シソーラスの構築，応用，ビジネス展開

Construction, application, business development of thesaurus.

国分芳宏

Yoshihiro KOKUBU
Email:kokubu@gengokk.co.jp

構文解析や用語標準化などの自然言語処理を目的とするシソーラスを開発した．このシソーラスでは，各用語の持つ関係語の数が膨大なため，観点（ファセット）を導入して分類し，用語を探しやすくしてある．また，差別語，表記の揺れなども区別している．シソーラスの使用法，シソーラスを用いたビジネス展開についても述べた．

We developed a thesaurus for the purpose of natural language processing such as parsing or the term standardization. Because each entry term has a large number of terms with various semantic relations, we introduce a facet and classify them for finding relative terms easily. Furthermore, we distinguish discriminatory terms, and fluctuating Japanese spellings. Thesaurus usage and business development using a thesaurus were also mentioned.

キーワード：シソーラス，観点，用語標準化，ネガポジ
Key Words:Thesaurus, Facet, Term standardization, Positive and Negative

1　はじめに

本シソーラスを作った株式会社言語工学研究所は 2012 年に事業譲渡してしまっている．その結果今回の話は過去の話になってしまうが，皆様がこれからシソーラスについて検討するときに何かの役に立つことを期待して述べる．

これまでのシソーラスは，主として，情報検索のキーワードを選択するための支援ツールとして開発されてきた．登録されている用語は該当する分野の専門用語が主体で，さらに品詞は名詞だけであった．そのため，情報検索を越えて，文書整理や統計処理などのために必要な構文解析や用語の標準化など，自然言語処理に利用することは難しかった．

筆者らのシソーラスは，自然言語処理を目的とした一般語を主とするシソーラスである．いわゆる名詞だけでなく，動詞，形容詞，形容動詞，副詞，代名詞，擬態語さらに慣用句までを登録している．これまでのシソーラスでは，作成者の考え方で分類してあった．使用者は，作成者の分類基準に従ってたどって探さなければならなかった．また紙面の物理的な制約もあって意味空間を１次元的に整理してあった．本来意味分類は多次元空間のはずで，筆者らのシソーラスでは，複数の観点で多次元的に分類してある．

2　用語の収集とシソーラスの構造

用語の収集と分類の仕方を述べる．

2.1　用語の収集

市販の辞書は語義が分からない用語を調べるためのものである．筆者らのシソーラスは自然言語処理で使うのが目的なので，市販の辞書に記載されている用語よりも，テレビや新聞などで頻繁に使われる用語を中心に登録している．正しい表記だけでなくよく使われるのであれば「キューピット」(Cupid)のように誤った表記や表記の揺れも登録している．

2.2　シソーラスの構造

複数の観点での分類

意味空間は１次元ではなく多次元である．どの属性に注目して（観点で）分類するかによって，いろいろな分類の仕方が考えられる身近な例で「料理」について考えてみる．古今東西の料理の種類は相当な数になり，分類の仕方も人によって異なる．ここで調理法，材料，地域の３つの観点で分類するとつぎのようになる．

調理法類　生もの，煮物，焼き物
材料　　魚料理，肉料理，野菜料理
地域　　和食，中華，洋食

例えば「刺し身」は，料理を３つの観点によって分類した結果，連想された用語「魚料理」「生もの」「和食」の狭義語である.逆に「刺し身」の広義語が「生もの」「魚料理」「和食」の３つあることになる．その結果，網構造になる．
これを図にすると，図１のようになる．

