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

保田　祥（国立国語研究所コーパス開発センター）†

Can We Recognize Objects Using the Texts in Corpora?

Sachi Yasuda (National Institute for Japanese Language and Linguistics)

１．はじめに

テキスト情報による経験のみで，人はテキストの示す対象物がどのようなものか認識することができるのだろうか。

本研究では，対象物に関する文脈情報をコーパスから抽出し，用例からの対象物の推定実験を試みる。コーパスから取得可能な用例をまとめるとともに，対象物を同定するために有効な要素・用例について考察する。また，大規模なコーパスを用いるなど用例数を増やすことによって，取得可能な情報量もまた増加するのかという問題についても検証を行う。調査は，コーパスから取得した用例で対象物がどの程度同定可能であるのか，現代日本語書き言葉均衡コーパス（以下BCCWJ）を用いた被験者実験を行った（調査①）ほか，Google-日本語Ngram[1]（工藤・賀沢,2007）から百科事典的意味（意味的な用例）を取得することを試み，BCCWJから取得可能な用例との異同を確かめた（調査②）。これらをもとに，コーパスから取得可能な，対象物の認識を可能とするテキスト情報を考えた。

２．先行研究と本研究

たとえば辞書は，対象物の説明として適切な質・量の情報を提供していると考えられる。しかし，Fillmore & Atkins (1994) では，辞書語釈で解釈できない用例が見つかることが示されている。辞書語釈として記述された要素から対象物を同定する実験を行うと，読み手が知識を有している動物に限っても，半数程度（52%）しか同定できなかった[2]（保田ほか，2013）。

Sinclair（1991，1992）は，学習者に必要とされるのは用例であるとし，コーパスに基づく用例を重視した新たな辞書（COBUILD）作りを行っている。対象物に関する多くの用例こそが，対象物の経験を得るテキスト情報であるとも考えられる。

それでは，コーパスから取得した用例を提示することで，対象物を認識することができるのだろうか。また，コーパスから取得できる要素・用例は，対象物の同定に質・量的に十分だろうか。本研究は，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を調査する。

３．調査①：BCCWJ 用例から対象物を同定する

BCCWJ から取得した要素・用例で対象物がどの程度同定可能であるのか，動物を対象として被験者実験を行い，結果の分析を行う。

３.１　データ①：BCCWJ からの用例取得

BCCWJから動物（一般に知識があると考えられる単語親密度 5.000 以上[3]）に関する要素・

---

† yasuda-s@ninjal.ac.jp

[1] http://www.gsk.or.jp/catalog/gsk2007-c/

[2] 日本国内で出版されている国語辞書 10 種類を用い，5 種以上の辞書に語釈として記述されていた内容を用いた。

[3] 最大値は 7.000（天野・近藤編，1999)。

用例を抽出した。中納言[4]を用いて検索語の前後 50 文字を取得し，手作業で整理を行った。なお，文意の読み取りに文脈情報不足している場合や用例が文字数の制限によって途切れている場合などは，前後 500 文字を再取得して同様に抽出を行った。さらに，実験協力者に提示する際には，抽出した用例が句などの場合に文へ整形[5]したほか，意味的に同種と判断される用例については，以下の例のように作業者の判断でまとめた。このような複数用例をまとめた例を，本稿では意味的な用例と呼ぶ。

取得用例）

A. カモシカの被害防止対策調査　カモシカの食害発生機構の解明　カモシカの林業被害が近年，特に問題になっており～　（「環境白書」）

B. カモシカが増えたため，ヒノキの幼木を食い荒らされるという被害を受けている～　（中村幸昭「鳥羽水族館館長のジョーク箱」）

C. 最近カモシカに食われる被害が出ている『会津の伝統野菜を守る会』によって選ばれた野菜は，現在十四品目。　（丹野清志「やさい畑」）

A～C のまとめ例：意味的な用例）

この動物による林業（ヒノキの幼木）や農業（野菜）などの食害が問題とされている。

３.２　実験：BCCWJ から取得した用例を用いた対象物同定実験

10 種類の動物について，実験協力者 12 名（日本語母語話者）が，提示された用例群（対象物名は「この動物」とマスク）から対象物の同定を行った。実験協力者は，回答が可能であった場合には，どの用例が有用であったのかをあわせて記した。

