読み時間と述語項構造・共参照情報について

浅原 正幸 *
国立国語研究所

1. はじめに

本稿ではゼロ代名詞を含めて述語項構造・共参照情報が読み時間にどのような影響を与えるかについて、データに基づいて調査したので報告する。一般に係り受けの数が多ければ多いほど、述語要素が予測でき、読み時間が短くなるということが報告されている (*anti-locality*: Konieczny (2000))。そこで本稿では、1 つ目のリサーチクエスチョンとして、述語項の各要素が読み時間にどのような影響を与えるのかについて検討する。また、日本語は主語を含めて代名詞による項要素を省略することを許す。ある述語に対する省略された項要素をゼロ代名詞と呼ぶ。ゼロ代名詞が出現するということは、陽に表現されるべき項が消失することを表すが、日本語母語話者は問題なくゼロ代名詞が出現する文を読むことができる。2 つ目のリサーチクエスチョンとして、ゼロ代名詞の存在が読み時間にどのような影響を与えるのかについて検討する。具体的には、『現代日本語書き言葉均衡コーパス』(Maekawa et al. 2014) に対する読み時間データ (Asahara et al. 2016) とそれに対する述語項構造・共参照データ (植田ほか 2015) の重ね合わせを行い、一般化線形混合モデルにより分析を行う。結果、述語項構造・共参照情報のある特定のパターンで読み時間が変化することを確認したので報告する。

2. データと手法

まず、視線走査のデータ BCCWJ-EyeTrack について簡単に説明する。詳細については Asahara et al. (2016) を参照されたい。自己ペース読文法 (SELF) と視線走査法に基づき、後者は first fixation time (FFT)・first pass time (FPT)・regression path time (RPT), second pass time (SPT),・total time (TOTAL) の 5 つの指標に変換されている(1) 。

`surface` は表層形である。読み時間は (`time`) 対数に変換したもの (`logtime`) を用いる。`measure` は読み時間指標のタイプ {SELF, FFT, FPT, RPT, SPT, TOTAL} を表す。`sample, article, metadata_orig, metadata` は記事情報、`space` は文節単位に空白を入れて呈示したか否か、`length` は表層形の文字数、`is_first,is_last,is_second_first` は画面上のレイアウト情報、`sessionN, articleN, screenN, lineN, segmentN` は呈示順序情報を表す。`subj` は被験者情報で、統計処理においてランダム効果として用いる。`dependent` は、人手で付与された (Asahara and Matsumoto 2016) である。

次に述語項構造・共参照アノテーションデータについて説明する。アノテーション基準は NAIST テキストコーパスに準ずる(2)が、本研究は文節単位に割り当て直したものを用いる。`type_pred` は述語の型を表し、名詞述語か (NOUN)、動詞・形容詞述語か (PRED)、それ以外を表す。`type_subj, type_dobj, type_iobj` は、ガ・ヲ・ニの述語項関係がどのような状態にあるかを表す：'DEP' は当

* masayu-a@ninjal.ac.jp

(1) First Fixation Time (FFT) はその注視領域に初めて視線が停留した際の注視時間である。First-Pass Time (FPT) は、注視領域に初めて視線が停留し、その後注視領域から出るまでの総注視時間である。出る方向は右方向でも左方向でも構わない。Second-Pass Time (SPT) は、注視領域に初めて視線が停留し、注視領域から出たあと、2 回目以降に注視領域に停留する総注視時間である。Regression Path Time (RPT) は、注視領域に初めて視線が停留し、その後領域の右側の境界を超えて次の領域に出るまでの総注視時間である。視線が領域の左側の境界を超えて戻った場合の注視時間も、元の注視領域の RPT として合算する。Total Time (TOTAL) は注視領域に視線が停留する総注視時間である。

(2) `https://sites.google.com/site/naisttextcorpus/ntc-annotation-scheme`

該項が述語と係り受け関係にあることを表す；'ZERO' は外界照応以外で当該項が係り受け関係にないが文内に項要素が出現することを表す；'EXO1' は外界照応が 1 人称（書き手）をさすことを表す；'EXO2' は外界照応が 2 人称（読み手）をさすことを表す；'EXOG' は外界照応が 1 人称・2 人称以外をさすことを表す；'NIL' は当該項が存在しないことを表す（自動詞など）(3)。

