読み時間と情報構造について（ちょっとみじかめ）

浅原 正幸 (人間文化研究機構 国立国語研究所)　　masayu-a@ninjal.ac.jp

1　はじめに

情報構造は和文翻訳時の冠詞選択や文書要約において重要であるが、言語処理の分野ではあまり研究されてこなかった。統語・意味情報処理が完全に行われたとしても、テキストの表層のみから機械学習に基づき情報構造の推定を行うことは困難であろう。一方、人間の文処理機構において、情報構造が文処理速度の促進・阻害の双方に影響を与えられることが考えられる。本稿では、**人間の読み時間に基づく情報構造の推定手法を確立するため**に、情報構造が人間の文処理時間のどのように影響を与えるかについて分析する。

日本人母語話者の読み時間データとして『現代日本語書き言葉均衡コーパス』(以下 BCCWJ)[3] に対して読み時間を付与した BCCWJ-EyeTrack [2] を用いる。同データに対して、情報構造アノテーション [4] を重ね合わせたうえで、線形混合モデルに基づく統計分析を行う。

結果、いくつかの情報構造について、読み時間の測定値に差が出ることを発見したので報告する。

表 1: Data format

列名	データ型	摘要
surface	factor	出現書字形
time	int	読み時間
logtime	num	読み時間 (常用対数)
measure	factor	読み時間の種類
sample	factor	サンプル名
article	factor	記事情報
metadata_orig	factor	文書構造タグ
metadata	factor	メタデータ
length	int	文字数
space	factor	文節境界空白の有無
subj	factor	実験協力者 ID
setorder	factor	文節境界空白の呈示順
dependent	int	係り受け関係
sessionN	int	セッション順
articleN	int	記事呈示順
screenN	int	画面呈示順
lineN	int	行呈示順
segmentN	int	文節呈示順
is_first	factor	最左要素
is_last	factor	最右要素
is_second_last	factor	右から 2 つ目の要素
infostatus	factor	情報状態
definite	factor	定性
specificity	factor	特定性
animacy	factor	有生性
sentience	factor	有情性
agentivity	factor	動作主性
commonness	factor	共有

2　分析手法

分析データとして BCCWJ-EyeTrack [2] に情報構造アノテーション [4] を重ね合わせたもの (表 1) を用いる。以下、データの詳細について説明する。

2.1　読み時間データ

自己ペース読文法は、他の文節をマスクしたうえで 1 文節単位に逐次的に呈示する読み時間測定手法である。読み戻しができないために、文節単位の読み時間がそのままデータとなる (SELF)。視線走査法で取得したオリジナルのデータから文字の半角単位に Start Fixation Time（注視開始時刻）と End Fixation Time（注視終了時刻）と Fixation Time （注視時間）を得る。このデータを国語研文節単位でグループ化しなおした注視順データを集計して、テキスト生起順データに加工する。テキスト生起順データは以下の 5 種類からなる。

- First Fixation Time (FFT)
- First-Pass Time (FPT)
- Regression Path Time (RPT)
- Second-Pass Time (SPT)
- Total Time (TOTAL)

これらの読み時間情報 (time, logtime) に対して、出現書字形 (surface)・記事情報 (sample, article)・文書構造 (metadata_orig, metadata) のほか、出現書字形文字数 (length)、文節単位の空白の有無 (space)、実験協力者 ID (subj)、係る文節数 (dependent)、実験協力者ごとの呈示順序 (sessionN, setorder, articleN, screenN, lineN, segmentN) 、画面水平方向の位置 (is_first,is_last,is_second_first) を付与したデータを分析に用いる。係る文節数は BCCWJ-DepPara [1] のものを用いる。

