上位語階層データ (Verson 1.0.1)

Created: 2009/11/25
Updated: 2009/11/25,26,30, 12/21, 22, 24, 25
2010/01/13

目次

はじめに

0.1 公開履歴

(1) 2009 年 11 月 25 日: Version 1.0 公開 [上位語パスの数: 68,908]
(2) 2010 年 01 月 13 日: Version 1.0.1 誤りタグの修正 [上位語パスの数: 不変]

0.2 ファイルの一覧

- typed-paths-v1.sjis.csv.zip [文字コード Shift-JIS (Windows 環境向け)]
- typed-paths-v1.eucj.csv.zip [文字コード EUC-JP (Unix/Linux 環境向け)]
- typed-paths-v1.utf8.csv.zip [文字コード UTF-8 (汎用)]

0.3 利用条件

本データベースの利用には，(独) 情報通信研究機構と利用許諾契約を結ぶ必要があります．詳しくは，

```
http://www.alagin.jp
```

をご覧ください.

0.4 利用に関する注意

本データベースは，インターネットホームページ等，（独）情報通信研究機構以外の第三者が作成した文書等のデータから，語彙の抽出及び統計処理等によって作成されたものです．そのため本データベースの内容は，（独）情報通信研究機構の主体的な意思決定・判断を示すものではありません．本データベースの生成は，電子的な方法又は一様の選別基準による機械的判定によって行われています．そのため本データベースの内容の正確性，真実性及び相当性は一切保証されません．以上の理由により，（独）情報通信研究機構は，本データベースの内容について，責任を負いかねます．本データベースの使用に関連して生ずる損失，損害等についても，一切責任を負いかねます．本データベースには，意図せず，第三者への誹謗中傷，差別用語，個人情報などが含まれている場合があります．本

データベースを利用の際はこれらによる権利侵害に十分な注意をお願いいたします．利用者においては，本データベースの以上の特質をよくご理解の上で，本データベースをご利用下さい．

0.5 本データベースに関する問い合わせ先

独立行政法人情報通信研究機構
知識創成コミュニケーション研究センター
MASTAR プロジェクト 言語基盤グループ
email: alagin-lr@khn.nict.go.jp

以下，1 ではデータの基本的な特徴を説明し．2 でデータ構造と付与情報の詳細を説明する．3 でデータの構築法の詳細を解説します．

1　データの簡単な解説

本データは約 69,000 名詞句からなる階層的シソーラスである．日本語 Wikipedia (2007/03/28 版) から自動的な処理で得られた上位下位関係のうち，上位語として現われた約 95,000 名詞句を形態素解析の結果に従って分解，階層化し，階層を構成する名詞句のすべてについて，それらの指示対象が十分に特定されるかどうかのタグづけを行なうことで構築された[1)]．例えば「成分」という語はそれだけでは指示対象が特定されず，「食品の成分」のように「A の」の表現を補わないと適切な上位語とは見なせない．我々は，この例のように，表現を補わないと指示対象が特定されない名詞句を「未飽和」であると言うが，本データは階層的に整理された名詞句について，それが未飽和であるか否か判断をし，それによって上位下位関係における上位語として適切であるか否かを示すものである．本データの利用によって，Wikipedia から抽出された上位下位関係から不適切な上位語が取り除かれ，上位語の階層化によって，より柔軟な活用が可能となる．特に，Wikipedia から抽出された上位下位関係を日本語 WordNet (Bond et al. 2009) に接続する際に有効であることが確認されている．

なお，本データで言う「上位語」は複合名詞句を含む．例えば「市販頭痛薬の成分」のように，通常は「語」とは言われないものも単純化のために「上位語」と呼んでいる．

1.1　データの見本

次に 30 例ほど見本を示す:

(1)　a. <type=L>学</>, <type=D>科学</>, <type=G>外科学</>, <type=G>胸部外科学</>

[1)] 初めの約 95,000 名詞句から仕上がり約 69,000 名詞句への減少は，主に次のような上位語を処理から除外したことによって生じている: (1) 最上位の上位語が「達」「たち」「等」「など」「ほか」「他」「類い」「分類」「もの」「モノ」「物」「こと」「事」「コト」「名」「呼称」「総称」「通称」である行 (2) 最下位の上位語が「・」を含む行 (3) 最下位の上位語が「及び」「および」「並び」「ならび」「あるいは」「又は」「または」「とその」を含む行 (4) 最上位語と最下位語とが共有される場合，短い行．(2,3) の事例を除外した理由は，最下位語で A or B or . . . の選言 (disjunction) が含まれる名詞句は，修飾語の除去によって指示する集合が常に小さくなるという問題があったからである．上位語の階層化では，修飾語句を取り除くことで概念の一般化が起らなければならないが，選言を含む最下位語の場合にはそれが満足されない．これに関連した事態はで §3.2.6 で説明する「A 兼 B」にも起こっているが，「A 兼 B」は該当例が少なかったので，除外しなかった．時間に余裕があれば，選言を含む名詞句のすべてについて，「A 兼 B」に対して行なった上位語パスの複数化を試みるべきである．

