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

令和　５ 年　 ６ 月 ３０ 日現在

機関番号：３２５０６
研究種目：基盤研究(C)（一般）
研究期間：2016～2022
課題番号：１６Ｋ０２６９７

研究課題名（和文）形態・統語情報を考慮した多層的語彙ネットワークの描出とその応用に関する研究

研究課題名（英文）A study on visualizing lexical networks based on the multilayered linguistic information: a case of modern Finnish

研究代表者
千葉　庄寿（Chiba, Shoju）

麗澤大学・外国語学部・教授

研究者番号：７０３３７７２３

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究は、語の形態統語的特徴や、コロケーション情報、統語情報、情報構造等、多様なレベルの情報をふまえて語彙の類似度を分析するための枠組みの開発と、語彙の類似性の可視化を目的に、データベースの構築と分析アルゴリズムの開発をおこなった。
さらに、このシステムの応用に関して、語彙ネットワークによる分析がサイズの異なるコーパスデータを含めたコーパス間の比較分析にどの程度応用できるか、辞書記述の改善にどう役立つか、さらにはフィンランド語以外の言語の記述に適用可能かどうかを検証することを目的に、学習者データの収集と整理・分析、既存辞書の評価と辞書参照サービスへの応用、日本語の分析への適用を試みた。

研究成果の学術的意義や社会的意義
サイズの異なるコーパス間の比較においても本研究の手法が有効であることがフィンランド語のコーパス分析から明らかになった。また、レマ lemma だけでは必ずしもフィンランド語の語彙間の関係が適切に可視化されず、レマ＋適切な形態統語的情報による可視化が有効であることが示された。統語情報など他の層の情報を用いた語彙記述においても、統語情報等のスロットに形態統語的情報を付与することに効果があることが判明し、語彙の類似度の多層的比較のための方法論の構築という観点で有益な示唆が得られた。本課題の応用として、学習者コーパス構築、辞書記述の評価、辞書検索アプリケーションおよび学習語彙集の構築をおこなった。

研究成果の概要（英文）：In this study, we developed a framework for analyzing lexical similarity with multilayered linguistic data, including morphosyntactic features, collocation information, syntactic information, and information structure. Based on the framework, we developed a database with an algorithm to visualize the similarity of lexical items using those multilayered data.
In addition, we conducted a pilot study to apply the system to concrete applied linguistic and computational linguistic tasks: We collected learner Finnish data and analyzed them with our framework, evaluated existing dictionaries and developed a prototype dictionary reference service, and analyzed a Japanese morphological construction with the algorithm.

研究分野：言語学

キーワード：言語学　フィンランド語　コーパス言語学　語彙ネットワーク　可視化　学習者コーパス　学習語彙集

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

本研究は、研究代表者がこれまで実施してきた、現代フィンランド語の大規模コーパスをもとに構築した形態統語データベースを援用した研究の応用事例として、新たに注目されてきている語彙ネットワークの描出にとりくむものである。大規模データがもつ語彙の使用パターンの重なりをネットワークとして可視化し、語彙間のつながりを多層的に記述することで当該言語の語彙の全体像を描出する「多層的語彙ネットワーク」(MDLN)の分析をおこなう。「多層的」とは、形態統語情報に加え、統語情報、語用論情報、談話構造や意味論情報といった、レベルの異なる言語情報を総合的に扱い新たな枠組みでの可視化を試みる、というものである。これまで、研究代表者は派生語の分析や構文の生産性の評価といった課題にコーパスの情報を応用し、ある程度の成果を収めてきた(千葉、2011-2015)。

２．研究の目的

本研究は、前節の背景をふまえ、語の形態統語的特徴や、語のもつコロケーションを含む統語的多様性をふまえ、語彙の類似度をベクトル化し、語彙間の距離を多様なレベルで記述し比較する枠組みを開発するとともに、コンピュータの可視化技術による語彙間の「つながり」の描出手法を整備し、語彙の類似性・特徴を多層的に捉えることが大きな目的であった。このための手法には、言語情報処理のなかで研究計画立案当時急速に開発が進んでいた分散意味論の時限圧縮の手法を念頭に置いていた。

