様式 C -19

科学研究費補助金研究成果報告書

平成 21 年 6 月 9 日現在

研究種目：基盤研究（A）
研究期間：2006～2008
課題番号：18202010
研究課題名（和文） 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

研究課題名（英文）Corpus building and corpus-based studies in non-Slavic languages of Russia and neighboring countries

研究代表者
松村　一登（MATSUMURA　AZUTO）
東京大学・大学院人文社会系研究科・教授
研究者番号：40165866

研究成果の概要：ロシア連邦およびその周辺諸国で話される非スラブ系の小規模言語・少数言語(ウラル諸語，チュクチ・カムチャツカ諸語ほか)のいくつかの言語資料を現地で収集し，Unicode のエンコーディングの電子的な文字データとしたものを，コーパスとして利用可能なデータ形式に整形加工した。また，電子化された言語資源が乏しい(ないし，十分とはいえない)言語の研究において，電子的な言語資料を利用した研究方法が果たしうる役割を示すことを目的として，作成したコーパスを用いてこれらの言語の記述的・歴史的な実際に研究する試みを行った。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	6,200,000	1,860,000	8,060,000
2007 年度	5,800,000	1,740,000	7,540,000
2008 年度	6,100,000	1,830,000	7,930,000
年度			
年度			
総　計	18,100,000	5,430,000	23,530,000

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：コーパス，言語資料，少数言語，ロシア，エストニア語，イテリメン語，ウラル諸語，チュクチ・カムチャツカ諸語

１．研究開始当初の背景

(1) 人文科学の諸分野に共通するのは，言語資料を読み(聞くことを含む)，そのテクストから情報を引き出すことが研究の出発点であることである。同じ言語資料であっても，それをどのように読み，どのような情報を引き出すかは，それぞれの分野によって異なる。たとえば，歴史学者なら歴史上の人物に関する情報を読み取るであろう日記のテクストから，言語学者は，当時の日本語の動詞に関する文法的事実を読み取り，社会学者は，噂の広がり方が昔も今も同じであるかどうかを知るための手がかりを読み取るかもしれない。テクストをもっぱら目で追うことによって情報を読み取っていた従来の人文科学は，今，電子化されたテクストからコンピュータを使って情報を引き出す研究方法に移行しつつあると考えられる。中規模言語(話者人口数百万人)，小規模言語(話者人口 100 万人以下)の言語資料の研究の今後もこのよ

うな流れの中で考えるべきであり，電子化された言語資料の蓄積に本格的に取り組む必要がある。

(2) コーパスを用いた言語研究の点で一歩リードしているのは，電子化された言語資料の量の点で，また，言語教育という観点からの需要の大きさの点で他の言語を圧倒的に凌駕する英語であるが，日本でも，とくに今世紀に入ってから，コーパスを用いた日本語研究が急速に盛んになってきている。話者が多く，文字化された言語資料が豊富にある大言語の場合，延べ語数が億の単位で語られる，いわゆる大規模コーパスが注目されている。しかし，6000 以上といわれる世界の言語の中で，大規模コーパスについて語ることができるのは，ほんの一握りの大言語にすぎないのが実情である。この「格差」をこのままに放置するのは好ましくない。

(3) ロシア国内やその周辺諸国の言語は，書記体系を発達させている場合には，非ラテン文字系の文字を用いている言語が多いし，ラテン文字系の書記体系をもつ場合でも，Latin-1 と呼ばれる西ヨーロッパ系の文字のセットには含まれない，特別なアクセント記号の文字を用いて表記している言語が大部分である。また，書記体系があるとされていても，社会的には十分に機能していない言語や，書記体系が定まっていないと見なすのが適当な言語（いわゆる「文字のない言語」）も少なからず存在する。このような言語の場合，言語学者によって音声表記で記録された言語資料が存在することもあるが，一般には十分ではなく，新たに母語話者から現地で聞き取り，音声表記のデータにする必要がある。

(4) Unicode によるエンコーディングとテクストのマークアップ方式としての XML が，国際標準と普及したことにより，非ラテン文字系の書記体系を持つ言語の言語資料や，音声記号などにより音声表記された言語資料を，文字データとしてコンピュータで処理することが，言語学者にとって現実的になってきた。様々な文字体系の文字を収録した Unicode フォント，Unicode による文字入力や文書作成に対応したテキスト・エディタ，Unicode でエンコードされたテクスト処理に対応したプログラミング言語など，ソフトウェアの面でも，Unicode によるテクスト処理が標準に成りつつある。また，XML でマークアップされた文字データを検索するためのツールも普及し始めている。

