自然言語処理 130−12
(1999. 3. 5)

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 —

村田 真樹　神崎 享子　内元 清貴　馬 青　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

〒651-2401 神戸市西区岩岡町岩岡 588-2

TEL:078-969-2181 FAX:078-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,kanzaki,uchimoto,qma,isahara}@crl.go.jp

あらまし

本論文では単語の羅列を意味でソートするといろいろなときに便利であるということについて記述する．また，この単語を意味でソートするという考え方を示すと同時に，この考え方と辞書，階層シソーラスとの関係，さらには多観点シソーラスについても論じる．そこでは単語を複数の属性で表現するという考え方も示し，今後の言語処理のためにその考え方に基づく辞書が必要であることについても述べている．また，単語を意味でソートすると便利になるであろう主要な三つの例についても述べる．

キーワード　意味ソート，意味的並べかえ手法，辞書の構築，タグつきコーパスの作成，情報提示

Meaning Sort MSORT
— Three Examples: Dictionary Construction, Tagged-Corpus Construction, and Information Presentation System —

Masaki Murata　Kyoko Kanzaki　Kiyotaka Uchimoto　Qing Ma　Hitoshi Isahara

Intelligent Processing Section, Kansai Advanced Research Center,

Communications Research Laboratory, Ministry of Posts and Telecommunications

588-2, Iwaoka, Nishi-ku, Kobe, 651-2401, Japan

TEL:+81-78-969-2181 FAX:+81-78-969-2189 http://www-karc.crl.go.jp/ips/murata

{murata,kanzaki,uchimoto,qma,isahara}@crl.go.jp

Abstract

It is often useful to sort words by their meanings like when using a thesaurus. In this paper, we introduce a method of arranging words semantically and show how to implement this method by using various types of dictionaries and thesauruses. We also examine an ideal dictionary that could be used for future natural language processing. Finally, we describe three main ways to use this method.

key words　Meaning Sort, Semantic Arranging Method, Dictionary Construction, Tagged-Corpus Construction, Information Presentation

1 はじめに

本論文では単語の羅列を意味でソートするといろいろなときに効率的でありかつ便利であるということについて記述する．この単語の羅列を意味で並べかえて考察するという考え方は一部の研究者で用いられているようであるが，その手法の有用性のわりに一般にはあまり知られておらずまた論文として明文化されていないもののようなので[1]，本論文では単語の羅列を意味でソートするという考え方について述べることとしたものである[2]．

本論文ではこの単語を意味でソートするという考え方を示すと同時に，この考え方と辞書，階層シソーラスとの関係，さらには多観点シソーラスについても論じる．そこでは単語を複数の属性で表現するという考え方も示し，今後の言語処理のためにその考え方に基づく辞書が必要であることについても述べている．また，単語を意味でソートすると便利になるであろう主要な三つの例についても述べる．

2 意味ソート

単語を意味で並べかえるという考え方を本論文では**意味ソート** Msort(meaning sort) と呼ぶことにする．この意味ソートは，単語の羅列を表示する際には 50 音順 (もしくは EUC 漢字コード順) で表示するのではなく，単語の意味の順番でソートして表示しようという考え方である．意味の順番の求め方は後節で述べる．

例えば，研究の途中段階で以下のようなデータが得られたとしよう[3]．

行事　寺 公式 母校 就任 皇室 学園 日本 ソ連 全国 農村 県 学校 祭り 家元 恒例 官民 祝い 王室

これは，行事という単語の前に「A の」という形で連接可能な名詞のリストであるが，このような情報が得られたときにその研究者はこのデータをどのような形式にすると考察しやすいであろうか．

まず，50 音順で並べかえる．そうすると以下のようになる．

行事　家元 祝い 王室 学園 学校 官民 県 公式 皇室 恒例 就任 全国 ソ連 寺 日本 農村 母校 祭り

これではよくわからない．

次に頻度順で並べてみる．

行事　恒例 学校 公式 日本 県 全国 寺 農村 王室 ソ連 祭り 学園 就任 祝い 母校 皇室 官民 家元

それもよくわからない．

これを単語の意味の順番 (ここでは，人間，組織，活動の順) でソートすると，以下のようになる．

行事　(人間) 皇室 王室 官民 家元
　　(組織) 全国 農村 県 日本 ソ連 寺 学校 学園 母校
　　(活動) 祝い 恒例 公式 就任 祭り

これは非常にわかりやすい．行事にはいろいろなものがあるが，ある特別な人を中心とした行事の存在，また，ある組織を中心とした行事の存在，さらに行事のいくつかの形態をまとめて一挙に理解することができる．

