Encyclopaedic Descriptions That are Useful for Identifying Entities: A Case Study of Descriptions of Animals

国立国語研究所論集 (NINJAL Research Papers) 9: 23–50 (2015)
ISSN: 2186-134X print/2186-1358 online

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――

加藤　祥

国立国語研究所 コーパス開発センター プロジェクト研究員

要旨

テキストの示す対象物を認識するために，どのような内容を記述することが有用か。本稿では，動物を例にした 3 種類の実験に基づく考察結果を報告する。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（現代日本語書き言葉均衡コーパス・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 *。

キーワード：百科事典的知識，対象物認知，コーパス，カテゴリ化，意味記述

1. はじめに

ある対象物について，我々はテキストから様々な知識を得ることができる。しかし，各種のテキストに記述された対象物を，我々がはたして正しく認知することができるのかという点においては疑問が生じる。たとえば，以下の（1）から（4）は「兎」の出現テキスト例（下線は著者による）であり，（5）は国語辞書の語釈文例である。

（1）　むしろ、奥山に実がなる樹木や好物の山芋などがなくなり、里山に下りてきたのが、食害の原因とみる。イノシシだけでなくタヌキ、ウサギ、猿などの被害も深刻だ。豊かなはずの本県の山々に兆す異変である。元凶は奥山を変えた人間なのかもしれない。

（PN1m_00001:『高知新聞』[1]）

* 本研究は JSPS 科研費 26770156 の助成を受けたものである。また，保田祥・浅原正幸・前川喜久雄「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，2013 年 9 月，於玉川大学）および，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，2014 年 3 月，於国立国語研究所），Yasuda, S. “Which features of encyclopaedic descriptions are useful for identifying entities? A case study of animals”（5th UK Cognitive Linguistics Conference. 2014 年 7 月，於 Lancaster University）における発表内容をもとに，データと新しい実験を加えて分析を行っている。

[1] 用例の出典は（BCCWJ サンプル ID：執筆者「書名等出典」）と示す。以下同様。

(2)　ラーブル・ド・ラパン・ソース・ムータルド
ラーブルと呼ばれるうさぎの背肉の部分を、骨付きのままローストした料理。ここで使われているうさぎはラパンと呼ばれる家畜のうさぎなので、肉の色も白く、味も鶏肉のようにマイルドで食べやすい。
（LBl5_00033：さらだたまこ・谷あつこ「レストランのメニュー」）

(3)　店長に話しかけ、ハットの素材のフェルトに、ウールとファー（ウサギ）があることや、実際に、基本のかぶり方、合うサイズなども話を伺う。　（PM21_00320:『POPEYE』）

(4)　和の雑貨には、時代を超えた美しさと愛らしさがあります。長い伝統と確かな技術の中に見え隠れする、ちょっとしたユーモアやセンスがたまらなくいとおしい。ウサギや辰、鳥獣戯画、鳥など、可愛らしいモチーフも沢山あるので、お部屋が動物だらけになりそうです。
（PB35_00262: 柳沢小実「ていねいな暮らし」）

(5)　うさぎ目の哺乳動物の総称。ふつう耳が長く、よくとびはねる。上くちびるは兎口。肉は食用。毛皮はえりまきなどにする。　（『岩波国語辞典』第 5 版）

国語辞書（以下辞書）を含め，そもそも読み手が兎を知っているものとして記述されたテキストからは，兎の形状などについて具体的な記述が得にくい。このようなテキストから得られる情報のみで，テキストに記述された対象物を認知することが容易だろうか。

本稿は，テキストが示す対象物はテキスト情報によってどの程度認識できるか，また，何が記述してあればテキストが示す対象物は認識できるのか調査する。

まず，適度に対象物について記述したテキストと考えられる辞書の語釈文を用い，対象物を認識するためにどのような要素の記述が不足とされるか，被験者実験によって確かめる。そして，不足していると考えられた情報が加われば認識可能であるか検証する。次に，コーパスから対象物の用例を収集し，一般的なテキストから対象物が認識できるか被験者実験を行う。これらの実験結果から，テキストから対象物を認識するためにどんな記述が取得しやすくどんな記述が不足しがちであったのか分析を行い，対象物の認識に有用な記述がどのようなものか考察する。

2. 関連研究：テキストから得られる情報

ある対象物についての情報をある個人の内省によって書き尽くすことは困難といえる。たとえば，辞書の語釈は専門家の内省によって記述されたテキストであると考えられる。Fillmore & Atkins（1994）は，辞書の項目例を挙げて様式や語義が辞書毎に異なることを指摘し，動詞 crawl の用例について，6 種の辞書の crawl の項目内にそのすべての用例を説明可能な記述があるのではないことをいい，コーパスに見える意味の区別が辞書よりも多様と示した。例として扱われた crawl は, 辞書においては「虫」と「手足のない無脊椎動物」に限定されていたが, 多様な「人間ではない生物」でも用いられるほか, メタファーやメトニミーの用例も現れることがいわれる。すなわち，誤用や省略例をはじめ，辞書にはない用例が多く見られるということが示されている。同時期の日本語では，後藤（1993）による名詞「神話」についての各種国語辞典における語義記

述の異同調査と「朝日新聞記事データベース」における用例調査の対照により，同様の結果が見られる。このほか，奥村・白井（2008）が，用例の語義が辞書項目に見られないことの機械的判定を目標とし，あらかじめ定義した語義だけでは新しい語義や用法に対応できなくなる例「ネタ」を示す。また，Sinclair（e.g. 1991）は辞書が用例に対応できないのは，それぞれの意味が特徴的な形式の類型と関係しているためであるという。Sinclair が編集主幹を務めた学習者用辞書の *COBUILD*（1987 〜）は，それぞれの語の意味を顕著と見なされた最小限の細目とし（Sinclair 1992），語義を構文や連語情報を含んだ文とするほか，コーパスに近い例文を掲載する試みが為される（*COBUILD* 2009: xi）。但し，これらの研究に見られるように既存の辞書の語釈文が情報不足だとして，コーパスから見つかる用例が十分な情報を提供するのかという疑問は残る。もちろん，対象物の的確な記述という点においては，用例に対応していないとしても，辞書の語釈が不足しているとは言い切れないであろう。

また，複数人の内省があれば対象物に関する十分な情報が得られるのではないかという期待もある。McRae ら（2005）は，大規模な連想実験（被験者数 725 名）によりベーシックレベルのコンセプト（541 種類）に関する意味特徴（semantic feature; 物質（知覚できる）特性，機能特性，その他の属性や百科事典的要素）を収集している。十分な意味特徴が得られている例もあると考えられるが，たとえばナイチンゲールの意味特徴が「鳥である」「飛ぶ」「歌う」「嘴と羽毛と翼がある」であるように，上位カテゴリとの別が不明な情報に限られている場合もあるため，対象物を特定するに十分な情報がどのコンセプトに関しても得られているとは言い難いだろう。

国広（1997）の「辞書の意味記述」における必要項目は，対象物を十分に説明する試みの例と考えられる。国広（1997）は，一般的な国語辞書の記述に現れにくいものとして，「語義的位置（語彙体系の中の位置）」「語義の対義的定義（対義語を示す）」「現象素[2]（認められる場合には図示）」「用例（広く実例を観察した上で適当にまとめる）[3]」「連想（動物名であれば､その動物の習性や故事来歴など（百科的知識））」を挙げる。よって，たとえば名詞の兎の意味として，一般的な国語辞書の記述に加え､「連想」情報の「《物語》「因幡の白兎」「兎と亀」「カチカチ山」。《俗信》兎は月夜に逃げる。《小学唱歌》「うさぎ、うさぎ、何見てはねる。十五夜お月さま見てはねる。」（兎）「兎追いしかの山、小鮒釣りしかの川、夢は今もめぐりて、忘れがたき故郷」（故郷）」を記述する。但し，記述内容については有用性の検証が求められよう。同様に，このような百科事典的知識（folk-knowledge; Wierzbicka 1996）は，Natural Semantic Metalanguage (NSM) theory（e.g., Goddard and Wierzbicka 2014）においても記述される。たとえば，Wierzbicka（1985）は，dog が認識可能な形や形態的な特徴を持たないため，必要十分な特性ではなく特徴的な特性のリストによって概念が定義されるとする。この際，dog の認識可能な特徴は振る舞い（とくに，吠える・唸る・尾を振る）であり，dog は「人とともに生き，献身的で従順，信頼し得る仲間，よき学習者，勤勉な労働者である」というような，人との関係において概念化されるという。この記述で

[2] 国広（1994）は，現象素を「人間の認知作用を通して､ひとまとまりをなすものとして把握された現象」と呼ぶ。
[3]「適切な用例が見付かるとは言い難いという問題がある」と指摘する。

は，dog との関係性などの文化的な前提が必要となるほか，特徴的振る舞いに関わりのない特徴（たとえば dog の上位カテゴリの有する特徴（四足，走るのが速い，雑食など））も現れにくい。具体的に dog の「人との関係」におけるどの情報が有用であるのかという問題も解決されていないため，検証が必要であろう。

それでは，対象物を認識するための記述において必要な情報とはどのようなものなのか。辞書の語釈は不十分であるのか。不足しているとすれば何が不足なのか。また，用例から取得できる情報はどのようなもので，それらが対象物の認識に十分であるのか。本稿は，これらの疑問について被験者実験を行い，テキストから対象物を認識するにあたって有用な記述内容を探る。

