テキストに対する印象評定データベース

浅原正幸・国立国語研究所
加藤祥・目白大学

はじめに

言語の生産実態の集積としての言語資源整備

- テキスト・音声の生産実態の集積によるコーパス構築
  - 書き言葉コーパス：　書籍・雑誌・新聞などの集積
  - 話し言葉コーパス：　読話・対話などの収録
  - ウェブコーパス：　ウェブテキストのクロール

生産実態の規範としてのデータ整備

- データ量：　大量のデータから多様な言語表現を収集
- 使用域：　メタデータによる好ましい言語表現の絞り込み

言語の解析モデルの規範としての言語資源整備

- テキスト・音声に対するアノテーション
  - 文字化：　音声書き起こし・漢字の包摂・誤字の記録
  - 形態論情報：　わかち書き単位・品詞・活用情報・レンマ（語彙資源との連携）
  - 統語情報：　ツリーバンク（句構造・依存構造）・述語項構造・節
  - 語義情報：　語義・語義の転換

解析系のための訓練データ→評価データ→言語学的な知見の符号化

言語の受容実態の集積としての言語資源整備

- 受容実態の集積
  - 読み時間：
    - 眼球運動　BCCWJ-EyeTrack
    - 移動窓方式自己ペース読文法　BCCWJ-SPR2, NIKKEI-SPR
  - 脳活動：
    - 固定窓方式（文字刺激）・音声刺激・動画刺激
  - 質問紙調査
    - 単語親密度（単語刺激）
    - 印象評定情報（単語刺激・文刺激）

テキストに対する印象評定データベース

研究の目的
読み時間を変化させる様々な要因

- 言語学的な要因
  - 文節単位に半角空白を入れると読み時間が短くなる
  - 係り受けの数が多いと読み時間が短くなる
  - 動詞 < 形容詞・副詞 < 名詞
  - 関係節内の関係　＜　関係節外の関係
  - 旧情報　＜　新情報
  - サプライザル（Perplexity）

- 表現上の要因
  - 自然さ
  - わかりやすさ
  - 古さ（obsolete）
  - 新しさ（innovative）
  - 比喩性

テキストに対する印象評定データベース


以下の表現について判定してください。

日程も保護者に希望を【募る】学校が多い。

1. 自然な表現ですか。
○ 0：まったく違う　○ 1
○ 2　○ 3
○ 4　○ 5：そう思う

2. わかりやすい表現ですか。
○ 0：まったく違う　○ 1
○ 2　○ 3
○ 4　○ 5：そう思う

3. 古い表現ですか。
○ 0：まったく違う　○ 1
○ 2　○ 3
○ 4　○ 5：そう思う

4. 新しい表現ですか。
○ 0：まったく違う　○ 1
○ 2　○ 3
○ 4　○ 5：そう思う

5. 何かを他の物事でたとえ（比喩）ていますか。
○ 0：まったく違う　○ 1
○ 2　○ 3
○ 4　○ 5：そう思う

【BCCWJ PN1a_00002 21290】

言語表現を示して
・自然さ
・わかりやすさ
・古さ
・新しさ
・比喩性
を 0-5 の 6段階評価により
クラウドソーシングを用いて評定（1表現 20人以上評定）

すでに収集したデータ

『現代日本語書き言葉均衡コーパス』
　文節単位：書籍コア 84736文節・教科書（国語） 50606文節
　　読み時間が付与されているもの
　長単位自立語：書籍コア 33546単語・雑誌コア 32037単語・新聞コア 27155単語
　短単位動詞：書籍コア 14236動詞・雑誌コア 13018動詞・新聞コア 10749動詞
　　係り受け・述語項構造・分類語彙表番号が付与されているもの
『日本経済新聞記事オープンコーパス』
　長単位自立語 11074単語・10628文節
　　係り受け・分類語彙表番号・読み時間が付与されているもの

