

固有表現認識課題におけるアノテーションの視線分析

徳永 健伸^{1,a)} 西川 仁¹ 岩倉 友哉² 湯上 伸弘²

概要: コーパスに基づく自然言語処理では、与えられた課題に対してどのような素性を利用するかが重要となる。従来、主に内省によって決めた様々な素性が利用されてきたが、我々はコーパス作成時のアノテーションの振舞いから有効な素性の手掛りを得ることを目的とし、主にアノテーションの視線を中心にデータ収集をおこなってきた。本稿では、固有表現認識課題を対象とし、アノテーション中のアノテーションの視線を収集した結果とその予備的な分析結果について報告する。

1. 背景と目的

コーパスに基づく自然言語処理では、特定の課題を解くために、その課題の正解をアノテーションしたコーパスを構築し、それを訓練データとして機械学習の技術を使って課題を解くという手法で多くの成果をあげてきた。従来、専門家ができるだけ広範な言語現象を網羅することを考えながら人手で規則を記述していたのに対し、この手法では、必ずしも専門家ではない複数のアノテータがコーパスに情報をアノテーションし、アノテーションされた事例数を増やすことによって言語現象の網羅性を高めることを目指している。専門家の役割は、複数のアノテータが一貫性を維持しながら大規模なコーパスにアノテーションができるようなアノテーションのガイドラインを整備することになる。コーパスから規則性を抽出するのは専門家ではなく、機械学習によっておこなわれる。

機械学習に用いる素性としては、主にテキスト中の局所的、表層的な情報が用いられている。しかしながら、課題解決のために利用する情報を考慮する際、アノテータが実際に利用している情報をふまえて検討がおこなわれることはほとんどなかった。これはコーパスのアノテーションにおいてはアノテーションの結果のみが重視され、アノテータがアノテーションの過程において利用している情報にはほとんど関心が払われていなかったことを意味する。コーパスに基づく自然言語処理の手法をさらに洗練し、高性能化するためには、アノテータがアノテーション過程に参照する情報も積極的に利用することが考えられる。アノテ-

ション過程のアノテーションの振舞いに注目した研究として Tomanek らの研究がある [16]。Tomanek らは、能動学習のための訓練データの選択基準として使うことを前提として、個々の事例の難しさを推定するモデルを構築している。具体的な課題として固有表現認識を設定し、まず、アノテーションの視線情報から個々の事例にアノテーションする際のアノテーションの認知負荷を推定し、それを基にアノテーションが難しい事例の言語的特徴を推定し、それを素性として回帰モデルに取り入れることによりモデルの性能を改善できることを示している。しかしながら、視線情報は対象となる固有表現の上下左右という非常に粗い粒度でしか利用しておらず、得られた知見も対象の固有表現の意味的な複雑さが認知負荷に影響するという粗いものにとどまっている。

このような背景をふまえ、我々はコーパス中のアノテーション結果とともにアノテータがそのアノテーションをおこなう際の振舞いを同時に記録し、それを分析することによって、計算機にとっても有用な情報を得ることを目指している [5], [10], [15]。これまでは、主に述語項構造解析課題のアノテーション過程についてデータを収集してきたが、本稿では、Tomanek らと同様、テキスト中の固有表現を同定する固有表現認識 (Named Entity Recognition: NER) 課題を取りあげ、アノテータがテキスト中の指定された固有表現の種別を同定する過程における視線情報とマウスの操作情報を時間同期して記録したデータを収集し (2 節)、その分析結果 (3, 4 節) について報告する。Tomanek らの実験と異なり、我々は視線情報の空間解像度を単語レベルまで上げ、アノテータが対象固有表現の上下左右のどの領域を見ていたかだけでなく、どの語を見ていたかまで考慮し、より詳細な分析をおこなう。

¹ 東京工業大学
Tokyo Institute of Technology

² 富士通研究所
Fujitsu Laboratories Ltd.

^{a)} take@cs.titech.ac.jp

2. アノータタのデータ収集実験

アノテーション作業者が固有表現抽出の課題に取り組む際の視線とマウス操作の情報を収集するためにデータ収集実験を実施した。

2.1 課題データ

作業者に提示する課題として IREX^{*1} の Named Entity (NE) 抽出用の以下の開発データ^{*2} を利用した。

- NE_DRYRUN：毎日新聞 CDROM 94 年版の中に含まれる 36 記事
- NE_SAMPLE：毎日新聞 CDROM 94 年版の中に含まれる 46 記事
- CRL：毎日新聞 CDROM 95 年版の中に含まれる 1174 記事
- ARREST_TRAIN：毎日新聞 CDROM 94 年版の中に含まれる 23 記事

これらのデータには、IREX で定義された 8 種類のタグ、〈組織名〉、〈人名〉、〈地名〉、〈固有物名〉、〈日付表現〉、〈時間表現〉、〈金額表現〉、〈割合表現〉が人手で付与されている。

今回は計算機によってうまく処理できない事例を収集するために、まず、このデータに KNP 4.11^{*3} (形態素解析は JUMAN 7.0^{*4}) を適用し、固有表現抽出タグを付与した。KNP が正しく抽出できなかった固有表現のうち、正解が、〈人名〉、〈地名〉、〈組織名〉、〈固有物名〉の事例について人手で重複などを除き、最終的に 72 の固有表現を選択し、これらを作業者提示する課題とした。72 の固有表現はいずれも重複のないテキストに含まれている。72 の課題テキストの平均文字数は 315、最小は 78、最大は 498 である。

