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—

The effect of frequency information on object recognition

—A case study of animal body parts—

保田　祥†，浅原　正幸†
Sachi Yasuda, Masayuki Asahara

†国立国語研究所
National Institute for Japanese Language and Linguistics
yasuda-s@ninjal.ac.jp

Abstract

The subject of this study is to evaluate the contribution of text information to object recognition. It is difficult to recognize an object based only on a text. We can obtain the frequencies of words associated with an object from a corpus. We can assume that highly frequent words in corpora represent the characteristic features of an object. In this study, we investigated the relationship between characteristics features and frequency in corpora. Using crowdsourcing methods, we conducted two experiments in which frequency graphs of animal body parts were presented to participants, following which they were required to identify particular objects. Through the experiments, we found that participants tended to focus more on the distinctive features of the target object and other related objects than on the highly frequent features. Furthermore, we found that a cognitive gap exists between general knowledge and frequency data from corpora.

Keywords — Corpus linguistics, Word frequency, Text, Categorization

1. はじめに

テキストのみから対象物を認知するのは困難である．辞書語釈やコーパスから取得した用例そのものから，記述された対象物を同定することは難しい（保田ら 2013，保田 2014）．しかし，コーパスから取得可能な用例の情報には，それらの頻度についての情報もある．コーパスにおける用例の頻度情報を効果的に用いることで，対象物に関して言及されやすい情報，すなわち一般に特徴的と考えられている情報を示すことが可能と期待される．

そこで，本稿はコーパスに頻出する情報が対象物の同定に有用とされるかを調査し，頻度情報と対象物の特徴の関係について考察を行う．各々の動物について身体部位のコーパス中の用例頻度をグラフ化し，被験者が対象物を同定できるかどうかを評価する実験を行った．実験においては，提示したどの部分が有効であるのかについても調査した。結果，コーパス中に頻出する部分よりも比較対象物との特徴的な差異である部分が着目されるとわかった．

本研究の貢献は以下のとおりである．(1) コーパス頻度を提示する被験者実験を通して，テキストのみにより対象物を同定するために必要な要素を明らかにした．(2) クラウドソーシングを用いることにより，比較的大規模な被験者実験を行った．

2. 関連研究と本研究

2.1　外観的な情報について

10 種類の国語辞書の語釈において，200 種類の動物について 5 冊以上の辞書に記述のあった要素を調査した結果，32.1%と最も多く取得されたのは形態情報であった（保田ら, 2013）．動物であれば耳・口・尾などの外観的な要素に関する特徴などがそれにあたる．

また，McRae（2005）の連想調査結果の動物についてみると，「visual-form and surface」と分類された要素（feature）が 31.3%と最も多い．すな

わち，被験者の連想実験においても辞書記述と類した数値が得られているということであり，また，外観的に特徴的な情報がもっとも連想されやすかったと考えられる．

2.2 テキストからの対象物認知について

辞書の記述から対象物の同定可能性を調べた結果（保田ら, 2013），読み手が対象物の知識を有している動物であっても，半数程度しか同定できないことがわかっている．なお，テキストから対象物の同定が可能であった場合には，個人的な経験に結びつき得る情報（関連情報）が 57%と最も利用されたが，形態情報の利用も 39%に上った．

しかし，この実験では記述内容を提示するにとどまったため，たとえば耳や尾が特徴的であるとの記述からウサギをカンガルーと誤答する例も見られた．この理由として，対象物についてどの情報が最も特徴的であるのか，テキストのみからでは判断が困難だったことが考えられる．そこで，コーパス中の頻度情報を用いることにより，人間が対象物を正しく同定できる可能性が期待される．

