科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１３　日現在

機関番号：　１２６０８
研究種目：　基盤研究(C)
研究期間：　2010～2013
課題番号：　２２５２０４５８

研究課題名（和文）和歌形態素解析用辞書開発のための用語連接規則に関する基礎研究

研究課題名（英文）Basic research concerning adjacency probabilities in the development of a morphological analysis dictionary for classical Japanese poetry

研究代表者
山元　啓史（Yamamoto, Hilofumi）

東京工業大学・留学生センター・准教授

研究者番号：３０２４１７５６

交付決定額（研究期間全体）：（直接経費）　　2,300,000円、（間接経費）　　690,000円

研究成果の概要（和文）：代表者は2007年に和歌用の形態素解析ツールを開発した。その解析対象は八代集に限定されていた。本研究では八代集の解析済みデータを用い、連接規則をコンピュータ処理で獲得し、それにより二十一代集の解析を実行し、品詞タグづけを行うことを目的とする。KyTea（京都大学KyTeaプロジェクト）とそれに付属する点推定連接規則学習システムにより、ノートブック程度のマシンであっても数十秒で学習モデルの生成ができた。これを用いて、二十一代集の単位切りを行ったところ、ほぼ96%の高い割合で解析ができた。未知語の入力と未知語周辺の連接規則の学習はまだ必要であるが、二十一代集の単位分割を行う辞書は完成した。

研究成果の概要（英文）：The principal investigator has previously developed a tool for the morphological processing of waka poems in 2007. However, its range of applicability was limited to the Hachidaishu. The goal of the present research is to automatically segment and annotate part-of-speech tags for the Nijuichidaishu using the previously annotated segmentation data and token adjacency probabilities of the Hachidaishu. Using the KyTea (Kyoto Text Analysis Toolkit) morpheme segmentation toolkit, with its default L2 regularized SVM learning algorithm, model learning took less than a minute. This model also achieved a high segmentation accuracy of around 96% on the Nijuichidaishu. While there is some remaining work to be done concerning the addition of unknown tokens and the learning of adjacency probabilities around unknown words, the development of a dictionary that can segment the Nijuichidaishu with a high accuracy can be considered complete.

研究分野：　人文学

科研費の分科・細目：　言語学・日本語学

キーワード：　和歌　通時分析　古語辞書　形態素　ネットワーク分析　語彙論　連接規則　機械学習

1. 研究開始当初の背景

和歌の用語はどんな語と相互に結びつき、意味のまとまりを作っているのだろうか。たとえば、「梅」と「鶯」、「桜」と「時鳥」、「吉野」と「桜」、「龍田」と「紅葉」のように和歌特有のコンビネーションが思い浮かぶ。このようなコンビネーションはどれくらい、どんな種類が存在し、どの時代から使われはじめたのだろうか。このような特別な語の組み合わせはどれくらい存在するのであろうか。和歌研究者の直観や経験だけでは即答しにくいコンビネーションを実際に和歌データから得ることはできるのであろうか。地名の「龍田」は紅葉彩る秋の風景、「吉野」は桜をとりまく春の花模様として有名であるが、山元（2005）では、この「龍田」と「吉野」を詠んだ歌をコンピュータで分析し、可視化した。このモデルによって、「龍田」は「神の地」、「吉野」は「人間世界／世俗の地」というまとまりが見られた。このような可視化モデルを生成するには、辞書の整備、言語分析の単位、和歌のデータ表記方法、シソーラスの記述などさまざまな問題があるが、これらの問題を解決していくことにより、今後のコンピュータによる言語文化研究の方法の可能性が広がった。

2. 研究の目的

2009 年までに八代集用語について辞書とシソーラスを整備し (山元 2007, 山元 2009)、八代集限定の可視化モデルを生成するシステムを完成させた。本研究ではこれを基盤として八代集 (905-1205: 9440 首) の 300 年間だけでなく、二十一代集 (905–1439: 25648 首) の 534 年間のより大きな古典の知識を蓄積し、体系化をめざすのが究極的な目的である。