2.2 実験手続き

実験参加者は成人男性 6 名、成人女性 10 名の合計 16 名である。参加者には実験目的を説明した後、作業内容について説明した。作業はこの実験のために作成したツールを用い、画面上に表示されたテキストを読んで黄色でハイライトされた固有表現についてその種別を以下の 6 つから選択することである。

- 人名：人物あるいは擬人化されたものの名前
- 地名：土地の名前、山や川などの名前、住所など
- 組織名：企業、グループ、チームなどの組織の名前
- 人工物^{*5}：固有名詞のうち、製品、作品、サービスなど人によって作られた物
- その他：上記にあてはまらない物

^{*1} <http://nlp.cs.nyu.edu/irex/NE/>

^{*2} <http://nlp.cs.nyu.edu/irex/Package/IREXfinalB.tar.gz>

^{*3} <http://nlp.ist.i.kyoto-u.ac.jp/?KNP>

^{*4} <http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN>

^{*5} IREX の「固有物名」という表現はわかりにくいので実験説明の際には「人工物」という表現を用いた。

- 不明：どれを選んでよいかわからない場合

作業中の参加者の視線は視線計測装置 Tobii T60 によって 60Hz のサンプリング周波数で両目のスクリーン座標を計測した。Tobii T60 の画面サイズは 17 インチで画面解像度は 1,280 × 1,024 画素である。テキストのフォントは 24 × 24 画素の MS ゴシックフォントを用い、行間は 72 画素とした。また、上下左右に 96 画素のマージンを取ってテキストを配置した。規則処理はおこなわず、対象となる固有表現が 2 行にまたがることも許した。段落の開始は改行と空白で始まっている。

72 課題は 32 課題ずつの 2 つの課題セットに分割し、実験では最初の課題セットをおこなった後、休憩を入れ、残りの課題セットをおこなった。休憩時間は参加者が必要なだけ取れるようにした。参加者は作業内容の理解を確認し、操作に慣れるために課題を開始する前に 72 課題に含まれない 5 例について練習セッションをおこなった。各課題セットを開始する前に画面上の 5 点の参照点による視線計測の較正をおこなった。

各課題の提示前には画面中央に課題番号を含むマーカが表示され、このマーカをマウスでクリックすることにより、課題テキストが表示される。参加者がテキスト中のハイライトされた固有表現をマウスでクリックすると固有表現の種別一覧がポップアップメニューで提示されるので、そのなかから正しいと思う種別をマウスクリックによって選択する。これで一課題が終了し、次の課題番号を表示するマーカ画面に戻る。これを 36 回繰返すとひとつの課題セットが終了する。マウス操作に関してはマーカをクリックして課題が提示された時刻、対象固有表現をクリックした時刻、ポップアップメニューから固有表現種別を選択した時刻の 3 点を記録した。

3. データ分析：正解率・作業時間

3.1 視線計測のエラー率

Tobii は計測した視点ごとに左右の視点のスクリーン座標と計測に成功したかどうかの情報を出力する。両目の視点とも計測できなかった場合をエラーとみなし、各作業者のすべての課題の平均視線計測エラー率を計算した (表 1)。表 1 は、全 72 課題のうちエラー率がそれぞれ 0.1, 0.2, 0.3 を越えるものの課題数も示している。採用するデータに関するエラー率の絶対的な目安というものはないが、ここではエラー率の平均が 0.15 未満であり、エラー率が 0.1 を越える課題が全体の半分以上を越えない (36 未満) という条件で実験作業者のデータを選別した。したがって、以下の分析では、作業者 03, 05, 06, 07, 09, 16 のデータは廃棄し、残りの 10 名のデータを用いる。

3.2 正解率

表 2 に作業者ごとの正解率を、表 3 に各正解率の課題の

表 1 作業者の視線計測エラー率

作業者	平均 エラー率	エラー率別課題数		
		> 0.1	> 0.2	> 0.3
01	0.073	11	1	0
02	0.031	1	0	0
03*	0.13	39	8	3
04	0.094	12	5	5
05*	0.47	49	38	33
06*	0.43	62	57	49
07*	0.15	53	13	5
08	0.086	16	8	6
09*	1.00	72	72	72
10	0.084	22	2	0
11	0.094	8	5	5
12	0.04	2	2	1
13	0.10	24	3	1
14	0.072	11	6	5
15	0.052	7	3	3
16*	0.17	23	17	14

*が付いたものは破棄するデータ

表 2 作業者の正解率

作業者	01	02	04	08	10	11	12	13	14	15	平均
正解数	59	54	63	57	64	66	64	47	59	66	59.9
正解率	0.82	0.75	0.88	0.79	0.89	0.92	0.89	0.65	0.82	0.92	0.83

表 3 各課題の正解者数別頻度

正解者数	0	1	2	3	4	5	6	7	8	9	10
頻度	2	2	0	0	2	0	3	9	10	12	32

表 4 各作業者の平均作業時間 (msec)