b. <type=G>トンネル</>, <type=G>道路トンネル</>, <type=G>水底道路トンネル</>
c. <type=D>台</>, <type=G>天文台</>, <type=G>公開天文台</>
d. <type=D>ボール</>, <type=G>バレーボール</>, <type=G>イタリアのバレーボール</>
e. <type=D>料</>, <type=L>資料</>, <type=G>日本語資料</>, <type=G>スタートレックに関する日本語資料</>
f. <type=D> 手 </>, <type=L> 選手 </>, <type=L> 北九州の選手 </>, <type=G>ニューウェーブ北九州の選手</>
g. <type=D>口</>, <type=B>手口</>, <type=D>大手口</>
h. <type=G>市町村</>, <type=D>県の市町村</>, <type=G>鹿児島県の市町村</>
i. <type=D>品</>, <type=L>作品</>, <type=G>トランプに関わる作品</>
j. <type=L>成分</>, <type=G>薬の成分</>, <type=G>頭痛薬の成分</>, <type=G>市販頭痛薬の成分</>
k. <type=L>疾患</>, <type=G>消化器疾患</>
l. <type=G>車両</>, <type=L>ベース車両</>
m. <type=G>刀剣</>, <type=G>投擲用刀剣</>
n. <type=D>団</>, <type=G>財団</>, <type=G>文化財団</>, <type=L>歴史文化財団</>, <type=D>都歴史文化財団</>, <type=G>東京都歴史文化財団</>
o. <type=G>橋</>, <type=G>鉄橋</>, <type=G>正門鉄橋</>
p. <type=D> 業 </>, <type=G> 企業 </>, <type=L> 本社を置く企業 </>, <type=D> 市に本社を置く企業 </>, <type=G> 唐津市に本社を置く企業 </>
q. <type=L>セミナー</>, <type=L>啓発セミナー</>, <type=G>自己啓発セミナー</>,

一行はコンマで区切られた要素の列で，左端が最上位の上位語，右端が最下位の上位語となるような階層を表わしている．

1.2 本データの基本構造

1.2.1 表記の統一

データ処理で標準化のために全角英字数字文字を半角文字に変換してある．これは見かけの異なりを減らし，体系化の程度を増すためであるが，実際の使用 (例えば検索結果との照合) の際にはこの点を考慮に入れる必要がある．

1.2.2 階層性の表現

各行はコンマ (,) で区切られた要素の列 $X_1, X_2, \ldots X_{n-1}, X_n$ である．これは最上位語が X_1 で，最下位語の上位語が X_n であるような上位語階層を表す．一行に含まれる要素の数は行によって異なる．以下，便宜的に各行を「上位語パス」と呼び，コンマで区切られた上位語パスの要素を「(上位語パスの) パス要素」と呼ぶ．

上位語パスの右端の要素 $X_{n=\max}$，つまり最下位の上位語が Sumida et al. (2008) の手法で獲得した元の上位語である．

1.2.3 パス要素の未飽和性の指定

上位語パスの個々の要素には (未) 飽和性 (= (西山 (1990, 2003) が言う意味の (非) 飽和性) が指定してある．それを表すのに本データでは <type=*T* ...>X</> という表記を用いている．これは X の未飽和性の値が *T* であることを表わす．ラベルはパス要素ごとに付与される．

意味が未飽和な名詞を未飽和名詞と呼ぶが，未飽和名詞の正確な定義は現実にはかなり難しい．本質的なことだけを言うと，未飽和名詞は他の何らかの対象と特定の関係にある対象を表す名詞のことで，語単体では指示を特定しにくい名詞である．未飽和名詞は

