An Introduction to Mathematical Linguistics for Historical Text Analysis

Tokyo Tech

Catalogue of Linguistics

Hilofumi Yamamoto, Ph. D.
Tokyo Institute of Technology

言語学と日本語教育を研究するための教室

言語と文化

山元研究室カタログ

目次

第 1 章

科学研究費助成金

1.1　2010–13 年度

研究題目「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」

1.2　2014–17 年度

研究題目「和歌用語シソーラスの開発と用語空間記述に関する基礎研究」

1.3　2018–21 年度

研究題目「歌ことばの効果的可視化技術と通時的変化の記述に関する基礎研究」

1.4　資料

2018 年度申請書（抜粋）

１作研究目的、研究方法など作

本研究計画調書は「小区分」の審査区分で審査されます。記述に当たっては、「科学研究費助成事業における審査及び評価に関する規程」（公募要領１１１頁参照）を参考にしてください。
本欄には、本研究の目的と方法などについて、３頁以内で記述してください。
冒頭にその概要を簡潔にまとめて記述し、本文には、(1)本研究の学術的背景、研究課題の核心をなす学術的「問い」、(2)本研究の目的および学術的独自性と創造性、(3)本研究で何をどのように、どこまで明らかにしようとするのか、について具体的かつ明確に記述してください。
本研究を研究分担者とともに行う場合は、研究代表者、研究分担者の具体的な役割を記述してください。

（概要） ※１０行程度で記述してください。

【背景と問題】本研究は可視化モデルを利用して、古代語の通時的語彙構造の変化を分析するものである。下記モデル (図 1,2) は**「吉野」**と**「桜」**の関係を数理的手法により可視化し、300 年間の比較を行った。２者間の通時的関係の変遷については明らかではあるが、すべての語についても同様に実現するには、1) 単語の長さをすべて短い単位で分割したため、**語句の比較が明確でない**、2) 多義語であるはずの語も**一義的に分類されている**、などの問題が残されている。

【目的】可視化技術を活かした古代語の通時研究はあまり多くはなく、本研究では古代語通時的変遷を効果的に可視化するシステムを構築し、基礎研究を行うことを目的とする。

【どこまで明らかに？】これまで（基盤研究 C）のデータでは辞書という静的な方式で蓄積してきたが、本研究では和歌から直接動的に 1) 単語の類似性情報の計算、2) 語と語の関係データの生成を行い、これら動的データと静的データとの差分をとり、通時的変遷を可視化する要因を明らかにする。

（本文）歌ことばの効果的可視化技術と通時的変化の記述に関する基礎研究

山 元 啓 史 (東京工業大学)

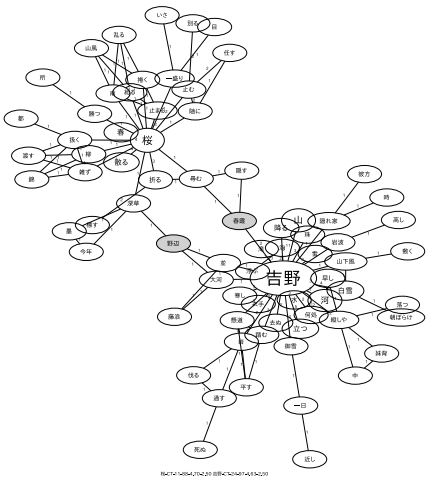


図 1: 古今集 (ca. 905) の「吉野」と「桜」: 古今の時代では吉野は桜の関係よりもむしろ雪と吉野の関係の方が強いことがわかる [20]。

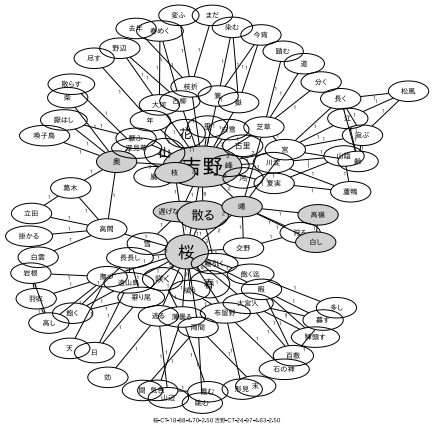


図 2: 新古今集 (1205) の「吉野」と「桜」: 専門家の間では桜と吉野の関係が一般的になるのは新古今になってからと言われている [20]。

【問題点】

現代語の理論的研究、自然言語処理の技術開発研究は、目覚ましく発展してきているが、古代語研究とりわけ通時研究は、コーパスのおかげで数こそ増え、古典資料の整備、人手による作業の多さ、評価の多様さなどの理由により、大規模かつ横断的な調査が実施された事例は多くない。

本研究の目的は、八代集（古今集 905 年頃から新古今集 1205 年）までの 300 年間の勅撰和歌集約 9500 首を対象に、古代語（和歌用語）を効果的に可視化するシステムの構築と通時的言語変遷の基礎研究を行うことである。これまでに基盤研究 (C) により、二十一代集（古今集 905 年から新続古

【1　研究目的、研究方法など（つづき）】

今集 1439 年）までの 20 巻を対象に、単語を抽出するためのシステムと辞書、および意味分類辞書（シソーラス）の開発を、機械処理と人手による目視修正を行ってきた。

近年、自然言語処理技術の成果により、人工知能を利用したテキスト処理が盛んに行われるようになり、これらの技術を駆使した応用が多く見られるようになってきた。しかしながら、言語の分析、中でも古代語の研究については、研究者のコンピュータ技術、数理的思考が直接的に人文科学領域になじまず、まだ十分に生かされてきていない。

【可視化システム開発の必要性と問題】

上記のモデルは、**吉野**と**桜**について、古今集 (図 1) から新古今集 (図 2) の 300 年間の通時的変遷を示したものであるが、このように可視化技術は通時的変化を要約・分析するのに便利である。しかしながら、他の語についても同様に比較・実行するには、

1）単語のサイズを一律的に決めたため、**当時の単語の成立が明確でない**、

2）多義語であるはずの語が**一義的に分類されている**、などの問題がある。

前者の解決法としては、Sentencepiece(ニューラル言語処理向けのトークナイザ; 教師なし、文脈依存、可逆式) 単語分割アルゴリズムの利用である。**後者**の解決法としては、Word2Vec などの Word Embedding[1,2]、分散表現と呼ばれている文脈から単語の意味特性を計算し、その意味次元をニューラルネットで圧縮（次元数を減らす）し、ベクトルの近さを類義語もしくは同語と判定する方法である。

【解決法】

機械学習を利用した言語分析の研究において、従来からも指摘されている問題は、単語の**類似性**と**関連性**がうまく区別されていないことである。単語の類似性と関連性というのは、たとえば、(梅, 桜) は類似している一方、(梅花, 枝) は関連してはいるが、類似はしていないということである。これらを区別して処理すれば、人間が行うモデリングに近くなるという報告があることから、和歌の処理においても良い効果が期待できる。

また、分散表現には単語の曖昧性が考慮されていないという問題点がある。単語にはさまざまな意味がある。たとえば、英語の “spring” という語は「温泉」「スプリング」「春」という意味がある。単語の多義性を考慮せず、１つの “spring” という形態を１つのベクトルで表すのには限界がある。むしろ、表記は同じであったとしても、ある文脈に挟まれた語の表記を一旦伏せておき、仮に x_i とし、文脈から得られたベクトルにしたがって、x_i の分散表現を与える方法を考えれば、異なるベクトルを同じ表記で示す必要はなくなる。これは、文脈の隔たりの大きいベクトルを x_i の添字毎に分割し、多義性のある単語を用法・文脈ごとに記述する方法である。その結果、単語の用法の弁別性能が向上したことが報告されている [3,4]。

【目的】

古代語は現代語とは異なり、従来より可視化技術を利用した通時的言語体系の研究は今までに多くなく、限られた資料から、目視によって丹念に分析していくものが大半であった。

一方で、現代語の分析に大いに利用と期待が寄せられている自然言語処理技術は目覚ましい発展があり、人工知能技術、ニューラルネット、ベイズ統計学、時系列分析などの基礎技術と融合し、近年大きな成果を収めている。

古代語は言語資料が限られており、現代語のような新しいデータが次から次と出てくるものではないが、今までの成果を利用し、蓄積を取り込みつつ、総合することで、少なくとも考え方を取り入れることで成果を収められると考えている。学術的創造性として注目するのは、古代語通時研究のための効果的な可視化システムと語彙データベースの開発である。

【古代語へのチャレンジ】

上記で述べた自然言語処理技術の応用が解決策として有力ではあるが、和歌（古代語）というテキストの特性として、1）現代語のように大量のデータがあるわけではなく、データ量は限られた上

【1　研究目的、研究方法など（つづき）】

で研究を進めなければならないこと、2）現存するテキストは何らかの理由（希少価値、読み継がれてきたほどの魅力、消失・散逸せずに残存している現状、長年にわたっても理解できる内容）で、テキストの内容、語の意味が限られている可能性はあること、から考えると、現代語でできることと、古代語にできることと隔たりがあることに注意すべきであり、簡単ではないことが予想される。ただし、上記の点が本研究の**チャレンジ**であり、可能性が見えれば、通時的言語研究への貢献は大きいと考える。

【ゴール設定：何をどこまで？】

二十一代集のすべてについて行うのではなく、基本的な八代集についてのみを対象とし、これをこの 4 カ年のゴールとして設定し、着実に成果をあげる計画を実行する。ただし、単語の単位の切り出し推定実験には、できるだけ多くのデータを用いた方が有利なので、万葉集、二十一代集の和歌本文データを利用する。

古代語も現代語と同様にデータは辞書形式（静的）で蓄積されてきたが、本研究では、和歌データのみから動的に 1) 単語の類似性情報の計算、2) 語と語の関係データの生成を行い、開発済みの静的データとの差分をとり、動的表示を可能にするための要因を明確にし、可視化を実現する。

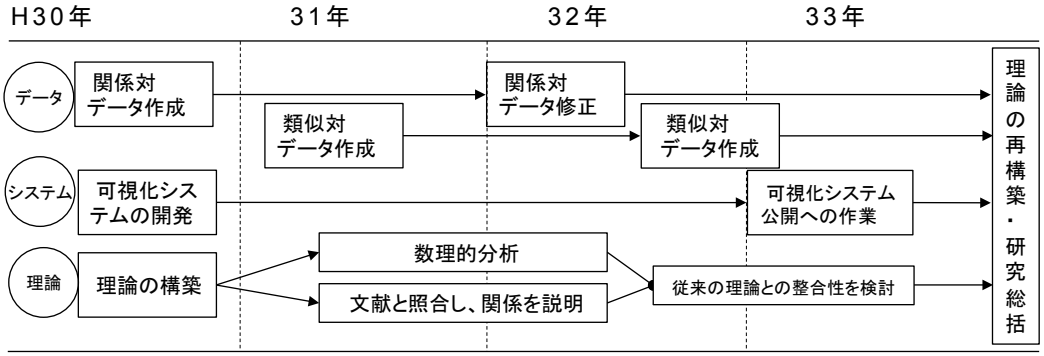


図 3: 研究計画・ロードマップ: データ、システム、理論の 3 要素で構成する。

参考文献

[1] Le, Quoc V. and Tomas Mikolov (2014) “ Distributed Representations of Sentences and Documents, ” CoRR, Vol. abs/1405.4053, URL: http://arxiv.org/abs/1405.4053.

[2] Tomas Mikolov, Quoc V. Le and Ilya Sutskever (2013) Exploiting Similarities among Languages for Machine Translation, CoRR.

[3] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013) “Efficient Estimation of Word Representations in Vector Space,” CoRR, URL: http://arxiv.org/abs/1301.3781.

[4] Řehuřek, Radim and Petr Sojka (2010) “Software Framework for Topic Modelling with Large Corpora,” in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50, Valletta, Malta: ELRA, May, http://is.muni.cz/publication/884893/en.

２作本研究の着想に至った経緯など作

本欄には、(1)本研究の着想に至った経緯 、(2)関連する国内外の研究動向と本研究の位置づけ、(3)これまでの研究活動、(4)準備状況と実行可能性、について１頁以内で記述してください。
「(3)これまでの研究活動」の記述には、研究活動を中断していた期間がある場合にはその説明などを含めても構いません。

【1. 着想に至った経緯】

研究代表者がこれまでに 15 年以上の歳月 (2001 年より) をかけて作成してきた和歌用形態素解析辞書とシソーラス（語彙体系用語集）を使い、和歌の語彙体系を効果的に可視化するための技術を開発し、さらに通時的言語記述として適切であるかどうかを検証してきた（尚、形態素辞書とシソーラスは、25 年度までの基盤研究 (C) で二十一代集対応版が完成している)。

【2. 国内外の研究動向と位置づけ】

Word2vec を始めとする Word Embedding（分散表現）に関わる研究も、ツールも数多く発表されており、自然言語処理研究においては大いに理論化がされており、技術の応用も多々行われている。ただし、言語学、とりわけ古代語や通時的変化を分析するまでには至っておらず、今までの古代語研究の成果を言語処理に利用する方法論の検討が待たれている。海外の日本語・日本文化の研究者についても、技術の導入は徐々に行われてはいるものの、テキスト処理・機械学習などの基盤となる技術を使った研究成果はほとんど行われていない。

【3. これまでの研究活動】

◎漸近的語彙対応推定法 [4][6]: 単語対相互情報量により推定した語対応の技術を取り入れ、今までの人間によるシソーラス作りの弱点を改善し、シソーラス体系作りの自動化と理論化を試みた。

◎二十一代集シソーラスの開発 [8]: 表 1 に示すように思いも寄らぬ表記が多数出現するため、シソーラスを開発した。

表 1: シソーラスなしでは同じ語として計算できない例（一部）

かな表記	実際に和歌に出現する実例
たつた	立田, 竜田, 龍田, ...
たつらむ	立つらん, 立らん, 立覧, ...
ちぎりけむ	契りけん, 契けむ, 契けん, 契剣, ...
おもふへふ	思ふてふ, 思てふ, 思ふ蝶, 思蝶, ...
えてしがな	得てしかな, 得てし哉, ...

◎語彙コミュニティの分析 [2]：R の Linkcomm を用い、単語をコミュニティとして、語群としての意味を検討した。

◎和歌用可視化システムの開発 [7]：D3.js を用いて、グラフ図形の生成とノードをクリックすることで、原典の和歌のリストが閲覧できる可視化システムを開発した。

◎自動タグ付けシステムの開発 [21]：当時、和歌用の辞書がなかったために、単語辞書、連接辞書を開発し、単語に切り分けるシステムを開発した。

【4. 準備状況と実行可能性】◎技術・材料・資料関係

これまでの基盤研究 (C) で培ってきた辞書・シソーラス、和歌本文データ、現代語訳データなどの基礎的な材料はすでに整備されている。一連の Word Embedding の理論と技術は広く公開されており、入手済みであるので、データ的にも技術的には実行可能である。

◎分担関係

ホドシチェク（阪大）は、機械学習技術を古代語の分析に応用、プログラミング、可視化技術を担当する。山元啓史（東工大）は、関連対・類似対データの開発、通時的分析手法の開発、研究総括を担当する。また、両者は、可視化システムをバイリンガルで表示するために、古代語の日英語対応関係を分析するシステムの開発を行う。

◎うまくいかない時の対応策

これまでの研究実績 [1][2] により機械学習で解析できる可能性はかなり高いが、和歌の根本的な限られたデータ量の都合により、うまくいかない時には、これまでの研究で利用したデータとの融合を考え、その上で、なぜうまく行かないのか、なぜシソーラスと連動させる必要があったのかを考察し、理論的な説明を構築し、研究の貢献とする。

以上を網羅した上で、**データ処理による通時的な視点での古代語の空間記述研究が開始でき、**この領域への貢献となるだろう。

３作研究代表者および研究分担者の研究業績作

本欄には、研究代表者、研究分担者がこれまでに発表した論文、著書、産業財産権、招待講演のうち重要なものを選定し、現在もしくは過去から発表年次の順に、通し番号を付して２頁以内で記入してください。なお、学術誌へ投稿中の論文を記入する場合は、掲載が決定しているものに限ります。

学術誌論文の場合、論文名、著者名、掲載誌名、査読の有無、巻、最初と最後の頁、発表年（西暦）を記入してください。以上の項目が記入されていれば、各項目の順序の入れ替えや、著者名が多数の場合、主な著者名のみ記入しその他の著者を省略することは問題ありません。なお、省略する場合は、省略した員数と、研究代表者、研究分担者が記載されている順番を○番目と記入してください。

研究代表者には二重下線、研究分担者には一重下線を付してください。

1. H. Yamamoto, and B. Hodošček. Relationships between Flowers in a Word Embedding Space of Classic Japanese Poetry, Doshisha University, JADH2017 Proceedings of the 7th Conference of Japanese Association for Digital Humanities “ Creating Data through Collaboration ”, Faculty of Culture and Information Science, Doshisha University, Vol. 2017, pp. 70-72, (2017) (査読有).

2. H. Yamamoto and B. Hodošček. “Development of the dictionary of poetic Japanese description”, Digital Scholarship in History and the Humanities, the 6th conference of the Japanese Association for Digital Humanities, Japanese Association for Digital Humanities 2016 pp. 44–46, (2016)（査読有）.

3. 山元啓史. “通時コーパスによる言語の研究”, コーパスと日本語史研究, ひつじ書房, pp. 17–35, (2015) (査読有).

4. 山元啓史, ホドシチェク・ボル, 村井源, “二十一代集シソーラスのための漸近的語彙対応システムの開発”, じんもんこんシンポジウム 2014, 人文科学とコンピュータシンポジウム論文集, Vol. 2014, No. 3, pp. 157–162, (2014) (査読有).

5. 山元啓史. “目で見てわかる歌ことば”, 日本語学, 明治書院, Vol. 33, no. 14, pp. 172–183, (2014) (査読無).

6. H. Yamamoto, B. Hodošček, and Hajime Murai. “Development of an Asymptotic Word Correspondence System between Classical Japanese Poems and their Modern Translations”, JADH Conference 2014, JADH Conference 2014 ABSTRACT, p.40, (2014) (査読有).

7. H. Yamamoto, B. Hodošček, and Makiro Tanaka, “A Visualization and Analysis System for Japanese Language Change: Quantifying Lexical Change and Variation using the Serial Comparison Model”, JADH Conference 2014, JADH Conference 2014 ABSTRACTS, p. 3, (2014) (査読有).

8. H. Yamamoto, and B. Hodošček. “Thesaurus of classical Japanese poetic vocabulary for the Nijuichidaishu (ca. 905–1439)”, 14th International Conference of European Association for Japanese Studies, 14th International Conference of European Association for Japanese Studies BOOK OF ABSTRACTS, p.86, (2014) (査読有).

9. B. Hodošček and H. Yamamoto, “A Diachronic and Synchronic Investigation into the Properties of Mid-Rank Words in Modern Japanese” The Japanese Association for Digital Humanities, the third annual conference at Ritsumeikan University, Kyoto, Japan, September 19-21, pp. 67–8. (2013) **査読有**.