従来の方法では手作業によって辞書を作成していたため、人間ならではの判断の揺れの問題、単位時間に処理する量的な問題、などがあり、一定品質以上の辞書を、数多くの古典作品にわたって作ることは難しい。そこで、本研究では、今までの研究で得られた形態素解析済みデータ（八代集辞書と分割済み和歌データ）を用い、和歌の連接規則を機械学習によって得ることにより、広範の辞書（ここでは、二十一代集用の単位分割辞書）を実際に開発し、その辞書による和歌の形態素解析が実務的なものであるかどうかを検証することを目的とする。

3. 研究の方法

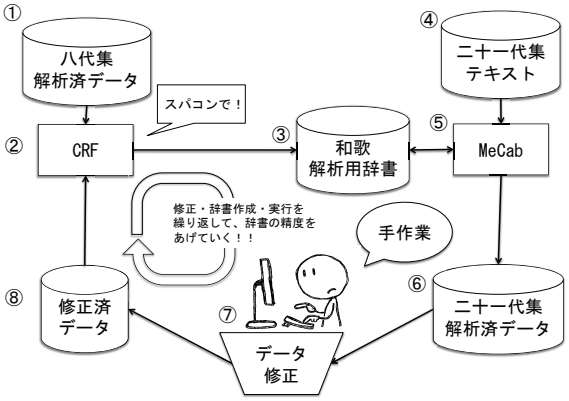

図１　辞書開発と連接規則獲得の手順

図１に沿って説明する。これまでに山元 (2007) が開発した①八代集用の辞書を②CRF（連接パラメタ推定プログラム）で処理することにより、③二十一代集用の初期辞書を作成する。別途電子化された④二十一代集テキストを⑤MeCab（形態素解析器）と上記辞書で処理し、⑥二十一代集を解析する。当然その結果には誤りや未知語があるので、それを手作業で⑦修正する。この⑧修正済みデータを用いて、再度②CRF で③ 辞書を作成する。②–⑧ までを何度も繰り返すと徐々に精度の高い辞書が得られる。この間、繰り返し行われるデータ修正において辞書の品詞体系の見直しと和歌の連接規則記述の理論化を試みる。

以上が当初の計画であったが、スパコンであったとしても、②CRF で連接パラメタ推定するのに大量の計算資源を必要とし、このサイクルを繰り返すには実務として現実的でないことが判明したため、作業は暗礁に乗り上げた。

根本的に連接パラメタ推定は別の方法で独自に考案することを検討しはじめたころ、2011 年になると、KyTea（Kyoto Text Analysis Toolkit: 京都大学 KyTea プロジェクト開発）がβ一般公開が開始された。Kytea 付属の点推定による連接学習プログラムを利用すると、八代集の解析済みの学習データすべてに見られる連接関係のモデルがノートブック程度のマシンであっても数十秒で生成できたことにより、図１の CRF の代わりに train-kytea を用い、連接規則を推定し、MeCab の代わりに KyTea を解析システムとして用いるよう、計画を変更した。

4. 研究成果

KyTea（京都大学 KyTea プロジェクト）とそれに付属する点推定連接規則学習システムにより、ノートブック程度のマシンであっても数十秒で学習モデルの生成ができた。これを用いて、二十一代集の単位切りを行ったところ、ほぼ 96 % (未知語の対応部分を除く) の高い割合で解析ができた。未知語の入力と未知語周辺の連接規則の学習はまだ必要であるが、二十一代集の単位分割を行う辞書は完成した。

```
00001 年のうちに春立ぬとや吉野山霞かゝれる峯のしら雲

00001/UNK/UNK
年/名:年:とし/とし
の/格助:の:の/の
うち/名:内:うち/うち
に/格助:に:に/に
春/名:春:はる/はる
立/タ四-体:立つ:たつ:立つ:たつ/たつ
ぬ/完-終:ぬ:ぬ:ぬ:ぬ/ぬ
と/格助:と:と/と
や/係助:や:や/や
吉野山/名-地名:吉野山:よしのやま/よしのやま
霞/名:霞:かすみ/かすみ
かゝれ/ラ四-已:掛かる:かかる:掛かれ:かかれ/かかれ
る/完-体:り:り:る:る/る
峯/名:峯:みね/みね
の/格助:の:の/の
しら雲/名:白雲:しらくも/しらくも
```

