべ	異なり	
CRL	1,174	593,763	505.8	19,254	7,153	16.4
毎日新聞	8,584	3,643,361	424.4	252,763	63,545	29.4
白書	400	2,340,364	5850.9	74,203	23,857	185.5

業務提携６２０の集約化が行われ，...

- **割合表現**
  ... 死亡率の低下により死亡数も減少したため，親と子の世代比がほぼ **1** になっている。...

地名に関する表現は，「市区町村名」と「都道府県州名」の曖昧性だけでなく，「駅名」，「スポーツ競技団体」，「学校名」，「企業名」の一部に地名を含み，その省略形として記述されることが多いためである．以下に，表現「**東京**」の例を示す．

- **市区町村名**
  ... 酸化物については **東京**，大阪等の大都市において環境基準を大幅に上回る汚染濃度を示しており，...
- **都道府県州名**
  ... 死亡者数を都道府県別に見ると，茨城，高知が１５．７で最も高く，以下，鳥取，山梨，三重，滋賀，徳島などが高率であり，**東京**，大阪は極めて低率となっている。...
- **駅名**
  ... 新幹線については、全国新幹線鉄道網の整備を図るため、東北（**東京**―盛岡間）、上越（大宮―新潟間）、成田（**東京**―成田空港間）の各新幹線の建設が、国鉄、日本鉄道建設公団によって施工中である。...
- **企業名**
  ... 邦銀は **東京**、富士、住友、日本興業の四行が各三億ドルの支援を求められた。...

また，同一表現に付与されたタグのペアとその表現例について表 6，表 7 に示す．毎日新聞では，最も同一表現に付与されやすいタグは，「競技組織名 その他」と「学校名」であった．これは，高校野球の結果について報道する記事が多いことに起因する．「競技名組織 その他」と「企業名」も同様である．また，食べ物に関する表現，地名に関する表現や数値表現に対する曖昧性も多いことが確認できた．一方，白書では，毎日新聞と同様に食べ物に関する表現，地名に関する表現，数値表現に関する表現が曖昧であった．白書特有の曖昧性として，「会議名」と「国際組織名」があった．この例として，欧州安全保障協力会議や UNCED(国連環境開発会議，地球サミット) などがある．

- **会議名**
  ... ，５８年３月の**臨時行政調査会**の答申の趣旨を体して，今後一層自主的な経営改善の促進 ...
- **組織名 その他**
  ... **臨時行政調査会**第一次答申において指摘された事項を実施するための措置として，...
- **政府組織名**
  ... **臨時行政調査会**及び臨時行政改革推進審議会による改革方策等の着実な実施を図るなど、...

結果から，タグの曖昧性は大きく以下の 4 種類の原因により生じていると考えられる．1 は，ある表現が複数のタグの意味を持つ場合である．例として「ニガウリ」，「コーヒー」，「まぐろ」などがある．2 は，前述した数値表現の曖昧性である．3 は，本来の表現が縮退され用いられるとき，他の表現あるいは縮退された表現と一致する場合である．この例は，地名によく見られる傾向である．4 は，換喩や提喩 [12] として表現が利用される場合である．特に組織名やブランド化された製品名がこれに当る．

1 および 2 により生じる曖昧性は比較的判別しやすいが，3 および 4 の曖昧性については人手タグ付けを行う場合に判別が困難であるようであり，同じような文脈においても異なるタグが付与される傾向にあった．

1. **真の曖昧性**
   例）ニガウリ → 「食べ物名 その他」「植物名」
   まぐろ → 「食べ物名 その他」「魚類」
2. **数値が表す対象により生じる曖昧性**
   例）1 件 → 「イベント数」「製品数」
3. **縮退により生じる曖昧性**
   例）広島 →「都道府県州名」「駅名」「プロ競技組織名」
4. **換喩・提喩により生じる曖昧性**
   例）欧州安全保障協力会議 →「会議名」「組織名」
   横浜高校 → 「学校名」「競技組織名 その他」
   ハリー・ポッターと賢者の石 → 「文学名」「映画名」