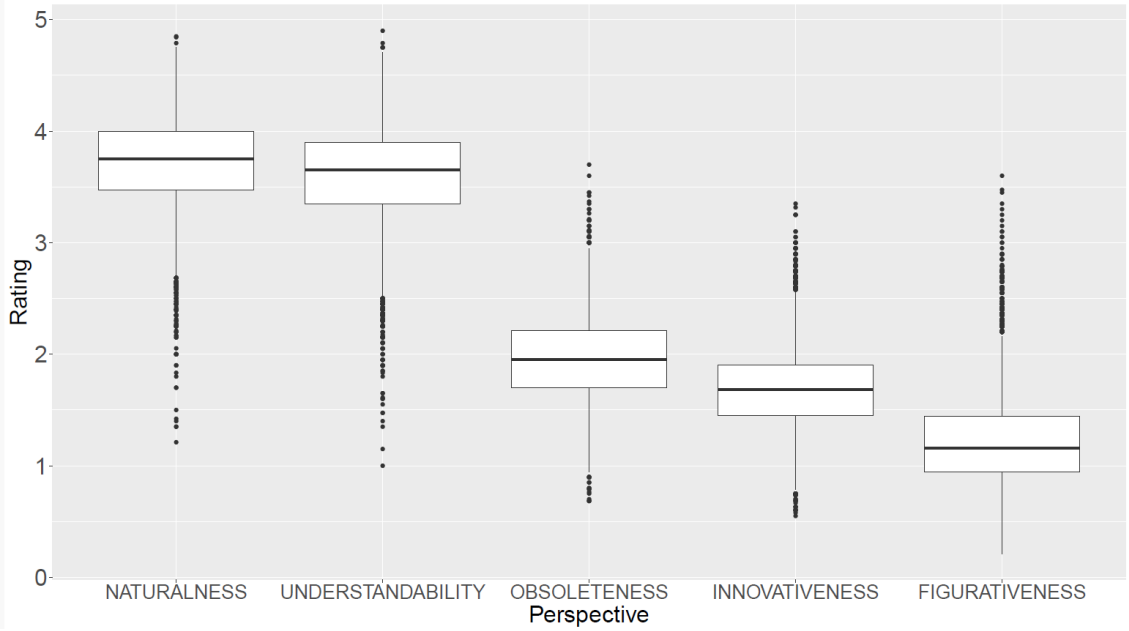
すでに収集したデータ

『現代日本語書き言葉均衡コーパス』の指標比喩	用例・結合　820件
『現代日本語書き言葉均衡コーパス』の結合比喩	結合（PB 2032，PM 3754，PN 5578）
『IPAL辞書』動詞・形容詞	用例　形容詞 2319件，動詞 5385件 最重要動詞 2806件，サ変 141件
『現代語の助詞・助動詞』	用例　535件
『日本語基本動詞ハンドブック』（接触動詞）	用例　1026件
『動詞の意味・用法の記述的研究』	用例・結合 4713件
『形容詞の意味・用法の記述的研究』	用例・結合 1438件
『比喩表現の理論と分類』	用例・結合 1799件

