英語前置詞句のメタファ表現に着目した名詞分類の試み

石井 康毅　佐野 洋
東京外国語大学 外国語学部

1　はじめに

語彙をその意味に基づいて分類した言語資料は有用である．自然言語処理でもそうした語彙分類を利用した研究成果があり，応用範囲も広がっているが，適用の効果は様々である．分析が難しい分野にメタファ(metaphor)表現の理解がある．語の分類は一般に分析者の分類視点に基づくため，語彙分類の規模が大きくなると細部の語の分類では，分析者の言語直観の影響が大きいことが予想され，客観性の確保が難しくなる．また，メタファは語の通常の意味を拡張させることで運用する言語表現であるから，語彙固有の意味をより正確に記述するとメタファ表現の理解には利用が難しくなる．メタファの理解は自然言語処理において重要な課題であり，メタファの説明が可能な語彙分類モデルの構築が必要である．

より人間の言語知識が反映された語彙分類を作り上げるためには，言語の運用体系全体を対象として分類視点を抽出することが必要である．また，内容語の視点からの一般的な語彙分類とは異なり，メタファを対象とする場合は，機能語の視点からの分析が有効である．そこで本稿では，メタファ表現を構成する前置詞句に着目して名詞を分類できるという仮説のもと，前置詞 (相当句)*in/inside/into/out of/outside* の補部の名詞の性質についての調査を行った．その結果，容器のイメージスキーマでとらえられる名詞をグループ化することができた．

2　メタファと語彙分類

2.1　メタファ

例えば，次の例文 [1][1] の view，sight，way，field が示すように，一般的な語彙分類では十分には記述されない関係 (この場合はある意味での類似性) がある．

(1) a. The ship is *coming into* view.
　　b. I *have* him *in* sight.
　　c. I can't see him — the tree is *in* the way.
　　d. He's *out of* sight now.
　　e. That's *in* the *center of* my *field* of vision.
　　f. There's *nothing in* sight.
　　g. I can't get *all* of the ships *in* sight at once.

例文 (1) の各文が「視界」について言っているということは，人間の英語話者ならば容易に分かることだが，ここに現れる view，sight，way，field が一般的な語彙分類で同一範疇に分類されているだろうか．

例文 (1) に現れる前置詞句の前置詞 (相当句) と名詞の関係は，メタファを通じた理解を必要とする．メタファとは，ある事柄を他の事柄を通して理解し，経験することであり [1]，単なる表層的な言語現象ではなく，人間の概念体系の本質を規定するものであり，言語運用のみならず，思考にも欠かすことができない．Lakoff らは様々なメタファ表現とその背後にある概念構造の分析を行い，人間の思考過程の大部分がメタファによって成り立っていると指摘している [1]．

例文 (1) の各文が理解される時には，VISUAL FIELDS ARE CONTAINERS[2] というメタファによって view，sight，way，field それぞれの語の意味が持つ容器の特徴に焦点が当てられ，同時にその他の面が隠されることによって，異種性と同時に類似性が際立っている．

2.2　容器のイメージスキーマ

ここで言う「メタファ」はより正確には，「イメージ・スキーマ」(image schema) と呼ばれる．「イメージスキーマ」とは，内・外，上・下などのごく基本的な物理的空間に関して人間が持つ，経験に基づく概念レベルでの理解のことであり，人間の概念体系の中でもより根源的なものであり [1]，中心的なものからより周縁的なものへの意味拡張の基礎を成す [2]．VISUAL FIELDS ARE CONTAINERS は容器のイメージスキーマの一つである．Lakoff らによれば，人間は最も基本的な習性の一つとして，自然界の物理的境界が存在しないところにも境界を想定し，内側と境界面を持つような領域に区分する (容器を規定する) という [1]．この基本的な認識に基づいて，人間は言語を含むあらゆる事象を同様に認識するということである．

