

TR-IT-0104

多元的類似度計算に基づく文脈依存のボトムアップパーサ MAC-BP

A Multi-dimensional, Analogy-based, Context-dependent,
Bottom-up Parser: MAC-BP

側嶋 康博

Yasuhiro SOBASHIMA

1995年3月

概要

本報告では、部分木用例との多元的な類似度計算のみでボトムアップに構文解析を行う手法およびプログラム (MAC-BP) の概要について述べる。本構文解析プログラムには、(1) 用例知識 (データ) に基づいて処理を行う「データ駆動式」であること、(2) アナログ数値で表わした、多元的な類似度尺度 (機能、意味、形態、および局所文脈) を用いて解析を行うこと、さらにそれらを (3) 統合して解選択の判断に利用する、という特徴がある。そのため、(1) インクリメンタルな性能向上、(2) 頑強で (3) 精度の高い構文解析が可能となる。

ATR 音声翻訳通信研究所

ATR Interpreting Telecommunications Research Laboratories

©(株) ATR 音声翻訳通信研究所 1995

©1995 by ATR Interpreting Telecommunications Research Laboratories

目次

1	はじめに	1
2	構文解析の概要	2
2.1	基本アイデア	2
2.2	言語モデル	3
2.3	類似度計算	5
2.4	類似度を用いたボトムアップ構文解析	6
3	実験時の設定と実験結果	8
3.1	用例データ	8
3.2	類似度の設定	9
3.3	新ノード属性	11
3.4	優先条件	11
3.5	類似度係数の決定	12
3.6	クローズテスト	13
3.7	オープンテスト	14
4	結果の考察とまとめ	15
	参考文献	16
	付録	
A	属性分類とデータ仕様	17
A.1	属性分類	17
A.2	用例ソースファイルの仕様	33
A.3	部分木用例データの仕様	35
A.4	パーサ入力データの仕様	35
B	処理の手順とパーサの操作方法	36
B.1	処理の概要	37
B.2	用例ソースファイルの作成	37
B.3	部分木用例データの抽出	37
B.4	部分木用例データベースの作成	37
B.5	構文解析の実行	37
B.6	構文解析の実行結果	38
B.7	構文木表示ツール	39

第1章

はじめに

自然言語の構文解析は、自然言語処理の中で最も基本的かつ重要な技術であるが、自然言語が持つ自由度(語彙的・構造的曖昧性、語句の省略、非文法的表現など)のため、高い精度で頑強に構文解析を行うことは依然困難である。従来のルールに基づく手法では、記述を詳細化して精度を上げようとするとう頑強さを失い、逆に記述を粗くすると多くの解釈から正解の選択が困難になるため、数多くのヒューリスティック・ルールを駆使して副作用に対処しながら複雑な制御を行わなければならなかった。

近年、用例との類似性に着目したアプローチが提唱^[1]され、主に機械翻訳の変換^{[2],[3],[4]}に応用されて成果を収めている。このアプローチでは、語のシソーラス階層関係から得られた、用例との意味的な類似度または距離^[3]を基に、尤もらしい変換事例を解として採用する。この手法では、アナログ的な数値を判断に使用して頑強さを向上させ、また具体的な用例を多く用意することにより精度を向上させる。すなわち、高い精度と頑強さを共に備えた手法として期待される。しかしながら、これまでの用例に基づくアプローチでは、類似度は「語の意味的類似度」に、適用対象はルールで曖昧性解消できない「変換パターンの選択」に限定されており、類似度尺度だけで自然言語を構文解析することは行われていなかった(注1)。

