NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所
NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

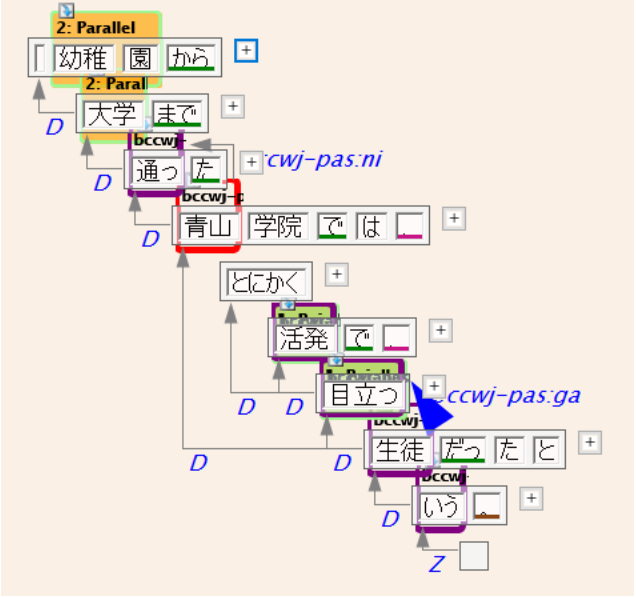
読み時間と情報構造

国立国語研究所　コーパス開発センター
浅原正幸

はじめに

はじめに

- Psycholinguistics × Corpus Linguistics



Reading Time × Balanced Corpus × Corpus Annotation

関連研究

関連研究

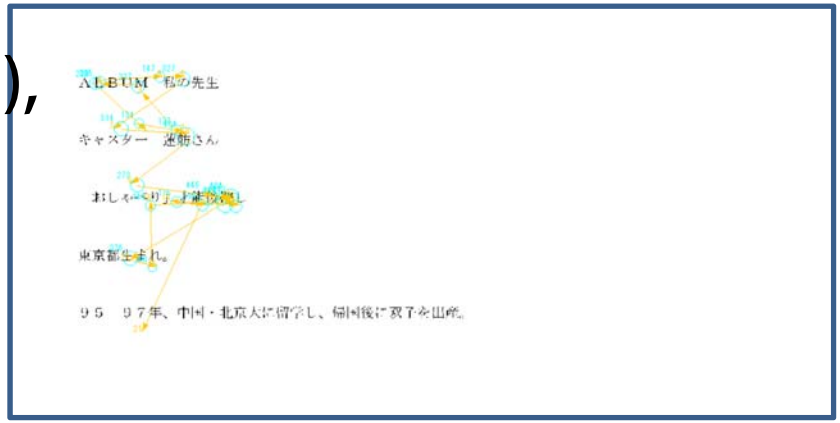
Dundee EyeTracking Corpus [Kennedy+ 2005]

- テキスト: 新聞記事社説 (英・仏)
  - 20 files with 40 screens / 5 lines
- 被験者:英・仏母語話者 (10 + 10)
- 2形式:
  - 視線走査順
  - 元文書の単語順

BCCWJ-EYETRACK

コーパスの設計
手法と環境

- 自己ペース読文法
  - linger
- 視線走査法
  - EyeLink 1000 (SR Research),
  - tower mount, 1000Hz



- ディスプレイ
  - EIZO FlexScan EV2116W
  - 1920x1080 (Full-HD, 1080p)
  - MS 明朝24pt
  - 最大 5 行 x 53 文字
  - 文節境界にスペースありとスペースなし