(2) a. **動詞派生名詞**: 「(A の) 残り」「(A の) 守り」「(A の) 始まり」「(A の) 終わり」
b. **サ変名詞**: 「(A による)(B の) 逮捕」「(A による)(B の) 探求」「(A による)(B の) 起訴」「(A による)(B の) 占拠」「(A による)(B の) 販売」「(A による)(B の) 購入」
c. **非動詞派生かつ非サ変の述語性名詞**: 「(A の)(B への) 不満」「(A の)(B への) 反感」
d. **関係名詞**: 「(A の) 父」「(A の) 母」「(A の) 妻」「(A の) 夫」「(A の) 親」「(A の) 子」「(A の) 恋人」「(A の) 友人」「(A(藩) の) 藩主」「(A(国) の) 王 (様)」「(A(社)

の) 社長」「(A(課) の) 課長」「(A(艦) の) 艦長」「(A(館) の) 館長」「(A(大 (学)) の) 総長」「(A の) 右」「(A の) 左」「(A の) 上」「(A の) 下」「(A の) 前」「(A の) 後ろ」(「A の」がないと指示が不安定)

e. **属性名詞**: 「(A の) 容姿」「(A の) 形」「(A の) 成分」「(A の) 体積」「(A の) 質量」「(A の) 位置」(「A の」がないと一般の文脈では指示が不安定)

f. **その他**: 「(A の) 弟子」「(A の) 姉妹都市」「(A の) 端」「(A の) 中心」「(A の) 峠」「(A の) 領土」「(A の) 領地」「(A の) 国土」「(A の) 基礎」「(A の) 要旨」(「A の」がないと指示が不安定)

などを含む，より大きな名詞のクラスである．それに対し，次の例は (「A の」がなくても指示が安定する) 飽和名詞である:

(3)「人」「男性」「女性」「都市」「魚」「人魚」「岩」「砂漠」「山」「川」「海」「空」「空気」「気体」「固体」

ここでの対比からもわかるように，未飽和名詞は異なる品詞にまたがるクラスであり，その認定は意味的判断に基づいて行われる[2)]．

ただし，飽和名詞と未飽和名詞の区別が常に明白とは限らない点はあらかじめ断っておきたい．例えば「神」が飽和名詞か未飽和名詞かを (文脈に依存しないで) 決めることは難しい (仮に定義ができたとしても，その定義をタグづけ作業者全員に理解させ，それを使った評定を徹底させることは，もっと難しい)．更に，語義によって未飽和性が変わる場合もある．例えば「子供」には (i) 関係名詞としての「(誰かの) 子供」の場合と (ii)「大人」の対極語としての「幼い人間」という場合があり，前者は未飽和名詞だが，後者は飽和名詞である．

未飽和名詞は，名詞単体では指示対象が一様に決まらないので，「A の」が伴わない未飽和な形では上位語としては不適切である．このため，飽和名詞と未飽和名詞の区別の情報は，名詞階層で有用だと考えられる (例えば「成分」は未飽和名詞である．これは大抵の辞書に載っている普通の語であるが，上位語として不適切である．「シャンプーの成分」

[2)] 本データの作成では西山 (1990, 2003) を参考にしているが，そこでの定義を踏襲してはいない．少なくとも二つの点で違いがある．第一に，西山は「非飽和名詞」という述語を使い，「未飽和名詞」という述語は使っていない．この用語の違いは，名詞の飽和名詞と非飽和名詞への二値分類が可能とは限らず，程度の違いがあるかも知れないと想定しているところに由来する．第二に，本データの作成ではサ変名詞と動詞派生名詞も未飽和名詞に含めたが，これは本データ作成者の独自の判断である．サ変名詞と動詞派生名詞は西山 (1990, 2003) では未飽和名詞ではないと明示的に規定されている．この二点に関連する詳細は Kuroda et al. (2009) を参照されたい．

と「頭痛薬の成分」のあいだには，ほとんど何の共通性もない).

日本語名詞句の未飽和性を明示化した大規模なデータベースは (先の「神」のような厄介な例の扱いもあって，現状では信頼性に難があるのは認めなければならないが)，本データが初めてと思われる．

(未) 飽和性 T の値は G[ood], L[ess Good], D[ubious], B[ad], C[onnector] のいずれかである．具体的に言えば，G 評価をもつ語はそれ自体が適切な上位語となる語句であり，L は未飽和性が解消されることで適切な上位語になる可能性のある語句である．これに対し，D や B は適切な上位語とはならない．これらのラベルの概要は §2.3.1 で簡単に説明する．例外的な処理に関しては，§3 で説明する．

1.2.4　階層化の効果

(1) に示した例にある上位語パスの最下位の要素=右端の要素を見ればわかるように，元の上位語の多くは記述の詳細度が高すぎて，検索などのデータ処理に向かない．この問題を解消するため，最下位の上位語から修飾語句を段階的に除去し，段階的な一般化階層を表したものが上位語パスである．階層化においては，飛躍がないように可能な限り中間レ階層の意味カテゴリーを人手処理で拾っている．その際，自動化では取りこぼされた中間レベルも追加した (取りこぼされた中間レベルの例は §3.2.7 を参照).