これはもともと名詞意味関係辞書の作成に「A の B」が利用できそうであるとすでにわかっている問題を例にあげたため，それはうまくいくでしょうといった感があるかもしれないが，後の節ではその他の問題でもこの意味ソートを用いるとうまくいく例をいくつか示している．われわれは，各研究の各段階でこの意味ソートというものを用いれば，ほとんどの問題がわかりやすくなり効率良く研究を進められるのではないかと考えている．

3 意味ソートの仕方

単語を意味でソートするためには，単語に対して意味的な順序づけを行なう必要がある．このためには分類語彙表[(3)] が役に立つ．分類語彙表とはボトムアップ的に単語を意味に基づいて整理した表であり，各単語に対して分類番号という数字が付与されている．電子化された分類語彙表データでは各単語は 10 桁の分類番号を与えられている．この 10 桁の分類番号は 7 レベルの階層構造を示している．上位 5 レベルは分類番号の最初の 5 桁で表現され 6 レベル目は次の 2 桁，最下層のレベルは最後の 3 桁で表現されている．

もっとも簡単な意味ソートの仕方は，単語に分類語彙表の分類番号を付与してその分類番号によってソートすることである[4]．

しかし，単にソートしただけではわかりにくい．数字ならば，順序関係がはっきりしているものなのでソートするだけで十分であるが，単語は順序関係がそうはっきりしたものではないので，ソートしただけではわかりにくい．ところどころに，物差しの目盛のようなものを入れた方がわかりやすい．

そこで，「人間」「具体物」「抽象物」といった意味素性というものを考える[5]．ソートした単語の羅列のところどころに意味素性のようなものをいれておくと，それ

---

[1] 有用な考え方の明文化は，各研究者に対する触発にもなり研究の進行速度の加速，学問の体系化などにも寄与すると考えられる．他にも重要そうな考え方が存在しているのならば，どんどん明文化するべきであると考えている．

[2] 筆者は過去に間接照応の際に必要となる名詞意味関係辞書の構築にこの意味ソートという考え方を利用すれば効率良く作成できるであろうことを述べている(1)．

[3] このデータは EDR 共起辞書のものを利用している(2)．

[4] 最近は便利な機能を持ったパソコンのソフトが多く出ており，Exel などに単語と分類語彙表の分類番号を入力しておいてソートすると簡単に意味ソートを行なうことができるであろう．
また，これとは別に筆者はホームページ (http://www-karc.crl.go.jp/ips/murata) において意味ソートを行なうツールを公開する予定である．

[5] ここであげる意味素性では目盛として荒すぎる場合は，分類語彙表の上位 3 桁目のレベル，上位 4 桁目のレベル，上位 5 桁目のレベルなどを目盛として用いてみるのもよい．

表 1: 分類語彙表の分類番号の変更

意味素性	分類語彙表の分類番号	変換後の分類番号
ANI(動物)	[1-3]56	511
HUM(人間)	12[0-4]	52[0-4]
ORG(組織・機関)	[1-3]2[5-8]	53[5-8]
PRO(生産物・道具)	[1-3]4[0-9]	61[0-9]
PAR(動物の部分)	[1-3]57	621
PLA(植物)	[1-3]55	631
NAT(自然物)	[1-3]52	641
LOC(空間・方角)	[1-3]17	657
QUA(数量)	[1-3]19	711
TIM(時間)	[1-3]16	811
PHE(現象名詞)	[1-3]5[01]	91[12]
ABS(抽象関係)	[1-3]1[0-58]	aa[0-58]
ACT(人間活動)	[1-3]58,[1-3]3[0-8]	ab[0-9]
OTH(その他)	4	d

を基準にソートした単語の羅列を見ることができ便利である．意味素性としては，IPAL 動詞辞書(4) の名詞の意味素性と分類語彙表の分類体系を組み合わせることによって新たに作成したものを用いる．このとき，分類語彙表の分類番号を名詞の意味素性に合わせて修正した．表1に作成した意味素性と分類語彙表での分類番号の変換表を記載しておく[6] [7]．表の数字は分類番号の最初の何桁かを変換するためものであり，例えば1行目の “[1-3]56” や “511” は，分類番号の頭の 3 桁が “156” か “256” か “356” ならば 511 に変換するということを意味している．([1-3] は 1,2,3 を意味している．)

表1に示した意味素性に目盛の役割をしてもらうわけだが，この目盛を意味ソートの際に入れる簡単な方法は，意味素性を単語のソートの際に混ぜてソートすることである．このようにすると，意味素性も適切な位置にソートされることとなる．

以下に意味ソートが実現される過程を例示する．ここでは，2節で示した名詞「行事」の前に「名詞 A の」の形でくっつく以下の名詞の集合を意味ソートすることを考えることとしよう．