10. H. Yamamoto, “Lexical Modeling of Yamabuki (Japanese Kerria) in Classical Japanese Poetry”, The Japanese Association for Digital Humanities, the third annual conference at Ritsumeikan University, Kyoto, Japan, September 19-21, 62–3, (2013) **査読有**.

【3 研究代表者および研究分担者の研究業績（つづき）】

11. H. Yamamoto, M. Tanaka, Y. 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, 381–384, (2012) **査読有**.

12. M. Tanaka, and H. Yamamoto, “Emotive Adjectives and Verbs of the Heian Japanese”, JADH 2012 conference abstracts, Vol. 2012, p. 52, (2012) **査読有**.

13. H. Yamamoto, M. Tanaka, and Y. Kondo, “Design of Serial Comparison Model for the Diachronic Corpus Study of Japanese”, JADH 2012 conference abstracts, Vol. 2012, 51–2, (2012) **査読有**.

14. 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, 86–94, (2011) **査読有**.

15. H. Yamamoto, and M. Tanaka, “Quantitative Analysis of Loanwords of Eight Literary Works in the Heian Period (794–1185)”, Osaka simposium on digital humanities 2011, Vol. 1, No. 1, 51–2, (2011) **査読有**.

16. H. Yamamoto, “Graph Representation of the Connotations of Classical Japanese Poetic Vocabulary”, Osaka simposium on digital humanities 2011, Vol. 1, No. 1, p. 42, (2011) **査読有**.

17. M. Tanaka, and H. 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, 496–505, (2011) **査読有**.

18. H. Yamamoto, and M. Tanaka, “Development of the thesaurus of classical Japanese poetic vocabulary”, Asialex 2011, Lexicography: Theoretical and Practical Perspectives, Vol. 2011, 576–585, (2011) **査読有**.

19. 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, 141–146, (2011) **査読有**.

20. 山元啓史, “八代集用語のモデリングシステム”, じんもんこんシンポジウム 2010, 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. 2010, No. 15, 247–254, (2010) **査読有**.

21. 山元啓史, “分類コードつき八代集用語のシソーラス”, 日本語の研究, 日本語学会, Vol. 5, No. 1, 46–52, (2009) **査読有**.

第 2 章

受賞

2.1　2015 年度情報処理学会山下記念研究賞

2015 年度情報処理学会山下記念研究賞を受賞。山下記念研究賞は、情報処理学会が主催する研究会およびシンポジウムにおける研究発表のうち、特に優秀な論文の発表者に授与される賞。初代情報処理学会会長の故山下英男氏寄贈の資金にて運営されている。

2.2　2017 年度じんもんこん 2017 ベストポスター賞を受賞

リベラルアーツ研究教育院の山元啓史教授らの研究が、じんもんこん 2017 ベストポスター賞を受賞。12 月 9・10 日に大阪市立大学杉本キャンパスで開かれた人文科学とコンピュータシンポジウム「じんもんこん 2017」（主催・一般社団法人情報処理学会人文科学とコンピュータ研究会）において、リベラルアーツ研究教育院の山元啓史教授らの研究が、ベストポスター賞を受賞した。

山元啓史准教授が山下記念研究賞を受賞

受賞・表彰　研究

2015.11.25

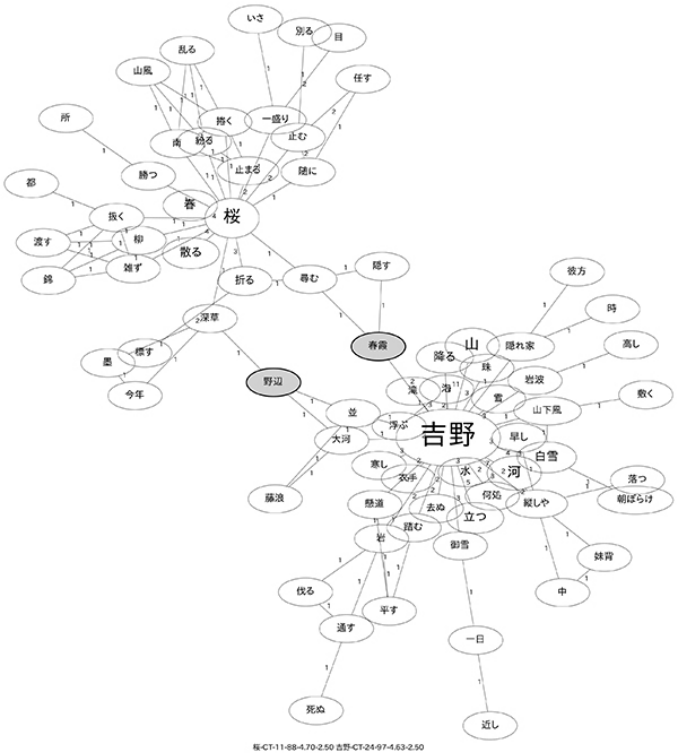
留学生センターの山元啓史准教授が、2015年度情報処理学会山下記念研究賞を受賞しました。山下記念研究賞は、情報処理学会が主催する研究会およびシンポジウムにおける研究発表のうち、特に優秀な論文の発表者に授与される賞です。初代情報処理学会会長の故山下英男氏寄贈の資金にて運営されています。授賞式は、来年3月情報処理学会全国大会で行われる予定です。

受賞テーマ

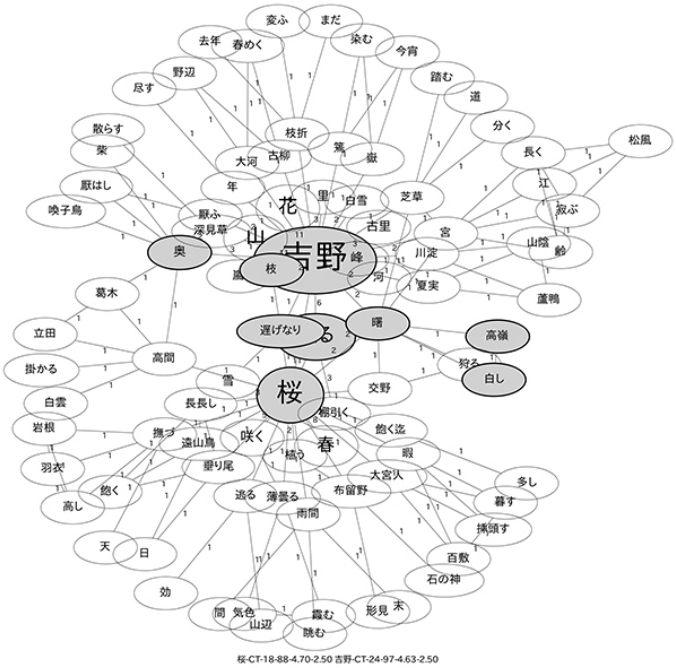
二十一代集シソーラスのための漸近的語彙対応システムの開発

受賞理由

本論文は、勅撰和歌集である二十一代集のシソーラス開発を介して、1000年以上前の日本語における語彙体系を明らかにすることを目的としています。和歌と現代語訳のパラレルコーパスを利用した単語対の推定法により、計算手続きだけで新規登録すべき語を抽出することが可能なアルゴリズムを提案・実証し、今後、計量国語学の分野で、他の古典籍研究にも応用可能な優れた業績であると認められました。また、本論文が、10年以上にわたって研究された古典における知識の体系化の実現のための取り組みの成果であり、その積み重ねは人文科学研究の基盤的な研究として位置づけられると同時に、今後の発展性が期待できるという点でも高く評価されました。



古今和歌集（905年ごろ）における桜と吉野山との関係：
「桜といえば吉野山」という関係はこの時代にはまだ成立しておらず、むしろ雪山、修験道、隠遁者と吉野の関係の方が強い。



新古今和歌集（1205年）における桜と吉野山との関係：
灰色のノードは2つ（桜、吉野）のネットワークで共有している用語であり、古今集のネットワークモデルと比較すると、古今集以来の300年間で、桜と吉野の関係が密になってきていることがわかる。一般的に、西行法師の時代には「桜といえば吉野山」が定着したと言われている。

今回の受賞を受けて、山元准教授は以下のようにコメントしています。

山元啓史准教授

長年に渡り、コンピュータによる日本語の言語の歴史を研究しております。
この研究は日本語を外国人に教えるために、入門期にざっくり「日本語はこんな形をしているんだよ」と伝えられたら、どんなに気が楽だろうと思ってはじめたことでした。それがいつの間にか日本語の普遍性を尋ねる研究となってきました。

普遍性を明らかにするには、言語の変化や変動を調べ、その相対的な違いを明らかにすることが必要になりました。一般的に「言語は、いったん形が定着すれば、その形は変わりにくいが、その意味はどんどん変化していく」と言われています。日本語は形式的には世界の言語の中でもごくごく一般的です。しかし、日本語のように多くの人間（たとえば1億人以上）1000年以上もの間、話させている言語は、ごくわずかです。これは、日本語は1000年以上言語をさかのぼって、調査できる言語であるということなのです。

「日本語は、どんな形をしているのか」という疑問については、まだまだ十分ではありませんが、今後も、日本語に限らず、言語というものはどんな変化をしていくのかを研究していきたいと考えています。

- 2015年度（平成27年度）山下記念研究賞詳細 情報処理学会
- 山元啓史研究室
- ひらめき ときめきサイエンス「目で見てわかる昔の日本語、今の日本語」開催

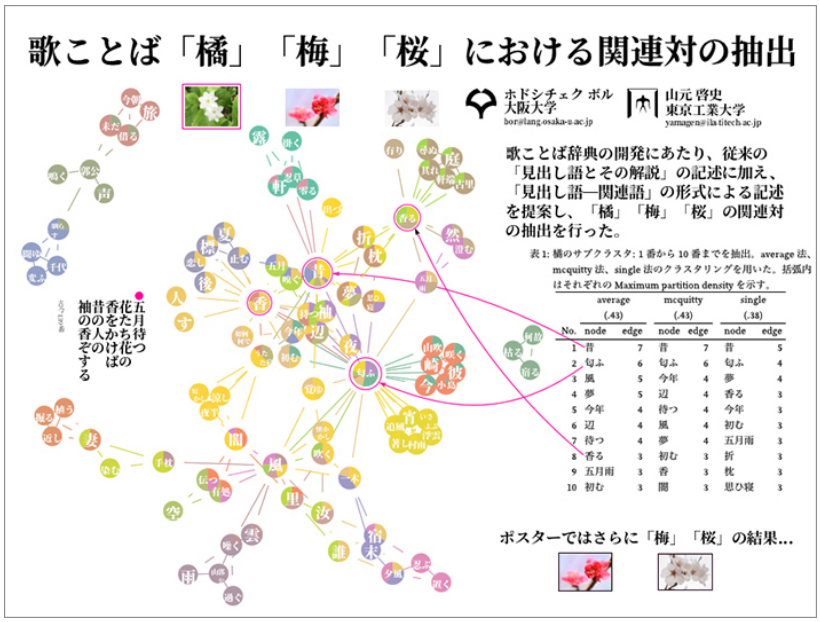
お問い合わせ先
留学生センター 山元啓史
Email : yamagen@ryu.titech.ac.jp
Tel : 03-5734-2324

The research of Classical Poetic Vocabulary won the best poster award of the conference of Computer and Humanities 2017

A study on the extraction of relational pairs of 'orange', 'plum', and 'cherry' flowers in poetic Japanese

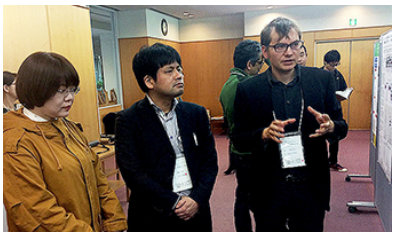
December 28, 2017

Research from Institute for Liberal Arts Professor Hilofumi Yamamoto was awarded the Best Poster Prize at the Computers and the Humanities Symposium, organized by the Information Processing Society of Humanities and Computer Studies, and held at the Sugimoto Campus of Osaka City University on December 9th and 10th.



Slide of Lightning Talk

The award-winning title was "A study on the extraction of relational pairs of ‘orange’, ‘plum’, and ‘cherry’ flowers in poetic Japanese." The research frames poetic Japanese vocabulary from the Heian period and onward in terms of leading and supporting characters in a play, extending the traditional dictionary entry description of leading words such as ‘orange’, ‘plum’, and ‘cherry’ flowers by showing their connection to their supporting words.
The selection of supporting words was achieved through objective algorithmic means devoid of the subjective interference of modern Japanese knowledge. For example, the leading role of the orange ‘tachibana’ flower, traditionally associated with summer, now inter-linked with the supporting roles of ‘mukashi’ (old), ‘ka’ (fragrance), ‘niofu’ (smell), ‘makura’ (pillow), and ‘yume’ (dream) allows an interpretation of ‘tachibana’ to be “reminded in a nostalgic dream of the smell of an old lover”. This linked structure was found to be closely related to the waka “五月待つ、花たち花の、香をかげば、昔の人の袖の、香ぞする” found in the 60th section of the Tales of Ise.


Presentation by Professor Hodoscek

The method used in this research does not come from traditional linguistic principles but is grounded in network community analysis, where the focus is on analyzing the community structure of connections between people. For example, while leading cast members Brad Pitt, Tom Cruise, and Johnny Depp rarely star in the same movies, supporting cast members such as Kevin Bacon (cf. Bacon Number) take on many roles and freely co-star with them. Just as there exist words that convey a strong and impressive meaning, there are words that are less memorable but that when connected with other words convey a potent message together. This connection was found to explain the different roles and meaning of polysemious words—words that can mean different things in different contexts.

Through this method, Professor Yamamoto's research group is trying to develop an analysis and machine description of ambiguity in language from the viewpoint of historical linguistics.

The research awarded was based on the JSPS Grant-in-Aid for Scientific Research (C) “Basic research on the development of Classical Poetic Thesaurus in terms of time-space description.” This content was also presented under the JSPS "Hirameki ☆ Tokimeki Science" program for junior high school students.


Award ceremony


Ceremony speech

The researcher collaborator on this project, Dr. Bor Hodošček, got a doctor of engineering degree at Tokyo Institute of Technology and is now an Associate Professor at the Graduate School of Language and Culture, Osaka University. The poster award is a recognition of the intuitive and effective presentation of conveying the ambiguity of language devised by Dr. Hodošček.

- Annual Conference for Computer and Humanities 2017 (in Japanese)
- Hilofumi Yamamoto Lab

第 3 章

サイエンス・カフェ神戸
「目で見てわかる歌ことばの姿」

3.1　サイエンス・カフェ神戸でのトーク

> サイエンスカフェ (Science Café) は、1997 年から 1998 年にかけて、イギリスとフランスで同時発生的に行われたのが起源とされる、カフェのような雰囲気の中で科学を語り合う場、もしくはその場を提供する団体の名前である。英国での呼称に倣ってカフェ・シアンティフィーク (Café Scientifique) と呼ぶこともある。
>
> サイエンスカフェ—Wikipedia

2008 年 3 月 24 日、神戸大学主催のようこそサイエンスカフェ神戸へで発表した。

3.2　開催報告

次ページ pdf。

開催記録 サイエンスカフェ神戸

「目で見てわかる歌ことばの姿」

ゲスト： 山元啓史 (オーストラリア国立大学アジア研究学部客員研究員)
場　所： 神戸花鳥園
日　時： ２００８年3月２４日（月）１５時～１７時
協　力： 神戸花鳥園

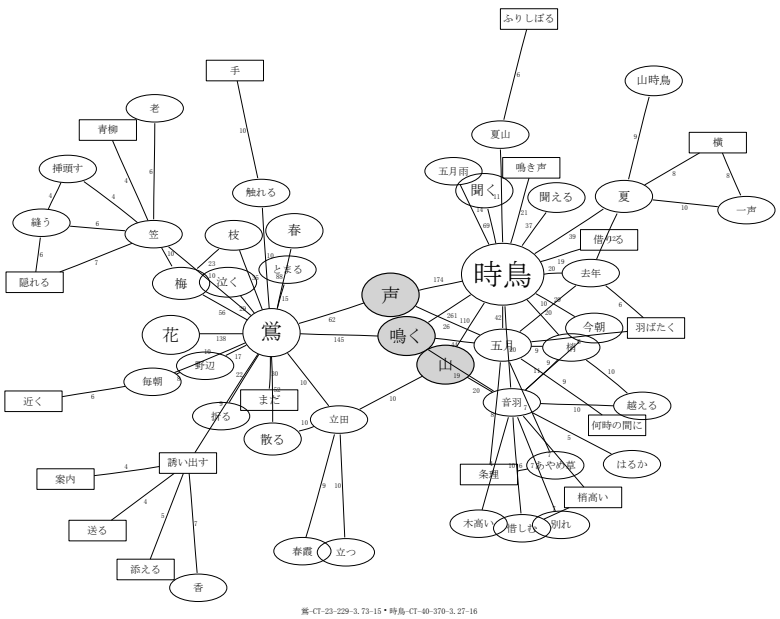
開催案内文の記録

今回は、古典（和歌）とコンピュータを結んで古の世界を垣間見る話題です。和歌は千年以上も前より日本に存在する伝統的な文芸です。この和歌に使われることばを「歌ことば」と言いますが、今回のサイエンスカフェでは、コンピュータを使って、歌ことばのネットワーク図（単語を網の目状に並べつないだ図形）を作り、和歌にはどのようなメッセージが示されているのか、目で見てわかる仕組みについてお話しをしていただきます。

当日の様子

2007 年度最後のサイエンスカフェは，光あふれる平日午後．花と緑と鳥 (!) に囲まれて，神戸花鳥園での開催となりました．写真をご覧下さい，もう，素晴らしいとしか言いようのない，花，花，花….. 花鳥園のスタッフのみなさまが，毎日愛情を込めて育てておられる花の中で，お茶を片手にゆったりとした気分で始まりました．

本日のゲストは山元啓史さん．歌ことば（和歌の言葉) の使われ方を，コンピュータを用いて客観的に解析し，グラフ理論からその潜在的な意味 (コノテーション)[*1]を引き出すことで時代の風景を捉えようとされている研究者です．従来の人文学的研究手法において，歌ことばの意味するところは，研究者 (受け手) の印象に依存しかねないのが現状．それに対し，山元さんは「客観的に捉える」事によって，新しく普遍的な見方を引き出しておられます．



鶯-CT-23-229-3, 73-15・時鳥-CT-40-370-3, 27-16

話題は「鶯」と「時鳥 (ホトトギス)」を例題として始まりました (まさに花鳥園らしい話題提供です！)．古今集において，和歌 (原文) と現代語訳のテキストの中に，それぞれ，この言葉が他の語とどんなパターンでどのくらい使われているかを調べます．この，和歌 (原文) での使われ方 (パターン) と現代語訳での使われ方 (パターン) の論理積というのは，辞書に書いてある「よくある使い方・意味」になりますから，それ以外のパターンに注目し，言葉のネットワーク図を作ります．

