辞書定義文を用いた複合語分割
— 語構成情報の抽出と考察 —

村田 真樹　内山 将夫　井佐原 均

郵政省 通信総合研究所 関西先端研究センター 知的機能研究室

1 はじめに

辞書定義文はこれまで数多くの研究[1, 2, 3, 4, 5, 6, 7]で用いられており，有益な情報をたくさん含んでいるものである．本稿では，この辞書定義文を単語の語構成[8]の考察に用いる方法について述べる．例えば，「アマチュア無線」という見出し語の定義文は，EDR辞書では「アマチュアによる無線通信」となっている．このことにより，「アマチュア無線」という語は，「アマチュア」と「無線」の二つの構成要素からなる語であり，その構成要素間の関係(意味関係)が「による」という語によって表されるものであると推察できる．本稿での語構成の考察手法はおおよそ上記のとおりで，見出し語と定義文を照合することにより，その語の構成要素とその構成要素間の関係を調べるものである．

本稿では実際に上記の照合を行ない，約8,000の構成要素とその要素間の関係の情報を抽出し，要素間の関係にどのようなものがあるのかを調べている．この抽出したデータの精度は97～99%という良質なものであり，様々な語構成の考察に利用可能だと思われる．本稿では，実際にこのデータが語構成の考察に役立つことを示すために，一例として意味素性を用いた考察を行なっている．そこでは，要素間の関係が「と」ならば，二つの構成要素には良く似た意味素性のものがきやすい，また，「で作った」ならば後ろの構成要素には生産物のものがきやすいなどといった結果が得られている．また，本稿で採用した構成要素間の関係は，「による」などの言語表現であり，場合によっては良く似た意味関係，例えば，「に用いる」「用の」が異なる意味関係として抽出されてしまう問題がある．本稿では，この問題を解決するためにクラスタリング手法を用いてよく似た関係をまとめることも行なっている．

以降，次節で見出し語と定義文の照合による構成要素とその要素間の関係のデータの抽出について述べ，その次の節でその抽出したデータを用いた意味素性による考察について述べる．

2 見出語と定義文の照合による語構成情報の抽出

見出し語と定義文の照合は以下の手順で行なう．(ただし，簡単のため，ここでは二つの構成要素からなる複合語しか扱わない．)

1. 見出し語を単語Aと単語Bに分割する．(つまり，単語Aと単語Bをくっつけると見出し語にもどる．) 単語A, Bともに辞書の見出し語にあることを条件とする．
2. 定義文をJUMAN[9]で形態素解析して，複数の形態素に分割する．
3. 最後に下記の評価式(1)の値が最も大きくなるように，単語Aと単語Bを定義文中の分割された各要素のいずれかと対応づける．ただし，単語Aと単語Bが同じ要素に対応することがないようにしておく．また，単語Aと単語Bに対応する要素がそのままの形で同順連接で出現しないことを条件とする．(これは，単語A,Bに対応する要素が単語A,Bの形のまま同順連接で出現する場合，分割箇所の確証が得られないためである．)

$$\text{評価式} = \text{類似度} + \text{近接度} \times 0.0001 + \text{同順度} \times 0.00000001 \quad (1)$$

ここで，「類似度」は単語Aと対応づけられた要素との類似度の二乗と，単語Bと対応づけられた要素との類似度の二乗の和の平方根の値で，「近接度」は，単語Aと対応づけられた要素と単語Bと対応づけられた要素の定義文中での距離に-1をかけたもので，「同順度」は定義文において単語Aと対応づけられた要素が単語Bと対応づけられた要素に先行する場合1で，そうでない場合-1である．単語間の類似度の計算にはEDR概念辞書を用いる[1]．

また，本研究では簡単のため，見出し語と定義文の対応づけでは，名詞同士しか対応させず，単語A,Bと定義文中の動詞を対応づけるということは行なっていない．

また，見出し語の分割が曖昧な場合や定義文が複数ある場合は，上記の評価式(1)の値が大きくなるものを選ぶことにする．

上記照合の結果，単語Aや単語Bに対応する要素が見出し語と同義語であるもの，単語Aと単語Bに対応する要素が同順連接で出現しているもの(ただし上述の条件で単語Aと単語Bの形のまま連接しているものは出現しない)については，評価式

