日英対訳文間の素性構造照合による統語的曖昧性の解消†

宇津呂 武仁†† 松本 裕治†† 長尾 眞††

自然言語処理の技術を実用上有効なものとするためには，計算機処理のための大規模な意味辞書の構築が必要不可欠である．特に，人間が読むための辞書や大規模なコーパスなどに含まれる自然言語の文章を解析して，大規模な意味辞書を半自動で構築する技術を確立することが重要である．しかし，自然言語の文章を解析する際には，統語的曖昧性および語彙の多義性という少なくとも二つの問題が生じる．われわれは，これらの問題を解消するために，英語と日本語のように統語構造および語彙が異なる二言語間の翻訳例を構文解析して，その結果を二言語間で比較するというアプローチをとる．二言語間で構文解析結果を比較することによって，多くの場合，統語的曖昧性と意味的曖昧性の両方が解消するものと考えられるからである．本論文では，統語的な知識だけを用いて対訳例の各単言語文を構文解析した結果を素性構造で表現し，対訳辞書の訳語の情報をもとにこの素性構造を二言語間で照合することによって，単言語文の統語的な曖昧性を解消する手法について述べる．二言語間で素性構造を照合するプロセスを実現するために，素性名およびアトミックな素性値の両立可能対に基づく素性構造照合演算を導入した．照合結果の素性構造中の各表層語は，二言語の訳語のペアによって表現されているので，単言語文における語彙の多義性を解消する際にも有用な情報となる．

1. はじめに

自然言語処理の分野においては，これまで構文解析を中心として解析技術が発展してきたが，これらの処理技術を実用上有効なものとするためには，計算機処理のための大規模な意味辞書の構築が必要不可欠である．そのような大規模な意味辞書においては，個々の言葉について，様々なタイプの意味的あるいは語彙的な知識を記述しておく必要がある．言葉の意味的あるいは語彙的な知識としては，主として次のようなものが挙げられる．

- 概念の集合および概念と表層語との関係．
- 概念間の階層関係あるいはシソーラス．
- 名詞概念の持つ属性概念や動詞概念の持つ格要素などの，概念間の非階層的関係．

このような自然言語処理のための意味辞書の構築を目的として，近年いくつかの研究が行われてきている．そこでとられているアプローチは大きく二つに分けることができる．一つは，意味辞書あるいは知識ベースなどを人手で構築するというものである[4)~6),8)]．このアプローチに伴う困難な問題としては，構築された意味辞書の記述に作業者の主観が含まれるために，辞書全体を通して記述が一貫しないなど，結果に不安定な面が出てくることが挙げられる．また，人手で行うために作業の量が膨大であり，意味辞書を拡張することも容易ではない．もう一つのアプローチは，計算機を利用してできるだけ自動的に意味辞書を構築するというものである．このアプローチの例としては，人間のために書かれた辞書から概念間の階層関係あるいはシソーラスを抽出するという研究[15),16)]や，大規模なコーパスの文を構文解析して統計的なデータを蓄積し，そこから語彙的な知識を抽出するという研究[1),3)]がある．計算機を利用したアプローチをとることによって，意味辞書に対する客観的な記述が可能になり，また，意味辞書構築の際の人間の負担が軽減されることが期待される．

自然言語処理技術を用いた知識獲得という観点からは，後者の計算機プログラムによって自動的に意味辞書を構築するというアプローチは非常に重要である．しかし，このアプローチには少なくとも次の二つの問題がある．

1. **統語的曖昧性の問題**
   構文解析の結果にはたいていの場合曖昧性が残るために，自然言語コーパスから正しい構文解析結果を自動的に得ることは難しい．
2. **語彙の多義性の問題**
   一つの表層語が複数の意味を持つことはよくあることであり，その場合，一つの表層語が複数の概念に対応し得ることになるが，文脈に応じて表層語を正しい概念に対応させることは難しい．

われわれは，これらの問題を解消するために，英語と日本語のように統語構造および語彙が異なる二言語間の翻訳例を構文解析して，その結果を二言語間で比

† Syntactic Disambiguation by Matching Feature Structures of Japanese-English Translation Pairs by TAKEHITO UTSURO, YUJI MATSUMOTO and MAKOTO NAGAO (Department of Electrical Engineering, Faculty of Engineering, Kyoto University).

†† 京都大学工学部電気工学第二教室

較するという方法を用いる．英語と日本語のように統語構造の異なる二言語の場合には，それぞれの言語が異なったタイプの統語的曖昧性を持つために，多くの場合，二言語間で構文解析結果を比較することによって統語的曖昧性が解消されると考えられる．また，概念が二言語間の中間言語的なものであると考えれば，二言語間の訳語のペアに対応し得る概念の集合は，単言語の表層語に対応し得る概念の集合の交わりとなると考えられるので，二言語間の訳語のペアを用いることによって概念レベルでの曖昧性が減少する[2),14)]．さらに，動詞と名詞の格関係のような概念間の関係についても，二言語間の関係のペアを用いることによって関係の可能性を絞り込むことができる．特に英語においては，主語・目的語のような統語的な特徴によって格関係を表現することがよくあり，これが有用な情報となる場合が多いと考えられる．例えば，例1の対訳例の場合には，統語的な曖昧性と意味的な曖昧性の両方が解消する．

例1

英語：I hung my coat on the hook.

