嗅覚語彙の網羅的収集

安藤豪啓 (1) 粟飯原俊介 (1) 小早川達 (2) 田中久美子 (1)
(1) 東京大学情報理工学系研究科　(2) 産業技術総合研究所
{ando,aihara,kumiko}@cl.ci.i.u-tokyo.ac.jp　kobayakawa-tatsu@aist.go.jp

1　背景

日本人の嗅覚に関する語彙を網羅的に収集、分類し、日本人の代表的な嗅覚十を抽出するプロジェクトの第一段階として、嗅覚語彙を網羅的に収集する研究に関する報告をする。

本プロジェクトの動機は、嗅覚同定能力キット [1] にある。このキットは嗅覚異常が疑われる被験者の検査に使われる。キットの中には、図 1 のようなカードが 10 枚入っており、各カードの左ページには、においを同定する物質、右ページにはそのニオイに関する語句が記載されている。被験者は左ページの物質を嗅ぎ、右ページの 4 択から正解を一つ選ぶことを 10 回繰り返す。正解の数が少ない場合には、嗅覚異常が疑われ、それはたとえばアルツハイマー病やパーキンソン病などの初期症状の診断に用いられる。本キットは日本で斉藤らにより発明され、医療現場において実用化されている。また、他の国でも同様のキットの開発が望まれている。

嗅覚同定能力試験は、20 名程度の日本人にアンケートを行って策定されたものである。アンケート参加者がにおいにまつわる語彙を網羅的に挙げ、語句間の近さなど数値を得て統計処理することにより百程度の嗅覚語彙を得た。これを元に、代表的な 12 種類のにおいと、それぞれのにおいに対して 4 つの選択肢が決められた。すなわち、嗅覚動的試験は、においに関する語彙のアンケートに基づいて設計されたものである。このキットは、少数の人に対するアンケートから作られていることから、日本のにおいの代表として適当であるか、また、若い世代でも用いることができるキットになっているかなどの問題がある。

このような手続きを経てキットが作られる理由は、次節でより詳しく述べるが、人間の嗅覚が具体物に基づいてしか同定され得ない、という特徴があるからである。嗅覚が語彙に拠るのであれば、現在では web などの莫大な量のデータがあり、そこには、ニオイに関する言及の全容が含まれておろう。そこで、本研究は、web のデータから日本人の嗅覚に関する語彙を網羅的に収集し、そこから次世代の嗅覚同定能力キットを設計するための基礎データを抽出することにある。本稿はその第一段階として、においに関する語彙を網羅的に収集する試みに関する報告である。このような語彙は、日本人の嗅覚の全容を表すばかりか、研究上の他の応用をさまざまに考えることができる。

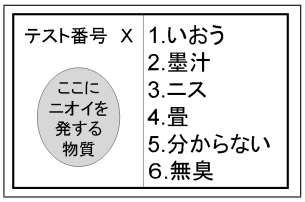

図 1: 嗅覚同定能力キット中のカード

2　嗅覚と語彙

人間の五感のうち、感覚を構成する要素に分解可能なものがあり、視覚や味覚が該当する。視覚に関しては、色相、明度、彩度の組み合わせとして決まり、色は R,G,B の 3 つの色味の強さが基底となる。また、明度や彩度は一軸上の数値として記述される。味覚は、甘味、塩味、苦味、酸味、うま味の 5 つが基本要素を構成し、複雑な味はそれらの複合として構成される。基本要素のいくつかは、身体の感覚受容体に対応することがあり、その場合には、感覚は感覚受容体にタイする刺激の強さの総合として認識される。このように、ある感覚を構成するものが基本要素に分解されれば、さまざまな感覚はそれらの複合として理解され、また、感覚空間の全容を知ることができる。

一方、嗅覚については、人間の嗅覚受容体は 388 あり、基本要素に相当するものが視覚、味覚と比べると多い [2]。しかも、受容体とにおいを発する物質の分子が 1 対 1 対応なら基本臭は 388 とも言えるが、実は

