英文中に出現する未登録語の特徴分析とその意味推定のための知識の有効性に関する考察

山田一郎[†] 山村毅[††] 佐川雄二[††]
大西昇[††] 杉江昇[†††]

自然言語の解析を行う場合，辞書に登録されていない単語（未登録語）が現れることがある．新語，外来語，入力ミスによる誤字，脱字のある単語などがこれにあたる．このため，実用的な自然言語処理システムの構築を考える際には，未登録語の処理をする必要がある．しかし，これは必ずしも容易ではない．例えば英単語などは多義，多品詞であることが多いので，組み合わせ爆発などの問題が生じる．このため，未登録語の意味推定法を考える前にまず未登録語の性質を明らかにし，効果的な手法を考える必要がある．本論文では，比較的大規模なコーパスと電子化辞書を用いることにより，実際に出現した未登録語を抽出し，それらの特徴を考察する．具体的には未登録語の出現頻度，種類別出現頻度，品詞別出現頻度についての調査とその考察を行う．この調査結果から，名詞未登録語の処理が最も重要であることを確認する．そして，対象を名詞に限定して意味推定のための知識をまとめる．知識は，形態素レベル，句レベル，文レベルのものを考え，どの知識が意味推定に効果的であるかを考察した結果，形態素レベルの知識が最も有効であり，文レベルの知識はあまり有効ではないことを確認する．また，名詞未登録語は陳述文中，並列句中に存在することが多い，という統語的特徴も示す．この特徴は未登録語の意味推定のための大きな手がかりになると思われる．

On Analysis of the Characteristic of Unknown Words and Examination of the Available Method for Inferring Meanings

Ichiro Yamada,[†] Tsuyoshi Yamamura,[††] Yuji Sagawa,[††] Noboru Ohnishi[††]
and Noboru Sugie[†††]

The existence of unknown words, not registered in a dictionary, causes a break on inaccurate result in natural language processing. For the practical system, it is necessary to suppose the existence of unknown words. But it is difficult to infer meanings of unknown words, because English words usually have many parts of speech and meanings. We have studied on estimation of meaning of an unknown word in order to construct a robust natural language processing system. First, we have investigated the occurrence rate of unknown words and classified the words using "Lob Corpus", a comparatively large corpus, and electronic dictionary. Second, based on the results we have obtained effective clues for inferring meanings in each level of morpheme, phrase and sentence. And, we have examined the availabilities of each method. Moreover, we have found the syntactical characteristics of unknown words, which seems important in inferring.

1. はじめに

自然言語の解析を行う場合，辞書に登録されていない語（以下，未登録語）が現れることが多い．近年では，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ており，登録単語数が十数万語を超えるものも珍しくはない．しかし一方で，毎年非常に多くの新語が作り出されており，こういった日々増え続ける新語のすべてを登録するのは，非常に困難である．この新語の問題は，計算機の記憶技術の向上のみでは解決できない．また，新語のほかにもシステムへの入力ミス等による誤字，脱字といったものも，システムが扱う上では，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

† NHK 名古屋放送局
NHK Nagoya Broadcasting
†† 名古屋大学工学部情報工学科
Department of Information Engineering, School of Engineering, Nagoya University
††† 名城大学理工学部
Faculty of Science and Technology, Meijo University

析，意味解析において，可能な限り，品詞，意味の推定を行い，正常に解析を終了させる方法を考えることが必要になる．

英語の単語は多義，多品詞であることが多く，また，未登録語の存在のため文の解釈が一通りに決まらないこともあり，未登録語の意味，品詞を特定することは困難である．同様のことは他の言語についてもいえる．実際に，人間が考えても全く意味を推定できない，あるいは，複数の解釈ができるような未登録語も存在する．従って，未登録語の意味推定には，得られる限りの情報を最大限かつ有効に利用することが必要である．

