様式J-Ⅰ

2016 年 9 月 30 日

博報財団 第 10 回「国際日本研究フェローシップ」成果報告書

Ⅰ. 研究成果概要

※定型フォーマット有（A4 1～2 枚） ※全て日本語で作成

氏名（在住国名）	HORN Stephen Wright（ホーン スティーブン ライト） （ イギリス（アメリカ） ）
所属・役職	無所属（元 オックスフォード大学 東洋学部・助教授）
招聘回（招聘研究期間）	第 10 回 （2015 年 9 月 1 日～2016 年 8 月 31 日）
受入機関	国立国語研究所
招聘研究テーマ	「近世以前の日本語の通時コーパスの統語情報付加：言語学研究の実用に向けて」
研究目的	● オックスフォード上代語コーパス（OCOJ） を構築した経験を生かして、国立国語研究所の歴史コーパス万葉集編（CHJ万葉集）のデータに構成素構造などの統語論情報を表すモデルを設計し、実際のデータの一部分に適用し、歴史言語学に役立つコーパスの開発に貢献するのが主な目的。上代語の共時的分析だけに留まらず、通時的研究に応用できるように、平安時代の仮名文学作品にも適用する。国語研のコーパスプロジェクト提案として、二つのサブコーパスとその規定集を纏める。 ● コーパスを使って上代語についての基本研究も行う。 ● コーパス検索ツールの開発を検討。 ● 今までできなかった照応関係、スコープ現象などのアノテーションの可能性を探る。

研究概要：

- 限量表現・スコープ現象などを記述するためのアノテーションモデルを設計：主に万葉集巻第一と巻第二に試行版として適用（97 実例）。
- 疑問詞の解釈についての上代語・平安時代日本語の比較のための基本研究（ビャーケ フレレッスビッグ教授と共同で）：上代語の疑問詞を網羅して採取済み、480 実例の分析（作業中）。
- 上代語の複合動詞の基本研究（国立国語研究所プロジェクト非常勤研究員 鴻野 知暁）：OCOJの検索の依頼を受けて、結果を Excel ファイルで出力
- 上代語のフレーズ投射についての基本研究（柳田優子とジョン・ホイットマン）：OCOJの検索の依頼を受けて、結果を Excel ファイルで出力
- Chaki.NET のラベル、セグメント。グループの種類の追加とその定義 （「CHJ 万葉集統語アノテーション」参照）
- CHJ 万葉集の統語論情報サブ・コーパス（第一から巻第四、792 首）の構築：1）係・受関係の修正、2）構成素の分析、3）文法役割の付与、4）特殊の文法構造の指定、4）照応関係、呼応関係、その他。GUI として Chaki.NET を使用。出来上がったデータが.Cabocha 拡張形式のファイル。合わせて 5,555 句、万葉集全体の 21.23%. 係・受ラベルで検索可能。
- 「CHJ 万葉集統語アノテーション」：万葉集巻第一から巻第四のサブコーパスの構築を詳しく説明する規定集。文法的振る舞いの難しい語彙についての取り扱い方、特殊な文法構造や助詞の品詞についての分析も含まれている。（63 pp.、作成進行中）
- 伊勢物語の統語論情報サブ・コーパス（第一話から第39話）の構築：1）係・受関係の修正、2）構成素の分析、3）文法役割の扶養、4）特殊の文法構造の指定、4）照応関係、呼応関係、その他。出来上がったデータが.Cabocha 拡張形式のファイル。
- 「CHJ 伊勢物語ドキュメンテーション」：上代語のサブ・コーパスについての「CHJ 万葉集統語アノテーション」が