3. 実験

まず，対象物を説明するテキストからどの程度対象物を同定できるのか，辞書の語釈をもとに調査する。同時に，どの記述が有用であったか，対象物が同定できなかった場合にはどんな記述があれば同定できたか訊ねる（実験 1）。次に，不足していたとされる情報の追加によって対象物の認識が可能となるのか確かめる（実験 2）。さらに，コーパスから用例を取得し，用例からテキストの示す対象物が同定できるか調べる。また，クラウドソーシングを利用することで，不足情報を検索する影響も調べる（以上実験 3）。

3.1 実験データ（実験 1・2）

まず，動物 200 種類について，国語辞書 10 種類（表 1）から語釈を収集した。

表 1　データを取得した国語辞書

辞書名	三省堂国語	新明解国語	岩波国語	明鏡国語	新選国語	集英社国語	角川国語	新潮現代国語	大辞林	デイリーコンサイス国語
出版社	三省堂	三省堂	岩波書店	大修館書店	小学館	集英社	角川書店	新潮社	三省堂	三省堂
版	5 版	6 版	5 版	初版	7 版	2 版	新版	2 版	Web 更新版	3 版
見出語	76,000 語	75,000 語	62,000 語	70,000 語	83,000 語	92,000 語	75,000 語	79,000 語	260,000 語	70,000 語

5 種以上の国語辞書に記述のある内容を以下のように分類し，記述内容を概観する。

- 分類：「イヌ科」「スズメ目」など種目
- 形態：「長い耳」「尖った口」など外観的特徴（※大きさを含める）
- 生態：「跳ねる」「托卵する」「池に棲む」など性質・動作（※「アジアに分布」のような生息や分布を含める）
- 人間との関係：「食用」など用途・「食害」など被害
- その他：上記分類外情報・フレーム知識など

たとえば，「狸」であれば，「分類：イヌ科の哺乳類」，「形態：尾が太い」，「生態：山地草原などにすむ」，「人間との関係：毛皮は防寒用。剛毛は毛筆用」，「その他：人を化かすと考えられた」が得られた。

200 種類の動物語釈において，5 種以上の国語辞書に記述のあった分類毎の割合と，各語釈における分類別記述割合の平均を表 2 に示す[4]。「分類」「形態」「生態」に分類される情報が 8 割以上の動物で記述されている情報である。また，「形態」情報は語釈の中で最も多量に記述され，4 割近くの記述量が割かれている。

表 2　国語辞書における動物語釈の分類別記述

	分類	形態	生態	人間との関係	その他
動物 200 種類の語釈における記述割合[5]	96.0%	87.5%	82.0%	52.5%	44.5%
各語釈における分類別の記述割合（平均）[6]	25.6%	36.7%	24.4%	23.3%	15.8%

3.2 実験データ（実験 3）

『現代日本語書き言葉均衡コーパス』（以下 BCCWJ: Maekawa ら 2014）と Web コーパスの一つである Google 日本語 n-gram を使用し，動物の用例を収集した。収集した用例を整理し，意味的な用例を抽出した。

3.2.1 BCCWJ を用いた用例収集

BCCWJ から 10 種類の動物（一般に知識があると考えられる単語親密度 5.000 以上の鳥獣虫魚をランダムに選択した）に関する要素・用例を収集した。タヌキ・カワウソ・オットセイ・ジャガー（以上獣）・テントウムシ・カナブン（以上虫）・スズキ・カマス（以上魚）・ジュウシマツ・ナイチンゲール（以上鳥）を選んだ。「中納言」を用いて検索語（語彙素）の前後 50 文字を取得し，手作業で内容の整理を行った。なお，文意の読み取りに文字数が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整理を行った。さらに，実験協力者に提示するため，収集した用例が句などの場合には文へ改変したほか，意味的に同種と判断される用例については，次のように作業者の判断でまとめている。このような複数用例をまとめた例を，本稿では意味的用例と呼ぶ。

[4] 国語辞書では日本人が一般的に知っているはずの常識的情報が記述されていない可能性が考えられたため，単語親密度（天野・近藤編 1999）の高低によって，記述される要素に差が見られると期待された。しかし，単語親密度と語釈の記述に顕著な特徴は見られなかった（補表 1 参照）。

補表 1　単語親密度と国語辞書における分類別語釈記述

単語親密度	動物種類(数)	形態	生態	人間との関係	その他
6.000 〜 7.000	78	35.7%	21.9%	26.5%	14.4%
5.000 〜	75	34.1%	27.5%	19.4%	17.9%
4.000 〜	23	43.4%	23.6%	22.1%	11.7%
3.000 〜	12	42.0%	21.2%	15.1%	30.2%
1.000 〜	12	42.5%	26.9%	35.3%	11.7%

[5] 当該分類における要素の記述があった動物数／ 200 種類
[6]（各動物の当該分類における要素数／各動物の全記述要素数）の合計／ 200 種類

・取得用例

(6)　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～　(OW2X_00172:「環境白書」)

(7)　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～　(PB24_00012: 中村幸昭「鳥羽水族館館長のジョーク箱」)

(8)　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。　(PM51_01452: 丹野清志「やさい畑」)

・意味的用例（(6)（7）(8）のまとめ例）

(9)　カモシカによる林業（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。

以下の表 3 に，各動物の検索結果（ヒット数・サンプル数）と意味的にまとめた結果（意味的用例数）を示す。同一サンプル内で複数ヒットする場合や，一つの用例に意味的用例が複数含まれる場合もあるため，必ずしも整理した意味的用例数が検索結果よりも少なくなるのではない。

表 3　BCCWJ から取得した用例数

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

なお，取得した意味的用例を 3.1 と同様に分類した結果が表 4 である。辞書語釈における記述（表 2，2 行目）と比べると，「分類」と「形態」が少なく，「人間との関係」情報が多く取得できていることがわかる。

表 4　BCCWJ から取得した動物 10 種の用例の分類別割合

	分類	形態	生態	人間との関係	その他
分類別の意味的用例割合（平均）[7]	7.4%	11.0%	21.5%	40.6%	19.6%

3.2.2 Google 日本語 n-gram を用いた用例収集

Google 日本語 n-gram（Web から抽出された約 200 億文（約 2550 億語）[8] の日本語データ）を用

[7]（各動物の当該分類における意味的用例数／各動物の全意味的用例数）の合計／ 10 種類
[8] 総単語数は 255,198,240,937，総文数は 20,036,793,177。

いて，取得可能な用例を調査した。

5 種類の動物（獣 2･鳥 1･虫 1･魚 1）について用例（n-gram データ（1 ～ 7 gram）･頻度 20 以上）の抽出と収集を行った。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），カナブン（異表記「かなぶん」を含む：105,200 件），スズキ（表記「鱸」：65,000 件）を取得した。

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，本コーパスが n-gram データであるため，文などが最大 7 gram で分割されている問題がある。そこで，用例件数をもとに手作業による整理を行い，最大例で 23 gram となる意味の把握が可能な長さとした。以下に整理例を示す。

（10）　老け顔アンパンマンおばさん狸顔だよね
（11）　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　（以上，Google 日本語 n-gram）

さらに，（12）（13）のような同内容と考えられる用例を，意味的用例（14）としてまとめた。

（12）　サックスのレース & 可愛いオットセイ柄のブラ & ショーツ
（13）　オットセイ柄のカットソー & スパッツ　（以上，Google 日本語 n-gram）
（14）　衣類の柄に用いられることがある　（（12）（13）などをまとめた例：意味的用例）

Web ベースの大規模コーパスにおいては，動物に関する用例を収集すると，固有の表現が多くを占める。固有名詞（個人のハンドルネーム・店名・商品名・キャラクター名など）と判断される用例は「固有名詞が多い」として対象物とは別扱いにした。完全な分類は困難であるが，作業者の判断によった。このほか，同 URL から重複取得されている用例[9]や，商品紹介など固有の表現の重複を除くと，用例数は取得数の 13% 程度の量となった。取得した用例数を表 5 に示す。

[9] 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」において「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」が 1,210 件あるが，これらの語が共起するのは，特定のブログ（http://suzusuzu.jugem.jp/）における説明部分の影響によることが確かめられた。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。その他，「タヌキ」では書籍タイトル「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）」などの種類も見られた。また，「【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，特定の質疑や説明等が別 URL から取得されている可能性のある場合もあった。これらの重複例は，「（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ている。重複ページの多さなどによって，本稿で示した Google 日本語 n-gram から取得した用例の頻度情報には均衡性が欠ける。

表 5　Google 日本語 n-gram から取得した用例数

	検索結果	意味的用例数
タヌキ	1,893,000	28
カナブン	105,200	21
スズキ	65,000	12
オットセイ	61,800	16
ジュウシマツ	33,500	11

取得した意味的用例を 3.1 と同様に分類すると，表 6 となった。取得した意味的用例の分布は BCCWJ（表 4）と類しているが，「人間との関係」情報が多いほか，「その他」に分類せざるを得ない情報が多く取得される傾向が見られる。

