自然言語処理 93－9
（1993. 1. 14）

英文における未登録語の意味推定の検討

山田一郎　山村毅　佐川雄二　大西昇　杉江昇
名古屋大学工学部
名古屋市千種区不老町

未登録語とは辞書に登録されていない単語のことである。多くの単語や文をコンピュータで処理する場合、このような未登録語の処理が必要になると思われる。本稿においては、この未登録語について、その処理の必要性を明示する。具体的な処理のための知識源として、形態素レベル、句レベル、文レベル、について整理を行なう。さらに、比較的規模の大きいコーパスを対象とし、電子辞書を利用して、実際に登録されていない単語を抽出し、それらについて、どのような特徴を持っているか、どのような知識が有用であるか、についての調査を行なう。この結果にもとづき簡単な未登録語の意味推定システムを実現する。

Inferring Meanings of Unknown English Words

Ichiro Yamada　Tsuyoshi Yamamura　Yuji Sagawa　Noboru Ohnishi　Noboru Sugie
School of Engineering, Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-01,Japan

Unknown words are those words that are not registered in a dictionary. To deal with many words,these words must be taken into consideration. In this paper, we point out how necessary it is to process these words. We apply an electronic dictionary to a comparatively large scaled corpus in order to pick out unknown words in practical use. We also analyze the charactristic of them and investigate the kinds of knowledge which are effective for infering the meanings. Based on the investigation, we construct a prototype system for inferring meanings of unknown wards.

1 はじめに

自然言語の解析を行なう場合、辞書に登録していない語（以下、未登録語）が現れることが多い。近年、技術の向上にともない機械辞書は大規模で充実したものとなってきた。登録単語数が十数万語を超える機械辞書も珍しくはない。しかし実用的なシステムにおいて、出現単語のすべてを登録している、ということは不可能に近い。その理由の一つとして、新語が挙げられる。毎年大変多くの新語が出現している。この新語による未登録語は、技術の向上のみでは解決できない。また、この他にも誤字、脱字、外来語なども、未登録語となってしまう。従って、実用的な自然言語処理システムの構築のためには、このような未登録語が存在する場合でも、構文解析、意味解析において、ある程度の処理を行ない、正常に解析を終了させる方法を考えることが必要とされる。

未登録語処理において、英語文は、分かち書きがされているので、その抽出、構文解析においては日本語文より容易に処理可能である。しかし、英語の単語は多義であることが多く、意味解析においてはきわめて処理が難しくなる。実際に、人間が考えても全く意味を推定出来ない未登録語も存在する。従って、得られる限りの情報を最大限に有効利用する必要性が挙げられる。

これまでにも未登録語を扱う研究は、英語文、日本語文ともに、数多くなされてきた [1] [2] 。しかし、これらの多くは、形態素解析や構文解析のみで意味処理まで考慮したものは少なかった。

本稿においては、まず実際の文章と辞書を用いて未登録語の特性について考察する。次に、未登録語についてその意味の範疇を推定する処理を可能とするための、知識源について整理を行なう。未登録語の品詞推定は宇佐美ら [3] によって既に行なわれているため、品詞は既知とし、その中の名詞について主に考察を行なう。知識源としては、接頭辞、接尾辞、語幹などに注目した形態素レベル、名詞句における並列や包含などの単語間の関係に注目した句レベル、動詞の格フレームや、主格ー目的格間などの関係に注目した文レベル、の３つのレベルについて考察し、これらの利用法、問題点、について整理する。また、得られた情報の統合について考察する。さらに、この知識源について、実際の文章における有効性を検討する。最後に、これらの知識源を利用した簡単なシステムを構築する。

2 未登録語処理の必要性

2.1 未登録語の定義

本研究において、未登録語を以下のように定義する。

【定義】そのもの自体が機械辞書に含まれていないような全ての単語。（ただし、複数形、過去形、比較級、最上級などの単純な語尾変化によるものは登録語とする。）

つまり、実際に自然言語を処理する際に、処理の過程が止まってしまう可能性のある単語の全てを未登録語とする。ただし、ハイフンによる明らかな複合語の場合は、切り出された語のうち一つでも未登録語があるときのみ、その複合語を未登録語とする。