本研究では日本語母語話者 24 人分のデータを統計分析に用いる。データの詳細については [2] を参照さ

表 2: Parameters of the linear mixed model for the self paced reading time (SELF) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.893	0.062	46.51
length	0.102	0.002	**42.31**
space=TRUE	0.003	0.004	0.86
dependent	-0.005	0.003	-1.61
sessionN	-0.021	0.022	-0.94
articleN	-0.023	0.007	**-3.23**
screenN	-0.032	0.002	**-11.19**
lineN	-0.014	0.002	**-6.10**
segmentN	-0.005	0.001	**-4.83**
is_first=TRUE	0.047	0.006	**7.19**
is_last=TRUE	0.040	0.008	**4.71**
is_second_last=TRUE	-0.011	0.005	**-2.11**
space=TRUE:sessionN	-0.019	0.044	-0.43
infostatus=discourse-old	-0.005	0.005	-0.98
definite=indefinite	0.004	0.015	0.30
specificity=specific	0.044	0.016	**2.78**
specificity=unspecific	0.001	0.010	0.16
animacy=inanimate	-0.000	0.050	-0.02
sentience=insentient	-0.105	0.067	-1.56
sentience=sentient	-0.098	0.050	-1.94
agentivity=both	-0.058	0.049	-1.18
agentivity=neither	-0.004	0.007	-0.69
agentivity=patient	-0.013	0.008	-1.63
commonness=hearer-new	0.025	0.007	**3.59**
commonness=hearer-old	-0.020	0.009	**-2.11**
commonness=neither	0.000	0.025	0.01

表 3: Parameters of the linear mixed model for the first pass time (FPT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.303	0.102	22.53
length	0.144	0.004	**33.61**
space=TRUE	-0.032	0.007	**-4.23**
dependent	-0.005	0.006	-0.89
sessionN	-0.041	0.028	-1.46
articleN	-0.001	0.009	-0.19
screenN	-0.023	0.005	**-4.76**
lineN	-0.018	0.004	**-4.64**
segmentN	-0.008	0.002	**-4.07**
is_first=TRUE	0.068	0.011	**5.94**
is_last=TRUE	0.021	0.015	1.40
is_second_last=TRUE	0.028	0.010	**2.84**
space=TRUE:sessionN	0.062	0.056	1.11
infostatus=discourse-old	0.005	0.010	0.50
definite=indefinite	0.024	0.026	0.90
specificity=specific	0.064	0.028	**2.26**
specificity=unspecific	0.031	0.018	1.70
animacy=inanimate	0.210	0.104	**2.01**
sentience=insentient	-0.001	0.129	-0.01
sentience=sentient	0.194	0.086	**2.25**
agentivity=both	-0.050	0.087	-0.57
agentivity=neither	0.014	0.012	1.19
agentivity=patient	-0.006	0.015	-0.43
commonness=hearer-new	0.024	0.012	1.95
commonness=hearer-old	0.000	0.017	-0.03
commonness=neither	0.002	0.043	0.05

れたい。

2.2 情報構造アノテーション

情報構造は、BCCWJ の短単位について以下の情報を付与したものを用いる。基準の詳細については [4] を参照されたい。

- 情報状態 (information status:speaker-new)
  談話中に同一指示名詞句が出現した (discourse-old) か否 (discourse-new) か。既存の共参照情報ラベルを見ながら判定する。想定可能 (bridging) は言語受容者側の判断として、共有性に委ねる。
- 定性 (definiteness:hearer-identify)
  言語受容者が外延の示す実体を認識できる (definite) か否 (indefinite) か。
- 特定性 (specificity:speaker-identify)
  言語生産者が外延の示す実体を認識できる (specific) か否 (inspecific) か。
- 有生性 (animacy)
  名詞句が指示しているものが生きているか (animate) か否 (inanimate) かを述語を見ないで判定する。
- 有情性 (sentience)
  名詞句が指示しているものが自由意志を持つ (sentient) か否 (insentient) かを述語-項の対を見て判定する。
- 動作主性 (agentivity)
  節レベルで動作主 (agent)・被動作主 (patient) になるかを判定する。従属節側と主節側の両方で検討するため、どちらも可 (both) も許す。
- 共有性 (commonness:hearer-new)
  共有性は、言語需要者側が既知であると、言語生産者が想定している (hearer-old) か否 (hearer-new) かを判定する。談話上、世界知識を利用して想定可能 (bridging) である場合を許す。

読み時間の分析が文節単位であるために、文節内の最右要素の情報構造を文節を代表する情報構造として用いることとした。

2.3 統計処理手法

まず、対象は情報構造が付与されている名詞句のみとする。データの前処理として、metadata が {authorsData, caption, listItem, profile, titleBlock} のものを除外した。さらに視線走査実