英文を対象とした未登録語に関する研究は，その品詞を推定するものと，その意味を推定するものとの二つの分野で研究がなされている．

品詞推定では，文中の未登録語に品詞を与えるために接尾辞を用いるほかに，語間の共起確率や語間の関係規則を用いることが提案されている[1)~5)]．これらの手法では，構文解析に先だって未登録語の品詞推定を行うが，文全体の統語構造を考えていないために，しばしば誤った推定がなされてしまう．また，品詞推定と構文解析とを切り離して考えているために，品詞推定の結果が誤っていると正しい統語構造が得られない．

一方，宇佐美らは構文解析による情報を利用する手法を提案している[6)]．従来までは，構文解析を行う前に未登録語の品詞を特定していたが，宇佐美らの手法では，構文解析終了まで未登録語の品詞を特定せずにいるため，比較的良い結果が得られている．

意味推定では，スクリプト[7)]を利用する手法，動詞の格フレーム情報を利用する手法などが提案されている[8),9)]．しかし，前者はスクリプト選択の問題が，後者は動詞の多義性の問題が残されている．このほか，派生語の意味推定に関する研究[10)]やSFと呼ばれる一種の動詞型を獲得する手法に関する研究[11)]もあるが，いずれも実用レベルには達していない．Jacobs と Zernik[12)]は，形態素や統語を用いて未登録語の意味を推定し，さらに格フレームや語用論的知識を用いることによりそれらの意味を限定・学習する方法を提案している．また，Zernik と Dyer[13)]は，例文から未登録語の語彙を獲得する方法を提案している．彼らは簡単なシステムを作成しているが，その方法が実用上どの程度有効であるかについては議論されていない．

日本語に関しては，日本語文がわかち書きをしないという特徴を持つために，効率の良い未登録語の抽出や形態素解析[14)]に関する研究やこれらの問題に絡んだ構文解析法[15),16)]に関する研究が行われている．

実用レベルで意味解析を行うことは，大変困難であると思われる．一般的な文は曖昧性が大きいため，意味推定のための情報がとても少なくなるからである．この情報を最大限有効に利用するために，本論文では，未登録語の特徴を明らかにすることを試みる．これまでにも，同種の調査が行われたことがあったが，それらは品詞の調査に関するものであった[2)~4)]．

以下では，まず最初に，比較的大規模なコーパスと電子化辞書を用いることにより，未登録語の性質について考察する．具体的には未登録語の出現頻度，形態上の種類別出現頻度，品詞別出現頻度についての調査とその結果に対する考察を行う．次に，名詞未登録語の意味推定を計算機上で処理を可能とするための知識を整理し，これらの有効性を検討する．最後に，未登録語の出現位置についての統語的な特徴を示す．

2. 未登録語の特徴調査

ここでは，比較的大規模なコーパスと辞書を利用して，実際に出現する未登録語の抽出を行い，その特徴を考察する．

2.1 未登録語の定義

未登録語を次のように定義する．

(定義) その語自体が機械辞書に含まれていないすべての単語．(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含まない)

つまり，実際に自然言語を処理する際に，処理が止まってしまう可能性のある単語のすべ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，切り出された語の中に，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ことにする．

2.2 特徴の調査

一口に未登録語といってもさまざまなものがある．ここでは，未登録語処理に対する必要性がどれくらいあるのか，また，どういった特徴（品詞等）を持った未登録語を考慮すべきかを明確にするため，実文章に対して，実用上の辞書を適用して，未登録語の抽出を行い，それらの整理を行った．

対象文は，一般的な英文が多数集められているLOB CORPUS☆を用いた．このコーパスには，約100万単

☆ 本論文で使用したコーパスは，*Norwegian Computing Centre for the Humanities* より入手した．

語，約5万4千文が含まれている．また，品詞，簡単な意味の情報が各語ごとにつけられている（**図1**）．

登録単語の偏りを防ぐために，辞書は，*EDR* 電子化辞書と，*UNIX* のスペルチェッカが参照する辞書の二つを利用した．これらには，併せて約 12 万6千種類もの単語が登録されているので，実用レベルの辞書と言ってもよい．

2.2.1 未登録語の出現率調査

コーパス中に出現する未登録語を抽出し，その数を調査した．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教えている．結果を**表1**に示す．

