テキストアナリティクスと音声解析と認知科学と検索エンジン

日英対訳コーパス CORPORA

田淵 龍二（ミント音声教育研究所）
E-mail:　tabuchiryuji@nifty.ne.jp

2018 年 9 月 6 日 13:35-14:00
第 13 回 テキストアナリティクス・シンポジウム
成蹊大学（東京都武蔵野市吉祥寺）

あらまし

1 はじめに　アメリカのリーダビリティ公式を**音声**解析による認知科学的視点から再評価し、**作動記憶**（ワーキングメモリ）との関連が極めて濃いこと示してきた。

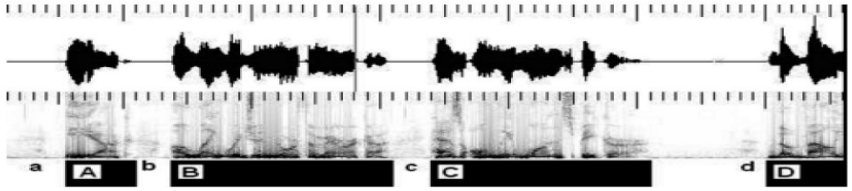


図 1. 連続音声の視覚化と呼気段落 [1]

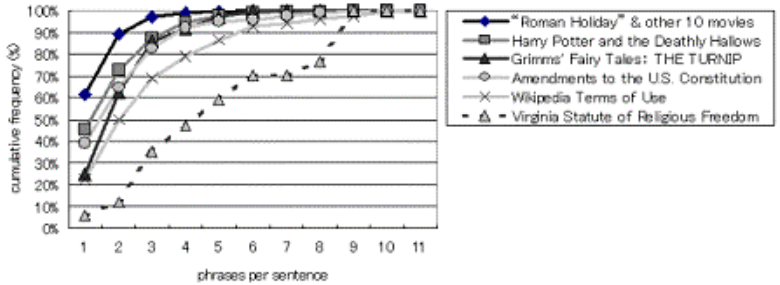


図 5. 句数で数えた文長ごとの累積 [6]