コーパスの設計
元テキスト

『現代日本語書き言葉均衡コーパス』

- 新聞記事コアデータ PN
- 5-6 記事x 4 sets {A, B, C, D}

Data	*文節*	文	画面
A	470	66	19
B	455	67	21
C	355	44	16
D	363	41	15

コーパスの設計
被験者のグループ化

- 24人の日本語母語話者
  - 2015年8月～12月に実験を実施

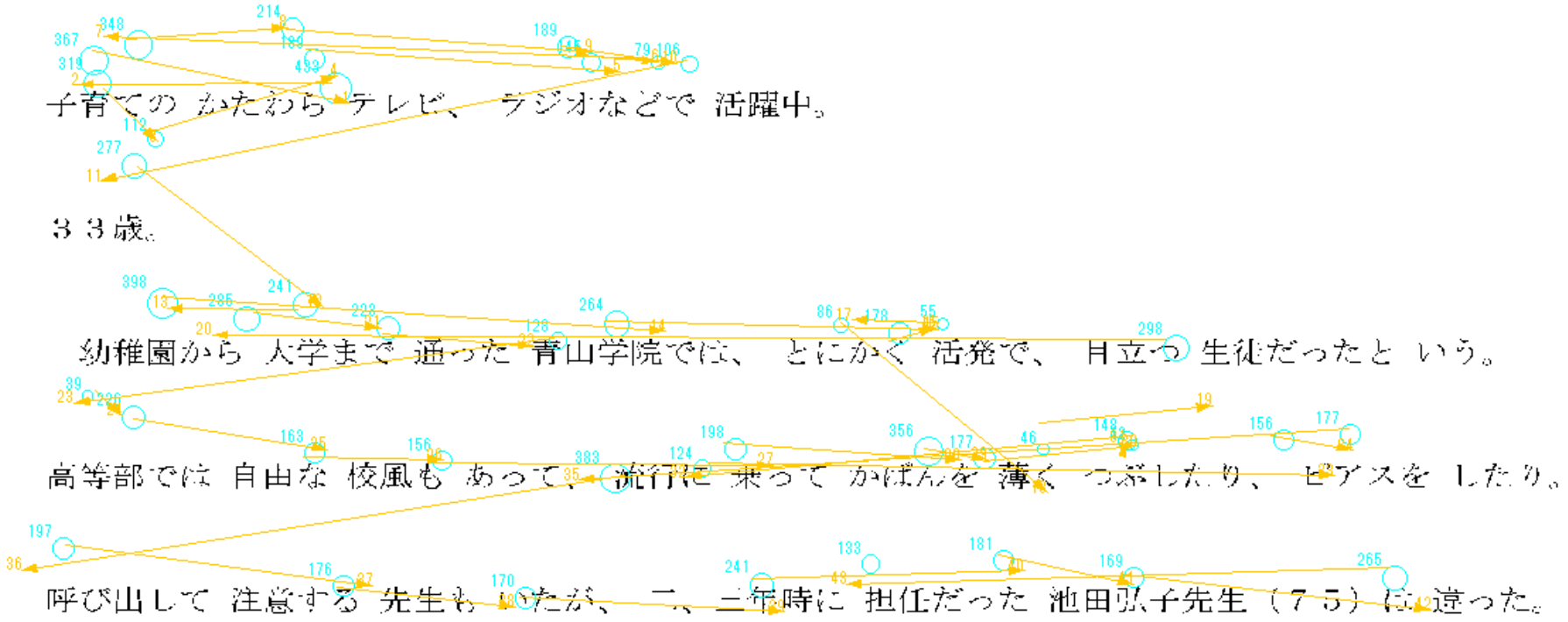
Group	視線走査法		自己ペース読文法	
1	A 境界なし	B 境界あり	C 境界なし	D 境界あり
2	A 境界あり	B 境界なし	C 境界あり	D 境界なし
3	C 境界なし	D 境界あり	A 境界なし	B 境界あり
4	C 境界あり	D 境界なし	A 境界あり	B 境界なし
5	B 境界なし	A 境界あり	D 境界なし	C 境界あり
6	B 境界あり	A 境界なし	D 境界あり	C 境界なし
7	D 境界なし	C 境界あり	B 境界なし	A 境界あり
8	D 境界あり	C 境界なし	B 境界あり	A 境界なし

コーパスの設計
被験者の言語背景情報

- アンケート
  - 年齢（5歳刻み）
  - 生年代 (5年刻み)
  - 性別
  - 出生地
  - 学歴（専門分野）
  - 言語形成地：0-15歳の在住地
  - 両親の出生地
  - 裸眼かソフトコンタクトレンズか

- テスト
  - リーディングスパンテスト [苧坂+ 2002]
  - 語彙数テスト [天野+ 1998]