実験に用いた用例については，実際に検索した結果としての件数（異表記を含むが，異義語（固有人名など動物でない）を除く）から，同サンプル内などの重複を除いたほか，同種の要素を有している用例については意味的な用例とした。以下に示す動物についての用例を提示した（異なり：提示した意味的な用例数／サンプル数／述べ：検索結果件数）。

タヌキ（41 例/372 サンプル/581 件）・カワウソ（23 例/38 サンプル/164 件）
テントウムシ（21 例/48 サンプル/69 件）・オットセイ（17 例/11 サンプル/67 件）
スズキ（12 例/36 サンプル/65 件）・カナブン（17 例/18 サンプル/36 件）
カマス（8 例/19 サンプル/26 件）・ジュウシマツ（10 例/7 サンプル/14 件）
ジャガー（5 例/9 サンプル/13 件）・ナイチンゲール（9 例/6 サンプル/8 件）

３.３　結果：BCCWJ から取得した用例を用いた対象物同定実験

10 種類の動物[6]について，BCCWJから抽出できる意味的な用例から対象物を同定した結果，動物毎の正答率のマクロ平均は 25%となった（人毎[7]の正答数のマクロ平均は 3.0 種類である）。

半数以上が正答する高正答率動物群（図１のⅠ）と，同定できないかごく稀に同定可能な低正答率動物群（図１のⅡ）に大別できる。

---

[4] https://chunagon.ninjal.ac.jp/

[5] 要素を複数含む場合など，一句を二文にした例もある。

[6] 保田ほか（2013）の実験によって，それぞれの動物に関する実際の知識率を調査している。知識率の平均は，87%であった。ナイチンゲール（20%）を除き，70%以上の実験協力者が当該塔物の知識を有していた（min：カマス（70%），max:タヌキ・テントウムシなど（100%））。

[7] 内訳は，1 種類：1 人，2 種類：5 人，3 種類：3 人，4 種類：2 人，8 種類：1 人である。

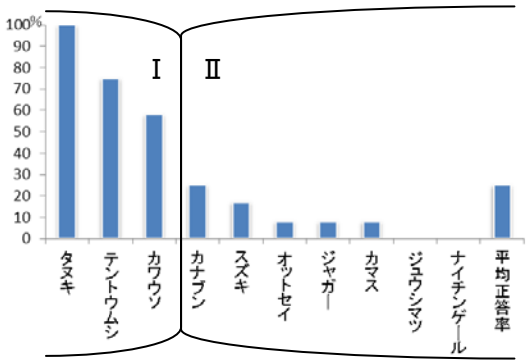

図１　BCCWJ 用例からの対象物同定正答率

３.３.１　高正答率動物群に見る判断に有用な情報

タヌキは 100%の正答が得られた。判断の際には，92%の実験協力者が「カチカチ山」(「タヌキ」全件の 2.5%と高頻度用例）を有用であったとした。同種の「ぶんぶく茶釜」は 33%にとどまるが，「カチカチ山」のように一般に知られた文化的知識レベルの情報は，タヌキを想起するトリガーとして重要であると考えられる。すなわち，用例としての出現頻度も高い人口に膾炙した物語名[8]などによって，高い正答率が得やすいことが予測される。このような文化的知識として経験した情報が，対象物の認識に最も有用である可能性があろう。

次に，75%の正答が得られたテントウムシで，判断に有用とされたのは，特徴的な形状に関する用例，とくに比喩表現でもある以下の用例（68%）であった。

例）（前略）お椀を逆さまにしてテーブルの上に伏せたような格好をしている。少し，てんとう虫にも似ている。　（垣根涼介「君たちに明日はない」）

また，一般的によく知られた情報と考えられる用例「アブラムシを捕食する。」が 42%で有用とされていた。このような用例は，テントウムシそのものの生態的な特徴説明というよりも，具体的な用途としての前提知識となっている例である。