日本語：私は上着をかぎにかけた．

1. **統語的曖昧性の解消**

   例1の英文においては，構文解析を行っただけでは前置詞句 "*on the hook*" が動詞 "*hung*" と名詞句 "*my coat*" の両方にかかり得るので，統語的曖昧性が残る．一方，日本文のほうでは，"かぎ，に" という句は動詞 "かけた" にしかかかり得ない．したがって，対訳辞書を用いることによって，〈*I*, 私〉，〈*hung*, かけた〉，〈*coat*, 上着〉，〈*hook*, かぎ〉という訳語に関する情報が利用できれば，構文解析結果を二言語間で比較することによって，英文の前置詞付加の曖昧性を解消することができる．

2. **意味的曖昧性の解消**

   例1の日本文における動詞 "かける" は，典型的な多義動詞である．例えば，収録語数約70,000語の国語辞典において，この動詞の小見出しは六つある．また，収録語数約50,000語の和英辞典において，この動詞の訳語は10個 ("*hang*", "*spend*", "*play*" など) ある．したがって，"かける" という動詞を適切な意味に対応させるのは容易ではない．しかし，例1のような対訳例を用いれば，英語の訳語 "*hung*" によって "かける" の意味を限定することができる．

このような方法を用いることによって，対訳例の文の統語構造を一意に決定することができれば，対訳例を曖昧性なく構文解析したデータを大量にしかも自動的に蓄積することが可能になる．その結果，構文解析済みのデータを大量に収集したものに対して統計的な処理を行うことによって，動詞の表層格などの語彙的な知識をある程度自動的に獲得することが実現可能になると考えられる．特に，この方法によって蓄積されたデータには，単言語だけの場合と比べて，多義動詞の意味の分離を行ったり名詞と動詞との間の正しい格関係を決定するために必要となる情報がより多く含まれているので，語彙的知識を獲得する際には有利である．

本論文では，このような考えに基づき，二言語間で素性構造を照合することによって，それぞれの言語の文の曖昧性を解消する手法を提案する．われわれの手法においては，まず，対訳例の各単語文を構文解析し，構文解析結果を表層語の依存構造とほぼ等価な素性構造に変換する☆．ここで，構文解析結果を素性構造で表現する理由は次の2点である．一つは，素性構造に対して，論理的な意味での形式化がある程度なされているので，二言語間で素性構造を照合するプロセスを定式化するのが比較的容易であるからである．もう一つは，将来，曖昧性の解消したデータを用いて，動詞概念の格フレームや名詞概念の属性などの語彙的知識を獲得することを考えた場合，それらの語彙的知識の表現形式として素性構造をそのまま用いることができ，しかも語彙的知識獲得の過程を，同じく素性構造上の演算として定義できるからである．

次に，対訳辞書の訳語の情報を参照することによって，二言語間で素性構造を比較し，照合する．対訳辞書の訳語の情報を参照した結果，英文におけるある単語が日本文における複数の単語に対応し得る場合がある．逆に，日本文におけるある単語が英文における複数の単語に対応し得る場合もある．このように，単語レベルにおいては二言語間で多対多の対応が可能な場合も含めて，二言語間で素性構造を照合するプロセスを実現するために，素性構造における素性名およびアトミックな素性値のうちで両立可能なペアを集めた集合（素性名およびアトミックな素性値の両立可能対集合，Sets of Compatible Pairs of Feature Labels and Atomic Values）に基づく素性構造照合演算（Matching Operation of Feature Structures）を定式化する．

☆ ここで素性構造によって表現されている内容は，具体的には主題構造 (thematic structure) または機能構造 (functional structure) に近いものである．

二つの素性構造を照合するための基本的な演算として単一化演算があるが，本論文で導入する素性構造照合演算は，従来の単一化演算を素性名およびアトミックな素性値の両立可能対集合に基づく演算に修正することによって定式化されている．第2章では，この両立可能対集合に基づく素性構造照合演算について説明し，さらに，二言語間の素性構造照合を実現する方法について述べる．第3章では，二言語間の素性構造照合によって統語的曖昧性が解消される例について述べる．今後，単言語の解析結果として素性構造によって記述された意味表現が得られるようになると，意味表現上で二言語間の素性構造照合を行うことによって意味的曖昧性の解消を行うことも原理的には可能である．ただし，現段階では，単言語解析の結果得られる素性構造は表層語の依存構造とほぼ等価なものであり意味的曖昧性の解消にまでは至っていないため，本論文では統語的曖昧性の解消についてのみ述べる．

第4章では，計算機上で利用可能な和英辞典（講談社学術文庫の和英辞典[13]）から抽出した簡単な対訳例に対して，統語的曖昧性解消の実験を行った結果を示す．現在のところ，この和英辞典から約40,000対訳例が取り出されている．また，統語的曖昧性解消の成功率は，63〜68%であった．前述したように，表現形式として素性構造を用いれば二言語間の素性構造照合の定式化が容易になる反面，演算方式が比較的粗いために曖昧性の解消に失敗するような例もでてくる．そこで，このような例も含めて，曖昧性の解消が失敗する例についてその原因を分析する．最後に，第5章で，全体のまとめおよび今後の発展について述べる．

2. 二言語間の素性構造照合

われわれの手法においては，まず，対訳例の単言語の文を構文解析する．構文解析の結果は，表層語の依存構造とほぼ等価な素性構造に変換される．その際，素性名に相当するのは，表層格ラベルになるものやその他の機能語であり，アトミックな素性値に相当するのは自立語である☆．また，機能語以外にも，時制(tense)・法(modal)・態(voice)などの文法的特徴を表す素性がある．例2のような対訳例について考える．

例2

英語：I wrote a letter with a pencil.