２．研究の目的

(1) 日本国内はもとより，外国でも，ロシア連邦およびその周辺諸国で話される非スラブ系の諸言語の電子化された言語資料は豊富に利用可能であるとは言えない。この状況を少しでも改善することを第１の目的とする。この地域の小規模言語・少数言語(ウラル諸語，チュクチ・カムチャツカ諸語ほか)のいくつかについて，言語資料を現地で収集し，Unicode のエンコーディングの電子的な文字データとしたものを，コーパスとして利用可能なデータ形式に整形加工する。また，このようにして作成した電子的な言語資料は，可能な限りにおいて，一般に入手可能な形で公開する。

(2) 人文科学の諸分野が，テクストを目で読む(耳で聞くことも含む)ことによって情報を読み取るという従来の方法から，電子化されたテクストからコンピュータを使って情報を引き出す研究方法に移行しつつある現状をふまえ，言語学において急速に発展しつつあるコーパス言語学の研究方法を，電子化された言語資料が比較的乏しい言語の研究において実際に適用する具体的な試みを行うことを第２の目的とする。

(3) 言語学者を主体にする研究プロジェクトではあるが，歴史学者にも研究組織に参加してもらうことによって，コンピュータを使って情報を引き出し，それをもとに議論を行い，結論を導く手法が，同じ言語資料に対して，異なる分野の研究者によって適用された場合に，引き出す情報や結論を導く過程にどのような共通性があり，どのような違いがあるのかの比較も試験的に検証することができると期待される。

３．研究の方法

(1) 書記体系が確立し，社会的に機能している言語については，まとまった言語資料を入手する。書記化された言語資料のない言語については，現地調査により，母語話者からテクストを入手し，音声表記などを用いて文字化する。いずれの場合も，Unicode でエンコードした電子テクストとしてコンピュータ入力し，最終的には(well-formed な)XML 文書として整形する。

(2) Unicode による文字入力，Unicode でエンコードされたテクストの整形・加工のためのツールとしては，以前の研究プロジェクトで開発したフォントやツール，あるいは，フリーウェア，シェアウェアなど，各自が使い慣れたツール類で十分対応できるので，とくに，この研究のためのツールの開発は行う必要はないと考える。

(3) コーパスを検索するためのツールについては，現状では，多言語対応を謳っていても，

この研究プロジェクトが対象とする地域の言語の書記体系を自在にカバーしているものは，残念ながら見あたらない。この研究の目的は，ツールを独自に開発することにはないので，コーパス検索のためのツールの開発，あるいは選択は，各自の裁量に委ねる。研究代表者は，Perl 言語で書いた自作のツールを，KWIC 索引作成や，検索結果の集計に用いる。

(4) コーパスを利用した言語学的な研究は，KWIC 索引の作成を基礎にして行い，共起関係の強さの判定など，コーパス言語学の標準的な手法を適用する。

４．研究成果

(1) 量的にまとまったものとして，エストニア語(ウラル語族)，イテリメン語(チュクチ・カムチャツカ語族)の言語資料が電子化された。

①エストニア語については，1920 年前後の言語資料として，「エストニア憲法制定会議議事録」(Asutawa Kogu protokollid, 1919-1920)のほぼ全文(193 万語)を電子化し，文単位に区切った(well-formed な)XML 文書とし，コーパスとして公開した。このコーパスは，エストニアのタルト大学のコンピュータ言語学研究グループから，エストニア語の歴史的なテクストのコンピュータ処理の観点から貴重なデータになると評価を受けた。このコーパス作成の過程では，印刷された原資料の前ページのマイクロフィルム撮影とマイクロフィルムからのデジタルスキャニング(電子画像変換)は日本で行った。その次の段階として，マイクロフィルムから変換した画像データをＯＣＲで読み取り，文字認識する作業は，エストニア国立図書館の電子化部門に委託して，エストニアで行った。同図書館では，いわゆる電子図書館のプロジェクトを進めており，この文字認識の結果は，そのプロジェクトでも活用される予定である。次の段階として，ＯＣＲによる文字認識の校正作業はエストニア在住の母語話者に委託して行った。最終段階のマークアップは，日本で，Perl 言語で書いたスクリプトによって行った。エストニア語のテクストは，歴史的な綴りで入力され，文レベルでは，次のようにタグ付けされている。

<s no="197"> Üks on wastu. </s>

②イテリメン語については，まず，既存の出版物(イテリメン語の教科書)を Unicode (UTF-8)によるエンコーディングで電子化した。また，ロシア・カムチャツカにおいて２度のフィールド調査を実施し，話者の協力のもとで新たなテキストを収集した。さらに録音音声データの電子テキスト化を進めた結果，多くの言語データを蓄積し，記述研究の基盤を充実させることができた。イテリメン語のテクストは，ロシア語アルファベットにない特殊なキリル文字も Unicode の文字として入力され，文に区切られている。文レベルでは，次のようにタグ付けされている。