4 おわりに

本論文では，新たな固有表現タグ付きコーパスの構築に向けて，固有表現タグを付与した新聞記事データと白書データの比較した結果について報告した．

「関根の拡張固有表現階層」の定義 (Version 7.1.0) に則って，毎日新聞 (8,584 記事)，白書 (400 文書) に対してタグ付けを行った．その結果，毎日新聞では，のべ 252,763 個，異なり 79,632 個のタグが付与され，白書では，のべ 74,203 個，異なり 23,857 個のタグが付与された．

毎日新聞に比べ白書はタグ付けされた固有表現数が少ない傾向にあった．毎日新聞では，数値表現はその数値の対象が異なることにより複数のタグが付与されやすく，地名は駅名，学校名，競技組織名などの他の表現の省略形として記述されることが多く曖昧になりやすいことがわかった．一方，白書では，数値表現に加え，国際会議名と組織名に対して複数のタグが付与されやすいことがわかった．

特定研究「日本語コーパス」プロジェクトでは，新聞記事や白書以外にも様々なジャンルのコーパスが構築されている．このプロジェクトの成果を利用して，新聞記事や白書以外のジャンルの文書に対して，コーパスを構築していく予定である．また，効率的にタグ付け作業を行うための仕様の作成，タグ付けツールや拡張固有表現抽出手法の構築を目指す．

謝辞

本実験を実施するにあたり，ニューヨーク大学の関根聡氏には，毎日新聞記事への拡張固有表現タグデータのご提供，およびタグ修正作業に対する多大なる助言をいただきました．ここに，心より感謝の意を表します．

参考文献

[1] Grishman, R. and Sundheim, B. Message Understanding Conference - 6: A Brief History, COLING-96 (1996).

[2] Sekine, S. Extended Named Entity Ontology with Attribute Information, In Proceedings of the 5th International Conference on Language Resources and Evaluation (2008).

[3] Sekine, S. and Isahar, H. IREX: IR and IE Evaluation project in Japanese, LREC2000 (2000).

[4] Sekine, S. and Nobata, C. Definition, Dictionary and Tagger for Extended Named Entities, In Proceedings of the Forth International Conference on Language Resources and Evaluation (2004).

[5] Sekine, S., Sudo, K. and Nobata, C. Extended Named Entity Hierarchy, LREC2002 (2002).

[6] 塩入寛之, 岡部正幸, 阿部洋丈, 梅村恭司固有表現自動獲得に向けての固有表現とコンテキストの関連度, 情報処理学会自然言語処理研究会 (2007-NL-186) (2008).

[7] 塩入寛之, 関根聡, 梅村恭司 拡張固有表現獲得の精度向上, 情報処理学会自然言語処理研究会 (2007-NL-180) (2007).

[8] 検索ログによる拡張固有表現辞書の整備 関根聡 and 鈴木久美, 言語処理学会第 13 回年次大会 (2007).

[9] 山田寛康 Shift-Reduce 法に基づく日本語固有表現抽出, 情報処理学会自然言語処理研究会 (NL-179-3) (2007).

[10] 山田寛康, 工藤拓, 松本裕治 Support Vector Machine を用いた日本語固有表現抽出, 情報処理学会論文誌, **43**, 1 (2004), 44–53.

[11] 新納浩幸, 関根聡 拡張固有表現タガーの作成とその問題点の考察, 言語処理学会第 12 回年次大会 (2006).

[12] 山梨正明 比喩と理解, 東京大学出版会 (1988).

[13] 浅原正幸, 松本裕治 日本語固有表現抽出におけるわかち書き問題の解決, 情報処理学会論文誌, **45**, 5 (2004).

[14] 中野桂吾, 平井有三 日本語固有表現抽出における文節情報の利用, 情報処理学会論文誌, **45**, 3 (2004).

[15] 渡辺一郎, 桝井文人, 福本淳一固有表現抽出ツール ＮＥｘＴの精緻化とユーザビリティの向上, 言語処理学会第 10 回年次大会 (2004).