コーパスの設計
視線走査実験データ

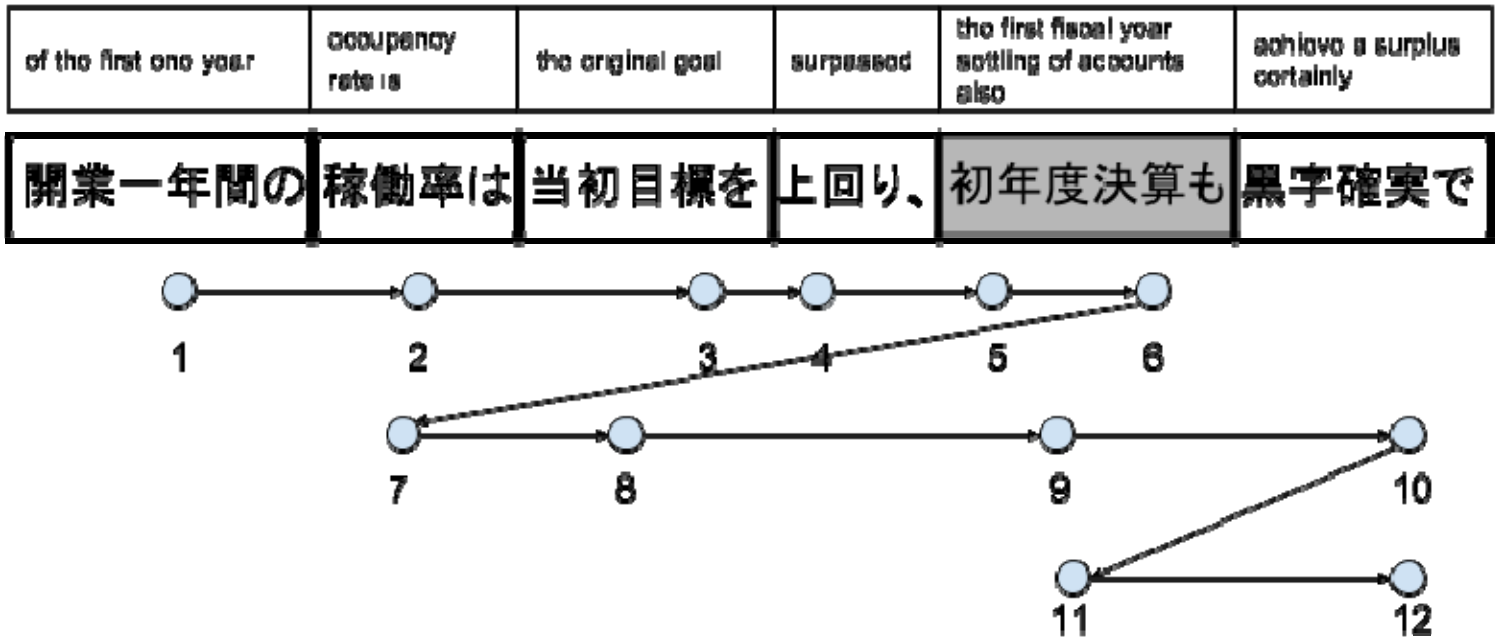


コーパスの設計
視線走査実験データ の言語資源化

視線走査順から元テキスト単語順への変換

読み時間の指標	概要
First Fixation Time (FFT)	Interest Area (文節の表示範囲) に最初に視線が入ったときの停留時間
First-Pass Time (FPT)	Interest Area に最初に視線が入ってから、左右どちらかに出るまでの停留時間の合計
Regression Path Time (RPT)	Interest Area に最初に視線が入ってから、右に出るまでの停留時間の合計
Second-Pass Time (SPT)	Total Time(Total) – First-Pass Time(SPT)
Total Time (Total)	Interest Area の停留時間の合計

コーパスの設計→視線走査実験データ の言語資源化
読み時間の指標（例）



measurements	「初年度決算も」
First Fixation Time (FFT)	5
First-Pass Time (FPT)	5, 6
Regression Path Time (RPT)	5, 6, 7, 8, 9
Second-Pass Time (SPT)	9, 11
Total Time (Total)	5, 6, 9, 11

コーパスの設計→視線走査実験データ の言語資源化
データ形式(1/2)