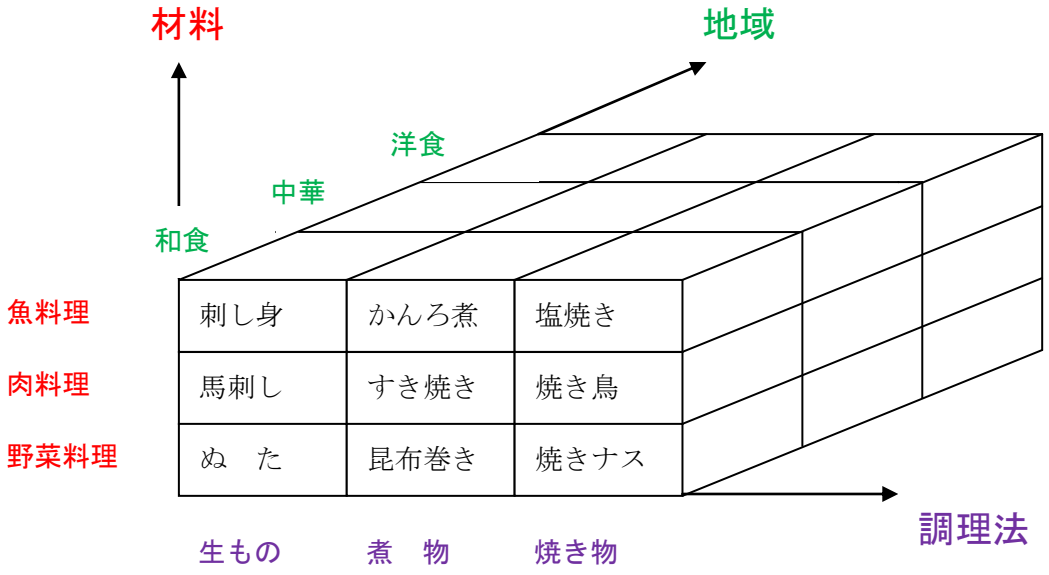


図 1.「料理」を「調理法」「材料」「地域」の３つの観点で分類した例

この他に「料理」のための観点としては「対象」（病人食，独身料理）「スタイル」（会席料理，飲茶）などが考えられる．いろいろな考え方で探す利用者がいるので，なるべく多くの観点で分類しておく必要がある．

2.3　シソーラス構造における補助的な記述

用語の持つ関係語の数が多いため，用語を３つの目的で，(｜)で区切って補助的な記述をつけている．

A．分類の観点を表示する．
　　狭義語の例
料理｜**材料**　肉料理，魚料理，野菜料理
料理｜**地域**　和食，洋食，中華料理

B．同じ分類に属する用語が膨大な数になるため細分したいときに，細分した分野に対応する適当な用語がなく，恣意的な用語になるのを防ぐ．

狭義語の例

肉料理｜**煮物**　シチュー
肉料理｜**薫製**　ビーフジャーキー

C．多義語を区別する．

狭義語の例

月｜**天体**　　満月，寒月，三日月
月｜**時間**　　正月，うるう月

用言

自然言語処理で使うには，名詞だけでなく用言（動詞，形容詞）や副詞も登録しておく必要がある．用言は語幹と活用形で登録してある．

慣用句

日本語では，慣用句が大きな意味的位置を占めている．慣用句はまとめた形で１語にして登録してある．

例　「水をあける」　＝　「引き離す」

「水をあける」は「引き離す」という意味で「水」の意味はまったくない．「水をあけ（る)」は１つの動詞にして「引き離す」の同義語として登録してある．

慣用句は用法によって間に挟まれる助詞までが変わるものがある．

例「山田は顔**が**広い」(叙述用法)
「顔**の**広い山田は」(限定用法)
それぞれ別の用語として登録した．

3.　用語同士の意味的関係

用語同士の意味関係として，表１のものを用意した．広義語－狭義語の関係は広義語に適用した規則が狭義語にも適用できるようにするため．同じ属性のものだけとした．原則として自立語だけとしたが，一部に接尾辞も採択してある．

表 1．用語同士の意味関係

同義語　　例　「犬」から見た「ドッグ」
表記の揺れも含む．
反義語　　例　「強い」から見た「弱い」
狭義語　　例　「犬」から見た「秋田犬」
広義語　　例　「犬」から見た「哺乳類」
関連語　　例　「犬」から見た「キツネ」「犬小屋」
品詞の異なる用語，自動詞－他動詞の対応なども関連語とした．
係り受け語　例　寿命（が)，延びる　良い
係り受け関係を構成する用語の組．係りと受けの間にはいる助詞およびネガポジも管理している．

3.1　同義語

英語で１人称単数は「I」だけであるが，日本語には「私」「僕」「我」「小生」「我が輩」「手前」「愚生」と数十あり，話者と相手との関係で使い分けられている．日本語にはなぜ同じ意味の用語，同義語がこんなに多いのか考えてみる．(表２参照)

