様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　５　月　２３　日現在

機関番号：　６２６１８
研究種目：　若手研究(B)
研究期間：　2014～2017
課題番号：　２６７７０１５６

研究課題名（和文）コーパスから取得しやすい情報と取得しにくい情報の研究

研究課題名（英文）The obtainability of the information from text corpora and its cognitive analytics

研究代表者

加藤　祥（保田祥）（KATO, SACHI）

大学共同利用機関法人人間文化研究機構国立国語研究所・コーパス開発センター・プロジェクト非常勤研究員

研究者番号：４０６２３００４

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本課題研究は、既存の言語資源から収集したテキストがどのような種類の内容であるのか、テキストからどのような情報が取得できるのかを調査した。また、読み手がテキスト内容をどのように認識するのかという実験によって、テキストと情報の関係を考察した。辞書の語釈文、知識のない人への説明文、様々な文章の適切な要約文などを作成するにあたり、情報伝達に有用な内容やそれらの提示順序を提案した。

研究成果の概要（英文）：In this study, we investigated what kinds of contents are collected from existing language resources; corpora and clarified what kind of information can be obtained from texts. We examined the relationship between text and information by experiments on how readers recognized the contents from text. We proposed the useful information and their practical presentation order for glossaries, explanation for people without knowledge, and appropriate summary of various documents.

研究分野：　認知言語学

キーワード：　対象物認知　コーパスから取得可能な情報　テキスト情報　情報提示順

１．研究開始当初の背景

研究開始当初、国立国語研究所において日本語コーパスの整備が進んでおり、現代日本語書き言葉均衡コーパス（以降 BCCWJ）が公開され（2011）、国語研日本語 Web コーパスの開発が進行中であった。

コーパスの活用法の一つに、コーパスを用いた辞書作りがある。海外では、Fillmore & Atkins（1994）によって、辞書の語義記述で解釈のできない用例が、コーパスに見つかると示されたのをはじめ、Sinclair（1991 など）らが、既存辞書が言語学習者の助けにならないとし、用法（共起語や構文など）を重視した COBUILD（1987～）を作っている。国内でも、三省堂がウィズダム英和（2003～）、同和英（2007～）を刊行している。現実的な用例を収集する目的のみならず、コーパスを活用した客観的な語義記述も期待されるといえる。しかし、コーパスは万能ではない。テキストの記述と現実的な対象物が一致しているのではなく、テキストに記述された情報から、確実にテキストの示す対象物が認識されるのでもないからである。申請者は、長期的にテキストから得られる情報の可能性と限界について調査を行っている。人はテキストから何を読み取り、何が書いてあれば情報の伝達が可能なのか。どのように書いてあれば、読み手はテキスト内容を認識可能なのか。これらを整理し、テキストから得られる情報を明らかにすることが、所属プロジェクトで開発を進めているコーパスへの情報付加や、今後の新たなコーパスの活用方法を考えるためにも肝要であると考えた。

＜引用文献＞

①Fillmore, C. J.、B. T. S. Atkins 、“Starting where the dictionaries stop: The challenge for computational lexicography.” In B. T. S. Atkins and A. Zampolli, eds., Computational Approaches to the Lexicon、1994、pp.349―393. Oxford: Oxford University Press.

②Sinclair, J.、 Corpus, concordance, collocation.、1994、Oxford: Oxford University Press.

２．研究の目的

(1)テキストから対象物を認識するために必要な情報を探る

テキストに記述された情報から、テキストの示す対象物は認識されにくい。しかし、反対に、動物事典の詳細なウサギの外観説明文（「ウサギ」を伏字にする）から描画再現を行っても、ウサギは描画されない現象もある。対象物を認識するために、テキストとして提供される情報は過剰であってもならない。そこで本研究は、テキストの示す対象物を認識するために必要な情報の過不足を調査し、対象物を認識可能なテキストの生成を目指した。本課題期間では、コーパスから取得できる対象物に関する情報について、その要素を明らかにするとともに、対象物の認識に必要な情報として十分であるのかを被験者実験によって検証し、語義記述のために必要な「要素」という観点で整理を試みた。あわせて、提供情報の伝達に効果的な提示順も明らかにし、情報の「提示順」を確かめることとした。

(2)テキストから取得可能な対象物の情報とその利用可能性を考える

テキストには記述されやすいことと記述されにくいことがある。テキスト情報と現実（対象物）との差異は、テキストから得やすい情報と得にくい情報があるということでもある。本研究は、テキストから取得可能な情報と取得しにくい情報を整理する。また、最大限に取得可能な情報について考察するため、表現に着目した調査も行うこととした。