本報告では、部分木用例データとの多元的な類似度計算により、類似度尺度のみでボトムアップに自然言語を構文解析する手法の概要と、この手法をインプリメントした構文解析プログラム(MAC-BP)の内容について解説を行う。本構文解析プログラムの特徴は、(1)用例知識(データ)に基づいて処理を行う「データ駆動式」であること、(2)アナログ数値で表わした、多元的な類似度尺度(品詞、活用、意味、形態、および局所文脈)を用いて解析を行うこと、さらにそれらを(3)統合して解選択の判断に利用するということである。具体的には、局所文脈を含む多元的な類似度の導入により、意味的に近い用例がない場合でも機能的(構文的)に近い用例を模倣し、また機能・意味的に類似する用例が複数ある場合は局所文脈の類似度を用いて文脈を考慮した解の選択を行うことができる。また、種々の解析(機能・意味・形態および局所文脈)を統合して実行するため、処理効率が高い。例えば、ボトムアップな解析の各段階でトップダウン制約(局所文脈の制約)がかかるため、全解探索やバックトラックなしに効率的に構文解析する。さらに、実験を通して、用例追加による副作用が極めて小さいことが示されている。

注1 TDMT^[4]では、解釈可能な構造はすべて求めた上で用例と意味的に最も近い解を選択する。

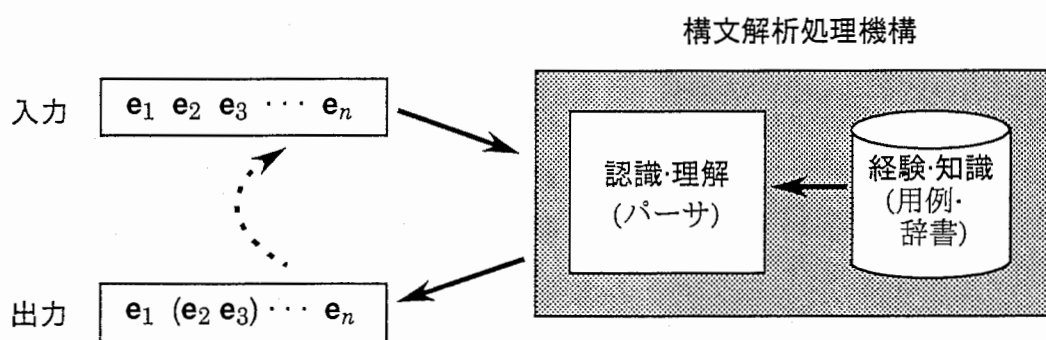
第2章

構文解析の概要

2.1 基本アイデア

本研究における基本的立場である「用例に基づくアプローチ」の利点の1つに、人間の経験的判断に近い処理を行うことがあげられる。構文解析は、人間が自然言語を認識、理解する過程に対応させることができるが、実際、人間が自然言語を認識・理解する場合、図1のように、言語的、あるいは非言語的な経験や知識を利用しながら、基本的にはボトムアップに意味的なまとまりを作りつつ表現全体を把握していると推察される。

この過程をシミュレートする構文解析では、語彙(辞書)と句の構成(部分木用例)に関する知識が必要である。パーサは、それら知識をアクセスしながら類似度計算を行い、尤もらしい意味的なまとまりを判断してノード作成を行う。パーサの処理は、構文解析の終了、つまり全体が1つのノードにまとまるまで繰り返す。ただし、曖昧性解消の必要(注2)から、部分木用例に隣接語句の情報、すなわち局所文脈情報も加えておくこととする。すなわち、局所文脈情報を含む個々の部分木用例との類似度計算をもとに行うノード作成を再帰的に実行することにより文全体の構造を作り上げることとする。



<凡例> () は意味的なまとまり、 e_i は i 番目の要素(語)を表す

図1 用例ベースの構文解析処理

注2 例えば「... make a room reservation」という入力に対して「a+room→a room」「room+reservation→room reservation」という2つの部分木用例の類似度だけでは選択できないが、左右語句(左端・右端マーカを含む)情報の追加により曖昧性が解消できる。

2.2 言語モデル

2.2.1 解析の単位

本研究の対象は、話し言葉の発話であり、翻訳を応用例に考えている。そのため日英表現の対応の便宜^[5]も考慮して、「文」すなわち句点で分けられる単位に限定しない解析単位を設定する。すなわち、応答表現(「はい」「yes/okay」など)から末尾の働きかけ表現(「~さん」「right?/please」など)までを構文解析の単位(メッセージ)としている。トップノードであるメッセージは、図2に示すように、プレ、メイン(必須)、ポストの各メッセージ要素と、それらを接続する要素(ジョイント)からなる。