行事　寺 公式 母校 就任 皇室 学園 日本 ソ連 全国 農村 県 学校 祭り 家元 恒例 官民 祝い 王室

1. まず初めに各語に分類語彙表の分類番号を付与する．「行事」と共起する名詞集合でこれを行なうと表 2 の結果が得られる．(書籍判の分類語彙表に慣れている人は注意して欲しい．書籍判では分類番号は 5 桁までしかないが，電子化判では 10 桁存在する[8]．) 表 2では「寺」が二つ，「公式」が四つ，存在しているが，これは多義性を意味しており，分類語彙表では「寺」に対し二つの意味が定義されており，「公式」に対し四つの意味が定義されていることを意味する．
2. 次に分類語彙表の分類番号の変換表の表 1に従って，付与した分類語彙表の番号を変更する．表 2のデータに対してこの番号変更を行なうと表 3の結果が得られる．例えば，表 2の一つ目の寺の最初の三桁は “126” であるがこれは表 1の三行目の “[1-3]2[5-8]” にマッチし，“536” に変換される．
3. 次に目盛用の分類番号つきの意味素性を 2 で得られた集合に追加する．表 3のデータに対してこれを行なうと表 4の結果が得られる．
4. 以上までで得られた集合を分類番号によってソートする．表 4のデータに対してこれを行なうと表 5の結果が得られる．

表 2: 分類語彙表の分類番号の付与例

1263005022	寺	1198007013	全国
1263005021	寺	1253007012	全国
1308207012	公式	1254006033	農村
1311509016	公式	1255004017	県
3101011014	公式	1263010012	学校
3360004013	公式	1336002012	祭り
1263013015	母校	1241023012	家元
1331201016	就任	1308205021	恒例
1210007021	皇室	1231002013	官民
1263010015	学園	1241101012	官民
1259001012	日本	1304308013	祝い
1259004192	ソ連	1336019012	祝い
右上につづく		1210007022	王室

表 3: 分類語彙表の分類番号の変更例

5363005022	寺	7118007013	全国
5363005021	寺	5353007012	全国
ab18207012	公式	5354006033	農村
ab21509016	公式	5355004017	県
aa11011014	公式	5363010012	学校
ab70004013	公式	ab46002012	祭り
5363013015	母校	5241023012	家元
ab41201016	就任	ab18205021	恒例
5210007021	皇室	5231002013	官民
5363010015	学園	5241101012	官民
5359001012	日本	ab14308013	祝い
5359004192	ソ連	ab46019012	祝い
右上につづく		5210007022	王室

[6] この表は現段階のものであって今後も変更していく可能性がある．

[7] 表では，体，用，相の分類を示す一桁の 1 ～ 3 の区別はなくしているが，これは文法的な分類の体，用，相の分類を行なわず意味的なソートをくしざし検索風に行なっていることになっている．もちろん，用途によっては体，用，相の分類を行なっておく必要があるだろう．その場合はそれに見合うように分類番号の変更を行なえばよい．例えば体，用，相の上位一桁目を a,b,c とするといったことを行なえばよい．

[8] ここでは KNP(5) に付属でインストールする分類語彙表の辞書を利用しているが，そこで用いられている分類語彙表は最新のものであってさらに桁が増えているが，KNP ではうまく 10 桁に変換しているようだ．

表 4: 目盛用の分類番号つきの意味素性の追加

5100000000	(動物)	ab70004013	公式
5200000000	(人間)	5363013015	母校
5300000000	(組織・機関)	ab41201016	就任
6100000000	(生産物・道具)	5210007021	皇室
6200000000	(動物の部分)	5363010015	学園
6300000000	(植物)	5359001012	日本
6400000000	(自然物)	5359004192	ソ連
6500000000	(空間・方角)	7118007013	全国
7100000000	(数量)	5353007012	全国
8100000000	(時間)	5354006033	農村
9100000000	(現象名詞)	5355004017	県
aa00000000	(抽象関係)	5363010012	学校
ab00000000	(人間活動)	ab46002012	祭り
d000000000	(その他)	5241023012	家元
5363005022	寺	ab18205021	恒例
5363005021	寺	5231002013	官民
ab18207012	公式	5241101012	官民
ab21509016	公式	ab14308013	祝い
aa11011014	公式	ab46019012	祝い
右上につづく		5210007022	王室

