様式 C-19

# 科学研究費補助金研究成果報告書

平成２３年　６月　１日現在

機関番号：32506
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・外国語学部・准教授
研究者番号：70337723

研究成果の概要（和文）：大規模なフィンランド語コーパスからサンプルを抽出し，詳細な言語学的情報を付与したデータベースを構築し，フィンランド語の諸構文の用例群が示す量的情報の有意性をサンプルデータベースと比較して評価する分析手法を整備した。

研究成果の概要（英文）：Using sampling techniques, about 10 million-sized textual database was extracted from the large corpora of written modern Finnish, and then linguistically annotated. Annotation ranges from lexical, grammatical to discourse-functional information. We also developed a quantitative profiling method alongside practical applications which compares the syntactic/morpho-syntactic/lexical profiles of Finnish grammatical constructions with the overall settings of the sampled database.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2007 年度 | 1,700,000 | 0 | 1,700,000 |
| 2008 年度 | 500,000 | 150,000 | 650,000 |
| 2009 年度 | 500,000 | 150,000 | 650,000 |
| 2010 年度 | 700,000 | 210,000 | 910,000 |
| 総　計 | 3,400,000 | 510,000 | 3,910,000 |

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・形態論・コーパス言語学

## １．研究開始当初の背景

(1) コンピュータの普及と高速化に伴い，電子化されたテキスト資料 (コーパス) を文法分析に利用している研究事例が急速に増えている。しかし，文法研究において，コーパスから得られたデータが文法記述の枠組み全体を変えるに至るような事例はこれまで殆どなく，多くはデータを「単なる」用例として，つまり既存の理論や仮説を検証する「質的な」証拠として引き合いに出しているに過ぎない。これは，コーパスから抽出された用例群がもつ数量的な情報(頻度など)がその言語本来の「ありよう」からどの程度逸脱しているかを厳密に評価するための基準がないことが原因であり，コーパスの理論的可能性を著しく狭める結果になっている。

(2) 研究代表者はコーパスデータを援用したフィンランド語の文法研究を進めてきた (千

葉 1998; Chiba 2001; 千葉 2005)。とりわけ，言語情報処理の技術を用い，形態・統語情報を付加したコーパスデータを文法分析に利用する手法は，平成 15 年度～17 年度の科学研究費補助金 若手研究 (B)「フィンランド語の動詞派生の名詞の統語論と語用論：大規模コーパスによる基礎研究」におけるフィンランド語の派生名詞の生産性の分析を通じフィンランド語研究に新しい視点をもたらすことができた。しかし，コーパスを用いた用例分析に精緻な言語学的情報を活用することで格段に豊かな量的情報が得られるようになった反面，用例の数量的特徴をその文法現象に特有のものと論証することの難しさを痛感した。

(3) 構文や形態統語的な現象を扱う場合，集めた用例から得られた量的な情報の解釈が研究者の裁量に任せられているという現状は，言葉を変えれば，たまたま見出された事実をあたかも重要な特徴と述べていることと同じと言わねばなるまい。コーパスに基づく文法研究が真に実証性と客観性を伴った成果を提供するためには，得られた事実の有意性を客観的に証拠づける仕組みが整備されなければならないと考えた。

## ２．研究の目的

コーパスから何らかの特徴により抽出された用例データが示すふるまいを，数量的に意味のある形で分析するためには，比較可能な基準データの構築が必要である。コーパスのサイズが数億語レベルまで肥大化した結果，コーパス全体に必要な情報を付与するには膨大な資金と時間がかかる。本研究では，大規模コーパスから一定の基準で用例をサンプル抽出することにより「あるコーパスのサンプルコーパス」を構築し，必要な文法情報を付与することでコーパス言語学の「代表性」の方法論を生かす，コーパス情報の「サンプル化」の手法を提案する。本研究は以下の射程をもつ。

