修士論文

テキスト分析によって明らかにする
新語によるコミュニケーションを可能にする複合語の語形成

大友和幸

主指導教員　　橋本　敬

北陸先端科学技術大学院大学
先端科学技術研究科
（知識科学）

令和 4 年 3 月

Abstract

In language, there are novelty and commonality. Commonality is the property that the meaning of a word is generally used as a fixed one, and its meaning is shared by users. In cognitive linguistics, semantic extension and language polysemy show that a single form of a word can have multiple meanings and multiple usages. On the other hand, not all of these meanings are used by all people. Therefore, the property of commonality is a necessary component of communication, but it is not high in all cases. On the other hand, we can also express new concepts in language using new forms that have never been used before (novelty). Indeed, prior research has shown the emergence of new forms of words to successfully express new concepts for the purposes of humor, irony, evaluation, amplification, mitigation, etc. The study also confirms that certain words are used many times as constituents of compound words, and that compound words are constructed by these words. While commonality tends to lower ambiguity through habitual patterns and allow for as accurate a transmission as possible, novelty tends to hinder accurate transmission because the new words generated are strictly known only by the speaker, and only he or she knows their meaning. Therefore, because of the opposing tendencies of commonality and novelty, it has not been possible to account for both the generation and communication of new words while ensuring both properties. By analyzing product names that require both novelty and commonality, this study aims to reveal the mechanism by which commonality and novelty are compatible. This will explain the mechanism how communication is possible through new words that have never existed before. First, we propose a method to measure how much a word is used with the same meaning, and a method to evaluate the emergence of the overall meaning of compound words that occurs when existing words are combined to form compound words. Then, using the proposed method, we evaluated the commonality and novelty of the newly emergent compound nouns in both form and meaning. The results showed that in terms of both form and meaning, the component words had more in common than the compound words. Formal novelty was found for four words. The analysis of the novelty of meaning for these words revealed that the analysis of frequency of occurrence detected meanings that were found only in compound words. From this result, it can be said that the word formation of a compound word, which can be considered a new word in terms of form, is secured by

combining constituent words that have something in common and giving them an emergent meaning (semantic novelty) that is not found in any of the constituent words. This commonality is not only the result of habitual use over a long period of time (formal commonality), but also the fact that many people use the word in the same way (semantic commonality).

目次

図目次

表目次

第1章 はじめに

本研究では、コーパス分析により複合語の語形成を観察することで、言語における共通性と新規性が両立するメカニズムの解明を試みる。研究背景として、新語の出現や、意味論における共通性を紹介し、両立するうえでの課題と研究目的について共有する。それと共に、手法を紹介し、今回使用するコーパス言語学について紹介する。最後に本論文の構成について説明する。

1.1 背景

1.1.1 言語の共通性と新規性

コミュニケーションに用いられる言語は、共通性と新規性という 2 つの性質を持つ。(Lehrer, 2003, p233)

共通性とは、語の意味がおおむね決まったものとして使用され、その意味が使用者に共有されている、すなわち使用者によって千差万別の意味を持たない性質を指す。認知言語学の特徴の一つともいえるプロトタイプ意味論においては、ある語の意味を考えるとき、必要条件と共に典型条件を定めている。また、認知意味論では拡張義も認めている。その一方で、複数ある意味の全てが全員で共有されているとは限らない。そのため、人によって特定の語の意味や使用法は異なる場合があり、共通性は全ての場合において高いとは言い難い。

また、信号理論(Shannon, 1948)に基づくコミュニケーションの見方では、話し手が言語化（外在化）したメッセージを聞き手が理解するためには、コミュニケーションを行う双方が語彙や文法に関する知識を共有していることが必要となる。したがって、共通性はコミュニケーションの前提と考えられる。

その一方で、我々は、今までになかった新しい形式を用いて新しい概念を言語で表現することもできる。すなわち、言語は新規性を生み出す創造的な性質を持つ。ファッションコミュニティを対象にしたコーパス分析(Camiciottoli, 2016)では、ユーモア、皮肉、評価、増幅、緩和等を目的として、新たな概念をうまく表現するために新しい形式の語の出現が示されている。このように、既存の習慣化された語ではなく、新しい語の形式が誕生することが現象として確認されている。

共通性が習慣化したパターンによって曖昧さを下げてできるだけ正確な伝達を可能にする傾向であるのに対し、生成された新語は厳密には話し手しかその意味を知りえな

いため、新規性は正確な伝達を妨げる傾向にある。使用される語や文法に関する知識が双方に共有されていない場合には、自分の意図を正確に相手に伝えることができず、発話に表現される意図を伝達するという意味でのコミュニケーションを成り立たせることが困難となる。このように、共通性と新規性には相反する傾向性があるため、両方の性質を担保したまま新語の生成とコミュニケーションの両方を説明することができていない。

1.1.2 複合語が持つ意味の創発性

新語を生むもっとも良く使われる方法のひとつが、既存の語を組み合わせて新しい形式を作る複合語である(Brown,, 2015)。複合語の意味は部分的合成性を持つと言われる(野田, 2011, p.8)。すなわち、複合語の意味は構成要素の意味と無関係ではないが、単純な他仕合せでもないため、個々の構成語の単純な総和として複合語全体の意味を説明できない。メンタルスペース理論における概念融合(Fauconnier& Turner, 1996)はその典型である。たとえば「butcher surgeon」という表現は「下手な外科医」という意味を持ち得るが(Grady, 1999)、butcher「屠殺業者」が「下手な」という意味は持つわけではなく、surgeon「外科医」と組み合わさったときに生じる創発性を持つと考えられる。創発とは、全体が要素の集合以上の性質を持つことであり、ここでは複合語の意味が構成要素の意味の単純な和ではないことである。

複合語が持つ意味の部分的合成性あるいは創発性から考えると、共通性と新規性が両立しつつコミュニケーション（伝達）を成立させるメカニズムを説明する仮説として、以下のものが考えられる。

習慣的に使用され共通性を持つ要素同士を組み合わせ、その組み合わせパターンによって複合語全体で新規性を持たせる。

すなわち、新しい概念を表現する際に、全く新しい語をあてがうのではなく、既存の語を組み合わせることにより新しい概念を表現する。このような語を使用することで、今までに聞いたことがない新語に対しても意味を理解し、コミュニケーションを成立させられる。

1.2 目的

本研究は共通性と新規性が両立するメカニズムの解明し、今までになかった新語によって、コミュニケーションが成り立つメカニズムを解明する。

具体的には、名詞として実際のコミュニケーションに用いられる複合語に限定し、新

規に出現した複合語の形式と意味の両方において、共通性を持った語を構成語として結合することによって新規性を持つ複合語が作られていることを示す。

また、そのためには、意味的な共通性と意味的な新規性を評価し、既存の語が結合され複合語になる際に生じる複合語全体の意味の創発性を評価する方法を提案する。

以上より、研究の小目的は以下のものとなる。

1. 意味的な共通性・新規性の評価の方法を提案する。
2. 形式的な観点で見た際、共通性を持った構成語を結合させることによって、新規性がある複合語が形成されている。
3. 意味的な観点で見た際、共通性を持った構成語を結合させることによって、個々の構成語には無い意味(創発的意味)が付与され、意味的に新規性がある複合語が形成されている。

1.3 手法

本研究では、コーパス言語学に則り、実際の語の使用ログデータをもとにした、実証的な検証を試みる。

上記の目的を達成するため、新規に出現した複合語や、それを分解した構成語について、出現の推移を計測し、形式面での新規性・共通性の評価を行う。形式面の評価では、ある語が登場する前から習慣的に使用されている性質に着目し、その性質の高さをもって共通性を、その低さをもって新規性の評価を行う。また、新規性では、ある語が作られただけではなく、「今までにはなかった新しい形式の語が普及している」という性質の評価をするため、もともと使われていなかったことに加え、新語の登場後には使用されるようになった性質も評価する。

意味の分析でも同様に、複合語・構成語について分析を行う。意味の共通性においては、多くの人が同じ語を使用している性質を評価する。

意味の創発性を評価するために、複合語にはあるが各構成語にはない共起語を分析し、複合語のみに存在する共起語の検出を行う。そして、その語を観察することによって、創発的意味の検出を試みる。また、LDA を使用した意味の検出手法を提案し、構成語が結合することによって付与した意味について分析手法を提案し、実際の分析を行う。

1.4 本章の構成

本論文では２章でコーパス言語学の概観や、言語における共通性と新規性について紹介する。また、それと共に、仮説を立てる上で参考になった、複合語における意味の創発を紹介する。３章では、複合語として収集した商品名の収集方法と、形式・意味のそれぞれにおいて、共通性と新規性を評価する。４章では３章で紹介した結果を共有するとともに５章で総合的なまとめをする。

第2章 関連研究と本研究の位置づけ

本章では、本研究における問題意識を述べるとともに、研究の仮説の参考となる研究を紹介する。問題意識としては、前章で述べた新規性と共通性に関する具体的な説明を行う。また、仮説の話にも登場する新語の語形成に関する研究ファッションコミュニティについての実際の調査も共有する。

2.1 コーパス言語学

本研究では、コーパス言語学による分析を試みる。ここでいうコーパス言語学とは以下のように説明される。

> コーパス言語学とは、電子化された大規模な言語テキストに基づき、コンピュータを駆使して主として実証的観点から言語の諸特性を観察・調査・記述・分析する研究実践の総称である(石川, 2012)

つまり、コーパス言語学においては、大規模言語データを分析することによって、実際の言語がどのように使用されているのかについて実証的な検証方法を採用した上で分析を行う研究である。

また、コーパス言語学では、言語学で伝統的に使用されていた作例や内省などといった手法を避け、頻度調査、共起語調査、共起ネットワークによる分析などの手法により、実際の言語使用を元に検証を行うことも特徴といえる。

