自己組織化マップを用いた助数詞の分析
— 助数詞「本」を例に —

李 在鎬（りじぇほ）* 井佐原 均†

1 はじめに

本稿では、現象分析の立場から助数詞「本」の意味記述を行う。とりわけコーパスベースに言語データを収集し、自己組織化マップで（意味情報の）可視化を試みる。

意味記述をめぐる二点の問題を中心に考察する。1）脱曖昧化におおいて表層の文字列には表れていない要素の問題をどう位置づけるべきか。2）意味のコード化はいかにすべきか。1）は直接には表れていないが、句や文の曖昧性の解消に貢献する要素の問題をいかに捉え、位置づけるべきかの問題である。2）は言語の意味をどのような方法で表現すれば良いかに関する問題である。前者は理論面の問題提起であるのに対し、後者は方法論面の問題提起になる。本稿では、1）の問題を認知言語学が主張するところの発話者の視点や捉え方（construal; Langacker [5]）の問題に関連づけ、一般化を試みる。2）の問題に対しては複数の要素に制約を分散する記述を与え、定量化に基づく評価で最適化する手法を提案する。なお、メタレベルの狙いとして自然言語処理の技術が従来の人文科学的な言語の意味記述に新たな一般化を可能にする点を示唆したい。

2 助数詞研究における記述的問題

2.1 従来の研究とその問題点

助数詞「本」に議論を限定した場合、これまでは、いわゆる「細長さ」という形状的特徴を中心的な意味とするカテゴリー化があると主張されてきた。例えば、 Matsumoto [7] などの場合、プロトタイプ意味論のアプローチから「際立って一次元的なもの」を「本」のプロトタイプ的な成員とし、三つの意味拡張があると指摘している。1）巻かれているものや輪状のものへの拡張、2）容器の形状のメトニミー的拡張、3）経験的な一次元性を形成するもの (例えば小説や論文など) や軌道を形成するもの (例えば電話の通話やホームランなど) へのメタファー的拡張があると指摘している。また、西光 [8] でも「細長さ (鉛筆など)」や「棒状のものの握り方 (バット)」といった助数詞「本」の具体的なプロトタイプ特性が抽象的領域に拡張されるとしている。

しかし、こうした先行研究には次のような問題点がある。提案された制約の多くが一般的すぎて、a）解釈次第でどのような分析も可能になってしまうこと、b）どの制約がどれだけ関与しているかについての問題があまり明確ではない。 例えば Matsumoto [7] が本の使用に対する制約として主張する「一続きの経験でなければならない」といった類の制約は経験事実に符号する形で判断し、定義づけることはほとんど困難である。なぜなら一続きの経験であるものとそうでないものの認定が不可能だからである。また、一続きの経験であってもそれを「本」で数えられるとは限らないのである。これらの問題点は記述モデルとして重大な欠陥であり、過剰般化の危険性を伴う。

2.2 記述的問題

原点に立ち戻って助数詞「本」の問題を考えてみた場合、二点の難題が存在する。一つ目は数える対象が必ずしも言語化されるとは限られない問題点がある。二つ目は規範的制約のみでは記述しきれない問題点がある。

* (独) 情報通信研究機構
† (独) 情報通信研究機構

(1) a. まずは 1 本の電話から。
b. 敗者復活戦では 1 本の魚しか検量してもらえない。

一つ目の問題として (1a) では物体としての「電話 (機)」そのものではなく通話 (回数) のことを数えている。二つ目の問題として (1b) では本来「匹」で数えるべき対象であるにも関わらず、「本」でも数えられる。さらに (2) のように発話文脈を考慮せずには、記述困難な現象も存在する。

(2) a. 漁師が 5 本のサンマを釣った。
b. ??母が 5 本のサンマを買ってきた。

本研究では、助数詞の記述的問題の背景には発話者の主観性の問題、すなわちどのような視点で対象を捉えているのかという問題が深く関与しており、そのことを明確にしない限り、助数詞の用法の充分な一般化ができないという結論に至った。

以上の事実を受け、本研究では (1) や (2) で示した記述的問題を認知主体の視点に関する問題として捉え直す（cf. 濱野 [2]）。さらにこの問題を反証可能な形で捉えるための方法論として (言語直感だけに頼るのではなく) コーパスデータに対する定量的分析手法を用いる。具体的には自己組織化神経回路網モデル (以下 SOM) を導入し、考察を試みる。