表 2．同義語の例

大和言葉	漢語（複合語）	片仮名語	英　字
打ち合わせ	会議	ミーティング	
しお	食塩	ソルト	ＮａＣｌ
	読み出し専用メモリー	ロム	ＲＯＭ

外来語

日本語のなかに奈良時代には中国，朝鮮から，最近は主に米国から輸入されて日本語の中に入ってきている用語がある．多少のニュアンスの違いはあるが，すべて同義語といえる．このような組み合わせが日本語のなかにたくさんあり，これが同義語を増やしている大きな原因である．大和言葉は親しみやすさを，漢語は権威を，片仮名語は近代的な感じをあたえる．また最近は「電子計算機」が「コンピューター」さらには「パソコン」に，「写真機」が「カメラ」さらには「デジカメ」になるといったふうに，漢語が片仮名語に置き換わり，さらには短縮された用語に置き換わる傾向がある．

通称

通称と正式名称が両方使われている．

例　「首相」＝「内閣総理大臣」

年号

わが国の問題であるが，年号が 2 種類ある．さらに漢数字とアラビア数字が両方使われる．

例「2018 年」＝「平成 30 年」

＝「平成三十年」

立場による用語の違い

立場によって同じことを違った用語で表す場合がある．例えば「税金」という用語を政府は「公的資金」という言い方をするが，納税者は「血税」という言葉を使う．検索者は「税金」という用語で探すだろう．このような傾向は社会科学の用語に多い．

省略語

「特別急行」→「特急」のようなものをいうが，「マスコミ」は「マス・コミュニケーション」の省略形であったというように，現在は省略形の方が 4 拍の新しい用語として定着してしまっているものがある．省略の程度も地域によって異なる．関東よりも関西の方が積極的に省略するようである．

例「弱冷房車」（ＪＲ東日本）

「弱冷車」　（ＪＲ西日本）

頭字語（英語の用語の先頭の文字だけを集めた用語：アクロニム）もこの省略形に入れる．

例　ROM　　Read Only Memory

表記の揺れ

日本語では標準とされている表記の他に複数の「表記の揺れ」が許されている用語がある．個人により，機関により，いろいろな表記が氾濫している．極端な場合には，同じ著者が書いた記事でも表記法が違うことがある．複数の機関の記事を検索しようとする場合には，考えられる「表記の揺れ」をすべてキーにして検索しなければならない．

漢字と仮名による表記の揺れ

例　　犬，イヌ，いぬ

漢字表記の揺れ

例　　沈殿，沈澱　　（「澱」の字が常用漢字でないので「殿」の字を代用した．）

例　超電導　　（ＪＩＳ）

　　超伝導　　（学術用語）

外来語をカタカナ書きするときの揺れ

例　インターフェース　　（新聞）

　　インタフェース　　（ＪＩＳ）

　　インターフェイス　　（学術用語）

　　インタフェイス

送り仮名の違いによる表記の揺れ

例　行う，行なう

　　打ち合わせ，打ち合せ，打合わせ，打合せ，打合

（内閣告示の「送り仮名の付け方」の中にも複数の表記が許容されている．）

3.2　反義語

意味が対立する用語の関係である．対立の仕方にいくつかある．

A．片方を否定すると対立する相手になる用語の関係である．
例　善　←→　悪
B．ある中間的な点を中心にして逆の方向になる用語の関係である．
例　上　←　中　→　下
C．一つの行為を対立する立場で捕らえた用語の関係である．
例　売る　←→　買う
D．さらには「兄」に年齢で対立する用語として「弟」がある．また性別で対立する用語として「姉」がある．どちらも反義語になる．
例　兄　←年齢的対立→　弟
　　↑
　性別的対立
　　↓
　　姉

3.3　広義語・狭義語

自然言語処理で広義語との関係が狭義語にも適用できるように広義語・狭義語の関係は，属性が同じものだけにした．「自動車」－「タイヤ」のような全体部分関係は関連語にした．
例１　　東京都　新宿区　（狭義語）
　　東京都　都庁　　（関連語）
「東京都に住む」，「新宿区に住む」は成り立つが，「都庁に住む」は成り立たない．
例２　　疾病　伝染病　　（狭義語）
　　疾病　発病　　（関連語）