例）毛虫は全部手で取り，アブラムシはテントウムシを連れてきて退治すると，無農薬で収穫できた。　（坂下まりあ「あたしは非定型精神病なのだよ」）

比喩表現をはじめ，前提知識として既知の情報という扱いをされている用例は，一般的な知識であると考えられている情報であろう。タヌキでも，「でっぷりした体（33%)」「寝たふりをする（42%)」のように，比喩表現や慣用的な表現についての用例が，他用例に比べて対象物の認識に有用性が高いとされていた。但し，知識や実際に経験する機会には個人差のある場合が考えられるため，テントウムシの例では，100%の正答には至らなかったようである。

また，カワウソは 58%の正答であるが，判断に用いられた用例は，正答した実験協力者のすべてが一致するということはなく，最大でも 3 人の一致であった。また，誤答は 1 人のみ（「ミンク」）で，4 人が無回答（空欄）であった。カワウソでは，複数の用例のいずれかが個人の有する個別的経験知識と合致した場合にのみ解答が可能となったと考えられる。テントウムシでも「落ち葉の裏などに集まって越冬する。」「手にとまることがある。」のような用例についても複数人（25%）が有用としていたが，経験の汎用性が下がり，個別的であるといえる。

[8]同様にオオカミならば「赤ずきん」，ウサギであれば「ウサギとカメ」のような用例が，それぞれ当該動物全件の 3%という高頻度の用例である。

このように，対象物の認識に有用とされたのは，文化的知識＞一般的経験知識＞個別的経験知識に関する用例の順であった。

なお，テントウムシの誤答は「蝶」であったが，その際「翅を広げた様子をリボンに喩えた例」が判断に有用だったとされていたことから，実験協力者によっては，印象に残る一文など，一つの経験が対象物の認識に強い影響を及ぼした可能性も残る。

３.３.２　同定の難しい動物群に見る判断に有用な情報

タヌキ・テントウムシ・カワウソ以外の動物については，正答率が 0%の動物もあり，いずれも低い正答率であった。

稀に同定が可能な場合には，「玄関や階段，バルコニーにいる」という個別的な経験に合致したと考えられる用例や，もしくは，「メソ・アメリカ文明の～」「プレ・インカ文明は～」といった専門性が高い知識と考えられる用例が判断に有用とされていた。個別的経験知識と一致する用例があった場合にのみ[9]，解答が可能であったことがわかる。

また，ジャガーの誤答は「トラ」「チーター」「ライオン」，オットセイの誤答は「アザラシ」「ビーバー」「トド」，カナブンの誤答は「タマムシ」「カメムシ」「コガネムシ」「セミ」，ナイチンゲールの誤答は「ヒバリ」「カナリア」などと，それぞれ多様であって統一性もなかった。

３.４　考察：対象物の同定に必要とされる情報がコーパスから取得可能か

知識があるはずの動物[10]であっても，その動物の情報が提示された際，その情報が実験協力者の知識と合致しなければ，対象物の同定は不可能である。

辞書（保田ほか,前掲）よりもコーパスから取得できた用例数は多く，情報量は豊富であると考えられるが，それらの記述が必ずしも対象物の同定に質的に十分とはいえない。

それでは，対象物の認識に求められる情報がコーパスから取得可能なのか。

３.４.１　経験知識を喚起する情報

経験知識を喚起する情報，「調理法」や「作品名」，「場所名（動物園・水族館など）」などは，コーパスから得られる場合が多い。

調理法例）スザンヌがよく作っていたのは，魚のクール・ブイヨン煮。ポワソニエールという魚が一匹まるまる入るような，大きな楕円形の鍋で煮た鱒や鮭，すずきなどをゆでたじゃがいもといっしょに食べた。（猪本典子「修道院のレシピ」）

作品名例）アンデルセンは，『天使』と『みにくいアヒルの子』と一緒に『ナイチンゲール』という作品を彼女のもとに送った。『ナイチンゲール』は，自然に歌う鳥の方が人工的なおもちゃよりもすばらしいということを述べた作品である。（アリソン・プリンス「ハンス・クリスチャン・アンデルセン哀しき道化」）