2.1.1 コーパス言語学を採用するメリット

コーパス言語学を使用するメリットとして、微細な意味の検出が挙げられる。コーパス言語学においては、類義語の違いなど、辞書的な意味では分析しにくいものについて分析をできることが挙げられる。「辞書」と「辞典」の違いについての分析(石川,2014)や、「イロイロな」と「サマザマな」の表記の差(石川,2012)などでは、類義語間の微細な意味の違いや、使用法の違いなどについて言及している。これは、母語話者による内省でも区別が困難とされるものである。また、このような意味の違いなどは、辞書に掲載されているものではない。この様な事例について扱うことができるのが、コーパス言語学の強みと言える。

2.1.2 コーパスとはなにか

上記の分析を行うために、コーパスは自然な言語を代表するものであり言語使用のデータベースとしての役割を担っている。

石川はコーパスについての定義をいくつか紹介したうえで、コーパスを定義する上で満たすべき条件を以下のものとしている。

1. 書き言葉や話し言葉などの現実の言語で構成されていること
2. 大規模であること
3. 基準に沿って網羅的に代表的に収集されていること
4. コンピュータ上で処理できるデータとして保存してあること
5. 言語研究を目的としていること(石川,2012)

「大規模」とは 10 万語程度でも足りるものでなく、全く出現しない語も多く存在する。コーパス全体のサイズが小さいと、出現頻度の低い語などの分析では、十分なコンコーダンス(対象の文字列が含まれるテキスト)が取得できず、分析に使用できない。そのため、(石川,2014)では、完成した世界発のコーパス(Brown Corpus)の規模にならって、おおよそ 100 万語程度を使用の最低ラインの目安としている。

2.1.3 ウェブ上のテキストデータを利用したコーパス言語学

語の意味や使用法の解明のためにウェブ上のテキストデータを使用する動きも存在する。ウェブ上のテキストがコーパスと見なせるかどうかについては「良いコーパスかどうか」の問いとは分けたうえで、「コーパス」の定義を「テキストの集合体」ととりWeb 全体はコーパスとしてみなされており (Kilgarriff&Grefenstette, 2003)、実際の検証にも使用されている。

ウェブ上のデータによるコーパス構築の一例として、幅広いテキストタイプや話題を包含する言語資源をウェブから自動収集することを目的として、WaC というプロジェクトがある。(石川, 2012) WaC(Kilgarriff&Grefenstette, 2003)は Web as Corpus の略で、ウェブ全体を巨大なコーパスとしてみなし、言語学の実証のために Web 上の言語データを使用する試みである。また、WaC を発展 TenTen プロジェクトも存在し、WaC よりもさらに大規模なコーパスの構築が進められている。

ウェブ利用のメリットとしては、その規模が挙げられる。語の出現回数の観察といった定量的な計算においては、データの信頼度が増し、統計的に推定精度が増す。また、語の不在証明を行う必要がある際にも、より広い範囲をカバーする大規模なコーパスで

ある方が信頼度が増す。それとともに、出現頻度が低い語を観察できる点も挙げられる。全体の規模が大きいほど低頻度語の出現回数も大きい事が期待されるため、小規模のコーパスでは十分な使用が認められなかった語に対して、分析を行うことができる。

その一方で、デメリットとして、大きすぎる規模では分析しきれないことが挙げられる。あまりにも大きすぎるコーパスでは、計量的な計算による検証はできるものの、実際の語の使用法を一つひとつ観察するなどといった定性的な分析は困難となる。

2.2 言語の共通性と新規性

目的にも示した通り、コミュニケーションに用いられる言語は、共通性と新規性という 2 つの性質を持つ。

2.2.1 共通性

共通性とは、語の意味がおおむね決まったものとして使用され、それが使用者に共有されている性質を指す。

信号理論(Shannon, 1948)に基づくコミュニケーションの見方では、話し手が言語化（外在化）したメッセージを聞き手が理解するためには、コミュニケーションを行う双方が語彙や文法に関する知識を共有していることが必要となる。したがって、共通性はコミュニケーションの前提とされているといえるだろう。

その一方で、共通性は全ての場合において高いとは限らない。語によっては使用する意味が人によって違い、一つの語でも同じように使用されない場合が存在する。その例として「期待値」と「進化」の事例を紹介する。

また、語の意味が一意に定まらないことが起きる候補の例としては、プロトタイプ意味論と多義性を紹介する。

共通性が高いとはいえない例：「期待値」

反対に共通性が低くなる例として、「期待値」の例を紹介する。広辞苑によると、期待値は「確率変数が取る値を、各値が生じる確率を考慮してへいきんしたもの確率変数が離散的な値を取るときは、各値に対応する確率をそれぞれ掛けて加えた値」(広辞苑)のみとされており、実際に統計学を中心とした幅広い分野では、この意味で使用されている。その一方で、「期待する基準の事で、最低限クリアしたいライン」(ライフコーディネート学校, https://lifecoordinate.com/skill/1913/)(広辞苑ではこの意味の記載は無い)のような意味でも実際に使用さる。広辞苑の意味(確率分布の加重平均)ではないものについても、同じ形式をもつ語が全く違う意味で用いられる。そのため、広辞苑の意味でしか使用していない人は、「最低限クリアしたいライン」としての期待値の意味を知らずにいるため、期待値の意味の一部のみしか使用していないということになる。

共通性が高いとはいえない例：「進化」

「期待値」と同様に、意味や使用法が人によって異なる語として「進化」の例が挙げられる。進化とは、Wikipedia では以下の通りに述べられている。

> 進化（しんか、羅：EVOLUTIO、英：EVOLUTION）は、生物の形質が世代を経る中で変化していく現象のことである

その一方で、「進化」を以下のような意味で使用することもある。

> 進化（しんか）とは、特定の条件を満たすことによってポケモンの種類が変わる現象のこと。（ポケモン Wiki「進化」, https://wiki.xn--rckteqa2e.com/wiki/%E9%80%B2%E5%8C%96）

これは、生物学をはじめとする多くの分野で使用される「進化」とは意味がことなっている。実際、このゲーム内では、幼虫が成虫になるかのように一つの個体がその姿を変える意味として使用されており、これは進化ではなく変態である。その一方で、学術的な意味での進化を知らないがポケモンをする人からすると、これは変態の意味での進化しか使用しない。そのため、学術的な意味のみを使用する人と、変態の意味のみを使用する人がいた場合、共通性が低い場合として考えられる。

プロトタイプ意味論

本研究の立場である認知意味論においては、プロトタイプ意味論を採用しており、語の意味については意味拡張も認めたうえで、ある程度一定である基本的な意味が存在するとしている。

プロトタイプ意味論の特徴として、語の意味の違いにはっきりとした条件があるわけではなく、あいまいな境界線によって語の区別がなされている点があげられる。語の意味を要素に分割し、その要素(意味特徴)の集合によって語の意味を記述しようとしたチェックリスト意味論では、ある語に対して、その語の意味に属するかを決める十分条件が存在することが前提であった。そのため、「queen」の意味を[+woman][-man][+king]のように記述すると、この[+woman][-man][+king]が「queen」の意味の十分条件にあたる。そのため、チェックリスト意味論においては語の意味は常に一定であり、十分条件があることによって規定されているといえる。

その一方で、プロトタイプ意味論では、語の意味については典型的な要素と周辺的な要素の存在を認める(松本, 2003)。そのため、語の意味を定義するときに、必要条件だけでなく、典型条件も存在する。そして、語の意味を表すプロトタイプの型は、fillmoreによると、「red」「bird」「climb」「bachelor」の４つに分かれるという(fillmore, 1982a)。

ここで、各意味論における共通性を考えると、チェックリスト意味論では、語の十分

条件が決まり、語と語の境界が明瞭であるため、語の定義を明確にすることができるとしている。その一方で、プロトタイプ意味論では、語と語の境界が明瞭でないため、完全に語と語の境界が決まったものとは言い難い。典型条件があり、使用者がある程度決まった意味で使用されているとしながらも、完全に決まりきったものとはとらえず、意味拡張などを通して同じ語が違う意味で使用される。大抵の場合、典型条件に入らない意味や、意味拡張によって新たに付与された意味は多くの語で共有されており、「使用者によって意味が異なる」という性質を必ずしも意味するものではないが、意味拡張などにより、一つの語の意味が変わる過程で共通性が低くなる例として考えうる。

多義性

語は多義性という性質を持ち、一つの語が多くの意味をもつことが報告されている。「襲う」という語に対してクラスター分析を行った研究(中本, 2004)では、「虐待」「抗争」「強盗」だけでなく、「自然災害」「不安」「疫病の流行」「活動への打撃」などの意味を持ったものであることが報告されている。このように、単一の形式の語が違った意味をもつことが報告されている。このような性質も人によって語の使用に違いはないとされている一方、意味は多くある一方で、全ての人が全ての意味を使うということは言い切れず、人によって一つの形式でも使う意味が分かれている可能性がある。

2.2.2 新規性

その一方で、我々は、今までになかった新しい形式を用いて新しい概念を言語で表現することもできる。すなわち、言語は新規性を生み出す創造的な性質を持つ。

ファッションコミュニティを対象にしたコーパス分析

Camiciottoli の研究では、ファッションブログでは、新たな意味をうまく表現するために新しい形式の語が出現していることが示されていることが分かった。(Camiciottoli, 2016)

Camiciottoli らは、ファッションブログ（Style.com）におけるレビュー文書を収集してコーパスを作成した。そして、ファッションブログ内に登場する複合語について、COCA(現代アメリカ英語コーパス)と比較して、1,000 語あたりの出現回数が大きい語の傾向をすることによって、ファッションブログで独自に使われる複合語を取得し、その語の傾向について分析した。その結果、COCA では出現が小さい語が多く発見され、ファッションコミュニティにおいて、通常使用では出現しない語が誕生していることがわかった。それらの語は、ユーモア、皮肉、挑発に関連した意味に加え、説明、評価、増幅、緩和等を目的として使用されている。

また、それらの語形成について調査を行った結果、構成語には「art-inspired」「menswear-inspired」のように、「-inspired」や「-included」のように何度も登場し

ている構成語が確認された。こういった語を含む複合語は多く、様々な品詞の複合語において多く使われる構成語が確認された。そのため、対照と比べて出現確率が大きい語でも、その構成語の多くは既存のものであったという。結論として、ファッションコミュニティでは他にない複合語が見られたこと、その語の語形成は、既存の形式の構成語の組み合わせで複合語を形成することで、今までになかった独自の語の形式を作り出していることが示された。