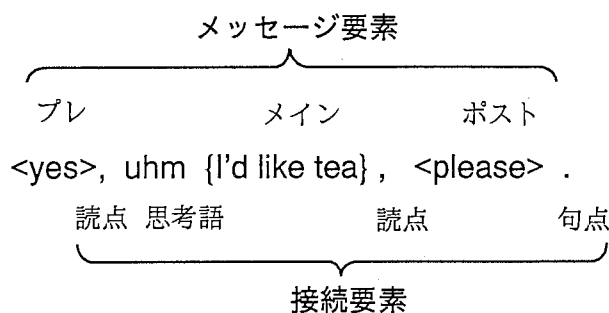


図2 メッセージの構成例

2.2.2 発話の言語的構成

プレ、メイン、ポストの各メッセージ要素は、語または複数の語で構成される句からなる。以下では語を要素、句を構造と呼ぶ。要素は「自立語的要素」と、他の自立語的要素と結合して構造を作る「付属語的要素」とに分類される。付属語的要素は、日本語では助詞や助動詞が対応し、英語では、冠詞、前置詞、助動詞などが対応する。したがって、本言語モデルにおいて、発話の言語的構成は図3のようになる。

本モデルでは、統一した類似度計算の都合上、自立語的構造だけを構造と認める。例えば、日本語の後置詞句、英語の前置詞句は構造と扱わない(注3)。

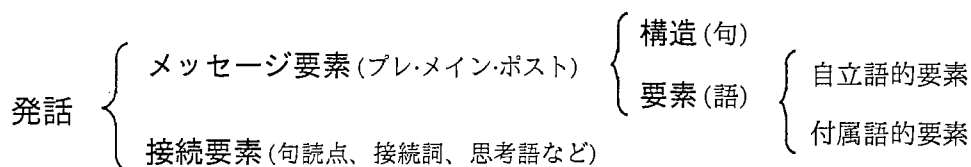


図3 発話の言語的構成

注3 断片的な発話で、本来の係り先が省略された前置詞句や後置詞句(「just for myself」「お名前は」など)は構造として扱う。

2.2.3 役割を用いた構造の記述

ある構造とその構成要素との関係を役割と呼び、ヘッド、サブ、フラグメントの3通りを設定する。ヘッドは構造の意味的中心、サブはその補語または付加語、フラグメントは構造を構成するために必須な付属語的要素である。また、ヘッドとサブは自立語的要素または構造である。さらにヘッドは、その構造に活用の制約を受けないストロング(S)ヘッドと、制約を受けるウィーク(W)ヘッドの2通りに分ける。図4に英文の構造と役割の例を示す。

この例では、名詞句を構成するヘッド(the sunのsunおよびthe eastのeast)は、構造から活用の制約を受けている(Wヘッドである)と解釈し、「the <sun>」「the <east>」のように表記する。ところが、「rises in the east」におけるヘッド「rises」はこの構造から活用の制約を受けていない(「rises」の3人称単数現在形の活用は時制と主語の数との関係で決まる)。したがって、この構造ではrisesはSヘッドであり「[rises] in (the <east>)」のように表記する。なお、この文全体では、「(the <sun>) <[rises] in (the <east>)>」のようになり、主語と述部との関係はサブとWヘッドの関係となる。この節以降では、語の表記を利用する場合、標準形(「rises」は「rise」)を用いることにする。