「科学は文化にかわるものではないけれど，科学を使うことで，こうやって文化に広がりを足すことができるのです」という山元さんの言葉が大変印象的でした．

「言葉」の世界に「解析のための “ 数式 ” が入ってきて，参加者のみなさんは最初戸惑い気味でしたが，山元さんの丁寧な説明と (何度も同じ質問が出ても，言葉を変えて，本当に丁寧に説明してくださいました) やわらかい語り口に，歌ことばの世界を堪能したカフェとなりました．（文責：橋口典子 2006 サイエンスカフェ神戸）

[*1] 「コノテーション」:「タコ (蛸)」という言葉ひとつでも，蛸そのものを意味するだけでなく，ある人は蛸から「やわらかさ」を連想し，別のある人は「気持ち悪さ」を連想し，別のある人は「タコ焼き」を連想するかもしれない．このように，言語記号の潜在的な意味―は文化によって違う場合があり，それゆえ，コノテーションを知ることは，文化を調べるのに重要になってきます (山元さん談)．

第 4 章

ひらめき☆ときめきサイエンス

略称「ひらとき」と呼ばれるセミナーは、大学や研究機関で「科研費」（KAKENHI）により行われている最先端の研究成果に、小学５・６年生、中学生、高校生が、直に見る、聞く、触れることで、科学のおもしろさを提供する科学日本学術振興会からの委託事業ある。研究費助成金と同じ取り扱いである。

整理番号	ＨＴ３００６８	分野	その他・人文	キーワード	言語学、情報科学

研究機関名	東京工業大学				
プログラム名	目で見てわかる昔の日本語と今の日本語：タイムマシンに乗らずに行ける昔の世界				
先生（代表者）	山元　啓史（やまもと　ひろふみ） リベラルアーツ研究教育院・教授				
自己紹介	大学卒業から今までずっと外国人に対して日本語を教えてきました。教えているうちに「ことばはどんな形をしているのだろう」と思い、言語学を勉強しはじめました。いろいろなことばを知りたくなって、アメリカとオーストラリアに留学しました。世界のことばと日本のことばを比較したり、昔のことばがどうして今の形になったかを調べたりするようになりました。ことばは誰もが使える楽しい宝物のように思います。さあ、みんなでことばについてお話しましょう。				
開催日時・募集対象	①平成 30 年 8 月 1 日（水） ②平成 30 年 12 月 26 日(水)	受講対象者	①中学生 ②小学 6 年生	募集人数	①20 名 ②10 名
集合場所・時間	東京工業大学西 1 号館 1 階ラウンジ	（集合時間）	9:50		
開催会場	東京工業大学大岡山キャンパス 住所：〒152-8550　東京都目黒区大岡山 2-12-1 W1-8 アクセスマップ URL：http://www.titech.ac.jp/maps/index.html				

内　　容

ことばは時代につれて変化します。私たちの知っていることばの意味は今の意味で、昔の意味とはまったく同じではありません。もしタイムマシンに乗って昔の日本語が聞けたなら「あれぇ～何か変だ！違うぞ？」と思うことでしょう。大昔の録音は残っていませんから、実際に聞くことはできません。しかし、昔の文章からことばの使われ方を図に描いて見ることはできます。そんな目で見てわかる昔のことばの世界についてお話しします。

スケジュール

09:50～10:00　受付（大岡山キャンパス西１号館１階ラウンジ集合）
10:00～10:20　開講式：あいさつ、科研費の説明
10:20～11:00　自己紹介：参加者、ご父兄の皆様、山元研究室学生
（終了後 15 分休憩）
11:15～12:00　講義：ことばの意味を図で見る仕組み
12:00～13:00　ランチタイム：サンドイッチ、みんなでおしゃべりしながら、楽しく食べましょう。
13:00～14:00　実習：コンピュータで自分のネットワークを描こう。
14:00～15:00　休憩：クッキータイム
15:00～16:00　お散歩：鳥人間コンテストのマイスターを訪問しよう！
16:00～16:40　発表会：みんなで意見と感想を述べよう！
16:40～17:00　修了式：アンケート記入、未来博士号授与、写真撮影
17:00　終了・解散：お疲れさま。
（8 月、12 月とも同じスケジュールで行います）

持　ち　物

筆記用具

特記事項

中学生：夏ですので、飲料水、汗を吸いやすいハンカチなど忘れずに。
小学生：冬ですので、暖かい服装で。

《お問合せ・お申込先》

所属・氏名 ：	リベラルアーツ研究教育院・山元啓史
住　　所 ：	東京都目黒区大岡山 2-12-1 東京工業大学 W1-8
TEL 番 号 ：	03-5734-2324
FAX 番 号 ：	03-5734-2324
E-mail ：	yamagen@ila.titech.ac.jp
申込締切日 ：	**中学生：平成 30 年 7 月 18 日（水）** **小学生：平成 30 年 12 月 5 日（水）**

※セミナーは参加者が話し合いをしながら、進めていきます。おしゃべりが大好きな皆さん、お待ちしております。ぜひご応募ください。

※当プログラムは先着順にて受付を行います。

※毎年かなり多数のご応募がございます。締切日前に締め切りとさせていただくこともございます。お早めにお申込みください。お申込みの際には、ぜひ日本学術振興会申し込みフォームのコメント欄に「参加の動機」をお書きください。応募者多数の場合は、その文面で選考させていただきます。楽しいコメントをお待ちしております。選考結果は、中学校（８月）・小学校（12 月）のそれぞれ開催 2 週間前に電子メールにてご連絡いたします。あらかじめご了承ください。

《プログラムと関係する先生（代表者）の科研費》

研究代表者	研究期間	研究種目	課題番号	研究課題名
山元啓史	H30-H33	基盤研究(C)	18K00528	歌ことばの効果的可視化技術と通時的言語変化記述に関する基礎研究
山元啓史	H26-H29	基盤研究(C)	26370530	和歌用語シソーラスの開発と用語空間記述に関する基礎研究
山元啓史	H22-H24	基盤研究(C)	22520458	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

★この科研費について、さらに詳しく知りたい方は、下記をクリック！

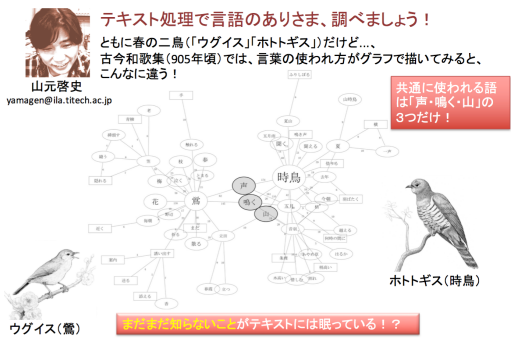
http://kaken.nii.ac.jp/

※国立情報学研究所の科研費データベースへリンクします。

平成29年度
ひらめき☆ときめきサイエンス～ようこそ大学の研究室へ～KAKENHI
（研究成果の社会還元・普及事業）
実　施　報　告　書

ＨＴ２９０８４

目で見てわかる昔の日本語と今の日本語：タイムマシンに乗らずに行ける昔の世界



開　催　日：	平成29年8月2日（水）
実施機関： （実施場所）	東京工業大学 （大岡山キャンパス）
実施代表者： （所属・職名）	山元　啓史 （リベラルアーツ研究教育院・准教授）
受　講　生：	中学生21名
関連ＵＲＬ：	https://cuckoo.js.ila.titech.ac.jp/~yamagen/hirameki2017.html

【実施内容】

■受講生に分かりやすく研究成果を伝えるために以下のような工夫をしました。

〇全員（参加者、協力者、保護者、事務局）で自己紹介をし、互いに話しやすい雰囲気を作りました。

〇聴講するのではなく、保護者も含め、３～４名のグループに別れ、ディスカッションで進めました。

〇ワークブックを作成し、参加者全員が自ら考え、自分で研究の要所が書き込めるようにしました。

〇ワークブックをすべて書き込んだら自由研究レポートができあがるようにしました。

〇大学の研究も中学校の勉強と関係づけながら、ディスカッションを進めました。

〇和歌については中学の国語便覧を用い、具体的にページ数を示し、後日復習できるようにしました。

〇言語学でも数学を使うことを示し、関数電卓を用いて、単語の重み計算を実習しました。

〇研究内容だけでなく、言語学の基礎（世界に言語はいくつあるか）、数学の基礎（対数とは、心理尺度とは）、研究の基礎（「特徴とは何か」、「似ている」と「同じ」「違う」とは）など皆で考えました。

〇散歩の時間をとり、学内の建物、ものつくりセンターを見学し、鳥人間コンテストで有名なサークルの協力を得て、人力飛行機、ものつくりの実物に触れることができました。



〇保護者の皆様のお席もご用意し、参加者と同じワークブックを使って討論に参加いただきました。
〇参加者が考えている間、保護者の皆様には研究内容や大学で行われている教育の紹介を行いました。
〇復習できるように、研究室のウェブの特設ページに、当日の記録と写真を掲載しました。

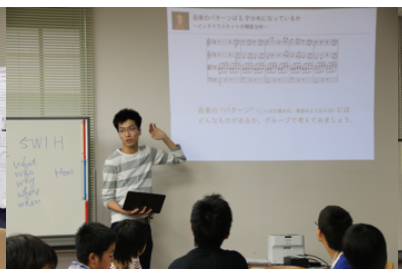
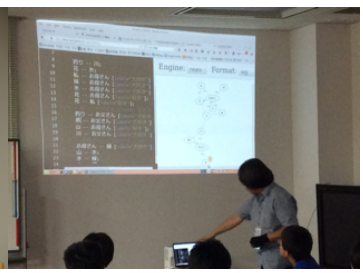


■当日のスケジュール

09:50～10:00　受付（大岡山キャンパス西１号館１階ラウンジ）
10:00～10:15　開講式：あいさつ、科研費の説明
10:20～11:00　自己紹介：参加者、保護者の皆様、研究室学生、研究企画課職員
11:00～12:00　講義：ことばの意味を目で見る仕組みとは何か。
12:00～13:20　ランチタイム（サンドイッチを食べました）
13:20～14:30　実習：コンピュータで自分のネットワークを描こう。
14:30～15:00　休憩：クッキータイム
15:00～16:00　お散歩：鳥人間コンテストのマイスターを訪問しよう
16:00～16:30　発表会：みんなで意見と感想を述べよう！
16:30～17:00　修了式：アンケート記入、未来博士号授与、写真撮影
17:00　終了・解散

■実施の様子

　実施はスケジュールのとおりですが、学術的には以下の内容を盛り込みました。
①言語学概論：世界の言語、日本語と外国語、昔の日本語と今の日本語の違い。
②計量言語学：頻度とは何か、文書頻度という考え方、重み付けとは何か。
③数学と言語：言語を数理的に捉える、数学の成果を言語学に利用する。
④研究方法論：仮説をもとに方法を考え、考察をまとめ、結論を導き出す。
⑤可視化技法：グラフ理論とグラフ記述言語を学び、モデルを作って目に見える状態を作り出す。

■事務局との協力体制

研究推進部研究企画課と事前に打ち合わせを行い、プログラム実施にあたって必要となる準備を確認してくださったほか、当日は事務担当者として研究企画課の２名が参加し、配布物の袋詰作業等の事前準備および受付・写真撮影等を担当していただきました。

■広報活動

東京工業大学のウエブサイトにて「東工大の夏休みイベント 2017」カレンダーに実施プログラムの情報を掲載したほか、リベラルアーツ研究教育院の Web サイトにも実施の告知を行いました。

http://www.titech.ac.jp/outreach/community/summer2017.html

http://educ.titech.ac.jp/ila/event_information/2017/054042.html

■安全配慮

保険に加入し、それを参加者に周知しました。昼食は夏場であることを考慮し、温度による賞味変化の少ないもの、中学生の分量として適切なものを選び、食物アレルギーが起こらぬよう、成分表示を行いました。水分補給には注意を促し、自分で飲み物を持参するようにお願いしました。

■今後の発展性、課題

ネットワークモデルを作る実習の他に、簡単なプログラムを書いて動かしてみる、小さいレポートを書いてみるなど、参加者同士のディスカッションを今回よりも多く活発にできればと考えています。

１回あたりの受講人数は限度がありますが、毎年、抽選に漏れる方が多いので、チャンス拡大のために回数を夏冬２回にすることを考えています。また、受講者の幅を広げるために、中学生だけでなく，小学生の部も検討に入れています。

【実施分担者】　該当なし

【実施協力者】　６　名

【事務担当者】　田中　愛彩美･齋藤　順子　研究推進部研究企画課･事務職員

第 5 章

人文情報学月報

5.1　巻頭言

巻頭言なるものをはじめて依頼され、執筆した。

5.2　資料

DHM 057 【前編】
土, 05/21/2016 - 19:05 — dhmadmin
2011-08-27 創刊 ISSN 2189-1621
人文情報学月報
Digital Humanities Monthly
2016-04-29 発行 No.057 第 57 号【前編】 628 部発行

◇ 目次 ◇

【人文情報学／ Digital Humanities に関する様々な話題をお届けします。】

◇《巻頭言》「言語学とコンピュータ」

（山元啓史：東京工業大学）

特別なことがない限り、論文以外の文章は書かないことにしています。もちろん、巻頭言を書くのははじめてです。論文ではないことをいいことに、今までボツになった本の内容について書くことにしましょう。

今まで本を出版しようとして、ボツになった企画が 2 つあります。1 つはコーパス言語学の入門書シリーズの 1 冊で、これは依頼された原稿でしたが、ボツになりました。もう 1 つは東工大の学生のために書いた言語学の教科書でした。

コーパス言語学の本は概論的なものを依頼されました。それを私はコーパス言語学概論と勝手に勘違いして、書き進めていきました。編集の方からは読者は Windows を使っていることを前提に、との注文はありましたが、私自身 Windows を使わないこともあって、Linux のコマンドやパイプ、簡単なスクリプトを中心に説明したテキスト処理の原稿を書きました。Windows であっても、cygwin を使えば大差ないと思ったからです。しかし、Windows 前提でないと本は売れないとのことで NG でした。「ディレクトリとは」「ファイルとは」「OS とは」などのコンピュータの基本用語を説明するように、と書き直しを告げられました。それらを説明した本はたくさんあるので、私自らがボツにしました。GUI のメニュー表示や用語が変わることはあっても、UNIX 由来のコマンドはずっと変わらないし、何をしているのかが、わかりやすいので、その方が息の長い記事になると思ったのですが、編集者さんはそうは思わなかったようです。

とにかく、テキスト処理は、手を動かさないことには、何も始まらないので、その本には次のような例題と練習を載せました。

1. 例題：文の長さのデータの平均値を求めよ。
2. 例題：任意の用語の文脈がわかるようにリストを作れ。
3. 例題：前後の文脈がわかるように文字順に並べ替えよ。
4. 例題：形態素解析器をインストールして、使ってみよ。
5. 例題：形態素解析器を使って名詞だけを選び出せ。
6. 例題：単語の頻度を計算せよ。

コマンドの基本的な原理を説明した上で、どのコマンドを使い、どのプログラム

を組み合わせれば、自分の意図する出力が得られるか、考えてもらう演習です。これの行き着く先は、いわゆるシェル芸というものです。シェル、キーボード・ショートカット、コマンドの組み合わせで、縦横無尽にテキストを料理するってやつです。誰もが同じことを考えるもので、この本を書いた後に、「言語処理 100 本ノック」(http://www.cl.ecei.tohoku.ac.jp/nlp100/) というものがあるのを知りました。私の方向性は間違ってはいないことはわかりましたが、同時にいまさら私が書く必要もないなぁとも思いました。

さて、もう 1 つのボツになった本は、「みんなで考える言語学」と題する教科書です。どうせ出版されないのだから「言語学の素」という調味料に似た題名をつけたこともありました。この本は東工大の大学院生に向けた授業が元になっています。ある出版社の担当者さんが「下書きでも良いので内容を見せてほしい」というので、お見せしたところ「オーソドックスな言語学でない」との返答でやんわり断られました。日本語教育能力試験などの検定試験対策になりそうなものを期待したのかもしれません。

- 1 章「ウィトゲンシュタインと言語ゲーム」
- 2 章「チューリングとチューリングマシン」
- 3 章「ジップとジップの法則」
- 4 章「ダニエル・ジョーンズの 18 の基本母音」
- 5 章「ソシュールと記号論」
- 6 章「フィルモアと格文法」
- 7 章「チョムスキーと生成文法」

確かに「オーソドックス」ではありません。ウィトゲンシュタインからはじまる言語学の教科書なんてありません。ウィトゲンシュタインは哲学者。チューリングは数学者。ジップでやっとトークンを取り扱うので言語学かな？とも。ダニエル・ジョーンズ（マイフェアレディのヒギンズ博士のモデル）が出てきたあたりから、言語学のようには見えます。音韻論を教えるのにダニエル・ジョーンズを出す教科書はほとんどないでしょう。たとえば、かの有名な George Yule の The Study of Language の索引でも、“ Jones, Danel ” の索引項目は見られません。おおむねアメリカの大学の教科書は版を重ねて、演習問題をどんどん新しくしていきます。演習問題はさまざまな観点から入れ替えられます。もっと勉強したい人のための Further Readings のリスト差し替えも頻繁です。どんどん版を重ねるので、古い版は面白いくらい安く入手できます。この本の第 3 版は 380 円（新品）でアマゾンから購入できます。

なぜオーソドックスでない構成になったのか？これにはいろいろな理由がありますが、一番の理由は、対象が東工大の大学院生だったということです。数理、計算、物理、化学などの専門家ではあっても、言語学は決して彼らの専門ではありませんし、彼らも言語学を自分の専門として勉強しようとは思っていません。こういう学生に「そもそも言語学とは」などと紋切り型で授業をはじめても眠くなるばかりです。言語学の知識はなくても、授業初日から、ディスカッションがしたくなるような授業を考えました。自分が話すことばと比べながら、「言語学の歩み」を教師が語るのではなく、ディスカッションによって学生さん自身に考えてもらう授業にしました。

どの章にも簡単な紹介・導入を記載しましたが、それ以外は「演習問題」です。これを 3、4 名のグループで「ああだ」「こうだ」とディスカッションしては、それをグループごとに発表していきます。

たとえば、1 章の演習問題（言語ゲーム）は、

> 「私が通りかかったとき、すでにゲームは進行中だった」の「私が通りかかったとき」を「私が生まれたとき」に、「ゲーム」を「言語」に言い換えたら、言語とはどんなものと言えるだろうか？
> チェスや将棋、ポーカーのルールを知らなくても、見ているうちにそのルールがわかり、なんとなくゲームに参加できるのは、なぜだろう？