３．研究の方法

(1)コーパスから収集可能な情報の調査とその整理、対象物認識に有用な情報の検証

読み手がテキストから対象物を認識するという観点で、コーパスから頻度情報を含めたどれだけの要素が取得可能であるのかを調査し、さらに取得可能であった要素の有用性を確かめる。コーパスから対象物に関する情報を収集するほか、頻度情報を取得した。収集データの整理を行い、提示したテキストから、あるいは提示した共起語の頻度情報から対象物が認識可能か、という被験者実験を行った。これらの調査の結果として、対象物認識に有用な情報がどのようなものか考察する。また、対象物の説明に用いられやすい比喩表現に着目した調査も試みた。

(2)読み手がテキストから対象物を認識するために最適な情報提供順序

(1)の成果として得られた情報を分類し、どのように提示することが有用なのか、あるいは提示順が対象物の認識にどのように影響を及ぼすのか、被験者実験によって調査した。(1)の結果とあわせ、対象物認識に有用な情報提示順序を提案する。

４．研究成果

(1)コーパスから収集可能な要素

BCCWJ と Google 日本語 n-gram から対象物群に関する要素や用例を収集し、手作業で意味的な要素に整理した。経験知識を喚起する情報（一般的な評価や俗説、商品情報、一般的経験）がコーパスから得られる場合が多い。また、慣用化した表現や比喩として用いられていると読み取れる用例もコーパスから取得しやすい。一般的に特徴的と考えられる性質や、経験知識源となり得る情報は、読み手から対象物認識にあたっても有用性が高いと判断された。しかし、コーパスから取得した情報からの対象物認識実験結果では、対象物認識が可能な動物と難しい動物に大別で

きた。テキスト情報は個別的あるいは専門的すぎることが多く、対象物そのものの情報が得られても、想定されるカテゴリにおける他メンバーとの差別化をする情報は得にくいためであるとわかった。また、大規模コーパスを用いることで、頻度情報を得ることが可能となり、頻度情報の有用性が推察された。

(2)テキストと頻度情報の関係

人間はテキストを記述する時、現実の世界をそのまま記述するのではなく、ある面を強調するかある面を無視する傾向にある。たとえば、テキストで出現率の高い部位は、目、手、顔で、それぞれ人の身体部位の用例の20%を超える。そのため、コーパス（Google日本語 n-gram）の部位頻度情報から再構築された人体は、手と顔と目 の割合が突出し、腿や腹、尻といった部位のテキスト出現率は1%未満（同上）であるため、極端に縮小された図（図１）となる。これらの頻度差は、現実世界とテキスト世界の間には３種類のずれが生じているためと考えられ、それぞれ「兎の耳現象 (Rabbit-ear bias)」、「兎のヒゲ現象 (Rabbit-whisker bias)」、「兎の角現象 (Rabbit-horn bias)」の３つに整理された。「兎の耳現象」は「正のプロトタイプ効果 (Positive Prototype-effect)」によって説明可能である。反対に、存在しないがゆえに記述されるという「兎の角現象」は、「負のプロトタイプ効果 (Negative Prototype-Effect)」と考えることで説明ができる。また、カテゴリ内の他メンバーとの差異でないがゆえに生じる「兎のヒゲ現象」は、「共通性効果 (Commonality Effect)」として説明することができる。

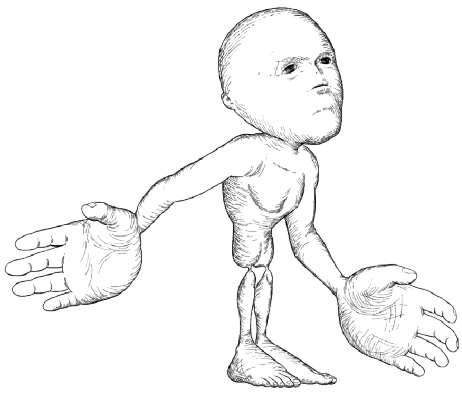

図 1　用例ホムンクルス

(3)頻度情報の対象物認識への有用性

Google 日本語 n-gram を用い、動物の身体部位の用例頻度を調査した結果を用い、①身体部位の用例頻度グラフから対象動物が同定できるか　②身体部位の用例頻度グラフにおいて、どの部分に着目することで解答しているのか調査した。平均 6 割の正答率が得られた。誤答の場合、順位相関の高い動物が選択される傾向があった。頻度情報からの対象物の同定においては、必ずしも高頻度の部分が用いられるのではなく、比較対象（同カテゴリーの他メンバー）との差異となる情報が利用されていた。この結果、類似性の高い語との差異情報を示すことで、対象物の同定がしやすくなることがわかった。