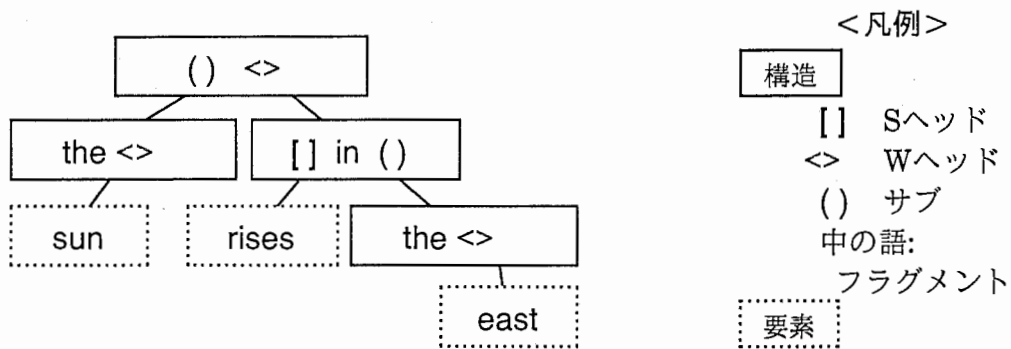


図4 役割を用いたメイン・メッセージ要素
「the sun rises in the east」の構造記述例

2.3 類似度計算

2.3.1 要素間類似度

ある言語的要素 **A** は、いくつかの属性に分解できるものとする。例えば、機能(品詞・活用などの構文属性)、意味、形態の3つの属性 $A_{fun}, A_{sem}, A_{gra}$ のセットとして eq. 1 のように表す。また、**A** と他の要素 **B** ($B_{fun}, B_{sem}, B_{gra}$) との要素間類似度 E-Sim を各属性ごとの類似度を用いて eq. 2 で定義する(注4)。

$$\mathbf{A}(A_{fun}, A_{sem}, A_{gra}) \quad (\text{eq. 1})$$

$$\text{E-Sim}(\mathbf{A}, \mathbf{B}) = \phi(\sigma_{fun}, \sigma_{sem}, \sigma_{gra}) \quad (\text{eq. 2})$$

ここで、 ϕ は要素間類似度を与える関数であり、 $\sigma_{fun}, \sigma_{sem}, \sigma_{gra}$ はそれぞれ、**A, B** 間の機能、意味、形態の類似度である。

2.3.2 構造間類似度

複数の要素 $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_n$ からなる列 **A** が構造を作る場合、同数の要素からなる列 **B** との間の構造間類似度を eq. 3 で定義する。

$$\text{S-Sim}(\mathbf{A}, \mathbf{B}) = (1/W) \sum_i \text{E-Sim}(\mathbf{A}_i, \mathbf{B}_i) w_i \quad (\text{eq. 3})$$

ここで、 i は要素位置を表す添字であり w_i は役割に応じた重みである。また、正規化係数 $(1/W)$ と w_i との間には $W = \sum_i w_i$ の関係がある。

2.3.3 文脈的類似度

図5に示すように、言語的表現 **A, B** が左右にそれぞれ a_1, a_2 と b_1, b_2 の語句を伴っている(注5)とする。

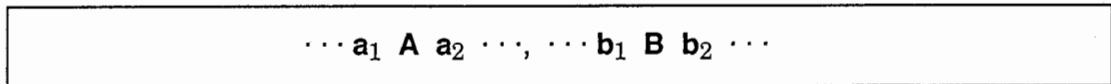


図5 2つの言語的表現 **A, B** の文脈的關係

この場合の **A, B** 間の文脈的類似度(局所文脈の類似度) C-Sim を、文脈的類似度を与える関数 ϕ_c を用いて eq. 4 で定義する。

$$\text{C-Sim}(\mathbf{A}, \mathbf{B}) = \phi_c(\text{E-Sim}(a_1, b_1), \text{E-Sim}(a_2, b_2)) \quad (\text{eq. 4})$$

2.3.4 統合した類似度

構文解析で解選択の尤度として使用する、構造 **A, B** 間の統合した類似度 Sim を、構造間類似度と文脈的類似度の関数 Φ である (eq. 5) と定義する。

$$\text{Sim}(\mathbf{A}, \mathbf{B}) = \Phi(\text{S-Sim}(\mathbf{A}, \mathbf{B}), \text{C-Sim}(\mathbf{A}, \mathbf{B})) \quad (\text{eq. 5})$$