多対多である [3]。すなわち、嗅覚において基本要素はないのと同然というのが認知心理学においては通説になっている。このことは、人間の語彙に嗅覚を直接表現する言葉がない、という現象として表れるとされる [4] たとえば、視覚では、R,G,B 相当には、赤、緑、青と色を構成する基本要素を直接表現する語彙がある。一方で、嗅覚には、嗅覚の基本要素に相当するものを表す語彙がない。「臭い（くさい）」などは、嗅覚を直接的に表す単語と思われるが、実は、人間にとって快くはない嗅覚物質の総体である。つまり、嗅覚は常に複合として認識され、嗅覚に働きかける物質を発する具体物に即して表現され、たとえば「～ のような匂い」「～ の香り」の形式として表れる。たとえば、「りんごのような香り」は、りんごの発する嗅覚にはたらきかける物質をりんごに即して表現する語句である。

このように、嗅覚が基本要素に分解されないことは、人間の嗅覚の全容はニオイに関する語句を集めることによってしか把握しえないということである。このことは、嗅覚の分節も自然言語同様に文化に依存するということである。たとえば、ラベンダーの香りはフランス人には一般的な南仏文化を代表する香りであるが、日本人の中には、その香りを同定できない人は多くいる。一方で、日本人のほとんどはひのきの香りがわかるが、フランス人の中にはひのきの香りを知らない人が多く、これを不快なものと感じる傾向にある。

このように、においの全容は言語の中に在る。その意味でこれらの言語処理技術は認知心理学分野へ大きな貢献を可能とする。本稿では、以下「ニオイ」により人間の嗅覚に働きかける物質を意味するものとし、日本人のニオイの全容を語彙を通じて収集することを試みる。

3 語彙収集

データの収集元として、本研究では Google n-gram 日本語版を用いた[1]。Google n-gram から 2 つのテンプレート「A の B」にマッチする出現を全部抽出した。ここで、B は「香り」「臭い」「匂い」「かおり」「におい」「ニオイ」六語ならびにこれらの前に「のような」を付けた 12 語句であり、A にマッチする嗅覚語句の全容を得ることが本研究の目的となる。延べ 153408 件の語句が得られ、うち、135628 件の異なり語句が獲得された。

[1] http://www.gsk.or.jp/catalog/GSK2007-C/catalog.html

表 1: Google n-gram の抽出例:「A の香り」

n-gram	出現頻度			
3-gram	13664			茶
	6764			パウダー
4-gram	41		ぴん	茶
	5939		ベビー	パウダー
5-gram	41	さん	ぴん	茶
	288	漂う	ベビー	パウダー

出現例を表 1 に示す。この例にも見られるように、抽出されたものには、不適当な語句が含まれる。また、異表記、同音異議語などの問題が散見された。そこで、A にマッチした部分をまず国語辞書 (goo 辞書) を用いて篩った。この工程の必要性は §4.3 でより詳しく述べるが、辞書に記載のある語句は、日本人の語彙として一般的なものであり、嗅覚同定能力キットにおいても一般的なものに限った方がよいとの観点から、このようにした。

結果、7012 件の語句が得られた。この語句集合を以下では母集合という。母集合は、web 上に表れる日本人の嗅覚に関する語彙を網羅しているが、不適当なものがある。母集合は以下の語句にだいたい分類することができる。

1. ニオイが顕著な具体物を示す語彙：バラの香り、汗の臭い、シンナーの臭いなど。
2. ニオイの顕著さが疑わしい具体物を示す語意：本のにおい、柿のにおい、カブトムシのにおいなど。
3. 具体物の上位範疇を抽象的に表す語彙：石けんの香り、魚のにおい、動物の臭いなど。
4. 文脈上、あるいは比喩的に用いられる語彙：犯罪のにおい、寺院の香り、京都の香りなど。
5. 語用論的な語彙：独特の匂い、系のにおい、一般の香りなど。

番号が大きくなるほど、実際の具体物からは離れる傾向にある。とはいえ、専門家が見てもこれらの弁別は難しく、その難しさは以下の 3 つに大別される。

第一の問題は、ニオイの顕著さの峻別の問題である。本は具体物である以上、図書室の書庫のニオイに代表されるような、本にもニオイがある。とはいえ、新書と古書ではニオイも異なり、紙質にも依存する。そもそも、本を顕著にニオイのあるものとして捉える人は多くはなかろう。嗅覚同定能力キットの策定上は、本は不適切な具体物といわざるをえないであろう。