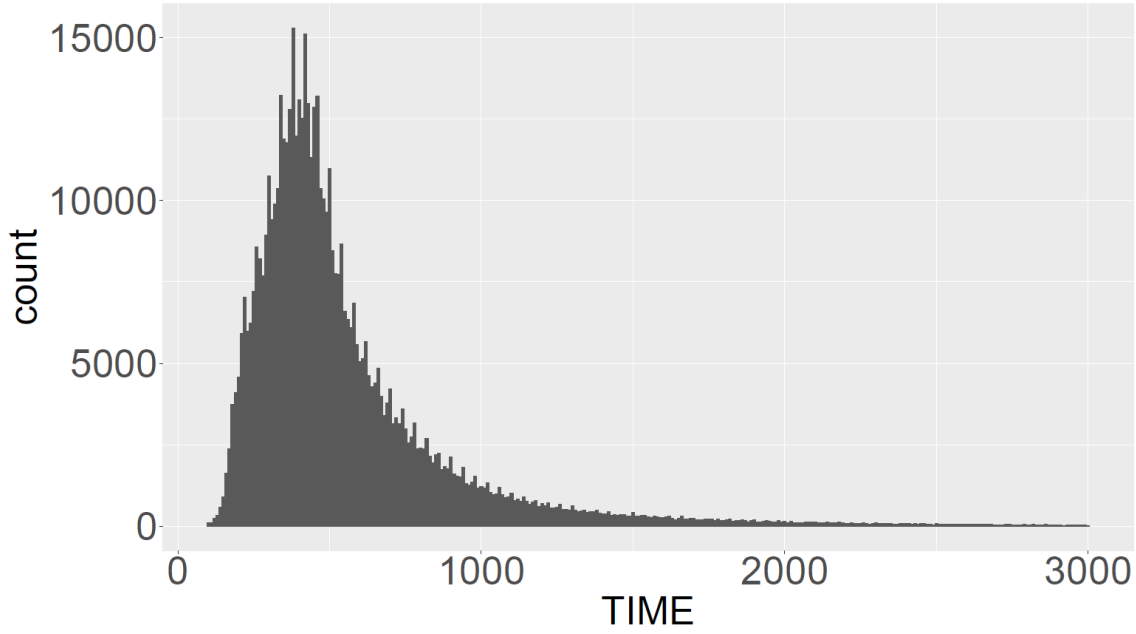
sample	sampleid	id	自然	わかりやすさ	古さ	新しさ	比喩性	sentence
PB	PB13_00021	8670	3.2	2.9	2	1.5	1.4	【自由度が】高く柔軟なエコシステムは、会社の元々のサイズも問わず、小さい規模で出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	8710	3.1	2.9	2.7	2.5	2	自由度が【高く】柔軟なエコシステムは、会社の元々のサイズも問わず、小さい規模で出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	8730	3.8	3.8	1.7	1.9	1.5	自由度が高く【柔軟な】エコシステムは、会社の元々のサイズも問わず、小さい規模で出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	8760	3.4	3.5	0.9	3.5	1.4	自由度が高く柔軟な【エコシステムは、】会社の元々のサイズも問わず、小さい規模で出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	8840	3.7	3.4	1.3	1.7	1.4	自由度が高く柔軟なエコシステムは、【会社の】元々のサイズも問わず、小さい規模で出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	8870	3.6	3.5	1.9	2.1	0.9	自由度が高く柔軟なエコシステムは、会社の【元々の】サイズも問わず、小さい規模で出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	8900	3.2	3	2.1	2	1.4	自由度が高く柔軟なエコシステムは、会社の元々の【サイズも】問わず、小さい規模で出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	8940	3.8	3.8	1.2	1.3	1	自由度が高く柔軟なエコシステムは、会社の元々のサイズも【問わず、】小さい規模で出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	8980	3.7	3.8	1.2	1	1	自由度が高く柔軟なエコシステムは、会社の元々のサイズも問わず、【小さい】規模で出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	9010	4	3.7	1.5	1.5	1.2	自由度が高く柔軟なエコシステムは、会社の元々のサイズも問わず、小さい【規模で】出発した会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	9040	4.2	4.4	1	0.8	0.9	自由度が高く柔軟なエコシステムは、会社の元々のサイズも問わず、小さい規模で【出発した】会社でも、速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	9080	3.8	2.9	1.2	1.3	1.3	自由度が高く柔軟なエコシステムは、会社の元々のサイズも問わず、小さい規模で出発した【会社でも、】速いスピードで業容を拡大できるといった大きなメリットがある。
PB	PB13_00021	9130	4.3	4.3	1.4	1.4	0.6	自由度が高く柔軟なエコシステムは、会社の元々のサイズも問わず、小さい規模で出発した会社でも、【速い】スピードで業容を拡大できるといった大きなメリットがある。