表 5: 分類番号の順番に並べかえ例

5100000000	(動物)	6200000000	(動物の部分)
5200000000	(人間)	6300000000	(植物)
5210007021	皇室	6400000000	(自然物)
5210007022	王室	6500000000	(空間・方角)
5231002013	官民	7100000000	(数量)
5241023012	家元	7118007013	全国
5241101012	官民	8100000000	(時間)
5300000000	(組織)	9100000000	(現象名詞)
5353007012	全国	aa00000000	(抽象関係)
5354006033	農村	aa11011014	公式
5355004017	県	ab00000000	(人間活動)
5359001012	日本	ab14308013	祝い
5359004192	ソ連	ab18205021	恒例
5363005021	寺	ab18207012	公式
5363005022	寺	ab21509016	公式
5363010012	学校	ab41201016	就任
5363010015	学園	ab46002012	祭り
5363013015	母校	ab46019012	祝い
6100000000	(生産物)	ab70004013	公式
右上につづく		d000000000	(その他)

5. 後は見やすいように整形すればよい．例えば，表5で分類番号を消し，意味素性ごとに一行にまとめ，語がない行を消去し，一行内にだぶって存在する語を消去すると表6のようになる．

前にも述べたとおり，表6の形になれば考察などに便利な状態になる．

表 6: ソート後の名詞集合の整形

(人間)	皇室 王室 官民 家元
(組織)	全国 農村 県 日本 ソ連 寺 学校 学園 母校
(数量)	全国
(関係)	公式
(活動)	祝い 恒例 公式 就任 祭り

4 意味ソートの諸相

4.1 分類語彙表以外の階層シソーラスを用いた意味ソート

今までの議論では分類語彙表を用いた意味ソートの仕方を述べてきた．意味ソートを行なうには意味の順序関係が必要であるが，分類語彙表はちょうど各単語に分類番号がついていたのでソートには最適であった．ここでは，EDR の辞書[2] のように，分類語彙表についていたような分類番号を持たない階層シソーラスを用いて，意味ソートはできないかを考察する．

前述したとおり，そもそも分類語彙表の 10 桁の分類番号は，7 レベルの階層構造を示している．EDR で意味ソートを行なう場合にも，分類語彙表と同じように上位桁から階層構造を作るような番号を各単語につけてやればよい．

しかし，番号をつけるのは面倒である．階層シソーラス上の各ノードにおける概念の定義文をそのレベルの番号のように扱ってやるとよい．こうすれば番号をあらためてふってやる必要がない．例えば，トップのノードから「母校」という単語に至る各ノードの概念の定義文を並べてみると以下のようになる．

概念
人間または人間と似た振る舞いをする主体
自立活動体
組織
組織のいろいろ
教育組織
学校という，教育を行う組織
数量や指示関係で捉えた学校
自分が学んでいる，あるいはかつて学んでいた学校

これを連結した“概念: 人間または人間と似た振る舞いをする主体: 自立活動体: 組織: 組織のいろいろ: 教育組織: 学校という，教育を行う組織: 数量や指示関係で捉えた学校: 自分が学んでいる，あるいはかつて学んでいた学校”を分類番号と見立てて意味ソートを行なえばよい．

先にあげた「の行事」に前接する名詞集合で EDR を用いた意味ソートを行なうと表7のようになる．表7では各行の出力のための目盛として上位三つの概念の定義文を用いている．

EDR では他の辞書に比べ多義性を設定する場合が多く，またシソーラスの階層構造においても複数パスを用いているので，同じ単語が複数の箇所に出ていて複雑なものになる．しかし，多観点から考察したいときには，ちょうどいろいろなとらえ方の単語を認識しやすいよう

表 7: EDR を用いた意味ソートの例

(概念：ものごと | ：もの)	寺 学校 県 家元 官民 祝い 公式
(概念：ものごと | ：事柄)	祭り 恒例 祝い
(概念：位置 | ：場所)	寺 学校 全国 県 農村 ソ連 日本
(概念：事象 | ：現象)	祭り
(概念：事象 | ：行為)	祝い 就任
(概念：事象 | ：状態)	官民 恒例 家元 寺 県 公式
(概念：人間または人間と似た振る舞いをする主体：自立活動体)		学校 学園 母校 寺 県 ソ連 日本 王室 皇室 家元 官民
(概念：人間または人間と似た振る舞いをする主体：人間)		寺 県 家元 官民

表 8: 単語に複数の属性を付与した辞書の例

単語	属性				
	種類	対象物	形状	サイズ	材質
うつわ	—	—	—	—	—
碗	和	—	深	—	陶磁
椀	和	—	深	—	木
湯のみ	和	緑茶 / 白湯	深	—	陶磁
皿	—	—	浅	—	—