表 6　Google 日本語 n-gram から取得した動物 5 種の用例の分類別割合

	分類	形態	生態	人間との関係	その他
分類別の意味的用例割合（平均）	3.1%	6.7%	19.0%	43.4%	27.8%

3.2.3 BCCWJ 用例と Google 日本語 n-gram 用例の差異

本稿の調査に用いた 2 種類のコーパスからは，それぞれ同じ意味的用例が取得できているのではない。BCCWJ から取得した用例と Google 日本語 n-gram から取得した用例にどのような差が見られるのかまとめておく。

Google 日本語 n-gram 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という問題がある。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。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。反対に，BCCWJ はサンプルによっては前後の文脈が十分に取得でき（例（15）参照），意味的な判断が可能となる場合が多い。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（例（16）参照）。

（15）　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：BCCWJ，下線部は Google 日本語 n-gram からも取得された）

（16）　平安時代は猫を用字「狸」で表していた。　（タヌキ：BCCWJ）

また，コーパスの規模によって取得可能となる用例に差異が生じる。単純に意味的な用例の種類が増えるというものもある（例（17）参照）ほか，共起情報の頻度が得られることで「〜が多い」という情報が取得可能となる（例（18）（19）参照）点で異なってくる。取得された用例が個別的であるか一般的であるかが，頻度情報によって分類可能となるためである。但し，BCCWJ のような均衡コーパスと異なり，Google 日本語 n-gram のような Web コーパスからは個人的経験・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は偏りが生じる（例（20）参

照）という制限もある。

(17)　うどんやそばにタヌキの名がついた種類がある[10]。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：Google 日本語 n-gram）

(18)　タヌキに喩えるのは，特に中年以上の男性や猫が多い。（タヌキ：Google 日本語 n-gram）

(19)　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：Google 日本語 n-gram）

(20)　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：Google 日本語 n-gram）

そのほか，本稿で使用した BCCWJ・Google 日本語 n-gram 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（例（21））。

(21)　世界じゅうのジュウシマツの展覧会が行われる。（ジュウシマツ：BCCWJ）

3.3 実験 1

対象物を説明するテキストからどの程度対象物を認識できるのか，対象物の認識にどのような情報が有用であるのか，どんな記述があれば認識できるのかを調査する。実験協力者は，提示された辞書の語釈が何のものであるか同定し，どの情報を用いて判断を行ったか，不足していた場合は何が記述されていれば正答できたか答える。

3.3.1 実験 1 の手順

3.1 で取得したデータのうち，単語親密度（天野・近藤編 1999, 天野・近藤・笠原編 2008）が 5.000 以上の動物を 20 種ランダムで選んだ（表 7 参照）。実験協力者は 30 代〜 50 代の男女（日本語母語話者）20 名であり，実験室で行った。動物名は「この動物」などとマスクした。

まず，実験協力者は，提示した記述から何についての説明であるのかを読み取り，テキストの示す対象物を回答する。これにより，正答率の評価を行う。

たとえば，ライオンでは，「ネコ科の猛獣である。」「黄褐色である。雄はたてがみがある。」「アフリカに分布している。」「百獣の王と呼ばれる。」という情報が提示される。実験協力者は，回答するために有用だった情報にマークする。

次に，正答を実験協力者に提示する。この際，当該動物に関する知識がなければ正答はあり得ないため，実験協力者の当該動物の知識有無を確認し，知識率を評価する。よく知っていた場合（100%），自信はないが知っていた場合（50%），まったく知らない場合（0%）の選択回答とした。誤答の場合，実験協力者は正答するために不足だった（記述があれば正答できたはずの）情報を自由に記述する。

[10] 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。

3.3.2 実験 1 の結果

結果を表 7 に示す。100% の正答率が得られた動物は，ライオンのみであった。20 種類の平均正答率は 52% にとどまる。なお，単語親密度の高さと対象物の知識が一致しないナイチンゲールの例が見られたが，実験協力者の知識率は平均 87% であり，100% の知識率（よく知られている）動物は，ライオン・キツネ・ロバ・オオカミ・ウサギ・テントウムシ・タヌキ・ウシ・エビの 9 種類あった。とくに，マムシ・エビ・オットセイは 85% 以上という高い知識率の動物であるが，正答率は 20% 未満である。提示された情報量では不足していたことが考えられる。

当該動物をテキストから同定するにあたり，主に「形態」情報，記述があれば「その他」の情報を利用することで正答を得られる傾向が見られた。

また，20 種類の動物それぞれについて，どんな情報が不足していたために正答が得られなかったのか，表 8 に「これがあれば正答できたはずである」と求められた情報を分類した。

表 7　実験 1 における正答率と対象物の知識率，正答に有用とされた情報（「—」は掲示情報なし）

	正答率	知識率	分類	形態	生態	人間との関係	その他
ライオン	100%	100%	10%	40%	10%	—	40%
キツネ	90%	100%	6%	33%	—	—	61%
ロバ	80%	100%	0%	56%	6%	38%	0%
ヤモリ	80%	88%	7%	43%	50%	—	—
オオカミ	75%	100%	7%	13%	33%	47%	—
ウサギ	75%	100%	7%	53%	27%	13%	—
テントウムシ	70%	100%	21%	57%	0%	21%	—
タヌキ	65%	100%	0%	0%	0%	8%	92%
ウシ	65%	100%	23%	23%	—	54%	—
カワウソ	57%	88%	0%	40%	10%	—	50%
カモシカ	53%	85%	0%	11%	22%	—	67%
ナイチンゲール	50%	20%	0%	—	100%	—	—
スズキ	35%	85%	50%	0%	0%	17%	33%
カマス	29%	70%	75%	0%	0%	25%	—
ジュウシマツ	25%	80%	0%	25%	—	75%	—
カナブン	24%	83%	0%	100%	—	—	—
ジャガー	19%	78%	0%	33%	67%	—	—
マムシ	18%	85%	33%	67%	—	—	—
エビ	15%	100%	0%	100%	—	0%	—
オットセイ	11%	88%	50%	50%	0%	0%	—
平均	52%	87%	14%	39%	23%	27%	49%

表 8　対象物認識にあたり，辞書語釈に不足していた（追加が求められた）情報

	分類	形態	生態	人間との関係	その他
不足情報	0%	28%	18%	32%	22%

すべての動物に対して「これがあれば正答できる」という情報が得られた[11]。また，不足して

[11] 100% 知識があっても正答率が 15% と低かったエビでは，全体の 16% にあたる多くの量（要素数）の記述があった。

いたとされる情報は「人間との関係」が 32% と高く，対象物の認識に有用とされた「形態」が 28%，「その他」が 22% と，「分類」以外で追加の情報が求められる傾向が見える。

3.4 実験 2

辞書の語釈に不足していたとされる情報を追加することで，対象物の認識が可能となるのか確かめる。

3.4.1 実験 2 の手順

対象物を認識するために求められた情報は有用か。先の実験 1 と同条件（但し，実験協力者は異なる）で実験を行うが，提示する情報に実験 1 で得られた情報を追加する。追加する情報は，複数人が記述した内容を整理したものである。たとえばライオンでは，実験 1 で提示した「ネコ科の猛獣である。」「黄褐色である。雄はたてがみがある。」「アフリカに分布している。」「百獣の王と呼ばれる。」という情報に，「多摩動物園ではバスで間近に見ることができる。」という情報が加わることになる。動物名はマスクし，「この動物」などとした。実験室において対面で実施した。

実験 1 で使用した 20 種のデータ（表 7 参照）のうち，80% 以上の正答率が得られた動物と実験協力者の知識率が 70% 以下であった動物を除き，オオカミ（実験 1 の正答率：75%），テントウムシ（以下同 70%），タヌキ（65%），カワウソ（57%），スズキ（35%），ジュウシマツ（25%），カナブン（24%），ジャガー（19%），マムシ（18%），エビ（15%）の 10 種を対象とした。

実験協力者は 30 代～ 50 代の男女（日本語母語話者）20 名である。実験協力者は，提示した記述から何についての説明であるのかを読み取り，テキストの示す対象物を回答し，回答に有用だった情報にマークする。

3.4.2 実験 2 の結果

実験 1 で平均 40% の正答率であった動物群であるが，実験 2 では平均 65% の正答率へと大きく上昇が見られた（表 9）。とくに，オオカミ・テントウムシ・タヌキ・マムシ・エビは，85% 以上の高い正答率となっている。しかし，情報を加えても正答率に改善の見られない種類もある。

表 9　実験 1・2 における正答率

	オオカミ	テントウムシ	タヌキ	カワウソ	スズキ	ジュウシマツ	カナブン	ジャガー	マムシ	エビ	平均
知識率	100%	100%	100%	88%	85%	80%	83%	78%	85%	100%	90%
実験 1（辞書語釈）	75%	70%	65%	57%	35%	25%	24%	19%	18%	15%	40%
実験 2（辞書語釈＋追加情報）	85%	95%	90%	40%	30%	15%	65%	45%	95%	90%	65%

実験 1 と 2 の正答率の変化は 3 種類に大別できる（図 1）。グループ 1 は，実験 2 で高い正答率が得られているが，もともと実験 1 でも 65% 以上の正答率が得られていた動物である。実験 1・2 ともに高い正答率が得られているグループであるといえる。グループ 2 は，正答するために求められた情報が加わったにもかかわらず，低い正答率にとどまったグループである。グループ 3 は，追加情報によって正答率が大きく上昇したグループである。