容器のイメージスキーマは，例文 (1) に見られるように，英語ではしばしば前置詞句によって表現される．本来は物理的な場所や位置を表すものであった前置詞が抽象的な状態の表現にも使われているのである．容器のイメージスキーマを表現する典型的な前置詞 (相当句) は，*in*，*inside*，*into*，*out of*，*outside* などである．これらの前置詞の補部になる名詞は，容器とみなすことができ，共通の意味素性を共有する範疇を形成すると予測できる．そのため，このような例では，内容語を中心にした一般の語彙分類とは異なり，機能語を中心にした分析が有効である．本稿では，これらの前置詞の補部の名詞を収集することで，容器のイメージスキーマでとらえられる名詞を収集できるという予測のもと，*in/inside/into/out of/outside* の各前置詞 (相当句) の補部名詞を収集し，分析する．

3　分析の方法

3.1　コーパスの利用

コーパス言語学 [3][4][5] の成果に見られるように，質的な分析に加えて，数量的な分析も可能なコーパスを分析対象とすることで，これまで見えなかった語と語の新たな関係が明らかになる．

本稿ではコーパスから前置詞の補部の名詞を収集し，その性質を検討し，考察を行う．具体的には整形済みのデータを一行ずつ読み込み，前置詞句 (本実験では *in/inside/into/out of/outside* が主要部となる句) がある場合は，構文解析を行って，補部を特定し，その名詞の前置詞句内および全コーパス内での出現度数を求めて検定・統計処理を行う．

本稿では，New ICAME Corpus Collection CD-ROM[3] に

An Experiment in Classifying Nouns Associated with "Metaphorical Prepositional Phrases"
ISHII, Yasutake, SANO, Hiroshi
Faculty of Foreign Studies, Tokyo University of Foreign Studies, 3-11-1, Asahi-cho, Fuchu-shi, Tokyo, 183-8534, JAPAN

[1] 斜体は原文のままだが，下線は著者が付した．

[2] メタファはメタ言語なので，通例大文字で表記される．

[3] http://www.hit.uib.no/icame/cd/参照．

収録されている Brown Corpus[4]，LOB Corpus[5]，Frown Corpus[6]，FLOB Corpus[7] を用いた．語数はいずれも約 100 万語，計約 400 万語である．

3.2 構文解析の利用

ある語句がどのような文脈の中で用いられているのかを示すプログラムであるコンコーダンサ (concordancer) には数多くのものがある [4][5]．しかしコンコーダンサは一般に統語構造は考慮しないため，前置詞の (修飾語などを除いた) 補部名詞のみを収集する本実験では利用できない．そこで，本実験では，前置詞の補部の抽出のために構文解析プログラム (Satoshi Sekine 氏製作の Apple Pie Parser(version 5.9)[8]) を利用する．構文解析処理は 100 %の解析率でなく，文が長いほど精度が低下することが Apple Pie Parser のマニュアルにも記載されているが，句のレベルでの解析はある程度精度が保たれており，本実験での利用には問題ないと判断した．

3.3 統計処理

前置詞の補部の名詞の頻度 (絶対度数ではなく相対的な出現の割合) と，それらの名詞のコーパス全体における頻度から，式 (1) を用いて差異係数を求める．

$$\text{差異係数} = \frac{Freq._A - Freq._B}{Freq._A + Freq._B} \quad (1)$$

$Freq._X$: x の頻度

式 (1) の差異係数は-1 から+1 までの間の値をとり，正の値で+1 に近いほど，その語が A の集合により多く出現し，負の値で-1 に近いほどその語が B の集合により多く出現するということを示す．

次に，集計結果の度数 (出現回数そのもの) に対して，式 (2) によりカイ 2 乗検定を行う．

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

O: 観測値，　E: 期待値

式 (2) より求めた値が，有意水準 0.01，自由度 1，集合の母数 ∞ の場合のカイ 2 乗値 6.6349 よりも大きい場合には，1%レベルの統計的有意性が認められる．この場合，差異係数に α を添えて示す (表 1・表 2・表 3 参照)．