この表からわかるように，未登録語は全単語数の約 1.5% を占めていた．このコーパスは 54,297 文からなっていたので，一文あたりの平均単語数は 18 個になる．従って，平均4文中に1個は未登録語が存在することになる．今回使用した辞書がとても充実していたことを考慮すると，この出来率は，かなり大きい数字といえる．この結果より，未登録語に対する処理の必要性が再確認できる．

2.2.2 未登録語の種類の調査

抽出された未登録語を，大文字で始まる単語，小文字のみからなる単語，小文字で始まり省略記号を含む単語，小文字で始まりハイフンを含む単語の4種類に分類し，その数を調査した．なお，対象としたコーパスは，文の先頭の単語でも小文字で書かれているので，それを大文字で始まる単語と誤認することはない．すなわち，大文字で始まっている単語は，真に大文字で始まる単語（固有名詞など）を指している（他のコーパスでは，文は大文字で始まっていることが多いので本論文と同様の調査を行うことは困難である）．結果を**表2**に示す．

大文字で始まる未登録語が多いことが目立つが，そのほとんどは固有名詞であった．実際に，4,971 種類中 4,786 種類（96%）が固有名詞であった．固有名詞は，その意味が，人，場所，組織，施設，生産物，理論に限定されるので，「大文字で始まる単語はこれらのいずれかであると仮定して，その中でもっともらしいものをその意味とする．」という方法をとるだけで十分である．限定する意味が，それぞれかなり離れた（相関性が少ない）意味となっているため，その処理は，容易であると思われる．

ピリオドによる省略形の単語の未登録語は，64種類中 32 種類（50%）が単位を意味する単語であった．単位を意味する単語は，そのほとんどが数字の後に位置しているので，この事実を利用することにより，容易に処理することができる．これを除けばピリオドによる省略形の単語の未登録語の出現頻度は 1% 未満になる．

ハイフンを含む未登録語は，214種類中88種類（41%）が，接頭辞（co-, neo-, pre-, un-, など）によるものであった．従って，接頭辞の処理を考慮することが，有効かつ必要なことがわかる．

表1 未登録語の出現率
Table 1 Occurrence rate of unknown words.

全単語数	1,013,851語
全未登録語数	14,892語

表2 未登録語の種類
Table 2 Occurrence rate of classified unknown words.

未登録語の種類	単語数	累計
大文字で始まる単語 例：Herold, Lytton, Nato	4,971種	11,151個 (83%)
小文字ではじまる単語 例：asphyxia, caesium, negro	948種	1,776個 (13%)
省略形の単語 例：approx., i.r.s., r.m.s.	64種	212個 (2%)
ハイフンのある単語 例：avant-garde, co-existence	214種	336個 (2%)
全体	6,197種	13,475個

```
D01   2    ^ with_IN so_QL many_AP problems_NNS to_TO solve_VB ,_, it_PP3
D01   2 would_MD be_BE a_AT great_JJ help_NN to_TO
D01   3 select_VB some_DTI one_CD1 problem_NN which_WDTR might_MD be_BE the_ATI
D01   3 key_NN to_IN all_ABN the_ATI others_APS ,_, and_CC
D01   4 begin_VB there_RN ._. ^ if_CS there_EX is_BEZ any_DTI such_ABL
D01   4 key-problem_NN ,_, then_RN it_PP3 is_BEZ undoubtedly_RB
D01   5 the_ATI problem_NN of_IN the_ATI unity_NN of_IN the_ATI gospel_NN ._.
D01   5 there_EX are_BER three_CD views_NNS of_IN the_ATI
D01   6 fourth_OD gospel_NN which_WDTR have_HV been_BEN held_VBN ._.
```

図1 Lob Corpus テキストの例
Fig. 1 Examples of Lob Corpus text.