図２　続後撰和歌集の１番歌とその解析結果

図２は続後撰和歌集の１番歌を入力として、それを KyTea で解析した結果である。続後撰集の１番歌の語はすべて八代集に出てきたものではあるが、歌そのものは八代集には見られない。すべてにおいて問題なく解析されている。このことは、(1) 八代集に見られる連接パターンだけで、二十一代集の連接パターンはほぼ構成されていること、(2) 八代集辞書の作成において、実際に公開されている和歌の表記（漢字表記、かな表記、藤原定家の表記、翻刻校訂段階による表記）のすべてを辞書に登録したため、その網羅的な作業が本研究において貢献していると考えられる。

以上、和歌の解析に実用的となったシステムを利用して、歌ことば可視化システムを試作した。このシステムにはまだ十分なシソーラスが整備されていないので、二十一代集の和歌すべてが可視化できるわけではないが、個別の用語の目的に限ってはネットワークモデルが生成できる。H. Yamamoto(2013) はその試作品による研究成果である。ある用語を含む和歌を材料に、語彙ネットワークを描画すると、通常は、ある用語をハブ（中心となる語）としてにネットワークが描かれる。しかし、時々、見出し語としては扱われにくいような語や一見トピックにあまり深く関連していないと思われるような語が、ハブとして描かれることがある。たとえば、「山吹」をトピックとしてネットワークを描画したところ、「蛙」「井出」「八重」の語が「山吹」に伴うハブとして描かれた（図３）。一般的に「山吹」には「蛙」「井出」が伴うことは和歌の世界では常識的であり、歌ことば辞典でともに詠まれる用語として解説されてはいるが、「八重」は用語集や辞典では、それ単独の用語としても取り上げられていない。このように、あまり辞典の見出し語としては取り扱われない語であってもネットワークによる分析において、ハブとして存在し、語と語の関係を表すのに、重要な役割を担う語が存在することがわかった。

従来、辞書編纂は執筆者が意識的に想起できる語のみが辞書に掲載されており、執筆者が意識できない語は当然辞書には掲載されない語となる。具体物ではなく抽象的な意味を表す語は説明しにくいのであろうか、一般辞書にはあったとしても和歌、歌ことば、俳句のような目的別辞書においては網羅されずに取り残されている可能性がある。この発見は本研究による成果である。

5. 主な発表論文等

〔雑誌論文〕(11 件)

① 田中牧郎, 山元啓史. 『今昔物語集』と『宇治拾遺物語』の同文説話における語の対応, 日本語の研究, 日本語学会, Vol. 10, no. 1, pp. 16–31, Jan. 2014. 査読あり

② ボル・ホドシチェク, 山元啓史. 不確かな情報が含まれる文の形式, 経済社会研究プロジェクト高度科学技術社会リスク・ソリューション 2012, 東京工業大学大学院社会理工学研究科, Vol. 2012, pp. 236–245, Mar. 2013. 査読なし