日本語：私は鉛筆で手紙を書いた．

☆ 英語の場合，各品詞のうち，アトミックな素性値に相当するのは，名詞・代名詞・動詞・形容詞・副詞・冠詞である．接続詞・前置詞・関係代名詞などは素性名に相当する．また，助動詞は法(modal)に関する素性を導入すると考える．

まず，この例の英文からは，前置詞付加の曖昧性によって次の二つの素性構造が得られる．

$$\begin{bmatrix} pred : write \\ tense : past \\ subj : [pred : I] \\ obj : \begin{bmatrix} pred : letter \\ spec : a \end{bmatrix} \\ with : \begin{bmatrix} pred : pencil \\ spec : a \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} pred : write \\ tense : past \\ subj : [pred : I] \\ obj : \begin{bmatrix} pred : letter \\ spec : a \\ with : \begin{bmatrix} pred : pencil \\ spec : a \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

日本文のほうからは，次の一つの素性構造が得られる．

$$\begin{bmatrix} pred : 書く \\ tense : past \\ は : [pred : 私] \\ を : [pred : 手紙] \\ で : [pred : 鉛筆] \end{bmatrix}$$

二言語間で素性構造の照合を行う際には，対訳辞書から得られる訳語の情報を用いることによって，片方の言語の素性構造中のアトミックな素性値がもう一方の言語の素性構造中のアトミックな素性値と同じ対象を表しているかどうかがわかる．素性構造中の素性についても同様のことが言える．つまり，対訳辞書から得られる訳語に関する知識は，素性およびアトミックな素性値の両立可能な対に関する知識とみなすことができる．この観点に基づき，本章では，素性およびアトミックな素性値の両立可能対集合に基づく素性構造照合演算を定式化し，さらに，二言語間の素性構造照合の実現法について述べる．

2.1 素性およびアトミックな素性値の両立可能対集合に基づく素性構造照合演算

データ構造

まず，われわれが用いている素性構造のデータ構造を定義する．素性構造の記述形式としては，Kasperらの *Feature Description Logic* (FDL)[7]の記法に基づき，それを拡張したものを用いる．

A および L を，それぞれ，アトミックな素性値および素性名を表す記号の有限集合とする．また，C_A および C_L を，それぞれ，素性構造におけるアトミックな素性値および素性名のうちで両立可能なペアを集

めた集合とする．すなわち，Aの任意の二つの要素a_i，a_jのうちで，お互いに矛盾がなく両立するペア$\langle a_i, a_j\rangle$を集めた集合をC_Aとし，Lの任意の二つの要素l_i，l_jのうちで，お互いに矛盾がなく両立するペア$\langle l_i, l_j\rangle$を集めた集合をC_Lとする．C_A，C_Lをそれぞれアトミックな素性値の両立可能対集合，素性名の両立可能対集合と呼ぶ☆．

両立可能対集合に基づく素性表現論理（*FDL with Sets of Compatible Pairs* (FDLC)）の式は，**図1**のように定義される．ここで，図1中のパス等式における各パス*pi*は，$C_L{}^*L^*$の要素である．これは，二言語間で素性構造を照合した結果に含まれるパスにおいては，二言語間でペアになった素性が任意個続いた後，単言語の素性が任意個続くという形をとるからである．

素性構造照合演算

両立可能対集合に基づく素性構造照合演算は，従来の素性構造単一化と同様に，素性構造に対する等式によって定式化することができる．素性構造に対する等式は，**図2**のように定義される．これらの等式は，左辺の式を右辺の式に置換することによって用いる☆☆．ここで，(15)，(16)式以外の式は，通常の素性構造単一化に用いる等式を拡張したものである．また，(15)式および(16)式は，二言語間で素性構造を照合する際に，単言語でのパス等式の制約を照合結果に反映させるためのものである．(15)式では，二言語間でペアになった素性を含むパスを見つけて，要素が一つのパス等式を新たに追加する．(16)式では，単言語でのパス等式の制約を利用して，二言語間でペアになった素性を含む二つのパス等式をマージする．

☆ これらの両立可能対集合は，必ずしもアトミックな素性値および素性の同値関係を定義するとは限らない．すなわち，推移律および対称律を満たすとは限らない．また，これらの両立可能対集合は，反射的であり，$\langle a, a\rangle$および$\langle l, l\rangle$は，それぞれ，aおよびlと等しいとみなされる．また，二言語間の素性構造照合においては，$\langle a_i, a_j\rangle(\in C_A)$における$a_i$は，片方の言語のアトミックな素性値であり，$a_j$はもう一方の言語のアトミックな素性値である．$\langle l_i, l_j\rangle(\in C_L)$における素性$l_i$と$l_j$についても同様のことが言える．

☆☆ この場合，置換の前後で，全体の式に含まれる素性およびアトミックな素性値の数の変化に注目すると，ほとんどの等式において，この数が減少する方向に等式が適用される．また，この数が増加する方向に適用される等式についても，適用可能な回数は有限回に抑えられる．したがって，前に現れた式に変形しないという条件のもとで等式の適用を行えば，この変換の過程は必ず停止する．さらに，可能な照合結果の数は高々有限個である．
また，可能な照合結果をすべて求めるための計算量は組合せ的になる．ただし，両立可能対集合に含まれていないペアは照合不可能なので，実際には組合せの数は少なくなる．さらに，一つの動詞の格の数は高々数個なので，実際の問題において計算量が深刻になることはないと考えられる．

NIL	何も情報がないことを表わす．
TOP	矛盾する情報を表わす．
a	アトミックな素性値を表わす．ただし，$a \in A$
(a_i, a_j)	アトミックな素性値のペアを表わす．ただし，$(a_i, a_j) \in C_A$
$[p_1, \ldots, p_n]$	パス等式を表わす．ただし，$p_i \in C_L{}^*L^*$
$l:\phi$	lという素性名がϕという素性値を持つ構造を表わす．ただし，$l \in L$かつ$\phi \in$ FDLC
$\langle l_i, l_j\rangle:\phi$	$\langle l_i, l_j\rangle$という素性名がϕという素性値を持つ構造を表わす．ただし，かつ$\langle l_i, l_j\rangle \in C_L$かつ$\phi \in$ FDLC
$\phi \wedge \psi$	FDLCの連言を表わす．ただし$\phi, \psi \in$ FDLC

図1 両立可能対集合に基づく素性表現論理（FDLC）の式の定義
Fig. 1 The domain of formulas of feature description logic with sets of compatible pairs.