`dist_inv_subj`, `dist_inv_dobj`, `dist_inv_iobj` は述語項関係が定義されている場合の距離の逆数を表す。距離の逆数は文節の数で表現し、述語の n 個前の文節に項が出現する場合に $\frac{1}{n}$ と定義する。項が述語と同じ文節に出現する場合 ($n = 0$；名詞述語など)、距離の逆数を 1.0 とする。項が前の文に出現したり、外界照応である場合、距離の逆数を 0.0(つまり、距離が ∞）とする。

統計分析は、対数読み時間 (`logtime`) に対して行う。前処理として `metadata` が {`authorsData`, `caption`, `listItem`, `profile`, `titleBlock`} のものを削除する。読み時間が確認されなかったデータポイントについても削除する。さらに 1 回目のモデルで ±3.0 標準偏差より外側のデータポイントを排除したうえで、2 回目のモデルを評価に用いる (Baayen 2008)。`subj` と `article` 記事をランダム効果としたうえで、次式により回帰した：

```
logtime ~ space * sessionN + length + dependent
        + is_first + is_last + is_second_last
        + articleN + screenN + lineN + segmentN
        + type_pred + type_subj + type_dobj
        + type_iobj + dist_inv_subj + dist_inv_dobj + dist_inv_iobj
        + (1 | subj) + (1 | article)
```

3. 結果と分析

表 1 に結果を示す。

`space`, `length`、レイアウト情報 (`is_first`, `is_last`, `is_second_last`)、進捗情報 (`sessionN`, `articleN`, `screenN`, `lineN`, `segmentN`) の結果は、過去の研究 (Asahara et al. 2016) と同じであった。また `dependency` の結果は、係り受けの数が多ければ多いほど、予測が効いて読み時間が短くなる (*anti-locality*: Konieczny (2000))。

`type_pred` の結果は述語の種類による違いを示している。`type_pred`=PRED (SELF, FPT, TOTAL) の結果は、動詞・形容詞述語において、読み時間データが短くなる傾向を見せている。これは、日本語が述部が項よりも後にくる言語であり、格助詞でマークされた先行する項の情報が読みを促進していることによると考える。`type_pred`=NOUN (RPT) の結果は、名詞述語においてその地点からより長く読み戻す傾向を示唆している。名詞述語の項要素は、格助詞でマークされていないことが多いため、項の確認のためにその地点から読み戻す傾向があるのではないかと考える。

`type_subj`=EXO2 (SPT) の結果は、主語のゼロ代名詞が外界 2 人称を指す場合に述語要素で、2 回目以降の読み時間が短くなる傾向を示唆している。読み手自身の経験を表出する述語は読み手に強い印象を与えるために、このような傾向がみられたのではないかと考える。ほかに、`type_subj`=EXOG (SELF, SPT) の結果は主語がその他の外界照応である述語でも、自己ペース読文法による読み時間と 2 回目以降の読み時間が若干短くなる傾向がみられた。