[1] X と Y の単語間の類似度の求め方について記述する．この類似度は，EDR概念辞書のトップノードと単語Xのノードの間の枝の数をnx，トップノードと単語Yの間の枝の数をny，単語X，単語Yからのトップノードへのパスで初めてパスが一致するノードをZとし，ノードZとトップノードの間の枝の数をnzとすると，$(nz + nz)/(nx + ny)$で与えられる．この式は，nx, nyに対してnzの値の大きさの割合をとったものである．nx, nyに対してnzの値が大きいとき，シソーラス中でのノードZの位置が相対的に下の方にあることになり，単語Xと単語Yの類似度が高いことを意味する．この手法は文献[10]の方法を利用している．

表 1: 対応づけの精度

	精度	積算総数
評価値の最上位の 100 個	97%	100 個
評価値 0.98 以上での下位の 100 個	99%	7,978 個
評価値 0.95 以上での下位の 100 個	82%	9,355 個
評価値 0.9 以上での下位の 100 個	26%	15,028 個

(1) の値を特別に 0 にして抽出しにくいようにしておく．この処理をしたのは，対応する要素が見出し語と同義語であるものは不自然であったり，また，同順連接で出現しているものは語形がかわったとしても照合失敗の可能性が高いものであったりしたからである．

上記の方法で実際に対応づける実験を行なった．入力の見出し語としては，JUMAN の名詞辞書 (Noun*.dic, 総数 187,109 単語) を用いた．このうち，EDR の単語辞書により二つの構成要素に分割可能であったものは，94,878 個であった．実験結果を表 1にあげる．表では，評価式 (1) の値の大きいところから適宜 100 個ずつサンプリングして人手で調べた対応づけの精度を示している．評価式 (1) の値の大きいところは照合の度合いが高く，精度よく対応づけがなされると思われるが，実験結果もそのようになっている．なお，評価値 0.98 以上は，今回の実験では評価式中の類似度の値が 1 のもののみ，つまり，語 A,B ともに，同義語かそのままの単語が定義文中にあらわれている場合のみであった．

対応づけを失敗した原因は主に以下の三つであった．

- EDR の定義文記述が「『見出し語』という」という記述をしており語構成を考察する余地がないもの．

  (例) 弓始め — 弓始めという武家の儀式

- 定義文中に対応する要素がない．

  (例) 切符 — 支払いの証拠となる札

- 定義文中に対応する要素があったが，それは動詞や複数の語で表現されており，本研究の手法では対応できなかった．

  (例) 尾 翼 — 飛行機の 胴体後部 についている翼

  「胴体後部」が複数の語で構成されており，現状の同義語辞書・シソーラスでは「尾」との類似性をはかる術がない．

次に上位で高精度に対応づけられたデータから，語構成の構成要素間の関係情報を抽出する．ここでは，評価値 0.98 以上のデータ (7,978 個，精度は 97 ～ 99%) を利用することにする．構成要素間の関係は，対応づけられた単語 A と単語 B の出現順序とその二語間の文字列によって表されるものと定義する．得られた構成要素間の関係を表 2にあげる．表中「逆－」のついた分類は単語 A と単語 B に対応する要素の定義文での出現が逆順であったことを示す．表では上位 20 個しかあげていないが，抽出された関係の数は 3,505 個であった．本研究では，人手で各語の語構成をわざわざ調べなくとも，定義文を用いることで表 2にあげるような構成要素間の関係が存在することを自動で得たことになる．ただし，本研究では分割されうる 10 万のうちの 8,000 のものしか用いていないので，関係の出現頻度などが片寄っている可能性がある．また，「が」などの助詞一文字が挟まっていたという関係では，「県有」(県 が 所有していること) のような関係の他に，「軍陣」(軍隊 が 陣を設けるところ) のような関係もあり，まだ構成要素間の意味関係を明確に区別できていない．このような場合は，単語 A,B の間の文字列だけでなくまわりの文字列も用いた方がよいと思われる．

表 2: 語構成の構成要素間の関係 (上位 20 個のみ)