2.3 コーパスの頻度情報について

コーパスを利用した辞書として，Sinclair（1991 など）らの COBUILD（1987～）や，三省堂の WISDOM（英和：2003～，和英：2007～）などをはじめとした多数の種類がある．これらの辞書は，独自の構築したコーパスにおける頻度に基づいた語釈あるいは用例の提示を行っている．特に外国語辞書では頻度順に語義が掲載されるが，現行の国語辞典についても，多義語の場合，頻度順に語義を掲載することが一般的である．しかし，このような頻度の利用は，多義相互間の意味関係が捉え辛いとして問題視されている（国広，2000 など）．対象物の認知という点においても，頻度情報がどのように役立つのか確認する必要があろう．

3. 調査

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．「A（動物名）の B（身体部位）」の組合せパターンを用いて頻度情報を取得し，身体部位について「尾」「しっぽ」「尻尾」などの異表記や同義語をまとめ，出現割合を示した円グラフ（図 1 参照）を動物毎に作成した．表示の際，2%未満の部位は「その他」としてまとめた．

実験は以下の①②の 2 種類を行ない，Yahoo! クラウドソーシングによって募集した実験協力者（15 歳以上，1,000 名）の回答を得た．

また，全ての動物対の組合せに対して身体部位比率に対する積率相関と順位相関を調査した．

3.1 実験①

身体部位の用例頻度円グラフから動物が同定できるか調査する．対象とした動物は，ライオン・ウサギ・コウモリ・ツバメなど[1]全 20 種類（図 2 内参照）である．実験協力者は，円グラフの示す動物を選択肢から選ぶ．選択肢は鳥や草食動物などのカテゴリー（群内の身体部位相関は各 0.6～0.8）とし，正答のほか 8 種示す．なお，事前実験により，円グラフから直接動物名を答えた場合，ほぼ正答が得られなかったため，候補群から選択することにした．

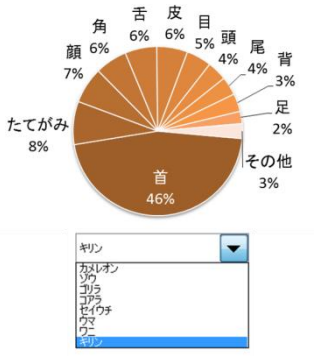

図 1 実験①の画面イメージ

[1] 20 種類の単語親密度（天野・近藤，1999，max:7～min:1）の平均は 5.88（max：ウサギ 6.56，min：セイウチ 4.53）である．調査対象として，概ね親密度の高い（一般に知られていると推測される）動物を選択した．なお，後述する実験①の結果においても，単語親密度の最も低いセイウチと最も高いウサギの正答率が 8 割程度でほぼ同じであることから，調査対象とした動物の親密度が正答に影響した可能性は低いと考えられる．

3.2 実験②

身体部位の用例頻度円グラフにおいて，どの部分に着目することで解答しているのか調査する．実験協力者に該当動物名とともに該当動物を含む 2 種類の動物の円グラフを呈示する．実験協力者はどちらが該当動物であるかを選択し，判断理由として着目した身体部位名を記入する．提示した円グラフは，カバ・ツルなど[2]の 12 種類（図 3 内参照）である．

4. 結果

4.1 実験①

正答率は平均 55%（max：カメレオン 96.1%, min：イタチ 11.5%）であった．正答率の高いカメレオンは「舌」が 48.3%と顕著であり，ペリカンは「口」と「嘴」で 84.3%と用例の殆どを占めている．用例頻度の高い特徴的な部位を有する動物であれば，対象物が同定しやすいことが予測される．

但し，正答率の低い動物でも 30%を占める高頻度な身体部位（例：コアラ（正答率 18.7%）における「鼻」「顔」，タヌキ（正答率 12.3%）における「尾」「皮」など）を有していることから，用例頻度の高い特徴的な部位のみが対象物の同定に有用とは言えない．

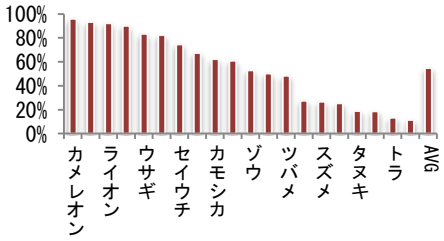

図 2 実験①の正答率

4.2 実験②

正答率は平均 64.7%（max：カバ 94.3%，min：鬼 14.9%）であった．提示したグラフは 2 種類であることから，グラフの類似性が正答率に関係することが考えられる．そこで，図 3 に正答率とともに，提示した 2 種類の動物ペアについて，それぞれの身体部位の出現率の積率相関・順位相関を示す．