このように、既存の習慣化された語ではなく、今までに見ることがなかった新しい形式を用いて、新しい意味を表現する性質が実際の現象として確認されている。

2.3 複合語の語形成と意味構造に関する研究

2.3.1 複合語

2 つ以上語を組み合わせて新語を作る仕組みとして、(竝木,2009)では以下の 2 つを挙げている。

- 派生　接頭辞、接尾辞といった接辞を結合することで、より長い語を作ること。happy の例では、unhappy: un(接辞) + happy(語基)や、happiness: happy(語基) + ness(接辞)などの派生語が存在する。
- 複合　複数の単語(構成語)同士を結合して別の語を作ること。接頭辞・接尾辞ではなく、個々の構成語を結合することによって、より長い語を作成する。例: bathroom(bath + room)

派生においては、接辞が語基を修飾し、それらの意味の総和によって全体の意味を成すことが多い。その一方で、複合語には意味の非合成性が存在し、個々の構成語の意味の総和として、全体の語が成り立つわけでは無い場合が存在する。

2.3.2 複合語の意味の非合成性

複合語は、意味の非合成性を持っているとされている。ここで言う非合成性とは、以下のものである。

> ２つ以上の単語がまとまりをなすとき、全体の意味の部分が意味から非論理的に推測できない場合、そのまとまりは複合語である。

つまり、個々の構成語と複合語全体の意味の関係を比較したときに、単純に構成語の意味の総和が複合語全体の意味とはならないという性質を指す。他の語と結合したとき

に構成語の意味が消えることや、構成語同士が結合したときに、どの構成語にもないような意味が付け加わることがある。その例として、(竝木, 2009)は以下の例を挙げている。

- dark room 暗い部屋
- darkroom 暗室

1 の場合には、dark が room を単純に修飾し、「暗い部屋」の意味となっている。その一方で、darkroom では、「(写真の現像や焼き付けなどをする)暗室」の意味となる。ここで、単なる暗いだけの部屋を darkroom と呼ぶことはなく、複合語になることによって、部屋の目的や現像用の設備の有無などに関する意味が、語が結合する際に付け加わっている。

意味の消失

複合語において構成語同士の修飾・被修飾関係が成り立たない例としては、(竝木, 2009)は以下の様な例を挙げている。

- ～音痴：　方向音痴、運動音痴、数学音痴、アニメオンチ

音痴の意味は「生理的欠陥によって正しい音の発生や記憶ができないこと」(広辞苑)である一方で、1 の例は「対象が苦手でよくわからないこと」「対象のことをよくわかっていない人」といったものの意味として扱っている。音痴とともに構成語となっている「方向」や「運動」と言った語は音痴の意味である「音の発生」の意味が消失している。また、この消失は「音痴」の 1 語のみの状態では起きないため、竝木はこれを、語が結合した際に起きる現象としている。

意味の創発

また、構成語が結合して複合語になった際、創発的意味が出現する場合も存在する。このような場合、創発したとされる意味は、複合語全体には生じているが、個々の構成語のどれにもない。このような例として、Grady et al.のものが挙げられる。

Grady et al.が例示した「butcher surgeon」は「butcher」「surgeon」という語が組み合わさることによってできていて、「下手な外科医」という意味で使用される。その一方で、「butcher」(肉屋)、「surgeon」(外科医)のどちらにも「下手な」といった意味はない。このように、複数の語が組み合わさることにより、個々の語以上のものが意味が出現することが確認されている。Grady, et al.に基づき、複合語全体の意味を各語の意味を部分的に構造を継承するだけでなく、各語を並べることによって新しく生じる意味を創発的意味とする(Grady, et al., 1999)と、ここでは創発的意味が各構成語の結合によって生じると考えられる。

このような創発的意味の出現の例は身の周りのものでも存在する。今回は「森の水だより」というコカ・コーラ社が販売するミネラルウォーターの例を挙げる。「森の水だ

より」は「森」「水」「便り」といった３つの構成語からなる複合語である。「森」「水」は多くのミネラルウォーターの語で実際に使用されていることや、この２語はミネラルウォーターの給水源と、その上位カテゴリを示しているため、そもそもミネラルウォーターは山や森などの水を使用することが多いことも納得できる。その一方で、「便り」は通常飲料水やミネラルウォーターと共に使用される語ではない。これは、水との関連性が低い語を商品名に用いながらも、「(郵便のように) 送った後、どこかに長期間保存されることなく直送で、受け手（消費者）の手元に届く」といった意味から、「新鮮」という意味を想起させようとしている。ここでいう「新鮮」という意味は、「森」「水」「たより」のどの構成語の意味でもなく、これらの語が結合することによって、「森の水だより」は創発的意味を持つ構成語といえる。

2.4 LDA について

本研究では、創発的意味の検出のために、出現頻度の比較と共に、LDA(Latent dirichlet allocation)を使用した検出方法を提案する。

LDA はトピックモデルの一種として、教師無し学習の生成モデルとして提案され、文書分類やレコメンドシステムなどに使用されている。

学習時には、指定数のトピックに分割する。文書内に出現する単語の種類と各単語の出現数がディレクリ分布として入力され、文書の内容の近さを元に指定数のトピックに分割される。また、トピック分類時には、各トピックについて特徴的な語のリストを作成する。そのため、トピックモデルの中身はこれによって解釈される。

そして、推論時には、推論を行う文書についてまたディレクリ分布を作成し、推論を行う。推論では、各トピックへの分配確率を計算することができる。また、学習時に各トピックへの分配確率の閾値を計算するため、閾値の上下を見ることで、指定されたトピックのうち、推論する文書はどのトピックに分割されるかを決めることができる。

ここで、LDA の特徴として、以下の点が挙げられる。

- トピックの分割数はハイパーパラメータとして与えられる。
- 文書内に出てくる語をトピックごとに分割し、各トピックに与えられる語の出現回数を元に推論を行う。
- 推論時には複数トピックに分割される。

第3章 手法

目的でも述べたとおり、共通性を持った既存の語によって、新規性がある複合語が生成されているということを示す。そのため、複合語・構成語の双方において、意味(form)、形式(meaning)の双方の面から語の新規性・共通性の評価を行う。本章では、新語の複合語の取得方法を紹介し、前章で述べた共通性・新規性をどのように評価すべきかを説明する。

3.1 新語の複合語の取得

3.1.1 商品名の取得

分析に当たって新語の複合語を取得する必要がある。また、新規性として「新たに出現した語がコミュニケーションに使用されている」という性質を評価するため、実際のコミュニケーションに使用される語を取得する必要がある。そのため、誰かひとりが単に発音しただけの語や、コミュニケーションを目的としない語は除外される。この要素を満たす語として消費者向け商品の商品名が適切であると考えた。まず第一に、商品名は消費者に対してその商品の特徴を説明する必要がある。その商品の特徴や属性を商品名からも訴求する必要があり、商品名には生産者の意図を消費者に伝える(共通性)必要性を持っている。その一方で、その商品の独自性を出し、消費者の目に留まる必要がある。この要求によって、多くの商品名は他の商品とは違った名前をあてがわれることになり、新規性が増す。以上の点から、商品名は共通性・新規性の双方を必要とされていると考え、本研究で採用する。

3.1.2 特許情報プラットフォーム

複合語の取得に際しては、特許情報プラットフォーム J-PlatPat(https://www.j-platpat.inpit.go.jp/)内にある、商標登録された商品名を使用した。

J-PlatPat は、独立行政法人工業所有権情報・研修館によって運営される、登録された特許を見るサービスである。これにより、検索によって商標の検索し、何らかのキーワードで検索することにより、キーワードと類似した特許の一覧を表示する。また、フィルターにより、表示される特許のうち、「特許・実用新案」「意匠」「商標」の 3 種類に分けることや、特定の期間に登録された特許のみを表示することができる。

3.1.3 取得

商標検索時には、消費者向けと考えられる語として、「菓子」「加工」「ゲーム」「おも

ちゃ」「飲料」を検索し、ヒットした商標を取得した。形式の分析をする際には、今回取得する複合語の商標や、それの構成語について出現の推移を評価する必要がある。そのため、近年出現したような、出現後のデータに乏しいものや、逆に昔のもので出現前のデータに乏しいものは適さない。今回の研究では取得開始の 5 年前となる 2017 年に商標を取得したものについて分析を行った。

また、表示された後の中から、消費者向けの商品と考えられる複合語(会社名・組織名や、画像だけものなどは除く)を取得し、分析に使用した。

3.1.4 構成語に分解

取得した複合語を構成語に分解する際には、MeCab を使用して一度形態素解析を行い、分割された語を確認しながら構成語への分解を行った。

3.2 形式分析の手法

本節では形式面において分析手法を説明する。複合語の語形成について形式面の分析を行う。そして、以下の仮説の検証を試みる。

3.2.1 形式の共通性・新規性

共通性・新規性について、「今までにも十分に使用されていた性質」(共通性)「今までにはなかったものが使用されるようになった性質」(新規性)の評価を行う。共通性は、今までも習慣的に使用され、語の形式が多くの人の間で共有されていたという点が特徴としてあげられる。その一方で、新規性では、今までにはない形式であったため、ほとんど使用されていなかったものが出てきたという特徴が挙げられる。そのため、これら 2 つの特徴は、今までの使用の量では、共通性と新規性で同じ評価軸で測定できる反対の性質といえる。それに追加して、本研究では新規性を、新語が世に普及する実態としての性質として見ており、誰にも使用されていない新しい形式の語は新語とみなさない。そのため、新語登場後に出現が十分に大きく、新語登場前とは出現が違っているという性質も新規性に必要な要素といえる。

3.3 評価データ準備

語の共通性・新規性を評価するためには、取得した複合語とその構成語それぞれについて出現の程度とその推移を調べる必要がある。すなわち、語の使用時期が特定でき、長期間に渡って観察できる分析データが必要である。また、今回扱う商品名はニュース記事や公文書等に出にくいが、商品名が使用される場面のデータである事が求められる。

