基本的意味関係の事例ベース（Version 1.2）

1. 公開履歴

2010 年 06 月 16 日	Version 1.0	収録数：17,275 対
2011 年 04 月 12 日	Version 1.1	収録数：39,561 対
2011 年 08 月 12 日	Version 1.2	収録数：59,618 対

2. Version 1.1からVersion 1.2の変更点

異表記対を20,057件追加収録しました。分類ラベル、ファイルフォーマットは、Version 1.1と同一です。Version 1.2における分類ラベル別の収録数を表 1 に示します。「収録数」欄の右カラムに、収録数のうち Version 1.1 に収録されていた語対の数を示しています。

分類ラベル	分類名	収録数	収録数のうちVersion 1.1に収録されていた語対数
v	異表記対	30,230	10,173
s	異形同義語対	24,273	24,273
a	略記対（*word1*が*word2*の略記）	621	621
A	略記対（*word2*が*word1*の略記）	576	576
d	対義語対	2,597	2,597
p	部分・全体語対（*word1*が*word2*の部分）	657	657
P	部分・全体語対（*word2*が*word1*の部分）	664	664
計		59,618	39,561

表 1： Version 1.2 における分類ラベル別収録数

3. Version 1.0からVersion 1.1の変更点

3.1 重複語対の集約

Version 1.0 では、重複語対がそのまま収録されていましたが、Version 1.1 では集約しています。下例に示すように、「s,竹飾り,七夕飾り」と「s,七夕飾り,竹飾り」は、*word1*,*word2* の並びが異なるのみで、組合せとしては同一であるため、一方を残して集約を行いました。集約の際、*word1*,*word2* の並びを UTF-8 コードに従い昇順にしました。

<Version 1.0>

```
s,竹飾り,七夕飾り
s,七夕飾り,竹飾り
```

<Version 1.1>

```
s,七夕飾り,竹飾り
```

3.2 分類ラベルの追加

異表記対や異形同義語対の関係においては、*word1* と *word2* は互いに他方の異表記や異形同義語であると言えますが、部分・全体語対の関係においては、*word1* と *word2* のどちらが部分であり全体であるのか、という区別が必要です。部分・全体語対については、Version 1.0 では以下のように位置関係によって区別していました。

```
p,word1（=部分を表わす語）,word2（=全体を表わす語）
```

Version 1.1 では、上記の区別を行うために、新しい分類ラベルを設けました。表 2 に、新設した分類ラベルを、従来の分類ラベルとの対照とともに示します。部分・全体語対に加え、略記対についても新しい分類ラベルを設け、どちらが略記であるかを区別できるようにしました。

旧分類ラベル	分類名	新分類ラベル	分類名と意味
A	略記対	a	略記対（*word1*が*word2*の略記）
		A	略記対（*word2*が*word1*の略記）
P	部分・全体語対	p	部分・全体語対（*word1*が*word2*の部分）
		P	部分・全体語対（*word2*が*word1*の部分）

表2： Version 1.1における分類ラベルの追加

3.3 収録語対の追加

分類の終わった語対を追加収録しました。Version 1.0 のデータを集約した後、追加収録分を加えた Version 1.1 における分類ラベル別の収録数を表 3 に示します。「収録数」欄の右カラムに、収録数のうち Version 1.0 に収録されていた語対の数を示しています。なお、Version 1.1 では、「表わすため」や「呼ばれているもの」のような、連用修飾節や連体修飾節の節内で目的語などが欠落しているため情報が不完全な語を含む場合、収録語対から外しています。

分類ラベル	分類名	収録数	収録数のうち Version 1.0に収録されていた語対数
v	異表記対	10,173	0
s	異形同義語対	24,273	9,417
a	略記対（*word1*が*word2*の略記）	621	332
A	略記対（*word2*が*word1*の略記）	576	205
d	対義語対	2,597	1,757
p	部分・全体語対（*word1*が*word2*の部分）	657	556
P	部分・全体語対（*word2*が*word1*の部分）	664	579
計		39,561	12,846

表3： Version 1.1における分類ラベル別収録数

4. 「基本的意味関係の事例ベース」について

4.1 概要

本データベースは、風間ら[1]の手法で構築された文脈類似語データベース[1]をもとに、文脈類似度の高い 2 語間の関係を分類し、ラベル付けした結果を収録したものです。

本データベースに収録される語対は、文脈類似語データベースの見出し語に対する類似語から、スコアが高い順に選択し、見出し語と組み合わせて生成しています。例えば、見出し語「りんご」に対する類似語として、スコアが高いほうから「みかん」「バナナ」「パイナップル」「ぶどう」「イチジク」があるとき、以下の語対を生成します。