作業者	T1	T2	T1+T2	正解率
01	4,862	2,090	6,952	0.819
02	13,115	2,998	16,114	0.750
04	7,753	1,383	9,136	0.875
08	8,198	3,373	11,571	0.792
10	11,197	3,037	14,234	0.889
11	5,759	1,734	7,493	0.917
12	6,525	2,749	9,273	0.889
13	4,257	3,105	7,362	0.653
14	9,965	2,439	12,404	0.819
15	3,580	2,067	5,647	0.917
平均	7,521	2,498	10,019	0.832

頻度を示す。正解率は作業者によってばらつきがあるが、平均で 8 割以上の正解率を示している。これらの課題がいずれも KNP が誤答したものであることを考えると、固有表現種別の判断に人間が利用している情報を固有表現認識処理にも活用できれば自動認識の性能をさらに改善できる可能性がある。

3.3 作業時間

表 4 に各作業者の平均作業時間を示す。ここで T1 は課

表 5 各課題にかかった平均作業時間の分布

区間 (msec)	T1	T1+T2	区間 (msec)	T2
(0, 4,000]	18	2	(0, 1,000]	0
(4,000, 8,000]	27	34	(1,000, 2,000]	30
(8,000, 12,000]	15	15	(2,000, 3,000]	19
(12,000, 16,000]	10	12	(3,000, 4,000]	15
(16,000, 20,000]	1	7	(4,000, 5,000]	6
(20,000, 24,000]	0	0	(5,000, 6,000]	0
(24,000, 28,000]	0	1	(6,000, 7,000]	1
(28,000, 32,000]	0	0	(7,000, 8,000]	0
(32,000, 36,000]	0	0	(8,000, 9,000]	1
(36,000, 40,000]	0	0		
(40,000, 44,000]	1	1		

表 6 課題の正解者数と平均作業時間の相関

	T1	T2	T1+T2
正解者数との相関係数	-0.47	-0.59	-0.54
テキスト長との相関係数	0.36	0.12	0.35

表 7 固有表現種別ごとの正解率

固有表現種別	人工物	地名	組織名	人名
作業者\課題数	17	15	29	11
01	13	13	23	10
02	14	9	23	8
04	13	13	27	10
08	14	13	23	7
10	17	15	23	9
11	17	12	27	10
12	14	13	27	10
13	6	12	20	9
14	15	14	21	9
15	16	13	27	10
平均正解率	0.82	0.85	0.83	0.84

題の開始から対象の固有表現をマウスでクリックするまで、T2 は対象がクリックされてからメニューの種別をクリックするまでの時間である。

表 5 に各課題にかかった平均作業時間の分布を示す。この表からほとんどの課題は 20 秒未満で完了していることがわかる。課題の正解者数と作業時間の相関を計算したところ表 6 に示すとおり、負の相関が見られた ($p < 0.00005$)。課題のテキスト長と相関については、固有表現の種別を判断する前の T1 とは弱い相関が認められた ($p = 0.002$)。これはテキストが長くなるとそれに応じて、より多くのテキストを読むためだと考えられる。これに対して固有表現の対象をクリックして、メニューから種別を選択する時間についてはテキスト長との相関はほとんどない。

3.4 固有表現種別ごとの正解率、作業時間

表 7 に今回付与した固有表現の種別ごとに作業者の正解数と平均正解率 (マクロ平均) を示す。正解率を見る限りでは、人工物はやや低いものの固有表現種別間での判定の難

表 8 固有表現種別ごとの作業時間

固有表現種別	T1	T2	T1+T2
人工物	7,428	2,843	10,271
地名	6,883	2,328	9,211
組織名	7,604	2,693	10,297
人名	8,316	2,859	11,175

表 9 固有表現種別の交差行列

正解\作業者	人名	地名	組織名	人工物	その他	不明
人名	92	2	6	1	8	1
地名	1	127	12	3	7	0
組織名	5	14	241	12	12	6
人工物	0	0	7	139	24	0

証券会社の代表的な貯蓄商品であるマネー・マネジスト・ファンド (MMF) の残高が三月に約三兆六千九百億円も激減し、三月末の残高は八兆九千三百億円に落ち込んだ。証券投資信託協会が十二日発表した。短期金利の低下で利回りが下がったため、決算期を控えた金融機関や事業会社が大量に解約したのが要因という。

MMFの平均利回りは、昨年九月の年率二・九〇七から今年三月末には二・二九九%へダウン。銀行の大口定期との金利差も縮小した。このため、三月には解約額が五兆二千九百億円にのぼり、新規設定の約一兆六千億円を差し引いた減少額は過去最高の三兆六千九百億円となった。ただ、四月に入ってから、MMFの残高は約一兆円ほど増加し回復傾向にあるという。

