様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１５　日現在

機関番号：　６２６１８
研究種目：　基盤研究(B)（一般）
研究期間：　2013～2016
課題番号：　２５２８４０８３

研究課題名（和文）言語コーパスに対する読文時間付与とその利用

研究課題名（英文）Development and Utilization of Reading Time Annotation on Text Corpora

研究代表者

浅原　正幸（Asahara, Masayuki）

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・コーパス開発センター・准教授

研究者番号：８０３７９５２８

交付決定額（研究期間全体）：（直接経費）　　12,900,000円

研究成果の概要（和文）：均衡コーパスに対して日本語母語話者24人分の読み時間を付与した。読み時間データに対して各種アノテーションを重ね合わせる作業を進めた。具体的には、統語情報としての係り受け・節境界情報、意味情報としての分類語彙表番号、談話情報としての情報構造アノテーションを重ね合わせた。
一般化線形混合モデルによる統計分析を行うことにより、次のような結果が得られた。たくさん係り受け関係を持つ文節の読み時間が短くなる現象が確認された。節境界において読み時間が短くなる現象が確認された。さらに体言よりも用言のほうが読み時間が長くなる傾向がみられた。

研究成果の概要（英文）：We annotated reading times of 24 Japanese native speakers. We overlaid several annotations on the reading time data. For example, syntactic dependency, clause boundaries, the category of `Word List by Semantic Principles', and information structure labels.
We got the following results by generalized linear mixed models: the subject participants read fast at the bunsetsu with many dependency, the clause boundaries.
It is also observed that the reading time of nominal phrase is longer than the verbal phrase.

研究分野：　心理言語学

キーワード：　読み時間

1．研究開始当初の背景

従来の視線走査法に基づく実験に基づく言語研究は作例に基づくものがほとんどであった。一方、コーパス言語学の分野では均衡コーパスが整備され、さまざまな統語・意味情報アノテーションが進められてきた。より自然な例文をコーパスからサンプリングし、読み時間を収集したうえで、アノテーションと比較する環境が整いつつあった。

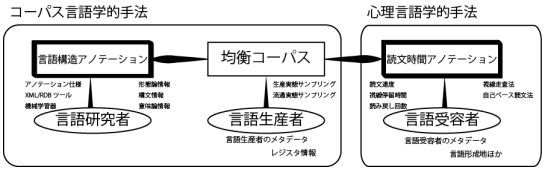
2．研究の目的

本研究では「現代日本語書き言葉均衡コーパス」に心理言語学的な手法で読文時間を付与する。人間の文処理過程を定量的に評価する読文時間情報を均衡コーパスに網羅的に付与し、心理言語学・コーパス言語学双方に寄与する言語資源を構築する。作成した言語資源と既存の言語構造アノテーションと重ね合わせることにより、既存の言語構造アノテーションの基準や仕様の妥当性を検証するとともに心理言語学で調査されている人間の文処理機構に対する仮説を検証する。さらに文章の可読性評価や言語教育研究などの応用研究に利用可能なコーパスコンコーダンサの設計・実装を行う。作成するコーパスコンコーダンサは読文時間情報と形態論情報・統語構造とを統合的に検索可能なものとする。

言語の生産過程の研究として、さまざまな出自のテキストデータを集積し、頻度・統計情報に基づき発見的に調査する方法がコーパス言語学の分野で用いられている。「現代日本語書き言葉均衡コーパス」(BCCWJ) などの均衡コーパスが整備され、人手や機械学習に基づく言語解析器により形態論情報・統語構造がアノテーションされ、頻度・統計情報を取るためのコーパスコンコーダンサが整備されてきた。しかし、アノテーションを行う過程も含めて、人間がコーパスを読む際に何が起きているのかを定量的に評価するような情報がコーパス上に残されることは極めて少なく、言語の受容過程研究に利用可能な言語資源は限られている。

一方、言語の受容過程の研究手段として、自己ペース読文法、視線走査法等の被験者実験に基づく手法が心理言語学の分野で用いられている。それらの多くが特定の言語現象に対する人間の文処理機構の仮説を証明するためのもので、作例に基づく刺激文を用いて実験が行われてきた。しかしこれら被験者実験は仮説の証明に重点が置かれ、刺激文作成が高コストであるうえに、扱われている構文の幅が非常に限定されているために再利用性が低く、言語資源として蓄積がなされていない。

我々はコーパス言語学的手法と心理言語学的手法を結びつけることで、言語の生産過程研究・受容過程研究の双方に寄与する再利用性の高い言語資源構築を試みる。日本語に対する初めての試みであり、心理言語学での幅広い研究が期待できる。言語情報アノテーションの基準・仕様の妥当性検証に役立たせることができ、言語情報アノテーションはどうあるべきかという本質的な問題に迫る第一歩となろう。さらに、本研究では、連続量である読文時間と離散量である言語構造の頻度情報を統合的に検索可能なコーパスコンコーダンサを構築するとともに、本データを利用した多角的な言語研究方法について検討する。