(1) フィンランド語の大規模な書き言葉コーパス「フィンランド語バンク」のサンプル化のとりくみを通じ，サンプル化の方法論を整備する。

(2) 主として文法研究に利用することを想定し，サンプル化されたコーパスに形態・統語情報を中心とした言語学的情報を付与しデータベースを構築する。

(3) サンプル化し，多層的な文法的・語彙的情報を付与したデータベースを構文研究に応用し，フィンランド語の諸構文の用例群が示す情報の有意性を評価するための分析手法を整備する。そのうえで，フィンランド語の諸構文の具体的な分析を通じ，サンプル化されたデータベースを利用した分析手法が文法研究の精緻化，客観化に貢献しうることを論証する。

## ３．研究の方法

(1) フィンランド語の構文研究において量的調査をおこなうための分析パラメータを選定する。① 語，句，節，文など文構造のレベル，② 句の主要部や句の間の文法関係のレベル，③ 指示関係，語順などの談話構造の結束性のレベル，④ 依存関係や一致の関係など形態統語構造のレベル，⑤ 語形や語形成など形態論的構造のレベル，⑥ 語種などの意味のレベル，など複数のレベルを想定し，フィンランド語学における統語論と形態論の記述 (Hakulinen, Karlsson & Vilkuna 1980; Vilkuna 1996; Hakulinen *et al*. 2004) を参照しながら整理する。

(2) フィンランド学術計算機センターのテキストコーパス「フィンランド語バンク」に収録されている１億語を超える書き言葉のコーパス群について基礎的な統計分析をおこない，サブコーパス毎に特徴の抽出をおこなう。

(3) サンプル化されたコーパスを生成する手法の検討をおこない，サンプル化プログラムを開発する。完成したプログラムを用いて「フィンランド語バンク」のサンプル化をおこなう。

(4) サンプル化したコーパスをデータベース化し，パラメータごとの統計情報を算出する。句構造や文法関係，形態・統語情報などの文法情報はフィンランド語統語解析ツール(Connexor 社製 Machinese Syntax)を用いて自動分析をおこない，手修正をおこなって精度の向上をはかる。

(5) 言語学的パラメータ間ならびに語彙的情報と各パラメータとの相関関係について，データベースを用いて言語学的な記述と分析をおこなう。

(6) 「フィンランド語バンク」全体から取得した構文の用例データをサンプル化したコーパスの量的情報と比較・対照するためのプログラムの開発をおこなう。

(7) (6) で開発した分析ツールを用いて具体

的な構文分析を試みる。

## ４．研究成果

(1) 各コーパスの分析に基づき，「フィンランド語バンク」のサンプル化をおこなった。「日本語話し言葉コーパス」(前川 2004) でのデータ構築の実績を踏まえ，付与する言語学的情報の量・質についてレベルを分けて対応した。

① 形態・統語解析ツールの解析結果の付与
② 形態・統語解析の結果の検証と校正
③ 文構造・談話構造のレベルの情報の付与

このうち，① は形態・統語解析ツール (Connexor 社製 Machinese Syntax)の解析結果を用いてサンプル化の手法で「フィンランド語バンク」から抽出された全データについておこなっており，2011 年 5 月末の時点で 2,000 万語あまりのデータが検索可能な形で整備されている。① の検証と校正をおこなう ② の作業については 2011 年 5 月末現在，850 万語分のサンプル化データベースへの実施を完了した。③ は ② の検証済みデータの一部 (50 万語) についてのみ，実験的に詳細な情報付与をおこなった。今後も内容の検証を通じ，当該情報のうち処理が単純で分析に有用なものをサンプル化データベース全体に付与する作業をすすめたい。

(2) サンプル化されたデータベースから形態・統語パラメータによる統計情報を算出するプログラムを開発した。データベースに収録されている付与情報は語彙レベル，形態・統語レベル，文・談話レベルと多岐にわたっており，データ構造は非常に複雑なものとなった。このうち語彙レベルの情報として処理できる主要な情報については事前に頻度計算をおこなってデータベースに格納し，処理速度の向上を図っている。また，データベースの用例検索プログラムは検索結果の表示や保存に XML を用いており，各レベルの情報を過不足なく記述できるようになっている。

(3) サンプル化したデータから統計情報を取得し，「フィンランド語バンク」全体から検索・抽出した用例群の特徴と比較するツールを開発し，比較の結果から用例群の数量的情報の有意性をマイニングできるようにした。解析手法の妥当性については，海外調査ならびに研究発表を通じフィンランド語学の専門家と意見交換をおこなっている。