関係	事例数	出現率	例 (括弧内は定義文)
の	1,589 個	19.92%	腕力 (腕 の 力)
と	354 個	4.44%	仏神 (仏 と 神)
つの	147 個	1.84%	八難 (八 つの 災難)
逆－の	138 個	1.73%	内もも (もも の 内側)
を	87 個	1.09%	氷献上 (氷 を 献上すること)
する	65 個	0.81%	記憶術 (記憶 する 方法)
で作った	62 個	0.78%	木刀 (木 で作った 刀)
用の	55 個	0.69%	偵察機 (偵察 用の 飛行機)
逆－を	53 個	0.66%	発熱 (熱 を 発生する)
で	44 個	0.55%	銃撃 (銃 で 攻撃する)
に	41 個	0.51%	頭熱 (頭部 に 熱があること)
に用いる	39 個	0.49%	揚げ油 (揚げ物 に用いる 油)
をする	39 個	0.49%	作業場 (作業 をする 場所)
にある	36 個	0.45%	外庭 (建物の外 にある 庭)
を入れる	33 個	0.41%	薬箱 (薬 を入れる 箱)
に関する	31 個	0.39%	文章論 (文章 に関する 論)
逆－が	29 個	0.36%	筒形 (形 が 筒状であること)
が	25 個	0.31%	県有 (県 が 所有していること)
製の	24 個	0.30%	金器 (黄金 製の 器)
のある	23 個	0.29%	雪月夜 (雪 のある 月夜)

3 語構成の意味素性による考察と分類

前節までの議論のように，辞書定義文を用いることで語構成の構成要素とその要素間の関係のデータを比較的容易に抽出することができることがわかった．本節では，この抽出したデータの有用性を確かめるために，一例として構成要素の単語 A,B の意味素性を用いて語構成を考察することを試みる．(この考察は，文献 [11] を参考にしている．) ここでは意味素性としては表 3のもの [12] を用いる．つまり，分類語彙表の分類番号の上位桁に応じて意味素性をふる．また，各意味素性は，さらに，名詞か動詞か形容詞かに応じて「体」「用」「相」の品詞的な分類もつける．この品詞的な分類の付与には，分類語彙表にもともとある一桁めの分類は用いず JUMAN の品詞推定機能を用いてい

表 4: 構成要素間の各関係における構成要素 A と構成要素 B の意味素性の組の出現の割合

「と」の関係 例：「母子」(母親と子供)			「する」の関係 例「補佐人」(補佐する人)			「で作った」の関係 例「麻縄」(麻で作った縄)			「用の」の関係 例「蝿取紙」(蝿取用の紙)		
意味素性		出現率	意味素性		出現率	意味素性		出現率	意味素性		出現率
単語 A	単語 B		単語 A	単語 B		単語 A	単語 B		単語 A	単語 B	
人間	人間	7.46%	用・活	活動	13.90%	未定義	生産物	19.35%	用・活	生産物	14.24%
現象	現象	5.66%	用・活	人間	7.23%	植物	生産物	17.47%	生産物	生産物	8.18%
活動	活動	5.24%	用・活	組織	6.79%	生産物	生産物	10.19%	人間	生産物	7.73%
未定義	未定義	4.24%	用・抽	数量	6.15%	現象	生産物	10.08%	未定義	生産物	7.27%
生産物	生産物	4.20%	用・活	数量	6.08%	植物	未定義	9.68%	活動	生産物	6.59%

表 3: 分類語彙表から作成した意味素性

意味素性	分類語彙表の分類番号の上位 3 桁
動物	[1-3]56
人間	12[0-4]
組織	[1-3]2[5-8]
生産物	[1-3]4[0-9]
体の一部	[1-3]57
植物	[1-3]55
自然物	[1-3]52
空間	[1-3]17
数量	[1-3]19
時間	[1-3]16
現象名詞	[1-3]5[01]
抽象関係	[1-3]1[0-58]
人間活動	[1-3]58,[1-3]3[0-8]
その他	4
未定義	分類語彙表にない語

る．(サ変名詞は「用」にしている．) この意味素性を用いて，表 2で得られた 20 の各分類において，単語 A と単語 B にどういう意味素性の語がきやすいかを求めた．このとき，分類語彙表において多義になっている語は，各語義が 1/(語義数) の頻度で出現したものとして扱った．この調査結果のうち，代表的なものを表 4にあげる．表では「用」のものは「用・」をつけている．これがないものは「体」のものである．

表から，「と」の関係の場合は単語 A と単語 B には同じ意味素性のものがきやすいことがわかる．また，「する」の関係の場合には，単語 A には「用」の名詞，つまりサ変名詞がくることがわかる．また，「で作った」「用の」の関係の場合は，単語 B に「生産物」がきやすいことがわかる．その他にも，この調査方法で様々なことがわかることだろう．