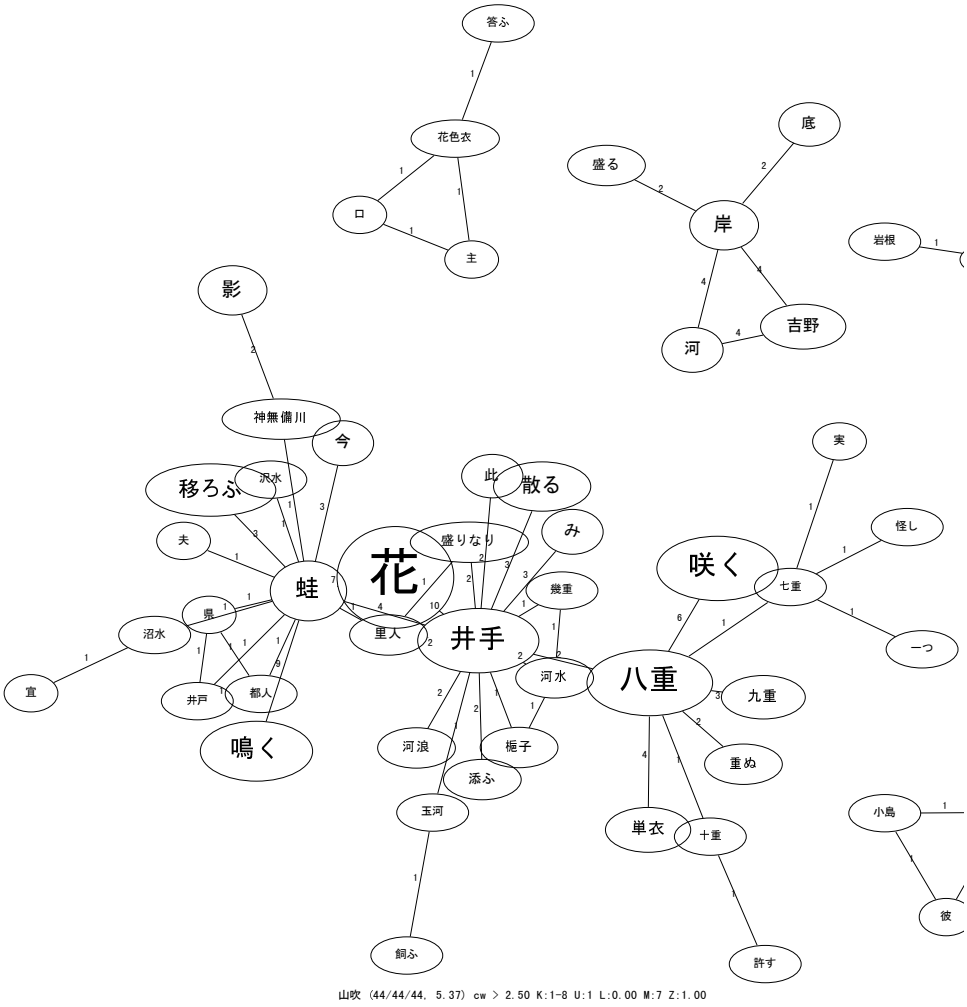

図３　「山吹（ここでは花）」のネットワークモデル:「蛙」「井出」「八重」の３つの用語のハブ（ノードを中心に放射状に他の用語と結びついている様子）が見える。

③ Bor Hodoscek, Hilofumi Yamamoto. Analysis and Application of Mid-Rank Lexicons of Modern Japanese, IPSJ Symposium 2013 Sig-CH, IPSJ Symposium 2013 Sig-CH, Pacific Neighborhood Consortium, Vol. 2013, No. 4, pp. 21–26, Dec. 2013. 査読あり

④ Makiro Tanaka, Hilofumi Yamamoto. A Corpus Study of Emotive Adjectives and Verbs of the Heian Japanese, SNPD2012, Proceedings 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE, Vol. SNPD.2012, No. 101, pp. 377–380, Aug. 2012. 査読あり

⑤ Hilofumi Yamamoto, Makiro Tanaka, Yasuhiro Kondo. Diachronic Corpus and Linguistic Space: New Methods for the Analysis of Language Change, SNPD2012, Proceedings 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE, Vol. SNPD2012, No. 101, pp. 381–384, Aug. 2012. 査読あり

⑥ 山元啓史. グラフを用いた集合演算による和歌用語の解析, 語彙研究, 語彙研究会, Vol. 9, pp. 86–94, Dec. 2011. 査読あり

⑦ Makiro Tanaka, Hilofumi Yamamoto. Quantitative Analysis of Loanwords of Eight Literary Works in the Heian Pe-

riod (794–1185), Osaka symposium on digital humanities 2011, Osaka simposium on digital humanities 2011, Vol. 1, No. 1, pp. 51–2, Sep. 2011. 査読あり

⑧ Hilofumi Yamamoto, Makiro Tanaka. Development of the thesaurus of classical Japanese poetic vocabulary, Asialex 2011, Lexicography: Theoretical and Practical Perspectives, Vol. 2011, 576–585, Aug. 2011. 査読あり

⑨ Makiro Tanaka, Hilofumi Yamamoto. An analysis of Sino-Japanese words of the Heian period for the development of the historical Japanese dictionary, Asialex 2011, Lexicography: Theoretical and Practical Perspectives, Vol. 2011, 496–505, Aug. 2011. 査読あり

⑩ 山元啓史. 「山吹」をめぐる和歌語彙の空間, じんもんこんシンポジウム 2011, 人文科学とコンピュータシンポジウム論文集, 情報処理学会, Vol. 2011, No. 8, pp. 141–146, Dec. 2011. 査読あり

⑪ 山元啓史. 八代集用語のモデリングシステム, じんもんこん 2010, 人文科学とコンピュータシンポジウム, じんもんこん 2010, 人文科学とコンピュータシンポジウム, 情報処理学会, Vol. 2010, No. 15, pp. 247–254, Dec. 2010. 査読あり

〔学会発表〕(11 件)

① Hilofumi Yamamoto. Lexical Modeling of Yamabuki (Japanese Kerria) in Classical Japanese Poetry, JADH2013 DH-JAC2013 Conference, JADH2013 DH-JAC2013 Conference Abstracts, Vol. 2013, pp. 62–63, Sep.19–21, 2013, Kyoto: Ritsumeikan University.