階層化前の段階では，最下位の上位語は日本語 WordNet (Bond et al. 2009) と 8% 程度の対応率しかなかったが，階層化処理により，95% の上位語が上位語パスのいずれかの要素と対応をもつようになった．これにより固有名と直接結びついている下位の上位語と上位語オントロジーに定義された抽象的な用語を結びつけることが可能になったと考えられる．

2　データの詳しい説明

2.1　概要

本データは，Sumida, et al. (2008) の手法で日本語 Wikipedia から自動抽出した約 2,400,000 万個の上位語と下位語の対の，上位語のみ (約 95,000 個) を人手で階層化したものである (本データは黒田ほか (2009) で報告した作業の最終産物である).

上位語の階層化は，素の上位語から段階的に修飾語句を取り除くことで実現した．こうして得られた名詞階層の要素の一つ一つに，(未) 飽和性を属性として与えている．

本データの階層の最下位の上位語は，ALAGIN で公開中の「上位下位関係抽出ツール」

(http://nlpwww.nict.go.jp/hyponymy/index.html) で獲得される上位語に相当する．このツールで獲得される下位語に固有名の割合が高いことを考えると，本データは Web 文書から獲得される固有名に有意味な上位語を与えるデータになると期待できる．

2.2 本データの利用法

本データの利用法は様々であるが，次のような用途が典型的だと考えられる．(市販頭痛薬の成分, X) という対が Web データから獲得されたとする．X の意味の推定するのにシソーラスを使うことが考えられるが,「市販頭痛薬の成分」という表記のままではたいていのシソーラスにはマッチしない．シーラスは原則として単語のみを登録しているからである．その一方,「市販頭痛薬の成分」から単に主要部名詞 (e.g., 成分) だけを取り出しても，有用性は低い．その理由は二つある．

第一に,「成分」という名詞はシソーラスにマッチしたとしても，一般的すぎる．「何かの成分であるもの」というカテゴリー C を作れば，(市販頭痛薬の成分, X) と (シャンプーの成分, Y) から X と Y とが共に C の要素であることが帰結する．だが，このような類似性は精度の低い分類しか与えない．X の類例を探すのであれば,「頭痛薬の成分」や「薬の成分」の方が精度が高くなる．

第二に,「成分」という名詞は一般的すぎるだけでなく，未飽和名詞 (= (西山 (1990, 2003) が言う意味の非飽和名詞) でもある．未飽和名詞というのは，それ自体では何を指示しているか決まらない名詞のことである (より詳しい定義は §2.3.1 で示す)．飽和名詞 (e.g,. 人, 魚) と未飽和名詞 (e.g,. 成分, 犯人) の挙動はかなり異なるので，それを考慮に入れた検索の方が精度はよくなると思われる．

2.3 パス要素の属性の表現

本データでは <type=T ...>X</> という表記を用いて名詞 X の (未) 飽和性が指定している．これは X の未飽和性の値が T であることを表わす．T の値は G[ood], L[ess Good], D[ubious], B[ad], C[onnector] のいずれかである．以下，G, L, D, B, C の認定条件を，幾つか例を示しながら説明する．

2.3.1 飽和名詞 G(=Good) と未飽和名詞 L(=Less Good) の定義

名詞 X がそれ単独で指示を決定できる場合に X は飽和していると言い，そうでない場合に X は未飽和であると言う．飽和名詞を G で，未飽和名詞を L と認識している．

§1 で触れたように，未飽和名詞とは指示対象を特定するのに何らかの項 (argument) を必要とする名詞である．項は通常「A の」で表わされる．例えば (2) に挙げた名詞は単体では外延を特定できないが,「彼の親」「彼の妹」「彼の友人」「彼の弟子」「彼の家族」「彼の容姿」「彼の探求」「彼の逮捕」という形になると一様な指示対象の集合が特定できるようになる．これに対し (3) の飽和名詞では「A の」による修飾でそういうことは起こらない．