次に，意味素性の情報を用いて表 2にあげたような構成要素間の関係をクラスタリングすることを試みる．これは，現状の構成要素間の関係が文字列によって表されるもので，例えば「用の」と「に用いる」は同じ意味だと思われるが，別個の分類とされている問題を解決するために行なう．つまり，クラスタリングによって，この「用の」と「に用いる」などの似た分類をまとめる．ここでは上位 30 個の構成要素間の関係を用いる．クラスタリングのアルゴリズムには，重心

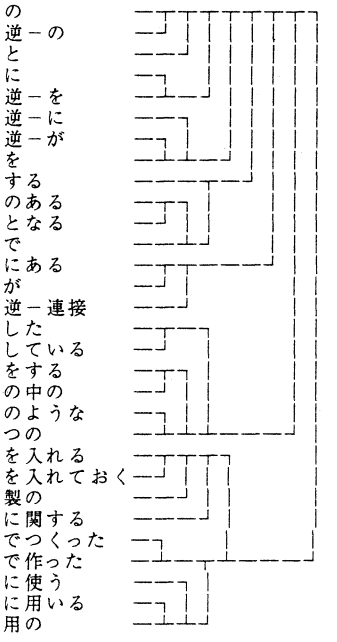

図 1: 語構成の分類のクラスタリング結果

間の距離を用いるボトムアップ法を用いる．ただし，重心の算出には各関係の事例数を重みとしてかけて行なう．また，各関係間の類似度は，単語 A と単語 B の意味素性の組を次元としたベクトル間の角度の逆数とする．この結果を図 1に示す．(図中の「逆－連接」は単語 A と単語 B が定義文で逆順で連接している場合の分類を意味する．)

図では，「に使う」「に用いる」「用の」や，「でつくった」「で作る」や，「した」「していた」など，それなりに似ていそうな意味関係がまとめあげられている．言語表現を意味関係に用いる場合，関係の種類が発散しがちになるが，そういう場合にはクラスタリングのテクニックが役に立つと思われる．なお，ここでは意味素性や分類を仲介にしてクラスタリングを行なったが，事例でクラスタリングしてもよい．

4 関連研究

本稿の2節の手法は，辞書定義文を用いるところ，構成要素の関係(意味関係)を「による」などの文字列で表すところ[2]が，黒橋らの辞書定義文を用いた「AのB」の意味解析[7][3]と似たものとなっている．しかし，黒橋らの方法ではBの語で辞書の定義文をひくために，Aの語が定義文に現れる必然性はなく，Bの語の定義文にAの語を意味するもの，もしくは，Aの語と関係のあるものが出現することを期待するものであるが，われわれの方法ではAとBの二つの語からなる語「AB」自体で辞書定義文をひくため，構成要素の二つのA,Bともに辞書定義文に出現することがほぼ期待されるものとなっている．

複合語的な語を辞書定義文で分割する先行研究として，徳田らのもの[6]がある．徳田らは，手話の翻訳において，「白波」という語が手話用辞書で未定義語の場合は，「白波」の定義文「白い波」を用い，手話用辞書にでものっている「白い」と「波」という語に分割して手話翻訳を実現している．語を定義文で分割するという意味で，本研究の手法は徳田らのものと非常によく似ている．しかし，徳田らのものは，日本語から手話への翻訳を目的としており，語構成を考察しようとする本研究の目的と異なる．

複合語をコーパスを利用して解析する研究として，久光らのもの[17]がある．この研究では「AのB」「Aに関するB」といったテンプレートを人手で用意しておき[4]，そのテンプレートにマッチするものを共起情報として蓄えて統計的に複合語解析を行なっている．われわれの方法では辞書に載っている複合語しか対応できないが，この久光らの方法は辞書に載っていない複合語にも対応できる利点がある．ただし，テンプレートを人手で用意する必要があるという問題があるし，構成要素間の意味関係を自動抽出することができない．また，コーパスを利用する方法では「AB」とそれと照合させる「A…B」の間に必ずしも意味の等価性が存在するとは限らないが，辞書を利用する場合「AB」とその定義文「A…B」には意味の等価性が約束されている．このため，意味関係を抽出することが目的の場合は，辞書を用いる方法の方が望ましいだろう．

5 おわりに

本研究では見出し語と辞書定義文を照合することにより語構成を考察する手法を示した．また，実際にこの手法を用いて語構成の関係を示す表現を抽出した．さらに，意味素性を用いて語構成の各関係には，どういった語が出現しやすいかを調べ，最後にクラスタリング手法を用いてよく似た関係をまとめることを行なった．