注4 付属語的要素間の要素間類似度は、一致した場合1、不一致の場合0と定義する。

注5 **A, B** が左端にある場合 a_1, b_1 は左端マーカ、右端にある場合 a_2, b_2 は右端マーカを使用する。

2.4 類似度を用いたボトムアップ構文解析

この節では、本章第1節の基本アイデアを実現する構文解析機構について解説する。構文解析の入力は、形態素解析の結果(単語のセグメントに分けられた要素の列)であり、それらの組と、左右の隣接する語句の文脈情報を含む部分木用例との間で類似度計算を行うことによりノード作成の判定を行い、順次構造を作り上げていく。

2.4.1 入力

構文解析の入力は、属性セットからなる要素の列とする。すなわち、単語に分割された要素の列で、個々の要素は機能、意味、形態の各属性値が定まっている。ただし、おのおの唯一である必要はない。図6は、 n 個の要素の列を表す。

$$e_1(e_{1fun}, e_{1sem}, e_{1gra}) \quad e_2(e_{2fun}, e_{2sem}, e_{2gra}) \quad \cdots \quad e_n(e_{3fun}, e_{3sem}, e_{3gra})$$

図6 入力する要素列

2.4.2 部分木用例データ

入力との類似度計算を行うための部分木用例データを用意する。部分木用例データは、構造を作る要素の数、構造全体の属性セット、各要素およびパース時の左右文脈要素の属性から構成される。図7は、「yes, uhm I'd like tea, please.」というメッセージから3つの部分木用例データが抽出される例を示す。

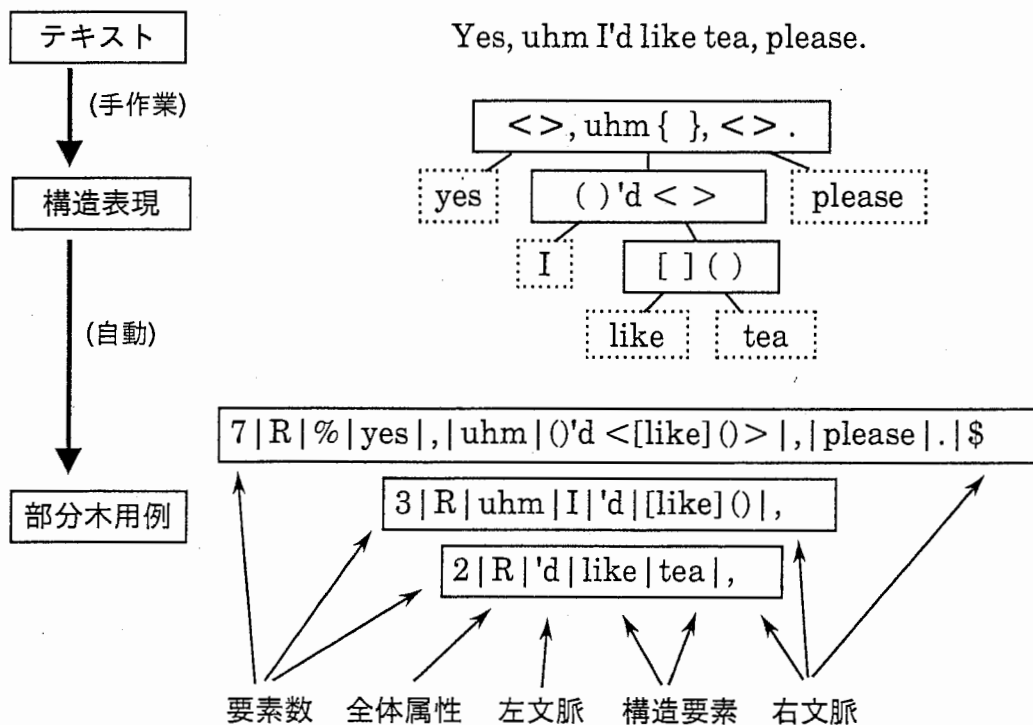


図4 部分木用例データの抽出

2.4.3 類似度計算によるノードの作成

構文解析では、連続する入力要素列の組合せを変えながら、部分木用例データベースから検索した個々の用例との間で類似度計算を行い、統合した類似度が最大となる要素列を選び用例の属性を模倣(3.3を参照)して新ノードを作成する(図8参照)。同点の他の要素列については優先条件(3.4を参照)を適用し、また新ノード属性候補が複数ある場合は、保存する。ノードを作成しなかった要素の組合せについての類似度得点は、後の再計算を避けるため保存しておく。この処理をノード数(要素数)が1になるまで再帰的に繰り返すことにより、メッセージ全体の構文解析を行う。