事前調査に基づき，(i) 個体間の関係を表す名詞 (e.g., 親，妹，友人，家族)，(ii) 個体の属性を表す名詞 (e.g., 容姿)，(iii) 事態を表わす名詞 (e.g., 探求，逮捕) を非飽和名詞と認定している．ただ，これらは決定的というより暫定的な取り決めである．未飽和名詞の方が認定が容易であり，現時点では，飽和名詞の定義は「未飽和でない名詞は飽和名詞である」という消極的なものである．

2.3.2 D(=Dubious), B(=Bad), C(=Connector) の定義

パス要素 X に成語性が認められる場合，X は G か L の評価をもつ．X に成語性がないと判断できる場合，X は B か C 評価をもち，X の成語性が不確実な場合，D と評価される．D, B, C のそれぞれの定義は以下に示す通りである．

2.3.3 D(=Dubious) の定義

パス要素が D と評価されるのは二つの場合がある．

(4) a. 絶対評価による D:　パス要素がそれ自体で成語性が怪しい場合
b. 相対評価による D:　パス要素自体は成語性をもつが，相対的下位語に対して有効な上位語になっているか怪しい場合

まず絶対評価による D の例を (5) に幾つか挙げる:

(5) a. <type=D>官</>, <type=D>事官</>, <type=G>参事官</>, <type=G>国務省参事官</>,<type=G>アメリカ合衆国国務省参事官</>
b. <type=D>官</>, <type=G>司令官</>, <type=D>隊司令官</>, <type=L>航空方面隊司令官</>, <type=G>北部航空方面隊司令官</>
c. <type=D>家</>, <type=D>世論家</>, <type=G>経世論家</>

「事官」や「隊司令官」や「世論家」の成語性は怪しいので，単体での評価は D である．

次に相対評価による D の例を (6) に幾つか挙げる:

(6) a. <type=L>人</>, <type=D>浪人</>, <type=G>流浪人</>
b. <type=L>家</>, <type=D>名家</>, <type=G>大名家</>, <type=G>本多氏の大名家</>

<type=D>浪人</> は単体では G，<type=D>名家</> は単体では L であるが，それぞれ <type=G>流浪人</> と <type=G>本多氏の大名家</> の上位語としては D である．この場合，<type=D>浪人</> と <type=D>名家</> は相対評価による D となる．

2.3.4 B(=Bad) の定義

D の場合と同様，パス要素が B と評価されるのは二つの場合がある．

(7) a. 絶対評価による B: パス要素がそれ自体で成語性をもたない場合
b. 相対評価による B: パス要素自体は成語性をもつが，相対的下位語に対して有効な上位語になっていない場合

原則として人手クリーニングの処理で絶対評価による B は除去した．上位語階層データに残っているのは，相対評価による B の場合の一部だけである．

(8) a. <type=D>屋</>, <type=B>酒屋</>, <type=G>居酒屋</>, <type=G>こだわり居酒屋</>
b. <type=D>達</>, <type=B>人達</>, <type=G>魔人達</>, <type=L>軍団の魔人達</>, <type=G>デルザー軍団の魔人達</>

これらの例で「酒屋」と「人達」は相対的に B 評価を受けている．

2.3.5 C(=Connector) の定義

C は，パスの最下位にある B を表わす．その意味では，C は B の特殊な場合である．このラベルは元データの上位語と下位語との接続を保証するために存在する．

(9) a. <type=D>家</>, <type=G>作詞家</>, <type=G>日本の作詞家</>, <type=C>出身日本の作詞家</>
b. <type=D>屋</>, <type=B>酒屋</>, <type=G>居酒屋</>, <type=G>個室居酒屋</>, <type=C>室個室居酒屋</>

2.3.6 ラベルの信頼性に関する注意

現時点でのラベルの値 G, L, D, B は定量的に定めたものではないので，どの場合でも十分に信頼性のあるものとは言えない．G と L の区別はかなり曖昧である．ただ，G, D, B, C の評価については信頼してよいと思われる．

現状では，G は完全に飽和しているという意味ではなく，十分に飽和しているという意味で使われる．今の時点では飽和性の指標は数量的なものではないので，主観性はなくなっていない．とはいえ，主観性をなくすことは常に意味のあることであるという保証もないので，データが有用である限りにおいては，評価の主観性は本質的な難点とはならないだろう．

2.4 属性タグで使われている (補助) 属性の説明

2.4.1 <type=T>X</>

<type=*T*>X</> の属性をもつ語句 X は，飽和性のタイプが *Y* であることを表わす．§2.3.1 で説明したように，*T* は G, L, D, B, C のいずれかである．