表 9: 左の属性からソートした結果

単語	属性				
	種類	対象物	形状	サイズ	材質
うつわ	—	—	—	—	—
皿	—	—	浅	—	—
碗	和	—	深	—	陶磁
椀	和	—	深	—	木
湯のみ	和	緑茶 / 白湯	深	—	陶磁

表 10: 右の属性からソートした結果

単語	属性				
	種類	対象物	形状	サイズ	材質
うつわ	—	—	—	—	—
皿	—	—	浅	—	—
碗	和	—	深	—	陶磁
湯のみ	和	緑茶 / 白湯	深	—	陶磁
椀	和	—	深	—	木

になっており，EDR を用いると有効だろう．

以上までの議論から階層シソーラスならばどのようなものでも意味ソートが行なえることがわかるであろう．ただし，階層構造での枝別れ部分においてどのノードから出力するのかは曖昧になっている．例えば，表7では概念の定義文の文字列の EUC コード順となっている．順序を人手であらかじめ指定しておければそれにこしたことはないが，無理ならば，定義文自体を他の辞書 (例: 分類語彙表) により意味ソートすることも考えられる．

4.2 単語を複数の属性で表現するといった形での辞書記述における意味ソート

単語に複数の属性を付与するといった形で単語の意味記述を行なうという考え方がある．例えば，計算機用日本語生成辞書 IPAL の研究[6] では，「器」を意味するさまざまな単語に対して表 8のような属性を与えている．表中の"—" は属性の値は指定されていないことを意味する．

このような形の辞書の場合でも意味ソートは可能である．各属性を階層シソーラスでの各レベルであると認識すればよい．この場合，左の属性から順に階層シソーラスの上位から下位のレベルに対応すると考えると，"種類','対象物","形状","サイズ","材質" とレベルがあると考えられるので，意味ソートに用いる便宜的な分類番号は EDR の場合を参考にすると，"種類: 対象物: 形状: サイズ: 材質" といったものとなる．例えば，「椀」は "和:—: 深:—: 木" という分類番号を持っていることになる．(厳密には，属性の値も意味ソートするために，この「和」「深」「木」も分類語彙表の分類番号に変更しておく．) このような分類番号をもっているとしてソートすれば意味ソートのできあがりである．この意味ソートを行なった結果を表 9に示す．これは，単純に左の属性から順にソートしていった結果と等価である．

今は左の属性をもっとも重要な属性として扱って意味ソートを行なったものであるが，複数の属性の間の重要度の関係はそれほど明確ではない．例えば，同じデータで右の属性から順にソートすると，表 10のようになる．このように複数の属性を付与する辞書ではどういった属性を重視するかでソートのされ具合いが異なることとなる．これは，ユーザが今興味を持つ属性の順番によってソートすることができることを意味しており，複数の属性を付与する辞書は非常に融通が効くものであるということがいえる．

これを階層シソーラスも交えて考察するとさらに面白いことに気づく．先にも述べたように，各属性は階層シソーラスの各レベルと見立てることができるので，階層シソーラスでのこの属性のレベルの順序を変更することで何種類もの階層シソーラスを構築できることとなる．例えば，属性を左から用いて意味ソートした表 9からは図 1のような階層シソーラスが構築できる．また，属性を右から用いて意味ソートした表 10からは図 2のような階層シソーラスが構築できる．図 1のシソーラスでは「碗」と「椀」の意味的な近さをよく理解できる．図 2のシソー