などです。人間が生まれたとき、すでに言語は存在し、いつのまにか、人間はそのルールを身につけ、それに参加し、それを発展させ、死んでいく。そして、つぎの世代の人間がその言語を使い、少しずつじわじわ形を変えていく。確かに言語は人間の口から出たものですが、人間が作ろうと思って作ったものではありません。何らかの力学によって、自然な仕組みで言語ができてきます。それは常に一定なものではなく、むしろ動的なものです。混沌としているようですが、その形には法則性があります。どういう例がわかりやすいでしょうか、あまりいい例ではありませんが、たとえば、人間の肘の関節は、内側には曲がるが、外側には曲がりませんよ！っていうような「なぁんだ」というような法則性です。その「なぁんだ」というものが本当は何であるのかがよくわからないので、それを見つける研究をしているのですね。

2 章の「チューリングマシン」では、

> 日常に見られるテープとヘッドに似たものを見つけて、それのどの部分がテープ、ヘッドに当たるかを述べよ。

というものです。ここでは、得体のしれない言語というものを、記述するには具体的に何をすればいいのか、そもそも記述するとはどんなことかなどを話し合います。言語もリニアなものであり、その抽象的な姿を整理するには計算機モデルが役立ちそうだというお話しです。

3 章ではジップの第二法則を紹介し、

> 人名の出現頻度、新聞記事に見られる単語の頻度がそれに従うのはなぜでしょう。また言語だけでなく、他の自然界にも見られるのはなぜでしょう。

と問いかけます。たとえば、人口の多い都市の数は少なく、人口の少ない町や村はめちゃくちゃ多い。ガラスの割れた大きい破片の数は少ないが、だんだん小さくなっていって、粉々になった破片の数はもう数えられないほどたくさんである。ジップ則を通して、単語の分布と自然の摂理にはどういう関係があるのかを議論してもらいます。実際に今も、なぜそれらがジップ則にしたがうのかはよくわかっていないものですから、この議論はそう簡単には終わりません。おそらく、とことんその理由を説明しなければ気がすまない理系の学生にはうってつけのトピックであったのでしょう。

東工大は伝統的に自然言語処理の研究者を多く輩出していることで有名です。その意味では東工大には、言語を扱う素地はあったと言えましょう。2016 年 4 月、東京工業大学は日本で初めての学部と大学院を一緒にした学院を設置しました。そして、この 4 月より東工大では、正式に学士課程の科目名として「言語学」を設け、理系の学生のための言語学の授業がはじまります。理学・工学を学ぶ新入生の目には、東

工大の言語学はどううつるのでしょうか。まだ始まったばかりです。非常に楽しみです。

執筆者プロフィール
- -

山元啓史（やまもと・ひろふみ）専門は言語学、言語変化、外国語としての日本語教育。オーストラリア国立学大学院博士課程修了。Ph. D in Linguistics。1993 年筑波大学文芸・言語学系留学生センター助手、1995 年同講師、1997 年カリフォルニア大学サンディエゴ校客員研究員、2006 年オーストラリア国立大学客員研究員、2009 年東京工業大学留学生センター准教授、2017 年東京工業大学リベラルアーツ研究教育院教授。著書は、“ Japanese A Comprehensive Grammar ” Routledge, 「コーパスと日本語史研究」ひつじ書房、などがある。

◆編集後記（編集室：ふじたまさえ）
- -

第 57 号前編、後編ともにいつも以上に読み応えのある内容となりました。巻頭言をはじめ、ご寄稿いただいた皆さま、ありがとうございます！

どの内容も素晴らしかったのですが、特に個人的な興味としては、特別寄稿をいただいた OMNIA のことが気になっています。また、巻頭言として掲載している山元先生の文章も、大変興味深い内容でした。

後編のイベントレポートの中では、国立国会図書館関西館の菊池さんがおっしゃっていた「DH の現状や課題などを体系的にまとめた日本語の解説書」が気になります。個人的な感想ですが、本メルマガが扱っている話題も含め、DH について体系的にまとめなおすにはどういった媒体が良いのか考えてみると、印刷物よりは Wikipedia のようなデジタルのもののほうが合っているようにも思いました。

◆人文情報学月報編集室では、国内外を問わず各分野からの情報提供をお待ちしています。
情報提供は人文情報学編集グループまで...
DigitalHumanitiesMonthly[&]googlegroups.com
[&] を@に置き換えてください。

人文情報学月報 [DHM057]【後編】 2016 年 04 月 29 日（月刊）
【発行者】"人文情報学月報"編集室
【編集者】人文情報学研究所＆ ACADEMIC RESOURCE GUIDE（ARG）
【 ISSN 】2189-1621
【E-mail】DigitalHumanitiesMonthly[&]googlegroups.com
[&] を@に置き換えてください。
【サイト】 http://www.dhii.jp/

第 6 章

大学研究室探検隊

大学研究室探検隊 Vol. 6: 東京工業大学　山元啓史研究室

6.1　取材

2018 年 2 月号「サクセス 15」pp. 16–19. グローバル教育出版より、インタビュー記事が発行された。日本学術振興会に取材があったことを報告。

中学生のみなさんにはあまりなじみがないかもしれませんが、多くの人が進むであろう大学の研究室では、文系・理系を問わず、日々さまざまな研究が行われています。このコーナーでは、そうした研究室や研究内容を紹介していきます。ここで見つけた研究がみなさんの視野を広げ、将来の目標への道標となるかもしれません。第 6 回は、言語の可視化に関する研究を行う東京工業大学の山元教授の研究室を紹介します。

6.2　資料

（4 ページ先から戻る順でご覧ください）

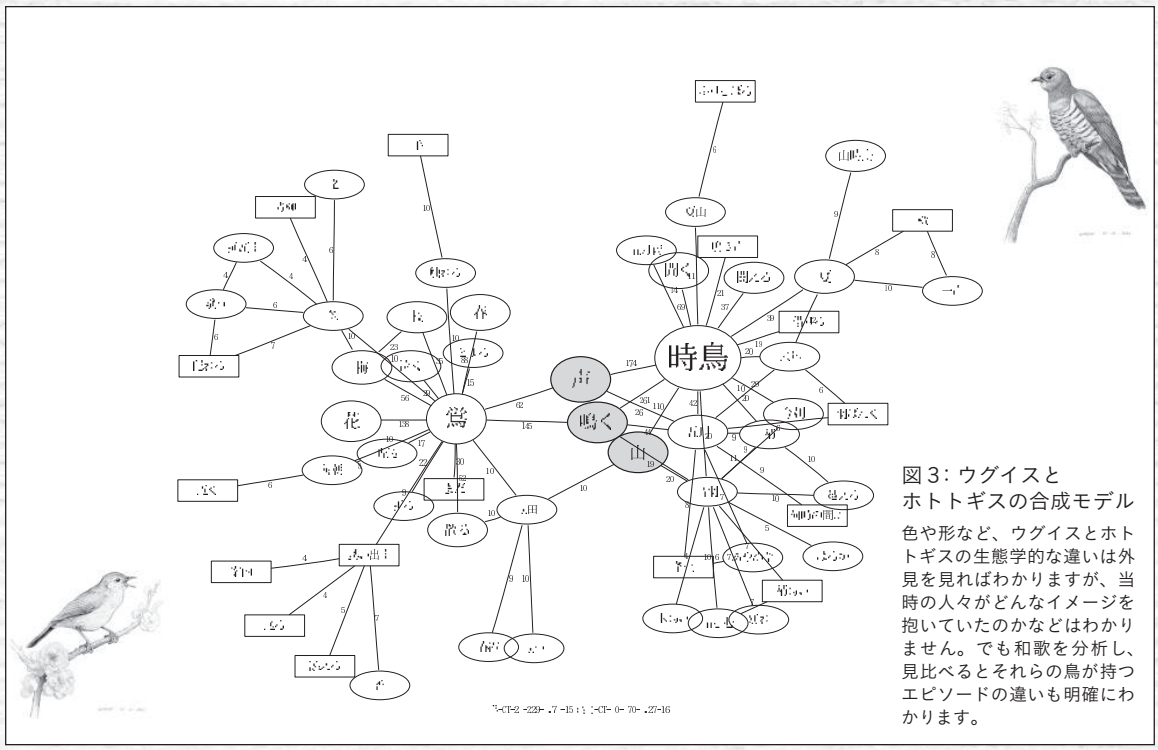


図3：ウグイスと
ホトトギスの合成モデル

色や形など、ウグイスとホトトギスの生態学的な違いは外見を見ればわかりますが、当時の人々がどんなイメージを抱いていたのかなどはわかりません。でも和歌を分析し、見比べるとそれらの鳥が持つエピソードの違いも明確にわかります。

研究の原点にあるのは日本語を教える難しさ

さて、意外なことに山元先生は、初めから言語学の研究者を志していたわけではありません。大学卒業後、日本語学校で留学生に日本語を教える仕事についたことがきっかけで言語学を勉強し始めたそうです。

「日本語は母語ですから、教えるのに苦労するなんて思いもしませんでした。じつは話せることと教えられることはまったく別で、流暢に話せるからといってうまく教えられるわけではありませんでした。そこで、日本語を教える技術を身につけようと、国立国語研究所の長期研修生となって第二言語としての日本語教育を学びました」（山元先生）

その後、筑波大学で再び留学生に日本語を教えることになり、コンピュータやインターネットを使った教育を実践するために教育機器の開発やことばと学習の関係について学ぶようになりました。留学生の学ぶ心を理解するには自分も留学しなければと思い、カリフォルニア大学（アメリカ）・オーストラリア国立大学へ行って学びました。アメリカで最新の記憶とモデルの研究をしていたとき、新しいことよりもむしろ古いことから新しいことを見つける方が将来・未来を分析するのに役立ちそうだということに気がつきました。そこで始めたのが古代語の可視化に関する研究でした。もし言語に形があるなら、留学生にその形を見せることで日本語が学びやすくなるのではないかと考えたそうです。

今回紹介した研究以外にも、言語に関するさまざまな研究を行う山元先生は、現在も東京工業大学で留学生に日本語を教えています。留学生にかかわるこの仕事をとても楽しいと語る山元先生からみなさんへのメッセージをいただきました。

「研究するときは、まず結果のありままの姿をとらえ、客観的に事実を把握することが大切です。重要なのは『自分でなにかを見つける』ことです。みなさんもだれかの情報だけで判断するのではなく、自分の目で見たこと、感じたことを大切にしてほしいと思います」

山元啓史研究室

メンバー
東京工業大学環境・社会理工学院社会・人間科学系大学院生3名、研究生1名

研究室
東京工業大学大岡山キャンパス　東京都目黒区大岡山2-12-1　西1号館
言語学・応用言語学研究室

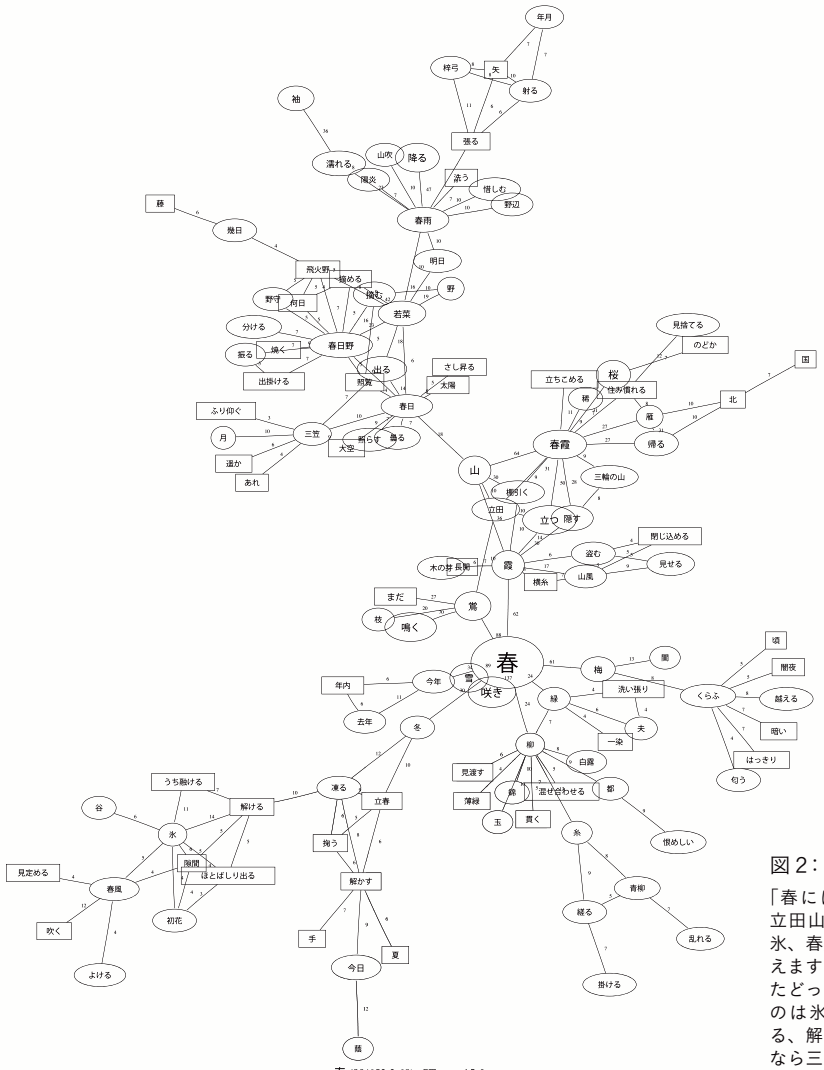


図２：春のネットワーク

「春には春日野、春霞、立つ、立田山、柳、青柳、立春、冬、氷、春風など色々なことばが見えますね。つながっている線をたどっていくと春風に関係するのは氷、氷は凍る、氷は解ける、解けるは立春…なんかさよなら三角の歌、ことば遊びみたいですね」（山元先生）

和歌と時鳥（ホトトギス）が出てくる和歌の分析結果です。２つに共通するのは色がついた「声」「鳴く」「山」の３つだけです。和歌の世界では２羽の鳥がかなり違うとらえ方で表現されているのがわかります。ウグイスには「梅」「春」など、ホトトギスには「夏」「五月雨」など、それぞれ異なる語が見られます。この図から２つの鳥を当時の人がどうとらえていたかがわかります。

　この研究に和歌を使う理由を伺うと「『枕草子』や『源氏物語』など当時の文章は色々残っていますが、これらの作品において文とはなにかを決めたのは現代の学者ですから、文の終わりは正確でない可能性があります。文の長さにもばらつきがあります。文の長さによって結果に影響が出ます。その点、和歌は一首でお話は完結していますし字数も31文字に決まっているので好都合です。

　また、平安時代に作られた二十一の勅撰和歌集は春、夏、秋、冬、恋など歌のジャンルが二十一集とも共通しています。そのため春なら春の歌を一番目の古今和歌集から二十一番目の新続古今和歌集までずっと比較することができるんです」と話されます。

山元啓史（やまもと・ひろふみ）／民間の日本語学校で教鞭をとった後、筑波大学文芸言語系助手、講師。カリフォルニア大学サンディエゴ校、オーストラリア国立大学で客員研究員を経て東京工業大学リベラルアーツ研究教育院、環境・社会理工学院　社会・人間科学系教授。夏休み中学生向けセミナーは毎年すぐに満員。2016年情報処理学会山下記念賞受賞。

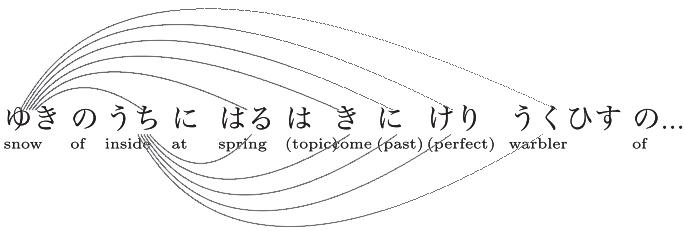


図1：共出現パターンの作り方

「雪のうちに　春はきにけり　鶯の　こほれる涙　いまやとくらん」という和歌のなかに出てくる単語でペアを作ります。このように同じ文に現れたペアを「共出現パターン」と呼びます。

筆を入れる人はほとんどいませんが、ことばの形はそのまま、意味だけが変化して現在も使われることばになっています。

「昔の文章に出てくる単語はいまと異なる意味を持っている可能性がありますが、昔の文章を読むときはどうしても現代の常識にあてはめて考えてしまいがちです。例えば『食べる』という単語の前にある単語は食べものをさすと思うでしょう。でも本当はまったく関係がないかもしれません。また、現存していないものの名前も私たちにはわかりません。人間はあくまで推測しか導き出せないのです」（山元先生）

そこで役立つのがコンピュータです。「はな（花）」という単語が花びらだけをさすのか、つぼみや茎も含めた全体をさすのかわからなくても、コンピュータで多くの和歌のなかから「はな」という単語がどんな単語といっしょに使われているかを探し出し、その結果を図に書き出せば、そこから花の意味はもちろん、香りや感触が確認できるというのです。コンピュータによる分析は客観的で信ぴょう性のある結果としてとらえられます。

約20名の小中学生が参加したワークショップの様子。和気あいあいとした雰囲気のなかで言語学のおもしろさに触れられるプログラムです。

「私たちは昔のことばを直接聴いたり、昔の食べものを直接味わったりすることはできません。でも、昔のことばを分析すれば、タイムマシンに乗らずに昔の世界を感じることができるのです」（山元先生）

ここでの「昔」とは、平安時代をさします。現存する平安時代の書物から、日本語は千年以上も前から使われていた言語だということがわかっています。ここまで古い歴史を持つ言語は世界でも日本語とアイスランド語だけだそう。日本語はそれだけ過去をさかのぼって調査できる貴重な言語なのです。

和歌を科学的に分析して ことばの形をとらえる

では実際にどう図を作っていくかというと、まず和歌から任意で2つずつ単語を取り出し、ペアを作っていきます。図1の「雪のうちに　春はきにけり　鶯の　こほれる涙　いまやとくらん」という和歌からは、「雪、うち」「雪、春」「雪、き」「雪、けり」「雪、うぐいす」…という形でどんどんペアができあがります。

次に言葉の重みを調べていきます。

「どこにも出てくる単語は『探す価値のない単語』です。例えば千首の和歌すべてに出てくる単語があったら、このような単語は検索しても、なにかを特定するのに役立ちません。つまり情報量0です。一方、たまにしか出てこない単語は当時の人がなにかを伝えるために使った『探す価値のある単語』です。こういう単語の情報量は多くなります。

このように単語の重みを計算していきます。そして2単語の重みの平均値が大きいペアから図に出力していきます。これらの作業はすべてコンピュータで行います。

図2は古今和歌集から『春』に関することばのペアを集めたものです。図にすることで、目に見えないはずの『春』の形が見えるようになってきます」（山元先生）

図3は鶯（ウグイス）が出てくる

中学生のみなさんにはあまりなじみがないかもしれませんが、多くの人が進むであろう大学の研究室では、文系・理系を問わず、日々さまざまな研究が行われています。このコーナーでは、そうした研究室や研究内容を紹介していきます。ここで見つけた研究がみなさんの視野を広げ、将来の目標への道標となるかもしれません。第６回は、言語の可視化に関する研究を行う東京工業大学の山元教授の研究室を紹介します。

