様式 C-7-2

自己評価報告書

平成　22 年　5 月　21 日現在

研究種目：　若手研究（B）
研究期間：　2007～2010
課題番号：　１９７２０１００
研究課題名（和文）　大規模テキストデータベースを用いた
フィンランド語の形態・統語情報のサンプル化
研究課題名（英文）　Building a Balanced Database of the Finnish Morpho-syntax Using Large Corpora of Finnish

研究代表者
千葉　庄寿（CHIBA Shoju）
麗澤大学
研究者番号：70337723

研究分野：人文学
科研費の分科・細目：言語学・言語学
キーワード：統語論・コーパス言語学

１．研究計画の概要

(1) 本研究は，電子化された大規模な言語コーパスから得られる大量の用例を用いて文法形式や文法構造の特徴を記述する際に不可欠な，「評価資料」となる言語データのサンプル化の方法論を整備することを目的としている。

(2) 具体的には，フィンランド語の大規模コーパス (フィンランド学術計算機センター提供の「フィンランド語バンク」) を用い，フィンランド語の語彙・文法情報をデータベース化し，(1) 文法情報の種類と統計分析に必要なサンプルサイズの関係の考察 (2) 文法情報のデータ構造の検討と最適な統計処理の手法の開発 (3) 最新のコーパス言語学で用いられるコロケーション分析の手法をはじめとする語彙分析とサンプル化による文法情報の分析の関係の考察をおこなうものである。

２．研究の進捗状況

(1) サンプル化したコーパスデータを格納する形態・統語情報データベースの仕様を平成 21 年度までに確定させた。サンプルデータベースへの文構造・談話構造レベルの情報付与についても，実装にむけ予備的な検証をおこなった。

(2) 使用域別に「フィンランド語バンク」のサンプル化作業を進め，サンプル抽出されたデータについて解析ツールを用いて形態・統語解析をおこない，データベースに格納した。

(3) 形態・統語解析結果の検証と校正作業を最終年度も引き続きおこない，データの精度向上をはかりたい。

(4) 構築の済んだサンプルについて，抽出したパラメータの量的情報を構文分析に利用する試みを開始し，得られたいくつかの知見について研究発表をおこなっている。ただし，平成 21 年度までに実施した分析で使用しているデータは限定的であり，精度も本研究が目標とするよりも粗いレベルにとどまっている。

(5) サンプル化されたデータベースの統計処理の手法に関する予備的な分析を開始した。その結果，分析対象の文法パラメータのサンプル数が非常に多い場合，そのパラメータを含む構文の分布特徴を評価し，記述するための適切な統計指標がないことが明らかになった。

３．現在までの達成度

(1) サンプル化とデータベース構築の作業はほぼ当初の計画どおりに進展している。最終年度でデータベースの構築を完了することを目標に引き続き作業を進める。

(2) 大規模コーパスデータの格納に必要なストレージ機器が当初予定されていたスペック(複数 OS のサポートおよび複数台の PC の同時アクセス)を満たさなかったため，Windows 環境によるデータベースシステム構築を余儀なくされている。Windows GUI によるアプリケーション開発に切り替えることで作業の遅延を最小限にとどめており，サンプル化データを用いた最終年度の分析

作業には支障がない見込みである。

(3) 現在最大の問題となっているのが，方法論的に最も未開拓であった，文法形式の出現環境の記述に必要な大規模データの解析手法の開発である。今後，統計学およびコーパス言語学の専門家と情報交換をおこない問題点を整理するとともに，統計処理に関する基礎研究の蓄積を積極的におこない，研究期間内で一定の方向性を見いだしたい。

４．今後の研究の推進方策

(1) データベースの構築および解析データの検証・校正を引き続きおこなう。

(2) データベースを用いて本格的な文法項目の解析と分析をおこなう。検証・校正が済んだデータを随時分析対象に追加していく。

(3) 平成 22 年 8 月にハンガリーPázmány Péter カトリック大学において開催される第 11 回国際フィン・ウゴル学会において本研究の研究成果を発表する。

(4) サンプル化の手法を用いた言語分析の方法論とサンプル化の具体的な作業過程，ならびに構築完了分のデータベースを用いた分析について報告書を作成する。

(5) データベースの公開方法については，本研究が依拠するフィンランド語コーパス「フィンランド語バンク」の運用をおこなっているフィンランド学術計算機センターの担当者との協議を続け，データベースの恒常的な運用と研究者間の共有にむけ引き続き調整をおこなっていく。

５． 代表的な研究成果

〔雑誌論文〕（計 3 件）

① 千葉庄寿,「フィンランド語記述文法とコーパスデータの役割」, 英語コーパス研究, 15，17-32，2008，査読有
② 千葉庄寿,「アノテートされた大規模コーパスを用いた分析ツールの現状と今後の方向性」,「ロシアおよびその周辺の少数言語のコーパスの構築と記述的・歴史的研究」研究成果報告書，55-70，2009，査読無
③ 千葉庄寿,「コロケーション，コリゲーションと携帯統語情報―類型論的観点から―」,「代表性を有する書き言葉コーパスを利用した日本語教育研究」研究成果報告書，91-107，2010，査読無

〔学会発表〕（計 5 件）

① 千葉庄寿,「フィンランド語記述文法とコーパスデータの役割」, 英語コーパス学会第 30 回大会，2007 年 10 月 6 日，立教大学池袋キャンパス
② 千葉庄寿,「コリゲーションの抽出における形態統語情報の役割」, 言語処理学会第 12 回年次大会，2008 年 3 月 20 日，東京大学駒場キャンパス
③ 千葉庄寿,「大規模コーパスの語彙統計情報の利用を支援する―語彙情報データベースを参照する API の構築と活用―」, 特定領域研究「日本語コーパス」平成 20 年度公開ワークショップ，2009 年 3 月 15 日，東京工業大学
④ 千葉庄寿,「フィンランド語の許可構文に現れる不定詞について―大規模コーパスにもとづく分析試論―」，第 36 回ウラル学会研究発表大会，2009 年 7 月 11 日，京都産業大学
⑤ 千葉庄寿,「大規模コーパスを用いたフィンランド語の分析的使役構文の語彙的・文法的特徴の記述」，日本言語学会第 139 回大会，2009 年 11 月 28 日，神戸大学

〔図書〕（計 0 件）

〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕

フィンランド・ヘルシンキ大学人文学部報「人文学者紹介」に研究に関する紹介が掲載されている（フィンランド語)。
http://www.helsinki.fi/hum/humanisti/2010/0310.htm