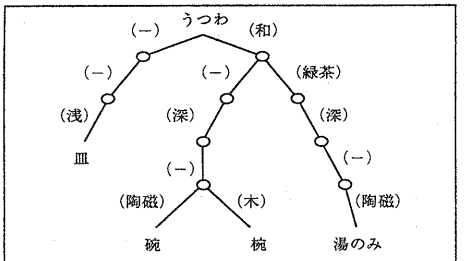

図 1: 左からのソート結果による階層シソーラス

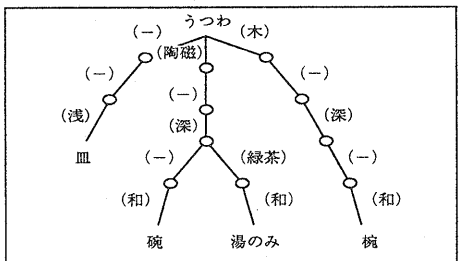

図 2: 右からのソート結果による階層シソーラス

ラスでは「碗」と「湯のみ」の同じ陶磁器としての意味的な近さをよく理解できる．この複数のシソーラスの構築は，種々の観点による階層シソーラスの研究にもつながるものである．観点によるシソーラスの必要性は文献[7]においても述べてある．それによると，「鳥」「飛行機」を上位で自然物と人工物と大別すると，「鳥」「飛行機」の意味的近さがわからなくなるとある．確かにそのとおりである．今後の言語処理を考えると観点による階層シソーラスの自由変形が可能な，複数属性を付与するといった辞書は非常に有用であり，生成用のみならず一般単語辞書，実用レベル的なもので構築する必要があると思われる．また，もともと単語の意味辞書を階層シソーラスの形にする必要があるのかという疑問も生じる．表8を見れば，「うつわ」の属性がすべて属性値を指定しない“—”になっていることから他の語の上位語であることが属性の集合の情報を見るだけでわかる．この属性の包含関係から上位・下位の関係が類推できるとすれば，階層シソーラスというものはわざわざ構築しておく必要はなく単語を属性の集合によって表現するというので十分なような気もしてくる．ただし，ペンギンは飛べないが普通の鳥は飛べるといった例外事象も扱えるような属性の定義などをしておく必要はある．また，複数属性で単語を表現する辞書では，一致する属性の割合などで単語間の類似度を定義するということも可能となるであろう．ただし，このとき属性に重みを与えるなどのことが必要になるかもしれない．単語の意味記述としては，さらに表現能力の高いものとして高階の述語論理で表現するもの，自然言語の文で定義するものなどが考えられるが，とりあえず現在の言語処理技術で扱えてそれでいて多観点を扱えるという意味で単語を複数属性の集合で意味記述するという辞書は妥当なところではないだろうか[9]．ちょっと脇道にそれて意味ソートと直接関係のない単語意味辞書のあるべき姿について議論をしてしまったが，単語を複数属性の集合で意味記述するという辞書ができれば，先にも述べたようにその辞書にはユーザが自分の好きな順番で属性を選んで意味ソートできるという利点があるので，意味ソートの立場としても非常に好都合である．

4.3 意味ソートとクラスタリング

以上までさまざまな議論を行なってきたが，意味ソートという操作は，結局のところ階層シソーラスを一次元に落したものに基づいてソートを行なうというものである．これは逆に見ると意味ソートした結果を木構造にすると階層シソーラスができあがるということを意味している．

ところで，階層シソーラスの構築には主にクラスタリング (クラス分け，分類の操作) という手法が用いられる．このクラスタリングは意味ソートとよく似た操作を行なうものである．クラスタリングを階層シソーラスの上位から下位まで綿密に行なうと意味ソートとほぼ同じ結果が出る．しかし，クラスタリングでは，階層シソーラスの各分岐点での兄弟関係のノードの順序関係というものを規定していない．ここの部分が意味ソートと異なっている．意味ソートでは，分類番号に基づいて並べかえるので，各分岐点での兄弟関係のノード間にも順序関係というものが存在している．大きな単語のまとまり同士では順序関係をわざわざ規定する必要がない場合も存在するかもしれないが，「春」「夏」「秋」「冬」などの単語の間のように順序関係が明確なものもあるので，兄弟関係のノードの間に順序関係を扱えないクラスタリングよりも順序関係を扱える意味ソートの方が望ましい．

また，意味ソートという操作はクラスタリングに比べるとかなり簡単な操作である．クラスタリングでは各クラスの箱を用意しておき入力された各要素をいれていくか，箱をあらかじめ用意しないならば動的に箱を用意して箱に各要素をいれていく必要がある．これにひきかえ，意味ソートではそれぞれ入力された各語に分類番号を付与すれば，あとはソートすればよいだけである．計算機実装でも意味ソートは簡便であるという利点を持っ

[9] 複数の属性を持つ単語辞書の作成には，国語辞典などの定義文が役に立つのではないかと考えている．例えば，定義文の文末から意味ソートを多段的にかけた結果を人手でチェックすることを行なえば，比較的低コストでこの辞書を作成できるだろう．

表 11: 「食べる」の格フレームの作成例

(a) ガ格の意味ソート結果

(動物)	牛 子牛 魚
(人間)	わたしたち みんな 自分 乳幼児 親 妹 お客 日本人 看護婦 作家

(b) ヲ格の意味ソート結果

(動物)	動物 貝 プランクトン
(生産物)	獲物 製品 材料 ベンキ 食べ物 えさ 和食 日本食 洋食 中華料理 おむすび 粥 すし ラーメン マカロニ サンドイッチ ピザ ステーキ バーベキュー てんぷら 空揚げ 穀物 米 白米 日本米 押し麦 キムチ カルビ 砂糖 ジャム 菓子 ケーキ ビスケット クッキー アイスクリーム
(体部)	遺骸 人肉 肝臓
(植物)	遺伝子 植物 牧草 ピーマン チコリ 桑 バナナ 松茸 昆布
(現象)	珍味 雪
(関係)	中身
(活動)	朝食 昼飯 夕食 夕御飯 おやつ 塩焼き

(c) デ格の意味ソート結果

(人間)	自分
(組織)	事務所 レストラン ホテル
(生産物)	しょうゆ シャトー 楽屋 便所 荷台 食卓
(空間)	現地 全域 車内
(数量)	ふたり 割合 複数
(活動)	研究 会議

ている.