特定の語彙の分布の集中度を確認するためのもう一つの手段として，積率相関係数を利用する．データが (x_1, y_1), (x_2, y_2), $\cdots$, (x_n, y_n) で与えられた場合の 2 変数 x と y の間の相関係数は，式 (3) で定義される．

$$R_{xy} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}} \quad (3)$$

$\overline{x}$: x の平均値

相関係数は-1 から+1 の間の値をとり，その絶対値が 1 に近ければ強い相関があり，0 に近ければ相関がないことになる．

[4] The Standard Corpus of Present-Day Edited American English の通称．1961 年のアメリカ英語のコーパス．

[5] The Lancaster-Oslo/Bergen Corpus of British English の通称．1961 年のイギリス英語のコーパス．

[6] The Freiburg-Brown Corpus of American English の通称．1990 年代のアメリカ英語のコーパス．

[7] The Freiburg-LOB Corpus of British English の通称．1990 年代のイギリス英語のコーパス．

[8] http://cs.nyu.edu/cs/projects/proteus/app/参照．

4 実験

4.1 分析処理

本実験で利用した四つのコーパスは，テキストに関するヘッダ情報を含み，一定の文字数を超えると強制的に改行され，文中にもタグを含む形式であるため，Perl スクリプトでタグを除去し，1 文/1 行にした．このデータを Apple Pie Parser で分析し，その結果をフィルタリングすることで，*in/inside/into/out of/outside* の補部の名詞 (句) を抽出し，集計する．その際の問題と対応を以下に示す．

- 単複の語形の差 … 手作業で吸収し，同一視した
- 複数語から成る固有名詞を構成する語の扱い … 集計対象外とした[9]
- 特殊文字の問題 … ウムラウトなどの記号がついた文字は記号のつかない対応するアルファベットに置き換えた

4.2 名詞の集計

頻度の計算のためには，コーパス中の名詞の語数を数える必要がある．そのために，コーパスの品詞タグ付き版[10] に対してタグを利用して正規表現でパターンマッチを行って名詞の数を数えた．

その結果，Brown Corpus の名詞は 210341 語，LOB Corpus の名詞は 176571 語，計 386912 語であった．

Frown Corpus と FLOB Corpus には品詞タグ付き版が存在しないため，両コーパスの名詞の数については，Brown Corpus と LOB Corpus の名詞の数の合計 386912 語と同じ出現の度合いであると想定し，全てのコーパス中の名詞の数を 773824 語として計算することにした．

4.3 実験結果

表 1 には，*in* の補部として出現した名詞のうち，200 度以上の頻度で出現したものを挙げている．複数形での使用の方が単数形での使用より多い語については，複数形で示している．なお，プログラムによって認定された *in* の補部名詞は計 68478 語で，相関係数は 0.2568 であった．定量的観点から見ると，概して差異係数が高く，“life”，“school”，“words” を除く全ての語で 1%レベルの有意性が認められることから，差異係数が高い特定の語彙の *in* の補部への集中が認められる．また，相関係数も低く，*in* の補部中とコーパス全体という二つの環境が独立性を持っていると言える．

表 2 には，*into* の補部名詞として出現した名詞のうち，20 度以上の頻度で出現したものを挙げている．プログラムによって認定された *into* の補部は計 5948 語で，相関係数は 0.0207 であった．定量的には，差異係数が高い名詞が多く，ほとんどの語で 1%レベルの有意性が認められことから，やはり特定の語彙が *into* の補部として分布していると認められる．相関係数もかなり低く，*into* の補部の特殊性が明らかになった．

表 3 には *out of* の補部として出現した名詞のうち，10 度以上の頻度で出現したものを挙げている．“bounds” は複数形のみでの出現であった．なお，プログラムによって認定された *out of* の補部は計 1810 語で，相関係数は 0.3019 であった．定量的には，ほとんどの名詞の差異係数が非常に高く，“life” を除く全ての語で 1%レベルの有意性が認められることから，特定の語彙が *out of* の補部として分布していると認められる．やはり相関係数も低い．

