自然言語処理 117－7
（1997. 1. 20）

# 意味分類の言語学的構成法とWWW上のシソーラス構築

緒方典裕　　橋本三奈子

筑波大学大学院文芸・言語研究科　情報処理振興事業協会 (IPA) 技術センター

ogata@stc.ipa.go.jp　　hasimoto@stc.ipa.go.jp

自然言語処理においては、どんな意味を持つ名詞がどんな意味を持つ述語と共起するかという情報について、きめ細かい記述が必要である。また、それを記述するための意味分類は、生物学的な分類や恣意的な分類ではなく、名詞と述語との共起関係や構文に基づいたものであることが望ましい。ＩＰＡ技術センターでは、カテゴリー誤謬やカテゴリー一致という自然言語の統語論的・意味論的現象に着目して、意味分類(カテゴリー)体系の作成を試みた。本稿では、カテゴリー体系の記述法および構成法を述べ、それをWWW上で実装した「ＩＰＡ ＳＴａＸ」の概観を示す。「ＩＰＡ ＳＴａＸ」では、上下関係・種名称体系・部分名称体系の自動構成を更新に対して頑強な形で実現した。

# On a Linguistic Theory-based Constructing Method of Semantic Taxonomy and its Application to Constructing a Thesaurus on WWW

Norihiro OGATA　　Minako Hasimoto

University of Tsukuba　　Software Technology Center、

Graduate School　　INFORMATION-TECHNOLOGY

of Linguistics and Literatures　　PROMOTION AGENCY

Since fine-grained information about cooccurence is needed in processing natural language, we have been developed a framework of taxonomy of semantic categories, namely, *IPA STaX*. This framework is independent of common knowledge or scientific knowledge, and dependent only on lexical knowledge of cooccurrence of words. This lexical knowledge of cooccurrence of words can be considered as the realization of semantic category of words, which can be described as phenomena of *category-mistake* and *category-agreement*. Furthermore, *IPA STaX* is implemented as a thesaurus onto WWW with automatic construction of relations among categories: *hyponymy*、*taxonymy* and *meronymy*.

# 1 序

IPA 技術センターでは、「計算機用日本語辞書IPAL」として動詞、形容詞、名詞+スル、名詞+ダが述語となる場合の文型を記述し、その述語がとる名詞句の種類を意味素性として記載してある。しかし、処理のためには、意味素性では分類が粗すぎて不十分な部分もあり、もう少し細かく分類した意味分類が必要となった。この意味分類は、生物学的な分類や恣意的な分類ではなく、名詞と述語との共起関係や構文に基づいたものであることが望ましい。このため、カテゴリー誤謬やカテゴリー一致という自然言語の統語論的・意味論的現象に着目して、意味分類(カテゴリー)体系の作成を試みた。本稿では、カテゴリー体系の記述法および構成法を述べ、それを WWW 上で実装した IPA STaX の仕様を示す。

# 2 カテゴリー一致・誤謬に基づいた意味分類構成法 *STaX*

## 2.1 カテゴリーとは

カテゴリー(*category*)とは述語の項・名辞・その他表現一般に割り当てられている**意味の種類**もしくは**型**で、プログラミング言語の型(*type*)に対応する概念である。[1]

## 2.2 カテゴリー誤謬・カテゴリー一致とは

(1) a. ＊愛が5台並んでいる。

b. ＊素数は赤くない。

上の文は「文である」ことは理解できるが，何を意味しているかは理解できない。これを**カテゴリー誤謬 (category mistake)** という。

また、インド・ヨーロッパ系の言語等では、主語と述語、修飾語と被修飾語、先行詞と照応詞の間で性・数・人称などの文法的な一致が強制される。一方、日本語ではこのような文法的一致はないが、意味的な一致が強制される。

(2) a. The number of planets is necessarily nine.

b. 惑星の数（かず）は必然的に＊九/九つ/**九個**である。

(3) a. She came crying up to me.

b. 彼女は泣きながら＊私/私のところ/私の方に来た。

このように英語を逐語訳したら，日本語として不自然だが，カテゴリーを表す接辞や名詞を付加すれば自然になる場合が多い。このような意味的な一致を**カテゴリーの一致**と呼ぶことにする。特に、次のような定義的な名詞文ではカテゴリーだけをあらわす名詞が使われる。

(4) 図書館とは多くの人が閲覧するために書籍や文書を貯えてあるところ/**場所**/**建物**です。[2]

