『言語による時間生成』論文集・報告集　pp. 142–148

時間情報アノテーションデータ

浅原正幸[*1]
越智綾子[*2]
鈴木彩香[*1]

要　旨

A01 班では，他班に日本語 BERT を含めた人工神経回路を提供するとともに，言語表現の分析データとして，アノテーションデータを提供した。本稿では，提供したアノテーションデータの仕様などについて報告する。

キーワード：時間情報，アノテーション

1. はじめに

A01 班では，時間生成学の中心的な役割として，他班に 日本語 BERT（Devlin et al. 2019）を含めた時間情報を推定するための人工神経回路を提供した。時間情報の推定のためには，人工神経回路の訓練データとして時間情報アノテーションデータが必要である。国立国語研究所では，各班から提供されるさまざまなデータに対して時間情報アノテーションを進めてきた。B01 班の脳活動データ収集時に刺激として呈示された DVD 動画・日本語話し言葉コーパスや，D01 班が収集した「過去」「未来」を主題とした作文データに対して，絶対時制・時間的順序関係・時間幅・時間間隔・{事実，仮想} の情報を付与した。本稿では作成したアノテーションデータの仕様について報告する。

2. アノテーションデータ

2.1. 時間情報アノテーション

本研究では，文法的に厳密な調査ではなく，一般の方がどのようにテキスト・音声・動画を捉えるかを記録することを目標とする。出来事を表す表現として，形態素解析結果の「動

*1 国立国語研究所　*2 元国立国語研究所

詞」に相当する表現のみを扱うこととした。「する」が後置するサ変名詞については，「する」ではなくサ変名詞を情報付与対象とした。

出来事を表す表現には「絶対時制」「時間幅」「事実性」を付与した。「絶対時制」は出来事を表す表現が，「過去」・「現在」・「未来」を表すのかを付与した。「時間幅」は表現が表す時間幅が，「瞬時〜1 秒未満」・「1 秒以上〜1 分未満」・「1 分以上〜1 時間未満」・「1 時間以上〜1 日未満」・「1 日以上〜1 年未満」・「1 年以上（常に成り立つは除く）」・「常に成り立つ」のいずれかを付与した。「事実性」は，表現が表す出来事が「現実」か「仮想」かを付与した。未来の表現は基本的に「仮想」として，過去の表現は反実仮想のものを「仮想」とした。

2 つの隣接する出来事表現については，「時間的順序関係」「時間間隔・重なり」を付与した。「時間的順序関係」は「A<B：A が B より前」・「A<=B：A が B より前だが重なりあり」・「A==B：重なりあり」・「B<=A：B が A より前だが重なりあり」・「B<A：B が A より前」のいずれかを付与した。「時間間隔・重なり」は時間幅と同様に「瞬時〜1 秒未満」・「1 秒以上〜1 分未満」・「1 分以上〜1 時間未満」・「1 時間以上〜1 日未満」・「1 日以上〜1 年未満」・「1 年以上（常に成り立つは除く）」・「常に成り立つ」のいずれかを付与した。

なお，与えられた文脈だけから判断がつかない場合には「わからない」を付与することをゆるした。作業は一般の作業者に付与を依頼したあと，専門的な知識を持っている者により全体の整合性をとる方向性で修正をおこなった。

2.2. アノテーション対象：DVD 動画

B01 班から提供された動画データに対して時間情報アノテーションを行った。B01 班では，同動画データを刺激としたうえで，fMRI による脳活動データ収集と頭蓋内電極による脳活動データ収集が行われている。B01 班動画データを書き起こして，形態素解析を行った。対象は以下のとおりである：

●BreakingBad（シーズン 1：1 話）
●glee/ グリー（シーズン 1：1 話）
●HEROES（シーズン 1：1 話）
●ザ・クラウン（シーズン 1：1 話）
●SUITS/ スーツ（シーズン 1：1 話）
●ドリームガールズ（映画）
●ビッグバン★セオリー / ギークなボクらの恋愛法則（シーズン 1：1 話〜3 話）
●メンタリスト（シーズン 1：1 話）
●攻殻機動隊　STAND ALONE COMPLEX（1 話・2 話）

出来事を表す表現には DVD 上の Timestamp を付与した。これにより脳活動データとの時間対応をとることができる。

2.3. アノテーション対象：日本語話し言葉コーパス

日本語話し言葉コーパス（Maekawa 2003）に収録されている模擬講演のうち，18 ファイルを対象に時間情報アノテーションを行った。同データを音声刺激で呈示した脳活動データがB01 班により収集されている。日本語話し言葉コーパスには，絶対時制・時間幅・{現実，仮想} のみを付与した。

2.4 アノテーション対象：「過去」「未来」を主題とした作文

D01 班が収集した「過去」「未来」を主題とした作文データについても，絶対時制・時間幅・{現実，仮想} の情報を付与した。同データについては統計分析を行い，『計量国語学』で発表した（浅原ほか 2023）。

3. 基礎統計

以下では，他文献で発表していない DVD 動画と日本語話し言葉コーパスの基礎統計について示す。

3.1. DVD 動画の基礎統計

表 1 に DVD 動画の絶対時制情報と {現実，仮想} の基礎統計を，表 2 に時間幅の情報を示す。集計においては「わからない」と記載されていたものについては計数しなかった。物語の進行を支える言語表現が多いために，他の媒体と比較して未来の表現が多い傾向にある。結果として，仮想の表現が多い傾向も確認された。DVD 動画は文脈が映像によって示され，発話される出来事は場面による変化の言及が多いために瞬時・1秒・1分などの短い出来事が多い傾向にある。