場所例）話が変わりますが，写真は海遊館のカワウソです。（「Yahoo!ブログ」）

実際に，コーパスから取得した要素から対象物を同定することができた場合，上記の用例のような経験知識源となり得る情報は有用性が高いと判断されている。

しかし，上記の作品名例で見たアンデルセンの作品名のほか，以下の例のような場所情

[9] なお，本実験において正答率が 0%であった「ジュウシマツ」と「ナイチンゲール」についても，ジュウシマツを飼育していた経験のある場合や，アンデルセン童話などの知識のある場合には，正答が可能であったことが別途確認された。

[10] 注 6 を参照。知識率はナイチンゲール（20%）を除き 70%以上であった。

報と関わる例などは，動物名がマスクされた場合，対象物の同定に有用とされない。

例）わが国では絶滅が心配されるカワウソを使った漁が，中国南部で行われていることが，
周達生・国立民族学博物館助教授の最近の現地調査でわかった。（花輪莞爾「猫学入門」）

「ウサギのヒゲ」や「ダチョウのクチバシ」のように，一般性が高いが比喩などに用いられるのではない（当該対象物に特徴的ではない）情報は，テキスト化されにくい（Yasuda ほか, 2012）。すなわち，一般に未知であろう情報を発信している用例が多いことから，個別的経験あるいは専門性が高い情報となりがちであり，読み手の経験知識と合致しない例も多いと考えられる。

３.４.２　想定されるカテゴリーの他メンバーとの差異情報

想定されるカテゴリーにおける他メンバーとの差異を示す情報が，コーパスから取得しにくいといえる。対象物の同定の難しい動物群においては，提示された用例では，「大型肉食獣」や「水棲動物」などと想定されたであろうカテゴリーにおいて，他メンバーとの差別化ができなかったことが推測されるためである（前掲，３.３.２）。

カテゴリーの他メンバーとの差異としては，特徴的要素のほか，大きさなどの情報が役立つ場合が考えられよう。対象物が「大きい」「小さい」のような情報はコーパスから取得しやすく，たとえばオオカミであれば，「大きい」が全用例中 2%出現する高頻度の要素として取得されるのである。しかし，用例を見ると，「大きい」はオオカミカテゴリーの中で「大きい」という情報である。対象物そのものの情報ではあるが，哺乳類や猛獣，野生の獣などのカテゴリーにおける大小の差別化情報ではないため，対象物が何であるのか認識するために有用なのではない。

例）森の主ならかれらは大きなオオカミだというだろうよ。さもなければ神聖な白いシカだと！　（栗本薫「マルガ離宮殺人事件」）

３.４.３　比喩表現

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は，コーパスから取得しやすい情報であり，対象物の同定への有用性が考えられる。一般的な読み手が「当該動物に特徴的である」と想定可能な性質が，焦点とされているはずだからである。テントウムシの用例において，対象物の同定に 68%の実験協力者に有用とされたのも比喩表現であった。

動物によって異なるが，たとえば知識率が 100%と一般によく知られていると考えられるタヌキであれば，全用例中の 11%（63 件/581 件）が「みたい」「よう」などの指標を含む直喩表現，あだ名などをはじめ隠喩表現と考えられる用例である。このようにタヌキ・マムシ・オオカミなど知識率の高い動物は比喩用例が全用例の一割以上と多いため，典型的と考えられる喩えの用例が取得可能である。以下に例を示す。

例）婿には人間出身ではなく猛々しい蝮のような悪い男を待つがいい。翼を持って虚空を自在に飛行する，炎と剣をもってすべてのものを懲らしめる男を。
（高橋康雄「ギリシアの女神の物語」）

例）あんまり無防備だと，狼に食べられちゃうぞ。ま，その狼も，今は病人だからおとなしいけどさあ。でも，あんまり可愛いことばっかりしてると―わからないんだからな。
（七海花音「僕らのロビン・フッド宣言」）

但し，これらの用例は喩えられた人の特徴から当該動物の特徴が類推されるのであって，動物の特徴から対象動物が想定されるのではない。たとえば，「したたかな人や腹黒い人を

喩えて言う」という用例がタヌキの同定に有用な情報とは必ずしも判断されなかった（50%）。「体を硬直させる様子を喩えて言う」という用例が，オットセイの判断に結びつくこともなかった（0%）。必ずしも，比喩表現に用いられた特徴が，対象物の認識に役立つのではない。