5 意味ソートの三つの利用例

5.1 辞書の作成

名詞と名詞の間の意味関係を示す名詞意味関係辞書の作成に意味ソートが利用できる例はすでに文献(1)において述べている. 名詞と動詞, 名詞と形容詞の間の関係辞書の作成も格フレームや多義性などを考慮に入れながら同様にできることだろう. ここでは, 例として表 11に動詞「食べる」の格フレームの作成例を示しておく.

表 11は受身文など考慮して「食べる」の各格要素にくる名詞をそれぞれ意味ソートしたものである. 表 11の形になれば人手で格フレームを作成するのも容易であろうと思われる[10]. ガ格は動作主になりうる動物や人間が入ることがわかるし, ヲ格には様々な食べ物になりうる名詞が入ることがわかる. また, 任意格のデ格を見ると, 「自分で」や「事務書で」や「しょうゆで」などデ格の

[10] 最近では, 格フレームは多項関係でとらえる必要があることがいわれてきている. 例えば, 魚はプランクトンを食べるが牛は食べず, また牛は牧草を食べるが魚は食べない. 表 11の形に各格要素ごとまとめてしまうと魚とプランクトンの関係, 牛と牧草の関係が見えなくなり, よろしくない.

表 12: 名詞句「A の B」の意味解析用のタグつきコーパスの作成例

名詞 A	名詞 B	意味関係
パナマ	事件	場所
中学校	事件	場所
軍	事件	場所
アルバム	事件	間接限定
タンカー	事件	間接限定
最悪	事件	形的特徴
最大	事件	形的特徴
周辺	物件	場所
両国	事項	主体対象
文献	事項	分野限定
総会	事項	主体対象
上院	条項	分野限定, 主体対象
新法	条項	分野限定, 全体部分
条約	条項	分野限定, 全体部分
協定	条項	分野限定

意味関係が多種多様なものであることまでわかる. たとえば, "(人間)" の語は主体, "(組織)" "(空間)" の語は場所, "(具体物)" の語は道具の場合と場所の場合があることまでわかる.

ここであげた例は動詞の格フレームであるが, このようなことは形容詞に対してもさらには名詞述語文に対してもその他の単語間に対しても容易に行なえることを考えれば, 意味ソートの汎用性, 有用性を理解できるであろう. これは, 辞書の作成に限った話しではなく, 言語現象の調査におけるデータの整理, 有用な情報の抽出にも役に立つ. また, 近年いろいろな知識獲得の研究が行なわれているが, 知識獲得で得られたデータの整理にも, 同じようにこの意味ソートが役に立つ.

5.2 タグつきコーパスの作成 (意味的類似度との関連)

近年, さまざまなコーパスが作成されてきており(8)(9)(10), コーパスベースの研究も盛んになっている(11)(12). ここでは, コーパスの作成にも意味ソートが役に立つことについて述べる.

例えば, 名詞句「A の B」の意味解析を用例ベースで解析したいとする. この場合, 名詞句「A の B」の意味解析用のタグつきコーパスが必要となる. 具体的には, 名詞句「A の B」の各用例に対して「所有」「属性関係」といった意味関係をふっていくこととなる. このとき名詞句「A の B」を意味ソートしておけば比較的よく似た用例が近くに集まることになり, 意味関係をふる手間が軽減される.

表 12は文献(13)において作成されたタグつきコーパスの一部分である. ここでは名詞句「A の B」のうち名詞 B の方が重要であろうとして名詞 B を先に意味ソートとしたのち名詞 A で意味ソートを行なっている. 表中の意

味関係の用語は少々難しいものとなっているが意味ソートの結果近くにあらわれている用例同士は比較的同じ意味関係がふられていることがわかるだろう．このように意味的に近い用例が近くに集まるとタグの付与の手間が軽減されることが理解できるであろう．

ところで，用例ベースによる手法では入力のデータと最も類似した用例にふられたタグを解析結果とする．意味ソートという操作は単語を意味の順番にならべかえるわけだが，そのことによって類似した用例を集める働きをする．用例ベースと意味ソートは単語の類似性を用いるという共通点を持っている．この類似性を用いるという性質が用例ベースによる手法と意味ソートの共通した利点となっている．