小文字のみの単語の未登録語については，その特徴，有効な処理方法が曖昧なため，さらに次節以降に示すような調査を行った．

2.2.3 未登録語の品詞調査

小文字のみからなる未登録語を対象に，その品詞を調査した．結果を**表3**に示す．ここで，英語以外の単語については品詞を考えず，外来語として分類した．

この結果より，名詞，形容詞，外来語（英語以外の語）の頻度が高いことがわかる．しかし，英語以外の単語を英文中に引用するときは，イタリック体にするという約束があるため，この語の抽出は容易である．また，その単語の意味を必要とすることは少なく，特に意味推定する必要はないと思われる．

2.2.4 派生語の調査

小文字のみからなる未登録語を対象に，語幹が登録語であるかを調査した．結果を**表4**に示す．

この結果より，小文字のみからなる形容詞と副詞の未登録語の大部分は，登録語に接頭辞，接尾辞を付加した単語，つまり登録語からの派生語であることがわかる．つまり，形容詞と副詞の未登録語は，派生に関する処理が最も有効で，また，この処理だけでほぼ十分であることがわかる．

名詞と動詞の未登録語も，登録語からの派生語が多く，派生に関する処理は，有効な一手段であるといえる．

また，表4の結果から，登録語からの派生でない未登録語は94%が名詞と動詞に含まれるので，新たな意味の単語を作り出すときは，名詞か動詞になりやすいことがわかる．その絶対数が多い名詞には，言葉の新造能力があり，新語は名詞が多いということが，このことからわかる．

3. 名詞未登録語の意味推定のための知識

名詞の意味を考えるためには，意味概念の分類が重要な課題となる．意味概念といってもさまざまなものがあるが，本論文では，科学技術庁のMuプロジェクト[17]においてなされた分類（**表5**）を利用する．これは12個のファセット（上位概念）と48個の意味マーカ（下位概念）からなる．我々は，名詞の意味が，これらのいずれに属するかを決定することを「意味の推定」と呼ぶことにする．意味の推定には，形態素レベル，句レベル，文レベルの三つのレベルが考えられる．ここでは，これら各レベルについて述べる．

3.1 形態素レベルの推定

3.1.1 接頭辞による派生語

通常，接頭辞は，単語に意味の変化を与える．しかし，表5の意味マーカはあらい分類であるため，接頭

表3 小文字で始まる未登録語の品詞
Table 3 Parts of speech of unknown words with no capital.

品詞	単語数	累計
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%)	214個
(外来語)	326種 (34%)	571個
その他	19種 (2%)	31個
計	948種	1,776個

表4 形態素により処理が可能な割合
Table 4 Rate of derivative from registered words.

品詞	単語数	登録語からの派生
名詞	337種	127種 (38%)
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	341種 (57%)

表5 名詞意味マーカ体系[4]
Table 5 Organization of noun meanings.

ファセット	意味マーカ
国・機関・組織	
生物	人，動物，植物，その他
無生物	自然物，部品および材料，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，知的抽象的生産物，その他
部分	部分・要素，生物の器官および構成要素，その他
属性	属性名，関係，形態，状態，構成，特徴，その他
現象	自然現象，力・エネルギー，生理的現象，社会的現象，物象，制度・慣習，その他
心得	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

辞がつくことによってその範囲を越えて意味が変化することは少ない．従って，接頭辞とその基本が登録語であるなら，意味は，その基体の意味マーカと同じであると推定できる．

(例1) "unawareness"

この例で，"unawareness" は "awareness" と同じ意味マーカ「認知」と推定できる．

3.1.2 接尾辞による派生語

接尾辞は意味の変化を与えない．主として，品詞の変化を与えるだけである．しかし，接尾辞は，接続できる語幹との間に意味的な制限があるため[18]，これを用いることにより，その単語の意味を限定することができる．名詞をつくる 26 個の接尾辞について，基体の品詞により単語の意味を分類し，この知識を利用する．**表6** はその知識の一部である．

(例2) "expressionism"

この例で "expressionism" は，接尾辞が "ism"，基体の品詞が名詞ということから「行為」「生理的現象」「理論」と推定できる．

3.1.3 複 合 語

複合語の場合，登録語の切り出しに成功したものは，その主要語，つまり最後の部分にある名詞と同じ意味マーカであると推定できる．

(例3) "applecake"

この例で，"applecake" は "cake" と同じ意味マーカ「生産物」と推定できる．

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[18]．後者は，この手法では誤った推定がなされてしまう．しかし，未登録語となりうるような，新たにつくり出される単語は，ほとんど前者に含まれる．実際に抽出された未登録語 57 種類中 56 種類（98%）が内心複合語であった．従って，この手法で十分であろう．

表 6 接尾辞による情報（一部）
Table 6 Informations of suffix.