そのため、分析には Google Trend 人気度を使用した。

Google Trend 人気度

語の出現の計測に Google Trend 人気度を使用した。ある語の Google Trend 人気度は、その語の検索数に応じた相対的な値を[0,100]の範囲で表示する。すなわち、その語について十分な検索クエリがない場合には 0、指定した期間内における最高値は必ず 100 となるよう算出される。Google Trend 人気度は月ごとに算出される。また、Google 検索は一般的に商品の情報取得や通信販売などで利用されている点から、商品名の出現の推移を計測するのに適していると考えられる。

なお、算出対象とする全期間で十分な検索が無い場合はエラーとなるが、本研究では全期間の人気度を 0 とした。

3.4 評価

形式面の分析では、出現の推移を年スパンで見るものとし、新語の登場を商標取得年とする。また、商標取得年の前後 3 年を、それぞれ商標登録前期間・商標登録後期間として考える。今回は、調査開始時から５年前の 2017 年に商標登録された商品名の複合語を取得し、2014-2016 年を商標取得前期間、2018-2021 年を商標取得後期間として、分析に使用した。

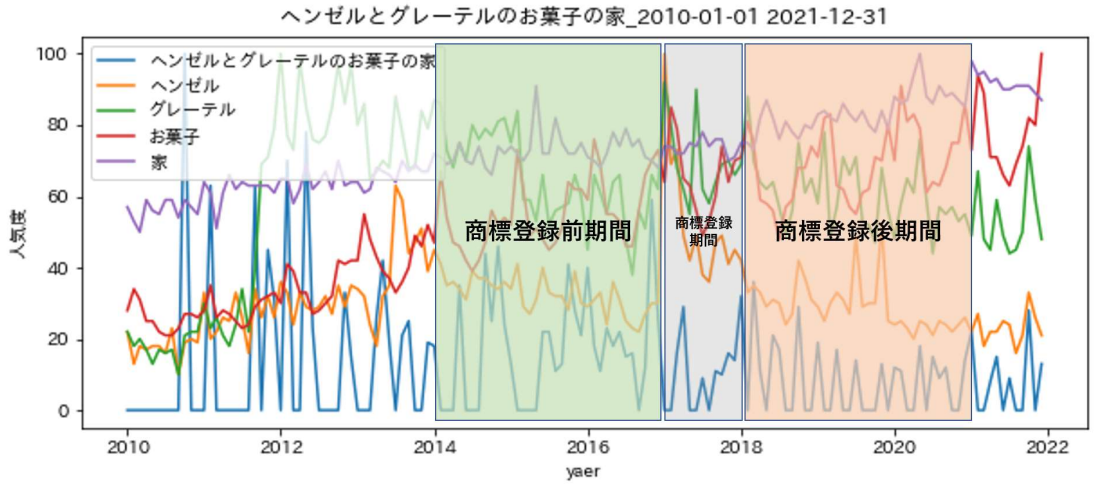

図 1 形式評価のための期間設定

3.4.1 形式の共通性

形式の共通性の評価では、対象となる語が新語登場前から習慣的に使用されていた性質を評価する。つまり、ある語が長期間に渡って使用されていることが求められ、ほと

んどの期間で使用されていなかったり、全く使用されていないようでは、その語が習慣的に使用されているとはいえない。そのため、商標登録前期間において、人気度が 0 であった月をカウントし、その月数の商標登録前期間全体における割合を計算した。

この指標では、共通性の低さを[0,1]で示し、0 に近いほど共通性が高く、1 に近いほど共通性が低いといえる。

3.4.2 形式の新規性

新規性については、①今までに使用されていなかった語が②新語登場後に普及している性質を評価する。

そのため、①の条件は共通性の低さを以て評価を行う。これは、商標登録前期間での、共通性評価で使用した人気度 0 の月の割合が高い必要がある。

また、②の条件は、新語登場後において、登場前よりも使用されている性質を評価する。これは、商標登録後期間と商標登録前期間の人気度の平均を比較して定量化を行う。また、T 検定を行い、２つの期間で統計的有意な差があるかを検証する。

3.5 意味分析の手法

本章では、意味面から複合語についての分析を行う。意味面の分析では、複合語の意味の新規性と共通性についての評価を目的とし、以下の仮説を分析する。

意味においても複合語の新語では、(共通性)意味がある程度共通して使用されている構成語を組み合わせることで、(新規性)各構成語には存在しない意味が複合語全体で創発する。

検証のために、語が持つ新規性・共通性を以下のものとして、上記の仮説立証を試みる。

3.5.1 意味の共通性

その一方で、共通性とは、ある程度決まった形で使用されており、使用者によって大きな違いがない性質を指す。これは、ある語の意味が同じものとして使用されており、同じ語を別の意味で使用されていないことを指し、一つの語が違った意味で使用されていた場合、共通性は低いと言える。共通性が高くなると考えられるものとして、「」が挙げられる。

このような形式と意味の不一致は、対象の 1 語を聞いただけで 1 つの意味に定まることが無いため、意味理解に支障をきたすものである。そのため、共通性が高く、多くの人が同じ意味で使用している語ほど正確な意味理解ができるものと考え、共通性を評価する。

3.5.2 意味の新規性

ここでいう、新規性とは、複合語全体の意味が各構成語には無い性質を指し、構成語が結合した際に生じる意味の創発性を指す。「butcher surgeon」の例(Fauconnier& Turner, 1996)のように、語が結合する際に各構成語にはない意味が出現する性質を指す。

逆に、意味の新規性が低い語の例としては、「二等辺三角形」の語が挙げられる。「二等辺三角形」は「三角形の一種で、3 本の辺のうち 2 本の辺の長さが等しい図形」(Wikipedia より)であり、複合語「二等辺三角形」の意味は構成語となる「二等辺」と「三角形」のみによって成り立つ。このような語においては、新規性は低いといえる。

以上より「ある語がどれだけ決まった意味で使用されているか」と「語が結合した際、個々の語には無い意味がどれだけ出現するか」を以て意味の共通性・新規性を評価する。

3.6 意味分析のためのコーパス作成

ある語の意味はその使用のされ方に反映されると考え、仮説演繹の推論手法を用いながら、共起語の分布によって、語の意味の分析を行う。意味分析のデータには、以下の点が求められる。

- 分析する商品名が出現する場面のデータである。
- 分析する語が書き言葉や話し言葉等の使用の中で見られ、共起語の取得が可能である。(Google Trend では単一の語の人気度を出力するため使用できない。)
- 出現時期のタイムスタンプは必ずしも出力される必要はない。

また、今回提案する LDA を使用意味分析以下の条件も追加して求められる。

- 分析する複合語と似たカテゴリに属する文書全般を取得できる。

以下の点を満たすものとして、後述の出現回数による新規性評価と、共通性の評価では Twitter のツイートログデータ、LDA を使用した新規性評価では Amazon 商品レビューが適切と考え、本研究ではこれらを使用した。

Twitter ツイートログ

共起語の出現回数分析では、Twitter の投稿ログデータを収集し、同ツイート内にある語を共起語と見なした。

Twitter(twitter.com)は Twitter 社が運営する SNS で、企業の商品紹介や選挙の宣伝などの投稿(ツイート)や、空き時間が多い人がする日常のツイート、他ユーザーに言及

ての話し合いなど多目的に使用される。ここには商品の紹介なども含まれており、その商品の感想などの共有が行われている。2006 年にサービスを開始しており、API(academic)を使用して、サービス開始から現在までの全ての期間のツイートログを取得することができる。

取得した期間は 2022 年 12 月 31 日以前のもので、直近の 10,000 件を取得した。また、全体での出現回数が 10,000 件に満たないものについては取得できる件数を全て取得した。

Amazon レビュー

また、Twitter ログデータと共に Amazon(amazon.co.jp)の商品レビューを取得した。Amazon では、日用品や、書籍、キャンプ用品など、幅広い商品の販売や、映画や音楽のストリーミングサービスを取り扱っている。商品レビューでは、商品の感想共有や、製造者や販売者の評価を目的として、ユーザーがその商品についてのコメントを投稿する。また、Amazon.co.jp には、商品がカテゴリ別に分かれており、特定のカテゴリの商品レビューの取得が可能である。

以上の点から、本研究では、Amazon の商品レビューを取得し、コーパスを作成した。取得に際しては、分析する商品名に関連したトピックとして、「菓子・スナック」と「ドラッグストア」からカテゴリ指定で取得できる 400 ページ分(約 6,400 商品)の商品のレビュー文書を全て取得し、コーパスを作成した。

データ前処理

取得した文書は Python, MeCab を使用して、形態素解析を行った。その際、助詞や助動詞などにより、意味判別ができない語が含まれるのを防ぐため、以下の語だけを取得した。

・名詞、動詞、形容詞、形容動詞、副詞

Mecab の誤認識などにより、明らかに上記のものでない品詞のものが入ることが確認された。出現数上位 100 語については、上記以外の語が入っていないことを確認し、見つけた場合には、品詞がどのように判別されていたかは関係なく除外されるよう設定した状態で再度形態素解析を試みた。また、絵文字、数値、人称代名詞についても、意味判別が難しいため削除した。

それと共に、Twitter データセットでは外部のサービスを利用した自動投稿が可能であり、同じツイートが複数回投稿されることが可能となっている。実際に全く同じ文書のツイートが幾度も投稿されることが確認されている。そのため、これらも語の使用データを歪める原因となりうるため、最初の 1 度のみ記録し、残りのものは排除した。

3.7 意味分析の評価

3.7.1 意味の共通性の評価

意味の共通性の評価に当たっては、ある語の意味がどの程度限定的な意味で使用されているかの検出を試みた。

ある語が限定的な意味で使用されていた場合、共起する共起語も限定的な意味となる。つまり、もしも「期待値」という語が確率分布の加重平均の意味のみで使用されていた場合には、共起語は統計などを中心とした数学用語や、それらを使って行う操作などに関する用語に限られる。その一方で、「期待値」という語は「仕事の最低限クリアしたいライン」という意味でも使用されているため、「期待値」の共起語として仕事に関連した語が付け加わると考えられる。このように、意味が限定されている場合には偏っていた共起語の分布が、違う意味が付け加わるにしたがって、多く出現する共起語の数自体が多くなる。