3 方法論

3.1 SOM

SOM は多次元の入力データに対して、二次元配列のノードで構成され、自己組織化によってデータの特徴を（非線形的にマッピングし）可視化する[1)]。この技術は、学習に基づくクラスタリングをすることで一般の多変量解析に比べ、高い精度を持つものとして知られており、様々な分野で利用されている（cf. 徳高・岸田・藤村 [9]）。自然言語処理の分野においても自動化の利点を活かし、馬 [6] や Kanzaki *et al.* [4] などでは名詞句の類似・非類似関係を明らかにしており、非常に興味深い。

[1)] アルゴリズムや解析手順の詳細は Kohonen [3] を参照されたい

3.2 学習データ

さて、上記の手法を本稿の言語現象に応用する場合、まず数える対象の意味的関係を定義した学習データが必要である。そこで「日英新聞記事対応付けデータ (Utiyama and isahara [10])」を利用し、データを集めた。データ収集および事前調査においては樋口耕一氏作成の KH Coder(Win 版) を利用した。なお、KWIC データ抽出においては構文バイアスを抑えるため、統語パターンを限定した。プレテストにおいてもっとも生産的であった「*本の*」(e.g., 3 本の鉛筆) で KWIC 検索を実行し、91 例のサンプルを得た。

3.3 コーディング

次に、個々のサンプルに人手によるコーディングを行った。しかし、この作業には本質的な問題点が存在する。それは本稿の代替案が主張する「認知主体の視点」と呼ばれるものの中身の問題である。というのは視点と呼ばれているものの明確な定義が難しいため、客観性を保ったコーディングが困難である。一例として大きさや細長さといった問題はスケール性（scalability）を有するため、その判断は決して容易ではない。

この種の難題を受け、私たちは視点を限定的に捉える必要があると判断した。アフォーダンス（affordance; Gibson [1]）からヒントを得て「身体を介した経験可能性」の点から視点の問題を捉えることにした。例えば、対象の大きさや細長さをめぐる視点の問題を [人が手を使ってそれを握れる] か、それを [もって振り回せる] か、その [中に人が入れる] か、などといった意味特徴を変数として導入した。最終的には三種類の変数セットを使用し、1/0 で判定を行った。

- 変数セット 1: 有情物, 場所, 人工物, 自然物, 抽象物, 形状 [線]・[面]・[点]・[面]
- 変数セット 2: 移動性, 握れる, 振り回せる, 人が入れる, 乗機性, 物体投入性, 情報投入性
- 変数セット 3: 対象の格 [が]・[を]・[に]・[で], 主語名詞による制御可能性

変数セット 1 はカウント対象の外的な属性に基づ

く特徴づけで、NTT の日本語語彙大系に部分的に準拠して判定した。変数セット 2 はカウント対象に対する主体の関わり方に関する特徴づけで、作業者間で判定した。変数セット 3 は生起文脈に関する特徴づけで、例えば (1b) であれば、対象はを格名詞句として生起し、主語となる漁師は魚を何らかの形で支配し、制御できることをコーディングした。

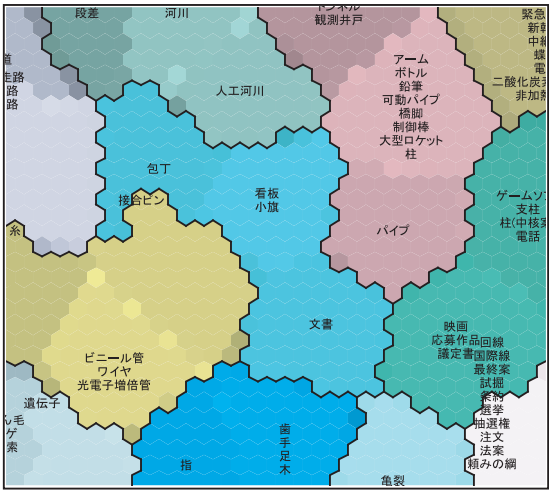
3.4　計算機実験の手続き

実験においては、同一のパラメータで異なる変数セットを投入し、複数回にわたって解析を行った。具体的には 32 × 31 の二次元配列のノードでマップを構成し、トレーニングサイクルは 27 回、テンションは 0.5 に設定した。解析は三回行った。一回目は外的属性を示す変数セット 1 のみで行った。二回目は変数セット 1 と変数セット 2 で行った。三回目はすべての変数を投入し、解析した。このような手続きを通して、どのような条件において最適なクラスタリングが得られるかを考察した。なお、マップの作成においては市販の Viscovery SOMine 4.0 を利用した。