2.2 未登録語の特性調査

調査は、一般的な英文が多数集められている"*LOB CORPUS**" を対象に行なった。このコーパスには、品詞の情報が各語ごとにつけられている。

辞書として、*EDR*電子化辞書と、*UNIX*にある辞書を利用した。これらには、約１２万６千種類もの単語が登録されており、実用レベルの辞書と言える。

調査目的は、未登録語処理に対する必要性の考察をすること、また、そこでどのような特徴（品詞など）を持った未登録語の考察が必要か、について明確にすることである。

調査内容、結果は以下に示す通りである。

1．未登録語の出現率調査

コーパスに出現する未登録語を抽出し、その数をカウントする。ここにおける数とは単語出現数の累計であり、同じ単語も重複して数えている。結果を表１に示す。この表よりわかるように、未登録語は全単語数の約1.5%を占めていた。計算すると、平均４文中に一個は未登録語が存在しており、未登録語の処理の必要性が再確認された。

*Norwegian Computing Centre for the Humanities より入手した。

2．未登録語の種類の調査

未登録語を、単語の先頭が大文字である単語と、小文字については記号の含まれない単語、ハイフンによる複合語、ピリオドを含む省略形の単語、に分類する。結果を表2に示す。大文字で始まる未登録語が多い理由は、固有名詞の登録に限界があるからである。実際に、人名、地名、企業名などがそのほとんどを占めていた。従って、品詞の決まらない小文字のみの単語の処理が、最も重要と思われる。

3．未登録語の品詞

小文字で始まる未登録語を対象に、その品詞を調査する。結果を表3に示す。この表から、名詞、形容詞、外来語（英語以外の語）の頻度が高いことが分かり、それらに対する処理の必要性が挙げられる。

4．派生語の調査

小文字で始まる未登録語を対象に、登録語を語幹として含むかを調査する。結果を表4に示す。この表から、形容詞、副詞は、形態素によりほとんど処理が可能であることがわかる。従って、名詞、動詞に対する考察が重要と思われる。

以上により、未登録語処理は重要であることがうかがえる。また、小文字ではじまる単語の名詞、動詞、外来語、についての処理の必要性があると言えよう。

表 1: 未登録語の割合

全単語数	１０１３８５１語
全未登録語数	１４８９２語

表 2: 未登録語の種類

未登録語の種類	種類	数
大文字で はじまる単語	４８６３種	１０９９３個 （８２%）
小文字で はじまる単語	９４８種	１６７８個 （１３%）
省略形の単語	６４種	２１２個 （２%）
複合語（ハイフンによる）	３２２種	４９４個 （４%）
全体	６１９７種	１３３７７個

表 3: 未登録語の品詞

品詞	種類	数
名詞	３３７種（３６%）	６３６個
動詞	５５種（ ６%）	９１個
形容詞	１５３種（１６%）	２３３個
副詞	５８種（ ６%）	１１６個
外来語	３２６種（３４%）	５７１個
その他	１９種（ ２%）	３１個
計	９４８種	１６７８個

表 4: 形態素による処理が可能な割合

品詞	種類	形態素による推定可能
名詞	３３７種	１４８種（４４%）
動詞	５５種	１９種（３５%）
形容詞	１５３種	１４７種（９６%）
副詞	５８種	４８種（８３%）
計	６０３種	３６２種（６０%）

3 意味推定のための知識源

名詞の意味を考えるためには、意味概念の分類が重要である。本稿においては、表5に示す科技庁のＭｕプロジェクト[4]においてなされた分類を利用している。

3.1 形態素レベルの知識源

表4より、未登録語は、登録語の接頭辞、接尾辞による派生語であることが多いと言える。このような情報について整理したものが形態素レベルの知識源である。具体的には以下の通りである。

接頭辞は、単語に意味の変化を与える。しかし、表5の意味マーカの範囲を越えることは少ない。従って、接頭辞とその基体が登録語であるなら、意味は、その基体の意味カテゴリーと同じであると推定可能である。

接尾辞は、前述のカテゴリーで考えるのであれば意味限定可能と言える。確実に接尾辞が抽出出来れば、意味も限定可能である。ところが、登録されている接尾辞の存在のみの情報からでは、誤った推定がされる場合がある。