図 1 停留の例

しさに大きな差はないと考えられる。同様に固有表現種別間での平均作業時間には大きな差は見られなかった (表 8)。

表 9 は各固有表現種別間の交差行列である。組織名を地名や人工物に混同する誤りが多い傾向を示している。

4. データ分析：停留数・停留時間

4.1 視線データの前処理

Tobii で収集した視線データは 60Hz でサンプリングされた時刻付きの画面上の画素座標の系列である。「視線が語に留まっている間、その語に対する認知的処理がおこなわれている」という eye-mind 仮説 [8] に基づき、各視線座標を停留 (fixation) に変換する。停留の計算には、Dispersion-Threshold Identification (I-DT) アルゴリズム [14] を用いた。I-DT アルゴリズムでは、視線の速度が停留の近傍では遅くなるという観察に基づいて、特定の分散の範囲内で連続して出現する視線の集合を停留として同定する。I-DT アルゴリズムによって停留を計算するためには、視線座標間の最大距離 (空間閾値) と停留の最小時間間隔 (時間閾値) という 2 つのパラメタを決めなければならない。本稿では、(i) 画面サイズとその解像度、(ii) 作業者と画面との距離、(iii) 視線計測装置の解像度という 3 つの実験設定に基づいて、空間閾値を 24 画素、時間閾値を 100 msec に設定した [13]。

視線計測では一般に水平方向の計測誤差より垂直方向の計測誤差が大きくなることが知られており、視線計測誤差

表 10 テキスト中の停留の局所分布割合 (平均/標準偏差)

両側窓幅 異り/延べ	1 文節		2 文節	
	異り	延べ	異り	延べ
停留数	0.20/0.084	0.29/0.12	0.27/0.11	0.36/0.14
停留時間	0.20/0.084	0.31/0.13	0.27/0.11	0.37/0.15

表 11 テキスト中の停留の局所分布と正解率の相関

両側窓幅 異り/延べ	1 文節		2 文節	
	異り	延べ	異り	延べ
停留数	0.28	0.21	0.28	0.21
停留時間	0.28	0.22	0.28	0.22

を補正する手法がいくつか提案されている [1], [3], [9]。しかし、これらはいずれも対象とする実験課題の特徴に基づくヒューリスティクスを利用しており、必ずしも今回の固有表現認識課題に適しているとは限らない [2]。そこで、今回は行間を 3 文字分と大きめにとり、停留の中心の y 座標がテキストの行間にある場合は、その停留の y 座標をより近い行の中心線上に補正することによって垂直方向の計測誤差を補正した。図 1 にテキスト上に補正後の停留を重畳表示した例を示す。円の半径は停留時間に対応し、円の中心が停留に含まれる視線座標 (左右視線座標の平均) の重心に対応する。

4.2 停留の分布

固有表現の自動認識では、対象となる表現の前後二語程度の局所文脈の特徴を主な素性として、固有表現の種別を同定することが多い。たとえば、英語 [12]、アラビア語 [4]、日本語 [7] において、近年、高い精度が報告されている手法においても、対象単語および前後二単語の単語表層文字列や品詞、それらの単語と辞書との照合結果などが主な素性として使われている。作業者の停留が実際に対象表現の両側の局所的な領域に集中しているかどうかを調査した。まず、各停留とテキストとの対応を明らかにするために、テキストを CaboCha*⁶ によって文節に区切り、その文節のパウンディングボックス内に補正した停留の中心がある場合に、その文節に視線が停留しているものとした。CaboCha で文節に分割した際の平均文節長は 4.7 文字であった。日本語読解における周辺視野は 5 文字から 7 文字とされていること [6], [11]、日本語では文節が述語に対する文法役割を持つ単位であることから、文節を停留の対象とするのは妥当であると考えられる。表 10 は、対象固有表現も含め、その両側 1 文節あるいは 2 文節上にある停留数/停留時間の割合の平均と標準偏差を示している。この表から異りで 8 割以上の停留は対象固有表現の前後 1 文節以外の文節にあることがわかる。範囲を 2 文節に広げても局所文脈上の停留は 3 割程度である。つまり、人間は広範な文脈を参照して固有表現の種別を決めているということが示唆さ

*⁶ <http://taku910.github.io/cabocha/>

表 12 正解率と停留数/停留時間の相関

停留\正解率	作業者	課題
停留文節の異り総数	-0.061	-0.46
停留文節の延べ総数	-0.21	-0.51
総停留時間	-0.28	-0.53
平均停留時間	0.30	0.21

れる。

表 11 は各固有表現の正解率、つまり、何名の作業者がその固有表現の種別を正しく同定できたかと局所文脈にある停留の割合の相関を示している。この表から両者の間に強い相関はなく、広範な領域を見ているから正解率が高いというわけではなく、参照すべき領域は事例に応じて異なるということが示唆される。この点については Tomanek らの分析でも言及されている。

4.3 正解率と停留数/停留時間

文節への停留数/停留時間と作業者/課題の正解率の関係を調査するために、両者の相関係数を計算した。その結果を表 12 に示す。作業者の正解率についてはサンプル数が少ないので p 値がいずれも高く、信頼性のある値は得られなかったが、課題の正解率については、停留文節の総数、総停留時間とは負の相関が見られる。

4.4 正解者と不正解者の停留傾向

我々は停留した文節まで同定しているので、文節レベルでの停留の分析が可能である。以下では、課題に正解した作業者と不正解の作業者の視線が停留した文節にどのような差があるかを分析する。個別の作業者間の比較は組合せの数が多くなりすぎることと、個人差の問題があるので、今回は各課題について正解グループと不正解グループに二分し、その間の停留の差を調査した。ただし、表 2 からわかるとおり、72 課題中 44 課題については 9 名以上の作業者が正解しているため、以下の分析では最低でも一方のグループに 3 名が含まれるような 14 課題 (表 2 中で正解者数が 4, 6, 7 の課題) を分析対象とする。これら 14 課題のテキストを付録に示す。