FDLCでは両立可能対集合に基づいて素性構造の照合を行うため，二つの素性構造の照合結果が一意に求まるという形にはならない．二つの素性構造の可能な照合結果を一つ求めるには，素性構造に対する等式を用いて二つの素性構造の連言の可能な変換結果を一つ求めればよい．二言語間で素性構造の照合を行う際には，あらゆる可能な変換結果を求め，後で述べるスコア関数（Scoring Function）を用いて最大重複照合（Most Overlapping Matching）を求める．

2.2 二言語間の素性構造照合

アトミックな素性値（Atomic Value）の両立可能対集合

アトミックな素性値の両立可能対集合C_Aを構成するために，訳語に関する知識を対訳辞書から求める．まず，対訳例の英文中の各単語（ここでは，自立語だけについて行う）W_{eng}について，その日本語の訳語を英和辞典から求める．次に，対訳例の日本文中の各単語（自立語）W_{jap}についても，その英語の訳語を和英辞典から求める．これらの訳語に関する知識を用いることによって，対訳例の片方の言語の文における単語が相手言語の文におけるどの単語に対応し得るかがわかるので，与えられた対訳例に関して，二言語間で両立可能な単語（自立語）のペア$\langle W_{Deng}, W_{Djap}\rangle$の集合を構成し，これを$C_{AD}$（$AD$ は Atomic value pairs which exist in the bilingual Dictionary の略）とする．さらに，与えられた対訳例に現れる英語の単語（自立語）W_{eng}および日本語の単語（自立語）W_{jap}のうちで，C_{AD}に現れないものについて，あらゆる可能なペア$\langle W_{NDeng}, W_{NDjap}\rangle$の集合を構成し，これを$C_{AND}$（$AND$ は Atomic value pairs which do Not exist in the bilingual Dictionary の略）とする．最

素性構造 f および g を照合するためには，
以下の等式によって $f \wedge g$ の可能な変換結果を一つ求める．

(1) $l:TOP = TOP$, ただし, $l \in L \cup C_L$
(2) $\phi \wedge TOP = TOP$
(3) $\phi \wedge NIL = \phi$
(4) $a \wedge b = \langle a,b \rangle$, ただし, $a,b \in A$, $\langle a,b \rangle \in C_A$
(5) $a \wedge b = TOP$, ただし, $a,b \in A \cup C_A$, $a \neq b$, $\langle a,b \rangle \notin C_A$
(6) $a \wedge l:\phi = TOP$, ただし, $a \in A \cup C_A$, $l \in L \cup C_L$
(7) $l:\phi \wedge l:\psi = l:(\phi \wedge \psi)$, ただし, $l \in L \cup C_L$
(8) $l_a:\phi \wedge l_b:\psi = \langle l_a,l_b \rangle:(\phi \wedge \psi)$, ただし, $\langle l_a,l_b \rangle \in C_L$
(9) $\phi \wedge \psi = \psi \wedge \phi$
(10) $(\phi \wedge \psi) \wedge \chi = \phi \wedge (\psi \wedge \chi)$
(11) $\phi \wedge \phi = \phi$
(12) $E_1 \wedge E_2 = E_2$, ただし, E_1, E_2 はパス等式, $E_1 \subseteq E_2$
(13) $E_1 \wedge E_2 = E_1 \wedge (E_2 \cup \{zy \mid z \in E_1\})$,
ただし, E_1, E_2 はパス等式, $\exists x : x \in E_1, xy \in E_2$
(14) $E \wedge x:c = E \wedge (\bigwedge_{y \in E} y:c)$, ただし, E はパス等式, $x \in E$
(15) $E \wedge x:c = E \wedge \{x\} \wedge x:c$, ただし, E はパス等式,
$x = \langle l_{a1},l_{b1} \rangle \cdots \langle l_{ai},l_{bi} \rangle l_{i+1} \cdots l_n, (1 \leq i \leq n)$,
$l_{a1} \cdots l_{ai} l_{i+1} \cdots l_n \in E$, または,
$l_{b1} \cdots l_{bi} l_{i+1} \cdots l_n \in E$
(16) $E \wedge E_1 \wedge E_2 = E \wedge (E_1 \cup E_2)$, ただし, E, E_1, E_2 はパス等式,
$\exists x : x \in E_1, \exists y : y \in E_2$,
$x = \langle l_{a1},l_{b1} \rangle \cdots \langle l_{ai},l_{bi} \rangle l_{i+1} \cdots l_n, (1 \leq i \leq n)$,
$y = \langle l'_{a1},l'_{b1} \rangle \cdots \langle l'_{aj},l'_{aj} \rangle l'_{j+1} \cdots l'_m, (1 \leq j \leq m)$,
$l_{a1} \cdots l_{ai} l_{i+1} \cdots l_n,\ l'_{a1} \cdots l'_{aj} l'_{j+1} \cdots l'_m \in E$, または,
$l_{b1} \cdots l_{bi} l_{i+1} \cdots l_n,\ l'_{b1} \cdots l'_{bj} l'_{j+1} \cdots l'_m \in E$
(17) $l:E = \{lw \mid w \in E\}$, ただし, $l \in L \cup C_L$
(18) $\{\epsilon\} = NIL$, ただし ϵ は空パス
(19) $E = TOP$, ただし, $x, xy \in E, y \neq \epsilon$

図 2　素性およびアトミックな素性値の両立可能対集合に基づく素性構造照合演算
Fig. 2　Matching operation of feature structures based on sets of compatible pairs of feature labels and atomic values.