しかし，「比喩表現に適切と考えられる特徴を有した動物」という ad-hoc なカテゴリーにおいては，当該動物が選定されるための他メンバーとの差異が明らかになりやすい可能性はあろう。今後，比喩表現を含む用例を詳細に分析することで，対象物を認識するためのテキスト情報が得られることが期待される。

３.５　まとめ①：コーパスから取得した用例で対象物が同定可能か

本稿の調査対象とした動物においては，コーパスから取得した用例から対象物の同定が可能な動物と難しい動物に大別できた。

対象物を読み取るための情報は，文化的知識（当該対象が登場する著名な物語名など）＞一般的経験知識（比喩表現に用いられる知識・前提的とされる知識など）＞個別的経験知識（読み手と一致する場合）の用例の順に有用とされていた。対象物を同定するために有用と考えられる知識経験を喚起する情報は，コーパスから取得しやすい。しかし，読み手と共有可能な知識経験が多い動物であれば，用例から対象物の同定が可能となるが，異文化圏の知識を要する用例や，そもそも未知であろう情報が発信されている用例などが多い動物は，対象物の同定が困難となるのである。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，また，対象物そのものの情報は得られても，想定されるカテゴリー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。

４．調査②：大規模 Web コーパスから用例を取得する

調査①で用いた動物は，取得できた用例数が最大のタヌキで 41 例（372 サンプル/581 件）であった。1 億語規模の BCCWJ では，取得可能な用例が少ないために，調査対象とした動物の十分な情報が得られないという可能性は考えられる。ここでは，用例数が多ければ，多くの適切な情報が取得できるのであろうかという問題について，大規模なコーパスを用いて検証する。また，個別のコーパスの特性も確認する。

４.１　データ②

Google-日本語Ngram（Web から抽出された約 200 億文（約 2,550 億語）[11]の日本語データ）を用いて，取得可能な用例を調査した。

調査①で調査した動物のうち 3 種類の動物について用例（n-gram データ（1～7 gram）・頻度 20 以上）の抽出を行った。タヌキ（異表記「狸」「たぬき」を含む：1,893,000 件），オットセイ（61,800 件），ジュウシマツ（異表記「十姉妹」を含む：33,500 件）が取得できた。

但し，本稿のような意味的情報を取得する試みにあたっては，まず，文などが最大 7gram で分割されているという問題がある。そこで，手作業により，7gram で分割されている複数語の重なる同件数の用例を合わせることで，意味の把握が可能な長さとし，以下のように最大例で 23gram を取得した。

例）老け顔アンパンマンおばさん狸顔だよね
　　コンテンツの著作権はスタジオタヌキが所有しています
　　水族館といえばイルカやオットセイなどによるショーをして

さらに，同内容と考えられる用例を，調査①の作業（データ①参照）と同様に，意味的

[11] 総単語数は 255,198,240,937，総文数は 20,036,793,177。

な用例としてまとめた。

取得用例）A．サックスのレース&可愛いオットセイ柄のブラ&ショーツ
　　B．オットセイ柄のカットソー&スパッツ

A・Bのまとめ例：意味的な用例）衣類の柄に用いられることがある

また，当該動物そのものではない固有名詞（HN・店名・キャラクター名など）と判断される用例は分けた。但し，Webベースの大規模コーパスにおいては固有名詞や固有表現が大部分を占めているといえるが，完全な分類は困難であるため，分類は作業者の調査と判断によった。

なお，この作業では，同ページから重複取得されていると考えられる用例も散見されていた。たとえばジュウシマツにおける「食事と音楽，本，ジュウシマツとラブラドール等自身のアンテナが向いたもの」という用例が 1,210 件取得されるが，この句は特定のブログ（http://suzusuzu.jugem.jp/）における説明部分に含まれると確かめられた。