両グループの停留傾向の差を見るために、各グループについて各文節に対する停留回数と総停留時間をグループ内人数で正規化した値を計算し、正解グループの値から不正解グループの値を引いた差が 0 より大きな文節に注目し、分析をおこなった。

14 課題中 9 課題 (940413095, 940413096, 940911116, 940911122, 950103080, 950105198, 950105203, 950105208, 950106257) について、正解グループの視線が対象固有表現が項となっている述語やその述語の他の項に停留している傾向が観察できた。たとえば、課題 950103080 では、対象固有表現の「東」が「東」は/.../こぼれ球を/けり込

み/.../決勝点を/奪った/」という文中の主語となっており、正解グループではこの文の項や述語である「こぼれ球を」、「けり込み」、「決勝点を」、「奪った」などに視線がより多くの停留をしている。同様に、課題 950105203 でも「ガラス工場の中で/... モーレロース」が/.../名を/知られている/」という文中において対象固有表現の「モーレロース」は主語であり、その述語「知られている」や補語「ガラス工場の中で」などの文節への停留が多い。課題 940413096 はもう少し複雑な例である。正確グループでは対象固有表現の「MMF」という頭字語の正式名「マナー・マネジメント・ファンド」に視線が停留しており、また、その「残高」が主語となっている述語の文節「落ち込んだ。」や「MMF」がゼロ代名詞となってヲ格項となっている述語「解約したのが」などの文節にも停留する傾向がある。課題 940911116 では、対象固有表現「ペンスキー・イルモア」はドライバーの「アル・アンサー・ジュニア」の所属するチーム名であるが、正解グループでは、「アル・アンサー・ジュニア」やそれが主語となっている述語「狙う」への停留が多く、まず「アル・アンサー・ジュニア」の種別を同定しようとしていると推測できる。

一方、局所的な文脈が正解/不正解の原因になる例もあった。課題 940911057 では、対象固有表現が宛先となっており、直後に「～係」という宛先を示す手掛りがあるために、不正解グループはこれに引きずられ、誤った判断をしてしまったと考えられる。正解グループは前半の文脈にも出現する対象固有表現「フラパン・ラブレー・バカラ」を含む文節とその前後にも停留が多く、文脈から「フラパン・ラブレー・バカラ」の種別を判断しようとしているのに対して、不正解グループの停留は宛先以降の文節に集中している。課題 950103056 は「山梨」という表現が地名としての馴染が深いためか、不正解グループでは「山梨」という表現のみを見て地名だと判断し、誤っている。不正解グループの視線がほとんど対象固有表現の「山梨」の周辺にしか停留していないのに対して、正解グループの視線はテキスト広範に分布しており、文脈を読みとろうとしていることがわかる。課題 950109041 では、逆に正解グループの視線はほとんど対象固有表現の「ドゴール」周辺にしか停留していないのに対して、不正解グループの視線はより広範囲に停留している。これは「ドゴール」がフランスの元大統領であるという知識があれば簡単にわかるが、文脈からは「ドゴール」が人名であることを推察することが難しい例だと考えられる。

以上の例から正解グループは必要に応じてより広い文脈を参照していることがわかる。この点は Tomanek らの研究でも指摘されていたが、我々の実験により具体的にどのような情報を参照するのかがより具体的に明らかになった。

5. おわりに

本稿では、固有表現認識課題における人間の振舞いと正解率の関係について分析をおこなった。特に文節単位への作業者の視線の停留の分析を通して、正しく固有表現の種別を同定するためには、対象固有表現の周辺文脈だけではなく、必要に応じてより広い文脈を参照する必要があることを確認した。特に対象固有表現が項となる述語やその述語の他の項が固有表現の種別の同定に有用であることがわかった。そのためには場合によってはゼロ代名詞を含む照応解析も必要となってくる。

今後の課題としては、どのような状況でより広い文脈を参照情報として用いるかを明かにし、実際の固有表現認識の自動処理に取り込む手法を検討することである。また、今回の分析では時間情報は捨象し、停留回数、停留時間の積分值のみを用いたが、今後、スキャンパスなどの停留の移動に関する情報を用いてさらに詳細な分析をすることも必要である。

付録：4 節の分析で用いた課題

テキスト中で四角で囲まれた固有表現が種別を判定する対象である。回答分布は各種別を選択した作業者数を示す。下線を引いたものが正解の種別である。記事は日外アソシエーツから販売されている毎日新聞 CD-ROM1994 年版および 1995 年版による*7。

[940413095]

回答分布：〈地名〉：4，〈組織名〉：2，〈人工物〉：2，〈その他〉：2

バンコクのオフィス街で、ショッピング街でもある、シーロム通りに通称「ラライサブ」と呼ばれる市場がある。昼休みになると、若い女性で通勤ラッシュ並みの混雑になる。「ラライサブ」とは、タイ語でおカネが溶けてしまうという意味だ。

国産のスーツ着七百一千パーツ（約三千—四千三百円）、T シャツ五十パーツ、ブラジャー百パーツ、ショーツ二十パーツ。ブランド商品はこの二倍程度。バンコクっ子は市場で食料品、日用品、衣類まで調達してしまう。価格はデパートの半値から三割引き程度。デパートで最新流行のブランドを下調べしてから市場で探すのが買い物のコツという。