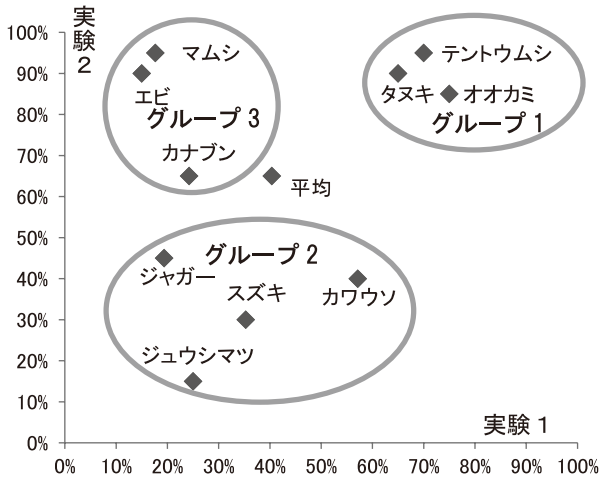

図 1　実験 1・2 における正答率

テキストの示す対象物を認識するために有用とされた情報を，実験 1 との対照で以下の表 10 に示す。実験 1 でほぼ使用されることがなく，求められた情報が全くなかった「分類」は実験 2 でも変化がないため省略する。

表 10 から，全般に，実験 1 で求められた「人間との関係」と「その他」の追加情報が利用されていることがわかる。概ね，実験 1 で利用された「形態」と「その他」に加え，「人間との関係」についての情報が，対象物の認識に有用とされていた。

表 10　実験 1・2 における正答に有用とされた情報

	正答率		形態		生態		人間との関係		その他	
	実験 1	実験 2	実験 1	実験 2	実験 1	実験 2	実験 1	実験 2	実験 1	実験 2
オオカミ	75%	85%	13%	65%	33%	50%	47%	30%	—	60%
テントウムシ	70%	95%	57%	60%	0%	10%	21%	45%	—	70%
タヌキ	65%	90%	0%	60%	0%	30%	8%	35%	92%	90%
カワウソ	57%	40%	40%	65%	10%	30%	—	20%	50%	20%
スズキ	35%	30%	0%	0%	0%	0%	17%	60%	33%	15%
ジュウシマツ	25%	15%	25%	45%	—	10%	75%	65%	—	10%
カナブン	24%	65%	100%	55%	—	15%	—	—	—	65%
ジャガー	19%	45%	33%	50%	67%	15%	—	—	—	55%
マムシ	18%	95%	67%	15%	—	55%	—	85%	—	10%
エビ	15%	90%	100%	60%	—	5%	0%	70%	—	65%
平均	40%	65%	44%	48%	18%	22%	28%	51%	58%	46%

3.5 実験 3

コーパスから取得した用例から，テキストの示す対象物が同定できるか調べる。辞書語釈よりも豊富な情報が取得できるとすれば，対象物の認識に有用となる情報が取得できている可能性が期待される。実際に，3.3 で見たように先の実験で有用とされた「人間との関係」と「その他」情報が多く取得できている。コーパスから取得した用例によって，対象物を認識するために十分な情報が得られるだろうか。

また，本実験は，実験室ではなくクラウドソーシングを利用したオンライン実験を行うこととした。オンライン実験では，個人の有する知識に限定されず，検索エンジン等を使用して自由に対象物について調べることが可能である。そのため，記述をもとに自発的な検索等を行うことでアクセスしやすい情報という観点でも，対象物を認識するために有用な情報を調査する。

3.5.1 実験 3 の手順

3.2 に示した BCCWJ と Google 日本語 n-gram から収集した意味的用例をデータとして用いた。

実験は，実験 1・2 と同様の手法によって行う。動物名はマスクし，「この動物」などとした。実験協力者は，提示した記述から何についての説明であるのかを読み取り，テキストの示す対象物を回答する。これにより，正答率の評価を行う。また，実験協力者は，回答するために有用だった情報にマークする。

本実験においては，Yahoo! クラウドソーシングを用いて募集した実験協力者（15 歳以上の男女）1,000 名の回答を得た [12]。

3.5.2 実験 3 の結果

BCCWJ 用例からの対象物同定（10 種）は平均正答率 50.5%，Google 日本語 n-gram 用例からの対象物同定（5 種）は平均正答率 64.1% となった（表 11）。コーパスから取得した用例で必ずしも対象物を十分に認識できるとはいえない。

[12] 実験 1･2 と同様に，30 代〜 40 代の男女（日本語母語話者）12 名の回答を得た結果との対照を補表 2 に示す。クラウドソーシング実験で正答率が上昇する傾向がある。オンライン実験では自由な情報検索が可能であるため，上位頻度の要素は検索サービスでヒットしやすい可能性が推測される。実際に，有用とされた情報は，実験室では使用されない情報であるという差異が生じている。

補表 2　同材料における実験室とクラウドソーシングの正答率

	タヌキ	テントウムシ	カワウソ	カナブン	スズキ	オットセイ	ジャガー	カマス	ジュウシマツ	ナイチンゲール	平均
実験室（男女 12 名）	100%	75.0%	58.3%	25.0%	16.7%	8.3%	8.3%	8.3%	0%	0%	25.0%
クラウドソーシング（1,000 名）	96.8%	85.0%	56.8%	64.3%	78.3%	48.1%	41.3%	1.6%	8.2%	24.3%	50.5%

表 11　実験 3 における正答率

	タヌキ	スズキ	カナブン	オットセイ	ジュウシマツ	テントウムシ	カワウソ	ジャガー	ナイチンゲール	カマス	平均
BCCWJ	96.8%	78.3%	64.3%	48.1%	8.2%	85.0%	56.8%	41.3%	24.3%	1.6%	50.5%
Google	97.0%	77.8%	74.6%	58.2%	12.9%	—	—	—	—	—	64.1%

また，個別の動物の正答率で見ると 2 種のコーパスに大きな差はないが，タヌキ・カナブン・オットセイ・ジュウシマツにおいて，Google 日本語 n-gram 用例による対象物同定で若干正答率の上回る傾向が得られている [13]。

Google 日本語 n-gram の用例は検索エンジン Google を使用した検索を行った際ヒットしやすい性質があり，本実験がオンライン実験であることから，回答のための情報が取得しやすかった可能性が考えられる（詳細については 4.3.3 で考察を行う）。

テキストの示す対象物を認識するために有用とされた情報を，BCCWJ と Google 日本語 n-gram 用例が対照可能な 5 種の動物について表 12 に示す。5 種の動物を平均すると，2 種のコーパスから取得された用例は，分類別に大きな差がないように見える。提示した情報数が多い（3.2.1 参照：最大 41 例）ため，特定の用例のみが有用とされていた場合，分類としての割合が低下しているためである（具体的な個別の用例については 4 節で議論する）。しかし，実験 1・2 との差異として，たとえば BCCWJ 用例として取得されたオットセイの「アシカ科の動物である」という情報が 99.8% の実験協力者（正答）に有用であるとされたなど，「分類」情報が用例から取得されている際には平均 4 割を超えて有用とされるという特徴が見られる。これは，クラウドソーシング実験を利用したため，個人の有する知識としての「分類」情報ではなく，検索によるカテゴリの絞り込みに有用だったと推測される（詳細は 4.3.3）。

表 12　実験 3 において正答に有用とされた BCCWJ・Google 日本語 n-gram 用例

	タヌキ		スズキ		カナブン		オットセイ		ジュウシマツ		平均	
	BCCWJ	Google	BCCWJ	Google	BCCWJ	Google	BCCWJ	Google	BCCWJ	Google	BCCWJ	Google
分類	26%	—	—	—	35%	40%	100%	29%	15%	52%	44%	40%
形態	33%	18%	4%	—	28%	28%	29%	—	27%	17%	24%	21%
生態	12%	22%	—	10%	7%	14%	43%	18%	30%	11%	23%	15%
人間との関係	19%	29%	12%	10%	18%	11%	18%	15%	35%	20%	20%	17%
その他	22%	19%	48%	34%	25%	13%	13%	21%	2%	31%	22%	24%

[13] 注 12 で示した通り，検索サービスを利用した用例の検索により，Google 日本語 n-gram 情報がもともと Web データであるためにヒットしやすい可能性が考えられる。

3.6 実験まとめ

ここまでの実験について，すべての実験で調査に用いた対象物 4 種（鳥獣虫魚各 1 例ずつ）の結果を表 13 にまとめる。本表では，対照のために実験 3 の予備実験として行った，BCCWJ 用例からの対象物同定実験を実験室で行った [14] 結果を含めて示す。