このような例は，ブログやサイトのタイトル，メニューなどの説明文に検索語が含まれているために，重複カウントされている場合が多いようだが，書籍タイトル「キツネとタヌキの大研究―人間との長くてふかーいつきあい（348 件)」や，演劇タイトル「ミュージカル吾が輩は狸である（3,670 件)」，商品紹介「劇場版どうぶつの森キャラポーチ全5種 タヌキ商店 DS 小物（282 件)」などの場合もあった。また，特定の質疑や説明等が，「【オンラインゲーム】トリックスターの狸育成方法について質問します（302 件)」「(名前が分からないのですが,) よく悪代官や悪徳商人する人で顔はタヌキ顔，ちょっと太りがちで強くはない（227 件)」「東京の多摩丘陵を舞台に，そこに棲むタヌキたちが人間に反旗をひるがえすべく（346 件)」のように，別 URL から複数取得されている可能性のある場合もみられる。

これらの特定例は，「(まめ) たぬきの雑記（24,200 件)」のように「たぬき」用例全件（842,000 件）の 3%を占める高頻度のものもあるほか，文を含むレベルでも「ここをクリックすると讃岐のタヌキのランキングポイントが加算されます（1,250 件)」「ぽんぽこ狸の考え方　社会問題等様々なことについてぽんぽこ狸が，独断と偏見で説教します。(1,010 件)」のように，多数の重複用例として取得される。

大規模であるが Web ベースのコーパスを用いると，Google-日本語 Ngram の場合，同 URL から重複取得されている用例が多いことや，固有名詞や商品紹介などの重複を除いたことから，用例数は取得数の 13%程度の量となった。また，このような重複ページの多さによって，本稿で示した Google-日本語 Ngram から取得した用例の頻度情報に正確性の疑問がある。

４.２　結果と考察②

当然 Ngram であるという取得可能な文字数の制限により，取得できない情報が多いことも予測されるが，単純な検索結果としては約 190 万件の規模が取得できたタヌキ用例（異表記「狸」「たぬき」を含む）であっても，取得できる意味的な用例の種類は 28 例となり，莫大になるということはなかった。そこで，取得した意味的な用例と調査①で BCCWJ から取得した意味的な用例（データ①）との対照を行った。データ①②には，タヌキで 22 例（調査異なり計 44 例中)，オットセイで 6 例（同 25 例中)，ジュウシマツで 4 例（同 17 例）の重複（意味的に同内容）があった。差異の生じた用例に着目し，Web コーパスと BCCWJ に差が見られた情報を分類した。それぞれ得やすいと考えられる情報をまとめておく。

４.２.１　大規模 Web コーパスから得やすい情報

まず，コーパスの規模によって取得可能となる用例がある（以下 A)。単純に意味的な用例の種類が増えるというものもいくらかあるが，特に，頻度が得られることで「多い」という情報の得られることが有用となる例である。取得された用例が個別的であるのか，一

般的であるのかが，頻度情報によって分類可能となるためである。
但し，均衡コーパスと異なり，個人的経験・評価，商品情報が多く取得されることから，取得可能となる用例に見られる偏りが考えられる（以下 B)。

A.　用例件数の量に関わる情報
A-1.　コーパスの規模が大きいため取得できる（別種類の）用例による
例）剥製にされていることがあり，置物になっている。(タヌキ)
シャチなどに襲われることがある。(オットセイ)
小斑があるものもいる。(ジュウシマツ)
歌に複雑な文法があるという説があり，書籍もある。(ジュウシマツ)
A-2.　個別のタイトルや固有名詞，複合名詞などの用例の多様さによる
例）三大伝説は，證誠寺・分福茶釜・隠神刑部とされる。(タヌキ)
うどんやそばにこの動物の名がついた種類がある[12]。冷やしたものもある。
丼や握り飯など米を用いたメニューもある。(タヌキ)
A-3.　共起情報の頻度による
例）この動物に喩えるのは，特に中年以上の男性や猫が多い。(タヌキ)
イタチ，河童，ウサギ，猪などと一緒に扱われることが多い。(タヌキ)
アザラシ，ペンギン，イルカなどの海獣類と並列されやすい。(オットセイ)
子供の頃など，昔飼っていたという人が多い。(ジュウシマツ)
A-4.　同種の用例が複数得られることによる（詳細情報）
例）タマネギを炒めるなどしてこの動物の色という表現がある[13]。(タヌキ)
青いこの形に似た猫型ロボットが著名である[14]。(タヌキ)