[9] 例えば “the United States の “States” は “state” としては集計しなかった．

[10] ICAME の CD-ROM の TEXTS/BROWNTAG/BTL.TXT と TEXTS/LOBTAG/ディレクトリ内の 15 個のテキストファイル．

表 1: *in* の補部となる語彙の出現度数と差異係数

語	*in* 補部中	コーパス中	差異係数
way	1015	2123	0.69α
case	990	2502	0.63α
years	828	7424	0.12α
fact	745	2051	0.61α
world	514	3104	0.30α
country	509	1140	0.67α
terms	472	1521	0.56α
area	465	1901	0.47α
place	459	1285	0.60α
front	396	965	0.65α
addition	388	570	0.77α
order	382	1626	0.45α
life	352	3505	0.06
mind	337	846	0.64α
sense	331	1242	0.50α
form	324	1989	0.30α
time	321	7800	-0.37α
hand	317	1501	0.41α
part	313	2735	0.13α
London	296	1184	0.48α
field	291	1164	0.48α
days	282	4281	-0.15α
century	277	1461	0.36α
direction	242	533	0.67α
England	242	830	0.53α
state	222	3677	-0.19α
room	220	958	0.44α
school	219	2241	0.05
words	217	2186	0.06
position	204	1105	0.35α
town	202	649	0.56α

表 2: *into* の補部となる語彙の出現度数と差異係数

語	*into* 補部中	コーパス中	差異係数
account	87	789	0.87α
room	84	958	0.84α
water	52	971	0.75α
house	43	1458	0.59α
world	43	3104	0.29α
hands	39	1501	0.54α
car	35	732	0.72α
eyes	33	2038	0.36α
bed	32	378	0.83α
life	31	3505	0.07
office	31	603	0.74α
place	30	1285	0.50α
air	29	1058	0.56α
kitchen	29	366	0.82α
something	29	1939	0.32α
force	28	1226	0.50α
pocket	28	159	0.92α
action	27	585	0.71α
bedroom	27	217	0.88α
groups	27	2064	0.26α
categories	26	220	0.88α
night	25	1722	0.31α
street	25	1179	0.47α
town	25	649	0.67α
darkness	24	171	0.90α
parts	23	2735	0.04
country	22	1140	0.43α
face	21	1801	0.21
hall	21	650	0.62α
space	21	621	0.63α
corner	20	431	0.72α
field	20	1164	0.38α
head	20	1843	0.17
trouble	20	528	0.66α

5 考察

5.1 結果の考察

in の結果を見ると，頻繁に用いられるイディオムやコロケーションの中で用いられることが多い名詞が多く挙がっている (“*in* fact”や “*in* addition”，“*in* order to”など) のは当然の結果として，直観的には内側と境界面を持つような領域を表現する語が多く挙がっていると言える．その典型的なものが “country”や “room”などの本当に容器のような空間を指す語であり，比喩的にある空間を指し示す “hand”や “position”，さらに比喩的な “case”や “mind”，状態を表す “life”や “state”といった語が見られる．同時に，典型的に容器のイメージスキーマでとらえられるとは言えない語 (“year”などの時を表す語，“order”，“direction”など) も挙がっている点が注目に値する．これは，*in* はその意味の幅が広いため，容器のイメージスキーマのメタファにばかり使われるわけではないためと考えられる．差異計数が *into* や *out of* の結果と比べると低めなのもこれが理由であると考えられる．

into の結果を見ると，こちらもコロケーションの中で用いられることが多い名詞が多く挙がっており (“take *into* account”など)，やはり直観的には内側と境界面を持つような領域を表現するととらえられる語が挙がっている．典型的なものは *in* の場合と同じく，“room”や “town”などの本当に容器のような空間を指す語であり，比喩的にある空間を指し示す “bed”や “office”，さらに比喩的な “category”や “darkness”といった語が見られる．

out of の結果を見ると，やはりコロケーションの中で用いられることが多い名詞が多く挙がっている (“*out of* sight, *out of* mind”や “*out of* date”など) のは当然として，直観的には