さらに、構築した MDLN をどのように応用するかも本研究の重要な射程であった。MDLN に基づく分析が、サイズの異なるコーパスデータの分析を含め、コーパス間の比較分析にどの程度応用できるか、辞書記述の更新といった具体的にアプリケーションにどのように応用することができるか、さらには類型論的に異なる言語の記述にも適用可能かどうかを検証することを目的に、フィンランド語の新たな学習者データ(日本語を母語とするフィンランド語学習者の作文コーパス)の構築、既存辞書の評価のこころみと辞書参照サービスへの応用、フィンランド語以外の言語への手法の適用を試み、分析手法の応用可能性を検証することとした。

３．研究の方法

研究開始後におこなった研究者との情報交換の中で、本研究における中心的な課題である、語彙ネットワークから得られたベクトル化された特徴量を重層的に比較分析するという手法について、語彙ネットワーク構築のためのモデルを検討するため専門家と情報交換をおこなった際、本研究が想定している語彙の類似度の多層的比較という手法は、応用を想定していた分散意味論のモデルのもつ射程とはかなり異なることが明らかになった。同様に、可視化するための技術的なポイントについては語彙ネットワークの多次元空間を 2 次元化するための事例は多数収集することができたが、主成分分析による情報の縮約がもたらす問題は本研究のもつ多層的視点と大きく関っており、次元圧縮による可視化は本研究の方法論的な問題への解決にならないことが判明した。

そこで、データ分析の手法としては word2vec を用いることとし、形態統語論、談話構造、語用論、意味論といった各層の情報のベクトル化と可視化を優先して実施し、各層のベクトル情報を含むデータベースの構築を優先する方針でシステムの開発をすすめ、多層化したデータの相互参照の方法、可視化をどのように実装するかについては早急な結論を出さず、最終的な方法論を確定するための基礎研究として、データベースの構築を優先することとした。

データ構築とともに、語彙ネットワークを応用した研究にも取り組んだ。まず、データベースと既存の辞書記述との比較分析をすすめた、2 つの辞書(Conlexis, 幡野)の記述との対照をおこなうことで、辞書記述の評価や整合関係の確認、既存辞書との役割分担の考察といった作業をおこない、分析の成果を検索できる用例データベースとして公開することとした。

さらに、本研究で構築する語彙ネットワークの応用実験のための大規模コーパスの比較対象としての言語データの収集と分析をおこなった。特に、本研究では、フィンランド語学習語彙集の構築と、その拡張・評価のための語彙ネットワークの活用、および小規模なデータへの手法の応用可能性を念頭に、大規模データとの比較分析を試行するためのフィンランド語学習者コーパス(日本語フィンランド語学習者コーパス、ICLFI-japan および成人フィンランド語学習者コーパス)のデータ収集作業をおこなった。

最後に、多層的語彙ネットワークに基づく分析が類型論的に異なる言語の記述にも適用可能かどうかを検証するため、フィンランド語以外の言語として日本語の分析への手法の適用を試みた。

４．研究成果

フィンランド語の 2 種類の大規模書き言葉コーパス(FTC, Suomi24)と小規模な学習者コーパス(ICLFI)を語彙ネットワークの分析枠組み(Word2Vec)により解析し、データベース化するとともに、結果の比較をおこなった。その結果、サイズの異なるコーパス間の比較において、本手法が有効であることが明らかになった。また、語彙ネットワークの可視化実験から、レマ lemma と表層形による語彙ネットワークの記述を比較し、レマによるネットワークが必ずしもフィンランド語の語彙間の関係を適切に可視化しないこと、レマと表層形の中間段階にあたる情報、具体的には適切な形態統語的情報を付与したコーパスによるネットワーク描出が最も有効であることが明らかになった。同様に、統語情報など他の層の情報を用いた語彙記述、つまりある語彙がどのような統語的な環境において出現し、またその環境は他のどの語彙とどの程度類似しているか、という観察においても、統語情報に形態統語的情報を付与することに効果があることが判明し、語彙の類似度の多層的比較の方法論の構築という観点において重要なデータが得られた。