接尾辞	基本の品詞	意 味 概 念
-(a)cy	名 詞	名・機関・組織，生産物，行為
	動 詞	行為
	形容詞	状態，特徴
-(e)ry	名 詞	施設
	動 詞	機能
	形容詞	状態，特徴
-ism	名 詞	行為，生理的現象，理論
	形容詞	状態，特徴，理論

3.2 句レベルの推定

句は，まとまって一つの意味をなす．形態素レベルで推定できなくても，あるまとまった句の中で，その前後関係等から意味を限定できることがある．このレベルの知識を利用する場合，句の抽出，係り受け関係の曖昧性が問題となる．この曖昧性を解消するため，係り受け関係に A. S. Hornby[19] の動詞型に基づく拘束条件を与え，構文解析が一意に決まらない場合にはすべての可能性を考えるものとする．

3.2.1 名 詞 句

「名詞句 1＋of＋名詞句 2」の型の名詞句では，名詞句1と名詞句2との間に，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．本論文では，名詞間の関係として，「提供」，「属性値」，「抽象的所有」，「部分」の四つを使用する．

(例4) ……and pains of the dromozoa.

この例で，"dromozoa" は "pain" の意味マーカである「感覚・反応」を属性値，抽象的所有，部分として持つ意味マーカに限定できる．

3.2.2 並 列 関 係

統語的に並列の関係にある二つの単語は，それぞれ同じ上位概念を持つ．従って，未登録語と並列関係にある語が登録語である場合，未登録語の意味は，その登録語の持つ上位ファセットに限定できる．

(例5) ……from pre-existing heart disease or from almost pure asphyxia.

この例で，"asphyxia" は "disease" の意味マーカ「生理的現象」の上位ファセット「現象」に限定できる．

また，異なる上位ファセットを持つ名詞が構造的に並列関係にあることも稀にある．その場合には，細かく記述されたシソーラス（名詞の意味概念関係）を利用した別の処理が必要となるが，本論文では取り扱わない．

3.2.3 前 置 詞 句

前置詞句中の名詞は，その前置詞により，意味限定が可能となる場合がある．この前置詞情報により，未登録語の意味が限定できる．

(例6) All the birds in my birdroom appeared….

この例で，"birdroom" は，"in" の可能な前置詞目的語「場所」「時間」「道具」を提供する意味マーカに限

定できる．

3.3 文レベルの推定

このレベルにおいては，動詞の格フレームによる情報を用いることにより意味の推定を行う．動詞の意味が，属性関係をあらわすもの，それ以外のもの，について考える．

3.3.1 属性関係をあらわす動詞

動詞が抽象的関係を意味する状態動詞である場合，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，動詞は，主体と対象の意味関係を規定しているだけである．このため，主体あるいは対象のいずれか一方から，他方の意味を推定できる．

(例7) …, drowning is not a simple asphyxia…

この例で，be 動詞の主体と対象との関係は，「同マーカ」「下位—上位」「対象—属性値」であるので，“asphyxia” は “drowning” の意味マーカ「行為」と上記の関係にある意味マーカに限定できる．

3.3.2 属性関係をあらわさない動詞

動詞が属性関係をあらわさない場合は，動詞の主体と対象は，動詞との関係のほうが密接である．この，動詞と表層的な格フレームの関係を，各動詞ごとに辞書に登録して，これを知識源とすることにより，主体や対象の意味が限定できる．

(例8) …the dancers began the calinda, …

この例で，“calinda” は，動詞 “begin” の対象格に入ることができる意味マーカに限定できる．

4. 知識の有効性

4.1 有効性の調査

3章で述べた知識の未登録語の意味推定における有効性を検証するために，調査を行った．調査は，BROWN CORPUS と LOB CORPUS の二つのコーパスを対象とし，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 LOB CORPUS から抽出した未登録語 337 種について，句，文レベルにおいては BROWN CORPUS と LOB CORPUS から抽出した小文字のみの未登録語 100 種，出現文 277 文について，それぞれ調査を行った．