（画像・資料提供：東京工業大学 山元啓史研究室）

今回取り上げるのは言語学に関する研究です。といってもみなさんがイメージする文系の学問としての研究とはひと味違います。お話を伺った山元啓史教授が在籍するのは理系トップレベルの大学として名高い東京工業大学。研究も「言語を可視化する」、つまり言語をコンピュータで分析して図として表現することで、言語の形を考えるというなんとも画期的なものなのです。

「言語を可視化する」と聞くとなんだか難しそうですが、山元先生が「ひらめき☆ときめきサイエンス（※）」の一環として中学生向けに開いているワークショップ「目で見てわかる昔の日本語と今の日本語～タイムマシンに乗らずに行ける昔の世界」で扱う内容はみなさんにもわかりやすいものとなっています。まずはその内容を見ていきましょう。

単語の意味は時代によって変わる

日本語に限らず、言語は一般的に一度形や音、つづりが定着するとそれらは変わりにくいのですが、その意味は時代によって変化していくといわれています。「下駄箱」や「筆箱」は、一般的にはもうそれらに下駄や

（※）日本学術振興会が科学研究費助成事業（科研費）の一環として主催するプログラム。大学や研究機関で科研費により行われている最先端の研究を、小５・小６、中学生、高校生が体験できる。全国の大学・研究機関で行われている。

第 7 章

JADH: 論文・ポスター・スライド

The field of humanities is undergoing a radical transformation in its encounter with rapid developments in the digital domain. In response to this situation, various efforts have been undertaken based on collaboration between the humanities and the information technologies in Japan and foreign countries. Recently, various related activities have been carried out under the rubric of Digital Humanities in Europe and North America. Progress in this area in Japan however, has been hindered in a couple of ways. For example, there have been limits to the extent of the collaboration between Japanese digital humanities specialists and their counterparts in the West brought about by the basic difficulties with the digitization of the characters and texts that compose Japanese resources. In general, the results of digitization efforts in Japan in the humanities disciplines have not been commensurate with the huge effort and expense made heretofore. To begin to resolve such issues, we intend to establish the Japanese Association for Digital Humanities (JADH), which aims to form an environment where international collaborative works are more fully realized. (13 Sep 2011) Japanese Association for Digital Humanities Japanese Association for Digital Humanities

7.1 OSDH2011

OSDH2011: Osaka Symposium on Digital Humanities 2011

1. Hilofumi Yamamoto, TokyoInstitute of Technology
   Graph Representation of the Connotations of Classical Japanese Poetic Vocabulary

7.2 JADH2012

"Inheriting Humanities" Program PDF

1. Hilofumi Yamamoto (Tokyo Institute of Technology / University of California, San Diego), Makiro Tanaka (National Institute of Japanese Language and Linguistics) and Yasu-Hiro Kondo (Aoyama Gakuin University / National Institute of Japanese Language and Linguistics), Design of Serial Comparison Model for the Diachronic Corpus Study of Japanese
2. Makiro Tanaka (National Institute for Japanese Language and Linguistics) and Hilofumi Yamamoto (Tokyo Institute of Technology), Emotive Adjectives and Verbs of the Heian Japanese

7.3 JADH2013

3rd Symposium JADH2013: "Bridging GLAM and Humanities through Digital Humanities"

The Japanese Association for Digital Humanities is pleased to announce its third annual conference, to be held at Ritsumeikan University, Kyoto, Japan, September 19-21, 2013.

The conference will feature posters, papers and panels. We invite proposals on all aspects of digital humanities globally, and especially encourage papers treating topics that deal with practices that aim to go beyond borders, for example, between academic fields, media, languages, cultures, and so on, as related to the field of digital humanities.

1. Lexical Modeling of Yamabuki (Japanese Kerria) in Classical Japanese Poetry
   Hilofumi Yamamoto (Tokyo Institute of Technology / University of California, San Diego)
   This project is a lexical study of classical Japanese poetic vocabulary through network analysis based on graph theory. The analysis is based on co-occurrence patterns, defined as any two words appearing in a poem.
2. A Diachronic and Synchronic Investigation into the Properties of Mid-Rank Words in Modern Japanese
   Bor Hodoek (Tokyo Institute of Technology)
   Hilofumi Yamamoto (Tokyo Institute of Technology / University of California,

San Diego)

The present study focuses on the role of mid-rank words in modern Japanese. Mid-rank words are defined as words having an average TF-IDF (term frequencyinverse document frequency) score. Mid-rank words are often overlooked for words with high TF-IDF scores, which act as reliable topic markers. Words with low TF-IDF scores are in turn seen as functional words and often discarded from analysis. Mid-rank words are thus words that do not lean heavily towards the two extremes of topic and function, but include a mixture of both. As such, their exact grammatical function is elusive and still relatively unknown.

7.4 JADH2014

4th Symposium JADH2014: "Bridging GLAM and Humanities through Digital Humanities"

The Japanese Association for Digital Humanities is pleased to announce its fourth annual conference, to be held at University of Tsukuba, Japan, September 19-21, 2014.

The conference will feature posters, papers and panels. We invite proposals on all aspects of digital humanities globally, and especially encourage papers treating topics that deal with practices that aim to go beyond borders, for example, between academic fields, media, languages, cultures, and so on, as related to the field of digital humanities.

GLAM (Galleries, Libraries, Archives, and Museums) have played prominent roles in the recent rapid evolution of the humanities in the digital environment. In this decade, digital methods have been illuminating new possibilities of their relationships. At JADH2014 we will be especially interested in hearing presentations that focus on these methodologies and practice in GLAM, but we nonetheless welcome papers on a broad range of DH topics.

1. A Visualization and Analysis System for Japanese Language Change: Quantifying Lexical Change and Variation using the Serial Comparison Model
2. Development of an Asymptotic Word Correspondence System between Classical Japanese Poems and their Modern Translations

7.5 JADH2015

5th Symposium JADH2015: "Encoding Cultural Resources"

The Japanese Association for Digital Humanities is holding its fifth annual conference at the Insitute for Research in Humanities, Kyoto University, Japan, September 1-3, 2015.

7.6 JADH2016

6th Symposium JADH2016: "Digital Scholarship in History and the Humanities"

The Japanese Association for Digital Humanities is holding its sixth annual conference at The University of Tokyo, Japan, September 12-14, 2016.

The main venue of the conference: Fukutake Learning Theater.

7.7 JADH2017

JADH2017: "Creating Data through Collaboration"

The Japanese Association for Digital Humanities is holding its seventh annual conference at Doshisha University, Kyoto, Japan, September 11-12, 2017.

The main venue of the conference: Ryoshinkan (Imadegawa)

7.8 JADH2018

JADH2018: "Leveraging Open Data"

The Japanese Association for Digital Humanities is pleased to announce its eighth annual conference, to be held at Hitotsubashi-Hall, Tokyo, Japan, September 9-11, 2018 hosted by the Center for Open Data in the Humanities jointly with the TEI conference 2018.

7.9 UCLA workshop in 2016 and 2017

UCLA Department of Asian Culture and Language invited me for the UCLA workshop of Computer and Humanities 2016 and 2017. I present two lectures for faculty members, Graduate students, and one class for undergraduate students.

7.10 Posters and flyers 2011–2018

Hilo Yamamoto
Ph. D. Linguistics

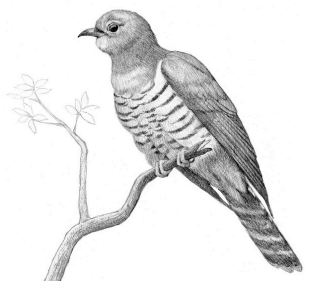
Graph Representation of the Connotations of Classical Japanese Poetic Vocabulary

Computer modeling using co-occurrence patterns

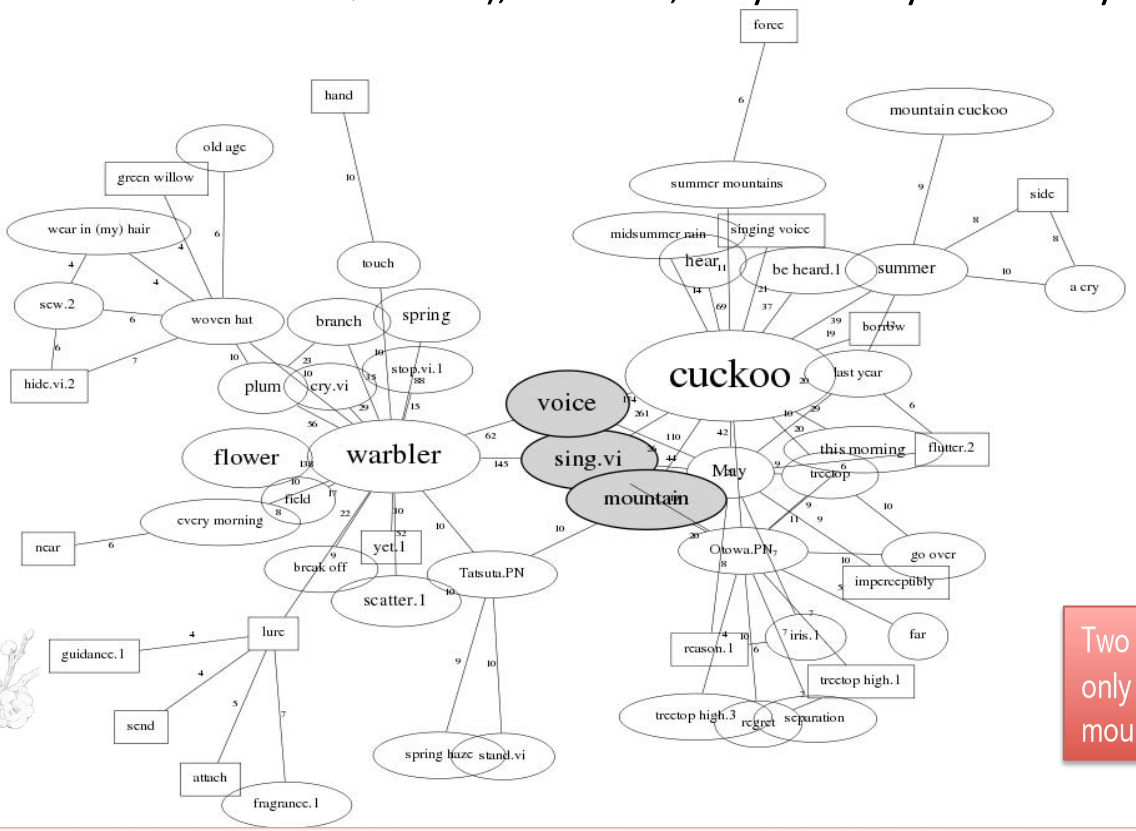
Two birds, warbler and cuckoo, come to Japan in spring.
In the *Kokinshu* (ca. 905), however, they are very differently used.

Tokyo
Institute of
Technology

Warbler



Cuckoo



Two birds' network shares only three words: i.e., sing, mountain, and voice.

Poem texts must contain more information we have not known yet!

Graph Representation of the Connotations of Classical Japanese Poetic Vocabulary

Hilofumi Yamamoto, Tokyo Institute of Technology

Can we define a connotation?

Octopus

weird
spooky ... VS *takoyaki* (fried octopus!) = delicious

'Connotation' depends on the receiver of a message.

Material

The *Kokinshū* (ca. 905)

the first anthology compiled by the order of the Emperors.

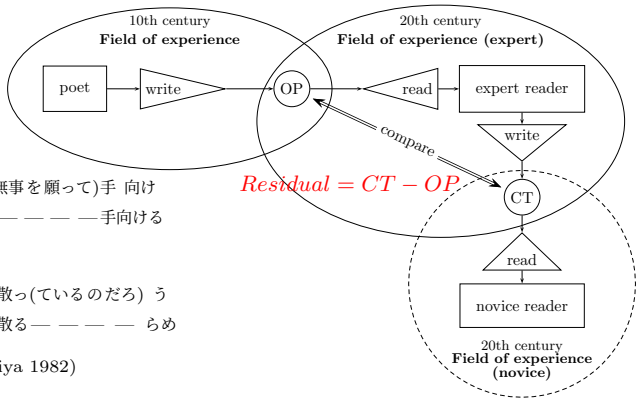
Targets: Tatsuta and Yoshino

Utamakura indicates:
1. a name of place
2. emotive notions.

→ HOW DIFFERENT!

Relationship between OP and CT

Based on Schramn's theory of communication [source]→ [encoder]→ (signal)→ [decoder]→ [destination]



Alignment of OP and CT

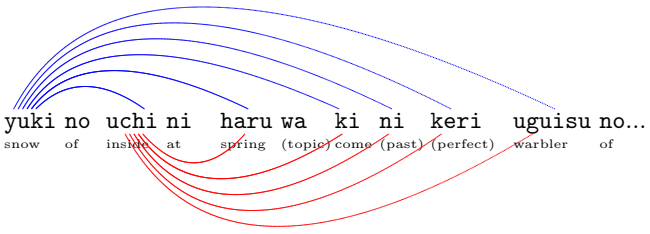
CT (秋の 末近くなって帰り道についた)龍田姫(が道中の無事を願って)手 向け
OP ———— — — — — — — —立田姫— — — — — — —手向ける

CT (をする)神があるからこそ秋の木の葉(が)幣(となって)散っ(ているのだろ) う
OP — — 神のあれ ば こそ秋の木の葉[の]幣 と — —散る— — — — らめ

CT298 translated by Teruhiko Komachiya (Komachiya 1982)

Cooccurrence Patterns

A pattern consists of any two words appearing in a text.



Patterns may number up to 5,000.

Co-occurrence weight

Rocchio (1971)

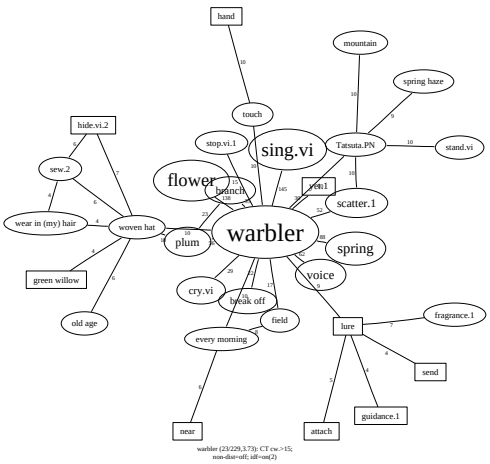
$$w(t,d) = (1 + \log\ tf(t,d)) \cdot idf(t)$$

Transform this for co-occurrence patterns.

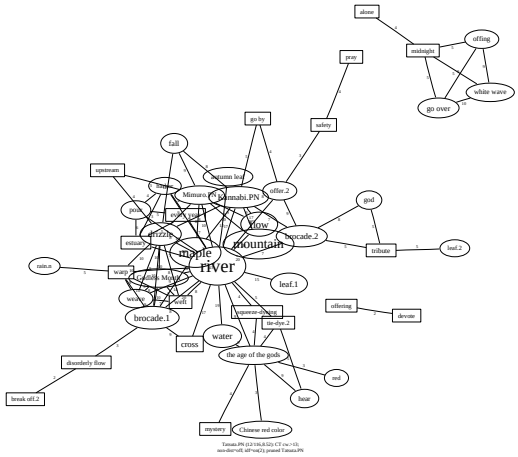
$$cw(t_1,t_2,d) = (1 + \log\ ctf(t_1,t_2,d)) \cdot \sqrt{idf(t_1) \cdot idf(t_2)}$$

document — frequency of co-occurrence — geometric mean of two weights

A Sample of Model —Warbler

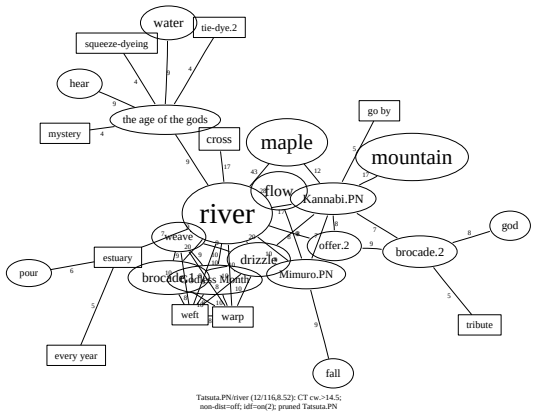


Model of Tatsuta after pruning



KKS 994
as the howling winds
keep white waves rising in the
offing in the deep
of night is my lord crossing
Tatsuta Mountain alone

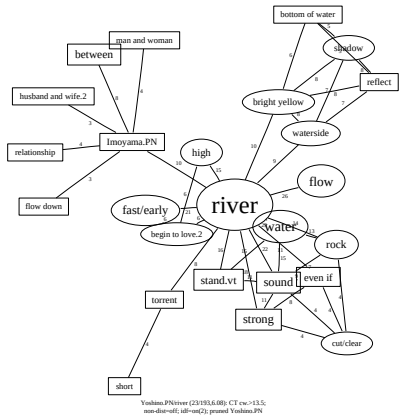
Model of Tatsuta (river)



KKS 294
unheard of even
in the stories of
the age of the awesome gods —
the waters of Tatsuta
stream dyed a Chinese red

KKS 300
because autumn is
returning crossing over
Kannabi Mountain
she offers crimson prayer
strips at Tatsuta River

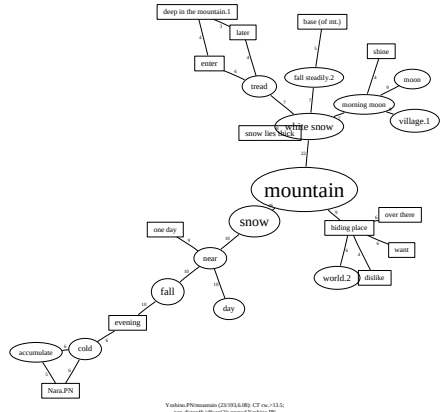
Model of Yoshino (river)



KKS 828
like the Yoshino
River cascading between
Husband and Wife peaks
obstacles forever part
lovers in this world of ours

KKS 471
so suddenly was
my heart dyed with passion for
you it leapt high as
the waves in the waters of
the swift Yoshino River

Model of Yoshino (mountain)



KKS 950
if only I had
a home on the far side of
fair Mount Yoshino
I would make it my refuge
in times of worldly sorrow

a place for hermits

Conclusion

- Visualisation by co-occurrence patterns
- Depending on the use of words, the clusters are formed separately e.g. white waves.
- Mathematical operations using models e.g. pruning, subtracting
- Displaying connotations using CT and OP e.g. gods→Tatsuta; secular→Yoshino
- Relative saliences
- Comparison of a word between two different periods e.g. cherry blossoms at Yoshino
- **Email**: `yamagen@ryu.titech.ac.jp`

Design of Serial Comparison Model for the Diachronic Corpus Study of Japanese

Hilofumi Yamamoto
Tokyo Institute of Technology

Makiro Tanaka
National Institute for Japanese Language and Linguistics, Japan

Yasuhiro Kondo
Aoyama Gakuin University

Development of Diachronic Corpus

Project by the National Institute for Japanese Language and Linguistics, Japan, NINJAL: 2009–13, 4 year project.