本稿の手法で抽出した語構成のデータは，構成要素となる二つの語とその関係を示す言語表現で，そのデータの精度は97～99%で良質なものであり，かつ，8,000というある一定量のものであるため，このデータを学習データとして用例ベースの方法や機械学習の方法を用いて，本稿の手法では語構成を調べることができなかったものについて構成要素の関係を推定するといったことも可能だと思われる．また，本研究の手法およびデータは複合語の解析[17, 18]にも利用できるものであると思われる．

[2]「AのB」の意味関係を動詞などの自然言語であらわそうとしている研究に文献のもの[13]もある．

[3]この黒橋らの研究は，文献[14]で示しているような名詞・動詞などの格フレームの情報やGenarative Lexicon[15]におけるQ-構造のような情報が，国語辞典の定義文に記載されていることに着目し，この定義文を格フレーム辞書やQ-構造のかわりに用いている研究ともいえる．同様の考え方を用いて，RWCにおいて辞書の定義文や例文の各語に見出し語との関係を意味するタグを付与する研究[16]もすすめられている(例：「見出し語の主語」「見出し語の目的語」「見出し語の係り先」「見出し語の唯一の引数」などのタグを定義文や例文にふる)．

[4]このテンプレート作成に，本稿で得た構成要素間の関係を利用することも考えられる．

参考文献

[1] 鶴丸弘昭，竹下克典，伊丹克企，柳川俊英，吉田将，国語辞典を用いたシソーラスの作成について，情報処理学会研究報告，Vol. 91, No. 37 (91-NL-83), (1991).

[2] 富浦洋一，日高達，吉田将，語義文からの動詞間の上位-下位関係の抽出，情報処理学会論文誌，Vol. 32, No. 1, (1991).

[3] 黒橋禎夫，長尾真，佐藤理史，村上雅彦，専門用語辞典の自動的ハイパーテキスト化の方法，人工知能学会誌，Vol. 7, No. 2, (1992), pp. 336–345.

[4] 渡辺靖彦，長尾真，図鑑の解説文から内容抽出を行なうための専門知識の構築，人工知能学会，Vol. 11, No. 3, (1996).

[5] 村田真樹，長尾真，用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定，言語処理学会誌，Vol. 4, No. 1, (1997).

[6] 徳田昌晃，奥村学，日本語から手話への機械翻訳における手話単語辞書の補完方法について，情報処理学会論文誌，Vol. 39, No. 3, (1998).

[7] 黒橋禎夫，酒井康行，国語辞典を用いた名詞句「AのB」の意味解析，自然言語処理研究会 99-NL-129, (1999).

[8] 斎藤倫明，石井正彦，語構成，(ひつじ書房，1997).

[9] 黒橋禎夫，長尾真，日本語形態素解析システム JUMAN 使用説明書 version 3.6, (京都大学大学院工学研究科，1998).

[10] 長尾真，佐藤理史，黒橋禎夫，角田達彦，自然言語処理，岩波講座ソフトウェア科学，第15巻，(岩波書店，1996).

[11] 横山晶一，佐久間一弘，意味素性を用いた複合名詞の生成による分析，計量国語学，Vol. 20, No. 7, (1996), pp. 304–314.

[12] 村田真樹，神崎享子，内元清貴，馬青，井佐原均，意味ソート msort — 意味的並べかえ手法による辞書の構築例とタグつきコーパスの作成例と情報提示システム例 —，言語処理学会誌，Vol. 7, No. 1, (2000).

[13] 田中省作，富浦洋一，日高達，統計的手法を用いた名詞句「NPのNP」の意味関係の抽出，言語理解とコミュニケーション研究会 NLC98-4, (1998), pp. 23–30.

[14] 村田真樹，長尾真，意味的制約を用いた日本語名詞における間接照応解析，言語処理学会誌，Vol. 4, No. 2, (1997).

[15] James Pustejovsky, *The Generative Lexicon*, (The MIT Press, 1995).

[16] 橋田浩一，岩波国語辞典のタギング，(1999), http://www.etl.go.jp/etl/nl/gda/iwanami.html，私信メールもあり．

[17] 久光徹，新田義彦，名詞間の意味的共起情報を用いた複合名詞の解析，言語処理学会誌，Vol. 5, No. 4, (1998).

[18] 小林義行，徳永健伸，田中穂積，名詞間の意味的共起情報を用いた複合名詞の解析，言語処理学会誌，Vol. 3, No. 1, (1996).