2.4.2 <type=...equals=Y>X</>

この属性は X が Y と意味的に等価であることを表わす．

(10) a. <type=D equals=者 > イスト </>, <type=G>アナキスト</>
b. <type=D equals=主義 > イズム </>, <type=G>ダダイズム</>
c. <type=D equals=主義 > イズム </>, <type=G>テロリズム</>

ただし実装は最小限である．

2.4.3 <type=...original=Y>X</>

この属性は X が Y の短縮形や省略形であることを表わす．

(11) <type=D original=金融 > 金 </>, <type=G>サラ金</>, <type=G>フリーダイヤルを設置しているサラ金</>

ただし実装は最小限である．

2.4.4 <type=...phonetics=Y>X</>

この属性は読みの異なりが type の異なりに対応している場合に，読みを指定する．例えば「大人」の上位語の「人」は読みに応じて L, D の二種類が区別される:

(12) a. <type=L>人</>, <type=L phonetics=タイジン > 大人 </>
b. <type=D>人</>, <type=L phonetics=オトナ > 大人 </>

ただし実装は最小限である．

2.4.5 <type=G correct=Y>X</>

この属性は X が Y の誤表記であることを表わす．

(13) <type=G correct=ヌクレオチド > ヌクレオシド </>, <type=G>含フッ素ヌクレオシド</>

ただし実装は最小限である．

3 (未) 飽和性を表すラベルの詳しい説明

§2.3.1 で説明したように，(未) 飽和性を表わすラベルは G[ood], L[ess Good], D[ubious], B[ad], C[onnector] の五つである．これらの詳細について情報を補足する．

3.1 飽和性の曖昧性

同一のパス要素に異なるラベルが付与されることがある．それは語義の別を反映して飽和性が変わるためである．「人」と「金」を例にして説明する．

3.1.1 「人」の場合

例えば「人」には <type=G>人</> と <type=L>人</> と <type=D>人</> の三つの場合がある．

(14) <type=G>人</>, <type=G>ベルギー人</>
(15) a. <type=L>人</>, <type=G>使用人</>
b. <type=L>人</>, <type=G>苦労人</>
(16) a. <type=D>人</>, <type=G>ケムール人</>

b. <type=D>人</>, <type=G>仙人</>

(17) a. <type=D>人</>, <type=L>一人</>

b. <type=D>人</>, <type=L>三人</>

(14) の例では生物種としての「人」の分類子なので，評価は G としている．(15) の例では「人」は役割を表す形態素であり，未飽和性もあるため，評価は L としている．(16) の例では G 評価をもった最下位の上位語は架空の生命体のタイプ名であり，人のタイプ名でないので，評価は D としている．(17) の例では L 評価をもった下位の上位語の単位を表す拘束形態素なので，評価は D としている．

3.1.2 「金」の場合

(18) a. <type=G>金</>, <type=G>塩化金</>

b. <type=L>金</>, <type=G>奨学金</>

c. <type=L>金</>, <type=L>料金</>

d. <type=D original=金属 > 金 </>, <type=L>合金</>, <type=G>共晶合金</>

e. <type=D equals=金融 > 金 </>, <type=G>サラ金</>

3.2 上位語パス要素の認定の特殊な場合

少数であるが，上位語パスの相対的上位語の認定に例外的な処理を行なった場合があるので，それらを説明する．

3.2.1 右端の要素が主要部になる場合

3.2.1.1 音便の処理: 最下位の上位語=元の上位語が音便を被っている時，上位語は非音便形とした．例えば，次のようにである:

(19) a. <type=G>タイ</>, <type=G>アマダイ</>

b. <type=G>サル</>, <type=G>テナガザル</>

c. <type=G> ハチ </>, <type=G>スズメバチ</>

d. <type=G>カメ</>, <type=G>ウミガメ</>, <type=G>アオウミガメ</>

3.2.1.2 架空の上位語: 最下位の上位語=元の上位語が日本語の語彙素として抽出できない場合は，D 評価をもつ架空の形態素を追加した．例えば，次の例で <type=D>イスト

</> と <type=D>イズム</> はそのようにして追加された最上位語である:

(20) a. <type=D equals=者 > イスト </>, <type=G>アナキスト</>
b. <type=D equals=者 > イスト </>, <type=G>ギタリスト</>, <type=G>ロックギタリスト</>
c. <type=D equals=主義 > イズム </>, <type=G>カニバリズム</>
d. <type=D equals=主義 > イズム </>, <type=G>ナショナリズム</>
e. <type=D>アイドル</>, <type=G>チャイドル</>