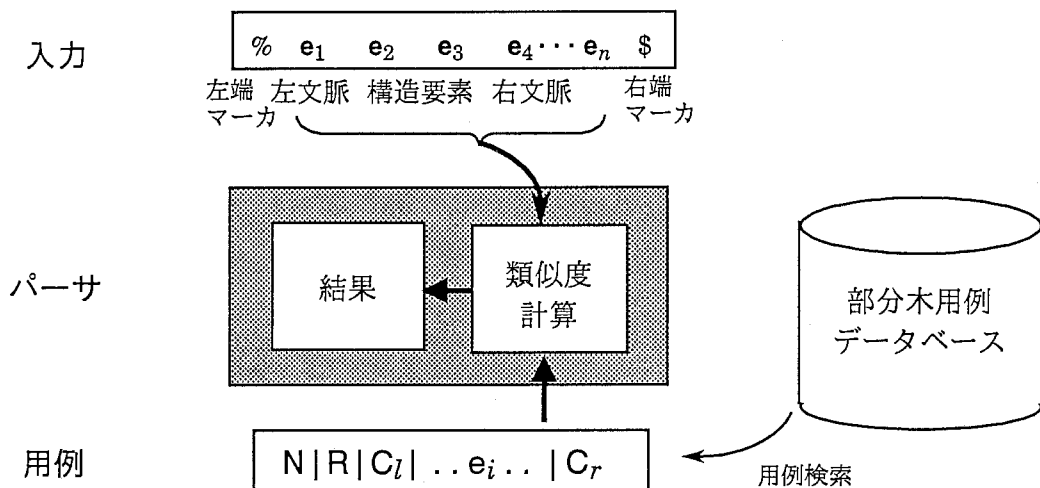


図8 文脈付き用例を用いた構文解析

2.4.4 曖昧性の解消

1要素に複数属性セットがある場合、最大の統合した類似度を与える属性セットを選択することにより、曖昧性を解消する。すなわち、構造的曖昧性を解消しながら語彙的曖昧性の解消を行う。図9は、「三日」という表現の意味属性が係り先(ヘッド)により選択される例である。すなわち、「三日」という表現が「サブ」という役割で「Sヘッド」の動詞と共に構造を作る場合、用例との類似度計算により、「滞在する」のような動詞では「期間」の意味属性が選択され、「到着する」のような動詞では「時点」の意味属性が選択されることになる。

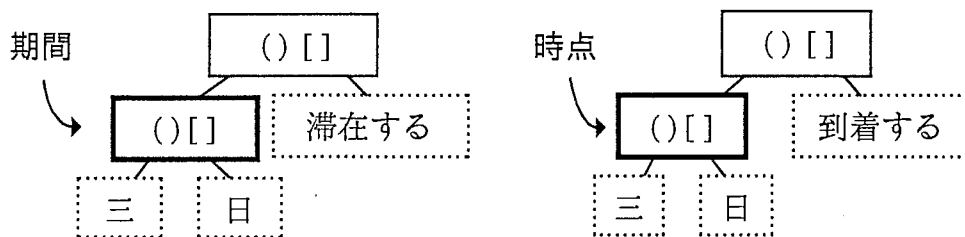


図9 語彙的曖昧性の解消例

第3章

実験時の設定と実験結果

3.1 用例データ

多元的類似度計算に基づくボトムアップ構文解析の実験を行うため、表1に示すように、(1) 動詞型別基本文[6]、(2) 会話基本文[7],[8]、(3) ATR 旅行模擬会話[9]の用例データを用意した。これらすべてのテキストは、タグを付与して構文木データとし、各構文木データから部分木データを抽出した。抽出された部分木の総数は、日本語 8887、英語 7591 であった。