Main purpose: Study of Japanese language
(sub) purpose: Study of Japanese (classic) literature

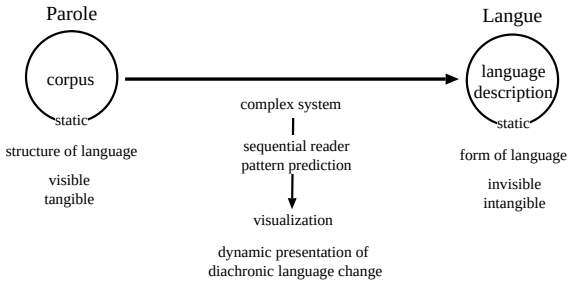


Figure 1: Corpus and Description, Langue and Parole:
The nature of language is dynamic and always changing while the phenomena of language might be static.
We should consider the dynamic change of language as a component comprised of various elements.
The feature of language we usually observe is a complex system and tangled with wide-ranging elements.

Contents of Diachronic Corpus

1. The Tale of the Bamboo-Cutter
   (ca. 890; Taketori monogatari; 12,583 tokens)
2. Tales of Ise
   (ca. 901; Ise monogatari; 15,900 tokens)
3. Tales of Yamato
   (ca. 950; Yamato monogatari; 26,733 tokens)
4. The Tosa Diary
   (ca. 935; Tosa nikki; 8,113 tokens)
5. The Pillow Book
   (ca. 996; Makura no sōshi; 79,861 tokens)
6. Tale of Genji
   (ca. 1100; Genji monogatari; 510,711 tokens)

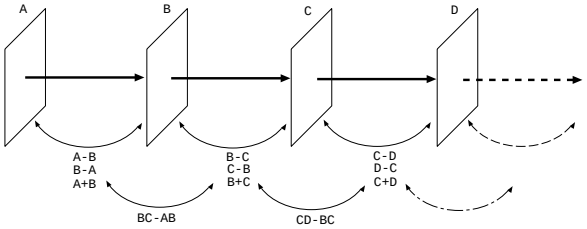
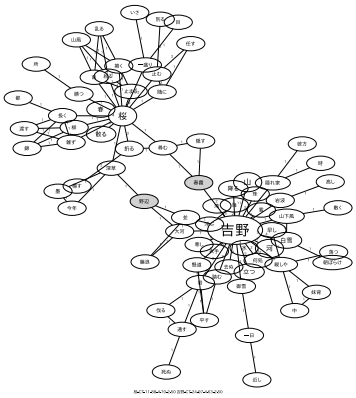


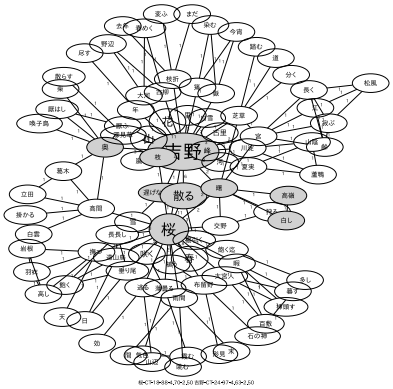
Figure 2: Extraction of delta from each synchronic layer: A, B, C and D are arbitrarily-assigned synchronic layers on the time axis. Examination of linguistic transitions is achieved through the comparison of lexical items in each layer with those in other layers, and the discovery of common principles appearing in the delta of data extracted from both systems as well.

A case study: use of SAKURA(cherry blossoms) in Mt. Yoshino → Kokinshū (ca. 905) vs Shinkokinshū (1205)



Sakura (桜) and
Yoshino (吉野), a place name in Nara prefecture

← Kokinshū (ca. 905)
Shinkokinshū (1205) →
during 300 years differences.



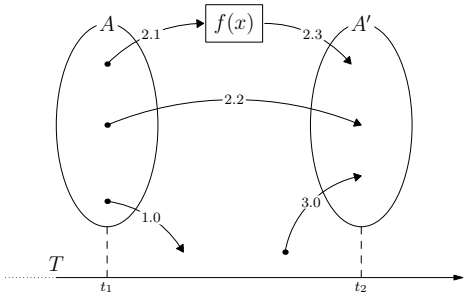


Figure 3: Serial comparison model; differential model of transitional linguistic elements of target texts;
A is a set of elements that occurred at Time t_1;
A' is a set of elements that occurred at Time t_2;
T is the time axis; $f(x)$ is a function for converting an element x of A into that of A'.

Future Task

- To define linguistic units suitable for each era
- To develop a dictionary for machine analysis

→ it allows us syntagmatic and paradigmatic anal

Conclusion

- Addressed basic concepts and framework of diachronic corpus
- Illustrated the serial comparison model for historical analysis

→ Lexical differences between any two groups of t

Lexical Modeling of *Yamabuki*, Japanese Kerria in Classical Japanese Poetry

Hilofumi Yamamoto / Tokyo Institute of Technology
yamagen@ryu.titech.ac.jp

Figure 1: The picture of "*Yamabuki To Kahazu*" (kerria and frog) by Hiroshige Utagawa (http://www.gekkanbijutsu.co.jp/shop/goods/030761011.htm).

Introduction

- We conduct a lexical study of classical Japanese poetry using network modeling.
- The terms *yamabuki* (kerria), *kahazu* (frog), and *Ide* (placename) are contained in some poetic dictionaries as entry items or collocations, and we have confirmed that they have strong relationships with each other.
- We have discovered the hub node term *yahe* in network models. The term *yahe* is, however, not recorded in any poetic dictionaries even as a single term.

Material: *Hachidaishū*

the eight anthologies compiled by the order of Emperors (ca. 905–1205), which contains about 9,500 poems.

Calculation methods:

$$w(t,d) = (1+\log\ tf(t,d)) \cdot idf(t)$$
$$cw(t_1,t_2,d) = (1+\log\ ctf(t_1,t_2,d)) \cdot cidf(t_1,t_2)$$
$$cidf(t_1,t_2) = \sqrt{idf(t_1) \cdot idf(t_2)}$$
$$idf(t) = \log \frac{N}{df(t)}$$

Result

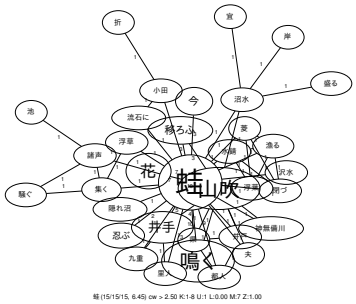


Figure 2: Graph model of *kahazu* (蛙, frog) before pruning node 蛙.

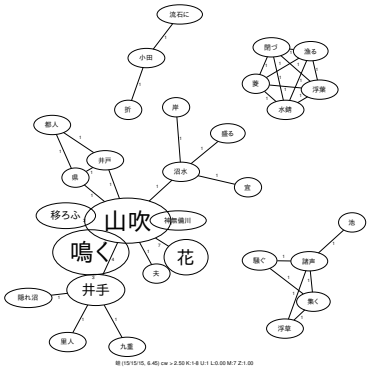


Figure 3: Graph model of *kahazu* (蛙, frog) after pruning node 蛙.

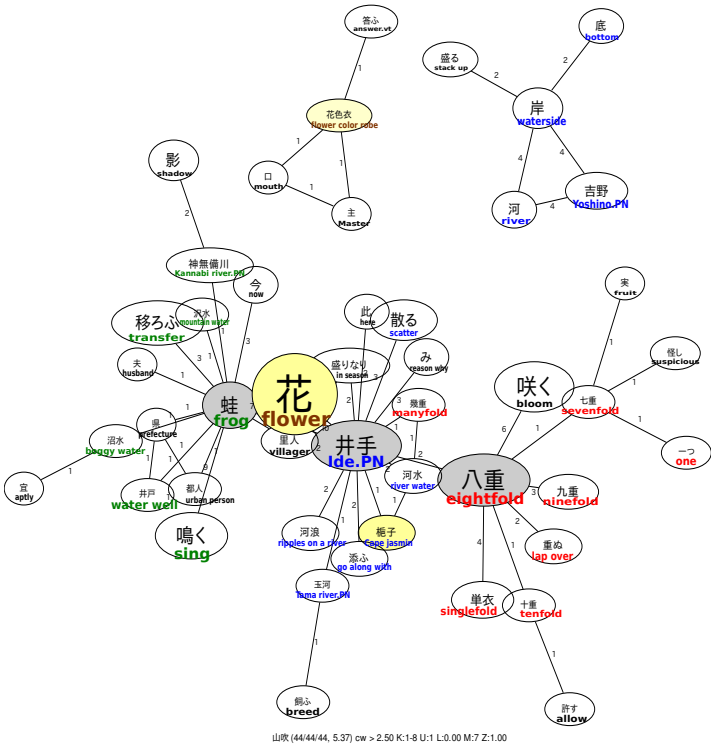


Figure 4: Graph model of *Yamabuki*: a core node, 山吹 *yamabuki*, is pruned. *kahazu* (蛙, frog), *Ide* (井手, place name, proper name), and *yahe* (八重, eightfold or double flower) are observed as hub nodes.

A minor term *yahe* (eightfold) can be shown as a hub node which plays a major role in connecting a topic word with other peripheral words which support/demonstrate poem stories. These minor words are not seen in poetic term dictionaries.

Conclusion

1. Discern not only patterns described by experts but also patterns yet undescribed, and
2. Identify not only specific or tangible words but also abstract or conceptual words which have a tendency to be left out of dictionaries.

Figure 5: Single petal (left), white petal (center), and plena petal (right) of *yamabuki*. (http://mkfarm.blog118.fc2.com/blog-entry-27.html)

Development of an Asymptotic Word Correspondence System between Classical Japanese Poems and their Modern Translations

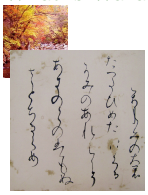
Hilofumi Yamamoto∗† **Hajime Murai**† **Bor Hodošček**‡
∗ **University of California, San Diego** †**Tokyo Institute of Technology** ‡**Meiji University**

Introduction

- This project will develop an automatic word alignment system for parallel texts comprising of Classical Japanese poems and their associated modern translations.
- By using these parallel texts, we will clarify the details of language change within Japanese in an objective procedural manner that is not influenced by human observations.
- Our aim is to develop a thesaurus of classical Japanese poetic vocabulary using the system.

Problem

What is Waka?



Tatsuta-Hime.. (5 syllables)
tamukuru KAMI no (7)
arebakoso (5)
aki no konoha no (7)
nusa to chirurame (7)

because Princess Tatsuta
has a god to whom she offers brocades,
the leaves of trees
in autumn will scatter as an offering.

1. Orthography Problem
   龍田, 立田, 竜田, たつた all indicate same placename: 'Tatsuta' in Nara pref.
2. Unit size Problem
   Does 卯の花 consist of one word or 卯/の/花 three words?
3. Attribution Problem
   Is 卯の花 the name of a flower or bean curd refuse?


4. Polysemy/PUN Problem
   海松藻 'mirume' a kind of sea weed; also means 見る目 (human eyes).

Methods

Material: ***Kokinshū*** a.k.a. *Kokinwakashū* is: the first anthology compiled by the order of Emperor Daigo (ca. 905), which contains about 1,100 poems. And 10 sets of **their Comtemporary Japanese Translations (CT)**

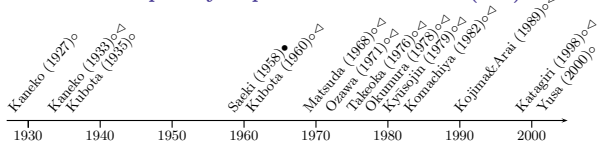


Figure 1: Dates of publication of annotations of the *Kokinshū*: ○ indicates that it has CT; ● indicates that it does not include CT; ▷ indicates that it is used in this project.

Mutual Co-occurrence Rate: Murai (2010)

$$mcr(o,t) = p(o\,|\,t)\;p(t\,|\,o)$$

where, o indicates a token in original texts; t, a token in translation texts; $mcr(o,t)$, the mutual co-occurrence rate; $p(o|t)$, the rate when a token o and t occur at the same time in corresponding texts which are original texts and translation texts.
→ when mcr is large enough, it will be estimated that token o and t are contextually equivalent.

Result

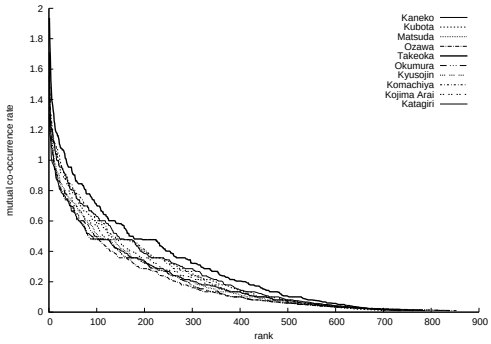


Figure 2: Distribution of Mutual Co-occurrence Rate: original text *Kokinshū* and ten sets of its translation texts.

Good or poor estimated pairs

Table 1: Good estimated pairs and poor estimated pairs; the values of good pairs are the first ten items (over 1.3); and the values of poor pair items are the last ten items (lower 0.01).

no.	good	pairs	poor	pairs
1	鳴く	鳴く cry	異なり	あの
2	風	風 wind	雫	どうして
3	世の中	世の中	此の	この
4	人	人 human	随に	まま
5	春	春 spring	匂ふ	美しい
6	秋	秋 autumn	見る	せい
7	時鳥	時鳥 cuckoo	連れ	つく
8	時鳥	ほととぎす	立ち返る	言う
9	散る	散る fall	有り	つく
10	見る	見る see	有り	まさしく

Conclusion

1. This project has already begun: the parallel corpus of the Kokinshū has been constructed.
2. We are now working on the development of computer software and the optimization of the calculation methods.

Reference

• Murai, Hajime. 2010 Extracting the interpretive characteristics of translations based on the asymptotic correspondence vocabulary presumption method: Quantitative comparisons of Japanese translations of the Bible. Journal of Japan Society of Information and Knowledge Vol. 20, No. 3, 293–310.

The differences of connotations between two flowers, plum and cherry, in classical Japanese poetry, 10th century.

Hilofumi Yamamoto Tokyo Institute of Technology

Introduction

- This project addresses an analysis of connotations of flowers in classical poetry: i. e., '*ume*' (plum) and '*sakura*' (cherry) .
- We will identify the characteristics of two flowers by computer modeling.
- Using parallel texts of original texts and contemporary translations of classical Japanese poetry, *the Kokinshū*, we will clarify the details of connotations in an objective procedural manner that is not influenced by human observations.
- The aim is to examine whether or not the residual of $CT - OP$ gives information on the non-literal elements of OP.

Problem

1. What is the difference between *ume* (plum) and *sakura* (cherry)?
2. What kind of connotations does each flower contain?
3. Which picture is that of cherry flowers?

a.  b.  c.  d.  e. 

Methods

Material: ***Kokinshū*** a.k.a. *Kokinwakashū* is: the first anthology compiled by the order of Emperor Daigo (ca. 905), which contains about 1,111 poems. And 10 sets of **their Contemporary Japanese Translations (CT)**

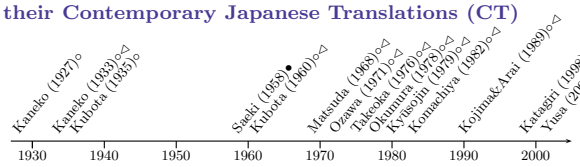
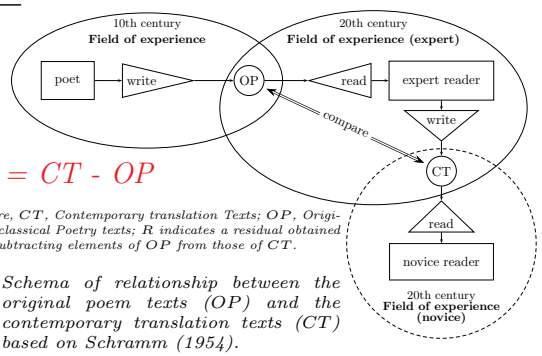


Fig. 1: Dates of publication of annotations of the Kokinshū: ○ indicates that it has CT; ● indicates that it does not include CT; ▷ indicates that it is used in this project.



$$R = CT - OP$$

where, CT, Contemporary translation Texts; OP, Original classical Poetry texts; R indicates a residual obtained by subtracting elements of OP from those of CT.

Fig. 2: Schema of relationship between the original poem texts (OP) and the contemporary translation texts (CT) based on Schramm (1954).

Result

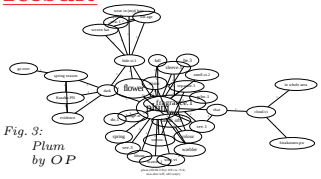


Fig. 3: Plum by OP

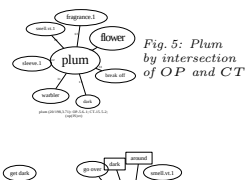


Fig. 5: Plum by intersection of OP and CT

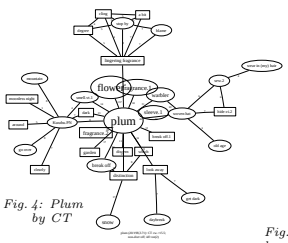


Fig. 4: Plum by CT

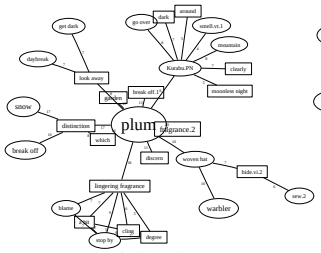


Fig. 6: Plum by subtracting OP from CT

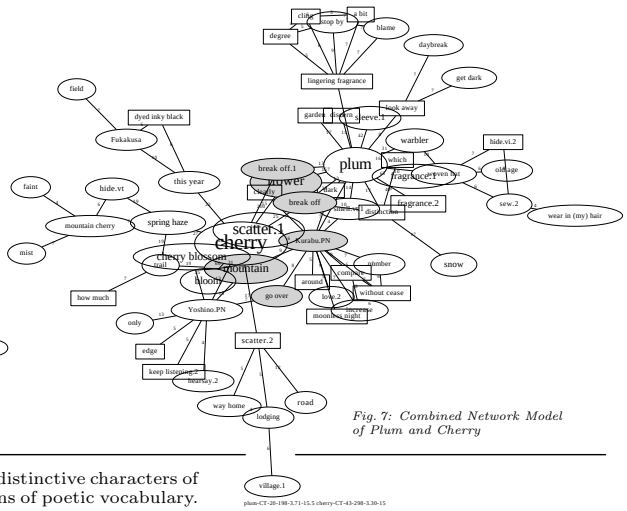


Fig. 7: Combined Network Model of Plum and Cherry

Conclusion

- It will be necessary to examine not only common nouns but also the distinctive characters of proper nouns in order to further examine the connotative associations of poetic vocabulary.
- We observed proper nouns such as place names, *Kurabu*, *Tatsuta*, *Otowa*, *Yoshino* in the network models of common nouns, and concluded that they seem to strongly influence the associations of poetic vocabulary.
- The relative salience clearly indicates that both *ume* (plum) and *sakura* (cherry) share *Kurabu yama* (Mt. Kurabu), which comprises a cluster of nodes in the sub-network.