コロナ禍による本研究課題の研究期間の延長が続いた結果、異例に長い研究期間での研究遂行となった。渡航自粛による研究活動の断絶のなか、教育用語彙集の構築・更新のための活動は本務校のカリキュラム変更のため、終了せざるを得なくなり、2020 年時点でのバージョンの語彙集を最終成果として公開した。

フィンランド語作文コーパスの収集と分析に関しては、本研究課題の開始後まもなく、口語フィンランド語、および学習者のフィンランド語の専門家と充分な情報交換と検討をおこない、緻密な計画のもとコーパスデータ収集システムを用いて 2019 年度までに日本語を母語とする学習者約 60 名から 70 あまりの作文データを収集した。残念ながら 2020 年以降、コロナ禍によりデータ収集が中断したが、その間データの整備と基礎的な分析をおこなうことができた。

コロナ禍の期間中最も大きな課題となったのが、言語解析の手法やトレンドの変更による、研究手法の大きな変化であった。

現在、フィンランド語の大規模コーパスの解析トレンドは、omorfi (Open morphology for Finnish)のようなオープンソースによるもので、CSC/kielipankki がホストする大規模言語データも解析データの移行が進んでいる。本研究では、本研究に先立つ研究課題で構築してきた統語解析データベースが用いてきた Connexor 社 fi-fdg (Machinese Syntax for Finnish)から Finnish-dep-parser への移行を余儀なくされ、一部データの書き換えとデータの再解析、データベース構造の修正、変換に大きな時間を割くこととなった。さらに、古いデータとの整合性の観点で、本研究で構築した日本語を母語とするフィンランド語学習者の作文コーパス(ICLFI-japan)はデータの統合ができない状況にあり、本研究では ICLFI の再解析という手段で分析を先にすすめることとなった。

同様に、辞書記述の評価分析のサンプルとして利用を予定していたオンラインフィンランド語辞書 ConLexis が研究開始後まもなく使用できなくなり、急遽フィンランド語日本語辞典(幡野)を使用することとなった。その後、ConLexis はサービス復活の運動もあり運用を再開したが、研究がその間中断してしまう結果となり、研究遂行に少なからぬ影響が出た。

本研究がそのデータベース構築に使用したベクトル化のアルゴリズム(word2vec)も、2013 年にそのアイディアが公開されて以来、長い時間が経過している。現在、このアプローチは Transformer などの新しい深層学習モデルに引き継がれ、自然言語処理のアルゴリズムの中ではかなり古いものに属しており、より新しいものに刷新する余地を残している。

本研究で構築したデータベースとその可視化アルゴリズムは、語彙ネットワーク可視化システムとして試験実装し、運用している Web サーバに試験公開し、研究者からのフィードバックを得た。

最後に、類型論的に近い言語として、研究協力者とともに日本語の解析に取り組み、複合語を含む例文をコーパスから取得するシステムへの応用をおこなった。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　0件／うち国際共著　1件／うちオープンアクセス　1件）

１．著者名 チャクマク＝ビルギル・ニハル、千葉庄寿	４．巻 1
２．論文標題 コーパスからの複合動詞の自動抽出の試み―近現代の文学作品からの用例抽出を例に―	５．発行年 2022年
３．雑誌名 Proceedings of Language Resources Workshop	６．最初と最後の頁 319-337
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

〔学会発表〕　計14件（うち招待講演　1件／うち国際学会　5件）

１．発表者名 チャクマク＝ビルギル・ニハル、千葉庄寿
２．発表標題 コーパスからの複合動詞の自動抽出の試み―近現代作家の文学作品からの用例抽出を例に―
３．学会等名 言語資源ワークショップ2022
４．発表年 2022年