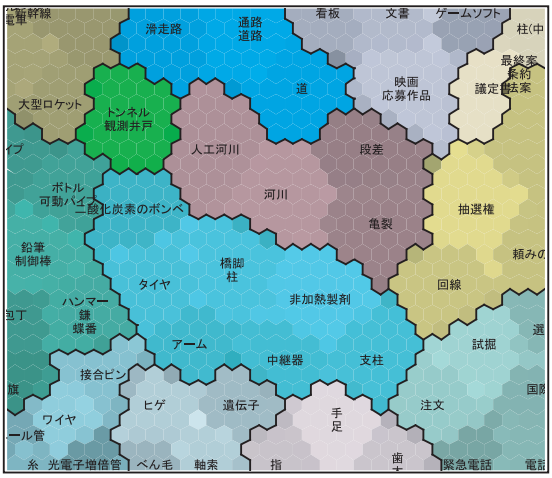
4　結果と考察

外的属性のみによる解析の結果、図 1（a）が示すような 11 個のクラスタで最適な分離が得られた。次に変数セット 1 に変数セット 2 を加え、解析した。その結果、図 1（b）が示すような 13 個のクラスタで最適な分離が見られた。最後にすべての変数セットを投入し、解析した。その結果、図 1（c）が示すように 7 つのクラスタで最適な分離が見られた。

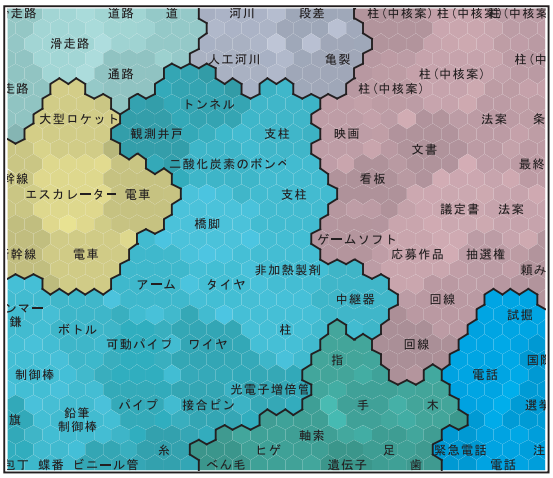
さて、最適な分類という観点から、図 1 の結果を観察した場合、（a）は人工物と自然物がうまく区別されてはいるものの、同じ人工物とは言え、本の動機づけとしては全く異なる性質のものが一つのグループになっている。例えば、文書と包丁のようなものが一つのクラスタに属している。この点において本の使用に関する動機づけを反映したクラスタリングとは言えないであろう。一方（b）を見ると右側に主として抽象物の意味グループが形成されているのに対して、左側は具体物の意味グループが形成



(a)



(b)



(c)

図 1　意味マップ

クラスタ (マッチング レコード)	ケース	貢献した変数
1 (28)	トンネル，観測井戸，支柱，二酸化炭素のボンベ，支柱，橋脚，非加熱製剤，アーム，タイヤ，中継器，ハンマー　鎌，ボトル，柱，可動パイプ，ワイヤ，制御棒，光電子増倍管，パイプ，接合ピン，小旗，鉛筆	人工物(1)、立体形(0.71)、手で握れる(0.64)
2 (25)	中核案，法案，条約，映画，文書，最終案，看板，議定書，法案，ゲームソフト，応募作品，抽選権，頼みの綱，回線	人工物(0.64)、抽象物(0.96)、創作物(0.84)、社会制度(0.72)、情
3 (7)	大型ロケット，新幹線，エスカレーター，電車	人工物(1)、立体形(1)、移動物(1)、人が入れる(1)、人が乗れる(1)、主語名詞による制御可能性(0.86)、を格(0.61)
4 (9)	指，手，木，軸索，ヒゲ，足，べん毛，遺伝子，歯	有情物の部分(1)、立体形(0.56)、握れる(0.54)、主語名詞による制御可能性(0.55)
5 (10)	滑走路，道路，道	場所(1)、人工物(0.9)、線形(1)、面(1)、人が入れる(1)、人が乗れる(1)
6 (8)	試掘，国際線，選挙，緊急電話，注文	抽象物(0.87)、立体形(0.5)、出来事(1)、が格(0.62)
7 (4)	河川，段差，人工河川，亀裂	自然物(1)、線形(0.75)、場所(0.75)、が格(0.75)