このタイで、ブラジャーとショーツを年間二千三百五十万枚生産しているのがワコールの現地法人タイワコール。タイでの下着の市場占有率六割を誇り、バンコクっ子ならだれでも知っている人気ブランドだ。昨年暮れ、日本でもヒットした、軽いワイヤで胸の位置を上げる「グッド・アップ・ブラ」を売り出した。二百二十パーツ、日本円にして千円弱。同タイプの商品が日本では四千円程度する。

[940413096]

回答分布：〈組織名〉：2，〈人工物〉：7，〈その他〉：1

証券会社の代表的な貯蓄型商品であるマネー・マネジメント・ファンド（MMF）の残高が三月に約三兆六千九百億円も激減し、三月末の残高は八兆九千三百億円に落ち込んだ。証券投資信託協会が十二日発表した。短期金利の低下で利回りが下がったため、決算期を控えた金融機関や事業会社が大量に解約したのが要因という。

MMF の平均利回りは、昨年九月の年率二・九〇七％から今年三月末には二・二九九％へダウン。銀行の大口定期との金利差も縮小した。このため、三月には解約額が五兆二千九百億円にのぼり、新規設定の約一兆六千億円を差し引いた減少額は過去最高の三兆六千九百億円となった。ただ、四月に入ってから、MMF の残高は約一兆円ほど増加し回復傾向にあるという。

[940911057]

回答分布：〈組織名〉：2，〈人工物〉：7，〈その他〉：1

サッポロビールは、コニャックの名門・仏フラパン社の最高級コニャック「フラパン・ラブレー・バカラ」＝写真＝45本を1本50万円で限定発売する。作家フランソワ・ラブレーの生誕500年を記念し、フラパン社が発売する500本の一部。平均熟成年数は50—60年。容器には24金で孔雀（くじゃく）などを装飾したバカラ社特製クリスタル・デカンタを使用。アルコール分40％で、700ミリリットル入り。予約受け付けは26日から来月31日まで、はがきで「〒150東京都渋谷区恵比寿4の20の1サッポロビール（株）ワイン洋酒事業部『フラパン ラブレー・バカラ』係」。申し込みが限定数量を上回った場合は抽選する。

[940911116]

回答分布：〈人名〉：2，〈組織名〉：6，〈人工物〉：1，〈不明〉：1

自動車レースのインディカー・シリーズ第14戦は9日、エルクハートレーク（米ウィスコンシン州）で第1回公式予選を行い、今季9勝目を狙うアル・アンサー・ジュニア（米国＝ペンスキー・イルモア）が平均時速134.948マイル（約217.131キロ）で暫定ポールポジション（PP）を獲得した。（UPI S時事）

[940911122]

回答分布：〈地名〉：7，〈その他〉：3

中嶋さんの「本職」は、修道士。社会福祉法人「聖フランシスコ会」に属し、山里相談室は、その一環として活動している。以前は三十年間、ソーシャルワーカーを務めたが、天職ではないという思いをぬぐえなかった。十年前、修道士になってから山谷に足を踏み入れて、自分の居場所を、ここに見いだした。助けを必要としている人々に寄り添い、ともにいること。それは、生きがいを探し続けてきた自分のためでもあった。

[950103056]

回答分布：〈地名〉：3，〈組織名〉：7

「まさか」の敗戦を喫した山梨学院大・上田誠仁監督は、予想を上回る早大の走りに脱帽するしかなかった。「区間記録を3人が連発するとは」。早大の主力3人の爆発力が、「本命」の走りを狂わせた。起爆剤となったのは、二区の渡辺。三区では主

*7 <http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html>

将、小林正が軽快なピッチ走法で一年生の中馬を追い、雪の富士山を正面に見る坂で一気に抜き去った。「渡辺さんの走りを見て興奮。オーバーペース気味になった」と言う四区・小林雅もそのまま押し切り、差を1分39秒に広げた。五区で山梨の主将、小椋がいったん40秒差にまで迫ったが、オーバーペース。ゴールでの差は1分47秒にまで開いていた。前回は「本命」と言われながら完敗。花田らエース級3人が卒業して迎えた今年は、**山梨**との立場が全く逆転した。

[950103080]

回答分布：〈人名〉: 7, 〈その他〉: 3

後半30分に清水市商・佐藤と一対一となるピンチを思い切りのよい飛び出しで防ぐなど、スーパーセーブを連発した。攻撃は、左サイドで一年生の中谷が、巧みなドリブルとスルーパスで攻撃のリズムを作る。後半18分に中谷の足がつると、同じ一年生で「一発がある」東が交代出場。**東**は34分、CKからGKがパンチしたこぼれ球をけり込み、最高のタイミングで決勝点を奪った。「やれると信じていた。三年生を中心にコツコツ我慢し、怖いもの知らずの一年生の結果を出した」。上間監督の勝因分析は冷静だった。「ウチにとっては最悪、相手にとっては最高の試合でしょう。勝負は結局頑張ったほうが勝つ」と大滝監督。清水市商におごりは、なかっただろうが、初戦にすべてをかけた「挑戦者」が、気持ちの面で、優勝を意識した「本命」を完全に上回った。

[950105198]

