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：　挑戦的萌芽研究
研究期間：　2015～2016
課題番号：　１５Ｋ１２８８８

研究課題名（和文）近代語コーパスに対する統語情報アノテーションの基準策定

研究課題名（英文）Design Guidelines of Syntactic Annotation Schema on Japanese Modern Language

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）　　2,600,000円

研究成果の概要（和文）：本研究では、近代語コーパスに対する統語情報アノテーションの仕様策定を行った。具体的には、文節係り受け・並列構造・述語項構造アノテーションを明六雑誌６サンプルに対して行い、問題点を明らかにした。このデータに基づき Universal Dependency に適応させたデータを開発した。同内容は2017年9月に開かれる国際会議 JADH-2017 で発表予定である。また、意味情報を含む他のレベルのアノテーションとして、節境界情報・分類語彙表番号アノテーションを試行的に行った。

研究成果の概要（英文）：We designed guidelines of syntactic annotation for Japanese modern languages. We performed annotation of syntactic dependency structures, coordinate structures, and predicate-argument structures on Meiroku Zasshi magazines 6 samples and explored the issues on the annotation procedures. We also performed annotation of clause boundaries and the label of `Word List by Semantic Principles'.

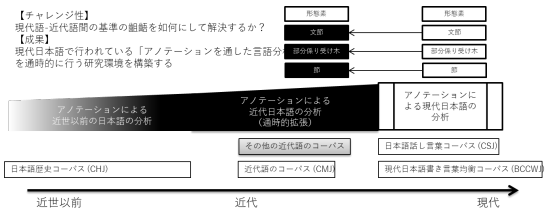
研究分野：　コーパス言語学

キーワード：　アノテーション　近代語

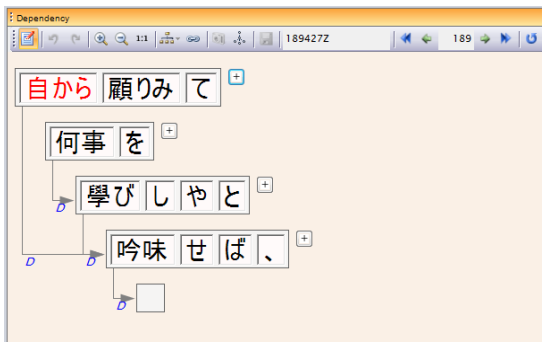
１．研究開始当初の背景

国立国語研究所において、現代日本語のコーパスとして「日本語話し言葉コーパス(CSJ)」「現代日本語書き言葉均衡コーパス(BCCWJ)」、近代以前の日本語のコーパスとして「近代語のコーパス(CMJ)」「日本語歴史コーパス(CHJ)」が整備されている。前者の現代日本語のコーパスに対しては形態論・統語論・意味論などのアノテーションが進められている一方、後者の近代以前のコーパスに対しては形態論レベルのアノテーションに留まっている。また、形態素解析用辞書として近代文語 UniDic・中古和文 UniDic が開発されたが、係り受け解析器などは整備が進んでいない。

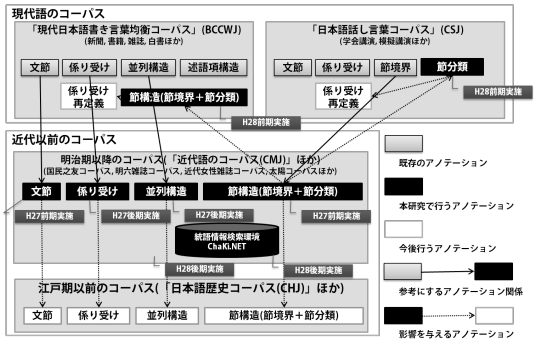
２．研究の目的



近代語コーパス(主に明治期／文語体・言文一致体~ 口語体) に対して統語情報分析のためのアノテーションを行うことを目標として、通時的に適用可能な統語情報（文節・係り受け・並列構造・節構造）アノテーション基準を策定することを研究目的とする。少量の近代語コーパスに対して実際にアノテーションを実施し、既に一部統語情報（文節・係り受け・並列構造・述語項構造）が付与されている現代語コーパスに対するアノテーションと対照分析を行いながら、アノテーション作業の要件定義を行う。作成したアノテーションデータをコーパスコンコーダンサ ChaKi.NET で検索可能にする。