読み時間と印象評定情報
『日本経済新聞記事オープンコーパス』

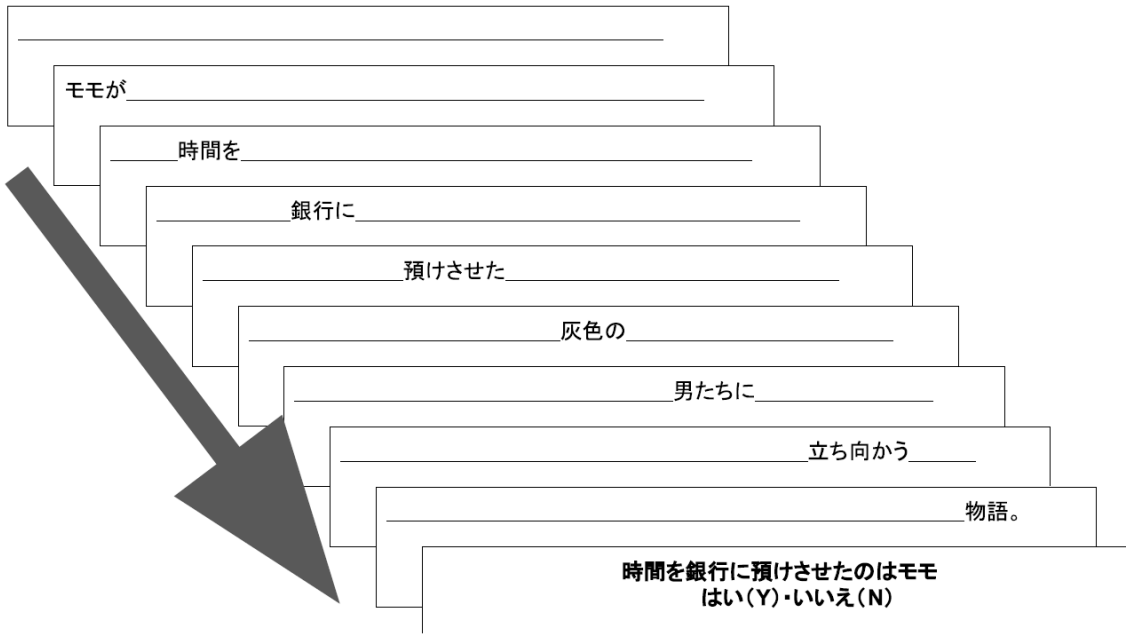
印象評定情報（文節単位）
(2023/03/15-17)



読み時間（文節単位）
(2023/03/01-15)



〔参考〕
『日本経済新聞記事オープンコーパス』の読み時間データ



自己ペース読文法による収集
【クラウドソーシング】
1記事あたり 50-200人分
異なり585人，延べ6828人分
782,475 データポイント

【データクリーニング】
内容把握回答が正しいもの
試行ごとの平均読み時間が 150ms < time < 2000ms のもの
データポイントごとの読み時間が
100ms < time < 3000ms のもの
→562,719 データポイント

読み時間と印象評定情報 分析結果

以下の線形式でモデル化
（線形混合モデル）

```
LOGTIME ~
 BIDC + CHARNUM + DEPNUM + NATURALNESS +
 UNDERSTANDABILITY + OBSOLETENESS +
 INNOVATIVENESS + FIGURATIVENESS +
 (1|SUBJ) + (1|ARTICLE)
```

Variable	Coefficient	Std. Error
BIDC	-0.001***	0
CHARNUM	0.045***	-0.0002
DEPNUM	-0.031***	-0.0004
NATURALNESS	-0.015***	-0.002
UNDERSTANDABILITY	-0.027***	-0.002
OBSOLETENESS	0.006***	-0.001
INNOVATIVENESS	-0.002*	-0.001
FIGURATIVENESS	-0.023***	-0.001
Constant	6.238***	-0.02

Observations	556,214
Log Likelihood	-183,728.50

読み時間と印象評定情報 分析結果

以下の線形式でモデル化
（線形混合モデル）

```
LOGTIME ~
 BIDC + CHARNUM + DEPNUM + NATURALNESS +
 UNDERSTANDABILITY + OBSOLETENESS +
 INNOVATIVENESS + FIGURATIVENESS +
 (1|SUBJ) + (1|ARTICLE)
```

読み時間を短くする要因:
自然さ，わかりやすさ，比喩性

読み時間を長くする要因:
古さ

Variable	Coefficient	Std. Error
BIDC	-0.001***	0
CHARNUM	0.045***	-0.0002
DEPNUM	-0.031***	-0.0004
NATURALNESS	-0.015***	-0.002
UNDERSTANDABILITY	-0.027***	-0.002
OBSOLETENESS	0.006***	-0.001
INNOVATIVENESS	-0.002*	-0.001
FIGURATIVENESS	-0.023***	-0.001
Constant	6.238***	-0.02

Observations	556,214
Log Likelihood	-183,728.50

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

今後の研究の方向性

- 結合（いわゆる格フレーム）の規範性・比喩性の分析
  結合同士の印象評定値による規範⇔比喩的の比較
  結合のみ見せた場合と文全体を見せた場合の印象の比較

- 言語生成における印象のパラメータ化
  「古い表現」もしくは「新しい表現」表現の出力
  「字義通りの表現」もしくは「比喩性をもつ表現」の出力