Column Name	Type	Description	Example
surface	factor	表層文字列	“初年度決算も”
time	int	読み時間 (ms)	750
logtime	num	対数読み時間	2.87
measure	factor	読み時間集計手法	“EyeTrack:FFT”, “EyeTrack:FPT” ...
sample	factor	サンプル名	{“A”, “B”, “C”, “D”}
article	factor	記事情報	“00001_A_PN1c_0001_A_1”, ...
metadata_orig	factor	文書構造タグ	“titleBlock”, “authorData”, ...
metadata	factor	修正文書構造タグ	“titleBlock”, “authorData”, ...

Analysis
Data format (2/2)

Column Name	Type	Description	Example
sessionN	int	セッション順	{1,2}
articleN	int	記事呈示順	{1,2,3,4,5}
screenN	int	画面呈示順	{1,2,...,21}
lineN	int	行(画面縦方向)	{1,2,3,4,5}
segmentN	int	文節番号(画面横方向)	1,2,3,...
sample_screen	factor	画面識別子	{“A_1”,“A_2”,..., “D_15”}
length	int	文字数	6
space	factor	文節間に空白をいれるかいれないか	{“0”, “1”}
subj	factor	実験協力者ID	“U100729”
dependent	int	係る文節数	0, 1, 2, ...

情報構造アノテーション

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所
NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

『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

共参照アノテーションの拡張としての情報構造アノテーション

7種類の情報を BCCWJ 新聞(PN)コアサンプル 16ファイル名詞句 2023 件に対して付与

今回はそのうちの595件のみ利用

- 情報状態 (information status)
- 共有性 (commonness)
- 定性 (definiteness)
- 特定性 (specificity)
- 有生性 (animacy)
- 有情性 (sentience)
- 動作主性 (agentivity)

宮内ほか(2017)
『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

情報構造アノテーション
情報状態 (information status)

- テクスト（談話）中に既出か未出か
- 共参照アノテーションから判断
  - 旧情報(discourse-old)　＝既出
  - 新情報(discourse-new)　＝未出

旧情報	新情報
228	367

情報構造アノテーション
共有性 (commonness)

- 情報を受容者側が既に知っていると発信者側が想定しているか否か (既知 or 未知)
  - 共有 (hearer-old)　＝既知
  - 非共有(hearer-new)　＝未知
  - 想定可能(bridging)

共有	非共有	想定可能	どちらでもない
337	109	143	6

情報構造アノテーション
定性 (definiteness)

- 指示対象を受容者が同定できるか否か
  本研究の基準：スコープとして前後3文を見る
  - 定(definite)
  - 不定 (indefinite)

定	不定	どちらでもよい
358	236	1

情報構造アノテーション
特定性 (specificity)

- 発信者が特定の事物を想定しているか
  本研究の基準：スコープとして前後3文を見る
  – 特定(specific)
  – 不特定 (unspecific)

特定	不特定	どちらでもよい
384	187	24

情報構造アノテーション
有生性 (animacy)

- 生きているか否か

  本研究の基準：名詞句レベルで判別

  – 有生 (animate)

  – 無生 (inanimate)

有生	無生
94	501

情報構造アノテーション
有情性 (sentient)

- 情意があるか否か

  自由意志による移動が可能か否か

本研究の基準：述語-項レベルで判別

– 有情 (sentient)

– 無情 (insentient)

有情	無情	どちらでもよい
91	502	2

情報構造アノテーション
動作主性 (agentivity)

- 事態に関わる人がその事態ではたしている役割

  本研究の基準：節レベルで判別

  – 動作主 (agent)
  – 被動作主(patient/theme)
  – どちらでもある (主節で agent, 従属節で theme)

動作主	被動作主	どちらでもある	どちらでもない
79	98	1	417

分析

分析
線形混合モデル

- データ処理
  - メタデータ “authorsData”, “caption”, “listItem”, “profile”, and “titleBlock” を排除
  - ゼロ秒データ（視線停留なし）を分析対象から排除
- 外れ値除去
  - ± 3-SD 以上のデータポイントを除去
- レイアウト要因の追加
  - is_first, is_last, is_second_last