[1] ALAGIN フォーラムにて公開されています。ご利用には独立行政法人情報通信研究機構との利用許諾契約が必要です。詳しくは、http://www.alagin.jp をご覧ください。

りんご,みかん
りんご,バナナ
りんご,パイナップル
りんご,ぶどう
りんご,イチジク

文脈類似語データベースの上位にある見出し語に対して同様に語対を生成し、関係分類の候補としました。ただし、解析誤りなどにより語として成立しない文字列(非単語)を含む語対は分類作業の対象外としています。

このようにして生成された語対の関係を人手によって分類し、分類が終わった語対から順次公開を行っています。

4.2 分類ラベル

語対の分類には、黒田ら[2]が報告した異表記認識のための分類基準をベースに、関係分類用に改編した分類ラベルを用いています。

Version 1.1 以降で使用されている分類ラベルを表 4 にまとめました。分類ラベルと分類基準は、作業を通じて変更されることがあります。そのため、Version 1.0 の分類基準と Version 1.1 以降の分類基準では、同じ語対であっても異なる分類ラベルを付与される場合があります。表 4 では Version 1.1 以降の分類基準に基づいて例を挙げています。Version 1.0 の分類基準については、Version 1.0 の説明書を参照ください。

分類ラベル	分類名	説明・分類基準	例
v	異表記対	読みが同じで意味が同じである語対。	[一リーグ制,1リーグ制] [100メートル,100m] [ゆううつ,憂鬱] [肩かけ,肩掛け] [アタリ,ATARI] [Kernel,kernel] [うみへび,海ヘビ] [問い合わせ,問合せ] [薮,藪] [ハードディスク,ハード・ディスク] [オーソリティ,オーソリティー] [バイオリン,ヴァイオリン] [町,街]
a A	略記対	一方の語の文字数または音節数が他方の語より少なく、他方の語の短縮形あるいは略称と呼ばれる語対。「word1/word2 は word2/word1 の略である」が言える。	[短大,短期大学] [ハードディスクドライブ,HDD] [ろうきん,労働金庫] [日本郵便,JP]

分類ラベル	分類名	説明・分類基準	例
s	異形同義語対	読み、形態素数、音節数のいずれかが異なり、異表記対にも略記対にも該当しない、同一の事象/事物を指す語対。原則的に「word1/word2 のことを word2/word1 とも言う」が言える。	[ご飯,食事] [ガウス分布,正規分布] [単独トップ,単独首位] [ワルキューレ,ヴァルキリー] [キネマ,シネマ] [お問い合わせ,問合せ] [山田,山田氏] [うすくち,うすくちしょうゆ] [しらたき,糸こんにゃく] [東京,江戸] [ディレクトリ,フォルダ] [キャリアサポートセンター,就職支援センター]
d	対義語対	互いに対義である語対。	[右側,左側] [低抵抗,高抵抗] [インフレ,デフレ]
p P	部分・全体語対	部分を表わす語と全体を表わす語との語対。	[手,親指] [椅子,背もたれ] [ジョン・レノン,ビートルズ] [太平洋戦争,第二次世界大戦] [大阪,近畿地方]

表4: Version 1.1以降で使用されている分類ラベル一覧

4.3 フォーマット

2種類のフォーマットを用意しています。一方を「フォーマットA」、他方を「フォーマットB」として説明します。

フォーマット A

カンマ(,)を区切り記号とする、一行一レコードのデータです。*word1,word2* をキーとして、UTF-8 コードに従い昇順ソートされています。

分類ラベル,word1,word2<RET>

以下に出力例を示します。

<出力例:フォーマット A>

```
v,あいさつ状,挨拶状
v,あいさつ程度,挨拶程度
v,あいずち,相槌
v,あいちゃん,愛ちゃん
s,あいつ,奴
v,あいつら,あいつ等
. . . (中略) . . .
s,チケット売り場,切符売り場
s,チケット売り場,切符売場
s,チケット売場,切符売場
s,チケット売場,券売所
s,チケット屋,金券屋
s,チタニウム,チタン
```

d,チチ,ママン
v,チッソ肥料,窒素肥料
. . .（中略）. . .
s,デザインサンプル,デザイン見本
s,デザインセンス,デザイン感覚
v,デザインパターン,デザイン・パターン
A,デザインフェスタ,デザフェス