*括弧内は平均値

図 2　SOM Ward 法によるクラスタリング

されており、Matsumoto [7] が指摘する抽象性の度合いを反映した分布を見せている。しかし、本の動機づけという観点からみた場合、少し細かすぎるクラスタリングになっており、法案と作品のようなものが別々のものとしてクラスタリングされている。最後に、(c) を見ると、二つの意味軸によってグループ化されていることが分かる。それは、横軸では具体性の相違として各々のケースがポジショニングされているのに対して、縦軸においては大きさの相違によってポジショニングされており、分かりやすい分布を示す。また図 2 に示すケースと変数を見ても偏りなくコーディングを反映した分布になっていることが分かる。

なお、元データ集合のデータ・ベクトルが、どれぐらい特定のノードによくマッチしているかを示す量子化誤差を見ると、全体の平均値として (a) は 8.73927E-33 となっているのに対して (b) においては 2.70715E-33、(c) においては 9.410000E-34 となっており、(c) がもっともよく適合していることが分かる。

5　終わりに

本稿の分析結果は次の点で注目に価する。コーディングに用いたいずれの変数も具体性や大きさを直接にはコード化していない。それにも関わらず情報の分散によって具体性と大きさの相違による分布事実をうまく捉えている。一方、Matsumoto [7] では指摘されなかった、大きさのスケールが本のポジショニングする際、重要であることが示唆された。

以上のことから SOM の現象分析への新たな可能性が示唆された。それは、人文科学的現象分析に対して分析結果の最適性を明確化し、データを探索的に観察できる点で大きな利点があると言える。

参考文献

[1] Gibson,J.J. *TheEcologicalApproachtoVisualPerception*. Lawrence Erlbaum Associates, 1979.

[2] 濱野寛子.「助数詞の意味拡張のメカニズムに関する認知言語学的考察－助数詞「本」の用法を中心に」(京都大学修士論文), 2005.

[3] Kohonen, T. *Self − OrganizingMaps*. New York; Springer, 1995.

[4] Kyoko Kanzaki, Qing Ma, Eiko Yamamoto, Masaki Murata and Hitoshi Isahara.”Objective Hierarchy of Abstract Concepts - Organization of Abstract Nouns via Distribution of Adjectives -,” *JournalofCognitiveScience*, Vol.4, No.2, pp. 201-225, 2003.

[5] Langacker,R.W. *FoundationsofCognitiveGrammar*, *Vol.I* California: Stanford University Press,1987.

[6] 馬青、神崎享子、村田真樹、内元清貴、井佐原均.「日本語名詞の意味マップの自己組織化」,『情報処理学会論文誌」, Vol.42, No.10, pp.2379-2391,2001.

[7] Matsumoto, Yo. ”Japanese numeral classifiers: a study of semantic categories and lexical organization.” *Linguistics* Vol.31, pp. 667-713, 1993.

[8] 西光義弘 「類別詞の認知様式の相関に関する理論的考察」, 西光義弘・水口志乃扶 (編),『類別詞の対照』, pp. 23-38, くろしお出版, 2004.

[9] 徳高 平蔵・岸田 悟・藤村 喜久郎. 『自己組織化マップの応用-多次元情報の 2 次元可視化』、海文堂, 1999.

[10] Utiyama, Masao and Hitoshi Isahara.”Reliable Measures for Aligning Japanese-English News Articles and Sentences”.*ACL* − 2003, pp. 72–79, 2003.