【例1】 “enhance” → 名詞の接尾辞として推定

この様に誤った接尾辞を抽出してしまうことを防ぐために、接尾辞を単語から取り除いたあとに残った部分が、本当に基体となりうるかについて判断を行う。

未登録語はまた、複合語の場合がとても多い。2章の調査における名詞の未登録語337種のうち、57種が複合語による未登録語であった。この処理手順は、前から一文字ずつ切り出し、一つ一つ辞書引きを行ないながら、名詞・動詞と名詞に限定した組み合わせをすべて考えるといったものである。複合語には、主要語を含む内心複合語（例：catfish〔ナマズ〕）と、含まない外心複合語（例：bootleg〔密造酒〕）がある。後者は、その要素のみからでは意味推定は困難である。しかし、実際、未登録語のような単語はすべて前者のような主要語を含むものである、と言っても良いであろう。従って、含まれる要素の切り出しに成功した単語の意味は、その主要語つまり、最後の部分にある語のカテゴリーと同じであると、推定される。

表5: 名詞意味マーカ体系

ファセット	下位分類
国・機関・組織	
生物	人, 動物, 植物, その他
無生物	自然物, 部品・材料, 生産物, 施設, その他
知的抽象物	理論・法則・学問, 知的抽象的道具・方法, 知的抽象的材料, 知的抽象的材料, その他
部分	部分・要素, 生物の器官, その他
属性	属性名, 関係, 形態, 状態, 構造, 特徴, その他
現象	自然現象, 物象, 力・エネルギー, 生理的現象, 社会的現象, 制度・習慣, その他
心情	感覚・反応, 認知・思考, その他
行動	行為, 動き, その他
測度	数, 数量名, 基準・標準, 単位, その他
場所・空間	
時間	時点, 時間間隔, 所要時間, その他

3.2 句レベルによる知識源

文に含まれる名詞句、前置詞句のみを見ただけでも意味を限定出来る時がある。そのような情報を整理したものが句レベルによる知識源である。しかし、このレベルにおいては、句の抽出、係受け関係の曖昧性が問題として挙げられる。そこで、

1. 句の始まりとなる語（接続詞，前置詞，助動詞，動詞，冠詞，句読点）を手掛かりとする。
2. A.S.Hornby[5] の動詞型により拘束条件を与える。
3. 構文解析が一意に決まらない場合はすべての可能性について考える。

という考えに基づいて、この問題を処理する。

動詞がA.S.Hornbyの動詞型において前置詞句をとることが出来る場合は、前置詞句は副詞的用法と形容詞的用法の2つの可能性がある。そこで、両方の可能性を考えるために、文レベルにおける情報も同時に考慮する。

3.2.1 名詞句

『名詞句1＋OF＋名詞句2』の型の名詞句になっている時、名詞句1と名詞句2との間には、ある種の意味関係が成立する。この関係を利用することにより、名詞句1あるいは名詞句2の意味を限定することができる。ここでは、次のような名詞間の関係の知識を利用する。この知識は次節で述べる文レベルにおいても利用する。

【名詞間の関係の知識】

表5に示す関係『下位』の他に、対象が被対象と同じ意味を提供する関係の『提供』、対象が持つ属性名に対する『属性値』、対象が抽象的に持っている、即ち動詞”have” のような関係の『抽象的所有』、そして『部分』の4つについての知識。

次の2通りに分けて処理する。

1. 名詞句1が未登録語の場合
   名詞句2が部分、属性値、所有として持っているカテゴリーすべてを候補とする。
2. 名詞句2が未登録語の場合
   名詞句1を部分、属性値、所有として持ってい

るカテゴリーすべてを候補とする。また、名詞句1の主たる名詞が対応する動詞を持つ場合は、その動詞の主体と客体の格フレームも、候補に加える。

【例2】…… and pains of the dromozoa.
→ "dromozoa" は、『感覚・反応』を部分、属性値として持っているカテゴリーに限定

コーパスから実際に未登録語を含む文を抽出し調査したところ、277文中36文（13%）の未登録語に、この知識が利用可能であった。

3.2.2 並列関係

統語的に並列の関係にある単語は、意味的にも同じ上位概念を持つと言える。統語的な並列関係の抽出方法はこれまでにも研究されている[6]。しかし、未登録語の存在より曖昧性が増加してしまうため、複数の可能性が残る場合がある。このような場合はすべてについて考えることにする。構文上、未登録語に対して並列関係が抽出される場合は、未登録語の意味は対応する名詞と同じ上位概念を持つと予想される。実際の処理においては、並列関係にある登録語が1つの場合は、その上位のファセットに限定する。並列関係にある登録語が2つ以上あり、それらがすべて同じ意味マーカを持つ場合は、その意味マーカに限定する。