フォーマット B

カンマ(,)を区切り記号とする一行一レコードのデータです。先頭カラムに*見出し語*があり、*見出し語*の*関連語*をカンマ区切りで列挙しています。*見出し語*をキーとして、UTF-8 コードに従い昇順ソートされています。

見出し語,*関連語* $_1$,... ,*関連語* $_n$ <*RET*>

*関連語*フィールドは、本節末尾の例にある通り、アンダースコア(_)で区切られた分類ラベルと、*見出し語*に対する*関連語*から成ります。

フォーマット B における分類ラベルは、*見出し語*を基準とした意味合いになります。分類ラベルの意味は、表 4 の通りです。

分類ラベル	意味
v	*関連語*が*見出し語*の異表記である。
a	*関連語*が*見出し語*の略記である。
A	*関連語*が（*見出し語*が略記である場合）略さない形である。
s	*関連語*が*見出し語*の異形同義語である。
d	*関連語*が*見出し語*の対義語である。
p	*関連語*が*見出し語*の部分である。
P	*関連語*が*見出し語*の全体である。

表4：フォーマットBにおける分類ラベルの意味

フォーマット B は、任意の語が、どの関連語とどのような関係でデータベースに収められているか、一覧することができる形式です。Version 1.1 では、分類ラベル v, a, A, s を同義語群、分類ラベル d を対義語群、分類ラベル p, P を部分・全体語群とし、各群ごとに出力を行いました。以下に出力例を示します。

<出力例:フォーマット B(同義語群)>

うろおぼえ,v_うろ覚え
うろこ状,v_ウロコ状,v_鱗状
うろこ雲,v_鱗雲,s_いわし雲,s_鰯雲
うろ覚え,v_うろおぼえ,v_ウロ覚え
うわごと,v_譫言
うわさ話,s_ゴシップ
うわずみ,v_上澄み
うわっつら,v_上っ面,s_うわべ,s_上辺
つくばエクスプレス沿線,a_TX沿線
ともだち,a_だち,s_お友だち,s_オトモダチ,s_友
のこぎり,a_ノコ

<出力例:フォーマット B(対義語群)>

アップロード時,d_ダウンロード時
アップロード機能,d_ダウンロード機能
アップロード等,d_ダウンロード等
アップ・テンポ,d_スローテンポ
アトランダム,d_順番どおり,d_順番通り
アドバンスコース,d_ベーシックコース
アナログデータ,d_ディジタルデータ

<出力例:フォーマット B(部分・全体語群)>

お手手,p_人さし指
けん,p_つか
ごみ焼却施設,p_ごみ焼却炉
ごみ焼却炉,P_ごみ焼却施設,P_ゴミ焼却施設
たし算,P_四則計算
つか,P_けん
のり面,P_堤体,P_護岸

4.4 ファイル

Version 1.2 フォーマット A
classified-pairs-v1.2a.zip
解凍後のファイル名: classified-paris-v1.2a.utf8
文字コードは UTF-8 です。

Version 1.2 フォーマット B
classified-pairs-v1.2b.zip
解凍後のファイル名: classified-pairs-v1.2b_s.utf8 (同義語群)
classified-pairs-v1.2b_d.utf8 (対義語群)
classified-pairs-v1.2b_p.utf8 (部分・全体語群)
文字コードはいずれも UTF-8 です。

5. 「基本的意味関係の事例ベース」ご利用に際して

利用許諾契約

「基本的意味関係の事例ベース」(以下、「本データベース」と呼ぶ)の利用には、独立行政法人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。

なお、すでに「基本的意味関係の事例ベース Version 1.0, Version 1.1」をご利用いただいている方は、新たに利用許諾契約を行うことなく「基本的意味関係の事例ベース Version 1.2」をご利用いただけます。

利用上の注意

本データベースは、インターネットホームページ等、独立行政法人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、独立行政法人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、独立行政法人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。

問い合わせ先

独立行政法人情報通信研究機構　ユニバーサルコミュニケーション研究所
情報分析研究室

Email: alagin-lr@khn.nict.go.jp

参考文献

[1] 風間淳一，デサーガ・ステイン，鳥澤健太郎，村田真樹．係り受けの確率的クラスタリングを用いた大規模類似語リストの作成．言語処理学会第15回年次大会発表論文集, pp. 84-87. 2009.

[2] 黒田航，風間淳一，村田真樹，鳥澤健太郎.Webデータに対応できる日本語異表記対の認定基準．言語処理学会第16回年次大会発表論文集, pp. 990-993. 2010.