結果を見ると，たとえばネコとクマのペアの積率相関は 0.80，順位相関は 0.84 であり，どちらも高い相関があるため類似したグラフと考えられるが，正答率は 63.5%と平均的である．また，正答率が 14.9%と極端に低くなった鬼とシカのペアについては，積率相関は 0.10，順位相関は 0.66 であり，類似したグラフとは言い難い．しかし，鬼とシカのペアも，グラフの類似度としては，正答率が 94.3%のカバとアザラシのペア（積率相関 0.24・順位相関 0.58）と大差がないように見える．よって，グラフの類似性のみが正答率に関係しているとは言い難い．なお，対象とした動物ペアの身体部位積率相関・順位相関と正答率との相関を調べると，それぞれ 0.3 であった．

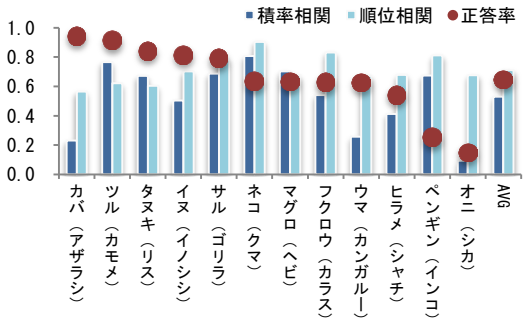

図 3 実験②の動物ペアの身体部位相関と正答率

また，判断の根拠として記述された身体部位名の回答数を図 4 に示す．正答・誤答ともに，ツルの首やヒラメの目など，判断時に着目される身体部位が集中する傾向があるとわかった．グラフ中の特徴的な身体部位に注目することで，対象動物の同定を行うことが推測される．

[2] 実験①と同様に，12 種類の単語親密度の平均は 5.89（max：ウマ 6.38，min：カバ 3.53）であり，概ね親密度の高い動物を選択した．なお，後述する実験②の結果においても，単語親密度の低いカバ（3.53）が最も高い正答率であることから，調査対象とした動物の親密度が正答に影響した可能性は低いと考えられる．