表 4: Parameters of the linear mixed model for the regression path time (RPT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.188	0.118	18.48
length	0.120	0.004	**24.79**
space=TRUE	-0.021	0.008	**-2.47**
dependent	0.001	0.006	0.18
sessionN	-0.048	0.028	-1.67
articleN	0.001	0.007	0.149
screenN	-0.014	0.005	**-2.50**
lineN	-0.012	0.004	**-2.69**
segmentN	-0.014	0.002	**-5.91**
is_first=TRUE	0.026	0.013	**2.00**
is_last=TRUE	0.063	0.017	**3.59**
is_second_last=TRUE	0.030	0.011	**2.65**
space=TRUE:sessionN	0.065	0.057	1.13
infostatus=discourse-old	-0.003	0.011	-0.30
definite=indefinite	0.041	0.030	1.34
specificity=specific	0.095	0.032	**2.95**
specificity=unspecific	0.038	0.020	1.82
animacy=inanimate	0.112	0.119	0.94
sentience=insentient	0.248	0.150	1.65
sentience=sentient	0.345	0.102	**3.37**
agentivity=both	-0.054	0.100	-0.54
agentivity=neither	0.013	0.014	0.91
agentivity=patient	-0.000	0.017	-0.01
commonness=hearer-new	0.001	0.014	0.09
commonness=hearer-old	-0.018	0.019	-0.94
commonness=neither	0.042	0.049	0.86

表 5: Parameters of the linear mixed model for the total time (TOTAL) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.500	0.105	23.69
length	0.135	0.004	**30.44**
space=TRUE	-0.043	0.007	**-5.51**
dependent	-0.001	0.006	-0.30
sessionN	-0.050	0.027	-1.82
articleN	-0.000	0.009	-0.09
screenN	-0.037	0.005	**-7.17**
lineN	-0.020	0.004	**-4.84**
segmentN	-0.015	0.002	**-7.27**
is_first=TRUE	0.061	0.011	**5.16**
is_last=TRUE	-0.007	0.016	-0.49
is_second_last=TRUE	0.027	0.010	**2.62**
space=TRUE:sessionN	0.062	0.054	1.14
infostatus=discourse-old	0.009	0.010	0.89
definite=indefinite	0.036	0.027	1.32
specificity=specific	0.070	0.029	**2.39**
specificity=unspecific	0.016	0.019	0.88
animacy=inanimate	0.177	0.108	1.63
sentience=insentient	-0.027	0.133	-0.20
sentience=sentient	0.130	0.089	1.46
agentivity=both	-0.025	0.091	-0.28
agentivity=neither	0.006	0.013	0.50
agentivity=patient	-0.011	0.015	-0.70
commonness=hearer-new	0.030	0.013	**2.34**
commonness=hearer-old	-0.000	0.017	-0.02
commonness=neither	0.033	0.045	0.74

験結果の 0 (fixation がない対象) のデータポイントを除外した。この時点でのデータポイント数は SELF が 6444 件、FFT・FPT・RPT・TOTAL が 5268 件、SPT が 2081 件である。

分析は常用対数時間に対して線形混合モデルに基づいて行い、最初に一度モデル化したうえで、標準偏差 ± 3.0 を超えるデータポイントを除外した。subj と article をランダム切片として、次のような式に基づき分析を行った。なお、ランダム切片に対する係数の組み合わせによるモデル選択は行っていない。

```
logtime ~ space * sessionN + length + dependent
  + is_first + is_last + is_second_last
  + articleN + screenN + lineN + segmentN
  + infostatus + definite + specificity + animacy
  + sentience + agentivity + commonness
  + (1 | subj) + (1 | article)
```

3 結果

自己ペース読文法 (SELF) と視線走査法 (FPT,RPT,TOTAL) の結果を表 2,3,4,5 に示す。紙面の制約上、情報構造に対して有意差が出なかった視線走査法 (FFT,SPT) の結果については省略する。

ここでは情報構造以外の傾向について確認しておく。文字数 (length) が多くなれば読み時間が長くなる傾向、および、実験が進むにつれて読み時間が短くなる傾向 (articleN, screenN, lineN, segmentN) がある。さらにレイアウト上、最左文節 (is_first)・最右文節 (is_last)・右から 2 番目の文節 (is_second_first) で読み時間が長くなる傾向がある。視線走査法においては、文節単位に半角空白を置いたほうが読み時間が短くなる傾向がある。一方、係る文節の数 (dependent) の効果は、対象を名詞句のみとした結果、有意差がなかった。