なお，equals=...属性によって「イスト」は「者」の，「イズム」は「主義」の類義語と見なせることを指定している．

3.2.1.3 省略語の補完

略語形の上位語を非略語形に変換した場合がある．例えば，次の例で「外科医」の上位語が「医師」に，「大学」の上位語が「学校」に，「高速」の上位語が「道路」に，「市電」の上位語が「電車」になっているのは，そのためである:

(21) a. <type=D>師</>, <type=G>医師</>, <type=G>外科医</>, <type=G>美容外科医</>
b. <type=D>校</>, <type=G>学校</>, <type=G>大学</>, <type=G>アイルランドの大学</>
c. <type=D>路</>, <type=G>道路</>, <type=L equals=高速道路 > 高速 </>, <type=G>放射状高速</>, <type=G>五放射状高速</>
d. <type=L>車</>, <type=G>電車</>, <type=G>市電</>

なお，補完はしないが，属性として元の語形を補った場合もある．例えば次のような場合である:

(22) <type=D original=金融 > 金 </>, <type=G>サラ金</>

3.2.2 左端の要素が主要部になる場合

3.2.2.1 「類」「の種類」「など」で終わる語句

「X 類」「X の種類」「X など」で終わる語句は，これらを取り除いた語形 X を最下位の上位語と見なして処理した．例えば，次の例で「化石鳥」「麺」「野菜」「焼酎」はこの基準で認定された:

(23) a. <type=G>鳥</>, <type=D>化石鳥</>, <type=D>化石鳥類</>
b. <type=G>麺</>, <type=D>麺類</>
c. <type=G>野菜</>, <type=D>野菜類</>
d. <type=G>焼酎</>, <type=G>焼酎乙類</>

3.2.2.2 「類」「の種類」「など」で終わらない語句の場合

「X 類」「X の種類」「X など」のような特殊な語や形態素で終わらない語句でも，意味を考えて語形 X を最下位の上位語と見なして処理した場合がある．例えば，次の例で「スーパー」「意味」「国」「女」「男」「役」「言」「機」はこの基準で認定された:

(24) a. <type=G>スーパー</>, <type=G>スーパーマーケット</>, <type=L>地域型スーパーマーケット</>
b. <type=L>アシスタント</>, <type=G>アシスタントディレクター</>
c. <type=L>意味</>, <type=L>意味合い</>
d. <type=G>国</>, <type=L>国家</>, <type=G>アラブ人国家</>
e. <type=G>女</>, <type=G>女性</>, <type=G>ユダヤ人女性</>
f. <type=G>男</>, <type=G>男性</>, <type=G>中年男性</>
g. <type=L>役</>, <type=G>役職</>, <type=G>自由民主党での役職</>
h. <type=D>言</>, <type=L>言葉</>, <type=G>一休宗純が遺した言葉</>
i. <type=D>機</><type=L>機械</>, <type=G>農業機械</>

これらの場合は，上の 3.2.2.1 の場合と異なり，規則化は難しい．

3.2.3 抽象的な上位語の追加

極く稀に上位語パスの最上位に，D 評価をもつ抽象的な要素を追加した．例えば，次の例で「版」や「法」はこの基準で認定された:

(25) a. <type=D>版</>, <type=G>サウンドトラック</>, <type=L>オリジナルサウンドトラック</>
b. <type=D>法</>, <type=L>ノウハウ</>

このようにして認定された最上位語のタイプは原則として D としている．

3.2.4 その他の注意事項

次のような条件をもつ語句を含むパス要素には D 評価を与えた．

3.2.4.1　進行形

「現在...」「...ている...」「...V 中の...」のように進行性を表わしている場合は，時間の経過によって記述が妥当でなくなる可能性があるので，評価は D とした.

3.2.4.1　比較を前提にするもの

「他の...」「(...) 以外の...」のように他の場合との比較を前提にしている場合は，比較対照が復元できないと事実性の判定ができないので，評価は D とした.

3.2.5　複合ラベル: タイプを絞り切れない場合の処理

次の例で「神経科学」の意味が医学の一分野である neural mecidine の意味か， neural science の意味か判別不能であり,「神経科学」の意味が医学の一分野である anesthetic medicine の意味か，anesthesia science の意味か判別不能である:

(26)　a. <type=L>学</>, <type=D-G> 科学 </>, <type=G>神経科学</>
　　b. <type=L>学</>, <type=D-G> 科学 </>, <type=G>麻酔科学</>

D-G が複合ラベルである．ただし，このような複合ラベルの該当例はデータ中にはほとんどない.