## 2.3 いかに分類するか

カテゴリー一致を反映する文法的特徴としては、次の項目が挙げられる。

(5) a. カテゴリー一致述語とのカテゴリー一致(特にカテゴリー固有述語)

b. 固有名称による言い換え可能性

c. 助数詞・カテゴリー形態素・照応詞とのカテゴリー一致

従って、カテゴリーの記述とは上の項目（これをカテゴリー一致情報と呼ぶ）を総合的に記述することであり、それが分類の基準となる。

[1] category という用語の歴史的概観については *IPA STaX* マニュアルを参照のこと。

[2] 但し，定義的な名詞文でも次のような語句は排除されなければならない。

- 図書館とは多くの人が閲覧するために書籍や文書を貯えてあるものです。

この「もの」は、どのようなカテゴリーの主語に対しても用いることができ、一種のモダリティ表現と考えるべきである。

### 2.3.1 カテゴリー一致述語とカテゴリー固有述語

あるカテゴリーを項に取る述語を，そのカテゴリーのカテゴリー一致述語と呼ぶことにする。また、あるカテゴリーだけを項に取るような述語を，そのカテゴリーのカテゴリー固有述語 (*proper predicate*) と呼ぶことにする。[3]例えば，

(6) a. 人が死んだ。

b. 人が絶滅した。

のように,「死ぬ」は個についてのみ陳述が可能で,「絶滅する」は種についてのみ陳述が可能である。従って，それぞれの「人」は違うカテゴリーを表しているのである。

### 2.3.2 固有名称による言い換え可能性

「人」は個と種の両方のカテゴリーをもちうるが，そのカテゴリーだけに使われる名詞もある。これを固有名称 (*proper nominal*) と呼ぶことにする。例えば，(6) は味次のようにそれぞれのカテゴリーの固有名称を用いて言い換えることができる。

(7) a. ある人物が死んだ。/田中さんが死んだ。

b. 人類が絶滅した。/ある人種が絶滅した。

しかし，固有述語はすべてのカテゴリーにそれぞれ存在するのに対して，固有名称は必ずしも存在するとは限らない。例えば,「トキ」も個と種のカテゴリーを持つが，その固有名称はもたない。

### 2.3.3 カテゴリー形態素

2.2で示したように、日本語では接尾辞がカテゴリーの表示に大きな役割を果たしている。また、

(8) a. ニューズウィークを購入した。

b. ニューズウィークを買収した。

上の文の「ニューズウィーク」は，それぞれ購読物、会社というカテゴリーをもち、曖昧であるが、次のようにある種の形態素を補うと曖昧性がなくなる。

(9) a. ニューズウィーク紙を購入した。

b. ニューズウィーク社を買収した。

-紙、-社のようなカテゴリーを表す形態素をカテゴリー形態素という。

しかし，カテゴリー形態素は必ずしもカテゴリー決定の必要条件でも、十分条件でもない。例えば，種を表す名詞は,「... の一種類である」という述語と共起できるかが一つの基準となる。

(10) a. 秋田犬は犬の一種類である。

b. * 野良犬は犬の一種類である。

このように秋田犬はこの基準を満たすのに対して，野良犬は満たさない。従って，前者は犬種を表しているが，後者は犬種を表していないということになる。後者はむしろ飼い犬・野犬・猛犬などの犬の性質を表す、つまり、犬タイプというカテゴリーに属する。

### 2.3.4 照応詞・指示詞

指示詞 (*demonstrative*)，照応詞 (*anaphora*) もまたカテゴリー一致を示す。次の例では「そこ」が部位というカテゴリーに対して用いられている。

(11) そこは胃だ。

しかし，次のように「それ」だとカテゴリーが変わってくる。

(12) それは胃だ。

(12) は肉の塊などを指しているような状況で使え、むしろ、個というカテゴリーに属する。

[3] 固有述語の概念は，Carlson (1977) の *kind-level predicate*、*stage-level predicate*、Bennet (1974) の *group-level predicate* などのカテゴリー特有の述語という概念と共通している。しかし，固有述語は,「猫」それだけに使われる述語「黒猫だ，白猫だ，どら猫だ，野良猫だ」などの述語のようなものも含まれ，さらに一般的な，あるカテゴリーに固有に用いられる述語という広い概念である。

### 2.3.5 カテゴリーのタイプと例

以上の観点から分類されるカテゴリーは次のようなタイプに分けられる。

| タイプ | カテゴリー -固有述語 | カテゴリー -形態素 |
|---|---|---|
| 抽象カテゴリー | 有 | – |
| full カテゴリー | 有 | 有 |

| タイプ | 助数詞 | 例 |
|---|---|---|
| 抽象カテゴリー | – | 乗り物 |
| full カテゴリー | 有 | 馬 |

抽象カテゴリーとはカテゴリー固有述語はもつが、カテゴリー形態素や助数詞など形態的情報をもたないカテゴリーである。たとえば、「乗り物」が属する**乗り物**（カテゴリー固有述語は「～に乗る、～から降りる」等）、「人工物」が属する**人工物**（カテゴリー固有述語は「～を作る」等）、などは抽象カテゴリーである。

full カテゴリーとは、カテゴリー一致情報をすべて持つカテゴリーで、「馬」が属する**馬**（カテゴリー固有述語は「～は子馬だ、～は愛馬だ」等）がその例である。