3.4　関連語

ある程度の意味的な関連性を持つ用語の関係を言う．大きく分けると同じカテゴリーの用語と異なるカテゴリーの用語との関係がある．
A．共通の広義語を持つ用語．
　広義語　　狭義語
　食材　⇘ 肉
　　　　⇗ 野菜
（「肉」と「野菜」とは関連語である．）
B．異なるカテゴリーであるが，意味的な関係のある用語．
　広義語　　狭義語
　食材　→ 肉
　料理　→ 肉料理
（「肉」と「肉料理」とは関連語である．）

3.5　多義語

英語は多義語が多いと言われているが，日本語，特に大和言葉も多義語が多い．
大和言葉での例
「うめる」　　穴をうめる．
　　お風呂をうめる．
　　借金をうめる．
　　時間をうめる．

外来語での例　英語の多義性の影響も受けている．
「ライト」　　光，照明，明るい，軽い
　　右，右翼手
　　権利
　　書く
多義語は，｜で区切って補助的な記述を付けてそれぞれ別の語として扱っている．

3.6　係り受け語

係り受けを構成する組み合わせを集めた辞書である．構文解析で係り先を決定したり，「ネガポジ」を決定したりするときに用いる．係り側の格助詞を含めて管理している．

4. 応用

4.1 用語の標準化

3.1 同義語のところで述べたように日本語は同じことを表すのにいくつかの表記が許されている．定量的な評価を含

めて有益な知識を得るためには，用語を標準化する必要がある．

このためにはまず同義語のグループのうちどの用語を標準の用語にするかを決めて標準の用語にマークをする．

例　　米，米国，ＵＳＡ，Ｕ．Ｓ．Ａ.，アメリカ合衆国，合衆国，アメリカ（新聞）

→アメリカ

ちなみにこの「米」という表記は複数の意味で使われているので注意する必要がある．

例　米（コメ）

（アメリカ）

誤った用語や差別語も標準の用語にならない．

誤った用語の例

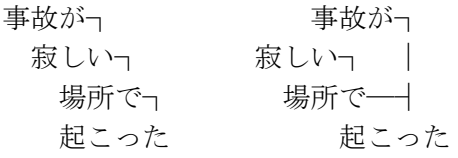
例　　ベットタウン、ピラミット

置き換え

データベースの記事の用語を推薦するものに置換する．

4.2 解析精度を上げる

構文解析に係り受け関係を正確にする目的に利用できる．例えば「事故が寂しい場所で起こった」という文を考えてみる．連体修飾格は体言に，連用修飾格は用言に係るという修飾関係を調べただけでは次の２つの異なった構造が考えられる．



誤った解析の例　　正しい解析の例

前の解析では「事故が」という連用修飾文節が「寂しい」という用言文節に係っていますが，後ろの解析では「起こった」という文節に係っている．「事故が」という文節が，「寂しい」「起こった」の２つの文節のどちらに係るかを考える必要がある．「事故　が　起こる」という係り受け語を登録しておけば，それを利用して「事故が」という文節は「起こった」という文節に係るようにできる．

さらに「事故」と「アクシデント」とは同義語であるという情報をシソーラスから得られれば，

「アクシデントが寂しい場所で起こった」

という文も解析できる．

しかし、同義語でも係り受けになれるものとなれないものがある．

係り受けの関係が狭義語にまで成り立つかどうかを調べておく必要がある．表２同義語の例のなかにある「塩」の同義語を調べると「食塩」「ソルト」「NaCl」がある．また「塩」を含む係り受けを考えてみると

表３　「塩」の許される係り受け

塩	食塩	ソルト	Na Cl	係り／受け
○				を送る
○	○			を振りかける
○	○		○	を加える
○	○	○	○	自然科学書

シソーラスによって構文解析の精度を上げるのは正攻法である．大変ではあるがやってみる価値は十分ある．

構文解析技術は行き詰っていてシソーラスに救いを求めている．

5. ビジネス展開

構文解析と組み合わせていくつかの展開が考えられる.

5.1 市場調査支援

市場の評判情報をツイッター, ブログなどから集められるようになってきた. シソーラスと構文解析を組み合わせてネガポジを判定して記事をネガポジで分類する. ただ記事のネガポジだけでなく, 何が良くて何が悪いかまでを判定する.
例を用いて説明するが説明の便宜上「良い」と判定する場合はP,「悪い」と判定する場合にはNと記入した.