Reference

- Schramm, W. L. 1954. How communication works. *The process and effects of mass communication.* 3–26. University of Illinois Press.
- Yamamoto, H. 2006. Extraction and Visualisation of the Connotation of Classical Japanese Poetic Vocabulary. Symposium for Computer and Humanities, 2006. The information processing society of Japan. Vo. 2006, No. 17, 21–8.

Development of the Dictionary of Poetic Japanese Description

Hilofumi Yamamoto
Tokyo Institute of Technology

Bor Hodošček
Osaka University

Introduction

- This paper proposes to further the development of a dictionary of classical Japanese poetry using pairwise term information (Yamamoto et al., 2014).
- Information on pairwise terms between an index and related term such as "flower–spring" is not included within traditional modern and classical Japanese dictionaries, even though this information connects terms with their contexts in a transparent way and thus offers an unbiased method for inferring the meaning of old Japanese terms.
- An R package for the analysis of linked communities in networks, linkcomm (Kalinka and Tomancak, 2011), is used to extract subordinate terms. Average, McQuitty, and single linkage methods are evaluated for the quality of their extraction of subordinate clauses of terms representing the 'cherry', 'plum', and 'orange' flowers. All methods extracted similar subordinate terms, which were quite natural in the context of classical Japanese poetry.

Problem

1. Many scholars of Japanese poetry have tried to explain poetic vocabulary based on their intuition and experience.
2. As scholars can only describe constructions that they can consciously point out, those that they are unconscious of will NEVER be uncovered. ⇒ In order to conduct more exact and unbiased descriptions:
   1) using computer-assisted descriptions;
   2) using co-occurrence weighting methods on corpora of Japanese poetry; and
   3) using linkcomm R package, extract the lists of words grouping sub communites.
   ⇒ allows one to BETTER GRASP the construction of poetic words.

Methods

Calculation: ***Linkcomm*** (Kalinka and Tomancak, 2011) for sub communities of three flowers: *ume* (plum), *sakura* (cherry), and *tachibana* (mandarin orange).

Material: ***Hachidaishū (ca. 905–1205)*** from *Kokkataikan* (Shin-pen Kokkataikan Henshū Committee, 1996), Nijūichidaishū database published by NIJIL (Nakamura et al., 1999), Shin-Nihon Koten Bungaku Taikei (Kojima and Arai, 1989), and Shin-kokinshū (Kubota, 1979).

Result

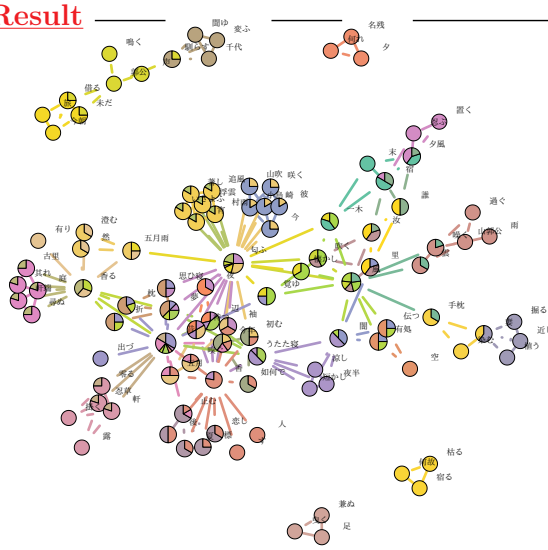


Fig. 1: Network of Words; mandarin orange.

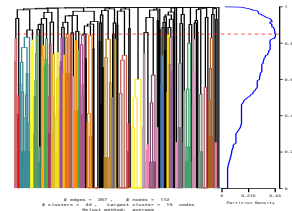


Fig. 2: Link Community Dendrogram.

Table 1: Sub-clusters of orange.

	average (.43)		mcquitty (.43)		single (.38)	
No.	node	edge	node	edge	node	edge
1	mukashi (old days)	7	mukashi	7	mukashi	5
2	nihofu (smell)	6	nihofu	6	nihofu	4
3	kaze (wind)	5	kotoshi	4	yume	4
4	yume (dream)	5	atari	4	kaoru	3
5	kotoshi (this year)	4	matsu	4	kotoshi	3
6	atari (aroud)	4	kaze	4	somu	3
7	matsu (to wait)	4	yume	4	samidare	3
8	kaoru (fragrance)	3	somu	3	ori	3
9	samidare (summer rain)	3	kaori	3	makura	3
10	somu (to dye)	3	yami	3	omohine	3

Conclusion

- Pairwise term information generated by the community centrality procedure works well.
- R package "linked communities" could extract proper sub cluster terms which contribute to the description of classical Japanese poetry.

Reference

- Kalinka, A. T. and Tomancak, P. 2011. linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type. *Bioinformatics*. 2011–2. **27** (14).
- Yamamoto, H., Hajime Murai, Bor Hodošček. 2014. Development of an Asymptotic Word Correspondence System between Classical Japanese Poems and their Modern Translations. Symposium for Computer and Humanities, 2014. The information processing society of Japan. Vol. 2014, No. 3, 157–62.

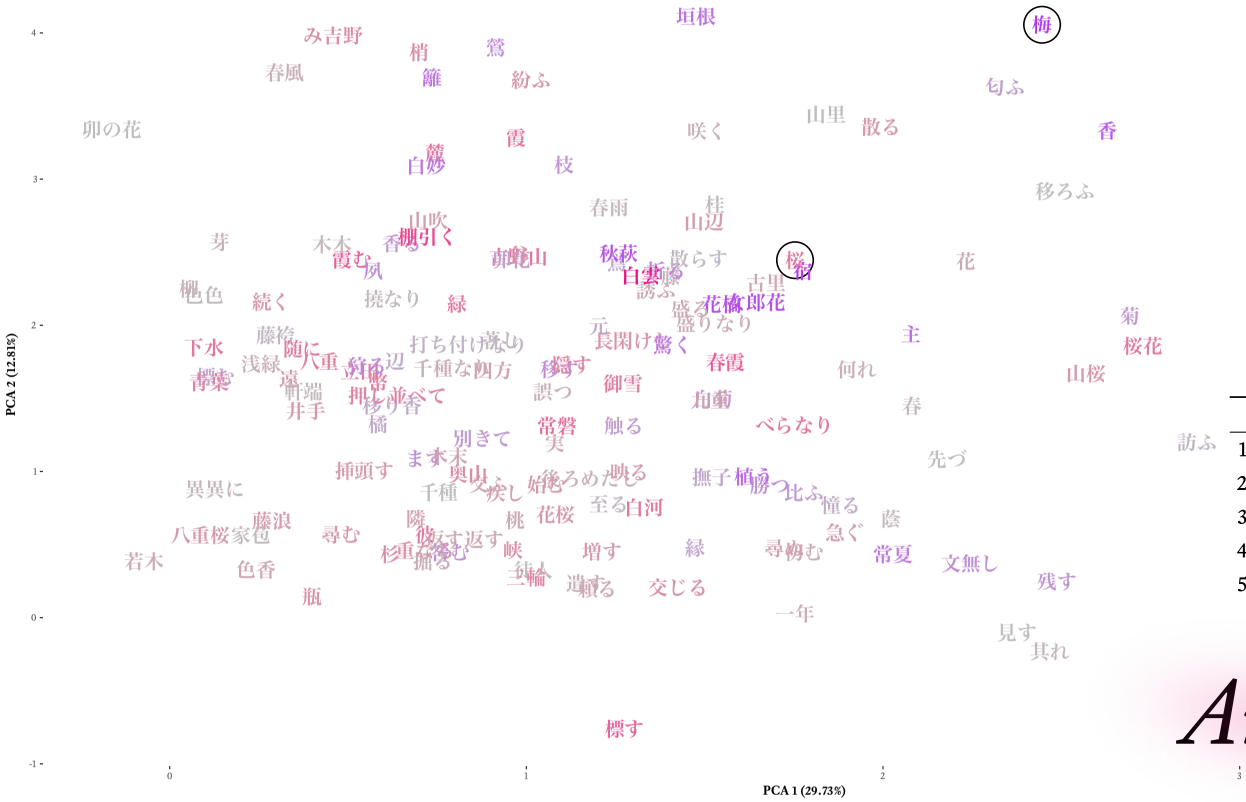
Relationships between Flowers in a Word Embedding Space of Classic Japanese Poetry

Hilofumi Yamamoto, Tokyo Institute of Technology
yamagen@ila.titech.ac.jp

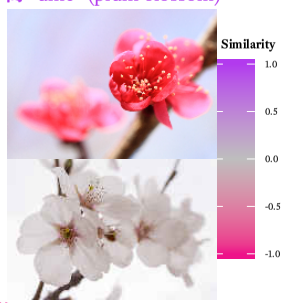
Bor Hodošček, Osaka University
bor@lang.osaka-u.ac.jp

@JADH2017
September 11

Examine the possibility of word embedding spaces (Word2Vec) to explain the semantic relationships between classical Japanese poetic terms within the *Hachidaishū* poem anthology. (ca. 905—1205)



梅 ‘ume’ (plum blossom)



桜 ‘sakura’ (cherry blossom)

	散る *chiru* (fall)	cosine similarity
1	盛る *moru* (raise)	0.95
2	桜花 ***sakura****bana* (cherry blossoms)	0.94
3	吉野山 *Yoshinoyama* (pn.)	0.93
4	八重 *yahe* (eight fold)	0.93
5	山桜 *yamazakura* (mountain cherry)	0.92

And more...

Figure 1: PCA of word embedding space (4157 words × 50 dimensions) filtered to include only top 100 similar words for each of ume and sakura (150 total). Similarity is represented by the difference in similarity scores between ume and sakura, scaled to [-1, 1].

Relationships between Flowers in a Word Embedding Space of Classic Japanese Poetry

Hilofumi Yamamoto, Tokyo Institute of Technology
yamagen@ila.titech.ac.jp

Bor Hodošček, Osaka University
bor@lang.osaka-u.ac.jp

Introduction

- Word embedding methods such as Word2Vec (Mikolov et al., 2013; Le and Mikolov, 2014) have been shown effective in extracting semantic knowledge from large corpora.
- Quantify the relationship between the content of a word and its word embedding vector.
- Examine the possibility of word embedding spaces to explain the semantic relationships between classical Japanese poetic terms.

Problem

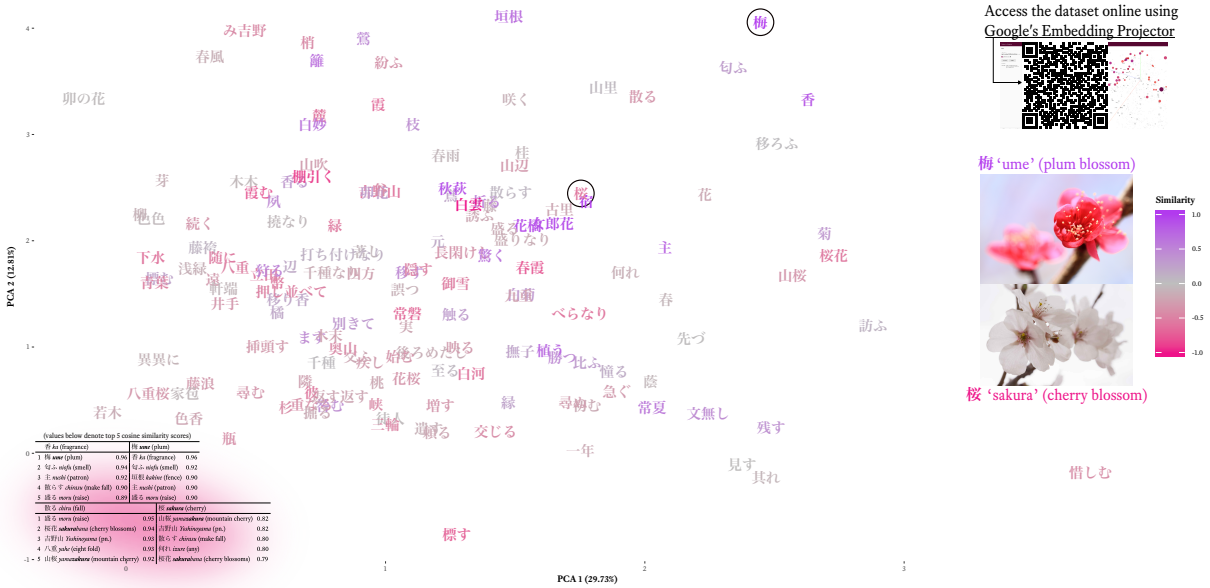
- Can word embeddings trained on the Hachidaishu encode enough semantic information to find subordinate words via their superordinate concept?

Materials

- *Hachidaishū*: classical Japanese poem anthologies compiled under decree by Emperors (ca., 905—1205), comprising approximately 9,500 poems and 159,183 tokens (Source: *KokkakaitanNijūichidaishū* database published by NIJIL).
- Each poem is tokenized into lemma forms by **kh** (Yamamoto, 2007) which divides poem texts into tokens using a classical Japanese dictionary.

Methods

- 50-dimensional skip-gram model with negative sampling, context window covering the whole poem using Gensim 2.3.0 (Řehůřek & Sojka, 2010).
- In order to examine the notable relationships between 'ka' (fragrance), 'chiru'(fall), we look at the cosine similarity scores between terms in the word embedding space generated by Word2Vec.



(values below denote top 5 cosine similarity scores)

香 *ka* (fragrance)		梅 *ume* (plum)	
1	梅 *ume* (plum)	0.96	香 *ka* (fragrance) 0.96
2	匂ふ *niofu* (smell)	0.94	匂ふ *niofu* (smell) 0.92
3	主 *nushi* (patron)	0.92	垣根 *kakine* (fence) 0.90
4	散らす *chirasu* (make fall)	0.90	主 *nushi* (patron) 0.90
5	盛る *moru* (raise)	0.89	盛る *moru* (raise) 0.90

散る *chiru* (fall)		桜 *sakura* (cherry)	
1	盛る *moru* (raise)	0.95	山桜 *yamazakura* (mountain cherry) 0.82
2	桜花 *sakurabana* (cherry blossoms)	0.94	吉野山 *Yoshinoyama* (pn.) 0.82
3	吉野山 *Yoshinoyama* (pn.)	0.93	散らす *chirasu* (make fall) 0.80
4	八重 *yahe* (eight fold)	0.93	咲く *saku* (bloom) 0.80
5	山桜 *yamazakura* (mountain cherry)	0.92	桜花 *sakurabana* (cherry blossoms) 0.79

Figure 1: PCA of word embedding space (4157 words × 50 dimensions) filtered to include only top 100 similar words for each of ume and sakura (150 total). Similarity is represented by the difference in similarity scores between ume and sakura, scaled to [-1, 1].

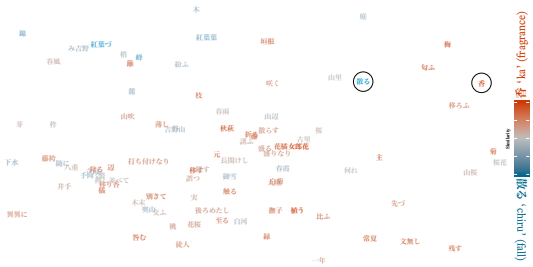
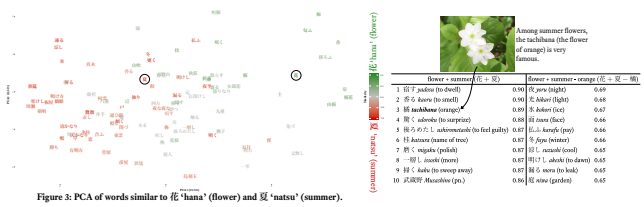


Figure 2: PCA of words similar to 香 'ka' (fragrance) and 散る 'chiru' (fall).



flower + summer (花 + 夏)		flower + summer + orange (花 + 夏 − 橘)	
1	宿す *yadosu* (to dwell)	0.90	咲 *saku* (to bloom) 0.69
2	香る *kaoru* (to smell)	0.90	光 *hikari* (light) 0.68
3	橘 ***tachibana*** (orange)	0.89	氷 *kōri* (ice) 0.67
4	驚く *odoroku* (to surprise)	0.88	顔 *kao* (face) 0.66
5	後ろめたし *ushiromotashi* (to feel guilty)	0.87	払ふ *harafu* (pay) 0.66
6	付 *kotsuke* (name of tree)	0.87	冬 *fuyu* (winter) 0.66
7	磨く *migaku* (polish)	0.87	涼し *suzushi* (cool) 0.65
8	一切し *issaki* (more)	0.87	明けし *akeshi* (to dawn) 0.65
9	棲く *kaku* (to creep away)	0.87	眺る *naru* (to look) 0.65
10	武蔵野 *Musashino* (pn.)	0.86	庭 *niwa* (garden) 0.65

Figure 3: PCA of words similar to 花 'hana' (flower) and 夏 'natsu' (summer).

Results

- 'ka' (fragrance) is related to 'ume' (plum) (replicating Mizutani, 1983).
- Falling flowers denote 'sakura' (cherry) and not 'ume' (plum); 'sakura' (cherry) relates to chiru (fall), which indicates that people at the time lamented falling sakura (falling cherry blossom petals) (replicating p. 84 in Katagiri, 1983).
- Subtracting tachibana out from the summer vectors reveals a vector space devoid of relationships between natsu (summer) and hana (flower). These relational expressions (summer + flower; summer + flower - tachibana) reproduce our current understanding of the relationships between flowers and seasons as well as some emotions associated with them in the word embedding space.

Conclusion

- Word embeddings allowed us to extract specific subordinate words based on the superordinate concept of classical terms → when the distance between two terms such as 'tachibana' (orange) and 'natsu' (summer) is close enough, the superordinate concept A indicates the subordinate concept *a*.
- We could therefore verify that it allows us to extract the concrete name from its superordinate concept.

References

Katagiri, Yoichi (1983) *Uhanakusa utakotoba jiten (Dictionary of poetic vocabulary)*, Vol. 35 of Kadokawa shojiten, Tokyo: Kadokawa Shoten.

Le, Quoc V. and Tomas Mikolov (2014) "Distributed Representations of Sentences and Documents," *CoRR*, Vol. abs/1405.4053, URL: http://arxiv.org/abs/1405.4053.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013) "Efficient Estimation of Word Representations in Vector Space," *CoRR*, URL: http://arxiv.org/abs/1301.3781.