表 2　拡張固有表現タグの頻度分布 1

拡張固有表現階層				毎日新聞		白書	
				表現数 (のべ)	表現数 (異なり)	表現数 (のべ)	表現数 (異なり)
名前	名前 その他			184	90	0	0
	人名			30979	11785	377	210
	神名			36	25	0	0
	組織名	組織名 その他		1696	778	246	147
		国際組織名		1351	226	1317	331
		公演組織名		359	204	0	0
		家系名		102	59	0	0
		民族名	民族名 その他	533	165	58	42
			国籍名	1072	113	124	21
		競技組織名	競技組織名 その他	2686	752	1	1
			プロ競技組織名	2108	225	0	0
			競技リーグ名	380	51	1	1
		法人名	法人名 その他	3078	1201	1189	409
			企業名	6448	1727	512	206
			企業グループ名	94	45	18	2
		政治的組織名	政治的組織名 その他	880	301	103	70
			政府組織名	5597	906	3629	462
			政党名	2968	186	52	41
			内閣名	124	31	10	8
			軍隊名	951	204	180	48
	地名	地名 その他		307	175	96	68
		温泉名		31	26	0	0
		ＧＰＥ	ＧＰＥ その他	759	71	158	54
			市区町村名	10010	2146	857	408
			郡名	212	126	23	13
			都道府県州名	6395	271	1531	128
			国名	14286	324	3965	253
		地域名	地域名 その他	6	6	32	20
			大陸地域名	1406	180	1096	167
			国内地域名	1342	426	1226	228
		地形名	地形名 その他	240	98	75	62
			山地名	0	0	0	0
			島名	339	127	78	44
			河川名	231	84	174	105
			湖沼名	64	32	59	21
			海洋名	312	98	174	49
			湾名	48	18	123	21
		天体名	天体名 その他	18	10	0	0
			恒星名	7	7	0	0
			惑星名	200	7	7	3
			星座名	4	3	0	0
		アドレス	アドレス その他	1	1	0	0
			郵便住所	971	831	7	7
			電話番号	680	554	3	1
			電子メイル	0	0	0	0
			ＵＲＬ	3	3	0	0
	施設名	施設名 その他		807	463	237	161
		施設部分名		952	439	153	88
		遺跡名	遺跡名 その他	41	26	8	7
			古墳名	26	13	2	1
		ＧＯＥ	ＧＯＥ その他	1415	725	255	198
			公共機関名	1131	505	338	169
			学校名	3174	1056	57	44
			研究機関名	162	113	340	120
			取引所名	159	40	9	5
			公園名	103	71	144	82
			競技施設名	513	225	1	1
			美術博物館名	238	144	5	5
			動植物園名	0	0	0	0
			遊園施設名	55	37	5	3
			劇場名	216	122	2	1
			神社寺名	345	191	2	2
			停車場名	24	18	5	5
			電車駅名	477	281	47	38
			空港名	276	51	176	71
			港名	92	52	66	52
		路線名	路線名 その他	19	13	12	11
			電車路線名	203	107	71	41
			道路名	184	99	105	67
			運河名	6	3	5	2
			航路名	1	1	0	0
			トンネル名	12	8	11	3
			橋名	41	27	33	15