３．研究の方法



「近代語のコーパス」(CMJ) やその他の明治期の共有可能なコーパスから、文語体のサンプルを抽出し、文節・係り受け・並列構造・節情報をアノテーションしながら、日本語に対して通時的に適用可能なアノテーション基準を策定する。随時、現代語のコーパス「現代日本語書き言葉均衡コーパス」「日本語話し言葉コーパス」のサンプルにも適用し、問題点を適宜解消しながら、多様な日本語表現を被覆する基準策定を行う。

近代語における文節・節情報に関する基準を確立し、上記近代語コーパスに対して、アノテーションを試みる。文節については BCCWJ, CSJ の長単位・文節の設計に基づき、近代語に対する文節アノテーション基準を確立する。

節構造については、節のスコープを新たにアノテーションし、節の意味分類を規定する。南の ABC 類・野田の節分類を付与して、表層格「ハ」と「ガ」の係り先を規定するだけでなく、節の統語的性質（連用節・連体節・引用節・並列節など）や意味的性質（理由節・条件節・時間節・目的節・逆接節など）や時制節性などを付与する。

、BCCWJ のアノテーション基準に基づき、近代語における節内の係り受け・並列構造アノテーション基準を確立する。BCCWJ の係り受けアノテーションから学習した係り受けモデルを用い、近代語を解析したものを修正することにより工数削減をはかる。

また現代語コーパス(BCCWJ, CSJ) の一部(5 万語規模) に対して節情報のアノテーションを試み、既存の係り受けアノテーションの問題点を検証する。

近代語に対する各アノテーションを一般公開する。全ての統語情報アノテーションを ChaKi.NET で可視化し、検索可能な環境を構築する。

４．研究成果

係り受け・述語項構造アノテーションについては、昨年度作成した明六雑誌データ６記事を統合し、コーパス管理システム『ChaKi.NET』で可視化する環境を構築した。同内容は日本語学会 2016 年度秋季大会で発表した。年度末に同データの Universal Dependency 対応を進めた。2017 年度中に対外発表を行う予定である。

節境界アノテーションについては、『現代日本語書き言葉均衡コーパス』の新聞記事データ 54 ファイルに対して、悉皆付与し、言語資源活用ワークショップ 2016 でアノテーション仕様について発表した。2017 年度に開かれる国際会議に投稿予定である。近代語の節境界アノテーションについては、問題の分析にとどまったが、今後現代語のデータに基づき、アノテーションを進める。

また、試行的に通時的に意味情報アノテーションが可能かどうかを検証するために、『現代日本語書き言葉均衡コーパス』『日本語歴史コーパス』に対して『分類語彙表』番号付与の検討を行った。具体的には、現代の新聞記事・狂言データ・竹取物語について、作業を行った。この作業環境を進めるために、形態素解析用辞書 UniDic と分類語彙表番号対応表を用いた、自動ラベリング環境を整えた。

さらに、作成した係り受け・述語項構造アノテーションの Universal Dependency 対応を進めている。

作業の全体について「通時コーパス」シンポジウム 2017 において、「『日本語歴史コーパス』に対する統語・意味情報アノテーション」というタイトルで発表を行った。今後、本挑戦的萌芽研究 15K12888 (H27-H28)の成果に基づき、基盤研究(A) 17H00917 (H29-H33)「日本語歴史コーパスに対する統語・意味情報アノテーション」を進めていきたい。

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 0 件）

〔学会発表〕（計 4 件）

1. 浅原正幸,“『日本語歴史コーパス』に対する統語・意味情報アノテーション”,「通時コーパス」シンポジウム 2017, 2017 年 3 月 11 日,国立国語研究所 （東京都立川市）
2. 松本理美・浅原正幸・有田節子,“『現代日本語書き言葉均衡コーパス』に対する節の意味分類情報アノテーション---基準策定，仕様書作成の必要性について---，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）
3. 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）
4. 浅原正幸・髙橋雄太,“近代語コーパスに対する統語アノテーション基準の検討”，日本語学会 2016 年度秋季大会，2016 年 10 月 30 日，山形大学(山形県山形市)

〔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

中田節子（Setsuko Nakada）
立命館大学・言語教育情報研究科・教授
研究者番号： 70263994

(3)連携研究者

なし

(4)研究協力者

なし