調査は，まず，知識の有効性を限定可能な意味マーカの数により以下の四つのランクに分類する．そして，一文ごとに前章で述べた知識源の各レベルが，有効性のどのランクに入るかを調べるものである．

1. 得られた意味マーカが4個以下
2. 得られた意味マーカが5個以上20個以下
3. 得られた意味マーカが21個以上48個以下
4. 知識利用不可能

調査結果を**表7**に示す．この表では例えば，句レベルの知識だけを用いた場合，全体の 22% の未登録語が4個以下，23% が5個以上20個以下，8% が21個以上48個以下の意味マーカに限定可能で，残りの 47% が知識利用不可能な位置にあったことを示す．

調査の段階で，未登録語を含む文には，次の三つの特徴がみられた．

(1)未登録語と並列関係にある名詞が同じ文中に出現していることが多い．

(2)未登録語が be 動詞の主体または対象の位置にあることが多い．

(3)「名詞1＋of＋名詞2」の型の名詞句において未登録語は名詞1より名詞2に位置することが多い．

(1)，(2)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同じ条件にある登録語(100種，277文)との比較調査を行った．結果を**表8**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．また，(3)についての同様の調査結果を，**表9**に示す．

4.2 考　　察

前節の知識の有効性の調査において，有効といえる

表7 未登録語の処理方法の有効性
Table 7 Availabilities of the method for inferring.

レベル＼ランク	1	2	3	4	計
形態素レベル	49%	6%	0%	45%	100%
句レベル	22%	23%	8%	47%	100%
文レベル	9%	20%	24%	47%	100%

表8 未登録語の特徴（登録語との比較）
Table 8 Syntactical characteristics of unknown words.

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

表9 名詞句中にある未登録語の位置の特徴
Table 9 Syntactical characteristics of unknown words in noun phrase.

名詞1	名詞2
8/35	27/35

知識は，有効性のランク1と2に分類されるものである．これらの和で各レベルを比較すると，低いレベル（つまり，文よりも句，句よりも形態素）ほど有効な情報が得られることが表7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単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の一つとして，多くの文において，未登録語と並列関係にある名詞が同じ文中に出現しているということが挙げられる．表8より，未登録語は登録語と比較すると，並列な関係にある語が2倍近く出現していることがわかる．これは，実際に人間が読む時にも未登録語は未知の概念となりやすいため，並列関係を作ることにより，意味を補っているものと考えられる．

また表8より，be 動詞文の主体また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，その2倍以上であった．主動詞が be 動詞である文は一般的に陳述文であるので，これも，未登録語の意味を人が理解しやすいようにしていると考えられる．

これらの特徴は，未登録語の意味推定を行う上で，非常に重要な知識となる．

文レベルにおいては，be 動詞による陳述文以外からは多くの情報は得られなかった．この理由としては，動詞の多義性が挙げられる．実際英語の動詞においては，一つの単語がさまざまな意味を持っていることが多い．そのため，すべての可能性を考えなくてはならず，使用する格フレームが増加するので，意味の限定が困難になる．従って，動詞の多義性の問題の解決は，未登録語の意味推定処理においても必要とされる．

また表9より，「名詞1＋of＋名詞2」の名詞句において，名詞2が未登録語であることは，名詞1より3倍以上多かった．この理由は，名詞1，名詞2の位置に入りやすい名詞の意味に偏りがあるからと考えられる．例えば，名詞句1には「属性」を意味する単語が入りやすい．このことより，「未登録語の意味は限定できる」という仮説が提案できる．これは未登録語の意味処理のための大きな手がかりとなるため，今後，この仮説を立証することは，重要な課題となるであろう．