回答分布：〈組織名〉: 1, 〈人工物〉: 4, 〈その他〉: 5

テーマは“花、華やかに”。12カ月の花言葉を主題に、古式ゆかしい伝統ある結婚式から新しい感覚のウエディングやオリジナリティー豊かな披露宴、レストラン・ウエディング、2次会パーティーまで、婚礼に関するすべてを多彩に披露する。アンケートによりプレゼントが当たるお楽しみ抽選会も。電話0473・55・5000・宴会事務所。弦楽四重奏演奏、ウエディングトーン演奏やメルヘンビデオの上映なども。「披露宴料理コーナー」特別プラン“光輝”メニューを展示、「衣装&フィッティングルーム」をはじめ婚礼に関する各種展示コーナーで専任スタッフがすべての相談に応じてくれる。電話03・3541・2411・宴会オフィス。料理を試食しながらプロの司会者による披露宴を体験。「紫苑」1人1万3000円、「**瑠璃**」1人1万1000円、先着100人。相談および展示会。当日予約および成約の方には、食事招待がつく。電話03・3447・3111・宴会承り係。いずれも4階宴会場「藤波」。豊富にそろえた華やかな婚礼衣装をゆとりあるスペースで試着できる。婚礼に関するすべての展示と相談会。電話043・245・6711・宴会予約課。

[950105203]

回答分布：〈組織名〉: 7, 〈人工物〉: 2, 〈その他〉: 1

スウェーデンの首都ストックホルムから南西部に位置するスモーランド地方は別名「ガラスの王国」とも呼ばれている。この地方にある二つの大きな町カルマルからベクショーにかけて、十六ものガラス工場が点在しているからだ。それぞれが独立し

た“村”を形づくっており、おとぎの国に出てくるような住宅街の中心にレンガ造りのガラス工場と、直売店がある。この地方でガラスの製造が盛んになったのは、十六世紀の中ごろ、グスタヴ・ヴァーサ王が産業振興の目的で、イタリアのベネチアからガラス職人を招いたことに始まる。古くからあった鉄工の技術と、火を起こすための木材となる広大な森林を生かして、ガラス生産の中心になった。数多いガラス工場の中で、「オレフォス」「コスタ」「**モーレロース**」といったところが特に名を知られている。

[950105208]

回答分布：〈人工物〉: 7, 〈その他〉: 3

日本の女優の草分け松井須磨子が1919年のこの日愛人の島村抱月の後を追って自殺した。長野県の生まれ。本名小林正子。1909年文芸協会演劇研究所の第1期生になり、11年の帝劇の「ハムレット」でオフィーリアを演じて注目された。続いてイブセンの「人形の家」のノラを演じ、一躍スターになった。この後抱月との恋愛関係が理由で協会から追われたが、抱月と共に芸術座を創立、「モンナ・ヴァンナ」「海の夫人」「**復活**」「サロメ」などに次々と主演し人気を集めた。とくに「復活」の劇中歌として歌った「カチューシャかわいいや」は全国で愛唱された。野生味のある情熱的な芸に特色があり、一部に反発はあったものの一時代を築いた女優であった。抱月が18年秋にスペインかぜで急死、その2カ月後「カルメン」の公演中に後を追った。

[950106125]

回答分布：〈組織名〉: 7, 〈人工物〉: 2, 〈不明〉: 1

トレンディードラマには欠かせない別所クン。洗剤のCMではちょっと軟弱な姿を見せているけれど、映画デビュー作では得意の語学力を生かした日系人役を演じた経歴を持つなど、器用な人なのだ。もちろん歌でも甘くソフトな雰囲気ファンを増やそう。今月スタートの金ドラでは初の父親役も。**ハマラジ**「シネマチック・ラブ」

[950106257]

回答分布：〈人工物〉: 7, 〈その他〉: 3

作家の故梁雅子さんの戦争体験を書いた自伝的小説「われ餓えたる民」が、執筆から約三十年経た今年、初めて出版される。ちょうどの区切りに、戦場に行かなかった女性や子供の味わった、食糧難や物資不足という戦場外での戦いを改めて知る一冊として関心呼びそう。梁さんは、老人ホームを舞台に老人の性をとらえた小説「悲田院」で六〇年に女流文学賞を受賞、作家としてデビューした。「**われ——**」は、の身の回りの暮らしをまとめた原稿用紙に約八百枚の作品で、六五年ごろ一年半かかって仕上げた。主人公は梁さんをモデルにした軍医の妻で、四四年五月二十一日の大阪・天王寺から始まり、四五年九月末の活気あふれる大阪のヤミ市で終わる。季刊誌「大阪春秋」で七四年から連載したものの、同誌の廃刊に伴い、すべてを収容しないまま打ち切り。原稿は大阪府吹田市の自宅の書庫で眠ることになった。昨秋、梁さんの長男でテレビプロデューサーの典雄さんが、知人の作家に雑談でこの小説の話をしたところ、

「来年は、本にするにはタイミングがいい」と出版が決まった。

[950108064]