`type_dobj`=NIL (SELF) の結果は、自己ペース読文法で自動詞（正確にはヲ格を持たない動詞）や形容詞で読み時間が短くなる傾向を示唆される。

(3) BCCWJ-EyeTrack では `subj` は被験者情報を表すが、述語項データにおける `*_subj` は、ガ格を表すことに注意。

表 1　線形混合モデルの結果

	Dependent variable:					
	logtime					
	SELF	FFT	FPT	SPT	RPT	TOTAL
space	−0.001	−0.006	−0.018***	−0.040***	−0.019***	−0.030***
logical	(0.002)	(0.004)	(0.005)	(0.009)	(0.006)	(0.005)
length	0.089***	−0.002	0.139***	0.023***	0.119***	0.134***
integer	(0.001)	(0.002)	(0.003)	(0.005)	(0.003)	(0.003)
is_first	0.050***	0.021***	0.091***	−0.028**	0.030***	0.069***
logical	(0.004)	(0.006)	(0.008)	(0.013)	(0.009)	(0.008)
is_last	0.035***	−0.012*	0.012	−0.050***	0.085***	−0.008
logical	(0.004)	(0.007)	(0.008)	(0.016)	(0.010)	(0.009)
is_second_last	−0.009***	−0.00001	0.036***	−0.006	0.047***	0.036***
logical	(0.004)	(0.006)	(0.007)	(0.012)	(0.008)	(0.007)
sessionN sessionN	−0.022	−0.022	−0.041*	−0.036**	−0.049*	−0.047*
integer	(0.021)	(0.016)	(0.024)	(0.018)	(0.025)	(0.024)
articleN	−0.028***	−0.005	−0.006	−0.003	−0.010	−0.004
integer	(0.006)	(0.004)	(0.007)	(0.007)	(0.007)	(0.008)
screenN	−0.030***	−0.004	−0.018***	−0.015***	−0.016***	−0.026***
integer	(0.002)	(0.003)	(0.003)	(0.006)	(0.004)	(0.003)
lineN	−0.011***	−0.010***	−0.018***	−0.017***	−0.008**	−0.019***
integer	(0.001)	(0.002)	(0.003)	(0.005)	(0.003)	(0.003)
segmentN	−0.004***	0.003***	−0.005***	−0.009***	−0.012***	−0.011***
integer	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)
dependent	−0.009***	−0.006**	−0.017***	−0.015**	−0.015***	−0.018***
integer	(0.002)	(0.003)	(0.003)	(0.006)	(0.004)	(0.003)
type_pred=NOUN	−0.044	0.046	0.017	0.166	0.357***	0.071
vs. =NONE	(0.054)	(0.074)	(0.093)	(0.138)	(0.112)	(0.097)
type_pred=PRED	−0.043*	−0.026	−0.091**	−0.069	−0.050	−0.099**
vs. =NONE	(0.023)	(0.036)	(0.044)	(0.090)	(0.054)	(0.046)
type_subj=EXO1	−0.025	0.035	0.014	−0.008	−0.061	0.025
vs. =DEP	(0.021)	(0.033)	(0.041)	(0.093)	(0.049)	(0.043)
type_subj=EXO2	0.016	−0.036	0.092	−0.499*	−0.038	0.025
vs. =DEP	(0.049)	(0.092)	(0.114)	(0.302)	(0.138)	(0.120)
type_subj=EXOG	−0.025**	0.002	−0.022	−0.069*	−0.020	−0.011
vs. =DEP	(0.010)	(0.015)	(0.019)	(0.037)	(0.023)	(0.020)
type_subj=NIL	−0.018	−0.014	−0.050	−0.039	−0.008	−0.046
vs. =DEP	(0.024)	(0.036)	(0.045)	(0.092)	(0.055)	(0.047)
type_subj=ZERO	−0.008	−0.0004	−0.002	−0.035	−0.017	−0.008
vs. =DEP	(0.006)	(0.009)	(0.011)	(0.023)	(0.014)	(0.012)
type_dobj=EXOG	0.002	0.030	−0.120	0.247	−0.233	−0.109
vs. =DEP	(0.048)	(0.127)	(0.158)	(0.301)	(0.192)	(0.166)
type_dobj=NIL	−0.020**	0.001	−0.010	−0.028	−0.022	−0.010
vs. =DEP	(0.010)	(0.016)	(0.019)	(0.037)	(0.023)	(0.020)
type_dobj=ZERO	−0.037***	−0.033*	−0.065***	−0.010	−0.062**	−0.049**
vs. =DEP	(0.012)	(0.019)	(0.024)	(0.047)	(0.029)	(0.025)
type_iobj=EXOG	−0.153***	−0.0001	0.076	−0.234	−0.016	−0.001
vs. =DEP	(0.050)	(0.094)	(0.117)	(0.306)	(0.142)	(0.123)
type_iobj=NIL	0.021	−0.002	0.074**	−0.050	0.054	0.058
vs. =DEP	(0.018)	(0.029)	(0.036)	(0.070)	(0.044)	(0.038)
type_iobj=ZERO	−0.005	0.017	0.085*	0.085	0.053	0.079
vs. =DEP	(0.022)	(0.037)	(0.046)	(0.090)	(0.056)	(0.049)
dist_inv_subj	−0.005	0.020*	0.014	−0.024	0.029	0.014
numeric	(0.008)	(0.012)	(0.015)	(0.030)	(0.018)	(0.016)
dist_inv_dobj	−0.027**	−0.014	−0.062***	0.003	−0.038	−0.053**
numeric	(0.011)	(0.017)	(0.021)	(0.040)	(0.025)	(0.022)
dist_inv_iobj	0.014	−0.026	0.024	0.013	0.017	0.040
numeric	(0.020)	(0.032)	(0.040)	(0.077)	(0.048)	(0.042)
space=TRUE:sessionN	−0.016	0.044	0.059	0.060*	0.061	0.061
logical× integer	(0.042)	(0.031)	(0.048)	(0.035)	(0.050)	(0.047)
Constant	2.805***	2.321***	2.525***	2.581***	2.583***	2.677***
	(0.038)	(0.052)	(0.066)	(0.123)	(0.079)	(0.069)
Observations	17,628	13,232	13,232	4,769	13,232	13,232
Log Likelihood	7,134.189	1,317.788	−1,555.933	−1,016.164	−4,105.689	−2,193.576
Akaike Inf. Crit.	−14,204.380	−2,571.576	3,175.866	2,096.327	8,275.378	4,451.152
Bayesian Inf. Crit.	−13,955.510	−2,331.883	3,415.559	2,303.364	8,515.071	4,690.844