B.　均衡コーパスとの性質的差異として得やすい情報
B-1.　一般的評価・俗説
例）可愛い。いたずらをすると思われている。(タヌキ)・愛らしい。(ジュウシマツ)
B-2.　商品情報
例）ぬいぐるみや人形に模られる。(タヌキ)
陰茎や睾丸，骨格筋から抽出したエキスが加工食品に用いられる。(オットセイ)
B-3.　一般的経験
例）同種の鳥類とあわせて，家族・ペアなどで複数飼いされる。(ジュウシマツ)
水族館でショーが見られる。(オットセイ)
鳴き声に特徴があるといわれ，鳴きまねをする人や動物がいる。(オットセイ)

４.２.２　BCCWJ から得やすい情報

本稿の調査に用いた Google-日本語 Ngram では，そもそも文単位の検索が不可能であり，文脈情報が得にくい。たとえば「タヌキみたいな猫」が何をもってタヌキに喩えられたのか，タヌキの情報を読み取るためには，前後の文脈が必要となる（以下 C)。
意味的な用例を取得するためには，文脈情報を必要とする例も多く，現実的に運用される大規模コーパスから得られる情報には制限があると考えられる。また，専門性のある情報についての用例は，Web コーパスからは得にくいものである（以下 D)。

[12] BCCWJ から取得される「タヌキ」料理は，「関西では油揚げいりのそば・うどんを示す」例のみであった。
[13] BCCWJ の「タヌキ色」の用例は，「タヌキ色のインキ（橋爪和子「ぼくはひとりぼっちじゃない」)」のみで，詳細情報はない。
[14] BCCWJ でも同種の用例は取得されたが，「猫型ロボットなのに（中略）昔の世界に行くとよく狸と間違われやすいですが（Yahoo!知恵袋)」という一例に留まっていた。

C.　文脈から得られる情報（下線部はGoogle-日本語Ngramからも高頻度で得られる情報）
C-1.　詳細・関連・補足情報
例）庭で見ることもある。餌付けされて野生に戻らなくなることがある。（タヌキ）
里山で食害の被害の原因となる。雑食性で人家の残飯をあさることもある。
荒毛の下に柔らかい上質の毛皮を持つ。長い毛1本に短い毛が約五十本もあり，保湿効果を高めている。（オットセイ）
鵺の胴がこの動物であると平家物語に描かれている。（タヌキ）
小型で11ｃｍ。（ジュウシマツ）
C-2.　文をまたぐと考えられる情報
例）高速道路などでよくはねられて死んでいる[15]。（タヌキ）
全身が硬直した様子などを，この動物になったと喩える例がある。（オットセイ）

D.　一般的に言及されにくい情報
D-1.　専門的知識[16]
例）ヒレ状の前後の脚はアシカより長く，水中の生活に適応している。（オットセイ）
プラスチックの悪影響の代表的例として，網絡まりがあげられる。（オットセイ）
平安時代は猫をこの用字で表していた。（タヌキ）
日本で作り出された。（ジュウシマツ）
D-2.　限定的知識
例）サイコロを二つ使うことをこう呼ぶゲームがある。（タヌキ）
夕食抜きの素泊まり客をこう呼ぶ。（タヌキ）
D-3.　経験が得にくい
例）アベル・タスマン国立公園では，人を恐れずじゃれついてくる。（オットセイ）
水面からひょっこり丸顔をのぞかせる。（オットセイ）
子育てが上手い。（ジュウシマツ）

４.２.３　個別のコーパスに依拠する情報

その他，本稿で使用したコーパス各々に依拠すると考えられるために，重複のない用例が見られている。サンプリングされたテキストの生産時期により生じた差異と考えられる。

E.　サンプリングされたテキストの生産時期の影響が考えられる情報
E-1.　催事など
例）世界じゅうのこの鳥の展覧会が行われる。（ジュウシマツ：Google 日本語 Ngram）
その他，演劇名・映画名などの固有名詞とそれに関する情報[17]
E-2.　流行など
例）タヌキケーキ。（タヌキ：Google 日本語 Ngram）
E-3.　その他時代的と考えられる情報
例）酒をよくのむといわれる。（タヌキ：BCCWJ）