現在の解析では，パラメータの相対頻度のほか，任意の出現要素について ① ダイス係数 ② MI スコア ③ 対数尤度比 および ④ エントロピーを用いた解析と評価をおこなっている。t スコア，z スコアなど他の主要な統計的指標を用いることができないのは，異なるサイズのデータを比較する本研究のような事例の場合に統計的に妥当な値を算出できる指標が限られるためであり，サンプル化したデータベースを用いる比較の手法のもつ制限となっている。

一方で，機能語や屈折形態素のような高頻度で多くの環境に出現しうるパラメータと，派生形態素や小辞，語彙項目などの比較的出現頻度の低いパラメータが統計的にかなり異なる様相を見せていることが明らかになっている。これらは，現在の指標で一般に用いられている「量的データの統計的補正」という考えでは統一して処理することが難しく，構文特徴の有意性を一意に検出できる独自の指標や算出式の提案など，今後も試みを続けていく必要性を示すものである。

(4) サンプルケースとして，選定した構文特徴をもつ用例群とサンプル化したデータベースとを文法情報のパラメータに基づき比較分析し，その結果をハンガリーで開催された第 11 回国際フィン・ウゴル学会において発表した。

プロトタイプのデータベースを用いた分析の結果，文法要素が現れる構文環境について，そのパラメータの構成要素の出現比率がサンプル化されたデータベースのそれと明らかに異なるものが数多く見いだされた。これらには，許可構文の場合など，Mayerthaler (1981)が述べる有標性の逆転（markedness reversal）現象といえる事例も散見される。しかし，有標・無標の対立とまでは言えないものの，有意な比率の違いがみられるものもあった。前者の場合，語彙的な特徴が比率の逆転に大きく貢献している事例が多い。現段階では未実施となっている複数のパラメータを複合的に分析しそれらの分布パターンを分析する作業とともに，構築されたデータベースを活用し今後さらなる分析をおこないたい。

(5) サンプル化の手法を用いた分析を日本語の大規模コーパスについても応用し，その成果を内外の学会で発表した。本研究の提案するサンプル化の手法を用いることで，フィンランド語以外の言語においても用例のもつ量的情報をより客観的に評価できることが明らかになった。

## 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７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，査読有，15 巻，2008，17-32
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，査読無，2009，55-70
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，査読無，2010，91-107
④ 千葉庄寿，素性情報を利用した，解析済み日本語コーパスからの語彙パターンの抽出，2010 世界日本語教育大会研究論文集（DVD-ROM），査読無，2010，1547/0-7
⑤ 千葉庄寿，大規模均衡コーパスを利用した語彙・文法情報の評価とその応用，言語処理学会第 17 回年次大会論文集，査読無，2011，675-676
⑥ 千葉庄寿，BCCWJ を用いた語彙・文法情報のプロファイリングとその応用，「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」平成 22 年度公開ワークショップ予稿集，査読無，2011，439-442
⑦ 千葉庄寿，HTML と XML，ウェブによる情報収集（IT と日本語研究 第 7 巻），明治書院，査読無，177-227

〔学会発表〕（計８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
⑥ 千葉庄寿，「素性情報を利用した，解析済みコーパスからの語彙パターンの抽出」2010 世界日本語教育大会，2010 年 7 月 31 日，国立政治大学(中華民国)
⑦ Shoju CHIBA , “Colligational frameworks in Finnish,” The 11th International Congress for Fenno-Ugric Studies (第 11 回国際フィン・ウゴル学会), 2010 年 8 月 12 日, Pázmány Péter カトリック大学(ハンガリー共和国)
⑧ 千葉庄寿，「大規模均衡コーパスを利用した，解析済み日本語コーパスからの語彙パターンの抽出」言語処理学会第 17 回年次大会，2011 年 3 月 9 日，豊橋科学技術大学

〔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

本研究で構築した「フィンランド語バンク」のサンプル化データベースの一般公開については，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協議していく。

６．研究組織

(1)研究代表者

千葉　庄寿（CHIBA　SHOJU）
麗澤大学・外国語学部・准教授
研究者番号：70337723