表 1　DVD 動画の絶対時制情報と {現実，仮想}

Filename	過去	現在	未来	現実	仮想
BreakingBad	90	167	225	406	76
glee	68	108	145	143	171
HEROES	91	146	155	208	199
ザ・クラウン	46	106	124	121	152
SUITS	172	189	297	281	389

ドリームガールズ	80	181	209	222	227
ビッグバン★セオリー1	46	93	89	107	109
ビッグバン★セオリー2	72	89	74	119	96
ビッグバン★セオリー3	46	84	99	88	128
メンタリスト	151	107	80	188	149
攻殻機動隊 1	48	46	47	65	73
攻殻機動隊 2	45	63	73	70	96

表 2 DVD 動画の時間幅

Filename	瞬時	1 秒	1 分	1 時間	1 日	1 年	常時
BreakingBad	39	120	119	31	53	13	93
glee	89	24	29	9	16	23	129
HEROES	191	21	34	5	13	16	87
ザ・クラウン	86	29	24	11	25	13	56
SUITS	347	94	25	18	24	34	81
ドリームガールズ	183	75	34	5	13	13	58
ビッグバン★セオリー1	106	41	7	0	0	10	35
ビッグバン★セオリー2	102	34	19	11	0	2	50
ビッグバン★セオリー3	135	40	3	1	6	0	11
メンタリスト	185	53	7	1	15	6	36
攻殻機動隊 1	53	14	6	1	9	1	32
攻殻機動隊 2	90	19	10	0	0	0	44

3.2. 日本語話し言葉コーパスの基礎統計

表 3 に日本語話し言葉コーパスの絶対時制情報と {現実, 仮想} の基礎統計と内容を，表 4 に時間幅の情報を示す。基本的に過去の経験について話しているデータが多いため，過去・現在・現実の表現が多い。S00M1046「電車のきままな旅」のように未来・仮想の話が全く出現しない話者もいた。一方で S04M0497「人間ドック」のように極端に過去の出来事が少なく，現在・未来が多い話者もいた。

表 3　日本語話し言葉コーパスの絶対時制情報と {現実, 仮想} と内容

Filename	過去	現在	未来	現実	仮想	内容
S00F0131	92	138	12	214	28	乗馬
S00F0374	31	149	4	171	13	日本酒について
S00F0458	125	54	3	167	15	方言について
S00F1396	126	63	6	189	6	中学生時代の思い出
S00M0117	94	56	2	150	2	地下鉄サリン事件
S00M1046	245	15	0	260	0	電車のきままな旅
S01F1707	141	55	39	203	33	アナウンサーの勉強の思い出
S02F0100	175	36	23	212	22	難病になったこと
S02F1109	159	51	8	208	10	交通事故と厄年
S02F1183	147	15	26	162	26	阪神大震災
S02F1704	210	40	18	235	33	実家がダムの底に
S03F0072	192	42	13	232	15	イランでの生活
S04M0497	8	192	91	202	89	人間ドック
S04M0790	44	134	45	177	46	都市型伝説について
S05M1110	174	47	34	208	47	66 歳でジャズ学校に通う
S06M0740	219	66	26	260	51	ストーカーについて
S08M1702	81	144	22	216	31	結婚式の挙げかた
S11M0472	207	89	25	265	56	アレルギー性鼻炎

表 4 の傾向から，日本語話し言葉コーパスには 1 年以上の出来事の描写が多い。長年続けている習慣的な経験が語られている傾向が見られた。

表 4　日本語話し言葉コーパスの時間幅

Filename	瞬時	1 秒	1 分	1 時間	1 日	1 年	常時
S00F0131	28	28	26	5	5	25	115
S00F0374	5	10	7	1	2	18	136
S00F0458	20	17	11	0	4	85	32
S00F1396	0	1	0	0	5	113	21
S00M0117	0	0	0	0	0	99	42
S00M1046	1	0	0	0	0	246	5
S01F1707	1	3	0	0	1	138	30
S02F0100	1	0	1	0	0	177	20
S02F1109	1	3	0	0	1	168	33

S02F1183	38	4	58	43	8	20	4
S02F1704	64	10	37	37	28	63	23
S03F0072	39	13	30	38	38	54	34
S04M0497	17	3	44	7	16	21	181
S04M0790	26	16	33	6	3	29	100
S05M1110	40	23	35	11	27	86	27
S06M0740	43	49	50	6	26	3	63
S08M1702	12	25	20	20	7	27	135
S11M0472	47	63	40	26	9	54	71

4. おわりに

本稿では,「時間生成学」の共同研究に資する言語データに対する時間情報アノテーションの仕様と基礎統計について述べた。同データは，A01 班お茶の水女子大学・京都大学の研究グループにより時間情報の自動推定のための訓練データとして利用されるほか，B01 班により脳活動データの分析にも利用されている。D01 班の作文データについては，D01 班–A01 班共同で執筆した論文（浅原ほか 2023）にて，主題差・年代差・性差による文体の分析をおこなった。

参考文献

J. Devlin, Chang, M.-W., Lee, K., Toutonova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 4171–4186.

K. Maekawa. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. In: ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition.

浅原正幸・川﨑采香・上原泉・酒井裕・須藤百香・谷口巴・小林一郎・越智綾子・鈴木彩香．2023.「過去」「未来」を主題とする作文の分析，『計量国語学』34 巻 1 号．17–30.

Temporal Information Annotation Data

Masayuki ASAHARA, Ayako OCHI, Ayaka SUZUKI

Abstract

Group A01 provided artificial neural networks including Japanese BERT to other groups and also provided annotation data as language expression analysis data. In this paper, we report on the specifications of the provided annotation data.

Keywords: Temporal Expressions, Annotation