---

[15] 「死骸」「死体」「～を見る」類は高頻度であるが，「高速道路」のような場所情報が得られておらず，文が別になっていることが考えられる。

[16] 雑学やクイズなどから専門的な用例が得られることはあり，「語源はアイヌ語といわれている。（オットセイ）」のような専門的知識が，Webコーパスからのみ得られている場合もある。

[17] 固有の作品が公開・放送などされることにより，個別的な表現をはじめとする関連情報のテキストが増加するほか，個人の感想や意見傾向がテキスト化される影響が考えられる。
例）「平成狸合戦ぽんぽこ（映画）」：1994 年 7 月 16 日公開で，テレビ放送が繰り返されている。
例）「吾輩は狸である（ミュージカルコメディ）」：2003 年 4 月 7 日初演。

５．まとめ

テキスト情報から対象物を同定する試みは，明確に対象物の認識が可能なテキスト作成のために必要な情報を整理する試みの一つである。最後に，調査①②の結果から，対象物を同定するためにコーパスから取得可能な情報について考察する。

コーパスを用いた用例からの対象物の同定は，平均正答 25%と低く，基本的には困難であるといえる。よって，テキスト情報による経験のみで人が対象物をどのようなものか認識することも当然難しいのだと考えられる。また，対象物の同定正答率は，動物によって高い群と低い群に大きく二分されたという結果でもあった。タヌキのように 100%の実験協力者が対象物を同定可能な場合には，コーパスによって提供された情報（意味的な用例）が，十分であったということである。但し，個人の有する情報と合致したかどうかという点が有用性の有無の判断ともなっていた。すなわち，対象物の同定には，文化的レベル＞一般的レベル＞個別的レベルの順に有用性があると考えられる。

これは，92%の実験協力者に有用とされた「カチカチ山」のストーリーおよび登場人物情報など，テキストに記述された情報外の情報を対象物の認識に求めるということではあろう。しかし，テキスト情報から対象物を同定するために有用とされた情報が，一般的，さらには文化的であることを求められていたという点から，大規模コーパスによって頻度情報が提供されることが望ましいともわかった。取得された用例が個別的なレベルの情報であるのか，一般的レベル以上であるのかは，ある程度，頻度情報によって分類可能となる。頻度が得られること（「多い」という情報の得られること）が有用と考えられる。

なお，コーパスからは，経験知識に関する情報は得やすいのではあるが，読み手の想定したカテゴリーの成員との差異を示す情報は得にくいこともわかった。これらの情報を取得するためには，今後，比喩表現などをはじめとする「人」の特徴と対象物との関係（人を基準とした異同情報），比較対象として他の動物と並列される場合の関係などに着目したい。比喩表現などの分析には文脈情報が必要となるが，文脈が不明の場合でも，用例数が増加することによって，並列される動物は取得できる可能性がある。対象物そのものの記述のほか，同カテゴリーの成員との異同によって情報を提供する方向性も考えられよう。

以上のように，テキストに記述された対象物の認識には，読み手が一般的に有していることが期待されるテキスト外の知識の喚起が求められている。コーパスから取得した用例から対象物を認識するためには，テキストそのものから得られる情報に留まらず，広く文脈が要されるということでもあろう。

文　献

天野成昭，近藤公久編（1999）『日本語の語彙特性』第 1 巻（単語親密度）三省堂

天野成昭，近藤公久，笠原要編（2008）『日本語の語彙特性』第 9 巻（単語親密度増補）三省堂

保田祥，浅原正幸，前川喜久雄（2013）「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 回大会大会論文集，pp.370-379.

Fillmore, Charles J. and B. T. S. Atkins (1994) “Starting where the dictionaries stop: The challenge for computational lexicography”, In Atkins, B. T. S. and A. Zampolli, Ed. *Computational Approaches to the Lexicon*, 349-393. Oxford: Oxford University Press.

Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1992) “Trust the text.” In M.Davies and L.Ravelli, Ed. *Advances in Systemic Linguistics* : recent theory and practice, 5-19. London: Pinter.

Yasuda, S., Okamoto, M., & Aramaki, A. (2012) “‘Mind the Gap’ between Text and Real World: A Corpus-based Study on the Prototype Effects of Animal Body Parts.” 4th UK Cognitive Linguistics Conference.