表 3　拡張固有表現タグの頻度分布 2

拡張固有表現階層				毎日新聞		白書	
				表現数 (のべ)	表現数 (異なり)	表現数 (のべ)	表現数 (異なり)
名前	製品名	製品名 その他		2205	951	1320	466
		材料名		229	41	329	126
		衣服名		493	170	36	20
		貨幣名		4057	2070	0	0
		医薬品名		123	69	46	10
		武器名		1606	187	11	9
		株名		21	11	1	1
		賞名		548	286	18	15
		勲章名		0	0	4	4
		罪名		1449	386	0	0
		便名		15	12	0	0
		等級名		275	126	37	24
		キャラクター名		54	28	0	0
		識別番号		133	79	0	0
		乗り物名	乗り物名 その他	15	9	34	15
			車名	193	127	0	0
			列車名	151	45	6	6
			飛行機名	51	32	33	26
			宇宙船名	58	28	24	23
			船名	135	81	18	13
		食べ物名	食べ物名 その他	1308	483	631	148
			料理名	1166	461	54	39
		芸術名	芸術名 その他	74	50	9	7
			絵画名	79	48	0	0
			番組名	325	238	7	6
			映画名	402	308	2	2
			公演名	309	213	0	0
			音楽名	322	251	0	0
			文学名	844	623	0	0
		出版物名	出版物名 その他	1404	294	65	51
			新聞名	445	147	1	1
			雑誌名	188	123	30	4
		主義方式名	主義方式名 その他	1956	813	2209	1080
			文化名	57	28	10	2
			宗教名	1008	86	33	15
			学問名	572	286	327	110
			競技名	1799	302	27	10
			流派名	37	20	0	0
			運動名	0	0	59	21
			理論名	46	39	4	4
			政策計画名	342	216	873	535
		規則名	規則名 その他	125	69	165	125
			条約名	399	98	306	179
			法令名	757	268	1225	508
		称号名	称号名 その他	7723	20	7	4
			地位・職業名	29441	2971	2563	669
		言語名	言語名 その他	46	36	9	8
			国語名	273	40	31	15
		単位名	単位名 その他	67	18	14	12
			通貨名	588	24	97	12
	イベント名	イベント名 その他		1458	544	330	78
		催し物名	催し物名 その他	598	382	164	128
			例祭名	277	80	7	6
			競技会名	2242	903	0	0
			会議名	1589	536	533	288
		事故事件名	事故事件名 その他	743	269	97	61
			戦争名	391	74	62	26
		自然現象名	自然現象名 その他	34	13	180	61
			自然災害名	14	12	116	51
			地震名	1262	41	43	15
	自然物名	自然物名 その他		13	8	30	16
		元素名		198	34	226	62
		化合物名		350	126	533	137
		鉱物名		38	13	101	29
		生物名	生物名 その他	114	45	36	15
			真菌類名	14	8	8	4
			軟体動物 節足動物名	113	41	0	0
			昆虫類	96	28	29	11
			魚類	195	82	118	42
			両生類	41	13	0	0
			爬虫類	23	12	0	0
			鳥類	182	54	27	9
			哺乳類	759	117	44	20
			植物名	992	375	369	112
		生物部位名	生物部位名 その他	80	18	22	6
			動物部位名	2023	380	28	14
			植物部位名	140	41	6	6
	病気名	病気名 その他		0	0	5	5
		動物病気名		1230	365	355	91
	色名	色名 その他		104	58	0	0
		自然色名		244	26	0	0

表 4　拡張固有表現タグの頻度分布 3

拡張固有表現階層				毎日新聞		白書	
				表現数 (のべ)	表現数 (異なり)	表現数 (のべ)	表現数 (異なり)
時間表現	時間表現 その他			0	0	1	1
	時間	時間 その他		0	0	5	5
		時刻表現		3901	915	97	68
		日付表現		18426	2851	13919	2729
		曜日表現		469	39	34	14
		時代表現		797	179	171	40
	期間	期間 その他		123	42	21	21
		時刻期間		853	353	242	144
		日数期間		795	165	0	0
		週期間		177	30	30	9
		月期間		541	131	145	66
		年期間		3058	373	1012	234
数値表現	数値表現 その他			676	248	168	113
	金額表現			0	0	2721	2305
	株指標			0	0	2	2
	ポイント			1612	423	163	92
	割合表現			3002	995	9157	1688
	倍数表現			254	108	492	240
	頻度表現			197	46	180	95
	年齢			4321	437	909	209
	学齢			945	194	360	47
	序数			3526	828	1124	276
	順位表現			2061	178	238	64
	緯度経度			6	6	4	4
	寸法表現	寸法表現 その他		227	143	146	122
		長さ		1358	821	285	191
		面積		244	201	376	309
		体積		122	99	154	139
		重量		419	308	387	353
		速度		110	68	11	9
		密度		2	2	0	0
		温度		166	127	2	2
		カロリー		8	7	7	6
		震度		176	16	10	6
		マグニチュード		57	48	5	4
	個数	個数 その他		902	456	941	630
		人数		5699	1324	2267	1332
		組織数		1282	542	620	396
		場所数	場所数 その他	497	312	1043	598
			国数	177	73	258	106
			施設数	642	480	665	551
			製品数	2020	1221	438	364
			イベント数	2668	917	650	414
		自然物数	自然物数 その他	14	8	10	9
			動物数	144	106	59	47
			植物数	37	34	9	9