表 3: *out of* の補部となる語彙の出現度数と差異係数

語	*out of* 補部中	コーパス中	差異係数
sight	35	366	0.95α
way	35	4428	0.54α
car	31	1379	0.81α
house	31	2699	0.66α
window	31	716	0.90α
room	26	1883	0.71α
place	22	2440	0.59α
bed	21	751	0.85α
hand	21	2793	0.53α
mind	20	1452	0.71α
town	20	1164	0.76α
pocket	16	292	0.92α
office	15	1259	0.67α
country	14	2210	0.46α
control	13	1158	0.66α
date	13	519	0.83α
touch	13	357	0.88α
water	13	1762	0.52α
business	12	1341	0.59α
money	11	1321	0.56α
reach	11	432	0.83α
action	10	1101	0.59α
bounds	10	33	0.98α
experience	10	1158	0.57α
life	10	3505	0.10

内側と境界面を持つような領域を表現する語が多く挙がっていると言える．その典型的なものが “car”や “house”などの本当に容器のような空間を指す語であり，その空間の境界面に焦点を当てた “window”や “bounds”，比喩的にある空間を指し示す “sight”や “reach”，さらに比喩的な “mind”や “action”，状態を表す “business”や “life”といった語が見られる．

in の結果と *into* の結果を比べてみると，後述する *into* と *out of* の比較に比べて両方に出現する名詞が少ないことがわかる．*in* と *into* の両方に出現する名詞は，country，field，hand，life，part，place，room，town，world の 9 語である．*in* も *into* もともに容器のイメージスキーマの表現に使われるのだが，基本的に *in* は静的な状態の表現に，*into* は動的な動作の表現に用いられることが多いために，両者の間に差が現れると考えられる．

続いて *in* の結果と *out of* の結果を比べてみると，両者に共通に出現する名詞は，country，hand，life，mind，place，room，town，way の 8 語である．差異計数については，両者を比較すると *in* と比べて *out of* の方は概してかなり高い．*out of* は意味の場が狭いために，その補部には容器のイメージスキーマでとらえられる名詞が集中するためと考えられる．

into の結果と *out of* の結果を比べてみると，両者に共通の名詞の多さが目に付く．具体的には，action，bed，car，country，hand，house，life，office，place，pocket，room，town，water の 13 語である．ともに動的な動作の表現に使われることが多いためと考えられる．また数値的には，*into* の方は *out of* の結果と比較すると差異係数が低めの語が多い．*out of* の反義語として，*into* と *in* の両方が用いられることから，分布が両者に分散したことが一つの理由と考えられる．また両者の絶対度数からも分かるように，「～の中から」という表現よりも，「～の中に」という表現の方が絶対的に多く用いられている．そのため，名詞の分布も *into* の方が多くなる，つまり *into* の補部になる名詞の方が多様であるということも理由の一つであると考えられる．

in/*into*/*out of* の全ての補部として高い頻度で現れる語としては，country，hand，life，place，room，town がある．これらの語を見ると，容器のイメージスキーマでとらえられる典型的な容器的空間を指す語が多い．

inside と *outside* の結果はデータの規模がかなり小さく，信頼できるデータではないので挙げていないが，やはり容器のイメージスキーマでとらえられる語が多い．注目すべき点は，他の前置詞の補部位置には出現しない “window”が *outside* と *out of* の補部位置にはかなりの頻度で出現しているということである．“window”は容器の空間の境界面に焦点を当てていると言え，容器の内側に視点の中心を置き，そこから容器の外を見ると通路となる “window”や “door”(*outside* の補部として多く出現) が目立つが，容器の外の大きな世界から容器を見る場合には，その容器の特定の部分よりも，容器全体の方が目に付くために，このような結果になったと考えられる．