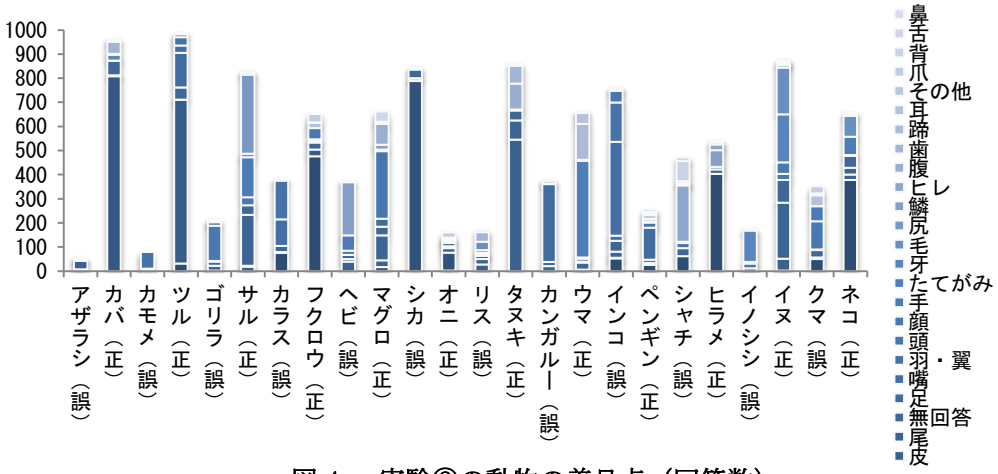

図 4　実験②の動物の着目点（回答数）

5. 考察

用例頻度から対象物を同定する実験の結果，用例頻度の高い特徴的な部位のみが対象物の同定に有用とは限らないが，着目される部位は集中していることがわかった．

5.1　対象物の同定に有用な情報

実験結果として得られた誤答を分析することで，対象物の認知における用例頻度以外の影響可能性を考えてみたい．

正答率とグラフの類似性に強い相関は見られなかったが（前述），誤答の原因としては，対象物と部位の用例頻度分布の類似した動物を選択した影響が考えられるであろう．たとえば実験①で正答率が低いコアラ（正答率 18.7%）の誤答はほぼゾウ（コアラとゾウの積率相関：0.39，順位相関：0.60）であった（図 5）．ほかに若干の回答があったゴリラ（積率相関：0.64，順位相関 0.48），セイウチ（積率相関：-0.03，順位相関 0.30）よりも順位相関の高い動物が選択されていたように見える．しかし，イタチ（正答率 11.5%）の誤答はキツネ（イタチとの積率相関：0.13，順位相関：0.62）が大部分であった（図 6）が，以下のタヌキ（積率相関：0.22，順位相関：0.79），クマ（積率相関：0.28，順位相関：0.66），オオカミ（積率相関：0.25，順位相関：0.62）などに比べて積率相関・順位相関ともにむしろ低い傾向があり，イタチとの身体部位が類似した頻度分布であったためにキツネの誤答が増えたとは言い難い．

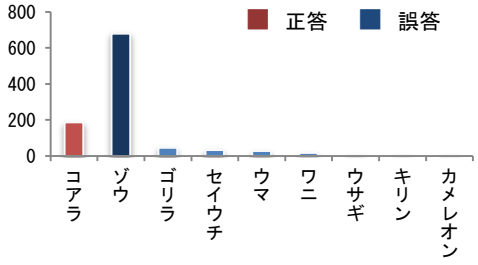

図 5　コアラの正答と誤答数

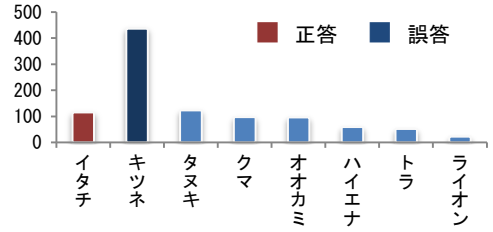

図 6　イタチの正答と誤答数

そこで，実験②の正誤ペアにおいて，頻度分布において何が着目されていたのか詳細を見てみたい．図 7 は，ネコ（正答率 63.5%と本稿実験において平均的な正答率）の身体部位用例の頻度分布（提示グラフ）と，正答（ネコ）・誤答（クマ）における着目部位をそれぞれ割合で示したものである．身体部位用例の頻度として最も高いのは「手[3]」であるが，正答において着目されているのは「目」であった．また，誤答では「顔」が着目されていることがわかる．「目」も「顔」も実際の頻度分布

[3]「猫の手も借りたい」のような慣用的な表現の影響が考えられる．

で最も多い(特徴的と考えられた)部位ではない．頻度の高い部位が着目されていたのではないということである．

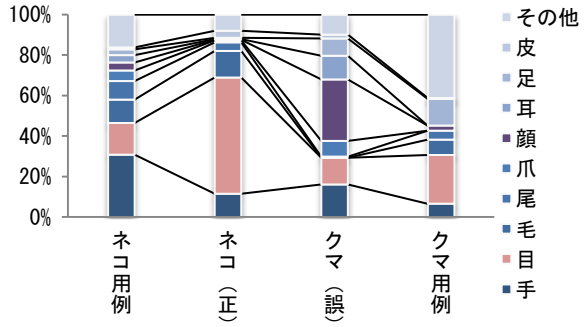

図 7　提示した身体部位用例比率と正誤答における着目点（ネコ）

もっとも，実験②においてカバは正答率が高く（94.3%），アザラシ（積率相関：0.24，順位相関：0.58）との選択において，高頻度（27.8%）の「口」が 84.1%の割合で注目されていた（図 8）．