表 5　同一表現に付与されたタグの種類数の分布

毎日新聞		タグ数	白書コアデータ	
例	表現数		表現数	例
(省略)	58,745	1	22,888	(省略)
神戸製鋼，タンポポ，核廃棄物問題，川口	1,828	2	770	○○に関する法律，いわし，○○協議会
オウム，日立，バレンタイン，西武	175	3	85	臨時行政調査会，カンボディア問題，茶
さくら，富士，観音寺，5 回，茶	54	4	26	米国，九州，赤潮，銅，2 件，16
名古屋，千葉，三回，一件，95	25	5	7	名古屋，4，15，40，2 年
神戸，14，五つ，五十，1 つ	10	6	2	大阪，米
17，東京，広島	4	7	1	2
2 つ，二つ，一，15，大阪	8	8	2	1，東京
7，金	2	9	0	
4，ゼロ，5，三, 3	5	10	0	
一つ，6，10	3	11	0	
1	1	12	0	
2	1	13	0	

表 6　同一表現に付与されやすいタグとその表現例 (毎日新聞)

拡張固有表現タグ		表現数 (異なり)	例
競技組織名 その他	学校名	265	高松商，帝京，大東大，宇部商，PL 学園高
食べ物名 その他	植物名	176	ニガウリ，スズナ，キウイ，小豆，カブ菜
製品名 その他	イベント名 その他	161	パレスチナ問題，被差別部落問題，減税問題
競技組織名 その他	企業名	153	チチヤス乳業，いすゞ自動車，住友，エスビー
人名	市区町村名	72	池上，神田，千葉，大和，西宮
年齢	ポイント	56	30，34
イベント数	製品数	51	2 件，一発，一つ
イベント数	序数	49	7 回，二期，5 戦
製品数	個数 その他	48	一つ，三種類，二本
日付表現	日数期間	39	16 日
イベント数	数値表現 その他	38	一つ，百件，二本
食べ物名 その他	魚類	33	メバル，アマダイ，トラフグ，さけ
都道府県州名	市区町村名	32	青森，熊本，東京，京都
ポイント	イベント数	32	20
年齢	イベント数	31	19
日付表現	年数期間	30	三年，65 年，九〇年
駅名	市区町村名	30	東京，横浜
人数	イベント数	29	3
競技組織名 その他	人名	28	秋田，北，山城

表 7　同一表現に付与されやすいタグとその表現例 (白書)

拡張固有表現タグ		表現数 (異なり)	例
個数 その他	イベント数	76	2 回，51 件，3 次
日付表現	年期間	61	上半期，4 年，昭和 61 年度，90〜91 年，四半期
食べ物名 その他	植物名	42	二条大麦，そば，コーヒー，コシヒカリ
個数 その他	製品数	30	2 回，12 品目，2 条約，71 件，3 類型
場所数 その他	施設数	23	60，7 カ所，15 港
個数 その他	施設数	21	1 件，1，12 路線
都道府県州名	市区町村名	21	富山，新潟，山口，大阪，熊本
倍数表現	割合表現	20	48.2%，1，2.3 倍
会議名	国際組織名	19	欧州安全保障協力会議，UNCED
日付表現	日数期間	18	24 日
食べ物名 その他	魚類	16	いとより，さけ，くろまぐろ，さば
製品数	イベント数	14	2 回，71 件，4 事業，約 320 万件
組織数	個数 その他	14	1，3 個，3 事業，8 つ，3 者
序数	順位表現	14	3 番目，第 2，二次
港名	市区町村名	13	大阪，広島
空港名	市区町村名	13	羽田，名古屋，新潟
製品名 その他	主義方式名 その他	13	列車集中制御装置，関税，商品借款
空港名	都道府県州名	13	広島，鹿児島
イベント数	施設数	12	1，2
植物名	材料名	12	ゴム，なたね，い (い草)，コルク