回答分布：〈人工物〉：6，〈その他〉：4

一九四五年に執筆を始めた形而上小説『死霊』が五十年を経てもまだ完結に至らず、“不可能性の作家”とも呼ばれながら、熱狂的なファンを持つ作家、埴谷雄高氏が本格的な番組としては初めてテレビに登場、九日夜八時から連続五夜、NHK教育テレビのE T V特集で「『死霊』の世界」を語る。埴谷氏は「足腰は弱ってしまったが、幸い口だけは元気です。僕は死ぬまでしゃべり続けるつもりです。『死霊』は無限といわれている宇宙の中では今の宇宙と違うものがあってもいいじゃないかと思って書いている小説です。こんな妄想の文学の作家がしゃべった番組なんか見る人はいませんよ」と笑い飛ばしている。

[950109041]

回答分布：〈人名〉：7，〈組織名〉：1，〈その他〉：2

四月二十三日に第一回投票、五月七日に第二回投票が行われる、仏大統領選挙をめぐる立候補者の最終的な絞り込み作業が、年明けとともに白熱してきた。保守、中道が乱立傾向を示す一方、ドロール欧州連合委員長の不出馬で、意気消沈する左翼陣営は土壇場の対応を迫られている。同大統領選に今のところ、正式な出馬表明をしているのは、**ドゴール**主義の継承を掲げる保守の共和国連合のシラク元首相ら四人。各種世論調査でリードを続けたドロール氏の不出馬表明後、独走態勢にある保守・バラデュール首相は今年二十日前後に、沈黙を破り正式出馬表明を行う可能性が強くなった。先月下旬のフランス航空機乗っとり事件の、手際の悪い処理で人気を高め、八日までに、ベルナル・ドブレ経済協力相ら十三閣僚がバラデュール氏支持を表明した。社会党は二月上旬までに候補者選出の予定だが、有力候補に次々と辞退され人選作業は難航している。この間げきを縫い、反欧州統合を唱える保守「もう一つの欧州」のドビリエ党首、中道のバール元首相などが独自の立場から出馬をうかがっている。

参考文献

- [1] Carl, M.: Dynamic programming for re-mapping noisy fixations in translation tasks, *Journal of Eye Movement Research*, Vol. 6, No. 5, pp. 1–11 (2013).
- [2] Carl, M., Jakobse, A. L. and Spakov, O.: Towards an annotation standard for eye tracking data, *Proceedings of Measuring Behavior*, p. 223 (2008).
- [3] Cohen, A. L.: Software for the automatic correction of recorded eye fixation locations in reading experiments, *Behavior Research Methods*, Vol. 45, No. 3, pp. 679–683 (2013).
- [4] Darwish, K.: Named entity recognition using cross-lingual resources: Arabic as an example, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pp. 1558–1567 (2013).
- [5] Iida, R., Mitsuda, K. and Tokunaga, T.: Investigation of annotator's behaviour using eye-tracking data, *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 214–222 (online), available from (<http://www.aclweb.org/anthology/W13-2326>) (2013).
- [6] Ikeda, M. and Saida, S.: Span of recognition in reading, *Vision Research*, Vol. 18, No. 1, pp. 83–88 (online), DOI: 10.1016/0042-6989(78)90080-9 (1978).
- [7] Iwakura, T.: A named entity recognition method using rules acquired from unlabeled data, *Recent Advances in Natural Language Processing, (RANLP 2011)*, pp. 170–177 (2011).
- [8] Just, M. A. and Carpenter, P. A.: A Theory of Reading: From Eye Fixations to Comprehension, *Psychological Review*, Vol. 87, No. 4, pp. 329–354 (1980).
- [9] Mishra, A., Carl, M. and Bhattacharya, P.: A heuristic-based approach for systematic error correction of gaze data for reading, *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*, pp. 71–80 (2012).
- [10] Mitsuda, K., Iida, R. and Tokunaga, T.: Detecting missing annotation disagreement using eye gaze information, *Proceedings of the 11th Workshop on Asian Language Resources*, pp. 19–26 (2013).
- [11] Osaka, N.: Size of saccade and fixation duration of eye movements during reading: Psychophysics of Japanese text processing, *Journal of Optical Society of America*, Vol. 9, No. 1, pp. 5–13 (1992).
- [12] Passos, A., Kumar, V. and McCallum, A.: Lexicon infused phrase embeddings for named entity resolution, *Proceedings of the 18th Conference on Computational Natural Language Learning (CoNLL 2014)*, pp. 78–86 (2014).
- [13] Richardson, D. C., Dale, R. and Spivey, M. J.: Eye movements in language and cognition: A brief introduction, *Methods in Cognitive Linguistics* (Gonzalez-Marquez, M., Mittelberg, I., Coulson, S. and Spivey, M. J., eds.), John Benjamins., pp. 323–344 (2007).
- [14] Salvucci, D. D. and Goldberg, J. H.: Identifying fixations and saccades in eye-tracking protocols, *Proceedings of the 2000 symposium on Eye tracking research & applications (ETRA '00)*, pp. 71–78 (online), DOI: 10.1145/355017.355028 (2000).
- [15] Tokunaga, T., Iida, R. and Mitsuda, K.: Annotation for annotation – Toward eliciting implicit linguistic knowledge through annotation –, *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, pp. 79–83 (2013).
- [16] Tomanek, K., Hahn, U., Lohmann, S. and Ziegler, J.: A Cognitive cost model of annotations based on eye-tracking data, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)*, pp. 1158–1167 (online), available from (<http://www.aclweb.org/anthology/P10-1118>) (2010).