(4)特徴的な要素と用例頻度の関係

(1)(2)(3)の成果から、我々は対照する他対象物との差異となり得る特徴的な要素に着目し、それらが高頻度であることを期待するとわかった。しかし、高頻度であることの期待される要素が、必ずしも高頻度で言及されていない場合がある。たとえば、馬と人との差異として角を有するユニコーンと鬼を見ると、ユニコーンの角は期待通りの高頻度で言及されるが、鬼の角は頻度が低い（図 2）。

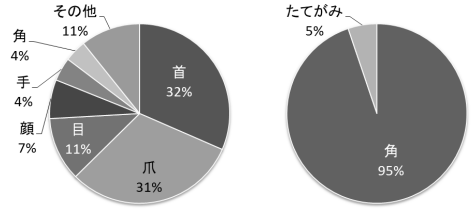

図 2　鬼（左：119,006 件）とユニコーン（左：695 件）の部位頻度分布

期待する頻度と実頻度に差の生じる一因は、用例において比喩表現に現れていた。特徴的な要素は、外観があれば形状を表す喩辞として用いられる傾向がある。ゆえに、固定的なイメージがない場合には比喩表現として用いられにくい。また、対照されやすい他動物が被喩辞となる比喩表現では、差異となる要素こそあえて言及する必要がない。調査の結果、特徴的な要素と用例頻度の関係には比喩表現のような表現形式が関わるため、頻度情報を用いる際には考慮が必要であると示した。

また、データの言語横断的な普遍性（図 3）と問題点についても確認した。

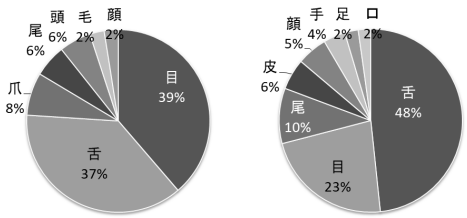

図 3　chameleon（左：1,689 件）とカメレオン（右：1,136 件）の部位頻度分布

(5)テキストの示す対象物を認識するために有用な記述

複数辞書に共通して記載のある語釈、辞書の語釈に不足しているとされた情報を追加したテキスト、コーパス（BCCWJ・Google 日本語 n-gram）から取得した用例を用い、そ

れぞれのテキストから対象物を同定する実験を行った。どの実験結果でも正答率は半数程度にとどまり、テキストのみからの対象物認識は困難であった。また、対象物の認識に求められた情報は、主に読み手の経験や知識を喚起する情報と、提示された情報によって設定したカテゴリにおける他メンバーとの差異に関する情報であった。我々が実際目にするテキスト（コーパス）からは、個別的一般的な経験や知識は取得しやすく、予め読み手の保有している知識と合致した場合には有用な情報となる。しかし、対象物に関する知識が読み手に不足している場合、対象物の認識には親カテゴリのプロトタイプとの差異を記述することが有用であり、あるいは誤認を避けるために他メンバーとの差別化が可能な記述を行うことが有用であるとわかった。テキスト情報の取得のしやすさと対象物認識への利用を表 1 にまとめる。

表 1　情報取得のしやすさと対象物認識

<table>
<tr><td rowspan="2">コーパスから</td><td colspan="3">対象物認識に</td></tr>
<tr><td>役立つ</td><td>役立ちにくい</td><td>利用可能</td></tr>
<tr><td rowspan="2">取得しやすい</td><td>一般的経験知識を喚起</td><td>個別的経験知識を喚起</td><td>一般知識でないが特徴的</td></tr>
<tr><td>読み手に対象物知識有</td><td>読み手に対象物知識無</td><td>追加情報の検索が可能</td></tr>
<tr><td rowspan="2">取得しにくい</td><td>個別的経験知識に合致</td><td colspan="2" rowspan="2">N/A</td></tr>
<tr><td>対象物の差別化が可能</td></tr>
</table>