後に，$C_{AD} \cup C_{AND}$ を C_A とする．

例2の場合には，C_{AD}, C_{AND} および C_A は次のようになる．この場合，対訳例に含まれるすべての自立語についてその訳語の情報が対訳辞書から得られるので，C_{AND} は ϕ（空集合）になる☆．

$C_{AD} = \{\langle write, 書く \rangle, \langle I, 私 \rangle, \langle letter, 手紙 \rangle, \langle pencil, 鉛筆 \rangle\}$,
$C_{AND} = \phi$, $C_A = C_{AD} \cup C_{AND}$

☆ ここで述べた C_{AND} の構成方法は，もっとも条件の強い方法である．もっとも条件の緩い構成方法としては，C_{AD} に現れた単語についても C_{AND} を構成するのを許すという方法が可能である．この方法によれば，あらゆる単語のペアについて C_{AND} を構成することになる．この際に，日本語と英語の間で品詞のチェックを行うなどして，制約を課すこともできる．これらの条件のうちどれを選ぶかは，対訳辞書から得られる情報がどのくらい信頼できるかによる．

素性名（Feature Label）の両立可能対集合

われわれの二言語間素性構造照合の枠組においては，素性名の両立可能対集合 C_L は統計的なデータに基づいて構成されることを仮定している．すなわち，C_L に含まれる各素性のペア（Feature Label Pair）$\langle l_i, l_j \rangle$ は，統計的なデータから計算される確率値 p_{ij} $(0 < p_{ij} \leq 1)$ を持つとする．この確率値 p_{ij} は，一方の言語のある素性構造における素性 l_i が表す意味的な属性（Semantic Role）が，もう一方の言語のある素性構造における素性 l_j が表す意味的な属性と両立する確率を表している．例えば，$\langle subj, が \rangle$ という英語-日本語の素性のペアが，英語-日本語の動詞のペア $\langle write, 書く \rangle$ に対してある確率値 $p_{subj,が}$ を持っているとすると，この $p_{subj,が}$ は，動詞のペア $\langle write, 書く \rangle$ に関して，英語の "*subj*" という素性と日本語の "が" という素性が意味的に同じ役割を果たす確率を表す．この素性のペア $\langle subj, が \rangle$ は，別の英語-日本語の動詞のペア $\langle read, 読む \rangle$ に対しては，別の確率値 $q_{subj,が}$ を持つと考えられる．

語彙的知識の獲得を始める前の段階においては，統計的なデータがないので，素性名のペアの持つ確率値を1に初期設定しておく．ただし，文法的な知識によって，意味的に同じ属性になり得ないことが明らかな素性のペアについては，その確率値を0とする．これらのペアは C_L には含まれないことになる．これによって，明らかに誤りである照合結果を除いて，すべての照合結果が得られることになる．語彙的知識獲得の観点から言えば，正しい照合結果が得られないと，正しい語彙的知識が獲得されないことになるので，現段階では，このような方法で素性名の確率値を初期設定することにする．

最大重複照合

照合の結果得られる素性構造 h の妥当性を計算するために，スコア関数 SCORE(h) を用いる．このスコア関数は実数の二つ組 $\langle x_1, x_2 \rangle$ $(x_1, x_2 \in R$（実数の集合）$)$ を返す☆．このスコア関数は，照合結果の素性構造に含まれる訳語のペアの数を調べるためのもので，x_1 は照合結果の素性構造に含まれる訳語のペア

☆ 現段階では，素性（Feature Label）の持つ確率値が1または0であるので，x_1 および x_2 は整数である．

のうち対訳辞書からの情報と一致するものの重み付きの数を表し，x_2 は照合結果の素性構造に含まれる訳語のペアのうち対訳辞書からの情報には含まれなかったものの重み付きの数を表す．正確に言えば，x_1 は照合結果の素性構造に含まれる訳語のペアのうちで C_{AD} の要素 $\langle W_{Deng}, W_{Djap}\rangle$ であるものの重み付きの数であり，x_2 は照合結果の素性構造に含まれる訳語のペアのうちで C_{AND} の要素 $\langle W_{NDeng}, W_{NDjap}\rangle$ であるものの重み付きの数である．

スコア関数によって得られるスコアの間に次のような全順序関係を定義する．

$\langle x_1, x_2\rangle > \langle y_1, y_2\rangle \overset{def}{\Leftrightarrow}$

$x_1 > y_1$ または $(x_1 = y_1$ かつ $x_2 > y_2)$

この全順序関係のもとでスコアが最大となるものを，最大重複照合 (Most Overlapping Matching) とする．スコア関数の詳細な定義は**図3**のようになる．

例

例2の対訳例について，二言語間の素性構造照合を行い，スコア関数によってスコアを計算すると次のようになる．

$score = \langle 4, 0\rangle$

$$\begin{bmatrix} pred : \langle write, 書く\rangle \\ tense : past \\ \langle subj, は\rangle : [pred : \langle I, 私\rangle] \\ \langle obj, を\rangle : \begin{bmatrix} pred : \langle letter, 手紙\rangle \\ spec : a \end{bmatrix} \\ \langle with, で\rangle : \begin{bmatrix} pred : \langle pencil, 鉛筆\rangle \\ spec : a \end{bmatrix} \end{bmatrix}$$

$score = \langle 3, 0\rangle$

$$\begin{bmatrix} pred : \langle write, 書く\rangle \\ tense : past \\ \langle subj, は\rangle : [pred : \langle I, 私\rangle] \\ \langle obj, を\rangle : \begin{bmatrix} pred : \langle letter, 手紙\rangle \\ spec : a \\ with : \begin{bmatrix} pred : pencil \\ spec : a \end{bmatrix} \end{bmatrix} \\ で : [pred : 鉛筆] \end{bmatrix}$$

一つ目の素性構造においては，前置詞句 "*with a pencil*" が動詞 "*wrote*" にかかっており，二つ目の素性構造よりも高いスコアになっている．これによって，動詞の訳語のペア $\langle write, 書く\rangle$ に関して，一つ目の素性構造が正しい格フレームの例として得られる．

関数 SCORE(h) は実数の2つ組 $\langle x_1, x_2\rangle$ $(x_1, x_2 \in R$(実数の集合)) を返す．
ここで h は，照合の結果得られる素性構造である．

1. If $h \in C_{AD}$, then return $\langle 1, 0\rangle$
2. Else if $h \in C_{AND}$, then return $\langle 0, 1\rangle$
3. Else if $h = l : a$ where $l \in L \cup C_L$ and $a \in A \cup C_A$ and SCORE(a) $= \langle x_1, x_2\rangle$, then return $\langle$SCORE$_L(l) \times x_1$, SCORE$_L(l) \times x_2\rangle$
4. Else if $h = h_1 \wedge h_2$ where $h_1, h_2 \in$ FDLC and SCORE(h_1)$= \langle x_{11}, x_{12}\rangle$ and SCORE(h_2)$= \langle x_{21}, x_{22}\rangle$, then return $\langle x_{11} + x_{21}, x_{12} + x_{22}\rangle$
5. Else return $\langle 0, 0\rangle$

end.