Mizutani, Sizuo (1983) *Goi (Vocabulary)*, Vol. 2 of Asakura Nihogo Shin-Kōza, Tokyo, Japan: Asakura Shoten.

Řehůřek, Radim and Petr Sojka (2010) "Software Framework for Topic Modelling with Large Corpora," in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta: ELRA, May, http://is.muni.cz/publication/884893/en.

Rodd, Laurel Rasplica and Mary Catherine Henkenius (1984) *Kokinshū - A Collection of Poems Ancient and Modern*, Boston MA USA: Cheng and Tsui Company.

Yamamoto, Hilofumi (2007) "Waka no tame no Hinshi tagu zuke shisutemu / POS tagger for Classical Japanese Poems," *Nihongo no Kenkyu / Studies in the Japanese Language*, Vol. 3, No. 3, pp. 33–39.


A
DIGI
WITH
YAMAMOTO
HILOFUMI
TAL
HUM
ANITIES
11/
29
9a
CLINIC
WEST ELECTRONIC CLASSROOM,
CHARLES E. YOUNG
RESEARCH LIBRARY
3p
RICHARD C. RUDOLPH
EAST ASIAN LIBRARY
THE YANAI INITIATIVE
UCLA LIBRARY
TOKYO INSTITUTE
OF TECHNOLOGY
Image courtesy of the
Museum of Fine Arts, Boston
www.mfa.org


JAPANESE DIGITAL HUMANITIES: A HANDS-ON INTRODUCTION
WITH YAMAMOTO HILOFUMI
Yamamoto Hilofumi:
"Literary Studies and Japanese Language Processing"
Pete Broadwell:
"HYU:MA: A Model for Library-Supported Projects in Japanese Digital History"
11/30
12p
3p
WEST ELECTRONIC CLASSROOM,
CHARLES E. YOUNG
RESEARCH LIBRARY
RICHARD C. RUDOLPH EAST ASIAN LIBRARY
THE YANAI INITIATIVE
UCLA LIBRARY
TOKYO INSTITUTE OF TECHNOLOGY
Image courtesy of the Museum of Fine Arts, Boston
www.mfa.org

JAPANESE DIGITAL HUMANITES WORKSHOP '17

11/28-30 2017

WORKSHOP on Japanese Text Mining

11/29, 11am-1pm @Presentation Room (YRL Rm 11348A)

- Japanese corpora for mining: NINJAL, etc.
- Digital text-mining & analysis tools: MeCab, etc.; word2vec, text reuse analysis, and topic modeling.

Consultation on JDH Projects:

11/28-30, 1-5pm @West Classroom (YRL Rm 23167)

- 11/28: Extensive reading
- 11/29: Ryukyuan dialects, Speech corpora, *Genji Monogatari*
- 11/30: Other topics

Cosponsored by
Tadashi Yanai UCLA-Waseda Initiative for Globalizing Japanese Humanities,
UCLA Terasaki Center for Japanese Studies,
Tokyo Institute of Technology, and UCLA Library

11/29/2017

Wednesday

11 am – 1 pm

FREE WOKRSHOP

Japanese Text Mining

Hilofumi Yamamoto
Linguist and Professor
Tokyo Institute of Technology

Peter Broadwell
Academic Project Developer
UCLA Digital Library

RESEARCH LIBRARY (CHALRES E. YONG)

@ Presentation Room

(YRL Rm 11348A)
Open to UCLA faculty, students and staff

Light Lunch Provided

RSVP required

Contact: East Asian Library
Tomoko Bialock

A study on the distribution of cooccurrence weight patterns of classical Japanese poetic vocabulary

Hilofumi Yamamoto
Tokyo Institute of Technology

Bor Hodošček
Osaka University

2018.5.8

1 Introduction

The present study focuses on ongoing work exploring the threshold values dividing words in classical Japanese text into three groups: content, functional, and in-between. Content or semantic based analyses usually employ some techniques of data cleansing, such as eliminations of tags, punctuations, or symbols, as a preprocessing step. Stop words are also a type of token to be eliminated since they contain comparatively less meaning for content analysis. In general, it can be said that the most frequent words will be common words such as 'the' or 'and', which help build ideas but do not carry any significance themselves (Rajaraman and Ullman 2012: 8). Lists of stop words are commonly used, but have some problems: 1) it is necessary to compile them in advance; 2) they necessarily change depending on the domains of analyses; and 3) it is not clear which words should be included when analyzing classical texts.

Our previous study grouped modern Japanese words into low-, mid-, and high-range groups according to their information content given by their term frequency-inverse document frequency (*tf-idf*) and found that low-range words corresponded to infrequent and highly topical words, and high-range words corresponded to functional words expressing the grammatical relations between words. The study did not find an automatic method capable of classifying tokens into low-, mid-, and high-range. Furthermore, we found that previous research almost exclusively ignored the properties of the mid-range (Hodošček and Yamamoto 2013).

One of the methods used in Hodošček and Yamamoto (2013) exploited the occurrence not of individual words but of pairwise or cooccurrence patterns such as ‘ fragranceflower ’ relationships and revealed that the distribution of cooccurrence weights in modern Japanese texts approximately fitted a Gaussian curve. In this study, we will attempt to expand this analysis to classical texts by utilizing the characteristics of the Gaussian distribution to automatically group words into three clusters of cooccurrence patterns. We will show the differences between each of the three by visualizing with graph models.

2 Methods

We use the Hachidaishū as the material of the present study, which comprises the eight anthologies compiled under order of the Emperors (ca. 9051205) and contains about 9,500 poems. We developed the corpus and a method of cooccurrence weighting similar to the *tf-idf* method, *cw* (Yamamoto 2006), which calculates the weight of patterns of any two words occurring in a poem sentence (Spärck Jones 1972, Robertson 2004, Manning and Schütze 1999, Rajaraman and Ullman 2012).

$$
\begin{aligned}
w(t,d) &= (1+\log\ tf(t,d)) \cdot idf(t) \\
cw(t_1,t_2,d) &= (1+\log\ ctf(t_1,t_2,d)) \cdot cidf(t_1,t_2) \\
cidf(t_1,t_2) &= \sqrt{idf(t_1) \cdot idf(t_2)} \\
idf(t) &= \log \frac{N}{df(t)}
\end{aligned}
$$

Where w is a weight, t a token, and N the number of tokens. The function *idf* is called the “inverse document frequency.”(Spärck Jones 1972, Robertson 2004, Manning and Schütze 1999) The function *cw* is called the “co-occurrence weight,” which allows us to examine the patterns of poetic word constructions through mathematical modeling.

As in Fig. 1, there is a concept (Losee 2001: 1019) of terms located in each layer being effective query terms. Luhn (1968) cuts the top and bottom words of the frequency and uses mid-range vocabulary for the development of an automatic outline generation system.(Fig. 1) Nagao (1983: 28) also mentioned mid-range vocabulary to be effective in generating automatic abstracts. Nagao (1983)’s viewpoint is slightly different with Luhn (1968) in that it allocates the distribution of word lengths around the Gaussian curve. The positions both upper cutoff and lower cutoff are, however, assumed to be empirical; it is not discussed where to cut them off.

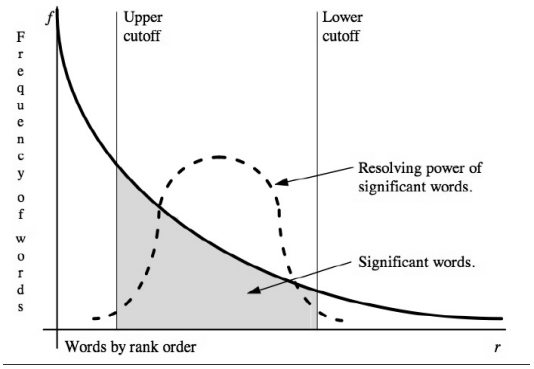


Fig. 1: Hyperbolic curve relating occurrence frequency with rank order; adapted from (Luhn 1968: 120)

3 Results

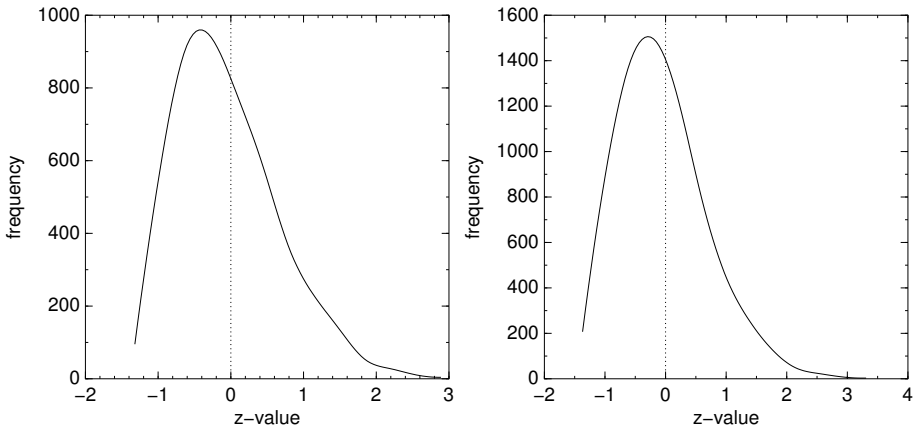


Fig. 2: The distribution of cw values *ume* (plum; left) and *sakura* (cherry; right) in Hachidaishū; The statistics of *ume* (plum): N=7016, min=-1.370, mean=0.138, max=3.700, SD=0.740, SE=0.009, CV=534.012%, Reliable interval low - upper = 0.116 - 0.161 (95%), skew=0.737, kurtosis=3.567, and that of *sakura* (cherry): N=4734, min=-1.320, mean=0.132, max=3.240, SD=0.716, SE=0.010, CV=544.116%, Reliable interval low - upper = 0.104 - 0.159 (95%), skew=0.740, kurtosis=3.345 indicate both approximately fit a Gaussian curve

The distribution of *cw* values is taken from the network model of both *ume* (plum) and *sakura* (cherry) and their curves belong to Gaussian curve as well as in classical texts (Fig. 2). Therefore we will attempt to divide this shape into three layers by inflection points.

The co-occurrence patterns of *sakura* (cherry) under -0.9 (near -1) *cw* value are adjacent patterns comprising function words, and over 1 *cw* value are those of the patterns with content words as we expected (Table 1 and

Table 1: Upper cutoff patterns of *ame* (sakura): *cw* = co-occurrence weight; *z* = z-value (normalized value of frequency). word annotations: ari(be), ba(cond.), ha(topic.), hana(flower), hito(human), keri(past.), ki(past.), koso(emphatic.), miru(see), mo (also), nasi(no exist), nu(neg.), o(obj.), omou(think), ramu(aux.will), su(do), te(p.), to(and), ware(we), zo(emphatic.), zu(neg.)

	cw	*z*	pattern		*cw*	*z*	pattern		*cw*	*z*	pattern
1	0.62	-0.91	mo–keri	11	0.59	-0.96	nasi–ha	21	0.52	-1.05	nu–o
2	0.62	-0.92	hana–o	12	0.57	-0.98	o–ramu	22	0.52	-1.05	o–zo
3	0.62	-0.92	o–koso	13	0.57	-0.98	mo–ramu	23	0.52	-1.05	miru–o
4	0.60	-0.94	zu–keri	14	0.57	-0.98	ha–ki	24	0.48	-1.09	ba–mo
5	0.60	-0.94	su–ha	15	0.56	-1.00	zu–mo	25	0.48	-1.09	o–keri
6	0.60	-0.94	to–ba	16	0.56	-1.00	o–te	26	0.43	-1.16	zu–ha
7	0.59	-0.96	ari–ha	17	0.55	-1.01	hito–mo	27	0.43	-1.16	to–o
8	0.59	-0.96	ari–mo	18	0.54	-1.02	zu–te	28	0.43	-1.16	te–ha
9	0.59	-0.96	ware–mo	19	0.52	-1.05	zo–ha	29	0.34	-1.27	o–ha
10	0.59	-0.96	nasi–o	20	0.52	-1.05	omou–o	30	0.34	-1.27	o–mo

2). As for the upper cutoff, we used an under -0.9 (near -1) σ value of *cw*, which could extract patterns of functional tokens: almost all patterns included functional words, while as lower cutoff, we used over 1 σ values, which could extract patterns of content tokens: almost all patterns included content words. Both under -1 and over 1 σ are regarded as inflection points which have mathematically interesting property.

4 Discussion

Inflection points are defined as the points on the curve where the curvature changes its sign while a tangent exists.(Bronshtein et al. 2004: 231) We consider the threshold values that separate upper cutoff, mid-range, and lower cutoff not as coincidental but as evidential points. It is, however, necessary to conduct further experiments and continue to discuss the mathematical traits behind the distributions of cooccurrence weights. In terms of removing the low-range (upper cutoff) and extracting the high-range (lower cutoff) from poetic texts, we found that we do not need to use any filters to eliminate terms, since *cw* values returned semantically cooccurring patterns. Apart from low-range and high-range, the characteristics of the mid-range lexical layer are still unknown.

5 Conclusion

Using the distribution characteristics of cooccurrence weights, we were able to classify cooccurrence patterns into three layers of cooccurrence patterns: high-, mid-, and low-range patterns.

Table 2: Lower cutoff patterns of *ame* (sakura) in Kokinshū: 30 out of 164 patterns extracted; *cw* = co-occurrence weight; *z* = z-value (normalized value of frequency) word annotations: ba(cond.), bakari(only), besi(should be), chiru(fall), fukakusa(deepgreen), hana(flower), isa(already), kakusu(hide), katu(win), koku(pull), komoru(go deep inside), magiru(mix), makasu(entrust), maku(wind up), manimani(as it is), masi(as), mazu(mix), me(eye), minami(south), miyako(city), mono(thing), nagara(even if), sakura(cherry), si(emphasic.), sumi(black ink), tatu(start,stand), tazumu(being around), tu(past.), uturou(change), watasu(give), yamakaze(mountain wind), yamu(stop), yanagi(willow), yononaka(world)

	cw	*z*	pattern		*cw*	*z*	pattern
1	3.86	3.18	yamu–manimani	106	2.38	1.31	si–fukakusa
2	3.75	3.04	minami–magiru	107	2.38	1.31	sakura–hana
3	3.67	2.93	minami–maku	108	2.38	1.31	sakura–isa
4	3.61	2.86	maku–magiru	109	2.38	1.31	sakura–ba
5	3.42	2.62	yanagi–koku	110	2.38	1.30	sakura–me
6	3.38	2.57	yamu–makasu	—			
7	3.38	2.56	mazu–koku	155	2.17	1.04	chiru–katu
8	3.27	2.43	yanagi–mazu	156	2.17	1.04	bakari–sumi
9	3.26	2.42	sakura–yamu	157	2.16	1.03	maku–besi
10	3.25	2.40	minami–yamakaze	158	2.16	1.03	tatu–maku
–				159	2.16	1.03	tatu–tazumu
101	2.40	1.33	uturou–komoru	160	2.16	1.03	tazumu–tu
102	2.40	1.33	sakura–watasu	161	2.16	1.03	miyako–sakura
103	2.40	1.33	katu–nagara	162	2.16	1.02	kakusu–si
104	2.39	1.32	sakura–masi	163	2.14	1.00	yononaka–sakura
105	2.39	1.31	sakura–makasu	164	2.14	1.00	mono–sakura

We found that 1) the distribution of classical texts fits a Gaussian curve as well as in modern texts; 2) the *cw* value can separate patterns into three layers (low-, mid-, and high-range) using inflection points (-1σ and 1σ); 3) of the three layers, the high-range could be extracted without a list of stop words; 4) the mid-range lexical layer might include mathematical traits not yet revealed in the present study.

References

Bronshtein, I.N., K. A. Semendyayev, G. Musiol, and H. Muehlig (2004) *Handbook of Mathematics*: Springer-Verlag, 4th edition.

Hodošček, Bor and Hilofumi Yamamoto (2013) "Analysis and Application of Midrange Terms of Modern Japanese", in *Computer and Humanities 2013 Symposium Proceedings*, No. 4, pp. 21–26.

Losee, Robert M. (2001) "Term dependence: A basis for Luhn and Zipf models", *Journal of the American Society for Information Science and Technology*, Vol. 52, No. 12, pp. 1019–1025.

Luhn, Hans Peter (1968) *HP Luhn: Pioneer of Information Science: Selected Works*: Spartan Books.

Manning, Christopher D. and Hinrich Schütze (1999) *Foundation of statistical natural language processing*, Cambridge, Massachusetts: The MIT press.

Nagao, Makoto (1983) *Gengo kogaku (Language Engineering)*, Jinkochino sirizu 2 (Series of Artificial Intelligence): Shokodo.

Rajaraman, Anand and Jeffrey David Ullman (2012) *Mining of massive datasets*, Cambridge: Cambridge University Press.

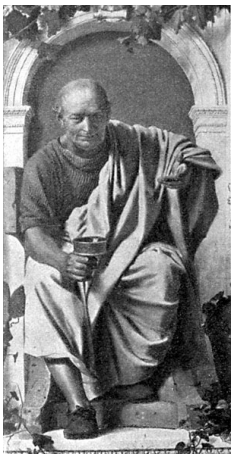
Robertson, Stephen (2004) "Understanding inverse document frequency: on theoretical arguments for IDF", *Journal of Documentation*, Vol. 60, pp. 503–520.

Spärck Jones, Karen (1972) "A Statistical Interpretation of Term Specificity and Its Application in Retrieval", *Journal of Documentation*, Vol. 28, pp. 11–21.

Yamamoto, Hilofumi (2006) "Konpyūta niyoru utamakura no bunseki / A Computer Analysis of Place Names in Classical Japanese Poetry", in *Atti del Terzo Convegno di Linguistica e Didattica Della Lingua Giapponese, Roma 2005*: Associazione Italiana Didattica Lingua Giapponese (AIDLG), pp. 373–382.

裏表紙について

写真の日時計にはラテン語で “Carpe Diem”（カルペ・ディアム）と彫ってあります。英語では “Seize the day”、日本語では「その日を摘め」と訳されています。そこには「その日を楽しみ、精一杯いきること」という意味があります。紀元前 1 世紀の古代ローマの詩人ホラティウスの詩に登場する句で、映画 “Dead Poets Society”（1989 年、邦題「いまを生きる」ロビン・ウィリアムズ主演）にも出てきます。



クィントゥス・ホラティウス・フラックス
Quintus Horatius Flaccus
BC.65.12.8–BC.8.11.27
古代ローマ時代の南イタリアの詩人

言語と文化、東京工業大学の取り組み
山元研究室カタログ
2018 年 4 月 12 日 第 1 版
著者: 山元啓史

Carpe Diem

東京工業大学
Tokyo Institute of Technology

HisaoTaki Japanese Education supporting project

科研費
KAKENHI