すると、共通性が高い場合においては偏っていた共起語の分布は、共通性が下がるにしたがって、偏りが小さくなると考えられる。そのため、この分布の偏りについてジニ係数での評価を試みる。ある語の共起回数でローレンツ曲線を計算すると、共通性が高いほど、共起語の分布の偏りが大きく、曲線は一様分布を前提とした完全平等線から乖離する。この乖離をジニ係数にて測定することにより、共通性を定量的に測定する。

それと共に、仮に共通性が低く、ツイートにより共起語の分布に大きな差がある場合、対象語が出現するコーパスを２つの群に分割したとき、各郡の共起語分布は異なるはずである。そのため、対象語を含むツイート全体をランダムに２分割し、各郡で共起語の分布を作成した。そして、その確率分布についてカルバックライブラー情報量(KL 情報量)を測定するとともに、KS 検定にて分布の一致度を測定した。また、ランダム分割と、指標測定を 100 回繰り返し、距離の平均と有意差の有無を観察した。

3.7.2 意味の新規性の評価

意味の新規性は、構成語には無く複合語全体にはある、創発的意味の検出を試みる。意味の検出には以下の２つの方法にて分析を試みる。

- 出現回数を利用した意味検出法
- LDA を使用した意味検出法

3.7.3 出現回数を利用した意味検出法

出現回数を利用した意味検出法では、ある単語の共起語の分布をその語の意味と見なし、複合語のみで出現する共起語を検出することにより創発的意味の検出を試みた。複合語、構成語それぞれに対して共起語を取得し、その共起語の分布をそれぞれの語の意味とする。そして、共起語の取得に際しては、Twitter(twitter.com)の投稿ログを使用し、検出したい語

複合語、共起語について、よく出現する共起語の違いを用いて意味の検出を試みる。

取得した共起語全てについて、出現回数をカウントし、上位 100 語を取得した。

3.7.4 LDA を使用した意味検出法

本研究では、出現回数による意味検出と共に、意味検出の方法として LDA を用いた手法を提案し、実際の検出を試みた。潜在ディレクリ配分法とは、トピックモデルの一種であり、文書分類やカテゴリ判別などに使用される手法である。

LDA の自分の研究への応用

LDA を本研究に応用するに、商品名の意味について分割された学習モデルを必要とする。そのため、Amazon(amazon.co.jp)の商品レビューを使用し、50 トピックに分割された学習モデルを作成した。学習に際しては、取得する商品名に合わせた Amazon 商品レビューを取得した。そして、レビュー1 件を 1 文書としてみなし、50 トピックで LDA の学習を行った。

LDA は文書のトピックを大まかに分割するための方法として確立されている。その一方で本研究では、単語 1 つの意味を分析したい。そのため、分析する 1 つの語を 1 文書として LDA モデルに推論させることで、分析したい語がどのトピックに分割されるかを観察した。

第4章 実験・評価

4.1 形式面の分析

4.1.1 共通性

形式面の分析では、語の形式について評価を行い、共通性と新規性を評価した。取得した語の GoogleTrend の推移について実例を紹介する。

図 2「おいしいお菓子の御神木」という語の例の人気度の推移を表している。このグラフでは、それぞれの語に対して人気度を算出しているため、ある 1 語の推移は分析できるが、同じグラフに掲載されている語同士の比較を行う事はできない。

ここから読み取れることとして、全体を通して構成語は人気度 0 となっており、商標登録の前後に関わらず、語が使用されていないことが伺える。その一方で、この語を構成する構成語「おいしい」「お菓子」「神木」は、全ての期間で 0 より大きい値となっている。「おいしい」の人気度は緩やかに減少すると同時に、「お菓子」の人気度は周期的な増加を伴いながら緩やかに上昇している。神木は全体を通した人気度の差はあまりないが、2015 年近辺で人気度が大きく上がっている。

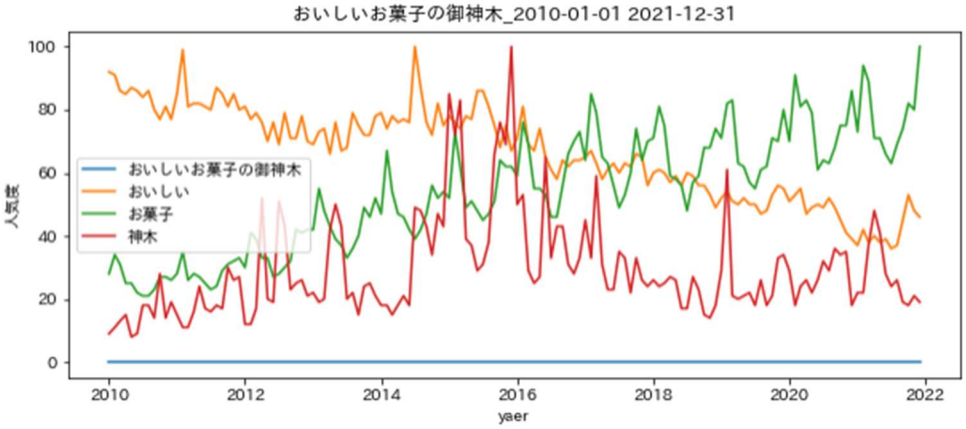

図 2 「おいしいお菓子のご神木」形式出現の推移

これらを集計した結果、縦軸を語数、横軸を人気度 0 だった月の割合として、ヒストグラムを作成した(図 3)。その結果、複合語全体としては、人気度 0 の月の割合が比較的高く、多くの複合語はほとんど使用されていない語であったことが言える。その一方で構成語は多くの語について、共通性が認められた。105 語中、人気度 0 であった月が全体の 1 割以下となり、それらの語はほとんどの月(9 割以上)である程度の人気度があ

ったということになる。

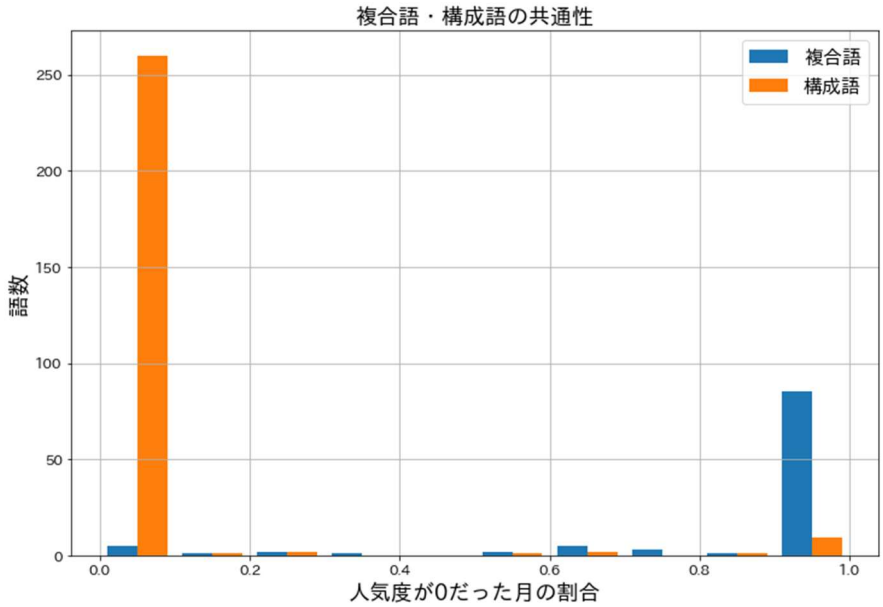

図 3 人気度が 0 だった月の割合

4.1.2 新規性

新規性の評価に際しては、商標登録前期間における人気度 0 の月の割合と、商標登録前後における増減を調査した。結果として、商標取得前期間においては、多くの語が人気度 0 だった月の割合が多かった。また、商標登録前後における人気度の差を T 検定を用いて調査したするとともに、登録前後の人気度の増減を観察した。その結果、105 語中、8 語が有意差ありという結果になり、そのうちの 4 語商標登録前後で人気度が増加していることがわかった。