表 13　実験 1 〜 3 の正答率と正答に有用な情報

対象物	知識率	分類名	辞書語釈（実験室）	求められた情報（実験室）	BCCWJ（実験室）	BCCWJ（クラウドソーシング）	Google（クラウドソーシング）	平均
タヌキ	100%	正答率	65%	90%	100%	97%	97%	90%
		分類	0%	—	25%	26%	—	17%
		形態	0%	60%	14%	33%	18%	25%
		生態	0%	30%	6%	12%	22%	14%
		人間との関係	8%	35%	14%	19%	29%	21%
		その他	92%	90%	21%	22%	19%	49%
スズキ	85%	正答率	35%	30%	17%	78%	78%	48%
		分類	50%	—	—	—	—	50%
		形態	0%	0%	67%	4%	—	18%
		生態	0%	0%	—	—	10%	3%
		人間との関係	17%	60%	21%	12%	10%	24%
		その他	33%	15%	22%	48%	34%	30%
カナブン	83%	正答率	24%	65%	25%	64%	75%	51%
		分類	0%	—	33%	35%	40%	27%
		形態	100%	55%	33%	28%	28%	49%
		生態	—	15%	11%	7%	14%	12%
		人間との関係	—	—	0%	18%	11%	10%
		その他	—	65%	22%	25%	13%	31%
ジュウシマツ	80%	正答率	25%	15%	0%	8%	13%	12%
		分類	0%	—	0%	15%	52%	17%
		形態	25%	45%	0%	27%	17%	23%
		生態	—	10%	0%	30%	11%	13%
		人間との関係	75%	65%	0%	35%	20%	39%
		その他	—	10%	0%	2%	31%	11%
平均	87%	正答率	37%	50%	36%	62%	66%	50%
		分類	14%	—	19%	25%	46%	26%
		形態	39%	40%	28%	23%	21%	30%
		生態	20%	14%	6%	16%	14%	14%
		人間との関係	27%	53%	9%	21%	18%	26%
		その他	54%	45%	16%	24%	24%	33%

テキストからの対象物同定は，これまでに行った 3 種の実験（実験 1：半数以上の辞書に記述のある語釈，実験 2:辞書語釈の不足情報を追加した記述，実験 3:コーパスから取得される用例）

[14] 実験とその結果については注 12 を参照。

すべてにおいて, 高い知識（表 13 の 4 種ではすべて 80% 以上;平均 87%）を有している動物に限っても，正答率の平均は実験室では 5 割，検索エンジンの使用が可能な環境となるクラウドソーシングであっても 7 割未満にとどまった。対象物の知識を有していればテキストから対象物を認識することができるとはいいがたい。

また, 正答に有用とされた情報は, 平均して「その他」と「形態」情報の割合が高い傾向があり,「人間との関係」が次ぐ。但し，クラウドソーシング実験では，実験室において対面で行う場合と異なり,「分類」情報が活用される傾向が見られる。

4. 考察

テキストから対象物を認識するにあたり，どのような記述が必要なのか。実験 1 〜 3 の結果をもとに，以下について考察を行う。

1. テキストから対象物を認識するために，辞書の記述では何が不足とされたか。何が記述されていればテキストからの対象物同定が可能なのか。（実験 1）
2. この記述があれば対象物が認識できるとされた情報を加えたテキストでも，対象物の同定ができないことがあるのはなぜか。何がまだ不足か。（実験 2）
3. 対象物を認識するために必要な情報がテキストから取得可能か。コーパスからどのような情報が取得しやすいか，あるいは取得しにくいか。不足していた知識を，記述された情報に基づく検索から補えるか。（実験 3）

4.1 考察 1：実験 1 結果に基づく考察

テキストから対象物を認識するために，辞書の記述では何が不足とされたか。

実験 1（3.3）において，テキストからの対象物同定にあたって実験協力者が利用していた情報は,「その他」と「形態」に分類された情報であった。しかし，正答できなかった場合には「人間との関係」と「形態」に分類される情報が不足していたとされる傾向があった。

では，具体的に何が記述されていればテキストからの対象物同定が可能なのか。

本節においては，対象物の認識にあたり辞書の語釈に不足していたとされる個々の情報を分析することで，どのような情報が対象物の認識に有用であるのか考察する。

4.1.1 個人の経験知識に関する情報

対象物を想起するために，具体的外観的な情報として「形態」に分類される情報は有用であろう。実際,「形態」に分類される情報が有用とされる割合は，これまでの実験すべてにおいて平均的に高い（表 13 など参照）といえる。しかし，テキストからの対象物認識には,「その他」に分類される情報が有用とされる傾向があった（表 7)。実験 1 で正答率が高いライオン（100%）やキツネ（90%）などは「その他」情報が用いられている（それぞれ 40% と 61%)。次いで正答率の高いロバ（80%）は,「形態」(56%）が最も有用とされているが「人間との関係」(38%）が次いで有用とされた。具体例を見ると，以下のような情報であった。

・キツネ：稲荷神の使いとされる（「その他」）
・ライオン：百獣の王と呼ばれる（「その他」）
・ロバ：農耕や運搬に用いる（「人間との関係」）

また，記述の求められた情報のうち「人間との関係」に分類される情報は，「どこで見ることができる」「どのように食べる」などの，個人の具体的な経験に関するものであった。同様に物語名や対象物をモチーフにした商品などの「その他」に分類される情報についても，文化的に個人の経験に関わる知識であるといえる。

たとえば，以下のような情報が対象物を認識するための記述に求められた。

・スズキ：お造り・寿司・カルパッチョ・グリルなどの料理名（「人間との関係」）
・オオカミ：「赤ずきん」に出てくる，悪役など（「その他」）
・オットセイ：水族館で見ることができる（「人間との関係」）
・ジャガー：このマークの自動車がある（「その他」）

これらは，個人的な経験や知識を喚起するために有用と考えられる情報である。具体的な料理名や場所名（寿司・水族館など），物語名や商標（赤ずきん・自動車）が示されることで，実験協力者が予め対象物の知識として個別に持つ経験と合致すれば，対象物が認識可能となる。

4.1.2 他メンバーとの差別化を行うための情報

正答率の高いロバ（80%）においては，「形態」（56%）が最も有用とされていた。実験 1 で提示したロバの「形態」情報は「ウマより小さい。耳が長い」であった。ロバでは，4.1.1 で見た「人間との関係」で絞り込まれたカテゴリのメンバーとしてウマが推測されるが，「ウマより小さい」が正答は「ウマではない」とウマを排除する情報であったため，有用とされたのであろう。

誤答と求められた情報の間には類似した傾向が見られた。たとえば，ウサギの誤答はカンガルーであった（複数の実験協力者の回答）。誤答した実験協力者がカンガルーを排除するために求めたのは「小さい」という情報である。これは，対象物の属する臨時的カテゴリ（Barsalou 1983）を考えるとき，同じカテゴリ（Taylor 1995 など）に属するメンバー間のいずれか判断しかねたために求められた情報であると推測される。

実験 1 で提示したウサギの情報は，以下である。

・哺乳類である。
・長い耳と長い後肢を持つ。
・よくはねる。
・毛皮を利用する。肉は食用である。

誤答のカンガルーは，実験 1 と同手法（3.1）で提示するならば以下であった。上記のウサギと重なる情報を下線で示す。

・哺乳類・有袋類である。
・オーストラリアやニューギニアに生息している。
・雌は腹にある袋に子を入れて育てる。
・大きな長い後肢と尾を持つ。
・よくはねる。

ウサギとカンガルーはどちらも「長い後肢」を持ち「よくはねる」「哺乳類」である。また，辞書の記述にないが，カンガルーは「長い耳」を持つことも排除しないであろう。実験協力者は提示された記述情報と保有する経験知識とをつきあわせることで想定されるカテゴリを狭めて行く。そのため，最終的なカテゴリメンバーにウサギとカンガルーがあったとき，ウサギとカンガルーを差別化するための情報が求められることになる。そこで，カンガルーと誤答した実験協力者の求める情報は「小さい」であったと考えられる。

このほかに，他メンバーとの差別化に用いると考えられる情報には，以下のような例があった。

・テントウムシ（誤答例：コガネムシ）：赤い斑点・星（模様）があるなど（「形態」）
・エビ（誤答例：カニ）：カニのライバル（「その他」），背が曲がっているなど（「形態」）
・カモシカ（誤答例：ヤギ）：すらっとした・きれいな脚があるなど（「形態」）
・タヌキ（誤答例：キツネ）：腹が出ている，腹に特徴があるなど（「形態」）

大きさや斑紋のパターンをはじめ，誤答を提示して否定することや，誤答との差異情報（カニとの差異としてエビは「背が曲がっている」，キツネとの差異としてタヌキは「腹に特徴がある」など）を示すことが求められていたのである。

4.1.3 対象物を認識するために求められる記述

本稿の実験 1 の結果から，対象物を認識するにあたり，個人的な経験や知識を喚起する情報（4.1.1）と対象物の属する臨時的カテゴリの他メンバーとの差異情報（4.1.2）が求められる傾向があるとわかった。

個人的な経験や知識を喚起する情報とは，「人間との関係」に関わることが多いという点で Wierzbicka（e.g. 1985）と類し，国広（1997）の示した「連想」記述にも類するものである。但し，本稿の実験では，料理名や場所名，物語名，商標などのように具体的であり，文化的にも一般的な情報であることが求められた。この種の「人間との関係」「その他」に分類される情報は，コーパスから取得されやすく（3.2.1, 3.2.2），一般的であることは頻度とも関係しやすいと推察される。

また，読み手の想定する臨時的なカテゴリにおける他メンバーとの差異とは，対象物に特徴的な情報を示すというよりも，想定されるカテゴリにおけるメンバーが等しく有する特徴とは異なる部分を示すということであった。読み手の想定する臨時的なカテゴリにおける他メンバーとの関係については，次節（4.2）でも考察する。