5. おわりに

本論文では，比較的規模の大きいコーパスと辞書を利用することにより，実際に出現する未登録語の抽出を行い，未登録語の性質を示した．この調査結果より，未登録語処理の必要性を，再確認した．また，未登録語の種類ごとに，その効果的な手法を検討した．次に，未登録語を名詞に限定して，システム上で実現可能な意味推定のための知識を整理した．さらに，その知識の有効性について調査を行い，未登録語を意味推定するための有効な手法を検討した．この調査結果から，未登録語の意味推定は形態素処理が最も有効な手法であることを示した．動詞の多義性の問題の解決が困難である現状において，文レベルの推定法は，それほど有効ではないこともうかがえた．また，未登録語の意味推定においてとても重要である統語的特徴も示した．

本論文の3章で述べた意味推定のための知識は，本来，統合的に用いるべきものである．本研究では調査しなかったが，統合的にこれらの知識を用いた場合の意味推定結果は，単独の場合（表7）よりも向上するものと思われる．今後これに関して，精密な調査が必要である．

今回の調査，考察は，実現可能な推定法について行ったため，文脈レベルは考慮しなかった．しかし，人間が知らない単語の意味を推定する際には，文脈からも情報を得ることができる．この文脈レベルについて，システム上で実現可能な知識を整理し，その有効性について考察する必要がある．

また，システムを実現するためには，文レベルにおける動詞格フレームの学習法が問題となる．近年のコーパスの公開など言語データの整備により，偏りのない学習がある程度は可能となったが，依然検討する必要がある．

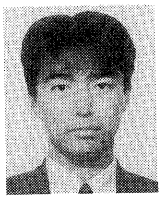
謝辞 本研究を進めるにあたり，日頃から有益な討論をいただいた名古屋大学工学部杉江研究室の皆様に深く感謝します．

参考文献

1) Klein, S. and Simmons, R. F.: A Computational Approach to Grammatical Coding of English Words, *J. ACM*, Vol. 10, pp. 334-347 (1963).
2) Greene, B. B. and Rubin, G. M.: *Automated Grammatical Tagging of English*, Dept. of Linguistics, Brown University, Providence, Rhode Island (1971).
3) Mashall, I.: Choice of Grammatical Word-Class without Global Syntactic Analysis: Tag-

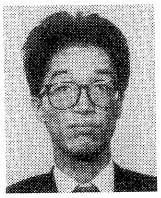
ging Words in the Lob Corpus, *Computers in the Humanities 17*, pp. 139-150 (1983).
4) Steven, J. D.: Grammatical Category Disambiguation by Statical Optimization, *Computational Linguistics*, Vol. 14, No. 1, pp. 31-39 (1988).
5) Church, K. W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol. 16, No. 1, pp. 22-29 (1990).
6) 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, NLC 90-49, pp. 1-8 (1991).
7) Schank, R. C., Riesbeck, C. K. (編), 石崎 俊 (監訳): 自然言語理解入門, 総研出版 (1986).
8) Granger, R. H.: FOUL-UP: A Program that Figures Out Meanings of Words from Context, *Proc. of IJCAI-77*, pp. 172-178 (1977).
9) Hirst, G.: Resolving Lexical Ambiguity Computationally with Spreading Activation and Polaroid Words, Small, S. I., Cottrell, G. W. and Tanenhaus, M. K. (eds.), *Lexical Ambiguity Resolution*, pp. 73-107, Morgan Kaufmann Publisher (1988).
10) Light, M.: A Computational Theory of Lexical Relatedness, Technical Report 421, The University of Rochester Computer Science Department (1992).
11) Brent, R. M.: Automatic Acquisition of Subcategorization Frames from Untagged Text, *Proc. of ACL-91*, pp. 209-214 (1991).
12) Jacobs, P. and Zernik, U.: Acquiring Lexical Knowledge from Text: A Case Study, *Proc. of 7th Conf. of AI*, pp. 739-744 (1988).
13) Zernik, U. and Dyer, M. G.: The Self-Extending Phrasal Lexicon, *Computational Linguistics*, Vol. 13, No. 3-4, pp. 308-327 (1987).
14) 吉村, 竹内, 津田, 首藤: 未登録語を含む日本語文の形態素解析, 情報処理学会論文誌, Vol. 30, No. 3, pp. 294-301 (1989).
15) 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, NL 73-6, pp. 43-50 (1989).
16) 大場, 元吉, 井佐原, 横山, 石崎, 板橋: 未定義語を含む文の多段階構文解析法, 情報処理学会自然言語処理研究会資料, NL 70-4 (1989).
17) 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, NL-84-46 (1985).
18) 竝木崇康: 語形成, 大修館書店 (1985).
19) Hornby, A. S. (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).
20) 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, NL 91-2, pp. 9-16 (1992).
21) 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, Vol. J 64-D, No. 9, pp. 815-822 (1981).
22) 山田, 山村, 佐川, 大西, 杉江: 英文における未登録語の意味推定についての検討, 情報処理学会自然言語処理研究会資料, NL 93-10, pp. 64-71 (1993).
23) Kay, M.: Algorithm Schemata and Data Structure in Syntactic Processing, Technical Report CSL-80-12, Xerox PARC (1980).
24) 井上, 山田, 河野, 成田: 現代英文法 6 名詞, 研究社 (1985).
25) 永井秀利: 文解析における確率の利用, 情報処理学会自然言語シンポジウム論文集, pp. 33-47 (1992).
26) Carter, D.: Interpreting Anaphors in Natural Language Texts, pp. 90-124, SRI International Cambridge Computer Science Research Center (1987).