１．発表者名 Shoju Chiba
２．発表標題 Depicting semantic similarities with valency: how syntactic information can enrich the lexical descriptions of verbs and nominals
３．学会等名 13th International Congress for Finno-Ugric Studies (CIFU XIII)（国際学会）
４．発表年 2022年

１．発表者名 千葉庄寿
２．発表標題 フィンランド語学習語彙集の開発と評価：大規模コーパスを用いた頻度情報をどう活用するか
３．学会等名 第46回日本ウラル学会研究大会
４．発表年 2019年

１．発表者名
Shoju Chiba

２．発表標題
Theory and practice of enriching a word list: A case study of building a student glossary of Finnish for Japanese learners

３．学会等名
Research Data and Humanities (RDHum 2019)（国際学会）

４．発表年
2019年

１．発表者名
Shoju Chiba

２．発表標題
Depicting semantic similarities with valency: how syntactic information can enrich the lexical descriptions of verbs and nominals

３．学会等名
13th International Congress for Finno-Ugric Studies (CIFU XIII)（国際学会）

４．発表年
2020年

１．発表者名
Shoju Chiba

２．発表標題
Emergence and development of collocational knowledge: How lexical networks capture cohesional developments of learner texts

３．学会等名
LinC Summer School, RWTH Aachen University, Germany（国際学会）

４．発表年
2018年

１．発表者名
千葉庄寿

２．発表標題
大規模コーパスに基づく単語ベクトル情報の教育的応用：フィンランド語の学習語彙リストの作成と評価を例に

３．学会等名
麗澤大学言語研究センター研究セミナー

４．発表年
2018年

１．発表者名
Shoju Chiba

２．発表標題
Theory and practice of enriching a word list: a case study of building a student glossary of Finnish for Japanese learners

３．学会等名
Research Data and Humanities (RDHum 2019), University of Oulu, Finland（国際学会）

４．発表年
2019年

１．発表者名
Jarmo H. Jantunen

２．発表標題
ICLFI Corpus and Annotation of Morphologically Rich Learner Language

３．学会等名
麗澤大学言語研究センターワークショップ「学習者コーパス研究の理論と実践」

４．発表年
2017年

１．発表者名
Shoju Chiba

２．発表標題
Developing a Learner Corpus of Japanese Learners of Finnish: a Prospectus

３．学会等名
麗澤大学言語研究センターワークショップ「学習者コーパス研究の理論と実践」

４．発表年
2017年

１．発表者名
千葉 庄寿, ヤルモ・ヤントゥネン

２．発表標題
多次元ベクトル分析の手法を用いたフィンランド語の語彙ネットワークの構築とその応用可能性について：フィンランド語学習者コーパスの分析を例に

３．学会等名
第44回日本ウラル学会研究発表大会

４．発表年
2017年

１．発表者名
千葉庄寿

２．発表標題
形態論的生産性を測る指標とコーパスサイズの関係について―現代フィンランド語の2種類の大規模コーパスを用いた考察―

３．学会等名
第43回日本ウラル学会研究発表大会

４．発表年
2016年

１．発表者名
千葉庄寿

２．発表標題
FU12フィンランド語関係セッション報告

３．学会等名
第43回日本ウラル学会研究発表大会シンポジウム「ウラル学研究の現在とこれから」

４．発表年
2016年

１．発表者名
Shoju Chiba

２．発表標題
Suomalais-ugrilaista kielentutkimusta ja akateemisen yhteiso"n haasteita Japanissa

３．学会等名
Ita"-Aasian alueen Suomen kielen ja kulttuurin opettajien yhteidtapaaminen 2017（招待講演）

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	ヤントゥネン　ヤルモ (Jantunen Jarmo)		

6．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	チャクマク＝ビルギル　ニハル (CAKMAK BILGILI Nihal)		
研究協力者	幡野　恒 (Hatano Hisashi)		

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計1件

国際研究集会	開催年
ワークショップ「学習者コーパス研究の理論と実践」	2017年～2017年

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
フィンランド	University of Jyvaskyla	University of Helsinki	CSC (IT Center for Science)	他2機関