表1 実験に用いた用例の種類と量 (M: メッセージ)

種類 (言語)	M数	文数	語数	語数/M
動詞型別 (日)	259	259	2639	10.2
基本文 (英)	259	259	1930	7.5
会話 (日)	835	852	6211	7.4
基本文 (英)	835	853	5527	6.6
ATR旅行 (日)	545	666	6997	12.8
模擬会話 (英)	545	660	6599	12.1
合計* (日)	1639	1777	15839	9.7
(英)	1639	1772	14056	8.6

3.2 類似度の設定

3.2.1 機能属性と機能類似度 (付録 A.1 参照)

機能属性は、大中小3階層に分類した品詞と、各品詞大分類ごとに共通な活用とを定めた。表2に品詞大分類、表3に英語の名詞類(注6)の分類を示す。品詞分類の総数は、日本語 86、英語 90 である。

表2 品詞大分類

日本語品詞	英語品詞
N 名詞類	N 名詞類
Q 数詞類	Q 数詞類
A 形容詞類	A 形容詞類
V 動詞類	V 動詞類 B be 動詞句 X 助動詞句
D 副詞類	D 副詞類
S 文	S 文
R 間投表現	R 間投表現

表3 英語名詞類細分

N10 名詞
N11 可算名詞
N12 不可算名詞
N20 名詞句
N30 代名詞
N31 主格代名詞
N32 目的格 "
N33 再帰目的格 "
N34 指示 "
N35 否定 "
N40 固有名詞
N50 単位
N51 数量単位
N52 位置単位

品詞分類、活用それぞれに3桁、1桁のコードを与え、それらを使用して、品詞類似度 σ_{pos} と活用類似度 σ_{inf} を eq. 6, eq. 7 により求める。

$$\sigma_{pos} = n/3 \quad (n = 0, 1, 2, 3) \quad (\text{eq. 6})$$

$$\sigma_{inf} = 0, 1 \quad (\text{eq. 7})$$

ここで、 n は品詞コードが上位から一致した桁数であり、 σ_{inf} はコードが一致した場合 1、不一致の場合 0 を与える。

また、機能類似度 σ_{fun} は、eq. 8 に示すように、品詞類似度 σ_{pos} と活用類似度 σ_{inf} の関数とする。

$$\sigma_{fun} = \Phi_{fun}(\sigma_{pos}, \sigma_{inf}) \quad (\text{eq. 8})$$

3.2.2 意味属性と意味類似度 (付録 A.1 参照)

意味属性の分類は、日英両言語で共通とし、語句レベルの意味分類とメッセージレベルの発話行為タイプの分類[5]の2通り、それぞれ大中小3階層の分類(3桁コード)とした(表4を参照)。細分化された意味属性の分類総数は323である。また、意味類似度 σ_{sem} についても、意味分類コードの上位からの一致桁数 n を用いて eq. 9 で与える。

$$\sigma_{sem} = n/3 \quad (n = 0, 1, 2, 3) \quad (\text{eq. 9})$$

注6 代名詞の格変化は品詞細分として扱う。

注7 例えば、英語の可算名詞 room には N11 という品詞を与え、単数には 1、複数には 2 という活用を与える。「余地」の意味の room は品詞 N12、活用 1 を与える。

表4 意味属性の分類

語句	発話行為タイプ(注8)
1 人	- 反応
2 物	-1 反応的応答(驚き・確認・理解)
3 抽象	-2 陳述的応答(判断・情報・約束)
4 時	0 思考
5 所	01 独り言(理解・判断・情報)
6 量	02 思考語(内容思考・表現思考・語思考)
7 様態	+ 作用
8 出来事	+1 情報提供(事象・心情・判断)
9 行為	+2 要求(判断・情報・行為)