4.2 考察 2：実験 2 結果に基づく考察

実験 1 で得た対象物を認識するに十分となるはずの情報を加えたテキストでも，対象物の同定ができない場合があった（表 9）。実験 2 において，3.4.2 の図 1 に示したグループ 2 は，正答するために必要とされた情報が加わったにもかかわらず正答率が低い。

ここでは，求められた情報を追加した記述に何がまだ不足していたのか分析し，対象物を認識するためのテキスト記述について考察を深める。

4.2.1 対象物の知識の不足

図 1 のグループ 2 に含まれる動物の知識率（カワウソ：88%，ジャガー：78%，ジュウシマツ：80%，スズキ 85%）は，表 7 の知識率が平均 87% であることを見るに，他グループに比べて僅かに低い傾向がある。この知識率は，多くの実験協力者が「知っている」認識であったとしても，中には「自信がない」と答えた実験協力者もいたということであり，実験協力者に対象物に関する十分な知識がなかったために，提示された情報から対象物が同定できなかった可能性が考えられる。

図 2 はスズキが正答である。

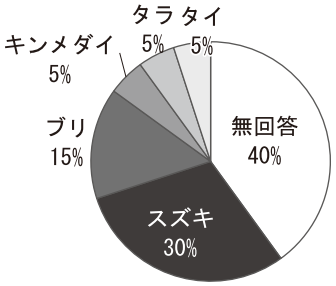

図 2　実験 2 におけるスズキ（正答）と誤答

図 2 では，最も多い回答が無回答（40%）である。実験協力者はスズキがどのような魚か知っている（知識率 85%）つもりであっても，一般に接するのは切り身など食材としてのスズキであり，求められて追加した情報は「白身で柔らかくあっさり」「寿司」「刺身」など食材としての「人間との関係」である。よって，「形態」や「生態」「その他」に関する具体的な知識（「セイゴ・フッコと名の変わる出世魚である」「口が大きい」「近海魚」など）が知識として保有されていないために，そもそもテキスト情報から特定ができなかった可能性が考えられる。なお，「出世魚」であることから，「出世魚」カテゴリのメンバーとしてブリの誤答が次いだようである。このように，対象物の知識が限定的か曖昧である場合，「食材の魚」のような大きなカテゴリの下位カテゴリへと絞り込むことが困難ということであろう。

4.2.2 対象物を絞り込む知識の不足

では，対象物同定において誤答の割合が多い場合はどのような原因によるか。図 3 にジュウシマツの回答を示す。図 3 では，過半数を上回る 60% がブンチョウの誤答である。

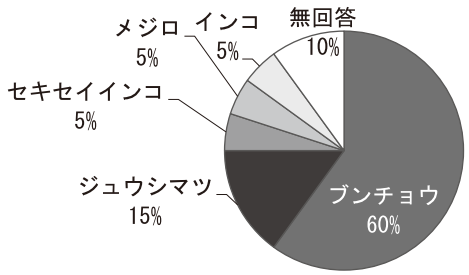

図 3　実験 2 におけるジュウシマツ（正答）と誤答

ブンチョウとジュウシマツは，概ね色味が少ない以外は，外観上に似た特徴があるわけではない。実験 2 では，ジュウシマツの「形態」として，「スズメよりやや小さく小形。羽色は豊富であるが，主に白く，茶の不定紋がある。」という情報が提示されている。「主に白く，茶の不定紋がある」は実験 1 で求められて追加した情報である。これらの情報はブンチョウの外観とそぐわず，ジュウシマツの知識があればブンチョウとの差別化が可能となることが期待された。

また，「人間との関係」における「飼い鳥」「手乗りにもできる」などの情報から「ペットの小鳥」というカテゴリが想定されたとも推測される。そこで，「ペットの小鳥（主に白く，茶の不定紋がある）」カテゴリにおいて，限定的か曖昧な知識しか有していなかった場合，同カテゴリのプロトタイプ（Rosch 1973, 1975ab, 1978, Rosch and Mervis 1975, 1981, Rosch ら 1976 など）としてブンチョウと回答した可能性が考えられる。ジュウシマツの記述から想定されたカテゴリのプロトタイプがブンチョウであった実験協力者の割合が高かったのであろう。

すなわち，対象物の属するカテゴリまで絞り込むことができない場合には，上位カテゴリのプロトタイプを回答する可能性がある。対象物の知識が限定的か曖昧であった場合に，4.2.1 で見たスズキのように「無回答」となるのでなければ，想定した臨時のカテゴリにおけるプロトタイプを回答することが考えられる。

但し，上位カテゴリのプロトタイプが回答されるとすると，提示されている「形態」情報などが無視されるという疑問が残る。ブンチョウがプロトタイプであったとすれば，ブンチョウは想定しやすいはずであり，外観の異なるジュウシマツについての「形態（茶の不定紋）」に関する記述は無視されたことになるからである。この原因としては，「ペットの小鳥（主に白い）」に関する知識が実際に小鳥を飼育する一部の人々のほかには一般的に得にくいものであり，一般にはブンチョウについての「形態」情報が曖昧であって，ブンチョウの名前や「ペットの小鳥（主に白い）」という知識のみであった可能性も考えられる [15]。よって，「ブンチョウ」という名を持つ

[15] 3.3.2 で見た単語親密度の高さと対象物の知識に大差のあるナイチンゲールの例もあり，現代日本では一般に名前（単語）を知っていても対象物の知識とは差のある鳥類は多いと推測される。

が形態の曖昧な「ペットの小鳥（主に白い）」は，ジュウシマツ固有の情報（茶の不定紋）を排除せず，誤答としてブンチョウが 60 % も回答されることとなったのであろう。

4.2.3 対象物を差別化する知識の不足

実験協力者の知識が曖昧で対象物を認識できなかった場合に，4.2.1 で見た無回答や 4.2.2 で見た上位カテゴリのプロトタイプという回答ではなく，誤答にバリエーションの見られる例もあった。図 4 にジャガーの回答を示す。

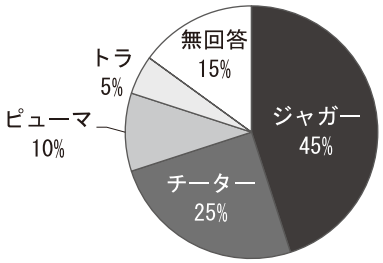

図 4　実験 2 におけるジャガー（正答）と誤答

「ネコ科の肉食獣」「ヒョウに似た斑紋がある」と対象物の属するカテゴリが絞り込まれた際に，4.2.2 同様プロトタイプが回答された結果，実験協力者毎にプロトタイプが個別的であったためにバリエーションが生じたとも考えられるが，絞り込まれたカテゴリに属するメンバーの差別化まではできなかった例と見ることもできる。ジャガーの例では，4.2.2 とは異なり「ヒョウに似た斑紋がある」など対象物の「形態」情報を用いて想定されるカテゴリの絞り込みが行われているためである。ジャガーという対象物についての認識は，一般にスズキやジュウシマツなどよりも高いと推測される。しかし，たとえば斑紋の種類などの詳細な知識までは有していない場合があり，その際実験協力者は，個人の有する曖昧な知識と一致するメンバーを，想定したカテゴリから選択したために，誤答が生じたのであろう。

4.2.4 対象物の認識に不足する知識

テキストの記述から対象物を認識するためには，対象物に関する知識をテキストから正確に取得することが必要となる。知識の不足は，まったくイメージすることができない（4.2.1）か，記述から臨時的に想定したカテゴリの上位カテゴリのプロトタイプをイメージする（4.2.2）か，同カテゴリ内の他メンバーをイメージする（4.2.3）という結果を生ぜしめる。誤認の生じた場合には，記述内容を無視する危険もある（4.2.2）。

対象物の知識が不足している読み手のためには，誤解なく対象物を認識するべく，上位カテゴリのプロトタイプとの差異を記述することや，類似した特徴を有する例を挙げた差異の記述が有用となるであろう。

4.3 考察 3：実験 3 結果に基づく考察

対象物を認識するために必要な情報は，既存のテキストから取得可能だろうか。実験 1・2 から得られた知見を実際のテキストで検証したい。

コーパスから取得した意味的用例から対象物を同定する本稿の実験（実験 3）では，検索が可能な状況でも 5 〜 6 割の正答率という結果にとどまっており，辞書語釈からの対象物同定に比して十分とはいいがたい。但し，実験 3 で有用とされた情報は，分類別には大差なかった（3.5.2）が，BCCWJ と Google 日本語 n-gram から取得した用例が同じであったのではない（3.2.3）。ここでは 2 種のコーパスから取得された用例について，対象物の認識に有用とされた情報を個別に分析することで，コーパスから取得可能な対象物認識に有用な情報はどのようなものか（人々に求められた情報（実験 1・2）が取得可能か），あるいは対象物認識に役立ちにくい情報はどのようなものか，また不足していた知識を調べて補えるかを考察する。

4.3.1 対象物認識に有用な情報

コーパスの種別によって取得される意味的用例は異なる（3.2.3）が，対象物の認識に有用な情報も異なるのか。

表 14 に，対象物がタヌキ（2 種のコーパスでどちらも正答率が 97%）であった意味的用例について，それぞれのコーパスで有用とされたものを示した（提示は文以上の単位で行ったが，表では大意のみ簡略化して記述する。2 種のコーパスで有意差があった場合は多い側を太字で示す）。表 14 はそれぞれ上位頻度（15 位まで）で有用とされた情報に限っているが，下位（16 位以下）の情報で 2 種のコーパスともに得られていた例はなく，共通して取得可能な情報が有用とされた傾向がわかる。有用とされた頻度にも大差のない例もある。また，個別の用例を見ても，「人間との関係」「その他」に分類される用例が上位で有用とされる傾向が見られている。なお，BCCWJ から取得した用例において「キツネと比較される」というキツネではないことを示した用例が最も有用とされており，想定されるであろうカテゴリ内の他メンバーとの関係性を示すことが有用であると考えられる。