関数 SCORE$_L(l)$ は素性 l の確率値を返す．ただし，$l \in L \cup C_L$ である．

1. If $l \in L$, then return 1
2. If $l \in C_L$, then return the probability of l

図3 スコア関数
Fig. 3 Scoring function.

2.3 部分フレーズ主導型戦略による素性構造照合

例3の対訳例においては，一文内の二つの動詞句の主従関係が英日で逆になっている．このように，対訳例の文がお互いに全く異なった統語構造の場合には，二言語間の素性構造照合をトップダウンにそのまま適用することは無意味である☆．

例3

英語：He abandoned the medical profession for writing a novel.

日本語：彼は医業を辞めて小説を書き出した．

このような場合には，対応する部分フレーズ同士を照合し，それぞれの照合結果のスコアを足し合わせる必要がある．例3の場合には，"*abandoned*" の素性構造と "辞めて" の素性構造，"*writing*" の素性構造と "書き出した" の素性構造をそれぞれ別々に照合し，それぞれの照合の結果のスコアを足し合わせる．この部分フレーズ主導型戦略については，今後素性構造照合のプログラムに組み込む予定である．

3. 統語的曖昧性の解消

本章では，統語的曖昧性の解消の例について述べる．ここでは，説明の都合上，片方の言語の文だけが曖昧な例だけについて述べるが，両方の言語の文が曖昧な場合も全く同じようにして統語的曖昧性を解消することができる．

日本語における典型的な統語的曖昧性の一つとして，「名詞句＋助詞」が複数の述語に係り得る場合がある．例4の場合，"私，は" および "彼，が" が二つ

☆ 今後素性構造によって意味表現が記述できるようになると，統語構造が異なっていても意味構造が照合できる場合も起こり得る．

の動詞“書いた”および“読んだ”の両方に係り得るので，構文解析の結果構文木あるいは素性構造が三つ得られる．

例4

英語：I read a letter which he wrote.

日本語：私は彼が書いた手紙を読んだ．

英文のほうからは，一つの構文木あるいは素性構造が得られ，二言語間で素性構造の照合を行った結果，次のような最大重複照合が得られる．

$$\begin{bmatrix} pred : \langle read, 読む \rangle \\ tense : past \\ \langle subj, は \rangle : [pred : \langle I, 私 \rangle] \\ \langle obj, を \rangle : \begin{bmatrix} pred : \langle letter, 手紙 \rangle \\ spec : a \\ relc : \begin{bmatrix} pred : \langle write, 書く \rangle \\ tense : past \\ \langle subj, が \rangle : [pred : \langle he, 彼 \rangle] \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

この照合結果の素性構造においては，「名詞句＋助詞」の係り受けの曖昧性が解消されている．照合結果の素性構造には動詞の訳語のペアが二つ含まれているので，素性構造全体を次の単位素性構造に分割することができる．

$$\begin{bmatrix} pred : \langle read, 読む \rangle \\ tense : past \\ \langle subj, は \rangle : [pred : \langle I, 私 \rangle] \\ \langle obj, を \rangle : \begin{bmatrix} pred : \langle letter, 手紙 \rangle \\ spec : a \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} pred : \langle write, 書く \rangle \\ tense : past \\ \langle subj, が \rangle : [pred : \langle he, 彼 \rangle] \\ \langle obj, \phi \rangle : \begin{bmatrix} pred : \langle letter, 手紙 \rangle \\ spec : a \end{bmatrix} \end{bmatrix}$$

ここで，関係節の先行詞〈*letter*, 手紙〉に相当する素性構造が〈*obj*, ϕ〉という素性によってマークされていることがわかる．日本語文からだけでは，〈*letter*, 手紙〉がどのような格要素になるかわからないが，英文から得られる情報によってこれが目的格になることがわかる．動詞の格フレームを獲得する場合には，このような情報は非常に有用である．

4. 実験および評価

本章では，和英辞典から抽出した簡単な対訳例に対して行った統語的曖昧性解消の実験の結果について述べ，さらに，統語的曖昧性が解消しなかった例について考察する．

4.1 実　　験

まず，単言語での解析の方法および二言語間の素性構造照合で用いる対訳辞書について簡単に述べる．解析の際には，各単語の統語範疇が何であるかという統語的な知識だけを用いる．英語品詞辞書の見出し語数は約55,000，日本語形態素解析システム（JUMAN[10]）の形態素数は約36,000である．構文解析は，DCGルールによって記述された文法規則と構文解析システムSAX[9]によって行う．素性構造を構成する規則はDCGルールの補強項に書かれている．現在の文法規則数は，英語について約135，日本語について約85である．対訳辞書については，現在のところ，見出し語数約50,000の和英辞典（講談社学術文庫の和英辞典[13]）のみを用いている．

実験は，計算機上で利用可能な和英辞典[13]から抽出した189対訳例に対して行った．まず，対訳例の英文および日本文をそれぞれ単言語で解析して素性構造を求め，得られた素性構造を二言語間で照合する．そして，文全体についてスコアのもっとも高い照合結果から，文中の動詞の格構造が一意に決定した場合に，一意の解が得られたとする．われわれの当面の目標は動詞の格フレームを獲得することにあるので，ここでは，文中の動詞の格構造が決定すればよいという立場に立つ．

実験の結果を**表1**に示す．まず，189対訳例のうち，文法の不備により，英文の解析に失敗したものが

表1 実験結果
Table 1 Results of the experiment.