(平成5年12月13日受付)
(平成6年9月6日採録)



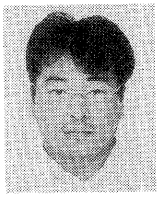
山田 一郎

平成3年名古屋大学工学部情報工学科卒業. 平成5年同大学院情報工学専攻博士前期課程修了. 同年NHKに入局. 現在, 名古屋放送局に勤務. 放送中継, 編集業務に従事. 在学中自然言語処理の研究に従事.



山村 毅 (正会員)

昭和39年生まれ. 昭和62年名古屋大学工学部電気卒業. 平成元年同大学院情報工学専攻博士前期課程修了. 平成4年同博士後期課程単位取得退学. 同年名古屋大学工学部助手. 自然言語処理の研究に従事. 工学博士. 電子情報通信学会, 計量国語学会各会員.

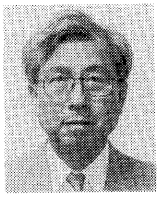


佐川 雄二 (正会員)

昭和60年名古屋大学工学部電子卒業. 昭和62年同大学院情報工学専攻修士課程修了. 同年(株)日立製作所入社. 平成元年名古屋大学大学院情報工学専攻博士課程入学. 平成4年単位取得退学. 同年同大学情報工学科助手. 平成6年同講師, 現在に至る. 自然言語処理の研究に従事. 工学博士. 計量国語学会会員.

大西 昇（正会員）

昭和 48 年名古屋大学工学部電気卒業．昭和 50 年同大学院電気系専攻修士課程修了．同年労働福祉事業団労災リハビリテーション工学センター研究員．昭和 61 年名古屋大学工学部講師．平成元年同助教授．平成 6 年同教授で，情報工学科に所属．コンピュータビジョン・オーディション，ロボティクス，生体工学，福祉工学などの研究・教育に従事．工学博士．電子情報通信学会，計測自動制御学会，ロボット学会，バイオメカニズム学会，日本神経回路学会，IEEE 等各会員．



杉江 昇（正会員）

昭和 32 年名古屋大学工学部電気卒業．同年通商産業省電子技術総合研究所入所．昭和37～39年カナダ・マギル大学客員研究員．昭和45年バイオニクス研究室長．昭和53年視覚情報研究室長．昭和54年名古屋大学大学院工学研究科情報工学専攻教授．昭和60年同大学工学部電気工学第二学科教授．平成2年同大学工学部情報工学科教授．平成6年名城大学理工学部電気電子工学科教授．現在に至る．バイオニクス，医用工学，コンピュータビジョン，自然言語処理などの研究・教育に従事．工学博士．電気学会，電子情報通信学会，計測自動制御学会，ロボット学会，エム・イー学会，テレビジョン学会，バイオメカニズム学会，日本神経回路学会，IEEE 等各会員．