3．研究の方法

(1) 読文時間を付与したコーパスの作成を主目的とする。均衡コーパスに対して、心理言語学的な手法に基づいた被験者実験による読文時間を網羅的に付与することにより、言語の受容過程の実態を定量的に評価することが可能な大規模な研究資料を構築する。『現代日本語書き言葉均衡コーパス』の新聞データ 20 ファイルを用いて呈示文書を準備し、日本語成人母語話者のみを対象とし、読文時間を取得する。読文時間は視線走査法と自己ペース読文法の 2 つの方法で付与する。視線走査法は視線走査装置を用い、画面に文を呈示した際に、被験者の視線が停留している個所と時間を直接評価する方法である。本研究では国立国語研究所にある Eyelink CL を利用する。

(2) 読文時間情報と形態論情報・統語構造を統合して検索可能なコーパスコンコーダンサを開発する。具体的にや ChaKi.NET の機能拡張により実現する。

(3) 読文時間情報と既存の言語構造アノテーションを重ね合わせることによりコーパスアノテーションの基準・仕様の妥当性を検証する

(4) 心理言語学の分野で調査研究されている人間の文処理機構についての仮説を作成した研究資料を用いて検証する。またそのために必要なデータの統計処理手法を確立する

4．研究成果

研究作業者を雇用し、読み時間データに対して各種アノテーションを重ね合わせる作業を進めた。具体的には、統語情報としての係り受け・節境界情報、意味情報としての分類語彙表番号、談話情報としての情報構造アノテーションを重ね合わせた。

一般化線形混合モデルによる統計分析を行うことにより、次のような結果が得られた。たくさん係り受け関係を持つ文節の読み時間が短くなる現象が確認された。同様に節境界についても読み時間が短くなる傾向があり、節ラベルで分析すると 並列節>>名詞修

飾節>>補足節，副詞節の順に短くなることが確認された。分類語彙表番号においては、統語的な性質を表す「類」に関して「体」>>「その他」 >> 「相」 >> 「用」と読み時間が短くなる傾向が確認された。意味的な分類を表す「部門」に関しては、「生産物」>「自然」 >「関係」　「主体」>「活動」という傾向がみられた。情報構造においては、特定性・有情性で読み時間が長くなる傾向がみられた。また、新情報と想定可能(Bridging)・旧情報が判別可能なレベルで読み時間に差があることが確認された。
得られた知見について、係り受けについてはCOLING-2016 で発表を行った。情報構造については 2016 年度内に言語処理学会年次大会で発表を行った。ほかのものについても 2017 年度中に対外発表を行う。
また、統計処理手法として、一般化線形混合モデルではなく、Bayesian 線形混合モデルに基づく分析を進めた。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）
1. 浅原正幸・加藤祥，“文書間類似度について”，自然言語処理，(23)5，463-498，(2016) [査読あり]

〔学会発表〕（計８件）
1. 浅原正幸，“読み時間と情報構造について（ちょっとみじかめ）”，言語処理学会第 23 回年次大会，2017 年 3 月 16 日，筑波大学（茨城県つくば市）
2. 浅原正幸・小野創・宮本エジソン正，“『現代日本語書き言葉均衡コーパス』の読み時間とその被験者属性”，言語処理学会第 23 回年次大会，2017 年 3 月 15 日，筑波大学（茨城県つくば市）
3. 宮内拓也・浅原正幸・中川奈津子・加藤祥，“『現代日本語書き言葉均衡コーパス』への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，2017 年 3 月 15 日，筑波大学（茨城県つくば市）
4. 浅原正幸，“読み時間と情報構造について（ちょっとながめ）”，言語資源活用ワークショップ 2016，2017 年 3 月 8 日，国立国語研究所　（東京都立川市）
5. 宮内拓也・浅原正幸・中川奈津子・加藤祥，“『現代日本語書き言葉均衡コーパス』への情報構造アノテーションの分析”，言語資源活用ワークショップ 2016，2017 年 3 月 8 日,国立国語研究所（東京都立川市）
6. Masayuki Asahara, Hajime Ono, Edson T. Miyamoto, “Reading Time Annotation for `Balanced Corpus of Contemporary Written Japanese”, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, 2016 年 12 月 13 日，大阪国際会議場（大阪府大阪市）
7. Masayuki Asahara, Yuji Matsumoto, “BCCWJ-DepPara: A Syntactic Annotation Treebank on the `Balanced Corpus of Contemporary Written Japanese’, Proceedings of the 12th Workshop on Asian Language Resources (ALR12), 2016 年 12 月 12 日，大阪国際会議場（大阪府大阪市）
8. 浅原正幸・小野創・宮本エジソン正，“BCCWJ-EyeTracking -- 『現代日本語書き言葉均衡コーパス』に対する読み時間アノテーション”，電子情報通信学会 IEICE-TL，2016 年 7 月 23 日，早稲田大学西早稲田キャンパス（東京都新宿区）

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計 0　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計 0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
浅原　正幸（Masayuki Asahara）
人間文化研究機構国立国語研究所・コーパス開発センター・准教授
研究者番号：80379528

(2)研究分担者
小野　創（Hajime Ono）
津田塾大学・学芸学部・准教授
研究者番号：90510561

宮本エジソン正（Miyamoto Edson Tadashi）
筑波大学・人文社会科学研究科・准教授
研究者番号：60335479

(3)連携研究者
なし

(4)研究協力者
なし