## 3 *STaX*の構造

前節の分類基準から*STaX*のもつカテゴリー間の関係が次のように構成できる。

- 種名称体系 (taxonymy): 'X は Y の一種(類)/一科目/一種目/一機種/... である' という句固有述語を利用。
- 部分名称体系 (meronymy): 'X は Y の一部分/一過程/一シーン/一場面/... である" という句固有述語を利用。
- 上下関係 (hyponymy): 次のようなカテゴリー一致述語の共有によって構成される。

(13) カテゴリー $C$のカテゴリー固有述語 $p$ が、カテゴリー $D$のカテゴリー一致述語ならば、$D < C$($C$は $D$ の**上位カテゴリー** (*super-category*), もしくは、$D$は $C$の**下位カテゴリー** (*sub-category*) という)。

## 4 IPA STaX for WWW = *STaX* + リンク・エンジン

IPA STaX for WWW は、*STaX*を html およびcgi プログラミングによってWWW上に実装したものである。IPA 日本語辞書グループにおいて利用可能なコーパス、およびいくつかの国語辞典から、所属する名詞・カテゴリー一致述語・カテゴリー形態素・助数詞を抽出し、各カテゴリーを構成した。このときに問題になるのが、分類体系の更新を繰り返すとカテゴリー間のリンクの管理が煩雑になるという点である。これを解決するために、3で示した、カテゴリー間の関係の構成を cgi プログラミングによって実現した。この結果、カテゴリー間のリンクは、共有されたカテゴリー一致述語を検索することにより、ダイナミックに構成される。このリンク構成に使われるカテゴリー一致述語を**リンカー** (linker) と呼び、リンカーによるカテゴリー間のリンク構成エンジンを**リンク・エンジン** (linking engine) と呼ぶ。また、リンカーを複数指定することにより、木構造的な分類体系ではなく、ネットワーク型の分類体系を構成することが可能となった。

現在、IPA STaX for WWW がもつリンク・エンジン、およびサーチ・エンジン[4]は次の通りである。

- 名詞カテゴリー・サーチ・エンジン（入力された名詞のカテゴリーを検索する。）
- 上位カテゴリー・リンク・エンジン
- 下位カテゴリー・リンク・エンジン
- 上位部分カテゴリー・リンク・エンジン
- 下位部分カテゴリー・リンク・エンジン
- 上位種カテゴリー・リンク・エンジン
- 下位種カテゴリー・リンク・エンジン

[4]現ヴァージョンでは、述語のカテゴリーの自動構成エンジンももつ。

**リンカーの選定の制約**　リンカーに選ばれるカテゴリー一致述語は、あるカテゴリーのカテゴリー固有述語でなければならない。さらに、多義性をもつようなものはリンカーとしてふさわしくない。

たとえば、「～に入る」という述語をリンカーとして選ぶと、次の例文が示すように、**空間**というカテゴリーと**組織**という上位カテゴリーの両方を同一化してしまう。

(14) a. 映画館に入る。

b. テニスクラブに入る。

従って、このような複数のカテゴリーを同一にしてしまう述語はリンカーとしてふさわしくない。代わりに、たとえば、**空間**に対して「～から出る」を、**組織**に対して「～を辞める」をリンカーとして選べば、**空間**と**組織**に関してはカテゴリーを同一化してリンクすることはなくなる。

## 5　まとめ

現在、IPA で進行中の意味分類体系 *STaX* の基本的アイデアとそのWWW上でのシソーラスとしての実装であるIPA STaX for WWW を紹介した。本稿執筆時の最新バージョンは2.2(図:IPA STaX version 2.2 参照) で、語彙数 3693 名詞+約 1000 述語、250 名詞カテゴリーを収録している。現在、さらに語彙数・カテゴリー数の増加を基本的に推し進めている。

また、IPA STaX for WWW の応用例である、DiSTaX(未定語のカテゴリーの自動付与、および、カテゴリーが曖昧 (ambiguous) な名詞に対する文脈中での曖昧性解消) の構築や、IPA STaX for WWW の基本仕様を internet 上に公開し、ユーザードメイン限定的なシソーラスを構築してもらい、それをリンクする *Open* STaX 等を構想中である。

## 参考文献


BENNET, M. 1974. *Some Extensions of a Montague Fragment of English.* PhD thesis, University of California at Los Angeles.

CARLSON, G. N. 1977. *Reference to Kinds in English.* PhD thesis, University of Massachusetts. published by Garland Publishing, New York, 1980.