【例3】…… from pre-existing heart disease or from almost pure asphyxia.
→ "asphyxia" は『生理的現象』と同じ上位概念を持つカテゴリーに限定

調査の結果、277文中65文（23%）の未登録語に、この知識が利用可能であった。

また、並列関係にあるが意味が全く異なることも稀にある。その場合には、別の処理が必要となるが、本稿では考慮しない。

3.2.3 前置詞句

前置詞を伴う名詞句は、その前置詞によって意味限定可能である。このような情報に対しては、前置詞に、伴うことのできる名詞に関する知識をつけて処理する。

【例4】All the birds in my birdroom appeared ….
→ "birdroom" は、『場所』を提供するカテゴリーに限定

調査の結果、277文中34文（12%）の未登録語に、この知識が利用可能であった。

3.3 文レベル

このレベルにおいては、動詞の格フレームによる情報より推定を行なう。動詞については高松[7]によって分類されたものを一部利用し、状態動詞、動作動詞についてそれぞれ考える。

3.3.1 状態動詞

状態動詞の場合、主体と対象は、動詞との関係よりもその相互関係のほうが密接である。つまり、主体と対象が共に存在し、未登録語がそのいづれかである時、このレベルにおいて推定を行なう。状態動詞は、意味により分類されたカテゴリーを利用する。

【例5】 ｂｅ動詞 （属性）
主体と対象との関係…『同マーカ』『下位－上位』『対象－属性値』

3.3.2 動作動詞

動作動詞の場合、動詞とその格フレームとの関係が深い。動作動詞は多義であることが多く、分類されたものを利用すると情報量が非常に少なくなってしまう。そこで、多少効率は悪いが、単語ごとについて、表層的な格（前置詞などによる）別に辞書にその知識を登録する。

しかし、格フレーム情報の学習法にも問題があり、現状の段階では1つ1つの動詞について調査を行なわなくてはならず、あまり実用的ではない。共起対象により分類された動詞の利用が課題として挙げられる。

3.4 知識源の利用

3.4.1 利用可能な知識源

以上のような知識源はいかなる場合でも利用できるというわけではない。つまり、未登録語の位置により、利用可能な知識源は異なる。形態素レベルはすべ

ての未登録語について調べる。句レベル、文レベルは次の条件を満たすときのみ調査を行なう。

句レベル　：構文木において、未登録語が、前置詞句のすぐ下にある名詞句の主要語となっている。

文レベル　：構文木において、未登録語が、動詞句、または動詞句のすぐ下の前置詞句の下にある名詞句の主要語となっている。

3.5 情報の統合

形態素、句、文のそれぞれのレベルにおける情報は、それらの積集合を取ることにより情報を統合することができる。同じ未登録語が複数の文において出現し、それらの情報を統合するときは、名詞の多義性について考える必要があるだろう。未登録語が異なる意味で使われているかもしれないからである。しかし辞書が充実している場合は、登録されていない単語が複数の意味を持っているとは考えにくい。つまり、複数の場所からの情報の統合も、ただ積集合をとるという手法で十分であろう。

4 知識源の有効性

4.1 有効性の調査

未登録語の意味を得るための有効な手法について、調査を行なった。調査は、*BROWN CORPUS**と*LOB CORPUS*を対象とし、単語辞書は2章で述べた辞書と同じものを利用した。形態素レベルにおいては未登録語337種について、句、文レベルにおいては小文字のみの未登録語100種、出現文277文について、それぞれ調査を行なった。

調査は、情報の有効性を限定可能な意味マーカの数により次の4つのランクに分類して、一文ごとに前章で述べた知識源の各レベルが、有効性のいづれのランクに入るかを調べる、と言ったものである。

1．4個以下の意味マーカに限定可能
2．4個以上20個以下の意味マーカに限定可能
3．21個以上の意味マーカに限定可能
4．意味限定不可能

*Norwegian Computing Centre for the Humanities より入手した。

調査結果を表6に示す。

また、未登録語には以下のような特徴が見られた。

1. 並列関係の名詞が同じ文中に出現していることが多い。
2. ｂｅ動詞の主体又は対象の位置にあることが多い。

この特徴を明確にするために、コーパスから無作為抽出した登録語（100種、277文）との比較調査を行なった。結果を表7に示す。この表は、未登録語は100種(累計277語)のうちその44種に並列関係がみられたことを示す。