表 1 商標登録前後での有意差(T 検定, $p<0.05$)の有無と増減傾向

		複合語(105 語)	構成語(273 語)
有意差あり		8	197
	増加	4	121
	減少	4	76
有意差なし		94	76

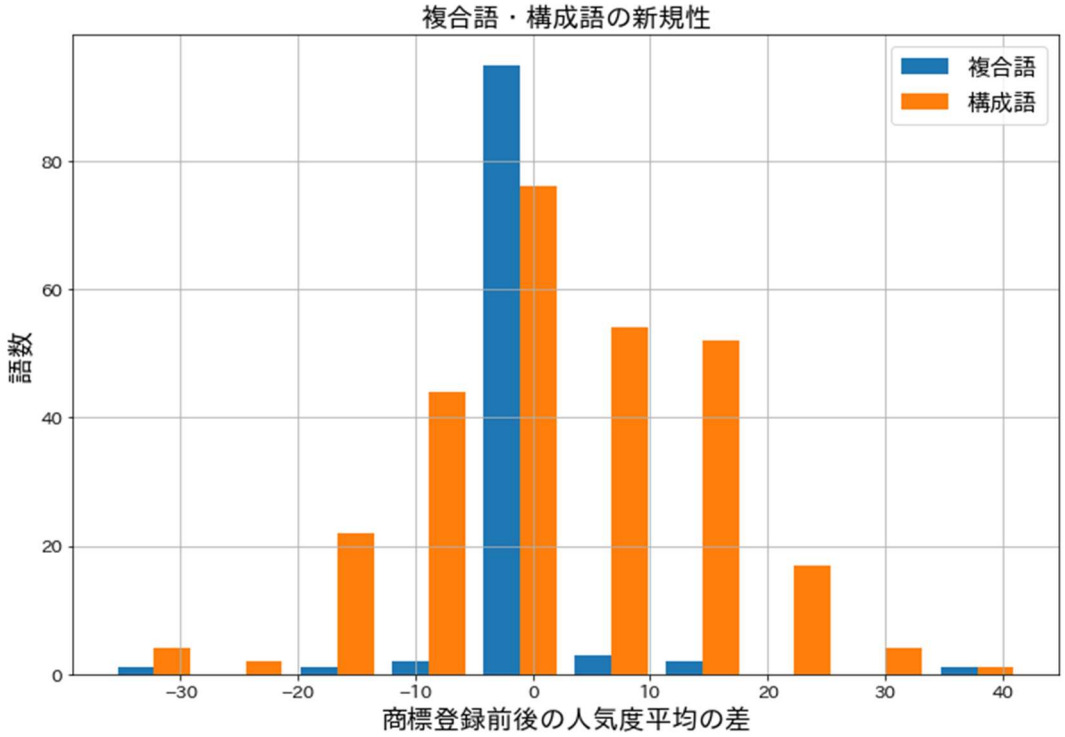

図 4 商標登録前後の人気度の平均の差

複合語で有意な増加が確認できた語は、「ワールドマスターズゲームズ」「インサイダー・ゲーム」「塩引鮭」「ナインタイル」の４語であった。（表 1）

また、有意差が商標取得前後の人気度について、有意差が出た語のヒストグラムは以下の通りになった。

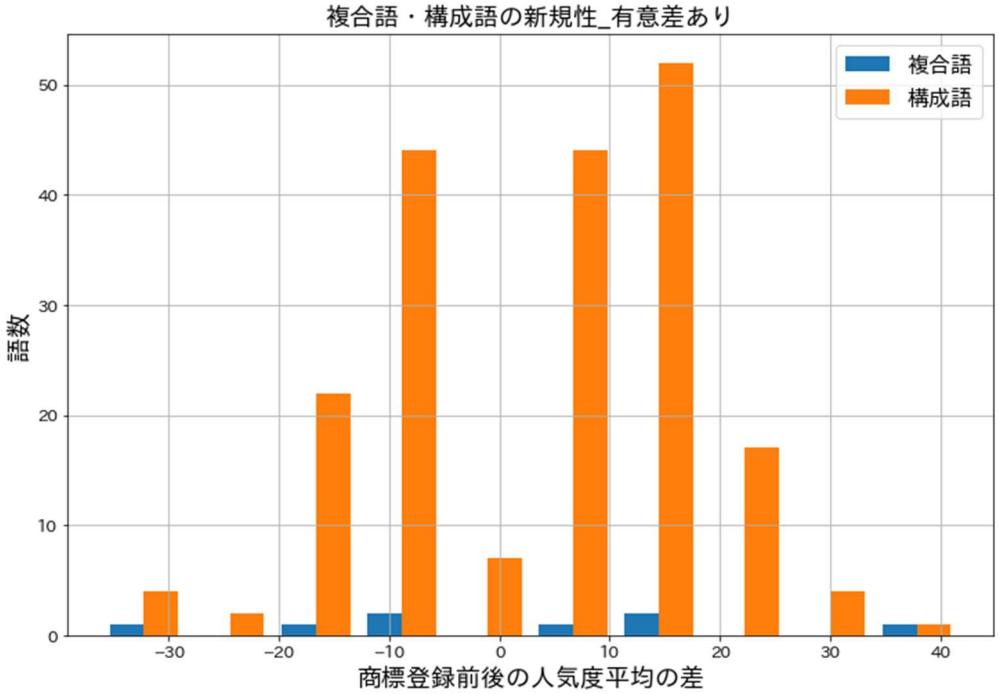

図 5 商標登録前後の人気度の平均の差(有意差あり)

4.2 意味面での分析

意味面での分析では、Twitter データによる共通性・新規性の評価に加え、Amazo レビューを利用した LDA による意味分析の手法を提案する。Twitter を使用した分析では、用意した複合語 105 語のうち、39 語においては、複合語のツイートが取得できなかった。そのため、複合語全体のツイートを取得できた 66 語(構成語:156 語)のみで分析を行う。

4.2.1 共通性

意味の共通性では、ジニ係数による比較と、語をランダムに２集団に分けた時の集団間の分布の違いについて計測した。

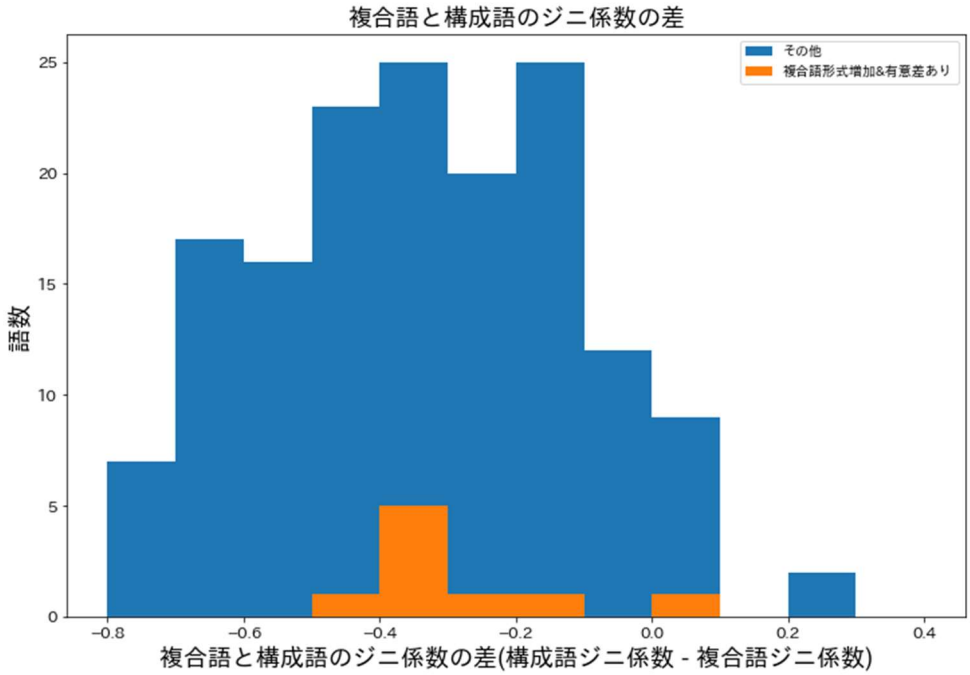

図 6 構成語と複合語のジニ係数の差の分布

全ての語について、ジニ係数を測定し、構成語と複合語の差(構成語のジニ係数-複合語のジニ係数)を求めた。その結果、図 1 では、多くの語について、差が 0 未満であることがいえ、複合語よりも構成語の方が共通性が高いことがいえる。この結果は、共通性が比較的高い構成語によって、共通性が低い複合語が作られているという結果を裏付けるものとなった。

また、形式面の分析で有意に増加していた語については、複合語と共起語に対して KL 情報量の測定を行った。これは、3 章で説明した方法を計 100 回繰り返し、KL 情報量の平均を比較した。

表 2 各語における KL 情報量と、KS 検定の結果

	KL 情報量平均	KL 情報量標準偏差	KS 検定 p 値平均	KS 検定 p<=0.05 の割合
ワールドマスターズゲームズ	55.25544487	10.65080079	0.427564675	0.38
ワールド	17.84162737	0.154435462	0.392413532	0.3
マスターズ	13.76541478	0.154348037	0.524709396	0.24
ゲーム	13.48095016	0.115059309	0.518138798	0.19

インサイダー・ゲーム	37.30449654	1.104555135	0.384995734	0.41
インサイダー	16.96571489	0.214165981	0.403835488	0.31
ゲーム	13.48095016	0.115059309	0.518138798	0.19
塩引鮭	45.09843079	2.53596323	0.352439898	0.41
塩引	47.73204681	2.568203974	0.315636531	0.51
鮭	10.79776977	0.089101611	0.565950096	0.14
ナインタイル	47.05497379	2.291063907	0.431976122	0.34
ナイン	14.1435727	0.121776934	0.481992292	0.23
タイル	11.59285389	0.101208616	0.511922655	0.18

表 4 より結果として、「塩引鮭」以外の 3 語は構成語は全て複合語よりも KL 情報量の平均値が小さい。そのため、これらの 3 語においては、複合語よりも共通性を持った構成語によって複合語全体が成り立ったいることがわかる。また、「塩引鮭」においては、「塩引」という語は共通性が低く、他の語では共通性が高いことが分かった。

4.2.2 新規性

出現回数の比較

出現回数の分析では、それぞれの複合語・構成語に対して最も多く出現した共起語 100 語を各々に対して選出した。そして、構成語には出てきていないが、複合語のみで出現する語を選出した。ある複合語が n 語の構成語から成り立つ場合、複合語の共起語 100 語の集合を C、構成語の共起語 100 語の集合を W_1、W_2 とすると、要素数を求める集合は以下のとおりになる。

$$C \cap \overline{W_1 \cup W_2 \cup W_3 \cup \ldots \cup W_n}$$

集合を数えた結果、以下の通りになった。全体として 、複合語のみに出現する共起語は 70 語近く持つことが多いことがわかる。また、形式の新規性の分析時、複合語の出現が有意に増加していた 4 語は橙とし、積み上げヒストグラムを作成した。