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

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所
NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

結果まとめ(一般)

+: t-value > 1.96
-: t-value < -1.96
0: others

Fixed Effect	SELF	FFT	FPT	SPT	RPT	TPT
length	+	-	+	+	+	
space=T	0	0	-	-	-	-
dependent	0	0	0	0	0	
sessionN	0	0	0	0	0	0
articleN	-	0	0	0	0	0
screenN	-	-	-	-	-	-
lineN	-	-	-	0		
segmentN	-	0	-	-	-	-
is_first=T	+	0	+	0	+	
is_last=T	+	0	0	-	+	
is_second_last=T	-	0	+	0	+	+
space=T:sessionN	0	0	0	0	0	0

文字列長
読み時間 +

空白入り
読み時間 -

係り受け影響なし

呈示順 読み時間 -

レイアウト要因

結果まとめ(情報構造)

+: t-value > 1.96
-: t-value < -1.96
0: others

Fixed Effect		SELF	FFT	FPT	SPT	RPT	Total
infostat=discourse-old	(vs. d-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=h-new	(vs. bridging)	+	0	0	0	0	+
commonness=h-old	(vs. bridging)	-	0	0	0	0	0
commonness=neither	(vs. bridging)	0	0	0	0	0	0

NINJAL 大学共同利用機関法人 人間文化研究機構 国立国語研究所 NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUISTICS

+: t-value > 1.96
-: t-value < -1.96
0: others

結果まとめ(情報構造

情報状態
有意差なし

情報状態
有意差なし

動作主性
有意差なし

Fixed Effect					SPT	RPT	Total
infostat=discourse-old	(vs. d-new)	0	0	0	0	0	0
definite=indefinite	(vs. definite)	0	0	0	0	0	0
specificity=specific	(vs. either)	+	0	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=h-new	(vs. bridging)	+	0	0	0	0	+
commonness=h-old	(vs. bridging)	-	0	0	0	0	0
commonness=neither	(vs. bridging)	0	0	0	0	0	0

+: t-value > 1.96
-: t-value < -1.96
0: others

結果まとめ(情報構造)

Fixed Effect		SELF	FFT	FPT	SPT	RPT	Total
infostat=discourse-old	(vs. d-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		0	0	0	0	0	0
agentivity=neither		0	0	0	0	0	0
agentivity=patient	(vs. agent)	0	0	0	0	0	0
commonness=h-new	(vs. bridging)	+	0	0	0	0	+
commonness=h-old	(vs. bridging)	-	0	0	0	0	0
commonness=neither	(vs. bridging)	0	0	0	0	0	0

FFT, SPT
有意差なし

+: t-value > 1.96
-: t-value < -1.96
0: others

結果まとめ(情報構造)

Fixed Effect		SELF	FPT	RPT	Total
specificity=specific	(vs. either)	+	+	+	+
specificity=unspecific	(vs. either)	0	0	0	0
animacy=inanimate	(vs. animate)	0	+	0	0
sentience=insentient	(vs. either)	0	0	0	0
sentience=sentient	(vs. either)	0	+	+	0
commonness=h-new	(vs. bridging)	+	0	0	+
commonness=h-old	(vs. bridging)	-	0	0	0
commonness=neither	(vs. bridging)	0	0	0	0

・特定性: 特定 SELF, FPT, RPT, Total で時間がかかる
・有情性: 有情 RPT で時間がかかる
・共有性: 非共有 SELF, Total で時間がかかる
共有 SELF で早くなる

FPT: 無生 +
FPT: 有情 +
おそらく打ち消しあう

強いて言うなら
<無生, 有情> が遅い

おわりに

まとめ

- 読み時間と情報構造の対照分析

  非共有でブリッジングより読み時間が遅くなる
  →情報抽出・自動要約にユーザ適応などの応用

  有情で読み時間が遅くなる
  →日本語では有生よりも有情が重要な要素なのか

  特定で不特定より読み時間が遅くなる
  →受容者側で特定のものを思い浮かべるために負荷がかかる？

今後

- 節境界情報 [松本+ 2017]との対照比較
- 分類語彙表番号情報 [加藤+ 2017]との対照比較
- 読み時間データの拡充