カバーしなかった、平安時代の散文の独特な構文などについての解説（作成進行中）。

- CHJ 万葉集の統語論情報サブコーパスの XML 形式ファイル：国立国語研究所のプロジェクト非常勤研究員岡照晃と共同に作成した Cabocha 拡張形式から XML 形式への変換。XPath や XQuery で形態素解析情報と統語論分析情報を組み合わせて複雑条件で検索可能(「CHJ 万葉集統語アノテーション」: Constituent Analysis 参照)。
- 上代語の名詞句内構造についての基本研究：助詞の分布の記述と分析 （Annotation manual for the NINJAL Parsed Corpus of Old Japanese, Chapter 7 参照）。
- NINJAL Parsed Corpus of Old Japanese （試行版）：Cabocha 拡張形式の CHJ 万葉集データを万葉集巻第1.1から万葉集巻第 2.97 までを Penn Historical Treebank の樹形図形式になるようにアノテーションを施す。量は762 句で万葉集全体の 02.91%. フレーズマーカーなら、Tregex で検索可能。XPath や XQuery で形態素解析情報と組み合わせて複雑条件で検索可能。
- Annotation manual for the NINJAL Parsed Corpus of Old Japanese, 英語版 pp.1–35（作業進行中）、日本語版 pp.1–29（作業進行中）。
- OCOJ を国立国語研究所へ移動するための交渉が成立、今年度以内のオンライン公開の準備に着手。
- 学術論文の出版：
  - Horn, Stephen Wright; Russell, Kerri L. “The Oxford Corpus of Old Japanese.” 共著（2 人） In『コーパスと日本史』. ひつじ書房. 177–195, 2015.
  - Frellesvig, Bjarke; Horn, Stephen Wright; Yanagida, Yuko. “A Diachronic Perspective on Differential Object Marking in Pre-modern Japanese: Old Japanese and Early Middle Japanese.” In Seržant, Ilja A; Witzlach-Makarevich, Alena; Mann, Kelsey (eds.) *The Diachronic Typology of Differential Argument Marking*. (Series: Studies in Diversity Linguistics) Language Science Press. In production.
  - Maruyama, Takehiko; Horn, Stephen Wright; Russell, Kerri L.; Frellesvig, Bjarke. “Multiple clause linkage structure in Japanese.” *Kobe Journal of Japanese Studies*, No.1. Universit Ca’ Foscari Venezia. In press.
- 口頭発表：
  - 「オックスフォードの上代語コーパス（OCOJ）について一万葉集の CHJ との比較を中心に」.「歴史コーパス科研」上代班・和歌班の研究会、奈良先端科学技術大学院大学（NAIST）、2015 年１１月７日.
  - 「万葉集コーパスの統語情報付与」, 137 回 NINJAL サロン, 国立国語研究所, 2016 年 3 月 1 日.

展望：

- Chaki.NET のインタフェースが不安定で、操作が限られている上、セグメントやグループの検索ができないことや、Cabocha 拡張ファイルの変換が煩雑なことなどを踏まえて、「CHJ 万葉集の統語論情報サブコーパス」の作成方針を変えることにした。国立国語研究所で行われている「統語・意味解析コーパスの開発と言語研究」（プロジェクトリーダー：プラシャント・パルデシ）の下でプロジェクト非常勤研究員のアラステア・ジェームズ・バトラー氏と共同で NINJAL Parsed Corpus of Old Japanese の試行版を Penn Historical Treebank の形式で作ったところ、言語学分析に役立つ統語論情報を表すための装備と、CHJ の形態素解析情報との紐付けを保ちながらツリー構造の位置づけやフレーズラベルで共時的分析を貫く方法論などに大きな可能性があるとみる。歴史言語学に広く使われたコーパスモデルなので、学界への大きいインパクトも期待できる。CHJ との互換性も高く、NINJAL Parsed Corpus of Old Japanese のモデルで CHJ のデータを多量処理していくことを薦めると同時に、自分の力で何とか追及するつもりでもある。万葉集のデータを全部部コーパス化できたら、世界有数のコーパスになると確信している。
- OCOJ が国立国語研究所の管理の下で公開される計画も大変有意義。古代語特殊音声と、体系的形態素分析と、統語論情報の備わった OCOJ が日本の学界で活用されるようになると、言語学ばかりでなく、国語学や国語教育にも大きな貢献ができるし、その分析を CHJ のデータに直接重ね合わせることが不可能だったにしても、その作り方の精神がこれからのコーパス構築の見本となることを望む。