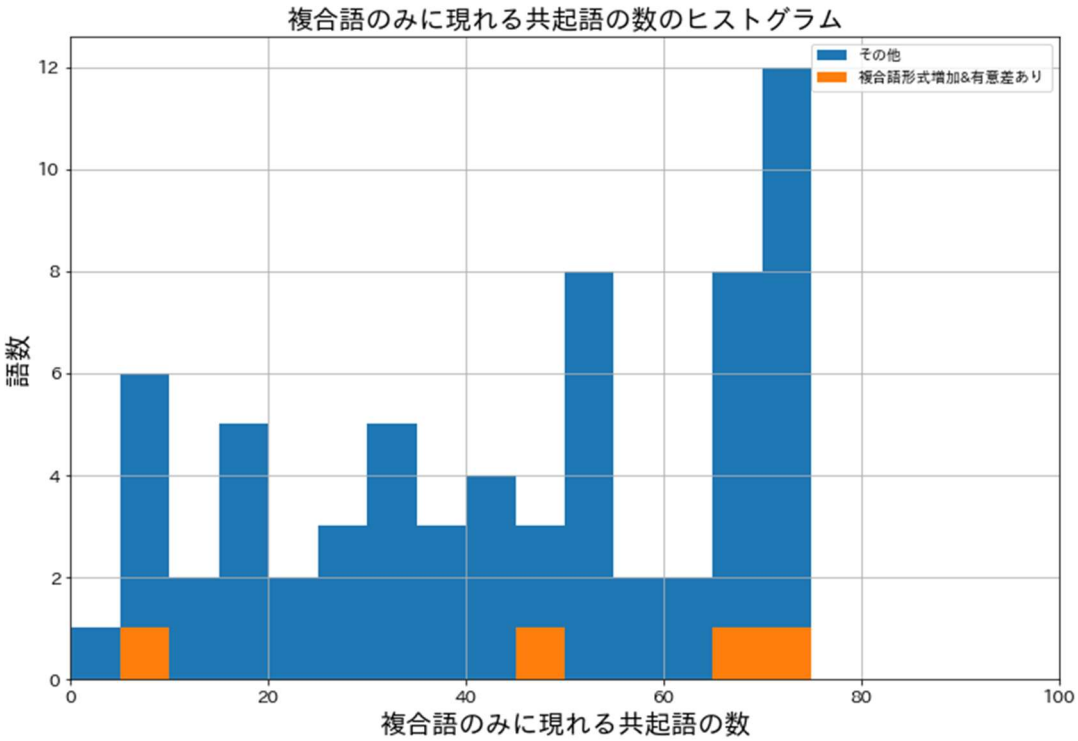

図 7 複合語のみに現れる共起語数の分布

また、形式面の分析で有意に増加していた語は、複合語のみに出現する共起語として、以下のようなものが見られた。

表 3 複合語のみに出現する共起語

	複合語のみの共起語数	複合語のみに出現する共起語
ワールドマスターゲームズ	71 語	和歌山, 委員, 県, 体験, バスケ, 関西, 車いす, 坂井, 素敵, わく, 公益, 財団, 教委, 茶室, 法人, 黄金, 組織, 新報, バレー, 福井, 実行, 市, 喜び, 決まる, 応募, 主催, ポチ, 制作, 再現, こんばんは, 企画, 動く, ｓｄｇｓ, 止まる, 未完, 頃, 厄, 天井, 府, イベント, ワクワク, アート, 身体, , でる, コロナ, 丸岡, ごみ, 地域, 共有, 今月, なん, 児童, キャンペーン, 当選, まつり, 城, ニュース, 福袋, 予定, 京都, 延期, スポーツ, 気づく, 誕生, 桜, 新年, ローカルニュース, 家族, プレゼント, 大当たり
インサイダー・ゲー	45 語	インサイダー, 回, ボドゲ, 題, ボード, よい, 質

ム		問,業界,会話,ｙｏｕｔｕｂｅ,ｉｔｏ,最高,動画,誰,操作,ルール,募集,くん,ｊｇｒ,ウミガメ,正解,遊べる,スクエニ,写真,後輩,友人,入れる,決断,あと,ワン,なぜ,メンバー,昨日,金,ゲームインサイダーゲーム,重岡,無い,ニムト,踊る,焼肉,ジャスト,新た,見つけ出す,役割,バンナム
塩引鮭	3 語	市,お供,バカ
ナインタイル	67 語	ポケモン,ボード,ａｍａｚｏｎ,絵,ルール,題,ボドゲ,面白い,欲しい,定番,並べる,ムーミン,とき,たった,大人,秒,版,たち,通り,会,参加,説明,者,木製,選ぶ,かわいい,歳,みつ,描く,サンリオ,入荷,出来る,難しい,なぜ,ハマる,動き,知育,背景,娘,ボードゲームカフェ,教える,僕,今度,子供,遊ぶ,裏,富豪,どこ,子ども,縛る,フシギ,遊べる,ゲー,勝負,スピードパズルボードゲームナインタイルポケモンドコダ,ナインタイルムーミン,ボードゲームナインタイル,持つ,さく,飲む,こども,単純,ぴき,速い,ネット

4.2.3 LDA を使用した意味検出法

LDA による意味検出法は、形式の新規性の分析で、語の出現が有意に増加した４語について、創発的意味の検出を試みた。４語のうち、「ワールドマスターゲームス」「インサイダー・ゲーム」「ナインタイル」については、Amazon の「ゲームソフト」カテゴリ(https://www.amazon.co.jp/s?rh=n%3A5121199051&fs=true&ref=lp_5121199051_sar)、「塩引鮭」の分析については、Amazon の「菓子・スナック」カテゴリ(https://www.amazon.co.jp/s?rh=n%3A71314051&fs=true&ref=lp_71314051_sar)より、一覧として表示される 400 ページに掲載されている商品の全ての商品レビューを取得して、カテゴリごとにコーパスを作成した。

また、作成したコーパスを LDA で学習させ、以下の文書の配分トピックを観察した。

表 4 「ワールドマスターゲームス」とその構成語に配分されたトピック

語	配分トピック
ワールドマスターゲームス	4, 28, 45, 52, 74, 80

ワールド	4，28，45，52，74，80
マスター	0，1，2，4，5，7，9，10，13，15，16，18，22，23，25，26，27，28，30，36，37，38，41，43，46，47，50，52，55，56，74，77，79，80，82，88，95，99
ゲーム	0，1，2，4，5，7，9，10，13，15，16，18，22，23，25，26，27，28，30，36，37，38，41，43，46，47，50，52，55，56，74，77，79，80，82，88，95，99

表 5 「インサイダー・ゲーム」とその構成語に配分されたトピック

語	配分トピック
インサイダー・ゲーム	4，28，45，52，74，80
インサイダー	4，28，45，52，74，80
ゲーム	0，1，2，4，5，7，9，10，13，15，16，18，22，23，25，26，27，28，30，36，37，38，41，43，46，47，50，52，55，56，74，77，79，80，82，88，95，99

表 6 「ナインタイル」とその構成語に配分されたトピック

語	配分トピック
ナインタイル	4，28，45，52，74，80
ナイン	4，28，45，52，74，80
タイル	0，1，2，4，5，7，9，10，13，15，16，18，22，23，25，26，27，28，30，36，37，38，41，43，46，47，50，52，55，56，74，77，79，80，82，88，95，99

表 7 「塩引鮭」とその構成語に配分されたトピック

語	配分トピック
塩引鮭	4，28，45，52，74，80
塩引	4，28，45，52，74，80
鮭	0，1，2，4，5，7，9，10，13，15，16，18，22，23，25，26，27，28，30，36，37，38，41，43，46，47，50，52，55，56，74，77，79，80，82，88，95，99

第5章　議論

5.1 まとめ

本研究では新規な商品名の複合語を取得し、その形式・意味に対して新規性・共通性を Google Trend 人気度の推移により分析した。その結果として、形式においては、構成語が新商品が現れる前から使われている共通性、複合語は新商品登場後に使われるようになる新規性があることが確認された。しかし多くの新規複合語商品名は調査を行った全期間で人気度が 0 であった。

意味の共通性に関しては、共起語分布の偏りによるものと、コーパスを 2 分して差を計算するものの２つの方法で分析した。その結果、分析対象とした語の多くでは一定の見られ、商標登録をされる複合語は共通性がある語によって複合語が構成されていることがわかった。

意味の新規性については、語が結合することで上位カテゴリ等以外の語が出現することが確認でき、構成語には無いイメージを複合語が持っている例があった。LDA による意味検出では、複合語のみに分配されたトピックを確認できなかった。

5.2　今後の課題

5.2.1 新語の定義

本研究では、形式の新規性を普及する実態としての性質として見ていた。そのため、実際に普及しなかった語に関しては新規性が無いものとして取り扱った。その一方で、商標登録までされるという事は、その商品名を新しい語として生み出し、商品化企画などにおいて何らかのコミュニケーションをしていたと考えられる。そのため、「実際にコミュニケーションで使用されたもの」としての新語ではなく、「今までにない語が商標取得されている」という性質をもって新規性とした場合、新規性を持った語が多く見られることになる。本研究では評価の際には前者、複合語取得の際には後者の基準を元に研究をしていたため、どちらかに統一することが今後の課題として求められる。

5.2.2 実際に普及した新語の取得

形式分析で新規性が評価できなかった理由として、ヒット商品の商品名だけを集めることの困難さが挙げられる。ある商標が一般消費者に広く使われるためには商標取得後に商品として販売し、その商品が広く普及することが必要になる。多くの商品は販売・

普及の前に開発・販売が中止になることがあり、販売された中でヒットする商品はごく一部となる。今回の分析では、普及していない商品名も取得したため、新語として認められたものが４語だけになったと考えられる。そのため、実際にヒットした商品名を取得して分析を行うことで、多くの新語について分析ができる事が可能となると考えられる。

5.2.3 Google Trend の問題

今回使用した GoogleTrend 人気度は、取得した期間における最大値を 100 とした相対的な値である。そのため、語の出現について絶対的な値を出していないため、出現の有無を言いにくい数値となっている。

また、粗頻度や数値の算出基準が明白でないため、語同士の単純比較ができない点は、この指標が抱える最も大きな問題点といえる。語の不在証明自体の信頼度が低いこと自体はコーパス言語学全体で抱える課題といえる(石川,2012)が、粗頻度での比較ができない点においては、Google Trend 特有の問題といえる。これには、商品名などのマイナーな語の出現の推移を特定できるコーパスの構築が必要となる。

5.2.4 均衡性の問題

今回使用したデータ(Google Trend は不明だが、少なくとも Twitter、Amazon レビュー)の全てを(石川,2012)の基準でコーパスと見なす際には、均衡性の問題が発生する。つまり、「日本語の話し言葉・書き言葉全体」や、Amazon レビューで取得しようとしたような「特定の話題に関する話し言葉・書き言葉全体」を母集団とした際、用意したコーパスは母集団全体を全て反映するとは限らない。