また，4.1 において考察した，対象物認識に際し辞書語釈に不足していたとされた情報（個人的な経験や知識を喚起する情報（4.1.1），対象物の属する臨時的カテゴリの他メンバーとの差異情報（4.1.2））は，表 14 における「カチカチ山（一般的経験知識の喚起）」や「キツネと比較される（他メンバーとの差別化）」などとして取得されているともいえる。

テキストから対象物の認識を試みる際に求められる情報は，コーパスの種別に関わらず，どのコーパスからもある程度は類似した傾向で取得される可能性があろう。

表 14　コーパス別対象物認識（正答）に有用とされた意味的用例
（タヌキ上位・複数回答）

	BCCWJ	Google 日本語 n-gram	有用（有意差[16]）
キツネと比較	60.20%	—	—
メニュー（そばうどん等）	39.00%	**63.70%**	あり
カチカチ山	53.60%	55.30%	なし
信楽置物	32.40%	**54.00%**	あり
化ける	49.40%	52.20%	なし
寝たふり	46.40%	46.90%	なし
皮算用	—	41.20%	—
三大伝説	—	39.40%	—
腹つづみを打つ	**53.20%**	28.60%	あり
ぶんぶく茶釜	40.90%	—	—
大きな腹・でっぷり体型	**31.80%**	24.70%	あり
八畳敷	**38.40%**	23.90%	あり
タヌキ顔・アイシャドー	11.26%	**19.90%**	あり
中年男性・猫を喩える	—	17.30%	—
ドラえもん	**29.10%**	15.60%	あり

4.3.2 対象物認識に役立ちにくい情報

対象物によっては，コーパスから取得した情報からの認識が困難で，有用な情報がコーパスからは取得しにくいと見える場合もある。実際にコーパスでは対象物認識のための情報が不足していたのか。あるいは何が正答の邪魔となったのか。誤答の際に有用とされた情報を見ておきたい。

表 15　コーパス別対象物認識に有用とされた意味的用例（ジュウシマツ上位・複数回答）

	BCCWJ				Google 日本語 n-gram		
意味的用例	利用[17]（正誤）	正答（8.2%）	誤答（91.8%）	意味的用例	利用[18]（正誤）	正答（12.9%）	誤答（87.1%）
日本で作り出した	正	63.4%	12.7%	歌に文法の本	正	53.5%	10.6%
手乗りにもなる	誤	37.8%	73.6%	小型鳥・フィンチ	正	51.9%	30.8%
飼い鳥	—	36.6%	46.3%	手乗りにもなる	誤	26.4%	64.6%
つぼ巣	—	35.4%	29.5%	複数飼い	—	19.4%	19.2%
11cm	正	26.8%	7.6%	昔飼った人が多い	誤	19.4%	38.2%
子育て上手	正	25.6%	7.7%	同じ餌であわせ飼い	—	17.8%	18.6%
多品種	—	14.6%	13.8%	小斑	正	17.1%	4.9%
女子供が珍重	—	3.7%	7.4%	展覧会	—	14.7%	8.8%

表 15 は，対象物がジュウシマツ（正答率は BCCWJ：8.2%, Google 日本語 n-gram：12.9%）であった意味的用例について，それぞれのコーパスで有用とされた情報の上位を，正答誤答別に示した（提示は文以上の単位で行ったが，表では大意のみ簡略化して記述する）。コーパスの意味的用例からジュウシマツを同定した実験（実験 3）の回答は図 5 に示す。

[16] 有意水準 0.1% 以下で頻度に有意差がある。
[17] 有意水準 0.1% 以下で頻度に有意差がある。
[18] 有意水準 0.1% 以下で頻度に有意差がある。

BCCWJ ではブンチョウ（誤答）が 60% と突出し，Google 日本語 n-gram ではブンチョウ（誤答）が 37%，次いでインコが 34% と分散されている。表 15 を見るに，誤答の際に有用とされたのはどちらのコーパスでも「手乗りにもなる」という情報が目立つ。また，Google 日本語 n-gram の提示情報においては「昔飼った」という情報がそれに次いでいることが着目される。4.2.2 で見たジュウシマツの分析と同じく，「ペットの小鳥」カテゴリを想定したとき，「手乗り」という情報によっては典型例として想起されたブンチョウは排除されない。しかし，Google 日本語 n-gram から得られた多くの「昔飼った」という「人間との関係」情報によって個別の経験知識が喚起され，実際に国内飼育数が多いと考えられる [19] インコが活性化されることになったと考えられる。3.2.3 で見たように，Google 日本語 n-gram から取得された情報は個別の経験知識に関わる種類の情報が多いこともあり，対象物の認識にあたって利用するにはカテゴリの絞り込みがミスリードされる危険を慮る必要があろう。

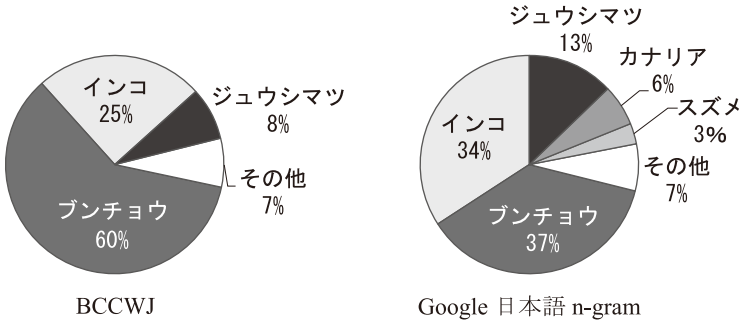

図 5　実験 3 におけるジュウシマツ（正答）と誤答

なお，正答で有用とされる割合の高かった「茶の小斑がある（Google 日本語 n-gram）」「日本で作り出した（BCCWJ）」「子育て上手（BCCWJ）」などは，ジュウシマツを「ペットの小鳥」カテゴリ内で差別化するに有用な情報であったといえる。しかし，誤答を招いた「昔飼った人が多い」のような一般的に得やすい経験とは反対に，これらは実験協力者個々人の有していた知識と合致した場合にのみ [20] 有用であったともいえる。

4.3.3 対象物認識に利用可能な情報

個人の有する知識が不足していたとしても，検索などの調査によって不足していた知識を補填し，対象物認識に役立てることの可能な情報がコーパスから取得できる可能性がある。実験 3 ではオンラインによるクラウドソーシングを用いており，実験協力者は実験中も常時自由に検索エ

[19] 国内飼育数は明確ではないが，一般的にインコの飼育数が多いと考えられ，最も飼育数が多いのはセキセイインコ（60%），次いでブンチョウ（25%）とする情報（http://www.pet-hospital.org/forvets-021.htm）もある。
[20] 内閣府の「動物愛護に関する世論調査（http://survey.gov-online.go.jp/h22/h22-doubutu/index.html，平成 22 年 9 月調査）」によれば，ペットを飼っている人の割合は 34.3%，そのうち鳥類をペットとして飼育している割合は 5.7% であり，「ペットの鳥」に関する経験知識を日常的に得ている割合が高いとはいいがたい。

ンジンなどを使用することが可能であった。

表 16 は，対象物がスズキ（正答率は BCCWJ：78.3%，Google 日本語 n-gram：77.8%）であった意味的用例について，それぞれのコーパスで有用とされた情報の上位 3 位を，正答誤答別に示した（提示は文以上の単位で行ったが，表では大意のみ簡略化して記述する）ものである。スズキは，表 13 などで見たように，実験室で行った対象物同定実験においては，正答率が 3 割程度と低かったが，オンライン実験ではほぼ 8 割という高い正答率の得られた対象物である。とくに，BCCWJ の用例から対象物を同定する実験の正答率を対照すると，実験協力者の知識にのみ頼る実験室では 17% であったにもかかわらず，検索などの可能なオンラインでは 78% と大きく差がある。この結果の差は，実験協力者が回答のために提示された情報（「出世魚」「セイゴ･フッコ」など）を用いて調査を行ったためと推測される。実際に表 16 を見ると，BCCWJ，Google 日本語 n-gram ともに，クラウドソーシング実験における正答に最も有用とされた情報は，「出世魚」であり「セイゴ・フッコ」と名の変わることであり，実験協力者の 9 割という大部分が有用と回答している。反対に，実験室における対面実験では有用とされていない。「セイゴ・フッコ」と名の変わることについては，0% とまったく利用されていなかったのである。

表 16　コーパス別対象物認識に有用とされた意味的用例（スズキ上位・複数回答）

	BCCWJ（クラウドソーシング）			BCCWJ（実験室）		Google 日本語 n-gram（クラウドソーシング）		
意味的用例	利用	正答（78.3%）	誤答（21.7%）	正答（16.7%）	意味的用例	利用 [21]	正答（77.8%）	誤答（22.2%）
セイゴ･フッコと変名	正	89.0%	11.5%	0　%	出世魚（セイゴ･フッコ）	正	94.6%	16.2%
出世魚である	―	49.0%	49.8%	16.7%	白身で淡泊な味	誤	22.5%	36.9%
白身で淡白な味	誤	22.3%	37.8%	25.0%	多くの調理法	誤	17.9%	43.7%