4.2 考察

前節の知識源の有効性の調査において、有効といえる情報は、有効性のランク1と2である。この和で比較すると、低いレベルほど有力な情報が得られることが表6よりわかる。形態素レベルが有効であるという事実は、未登録語は、登録語を派生させて新しくつくり出した語である可能性が高いことを示している。

句レベルが有効である理由として、未登録語の含まれている名詞句と並列関係で同じ意味を持つと推定できる名詞句がとても多いことが挙げられる。表7より、未登録語は登録語と比較すると、並列な意味関係の語が倍近く出現していることがわかる。これは、未登録語は、実際に人間が読む時にも未知の概念となりやすいため、文章を書く際に、並列関係をあえてつくり、理解しやすいように意味を補っているからだと考えられる。

また、ｂｅ動詞文の主体又は対象の位置にある未登録語の出現回数も、登録語と比較すると、倍以上であった。この理由も、理解のしやすさのためであろう。

表6: 未登録語の処理方法の有効性

レベル＼ランク	1	2	3	4
形態素レベル	６１％	０％	０％	３９％
句レベル	２２％	２３％	８％	４７％
文レベル	９％	２０％	２４％	４７％

表7: 未登録語の特性（登録語との比較）

	未登録語	登録語
並列関係	４４/１００	２４/１００
ｂｅ動詞文	１８/１００	８/１００

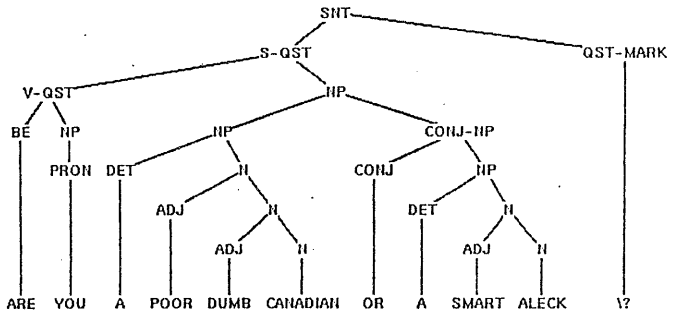

図 2: 構文解析木

```
(****** UNKNOWN WORD ******)
ALECK
(****** PROPERTIES OF NOUN *******)
((YOU ARE SUBJ 0) (CANADIAN ARE OBJ 1)
(ALECK ARE OBJ 1))
(****** PHRASE LEVEL ******)
(------ SAME CATEGORY -------)
(CANADIAN ALECK)
(---MEANINGS OF THE PAIR---)
CANADIAN --> (OH)
(---RESULT OF INFERENCE-----)
(OH OB OP OX)
(****** SENTENCE LEVEL ******)
(ALECK (OV OH AF AC AT AP AX))
(**** UNIFY INFORMATIONS ****)
(OH)
>
```

図 3: 意味推定結果

これら 2 つの特徴は、未登録語の意味推定を行なう上で、非常に重要な知識源となる。

文レベルにおいてはｂｅ動詞による説明文以外からは、多くの情報量は得られなかった。

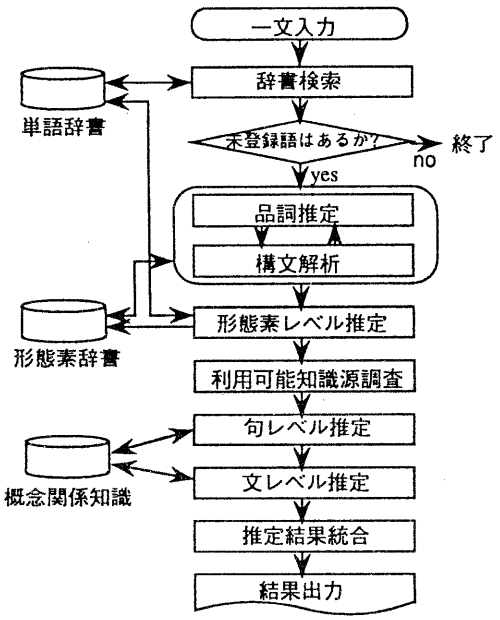

図 1: 処理の流れ

5 試作システムの概要

前章までの考えをもとに、未登録語の意味推定システムを SUN-4 Common Lisp 上で実現した。この処理の流れを図 1 に示す。構文解析はチャート法 [8] を用いた。

このシステムに、コーパスにおいて実際に出現した次の文を入力した。

入力文　：Are you a poor dumb Canadian or a smart aleck ?