ネガポジの判定

まず記事に使われている構文解析の結果の用語によってネガポジをきめる.

例　ポジ　　ネガ
　　美人　　瓦礫
　　涼しい

用語単独ではネガポジが決まらず, 係り受け関係でネガポジが決まるものがある.

例　「寿命が延びる」　P
　　「寿命が短い」　N

「寿命」「延びる」「短い」などの用語は単独ではネガポジの性質を持たないが係り受け関係になるとネガポジの性質を持つ. シソーラスの係り受け項目を調べて判定する.
このとき顔文字, 慣用句なども考慮する. 特に慣用句は教唆的なものが多いためかネガポジの性質を持つものが多い.

記事が書かれた状況によってネガポジが異なることがある.

例　「円が上がった」
　　輸出産業　N
　　輸入産業　P

業種ごとにカスタマイズが必要になる. 精度を上げていくためには業種別のコーパスを調べて分野ごとの係り受け辞書を充実させていく必要がある.

狭義語の展開

多義語についても用語を組み合わせた結果で判定する必要がある. 次の例では「甘い」という用語は多義語である.

例　果物が　甘い（甘味）　P
　　検査が　甘い（手ぬるい）　N

前の文では「甘い」という用語は「甘味」の意味でPの性質を持つ.

この関係は「果物」の狭義語にも成り立つ　お菓子, リンゴ, ミカン, ナシ・・・

後の文では「甘い」という用語は「手ぬるい」の意味でNの性質を持つ.

「検査」の狭義語にも成り立つ　考え, 詰め, チェック, 審査, 調査, ガード・・・

否定によるネガポジの逆転

否定によってネガポジは逆転するが, 日本語では「ない」と書いてあっても否定だとは決められない. このような自立語から後ろの付属語の並びを部分をモダリティーと呼ぶ.

「ない」を含んでいても否定にならない場合がある.

例	意味
飲まないか	勧誘
飲んだかも知れない	推量
飲まなければならない	義務
飲んでもらえないか	依頼
飲めないか	否定

特徴点の抽出

構文解析した結果の係り受けから何が良くて何が悪いかを判定する.

例　味が　良い　味がP
　　値段が　高い　値段がN

5.2 オントロジー構築支援

セマンティックウェブでWeb上のメタデータを記述する場合や, 知識ベースで

は共通する概念を提供する必要がある．共通概念のためにオントロジーというものがある．オントロジーとは共通する概念を体系化した辞書のようなものである．

オントロジーの中にはインスタンスとして具体的な用語が入っている．この具体的な用語を提供するためにシソーラスを用いる．

大阪大学でオントロジーを作成している溝口理一郎研究室でも使っていただいている．

6　おわりに

ネット上の記事が現在のペースで増えていくと，キーワードだけの検索ではノイズが多く早晩限界がくる．ノイズを減らすためは今後日本語解析などを高度化していく必要がある．そのためには意味の分野に立ち入らざるを得ないだろう．そのときにシソーラスは必須である．

ユーザーどんな使われ方をしていたかを聞いてみた結果を述べておく．

6.1　最適な用語を探す．

文書を書いていて自分の書きたいことが思い通りに書けないときにより適切な用語を探すために使う．現在筆者もシソーラスをこの目的で使っている．

6.2　検索で適切なキーワードを探す．

研究の初期段階には関係のありそうな記事を広く探す．

例えば「料理」の記事には「料理」という言葉は使われていない．

「煮る」「和える」「焼く」「下ごしらえ」などの関連語をキーにして検索する．場合によっては食材の名前などの関連語からキーワードを探す．

研究の最終段階ではごく限られた記事をピンポイントで探す必要がある．専門的な用語で探すことになるが，ここでも専門的な用語を探す支援をする．

謝辞

このような発表する機会を与えてくださった近畿大学田窪直規先生に感謝いたします．

参考文献

[1]国分芳宏他：複数の観点で分類した自然言語処理用シソーラス，自然言語処理，Vol17 No1

[2]溝口理一郎他：オントロジー構築入門，オーム社

著者　国分　芳宏
ブログ
http://kokublog.asablo.jp/blog/
ホームページ
http://www.asahi-net.or.jp/~wd2y-kkb/
メール
kokubu@gengokk.co.jp