3.2.6　同一の最下位語をもつパスの複数性

異なる上位語パスの最下位語が同一である場合が稀に存在する．例えば，次のように「原材料」という語が含まれる場合がそうである:

(27)　a. <type=D>料</>, <type=L>原料</>, <type=L>原材料</>, <type=L>使用されてきた原材料</>, <type=L>伝統的に使用されてきた原材料</>
　　b. <type=D>料</>, <type=L>材料</>, <type=L>原材料</>, <type=L>使用されてきた原材料</>, <type=L>伝統的に使用されてきた原材料</>

これは「原材料」が「原料や材料」という意味の選言的な上位語であるため,「原料」か「材料」かの一方に絞りこめないためである.

同様に,「X_1 兼 X_2」も二つの上位語パスをもつ:

(28)　a. <type=L>アナウンサー</>, <type=G>アナウンサー兼ディレクター</>, <type=G>女性アナウンサー兼ディレクター</>
　　b. <type=L>ディレクター</>, <type=G>アナウンサー兼ディレクター</>, <type=G>女性アナウンサー兼ディレクター</>

3.2.7 有用な上位語パス要素の追加

形態素解析では候補が生成されなかったパス要素を，監督者が気づいた限り追加した．追加されたの要素の大半は (29) のように重複形態素を含む場合である:

(29)
a. …, <type=G>画家</>, <type=L>原画家</>
b. …, <type=G>画家</>, <type=G>劇画家</>
c. …, <type=D>画家</>, <type=G>版画家</>
d. …, <type=G>画家</>, <type=G>洋画家</>
e. …, <type=G>画家</>, <type=G>童画家</>
f. …, <type=G>画家</>, <type=G>版画家</>, <type=G>銅版画家</>
g. …, <type=G>画家</>, <type=G>版画家</>, <type=G>木版画家</>
h. …, <type=G>小説家</>, <type=G>私小説家</>
i. …, <type=D>農家</>, <type=G>酪農家</>
j. …, <type=G>菓子屋</>, <type=G>和菓子屋</>
k. …, <type=G>菓子屋</>, <type=G>洋菓子屋</>
l. …, <type=G>料理屋</>, <type=G>小料理屋</>
m. …, <type=D>術師</>, <type=G>幻術師</>
n. …, <type=G>絵師</>, <type=G>浮世絵師</>
o. …, <type=L>影像</>, <type=L>投影像</>
p. …, <type=D>画集</>, <type=G>版画集</>
q. …, <type=D>楽隊</>, <type=G>軍楽隊</>
r. …, <type=D>楽隊</>, <type=G>音楽隊</>
s. …, <type=D>法律家</>, <type=G>魔法律家</>
t. …, <type=D>作家</>, <type=L>贋作家</>
u. …, <type=D>教師</>, <type=L>宣教師</>
v. …, <type=D>教師</>, <type=L>調教師</>
w. …, <type=D>楽師</>, <type=L>猿楽師</>
x. …, <type=D>兵隊</>, <type=L>歩兵隊</>

参照文献

- 黒田航・李在鎬・野澤元・村田真樹・鳥澤健太郎 (2009). 鳥式改の上位語データの人手クリーニング. 言語処理学会 15 回年次大会発表論文集.
- Kuroda, K., Murata, M., and Torisawa, K. (2009). When nouns need co-arguments: A case study of semantically unsaturated nouns. In *Proceedings of the 5th International Workshop on Generatvei Approaches to the Lexicon, Sep. 17-19, 2009, Pisa, Italy*, pp. 193–200.
- 西山 佑司 (1990). 『カキ料理は広島が本場だ』構文について: 飽和名詞句と非飽和名詞句. 慶応大学言語文化研究所紀要 22: 169–188.
- 西山 佑司 (2003). 日本語名詞句の意味論と語用論: 指示的名詞句の非指示的名詞句. ひつじ書房.
- Bond, F., Isahara, H., Uchimoto, K., Kuribayashi, T. and Kanzaki, K. (2009). Extending the Japanese WordNet. In 言語処理学会 15 回大会発表論文集, pp. 80–83.
- Sumida, A., N. Yoshinaga, and K. Torisawa (2009). Boosting precision and recall of hyponymy relation acquisition form hierarchical layouts in Wikipedia. In Proceedings of the Lexical Resources and Evalucation 2008.