	英　語	日本語	対　訳
全体数	189	189	189
解析結果複数	133	103	—
解析結果一つ	12	42	—
解析失敗	36	11	44
相手言語文の解析失敗	8	33	
格構造が一意に決定	96	112	86
複数の解析結果から格構造が一意に決定	84	70	—
格構造獲得率	50.8% (96/189)	59.3% (112/189)	45.5% (86/189)
格構造獲得率（解析成功中）	66.2% (96/145)	77.2% (112/145)	59.3% (86/145)
曖昧性解消率	63.2% (84/133)	68.0% (70/103)	—
解析結果数の平均（全体）	17.1	4.4	—
解析結果数の平均（格構造が一意に決定したもの）	13.3	3.1	—

36個，日本文の解析に失敗したものが11個，英文または日本文のどちらかの解析に失敗したものが44個あった．英文および日本文の両方の解析に成功した対訳例について，英文および日本文の解析結果の数をそれぞれ求め，解析結果が複数のものの数と解析結果が一つであるものの数を表に記述した．さらに，二言語間の素性構造照合によって，文中の動詞の格構造が一意に決定したものの数を求めた結果，対訳例のうちの英文の格構造が一意に決定したものが96個，日本文の格構造が一意に決定したものが112個，英文および日本文の両方の格構造が一意に決定し照合結果がただ一つ得られたものが86個あった．これから，対訳例が与えられた時に，二言語間で照合された両者の格構造が一意に得られる率を計算すると，約45.5%となり，さらに対訳例の単言語文の解析が成功した場合に，二言語間で照合された両者の格構造が一意に得られる率を計算すると，約59.3%となった．また，動詞の格構造が一意に決定したもののうち，もともと複数の解析結果が得られていたものは，英文について84個，日本文について70個あった．これから，複数の解析結果から格構造が一意に決定する率（曖昧性解消率とする）を計算すると，英文については約63.2%となり，日本文については約68.0%となった．また，単言語での解析が成功したものおよび二言語間の素性構造照合によって格構造が一意に決定したものについて，解析結果数の平均を表に記述する．

この結果から，和英辞典の例文程度の対訳例については，現在の解析プログラムの精度でも，5割弱の割合で，二言語間で照合された格構造を一意に得ることができることが分かる．さらに解析プログラムの精度を上げると，この割合を6割程度にまで上げることが可能である．なお，英文の解析プログラムと日本文の解析プログラムを比べると，この実験の結果からは，現在のところ英文の解析プログラムの精度のほうが劣るようである．また，統語的な知識だけを用いて解析を行った場合，英文の曖昧性のほうがはるかに大きいことが分かる．

4.2 失敗例の分析

両言語の解析が成功した145対訳例のうち，二言語間の素性構造照合によって統語的曖昧性が解消しなかった59対訳例についてその原因を調べ，失敗の主原因によって分類した結果を**表2**に示す．

まず，英文と日本文で曖昧性が全く同じ形になっているために，われわれの手法で統語的曖昧性を解消するのが原理的に不可能であると考えられるものが10個あった．例えば，次の例5では，英文中の句“*on the blackboard*”と日本文中の句“黒板に”の係先の曖昧性が全く同じパターンになっている．

表2 統語的曖昧性解消の失敗原因の分析
Table 2 Analysis of the failures of syntactic disambiguations.

原　　因	頻度
日英の曖昧性が同じ形	10
日英の統語構造のずれ	18
対訳辞書の不備	15
単言語解析での不備	11
そ の 他	5
合　　計	59

例5

英語：Write the kanji we've studied on the blackboard.

日本語：黒板に習った漢字を書きなさい．

したがって，二言語間の素性構造照合を行うと，係先が日英両方で正しいものと，日英両方で間違っているものが同じスコアになるために，統語的な曖昧性が解消しない．このような例については，われわれの手法によって統語的曖昧を解消するのは，原理的に不可能である．

また，英語と日本語との間で統語構造上のずれがあるために，二言語間で素性構造がうまく照合せず，統語的曖昧性が解消しないものが18個あった．例えば，次の例6の対訳例においては，日本文の“埋め草に”という慣用句が英文では“*to fill up the space*”という不定詞句で訳されている．現在の解析プログラムによれば，この対訳例の英文からは，“*to fill up the space*”の部分の曖昧性から54個の解析結果が得られ，日本文からは1個の解析結果が得られるが，“埋め草に”の部分と“*to fill up the space*”の部分の統語構造上の不一致や日本文で主語が省略されていることから，解析結果の素性構造の照合がうまく行われず，統語的曖昧性は解消しなかった．

例6

英語：I wrote this article to fill up the space.

日本語：この記事を埋め草に書いた．

その他に，対訳辞書の不備のために辞書中に適切な訳語が見つからず，二言語間の素性構造照合がうまく行われなかったものが15個，単言語解析における不備のために正しい解析結果が含まれていないことが原

因であると考えられるものが 11 個あった.

これらの他に考えられる問題としては，対応する訳文の対がそれぞれ相手言語の文にはない格を持っている場合に，これらの格が無理に照合してしまうことが起こり得る．また，片方の言語に省略がある場合，通常は，相手言語の文の構造の中に照合しない部分が残って，全体の照合結果が一意に求まるが，対訳辞書から得られる訳語の情報が不完全な場合には，本来照合してはいけない部分と照合してしまうことも起こり得る.