第二は、上位下位範疇のどこで線引きを行うかという問題である。たとえば、魚のニオイは、下位範疇と

して鰯のニオイ鯖のニオイなどがあり、鰯や鯖の方が魚よりは具体物に近い。とはいえ、鰯と鯖のニオイを弁別することができる日本人は専門家に限られるであろう。このように、ニオイの弁別能力は、人に依存し、また、上位下位範疇のどこで線引きが行われるかは、具体物の種に依存する。特に、後者については、認知意味論上のプロトタイプ理論の基本語彙との関連が疑われ、ニオイの弁別能力と基本語彙の関係を明らかにすることは今後の研究課題である。

第三は、語彙が直接表す具体物のニオイと、その具体物が即した別の具体物のニオイの弁別である。たとえば、寺院の香りは線香の香りである。一方で、寺院自体も具体物であり、木材など建材がある以上、それはニオイを発する。寺院自体のニオイが人間にとって顕著かの問題もあるが、寺院のニオイとして線香のニオイに言及することは嗅覚同定能力上は不適切である。

このようなさまざまな観点が複合的に相まって一つの嗅覚語句として表象される。臭覚同定能力キットの策定ためには、1 の集合を獲得し、その中の代表的な語句を選ぶことが必要となる。

4 嗅覚語彙策定

4.1 嗅覚語彙基準

嗅覚同定能力は、誰でも行うことが可能な検査でなければならない以上、日本人なら誰でも知っているニオイに基づき、なお、嗅覚異常を判定する上ではさまざまなニオイの種類を網羅するものでなければならない。

求める語句の集合、すなわちニオイを発する具体物を表象する語彙を「嗅覚語彙」と定義し、さらに、嗅覚語彙の基準として以下のものを定めた。本稿の第三著者は、人間福祉工学を専門とするが、この基準は関係する専門家と共に策定したものである。嗅覚語彙に含まれる語句は、

1. 一般的な用語であり、
2. ニオイに関する言及が顕著に見られ、
3. ニオイが特定される語句

である。今後この基準を関連分野において問うことになろう。本稿の目的の上では、この 3 つの条件を満たす語彙を網羅的に収集することになる。

4.2 嗅覚語彙に関する事前調査

事前調査として、人間の嗅覚語彙の認識に関する揺らぎをアンケートにより調べた。この揺らぎが小さい場合には、アンケート調査の結果をもとに嗅覚語彙の収集を行うことも考えることが可能である。

アンケートでは筆者らが適宜選んだ 496 の語句について、嗅覚語彙に含まれるかどうか (つまり語句にニオイがあるかかどうか) の○×を付けるものである。その際、上に述べた基準を示した。語句のうちの半分は嗅覚語彙に含まれる。アンケートの回答者は男性 3 名と女性 2 名であった。

5 名によるアンケートを集計したところ、全員のニオイの有り無しの意見が一致した語彙は 496 中 213 語しかなく、うち○が全員で一致したものは 36 語、×が全員で一致した語彙は 177 語である。また、任意の 2 名の平均一致率は 66% であった。また、第三著者との一致率は平均で 72%であった。

このように、たったの 5 名ではあっても揺らぎは大きく、一般人に嗅覚語彙の正解を作成することは難しい。そこで、専門家が正解を作成し、これに基づいて機械学習を行い、母集合から嗅覚語彙を抽出することとした。

4.3 嗅覚語彙策定の工程

正解データは、なるべくならば、嗅覚語彙かどうかが半々となるように作成した方が望ましいが、7012 のもののうち、明らかに正解であるものはそれほど多くはない。だいたい 2 対 5 の割合で嗅覚語彙とそうでないものがある。

§4.1 基準は、それぞれ言語処理の観点から捉えることができる。1 については、§3 で述べた「国語辞書にある」条件により代えることができる。2 については、顕著さは χ 二乗検定で臭いや香りに関する言及が顕著であるかを調べることができる。具体的には §3 で調査したテンプレートについて、A に対する B の顕著さを「A の B」「¬ A の B」「A の ¬ B」「¬ A の ¬ B」の四出現に対して google-ngram 上で頻度を計測し、検定を行う。χ 二乗値が高いものから上位 100 単語の正解率を調べたところ、嗅覚語彙であるものは 86 %であった。以下に χ 二乗検定の上位に上がった語句 15 を以下に挙げる。