5.2 意味群別に分類した語彙分類との比較

WordNet[11] を利用して，実験の結果として *in*/*into*/*out of* に共通に出現した名詞の概念区分を調べた．概念の上位関係は一つの語でも語義によって異なるが，最も代表的・一般的な語義ということで，各語の”Sense 1”として表示される語義を選択し，最上位からの分類を調べた．その結果を以下に示す．

country: / entity, something/ object, physical object/ location/ region/ district, territory/ administrative district, administrative division, territorial division/

hand: / entity, something/ part, piece/ body part/ external body part/ extremity/

life: / state/ being, beingness, existence/

place: / entity, something/ object, physical object/ location/ point/

room: / entity, something/ object, physical object/ artifact, artefact/ structure, construction/ area/

town: / entity, something/ object, physical object/ location/ region/ geographical area, geographic area, geographical region, geographic region/ urban area/ municipality/

メタファの観点からは容器のイメージスキーマでとらえられる語ということでグループ化できる語について，country，place，town については「場所」(location) の範疇に分類されているが，語彙分類データからはそれ以上の共通点項目を見出すことができない．逆に言えば，既存の語彙分類では同一範疇に分類されない語をグループ化することに，本実験では成功したということになる．

5.3 今後の課題

頻繁に用いられる他のメタファ(導管のメタファや道のメタファなど) について同様の調査を実施する予定である．また，MI-score[12] や t-score[13] などの尺度も研究に取り入れたり，イディオム/コロケーション/自由結合の別や実験の結果挙がる名詞群中の抽象化の度合いの差も考察の対象としたい．

構文に基づくクラスタリングの試みも提示されている [6] ことから，前置詞句以外のメタファ表現についても目を向け，構文・意味の両面から分析を行うことも有効だと考えられる．

これまで EU 圏外には公開されていなかった 1 億語の規模の The British National Corpus[14] が近々EU 圏外にも公開される予定であり，これを対象に実験を行うことで，まれにしか見られない例の分析もできると考えられる．

メタファ表現の処理手法の提案を目標としながら，メタファ表現の意味を説明可能な語彙分類モデルを構築し，クラスタリングを利用した，語の自動分類を行うことを目指したい．さらに，意味的類似性を距離とする連続した意味空間にマッピングした「意味マップ」の有効性とその自動構築法なども提示されており [7]，そういった別の視点についても調査をしながら研究を進めていきたい．

参考文献

[1] Lakoff, G. and Johnson, M. 1980. *Metaphors We Live By*. The University of Chicago Press.

[2] Lakoff, G. 1987. *Women, Fire and Dangerous Things: What Categories Reveal about the Mind*. The University of Chicago Press.

[3] McEnery, T. and Wilson, A. 1996. *Corpus Linguistics*. Edinbugh University Press.

[4] 齊藤俊雄・中村純作・赤野一郎 (編) (1998) 『英語コーパス言語学 -基礎と実践-』 研究社出版.

[5] 鷹家秀史・須賀廣 (1998) 『実践 コーパス言語学 英語教師のインターネット活用』 桐原ユニ.

[6] Hogenhout W. R and Matsumoto Y. 1998. *A Method for Syntactic Befavior Analysis*. 自然言語処理, 5(2). (pp. 25-46).

[7] 馬青・神崎享子・村田真樹・内元清貴・井佐原均 (2000) 『自己組織型神経回路網モデルを用いた日本語意味マップの構築』 言語処理学会第 6 回年次大会発表論文集 (pp.332-335).

[11] Princeton University で，同義語の集合によって語の意味を表現しようとするアプローチで製作されている辞書．http://www.cogsci.princeton.edu/~wn/参照．

[12] MI(Mutual Information)-score は，コーパス内での特定の 2 語の連想関係の強さを示す．一方が現れた場合に他方の出現をどの程度期待できるかということを表し，意味的な側面に焦点が当てられる．

[13] t-score は，特定の 2 語の共起頻度の変動が統計的に有意であるという確信度を示す数値で，文法的な側面も含めて，共起頻度に焦点が当てられる．

[14] http://info.ox.ac.uk/bnc/参照．