未登録語：aleck

図 2 に構文解析結果を、図 3 に意味推定結果を示す。図 3 においては、まず未登録語"aleck" を検出している。次に、名詞すべての格フレームを決定し、推定段階に入る。句レベルにおいて、aleck と Canadian が並列関係にあることを検出し、未登録語"aleck" の意味を、OH(人間),OB(動物),OP(植物),OX(その他) に限定している。さらに、文レベルでは、OV(生物),OH(人間),AF(形態),AC(状態),AT(構成),AP(特徴),AX(属性その他) に限定し、それらの情報を統合することにより"aleck" は OH(人間) であると推定している。

6 おわりに

具体的なデータをもとに未登録語処理の必要性を示した。また、未登録語の意味推定のための実現可能な知識源について整理を行ない、その知識源の有効性についての調査した。この調査結果より、未登録語は、

並列構造、ｂｅ動詞文の中に存在することが多く、この特徴が、意味推定において重要な部分を占める、ということ示した。さらに、この知識源を利用した試作システムを構築した。

今後の課題として、文脈レベルの考察が挙げられる。今回利用した知識源は、人間の思考の一部分であり、文脈に依存したようなものについては考えなかった。しかし、実際にはこのような思考は大変重要だと思われる。人間の思考における処理方法として次の３通りが提案できる。

１．定冠詞ｔｈｅによる指示関係の利用
２．スクリプトの利用
３．文章中にある関連の深い単語の利用

未登録語は、人間にも意味処理の難しい単語である。そのような単語は、それ自体を直接使用して繰り返さずに、その単語を一般化した単語にしてから繰り返すのではないか、と思われる。

【例】
Bill was working at a lathe the other day.
All of a sudden the machine stopped turning.

この例においては、未登録語"lathe" は、後ろの文においては"machine" という単語に一般化されている。しかし、この時照応の曖昧性が問題になる可能性がある。

２．は一連の動作についての知識とそれを選択する知識を利用する。実際に計算機システム上において実現するためには、かなりはじめから分野を限定しなくてはならないであろう。

３．は、単語の意味分類をさらに細かくし、それらについてさまざまな関係を明確に記述した概念関係辞書が、必要となる。

これら以外にも、動詞の意味を限定するという理由もあり、文脈レベルの考察は未登録語の研究においても、重要となるであろう。

最後に、日頃から研究に関して有益な御助言を賜わる研究室の皆様に謝意を表します。

参考文献

[1] Granger, R.H.: FOUL-UP: A Program that Figures Out Meanings of Words from Context, IJCAI(1977).

[2] 塚田、西野、小柳： 未登録語を含む文の一解析法, 情報処理学会自然言語処理研究会資料,73-6,pp.43-50(1989).

[3] 宇佐美、大西、杉江：未登録語を含む英文の構文解析システム, 電子情報通信学会技術研究報告, NLC90-49,pp.1-8(1991).

[4] 石川、坂本、佐藤：Mu プロジェクトにおける意味マーカの概念と体系, 情報処理学会自然言語処理研究会資料,84-46,pp.1-12(1985).

[5] A.S.Hornby 著、伊藤健三 訳：英語の型と語法, オックスフォード大学出版局 (1977).

[6] 武田紀子：英日機械翻訳システムにおける並列関係の検出, 情報処理学会自然言語処理研究会資料,91-2,pp.9-16(1992).

[7] 高松、西田：動詞パターンと格構造に基づく英日機械翻訳, 電子通信学会論文誌,64-d,No.9,pp.815-822(1981).

[8] Kay,M.: Algorithm Schemata and Data Structure in Syntactic Processing, Technical Report CSL-80-12,Xerox PARC,OCT(1980).

[9] Graeme Hirst: Resolving Lexical Ambiguity Computationally with Spreading Activation and Polaroid Words, in S.I.Small,G.W.Cottrell, M.K.Tanenhaus,Lexical Ambiguity Resolution, pp.73-107,Morgan Kaufmann Publisher(1988).

[10] 竝木崇康: 語形成, 大修館書店 (1985).

[11] 井上、山田、河野、成田：現代英文法６ 名詞, 研究社 (1985).