ラベンダー、グレープフルーツ、磯、金木犀、シトラス、ジャスミン、せっけん、花、シナモン、オレンジ、グリーンティ、果実、ココナッツ、コーヒー、線香

この例から、χ 二乗値が高いことは §4.1 に示した基準の 2 までを満たす根拠となりうるが、χ 二乗値の大きい順にどこまでを見るべきかは不明である。

また、基準の 3 から、嗅覚同定能力に不適切な語として、果実など、抽象的な上位概念の語が含まれてもいる。そこで、χ 二乗値に基づき、数百から成る正解データを人手で作成し、機械学習により二値で語句を篩い、それを最終的には専門家によりさらに篩い、嗅覚語彙を決定する。

χ 二乗値が高い 500 単語、逆に低いものからランダムに 500 単語を選び、専門家集団 2 名により嗅覚語彙の正解データを作成した。正解データは、嗅覚語彙であるものを 300 語句、そうでないものを 300 語句含む。

5 SVM による嗅覚語彙判定

正解データに対し、語句の共起単語を素性として SVM により分類器を構成し、母集合を篩う。

まず、正解データの各語句に対し、google-ngram から共起単語を得た。嗅覚語彙であるとされた語彙、されない語彙、各語彙において全ての単語に対する相互情報量を算出し、相互情報量の上位各 20 件の語彙を素性とした。そして、各語句に対して、各素性との相互情報量を値とする素性ベクトルを構成した。

SVM は SVM^{light}[2]を用い、カーネルは多項式カーネルを利用し、パラメータ d は 1 で、その他はデフォルトの値を用いた。

10 分割交差検定で精度は 86.8%であった。母集合から正解に属するものをのぞき、サポートベクトルから大きかった語句上位 30 を以下に挙げる。

> にんにく、酒、人参、はちみつ、しそ、にんじん、ねぎ、豆腐、ハーブ、大蒜、チョコ、マスタード、ヨーグルト、ドリアン、いちご、きなこ、きゅうり、ゆず、唐辛子、玉葱、オニオン、白桃、ごま、キンカン、金柑、みそ、ポテト、たまねぎ、ケチャップ、りんご

母集合の語彙を人手で篩って直接嗅覚語彙を策定することには、揺らぎなどの観点から無理がある。しかし、SVM で母集合を篩えば、1956 語に対象語句が限られた。χ 二乗検定と SVM の分類精度を比較すると、SVM の方が識別精度が高いと言える。χ 二乗検定は χ 二乗値の上位 100 のみに注視した場合に 86%であったが、SVM は常に 86.8%の精度で識別が可能である。現在は、SVM により得られた語彙を専門家によりチェックすることにより、日本人の嗅覚語彙を網羅する語彙群を作成している。

[2]http://svmlight.joachims.org/

6 まとめと展望

Google n-gram を用いて、日本語の嗅覚語彙を網羅的に収集する試みを行った。試行錯誤の結果、本稿で提案する嗅覚語彙は、以下の工程を経て残った語句の集合である。

1. web データから嗅覚に関する語句をテンプレートを元に収集する。
2. 語句を国語辞書で篩って母集合を作る。
3. このうち、χ 二乗検定により、嗅覚語彙に確実に含まれる語句を選び、それを元に専門家により正解データを作る。
4. 機械学習により母集合を篩い、最終的に専門家によりそれをさらに篩う。

今後は、得られた嗅覚語彙を機械的にクラスタリングを行い、日本語の嗅覚語彙代表十の選定を行って嗅覚同定試験の基礎データを抽出する。また、他の文化圏、特にフランス語圏で同じ作業を行うことや、人類汎用の嗅覚語彙を英語を通して得ることを試みたい。

参考文献

[1] 綾部 早穂斉藤 幸子. 環境臭気におけるにおいの質の評価のための記述語の選定. 臭気の研究 = Odor Control Association journal, Vol. 33, No. 1, pp. 1–12, 2002.

[2] Y. Niimura and M. Nei. Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. *Journal of human genetics*, Vol. 51, No. 6, pp. 505–517, 2006.

[3] 東原和成. 香りを感知する嗅覚のメカニズム. 八十一出版, 2007.

[4] 綾部 早穂斉藤 幸子. においの心理学. フレグランスジャーナル社, 2008.