Note: *p<0.1; **p<0.05; ***p<0.01

type_dobj=ZERO (SELF, FFT, FPT, RPT, TOTAL) は、ゼロ代名詞を有する述語が、より短い読み時間であることを示唆している。ゼロ代名詞を有するということは、文脈上、項を省略してもよい程度に顕現化していることが考えられ、後置する述語もある程度予測可能であり、読み時間が短くなるのではないかと考える。

ヲ格の述語項間の距離の逆数 dist_inv_dobj (SELF, FPT, TOTAL) は、負の値が割り当てられており、言い換えると距離が短いほど、距離の逆数は大きくなり、読み時間が短くなることを示唆している。ヲ格は一般的に述語に近いところにある。ヲ格と述語の距離が長くなるということは、談話的な理由（強調・情報の新旧）によりヲ格が先行していることが考えられる。ヲ格が先に来ることにより予測が効くということは、ヲ格の選択選好性が他の格よりも強い影響を述語に与えているのではないかと思います。

4. おわりに

本稿では日本語の読み時間と述語項構造・共参照情報との対照分析を行った。本分析は人間の文処理を促進するいくつかのパターンを、外界照応を含めた述語項構造・共参照情報の中から明らかにした。1 つ目のリサーチクエスチョンに対しては、自動詞述語 < 他動詞・形容詞述語 < 名詞述語の順で読み時間が短くなる傾向や、ヲ格が先行する文脈においては述語で読み時間が短くなる傾向が明らかになった。2 つ目のリサーチクエスチョンに対しては、例えば、外界照応においてガ格が二人称を指す場合、second-pass time (SPT) を短くすることが観察された。これは読み手の経験を指すような事象を表現する述語が、強く印象付けられることにより理解が促進され、2 回読む際の読み時間が短くなることを意味すると考える。

謝　辞

本研究は国立国語研究所コーパス開発センター共同研究プロジェクトおよび科研費 JP17H00917, JP18H05521, 18K18519 によるものです。

文　献

Lars Konieczny (2000). “Locality and Parsing Complexity.” *Journal of Psycholinguistic Research*, 29:6.

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

Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto (2016). “Reading-Time Annotations for ‘Balanced Corpus of Contemporary Written Japanese’.” *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 684–694.

植田禎子・飯田龍・浅原正幸・松本裕治・徳永健伸 (2015).「『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照アノテーション」 第 8 回コーパス日本語学ワークショップ, pp. 205–214.

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

R. Harald Baayen (2008). *Analyzing Linguistic Data: A practical Introduction to Statistics using R.*: Cambridge University Press.