(6)情報提示順が対象物認識に及ぼす影響、対象物認識に有用な情報提示順

(5)の成果を用いた実験を行い、同じ情報の提示順序が異なることで読み手の対象物同定率が変化する場合、どのような情報が読み手の認識を促進もしくは阻害するのか調査した。また、情報増加と正答率の関係、誤答に至った情報提示順の分析を行うことで、提示した情報のカテゴリとプロトタイプが認識に及ぼす影響についても考察した。この結果、テキストのみから対象物を適切に認識するための情報提示順序をこれまでの辞書一般とは異なり、以下のように提案する。まず一般的な読み手の有する経験・知識を喚起し、対象物の含有されるカテゴリが想定されるような情報を提示する。この情報から、一般的な読み手にとって対象物に近いプロトタイプが想起させられることが望ましい。続いてそのプロトタイプと対象物との差を示す情報を提示することで、対象物が認識可能である。また、未知の対象物であっても認知しやすくなる。

(7)対象物説明における比喩表現（隠喩と直喩）の差異

コーパスから取得された用例に現れていた比喩表現に着目し、隠喩と直喩における使い分けの実態を明らかにするとともに、両者の差異を考察した。まず、コーパスから同一の喩辞と被喩辞の組み合わせについて隠喩と直喩の用例を収集し、それぞれの産出傾向を明らかにした。同じ対象物の説明に用いる場合であっても、直喩が視覚的、隠喩が内面的な類似性を表すという用例分布が示される。次に、視覚的な情報を記述する作文実験では、直喩が頻出し、視覚情報が曖昧な喩辞の調査では隠喩が頻出するという傾向が明らかであった。

また、直喩は視覚的な特徴(形状など)の類似性が近接文脈に描写され、隠喩は喩辞と被喩辞の本質的な特徴(内面など)の類似性が長い文脈を経て集約されるという違いがある。

このように、表現の違いがテキスト内容や順序の他にも対象物の説明に影響する可能性が考えられた。今後、表現形式に着目した発展的研究を進めたいと考えている。

＜使用コーパス＞

①Kudo, Taku., and Hideto Kazawa. Japanese Web N-gram Version 1 LDC2009T08. Web Download. Philadelphia: Linguistic Data Consortium, 2009.

②国立国語研究所『現代日本語書き言葉均衡コーパス』

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　4 件）

①加藤　祥、テキストからの対象物認識に有用な情報提示順序―動物の説明文を用いた調査例―、国立国語研究所論集、査読有、15 巻、2018、ページ未定（採録済み）

②加藤　祥、隠喩と直喩の違いは何か―用例に見る隠喩と直喩の使い分けから―、認知言語学研究、査読有、3 巻、2018、pp.1－22

③加藤　祥、特徴的な要素と用例頻度の関係 : 角を例とした一考察、国立国語研究所論集、査読有、14 巻、2018、pp.55－72、DOI: 10.15084/00001412

④加藤　祥、テキストからの対象物認識に有用な記述内容―動物を例に―、国立国語研究

所論集、査読有、9 巻、2015、pp.23－50、
DOI: 10.15084/00000460

〔学会発表〕（計　9 件）
①加藤　祥、浅原正幸、読み手が共通の認識を得るための情報とその表現－小説のタイトルと帯から読み手が取得する情報－、社会言語科学会第 41 回大会（東洋大学）、2018

②加藤　祥、浅原正幸、テキストからの対象物認知における情報提示順序の影響、日本認知科学会第 32 回大会（千葉大学）、2016

③Kato Sachi、‘ Man becomes a dog ’ The difference between metaphor and simile in the corpus、6th UK Cognitive Linguistics Conference (Bangor University)、2016

④Kato Sachi、The effect of metaphor on frequency of usage: Horns are mentioned more for unicorns, less for devils、13th International Cognitive Linguistics Conference(ICLC13)(Northumbria University)、2015

⑤加藤　祥、象は鼻が長いか－テキストから取得される対象物情報－、第７回コーパス日本語学ワークショップ（国立国語研究所）、2015

⑥ Kato Sachi 、 Which features of encyclopaedic descriptions are useful in identifying the entities? A case study of animals 、 5th UK Cognitive Linguistics Conference (Lancaster University)、2014

⑦保田　祥、浅原　正幸、対象物の認知における頻度情報の影響－部位頻度を用いた動物の同定を例に－、日本認知科学会第 31 回大会（名古屋大学）、2014

⑧保田　祥、コーパスから取得した用例で対象物が認識可能であるのか、第５回コーパス日本語学ワークショップ（国立国語研究所）、2014

⑨保田　祥ほか、何が記述してあればテキストの示している対象物がわかるのか、日本認知科学会第 30 回大会（玉川大学）、2013

６．研究組織
(1)研究代表者
加藤　祥（KATO, Sachi）
人間文化研究機構・国立国語研究所・コーパス開発センター・プロジェクト非常勤研究員
研究者番号：40623004