CRUSE, D. A. 1986. *Lexical Semantics.* Cambridge: Cambridge University Press.

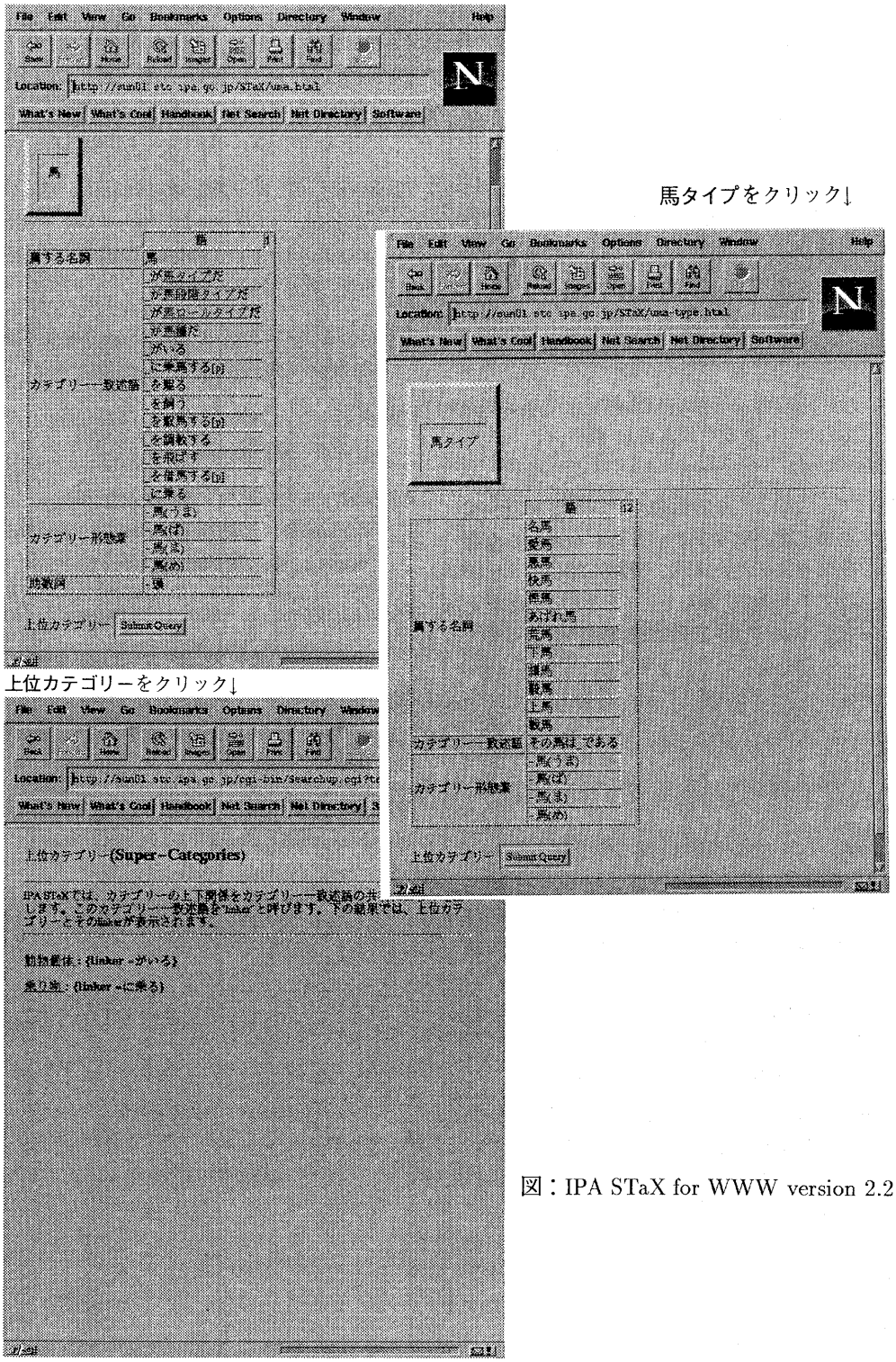
DRANGE, T. 1966. *Type Crossings.* The Hague: Mouton and Co.

LAPPIN, S. 1981. *Sorts, Ontology, and Metaphor: The Semantics of Sort Structure.* Berlin: Walter de Gruyter.

PUSTEJOVSKY, J. 1995. *Generative Lexicon.* Cambridge: The MIT Press.

RYLE, G. 1965. Categories. In A. Flew, ed., *Logic and Language*, New York: Doubleday, pp. 281–298.

VENDLER, Z. 1967. *Linguistics in Philosophy.* New York: Cornell University Press.




図：IPA STaX for WWW version 2.2