② Bor Hodoscek, Hilofumi Yamamoto. A Diachronic and Synchronic Investigation into the Properties of Mid-Rank Words in Modern Japanese, JADH2013 DH-JAC2013 Conference, JADH2013 DH-JAC2013 Conference Abstracts, Vol. 2013, p. 67, Sep.19–21, 2013, Kyoto: Ritsumeikan University.

③ Hilofumi Yamamoto, Makiro Tanaka, Yasuhiro Kondo. Design of Serial Comparison Model for the Diachronic Corpus Study of Japanese, JADH 2012, JADH 2012 conference abstracts, Vol. 2012, pp. 51–2, Sep. 15–17, 2012, Tokyo: University of Tokyo.

④ Makiro Tanaka, Hilofumi Yamamoto. Emotive Adjectives and Verbs of the Heian Japanese, JADH 2012, JADH 2012 conference abstracts, Vol. 2012, p. 52, Sep. 15–17, 2012, Tokyo: University of Tokyo.

⑤ 田中牧郎, 山元啓史. 平安時代日本語の感情形容詞と感情動詞: 『源氏物語』『今昔物語集』のコーパス分析を通して, 国立国語研究所国際シンポジウム「日本語の自他と項交替」, 2012 年 8 月 4–5 日, 東京: 国立国語研究所.

⑥ 山元啓史. IT を活用した日本語分析, 大阪電気通信大学情報学研究施設主催、公開ワークショップ「IT を活用した目的志向の日本語教育・運用支援」, 2012 年 3 月 20 日, 大阪: 電気通信大学.

⑦ 山元啓史, 田中牧郎, 近藤泰弘. 通時コーパスと言語空間論, コーパス日本語学ワークショップ, コーパス日本語学ワークショップ予稿集, 国立国語研究所言語資源研究系・コーパス開発センター, Vol. 1, No. 1, pp. 241–8, 2012 年 3 月 5–6 日, 東京: 国立国語研究所.

⑧ Hilofumi Yamamoto. Graph Representation of the Connotations of Classical Japanese Poetic Vocabulary, Osaka symposium on digital humanities 2011, Osaka symposium on digital humanities 2011, Vol. 1, No. 1, p. 42, Sep. 13–14, 2011, Osaka: Osaka University.

⑨ 山元啓史. BCCWJ 複合辞辞書の仕様・開発・評価, 特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研

究成果報告会）, 特定領域研究「日本語コーパス」平成 22 年度公開ワークショップ（研究成果報告会）予稿集, 文部科学省科学研究費特定領域研究「日本語コーパス」総括班, pp. 535–544, 2011 年 3 月 14–16 日, 東京: 国立国語研究所.

⑩ 山元啓史. ブーリアン演算による歌ことばモデルの解析, 第 16 回公開シンポジウム「人文科学とデータベース」, 第 16 回公開シンポジウム「人文科学とデータベース」論文集, 第 16 回公開シンポジウム「人文科学とデータベース」実行委員会, pp. 37–44, 2010 年 11 月 27 日, 京都: 花園大学.

⑪ 山元啓史. 通時コーパスで見る語彙論的トポロジーとトランジション, NINJAL 共同研究発表会・シンポジウム「通時コーパスの設計」研究発表会, 2010 年 3 月 3 日, 東京:国立国語研究所.

〔その他〕(1 件)

① 科学研究費助成金によるプロジェクト WEB ページ「和歌形態素解析用辞書開発のための用語連接規則に関する基礎研究」http://warbler.ryu.titech.ac.jp/~yamagen/waka/kaken2010.html

6. 研究組織

(1) 研究代表者
山元 啓史（Yamamoto, Hilofumi）
東京工業大学・留学生センター・准教授
研究者番号：30241756