このことにより、通販や SNS などをしない人々が話している内容や感想などを取得できないまま分析しており、結果に対しても影響する可能性がある。

5.2.5 LDA の問題

LDA が抱える限界として、限られた意味のまとまりのものを多数のトピックに分解するとき、うまく分割できない点が挙げられる。

また、用意したデータセットは、Amazon の特定のページにて感想などを書くものとなっている。つまり、商品名の形式を出すことなく、その商品にかんする話題を話すことができる性質を持っている。このことから、探そうとしている商品名がレビュー文書内に出現することが少なく、うまく判別できない原因になったと考えられる。

第6章 結論

6.1 本論文まとめ

まず、第 1 章では、本研究の目的と、背景となる共通性・新規性の特徴を大まかに紹介した。

2 章においては、背景となる共通性・新規性についてされている研究について紹介した。ここでは、語の多義性などにより、一つの語について多くの意味が存在している点などを紹介した。それと共に、一つの語が多くの意味を持つということには、多くの人がたくさんの意味を使用しているわけでなく、限られた意味で使用されるが、人によって使用する意味が異なるパターンについて紹介した。新規性については、今までになかった新たな語が生まれることや、それらの新語にもよく使用される語があることを紹介した。そのうえで、複合語においては、語が結合する際に個々の語には無い意味(創発的意味)が付与される事例を紹介した。

3 章では、実際に共通性・新規性を測定する方法について紹介した。形式の共通性・新規性では、共通性の「習慣的に使用され、その語が多くの人に使用されている」と、新規性の「今までにはなかった新しい形式の語が使用されるようになっている」という性質について評価を試みる。これについて Google Trend 人気度を使用した分析法について紹介した。

意味の共通性については、ある語がどの人にも同じ意味で使用される性質を評価した。これのためには、取得したコーパスを２分割し、その 2 群において共起語の分布に差がないかを検証した。

意味の新規性については、出現頻度の分析と、LDA による検出方法の 2 つの評価を試みた。出現頻度の分析では、複合語の共起語には出てくるが、どの構成語の共起語にも出てこない語の意味について分析した。

4 章では 3 章で紹介した手法の検証結果について紹介した。形式においては、共通性は見られたものの、新規性が確認できた語は少なく、認められたのは 105 語中 4 語だった。意味においては、形式の新規性が認められた 4 語について、分析を行った。その結果、意味面においても多くの語で共通性が見られた。新規性は、出現頻度の分析では、複合語の上位 100 共起語のうち、7 割近くが複合語のみに現れる語であり、複合語特有の共起語を見ることができた。その一方で、LDA の分析では、複合語のみに分割されるトピックは、見られなかった。

5 章では結果についての議論を行った。結果として、多くの構成語では共通性が認められ、今回調査した複合語では、広く知れ渡り、ある程度限られた意味で使用されてい

る語によって構成されていることがわかった。また、形式の新規性がある語については、出現頻度分析では創発的意味が検出された。ここから、それらの複合語の新語は共通性を持った語が結合し、創発的意味が付与されることによって、意味の新規性を保っているということが分かった。

6.2 結論

本研究では、新語のコミュニケーションにおける共通性と新規性の両立を可能とするメカニズムについての分析を行った。その結果、形式面で新語といえる複合語の語形成は、共通性がある構成語が結合し、どの構成語にもない創発的意味(意味面の新規性)を付与されることによって、語の新規性が確保されることを解明した。また、この共通性は、多くの期間にわたって習慣的に使用されている事(形式面の共通性)だけでなく、多くの人が同じような意味として使用していること(意味面の共通性)も含んでいることが分かった。

第7章 謝辞

本研究を進めるにあたり、ご指導いただいた橋本敬先生、副指導として本田弘之先生をはじめ、ご支援いただいた多くの方々に感謝申し上げる。

特に橋本敬先生には研究に対して多くのご指導いただき、深く感謝の意を申し上げる。入学当初、言語学についてあまり知識が無かった状態から丁寧にご指導いただいたことに加え、突飛なアイデアなどについての議論にも真摯にお付き合いいただき、研究活動を楽しむ上での必要不可欠な存在だった。また、研究以外においても、数々の奇行に目を瞑っていただき、楽しい学生生活を送る上での基盤となった。

本田先生には、副指導として、言語学者の立場から構築コーパスを検討するうえでご指導いただいた。また、水高将吾先生には、副テーマの指導教員としてご指導いただいた。これは、Twitter API の取得方法に関するアドバイスに加え、検定手法等について学習する機会となり、研究への貢献となった。ダム ヒョウ チ先生には、LDA に関するコメントをいただいた。コーパスの構築法などについてアドバイスいただき、非常に有益なものだった。

本研究室のメンバーにも大きく貢献いただいた。ゼミなどを通して議論に付き合っていただいたことに加え、休憩スペースなどで付き合っていただいたカジュアルな議論が研究を楽しく進める上での大きな励みになり、日頃からの何気ない会話が精神的には大きな助けになった。

黒川瞬先生、山本寛樹さんには、結果の分析や、検定・モデリングについてアドバイスいただいた。また、人生の楽しみ方を教えていただき、研究の息抜きを考える上でも多大なる影響を受けた。

博士後期課程の先輩である藤原正幸さんには、議論にお付き合いいただいたことに加え、日頃から研究室で声をかけていただき、研究面でも精神面でも大きな励みになった。研究テーマが似ているとは決して言い難い状況だったにも関わらず、どんな相談にも気さくに乗っていただき、非常に大きな助けになった。また、分析用のパソコンを貸与いただき、円滑な研究を行う上で貢献いただいた。

それと共に、石森宥佑君、笠野純基君をはじめとした多くの研究室メンバーに、研究生活において伴走してもらい、一緒に進めていくうえで多くの励ましをいただいた。研究を進める上で大きな助けになったとともに、大きな貢献だったといえる。また、博士後期課程への進学を予定する笠野くんに関しては、溢れんばかりのユーモアを研究に落とし込み、今後の研究生活をより一層充実させる事を祈り、エールに添えてこの 2 年間の感謝を申し上げる。

また、星宏侑さんにも研究についての議論に付き合っていただき、多くのアドバイスをいただいた。研究をうまくまとめるためのアドバイスなど、行き詰まっているときに的確なアドバイスをいただき、大きな助けになった。

実家にいた時に一緒に遊んでくれた烏骨鶏、野良猫の皆様に至っては、精神的な安定と論文執筆のエネルギーいただき、非常にためになった。とりわけ、野良猫のオカモトシンスケ、ニカイカンジチョウ、スガカンボウチョウカンより、研究への態度について多くの影響を受けた。また、研究室内の人間関係については、烏骨鶏から多くの影響を受けた。

最後に、両親をはじめ家族の皆様には多くのご支援をいただいた。母に連れて行ってもらった寿司屋では、研究で張り詰めていた緊張が、口の中に入ったシャリのように崩れ、大切なリフレッシュとなった。父が送ってくれたコカ・コーラゼロは生活を送る上でのライフラインであり、大切な支えとなった。

兄の広幸については、高い技術力・理解力を持ち合わせており、追うべき先輩として有難い存在であった。それに加え、そういったものは、常日頃からなる、他とは一線を画した量の努力に支えられているということを示してくれた。これら２つの事は、私自身が集中して研究を進める上で非常に大きな励みになるとともに、困難を乗り越える勇気をもらった。いつも遠くで温かく見守り、励ましてくれた両親に加えて感謝申し上げる。

参考文献

Blei, D., Ng, A., & Jordan, M. (2001). Latent dirichlet allocation. Advances in neural information processing systems, 14.

Brown, K. (2005). Encyclopedia of language and linguistics (Vol. 1). Elsevier.

Camiciottoli, B. C. (2016). “All those Elvis-meets-golf-player looks”: A corpus-assisted analysis of creative compounds in fashion blogging. Discourse, Context & Media, 12, 77-86.

Fauconnier, Gilles, Turner, Mark B. Blending as a Central Process of Grammar, In: A. Goldberg, (Ed.) Conceptual Structure, Discourse, and Language. Stanford: Center for the Study of Language and Information, pp. 113-130, 1996.

Grady, Joseph E., Oakley, Todd, Coulson, Seana. Blending and metaphor, In: G. Steen & R. Gibbs (Eds.). Metaphor in Cognitive Linguistics, Philadelphia: John Benjamins, 1999.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. In 7th international corpus linguistics conference CL (pp. 125-127). Lancaster University.

Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. Computational linguistics, 29(3), 333-347.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.

石川慎一郎. (2012). ベーシックコーパス言語学. Kabushiki Kaisha Hitsuji Shobō.

ISO 690

石川慎一郎. 共起ネットワーク分析をふまえた類義語定義の精緻化.

岩波書店. (2018). 広辞苑. 第７版. 「期待値」

中本敬子, 野澤元, & 黒田航. (2004). 動詞 “襲う” の多義性. 日本認知心理学会発表論文集, 019-019.

野田大志. (2011). 現代日本語における複合語の意味形成: 構文理論によるアプローチ (Doctoral dissertation, 名古屋大学).

松本曜 (Ed.). (2003). 認知意味論 (Vol. 3). 大修館書店.

吉村公宏. (1995). 認知意味論の方法: 経験と動機の言語学. 人文書院.

竝木崇康. (2009). 単語の構造の秘密−日英語の造語法を探る−. 開拓社.

ウェブページ

- Wikipedia「二等辺三角形」，

  https://ja.wikipedia.org/wiki/%E4%BA%8C%E7%AD%89%E8%BE%BA%E4%B8%89%E8%A7%92%E5%BD%A2 2023 年 1 月 30 日アクセス

- Wikipedia「進化」

  https://ja.wikipedia.org/wiki/%E4%BA%8C%E7%AD%89%E8%BE%BA%E4%B8%89%E8%A7%92%E5%BD%A2 2023 年 1 月 30 日アクセス

- 特許情報プラットフォーム， https://www.j-platpat.inpit.go.jp 2023 年 1 月 30 日アクセス
- ライフコーディネート学校「仕事の期待値とは？コントロールし、上回る為の考え方を紹介」，https://lifecoordinate.com/skill/1913/ 2023 年 1 月 30 日アクセス
- ポケモンWiki「進化」，https://wiki.xn--rckteqa2e.com/wiki/%E9%80%B2%E5%8C%96 2023 年 1 月 30 日アクセス