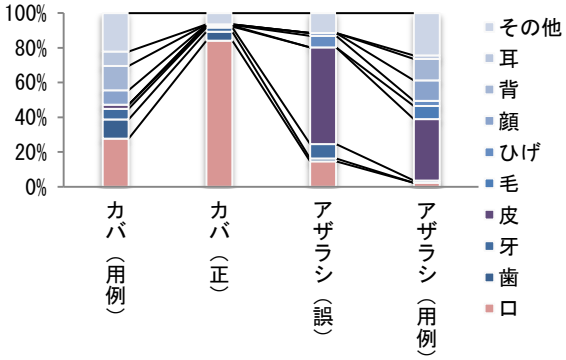

図 8　提示した身体部位用例比率と正誤答における着目点（カバ）

しかし，ウマ（正答率 62.5%）では，カンガルー（積率相関：0.27，順位相関：0.62）との選択において，用例頻度が 4.5%にすぎない「たてがみ」が 60.5%，次いで用例頻度 5.6%の「蹄」22.5%によって同定がなされていたのである（図 9）．

頻度の高い身体部位は特徴的であると考えられるだろうが，「耳」や「尾」のような特徴的な部位が類似した動物からウマを同定するためには，「高頻度の「耳」や「尾」が特徴的な動物」というカテゴリーにおいて，同じカテゴリメンバーである類似した比較対象との差異としての「たてがみ」と「蹄」が必要であったと推測される．反対に，他のメンバーの差異となる特徴的なウマの部位が「足」であると考えた場合には，誤答（カンガルーを選択）になったのだと考えられる．

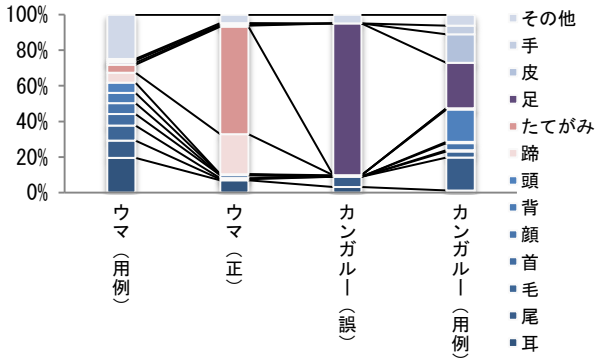

図 9　提示した身体部位用例比率と正誤答における着目点（ウマ）

このように，たとえば，ウナギとカンガルーは部位分布の相関を見れば類似性が高い（積率相関：0.66，順位相関：0.79）が，ウナギとカンガルーを区別するためには，頻度の 3.0%にすぎないウナギの「鰭」のような部位が，ウナギとカンガルーとの差異として注目されるのだと考えられる．もちろん，最も頻度の高い「口」が注目されていたカバ（口：84.1%）や「首」が着目されていたツル（首：69.7%）についても同様に，「背」「歯」，あるいは「嘴」「翼」などの他部分も含めた複合的判断がなされていたのであろう．

5.2　頻度情報と特徴的と考えられる情報

また，対象動物の特徴的な部位に関する用例は高頻度であると期待されるが，頻度上位の部位が必ずしも一般に特徴的であると考えられている部位ではない場合も有り得る．図 10 は，正答率が 14.9%と突出して低かったオニとシカのペアについて示したが，オニの用例においては「首（31.5%）」「爪（31.0%）」の頻度が高いのに対し，シカの用例では「角（63.1%）」の頻度が高いことが注目さ

れている．オニの用例では「角（3.9%）」の頻度は高くないが，誤答の場合の 93.3%が「角」に着目したと回答しているのである．オニは「角」が特徴的であると考えられているが，実際にはテキストにおいて出現する頻度が低いため，「角」の頻度が上位となっているシカが対照されたことで誤答が増加したものと考えられる．なお，正答の場合に「角」に着目したという回答には「鬼の角はこの程度の割合だと思う」「角ばかりではないはず」などとのコメントを加えた記述が散見され，単純に頻度上位の部位が特徴的部位とは考えられていないこともわかった．