本論文で述べた手法では，解析結果の素性構造は依存構造とほぼ等価なものであったが，解析結果として素性構造によって記述された意味表現が得られるようになると，本論文の手法をそのまま用いることによって意味表現上での二言語間の照合が可能になる．その結果，英語と日本語との間の統語構造上のずれのうちのいくつかは意味表現に吸収されることになり，統語的曖昧性の解消の精度も向上するものと思われる.

5. おわりに

計算機処理のための大規模な意味辞書を構築するためには，人間が読むための辞書や大規模なコーパスなどに含まれる自然言語の文章を解析して，意味辞書を半自動で構築する技術を確立することが重要である．しかし，自然言語の文章を解析する際には，統語的曖昧性および語彙の多義性という少なくとも二つの問題が生じる．本論文では，統語的な知識だけを用いて対訳例の各単言語文を構文解析した結果を素性構造で表現し，対訳辞書の訳語の情報をもとにこの素性構造を二言語間で照合することによって，単言語文の統語的な曖昧性を解消する手法について述べた．われわれの手法においては，照合結果の素性構造中の各表層語が二言語の訳語のペアによって表現されているので，単言語文における語彙の多義性を解消する際に非常に有用な情報となる.

現在われわれは，和英辞典から取り出した約40,000対訳例をもとに統語的曖昧性の解消したデータを蓄積し，その結果から動詞の格フレームを抽出する研究を行っている．われわれの手法によって得られたデータにおいては，素性構造中の各表層語が二言語の訳語のペアによって表現されている．したがって，単言語だけの場合と比べると，多義動詞の意味の分離を行ったり名詞と動詞との間の正しい格関係を決定するために必要となる情報がより多く含まれており，動詞の格フレームを抽出するのが容易になるものと思われる.

また，われわれの手法を応用することによって，機械翻訳のための二言語間の翻訳パターンを得ることができる．さらに，用例に基づく翻訳[11),12)]で用いられる構文解析済みの対訳例を蓄積するという目的に対しても，われわれの手法を利用することが可能であると考えられる.

謝辞 和英辞典のデータの使用を許可してくださった講談社編集部ならびに，和英辞典のデータを提供してくださった電子技術総合研究所の横山晶一氏および東京工業大学の田中穂積教授・徳永健伸氏に感謝いたします.

なお，本研究は，一部文部省重点領域研究「知識科学」の援助を受けている.

参考文献

1) Brent, M.: Automatic Acquisition of Subcategorization Frames from Untagged Text, *Proc. of the 29th Annual Meeting of the ACL*, pp. 209-214 (1991).
2) Dagan, I., Itai, A. and Schwall, U.: Two Languages Are More Informative than One, *Proc. of the 29th Annual Meeting of the ACL*, pp. 130-137 (1991).
3) Hindle, D.: Noun Classification from Predicate Argument Structures, *Proc. of the 28th Annual Meeting of the ACL*, pp. 268-275 (1990).
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) 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形容詞辞書 IPAL (Basic Adjectives) 説明書 (1990).
6) 日本電子化辞書研究所：概念辞書（第2版）TR-012 (1989).
7) Kasper, R. and Rounds, W.: A Logical Semantics for Feature Structures, *Proc. of the 24th Annual Meeting of the ACL*, pp. 257-266 (1986).
8) Lenat, D. et al.: *Building Large Knowledge-based Systems*, Addison-Wesley (1990).
9) 松本裕治，杉村領一：論理型言語に基づく構文解析システム SAX，コンピュータソフトウェア，Vol. 3, No. 4, pp. 308-315 (1986).
10) 松本裕治，黒橋禎夫，妙木 裕，長尾 眞：日本語形態素解析システム JUMAN 使用説明書，長尾研究室内部資料 (1991).
11) 佐藤理史：MBT 1：実例に基づく訳語選択，人工知能学会誌，Vol. 6, No. 4, pp. 592-600 (1991).

12) 佐藤理史：MBT 2：実例に基づく翻訳における複数翻訳例の組合せ利用，人工知能学会誌，Vol. 6, No. 6, pp. 861-871 (1991).
13) 清水 譲，成田成寿(編)：和英辞典，講談社学術文庫 (1979).
14) 徳永健伸，田中穂積：対訳辞書からの概念項目の自動抽出，人工知能学会誌，Vol. 6, No. 2, pp. 228-235 (1991).
15) 冨浦洋一，日高 達，吉田 将：語義文からの動詞間の上位-下位関係の抽出，情報処理学会論文誌，Vol. 32, No. 1, pp. 42-49 (1991).
16) 鶴丸弘昭，竹下克典，伊丹克企，柳川俊英，吉田将：国語辞典を用いたシソーラスの作成について，情報処理学会自然言語処理研究会，83-16 (1991)

(平成4年3月18日受付)
(平成4年9月10日採録)

宇津呂武仁（正会員）

1966年生．1989年京都大学工学部電気工学第二学科卒業．現在，同大学院博士後期課程在学中．自然言語処理の研究に従事．人工知能学会，日本ソフトウェア科学会，ACL 各会員．

松本 裕治（正会員）

昭和30年生．昭和52年京都大学工学部情報工学科卒業．昭和54年同大学院工学研究科修士課程情報工学専攻修了．同年電子技術総合研究所入所．昭和59～60年英国インペリアルカレッジ客員研究員．昭和60～62年(財)新世代コンピュータ技術開発機構に出向．昭和63年京都大学大型計算機センター助教授．平成元年京都大学工学部電気工学第二学科助教授となり，現在に至る．自然言語処理，論理プログラミング等に興味を持つ．

長尾 眞（正会員）

1960年京都大学工学部電子工学科卒業．1962年同大学院修士課程修了．京都大学工学部助手，助教授を経て，1973年より同教授，現在に至る．1976年より国立民族学博物館併任教授．パターン認識，画像処理，自然言語処理，機械翻訳等の研究に従事．