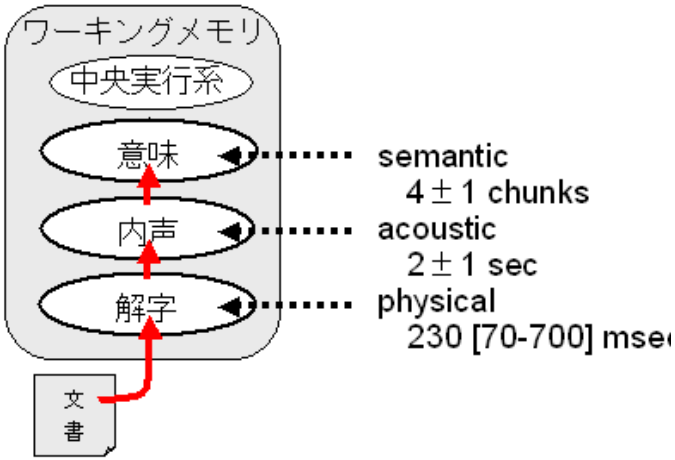


図 4. 読解プロセスの概要
（長期記憶は省略）[5]を改変

2 基礎研究 ： テキスト読解過程の認知科学的基盤、およびテキストから音声情報を抽出する方法。

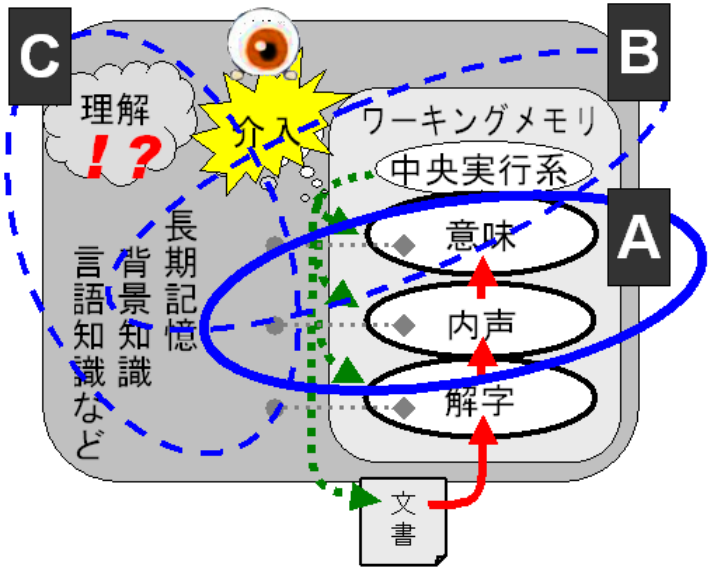


図 10. 読解プロセスとテキストアナリティクス

音韻符号化時間予測式

$$D = 120 \times Sy + 80 \times Cn \quad \cdots \quad ①$$

where

D: 時間（ミリ秒）

Sy: 音節数

Cn: 子音数

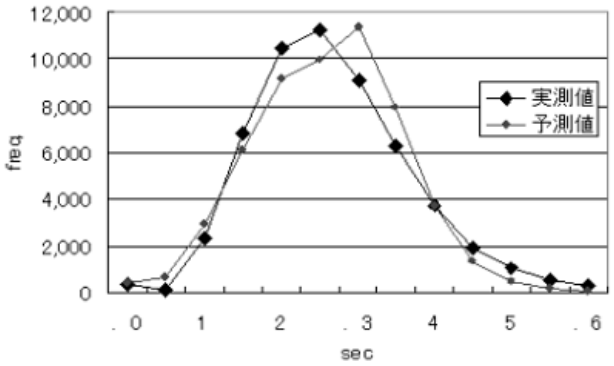


図 6. 字幕表示実測時間と字幕黙読予測時間の度数分布（n=54,253） [7]

3 応用　**適応学年**ごとに学習に最適なテキストや動画を選ぶ**検索エンジン**開発。

日英対訳コーパス
CORPORA（左）と
TED ビデオコーパス
selected360

4 まとめ　基礎研究から応用までの流れを紹介をすることで、単語や形態素以外の要素（音声）に注目したテキストアナリティクスの**多様な一面**を明らかにする。

キーワード　テキストアナリティクス，音声解析，認知科学，検索エンジン，リーダビリティ公式

日英対訳コーパス CORPORA

図 7. get away で検索した様子と QR コード

URL： http://www.mintap.com/talkies/pac/corpora.html

TED ビデオコーパス selected360

図 8. 文レベルを高 2 としたときの様子と QR コード

URL： http://www.mintap.com/talkies/?selected360

音声言語に埋め込まれていた法則の方程式

BG 長分布は対数正規分布に従う

$$f(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{1}{2}(\frac{ln(x)-\mu}{\sigma})^2}$$