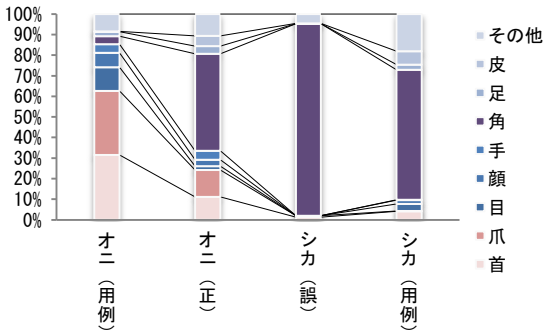

図 10　提示した身体部位用例比率と正誤答における着目点（オニ）

図 11 に示したのはテキストから取得されたコアラとゾウの身体部位比率であるが，実験①のコアラ（正答率 18.7%）の誤答としてゾウ（誤答率 68.0%；誤答中の 83.6%）の選択される割合が高かったのは，コアラについて「鼻」が特徴的と考えられていることによるほか，「耳」や「毛」といった部位も特徴的な部位であるとして排除されず，コアラの用例としては少ないが「背」に子を背負うイメージも考えられたなどの可能性が考えられる．しかし，ゾウの「足（6.4%）」や「牙（1.9%）」のようなコアラに特徴的と考えにくい部位が，ゾウをコアラではないとして排除する理由とはなりにくかったのだともいえる．前述のウマの「たてがみ（4.5%）」や「蹄（5.6%）」はカンガルーを排除したが，コアラについて「足」や「牙」がゾウを排除しないことは，あるはずのもの（特徴的と考えられる部位）について言及しないことよりも，ないはずのものについて言及することが，対象物の認識に関して影響の少ない可能性を示唆している．

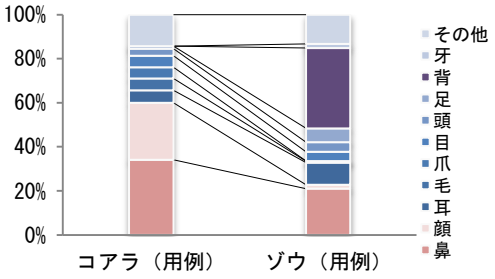

図 11　コアラとゾウの身体部位用例比率

6. まとめ

本稿の身体部位用例の頻度情報による動物の同定実験では，平均 6 割の正答率が得られた．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．対象物の認知のためには，単純に頻度情報を示すよりも，類似性の高い語との差異情報を示すことが有用であると考えられる．

また，対象物に特徴的と考えられている情報の頻度が高いとは限らず，特徴的と考えられる情報について言及していないことが着目される傾向が見られた．一般的な認識とテキスト頻度に差の大きい場合もあるため，頻度情報の扱いには留意が必要である．

本研究の応用として辞書の語釈文の自動生成があげられる．言語処理の分野で統計的手法やパターンによる語釈文の自動生成が提案されているが，多くのものは頻度に基づくものであった．本研究は頻度情報と人間の認知の乖離を明らかにするものである．今後の展開として，語釈文の自動生成においてどのような手法が有効なのか，認知言語学的な観点から調査を進めたい．

謝辞

本研究は JSPS 科研費 26770156 の助成を受けたものです．

参考文献

[1] 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

[2] 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

[3] 国広哲弥（2000）“日本語動詞の多義体系 (3)” 神奈川大学言語研究, 22, pp.1-12.

[4] McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005) “Semantic Feature Production Norms for a Large Set of Living and Nonliving Things.” *Behaviour Research Methods, Instruments & Computers*, No.37 (4), pp.547–559.

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

[6] 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.

[7] 保田祥（2014）“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予稿集, pp.117-126.