<s id="1433">ктхлэ.</s>

(2) コーパスを実際に利用して，エストニア語の記述的・歴史的な研究を行った。

① エストニア語の動詞 pruukima は，中世の低地ドイツ語の動詞 brūken(現在の標準ドイツ語の brauchen と同語源)が借用されたものである。pruukima は，現在のエストニア語で「～する必要がある」という法助動詞的用法と，「～を用いる」の意味の一般動詞としての用法がある。20 世紀末の新聞記事のコーパスで見ると，法助動詞的用法は，一般動詞用法と比べ使用頻度が圧倒的に高いほか，そのほとんどが否定的文脈で用いられている(＝否定極性語である)ことがわかる。「エストニア憲法制定会議議事録」での動詞 pruukima の用法を調べると，ほぼ同様の結果になり，すでに 20 世紀の初めには現在とほぼ同じ使用パタンが成立していたことを伺わせる。

② エストニア語の rahvas と rahvus(「エストニア憲法制定会議議事録」の綴りでは rahwas，rahwus)は，ともに英語の nation に対応するエストニア語であり，エストニアの歴史研究において，文献を読み解く際のキーワードになる。rahwas や rahwus が 20 世紀の初め頃意味していた内容は，それらの語が 19 世紀に意味していた内容とも，また，対応する語が現在意味している内容とも異なっていると考えられる。「エストニア憲法制定会議議事録」から実際にこれらの語の用例を抽出してみると，当時 rahwas は，民族的な意味での「エストニア人」をさす語であったことがわかる。

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 8 件）

①松村一登, Dictionary definition against corpus data: the polisemy of the Estonian verb jõudma, ウラリカ，15 号，2009, [印刷中], 査読有

②千葉庄寿，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 号， pp.17-32, 2008, 査読有

③小野智香子，イテリメン語の不定詞，環北太平洋の言語, 14 号, pp.1-18, 2007, 査読無

④松村一登，文字化された言語資源の少ない言語とテクストのマークアップ，国際セミナー TEI Day in Kyoto 2006 報告書，2006，pp.5-27，査読無

⑤松村一登，マリ語の言語資料とその電子化，ウラリカ，14 号，pp.45-56，2006，査読有

⑥滝沢直宏，コーパス利用のためのコンピュータ・リテラシー，日本語教育，130 号，pp.22-31，2006，査読無

⑦千葉庄寿，構造化された言語データが言語研究にもたらすもの，麗澤大学紀要，82 号，pp.43-65，2006，査読無

⑧小野智香子，イテリメン語テキスト 2，ユーラシア言語文化論集，第 9 号， pp.257-268，2006，査読無

〔学会発表〕（計 5 件）
①千葉庄寿，アノテートされた大規模コーパスを用いた言語分析のモデル：Xaira を例に，科学研究費特定領域研究「日本語コーパス」日本語教育班研究連絡会議，2008/12/21，早稲田大学

②松村一登，エストニア語の動詞 pruukima「必要だ；用いる」の多義性―コーパスと辞書の記述に基づく考察―，日本言語学会第 137 回大会，2008/11/29，金沢大学

③松村一登，エストニア語の動詞 jõudma「～できる；至る」の多義性について― 新聞記事コーパスに基づく研究 ―，日本ウラル学会第 35 回研究大会，2008/07/05，名古屋大学

④松村一登，９０年前のエストニア語の言語資料の電子化― コーパスによるエストニア語の歴史の研究を目指して ―，2007/07/07，東京大学

⑤松村一登，複合動詞の生産性といわゆる「統語的／語彙的」の区別，日本言語学会第 134 回大会，2007/06/17，麗澤大学

〔図書〕（計 1 件）
①松村一登(編)，電子化された言語資料と個別言語研究，東京大学大学院人文社会系研究科，2009，148pp.

〔その他〕
Web サイト：
http://www.l.u-tokyo.ac.jp/~kmatsum/kaken/

６．研究組織
(1)研究代表者
松村　一登（MATSUMURA KAZUTO）
東京大学・大学院人文社会系研究科・教授
研究者番号：40165866

(2)研究分担者
平成 18～20 年度
後藤　斉
東北大学・大学院文学研究科・教授
研究者番号: 90162156

千葉　庄寿
麗澤大学・外国語学部・准教授
研究者番号: 70337723

小森　宏美
京都大学・地域研究センター・准教授
研究者番号: 50353454

平成 18～19 年度
滝沢　直宏
名古屋大学・大学院国際開発研究科・教授
研究者番号: 60252285

平成 18 年度
畑野　智栄
東京大学・大学院人文社会系研究科・助手
研究者番号: 40376520

(3)連携研究者
平成 20 年度
滝沢　直宏
名古屋大学・大学院国際開発研究科・教授
研究者番号：60252285

小野　智香子
千葉大学・人文社会系研究科・講師
研究者番号：50466728