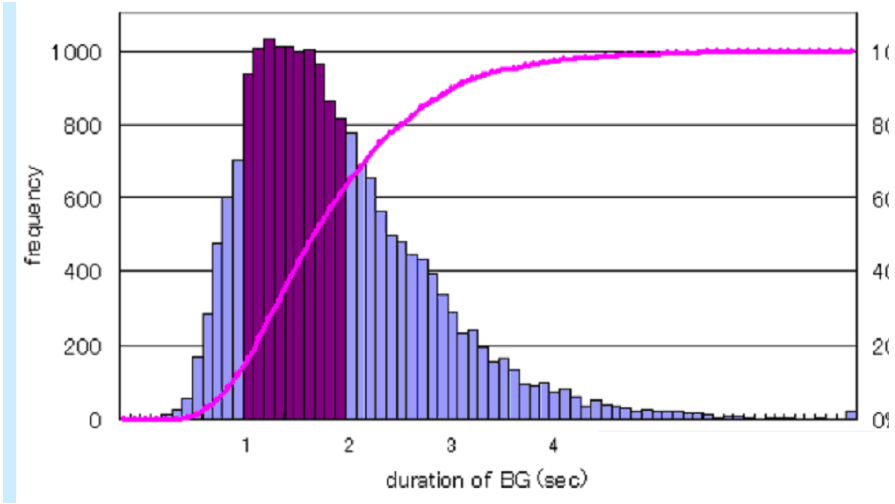


図 2. 呼気段落長の度数分布（n=19,551）[3] を改変

平均発話速度は対数分布に従う

$$wps = a \cdot \ln(w) + b$$

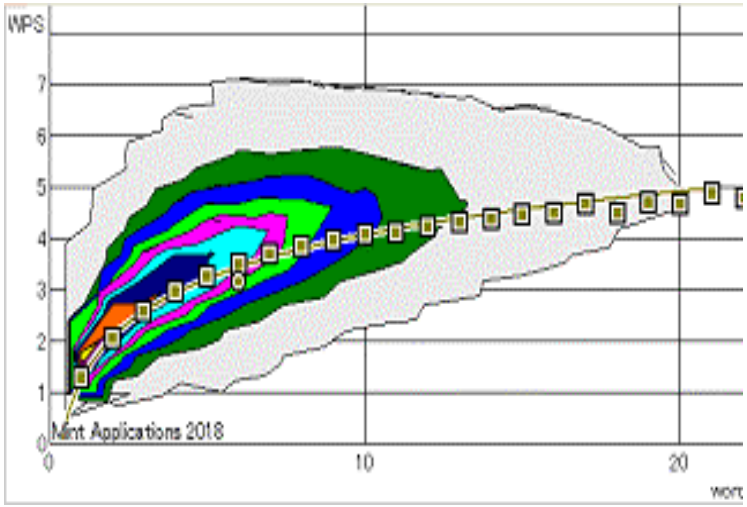
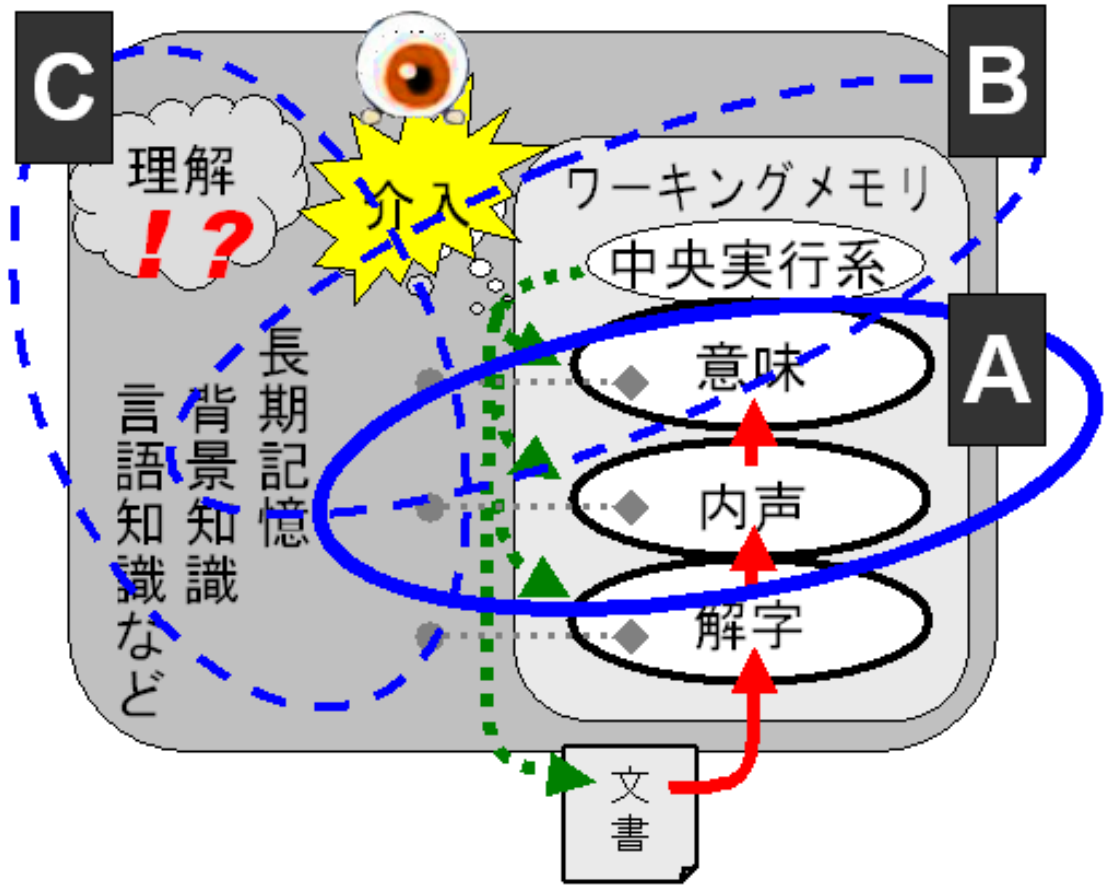


図 3. 単語数ごと発話速度の分布と平均値[3] を改変

読解プロセスとテキストアナリティクス



A 読みやすさ指標

B 形態素・構文解析と単語処理

C AI（人工知能）、アノテーション