スズキを他のメンバーから差別化するために，たとえば「出世魚」であることはブリなどを，「白身で淡白な味」であることはタイなどを排除するに十分ではなく，「セイゴ・フッコ」と名の変わることがスズキを特定するに有用であったに違いない。しかし「セイゴ・フッコ」の名は一般に知識としては有しておらず，検索するために用いることが有用であったと考えられる。このように，検索キーとして用いた折，追加情報を取得することができ，他のメンバーと明確に差別化が可能となる情報もまた，対象物の認識には有用な情報であるといえる。

4.4 考察のまとめ：コーパスからの取得情報と対象物認識におけるその有用性

テキスト情報と対象物認識について，表 17 にコーパスからの取得可能性（テキストからの取得しやすさ・しにくさ）という点から，ここまでの議論をまとめる。

[21] 有意水準 0.1% 以下で頻度に有意差がある。

表 17　テキスト情報の取得のしやすさと対象物認識への利用

コーパスから	対象物認識に		
	役立つ	役立ちにくい	利用可能
取得しやすい	一般的経験知識を喚起	個別的経験知識を喚起	一般知識でないが特徴的
	読み手に対象物知識有	読み手に対象物知識無	追加情報の検索が可能
	(4.3.1)	(4.3.2)	(4.3.3)
取得しにくい	個別的経験知識に合致	N/A	
	対象物の差別化が可能		
	(4.3.2)		

対象物をテキストから認識するという目的において，読み手に求められる経験知識に関わる情報はコーパスから取得しやすいといえるが，予め有している知識を要求しがちであるという点において，対象物の知識が欠如している場合にはカテゴリの絞り込みに適した情報が取得しにくいと考えられる。

5. まとめ

テキストから得られる情報のみでテキストに記述された対象物を認知することがどの程度可能かという疑問について，本稿は対象物の同定実験による調査を行った。複数（5 種以上／ 10 種）の辞書に記載のあった語釈，辞書語釈に不足していた情報を追加したテキスト，コーパスから取得した意味的用例の 3 種類のテキストを用いた実験結果から，何が記述してあればテキストが示す対象物は認識できるか，また実際のテキストから取得しやすいかあるいはしにくいかという点について考察した。

本稿におけるテキスト情報から対象物を同定する 3 種類の実験では，平均的に半数程度の正答しか得られなかった。対象物の認識に有用とされた記述は，具体的な「形態」情報のほか，「人間との関係」「その他」に分類される情報に関わる傾向があり，辞書語釈に不足とされたのもこの種類が大部分を占めた。これらは読み手個人の経験知識を喚起するために有用とされる。すなわち，対象物に関する個別的な知識を呼び起こす情報がある場合には対象物を認識しやすいということであり，読み手の経験知識と合致する情報がなければ，対象物をテキストのみから認識するのが困難であるとわかった。コーパスの種別に関わらず有用とされる情報が文化的一般的な情報であったという傾向は，読み手の経験知識に合致しやすいためであったと考えられる。もっとも，読み手の知識に合致せずとも，検索キーとして用いた折，他のメンバーと明確に差別化が可能となる情報が取得できる場合もある。また，情報が多いがゆえに，読み手の知識に一部のみが合致したことで，かえって誤った認識を招く場合もあり得た。対象物認識に有用な情報をコーパスから取得するためには，頻度情報の効果的な使用が期待される。

対象物認識に求められる情報には，もう一種類，対象物の属する臨時的カテゴリの他メンバーとの差異情報というものもある。対象物の誤認は，対象物の知識の欠如のほか，絞り込んだカテゴリにおいて他メンバーの排除が行えない場合に生じる。誤解なく対象物を認識するためには，上位カテゴリのプロトタイプとの差異と，同カテゴリに属する（類似した特徴を有する）メンバー

例を挙げた差異の記述が有用となろう。文化的な知識等を前提としない対象物記述を行う場合には，このような記述が効果的であると考えられる。但し，これらはコーパスから容易に取得することができるものでもない。今後，テキストからどのような情報が取得可能か，用例の頻度やカテゴライズのための提示順序など，テキストからの対象物認識に有用な情報とテキストの可能性について研究を進めたい。

参照文献

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）』東京：三省堂.

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）』東京：三省堂.

Barsalou, Lawrence W. (1983) Ad hoc categories. *Memory & Cognition* 11: 211–227.

Goddard, Cliff and Anna Wierzbicka (2014) *Words and meanings*. Oxford: Oxford University Press.

後藤斉（1993）「『神話』の比喩的用法について―コーパス言語学からのアプローチ―」『東北大学言語学論集』2: 1–16.

Fillmore, Charles J. and Beryl T. Sue Atkins (1994) Starting where the dictionaries stop: The challenge for computational lexicography. In: Beryl T. Sue Atkins and Antonio Zampolli (eds.) *Computational approaches to the lexicon*, 349–393. Oxford: Oxford University Press.

国広哲弥（1994）「認知的多義論―現象素の提唱―」『言語研究』106: 22–24.　日本言語学会.

国広哲弥（1997）『理想の国語辞典』東京：大修館書店.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka and Yasuharu Den (2014) Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation* 48(2): 345–371 (DOI 10.1007/s10579-013-9261-0).

McRae, Ken, George S. Cree, Mark S. Seidenberg and Chris McNorgan (2005) Semantic feature production norms for a large set of living and nonliving things. *Behaviour Research Methods, Instruments & Computers* 37(4): 547–559.

奥村学・白井清昭（2008）「現代日本語書き言葉均衡コーパスを用いた意味解析」『言語』37(8): 66–73.

Rosch, Eleanor. H. (1973) Natural categories. *Cognitive Psychology* 4(3): 328–350.

Rosch, Eleanor (1975a) Cognitive reference points. *Cognitive Psychology* 7(4): 532–547.

Rosch, Eleanor (1975b) Cognitive representation of semantic categories. *Journal of Experimental Psychology* 104(3): 192–233.

Rosch, Eleanor (1978) Principles of categorization. In: Eleanor Rosch and Barbara B. Lloyd (eds.) *Cognition and categorization*, 27–48. Hillsdale, NJ: Lawrence Erlbaum.

Rosch, Eleanor and Carolyn B. Mervis (1975) Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7(4): 573–605.

Rosch, Eleanor and Carolyn B. Mervis (1981) Categorization of natural objects. *Annual Review of Psychology* 32: 89–113.

Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson and Penny Boyes-Braem (1976) Basic objects in natural categories. *Cognitive Psychology* 8(3): 382–439.

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

Sinclair, John (1992) Trust the text. In: Martin Davies and Louise Ravelli (eds.) *Advances in systemic linguistics: Recent theory and practice*, 5–19. London: Pinter.

Taylor, John. R. (1995) *Linguistic categorization: Prototypes in linguistic theory*. 2nd edition. Oxford: Clarendon Press.

Wierzbicka, Anna (1985) *Lexicography and conceptual analysis*. Ann Arbor, MI: Karoma Publishers, Inc.

Wierzbicka, Anna (1996) *Semantics: Prime and universals*. Oxford: Oxford University Press.

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』370–379.

関連 Web サイト

『現代日本語書き言葉均衡コーパス』（国立国語研究所）
http://www.ninjal.ac.jp/corpus_center/bccwj/

コーパス検索アプリケーション「中納言」1.1.0，短単位データ 1.0，長単位データ 1.0
https://chunagon.ninjal.ac.jp/
Yahoo! クラウドソーシング
http://crowdsourcing.yahoo.co.jp/

例文出典

Kudo, Taku and Hideto Kazawa (2007) Web Japanese N-gram Version 1. Gengo Shigen Kyokai.
『現代日本語書き言葉均衡コーパス』（国立国語研究所）

Encyclopaedic Descriptions That are Useful for Identifying Entities: A Case Study of Descriptions of Animals

KATO Sachi

Postdoctoral Research Fellow, Center for Corpus Development, NINJAL

Abstract

This paper reports what features of encyclopaedic descriptions are useful for recognising entities based on the results of three experiments on target object identification from texts.

We used gloss descriptions of animals compiled from dictionaries (Experiment 1), texts with additional information not found in the dictionaries (Experiment 2), and usages acquired from corpora (Balanced Corpus of Contemporary Written Japanese, Google Japanese N-grams; Experiment 3). In all three experiments, the percentage of the entities which could be identified correctly from texts was only about half. Therefore, we conclude that it is difficult to recognise an entity based on its description in a text. The results of the three experiments suggest that the following information is important for the identification of the target: specific features selected based on participants' personal experiences and information that would distinguish the target animals from other members in the *ad hoc* categories were important for identification of the target. When readers have knowledge about the target entities, individual or general experience and knowledge are easily obtained from corpora, as the actual texts are useful for readers' recognition. In contrast, when readers have insufficient knowledge about the target entities, a description of the features that distinguish the target from the prototype of the superordinate category (*ad hoc* category) is useful.

Key words: encyclopaedic knowledge, object recognition, corpus, categorisation, description of word meaning