4 考察

表 6 に結果のまとめを示す。0 は読み時間に有意差がなかったものである。+ は読み時間が増える傾向にあり、- は読み時間が減る傾向にある固定要因である (いずれも有意差あり: $1.96 < |\text{t value}|$)。

定性 (definite) は読み時間に影響を与えることがない一方、特定性 (specificity) が、自己ペース読文法 (SELF) と視線走査法 (FPT, RPT, TOTAL) の読み時間を長くする効果があることがわかった。

また、有生性 (animacy) は視線走査法 (FPT) の読み時間を長くし、有情性 (sentience) は、視線走査法 (FPT, RPT) の読み時間を長く効果があることがわ

表 6: Summary: reading time and information structures

Fixed Effect		SELF	FFT	FPT	SPT	RPT	TOTAL
infostatus=discourse-old	(vs. discourse-new)	0	0	0	0	0	0
definite=indefinite	(vs. definite)	0	0	0	0	0	0
specificity=specific	(vs. either)	+	0	+	0	+	+
specificity=unspecific	(vs. either)	0	0	0	0	0	0
animacy=inanimate	(vs. animate)	0	0	+	0	0	0
sentience=insentient	(vs. either)	0	0	0	0	0	0
sentience=sentient	(vs. either)	0	0	+	0	+	0
agentivity=both	(vs. agent)	0	0	0	0	0	0
agentivity=neither	(vs. agent)	0	0	0	0	0	0
agentivity=patient	(vs. agent)	0	0	0	0	0	0
commonness=hearer-new	(vs. bridging)	+	0	0	0	0	+
commonness=hearer-old	(vs. bridging)	-	0	0	0	0	0
commonness=neither	(vs. bridging)	0	0	0	0	0	0

かった。日本語の文処理においては、他言語で言及される有生性よりも、日本語学で言及される有情性のほうが読み時間に影響を与えやすいことがデータから読み取れる。

さらに、共参照情報から得られる情報状態 (infostatus) は読み時間に影響を与えることがない一方、共有性 (commonness) については、言語受容者にとっての新情報 (hearer-new) が想定可能な要素 (bridging) に対して、有意に読み時間を長くする効果が自己ペース読文法と視線走査法 (TOTAL) に見られた一方、言語受容者にとっての旧情報 (hearer-new) が想定可能な要素 (bridging) に対して、有意に読み時間を短くする効果が自己ペース読文法に見られた。

最後に動作主性の差異は読み時間に影響を与えなかった。

このことから読み時間の差異により、特定性・(有生性・) 有情性・共有性について推定できる可能性があることがわかった。さらに、それぞれ読み時間の差異の出現傾向に違いがあることから、読み時間の測定値の組み合わせにより、各情報構造が推定できる可能性がある。

5 おわりに

本研究では、読み時間と情報構造の対照分析を行い、名詞句の特定性・有生性・有情性・共有性について、読み時間の差異が出現することを確認した。

本研究では『現代日本語書き言葉均衡コーパス』に対して付与された、文の読み時間データ『BCCWJ-EyeTrack』と、名詞句の定性などの情報構造アノテーションデータの対照分析を行った。日本語母語話者 24 人分のデータを線形混合モデルにより分析した結果、特定性 (specificity)・有情性 (sentience)・共有性 (commonness) が文の読み時間に影響を与え、それぞれ異なったパターンの読み時間の短縮・延長を引き起こすことがわかった。特に共有性においては新情報 (hearer-new)・想定可能 (bridging) が識別可能なレベルで異なった。このことは、ある名詞句が言語受容者にとって新情報なのか想定可能なのかを読み時間データから推定することができる可能性を示唆しており、文書要約のユーザ適応などの応用に利用することが期待できる。

謝辞

本研究は JSPS 科研費 基盤 (B) 25284083 「言語コーパスに対する読文時間付与とその利用」の助成を受けました。

参考文献

[1] Masayuki Asahara and Yuji Matsumoto. BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pp. 49–58, 2016.

[2] Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. Reading-Time Annotations for 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 684–694, 2016.

[3] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, Vol. 48, pp. 345–371, 2014.

[4] 宮内拓也, 浅原正幸, 中川奈津子, 加藤祥. 『現代日本語書き言葉均衡コーパス』に対する情報構造アノテーションの構築. 言語処理学会第 23 回年次大会発表論文集, 2017.