ここでは名詞句「A の B」を例にあげてコーパス作成に意味ソートを用いると効率的であることを述べたが，これは特に名詞句「A の B」に限ったことではない．単語が関係している問題ならば，その単語で意味ソートができるので，文字列レベルで扱わないと仕方がない問題以外はほとんど本論文の意味ソートが利用できる．また，もともと文字列レベルで扱わないと仕方がない問題では文字列でソートすればよいのである．

しかし，ここであげた例では名詞 B で意味ソートした後，名詞 A で意味ソートをするといった不連続性がある．名詞 A と名詞 B の両方を考慮することで，意味ソートで近くにくる用例よりも意味的に近い用例を持ってこれる場合がある．しかし，このような方法では一次元的に配列するのが困難で人手でチェックするのが難しくなってくる．

5.3 情報検索での利用

近年，インターネットの発展とともに情報検索の研究は非常に盛んになっている．この情報検索がらみの研究においても意味ソートの有効な利用方法が考えられる．

例えば，津田の研究[14]では文書データベースの特徴を多数のキーワードによってユーザに提示するということを行なっている．例えば，提示したい文書データベース A のキーワード群が以下のとおりであったとする．

検索 単語 文書 作成 候補 質問 数 キーワード 情報

この単語の羅列をランダムな順番でユーザに提示するのでは不親切である．ここで意味ソートを行なってやると，以下のようになる．

(数量) 数
(抽象関係) 候補
(人間活動) 検索
文書 キーワード 単語 情報 質問
作成

ここでは，分類語彙表の上位三桁が一致するものを同じ行に表示している．ランダムに表示するよりはこのように意味ソートを行なって表示した方がよく似た意味の単語が集まるので，ユーザにとってやさしいのではないかと思われる．

また，情報検索システムが検索式を作る際にユーザにキーワードを提示して適切なものを選んでもらう場合もある[14]．このような場合においても，キーワードを他に適切にならべかえる方法があればそれを用いればよいが，そういったものがない場合は上記と同様にとりあえず意味ソートを用いておけば少しはユーザに対してやさしくなる．

6 おわりに

本論文では意味ソートの有用性を様々な角度から述べた．意味ソートの利用により少しでも各研究の効率化がはかられることを望む．

謝辞

4.2節で述べた単語を複数の属性の集合によって表現するという考え方は国立国語研究所の柏野和佳子研究員との議論において御教示いただいた．ここに同研究員に感謝する．

参考文献

(1) 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，言語処理学会論文誌，Vol. 4, No. 2, (1997).
(2) 日本電子化辞書研究所，EDR 電子化辞書仕様説明書，(1993).
(3) 国立国語研究所，分類語彙表，(秀英出版，1964).
(4) 情報処理振興事業協会技術センター，計算機用日本語基本動詞辞書 IPAL(Basic Verbs) 説明書，(1987).
(5) 黒橋禎夫，日本語構文解析システム KNP 使用説明書 version 2.0b6，(京都大学大学院情報学研究科，1998).
(6) 村田賢一，石田直子，岡部了也，細井正樹，柏野和佳子，猪塚元，計算機用日本語生成辞書 ipal(surface/deep) の研究，IPA 第 17 回技術発表会論文集，(1998), pp. 149–158.
(7) 川村和美，宮崎正弘，語を種々の観点から分類した多次元シソーラス，情報処理学会第 48 回全国大会予稿集，3Q-2, (1994), pp. 75–76.
(8) 日本電子化辞書研究所，EDR 電子化辞書 日本語コーパス第 1.5 版，(1995).
(9) 黒橋禎夫，長尾眞，京都大学テキストコーパス・プロジェクト，言語処理学会第 3 回年次大会，(1997), pp. 115–118.
(10) 新情報処理開発機構，RWC テキストデータベース第二版，(1998).
(11) 森信介，テキストコーパスからの確率的言語モデルの推定，京都大学工学部博士論文，(1998).
(12) 村田真樹，内元清貴，馬青，井佐原均，学習による文節まとめあげ — 決定木学習，最大エントロピー法，用例ベースによる手法と排反な規則を用いる新手法の比較 —，情報処理学会 自然言語処理研究会 NL128-4, (1998).
(13) 矢田恭儀，用例とシソーラスからの決定木学習による名詞句「a の b」の意味解析，京都大学工学部修士論文，(1997).
(14) 津田宏治，仙田修司，美濃導彦，池田克夫，自動作成された単語間リンクによる検索質問作成支援語を種々の観点から分類した多次元シソーラス，情報処理学会第 48 回全国大会予稿集，4E-6, (1994), pp. 157–158.