3.2.3 形態属性と形態類似度

形態属性は、対象が要素の場合、見出し語表記を採用する。構造の場合は、語句構成の利用も考え、フラグメントの文字列と共にSヘッド、Wヘッド、およびサブの各要素の記号を加えて表す。ただし、抽象度を高めるため、サブ要素の表記は除外する。図10に形態属性の例(注9)を示す。

(日本文例)	"部屋の予約をしたいのです"
	形態属性 "<<()を[し]>たい>のです"
(英文例)	"I would like to make a room reservation"
	形態属性 "() would <[like] to ()>"

図10 文の形態属性例

形態類似度は、形態属性の文字列、形態属性中のヘッド文字列および基底パターン(ヘッド・サブ記号およびフラグメントの文字列(注10))の一致を用いて、eq. 10により与える。

$$\sigma_{\text{gra}} = \begin{cases} 1 & (\text{形態属性が完全に一致}) \\ 0.9 & (\text{ヘッド・基底パターンが一致}) \\ 0.6 & (\text{ヘッドのみ一致}) \\ 0.3 & (\text{基底パターンのみ一致}) \\ 0 & (\text{上記以外}) \end{cases} \quad (\text{eq. 10})$$

注8 発話行為タイプの分類の括弧内は小分類を示す。

注9 日本語の動詞は、名詞派生語との関連を考慮し連用形(ここでは「し」)を見出しに使用している。

注10 図10の例では「<>のです」「() would <>」が基底パターンとなる。

3.2.4 類似度関数・係数および重み

eq. 2 の要素間類似度は、各属性の類似度ごとの関数(類似度関数)の積(注11)で表わし(eq. 11)、各類似度関数 ϕ_x は、それぞれの類似度 σ_x の一次関数とする(eq. 12)。ここで、 μ_x (類似度係数, $x = \text{pos, inf, sem, gra}$)は要素間類似度関数 ϕ における各属性ごとの類似度 σ_x の重みを表し(注12)、実験的に求める(6.5を参照)。

$$\phi(\sigma_{\text{fun}}, \sigma_{\text{sem}}, \sigma_{\text{gra}}) = \prod_x \phi_x(\sigma_x) \quad (x = \text{pos, inf, sem, gra}) \quad (\text{eq. 11})$$

$$\phi_x(\sigma_x) = \mu_x \sigma_x - \mu_x + 1 \quad (\text{同上}) \quad (\text{eq. 12})$$

eq. 3 の構造間類似度 S-Sim の計算に必要な役割ごとの重みを eq. 13 で与える。

$$w = \begin{cases} 10 & (\text{ヘッド}) \\ 5 & (\text{メイン・メッセージ要素}) \\ 1 & (\text{サブ、プレ・ポスト各メッセージ要素}) \end{cases} \quad (\text{eq. 13})$$

eq. 4 の文脈的類似度(C-Sim または σ_{con})は、左右要素の類似度の積で与える(eq. 14)。ここで σ^l, σ^r はそれぞれ左・右要素の要素間類似度である。また、統合した類似度関数 Φ は、文脈類似度関数 ϕ_{con} および係数 μ_{con} を用いて eq. 15, eq. 16 で表す。

$$\phi_c = \sigma_{\text{con}} = \phi(\sigma^l) \phi(\sigma^r) \quad (\text{eq. 14})$$

$$\Phi = \text{S-Sim} * \phi_{\text{con}}(\sigma_{\text{con}}) \quad (\text{eq. 15})$$

$$\phi_{\text{con}}(\sigma_{\text{con}}) = \mu_{\text{con}} \sigma_{\text{con}} - \mu_{\text{con}} + 1 \quad (\text{eq. 16})$$

3.3 新ノード属性

ある要素列(ヘッド H_{in})に与える新ノード(N)の品詞・活用・意味属性は、採用された部分木用例データの構造ノード(E)とそのヘッド(H_{eg})の属性を使って、eq. 17, eq. 18 のように模倣する。また、形態属性は、6.2.3 の手順で作成する。

$$N_{\text{pos}} = E_{\text{pos}} \quad (\text{eq. 17})$$

$x = \text{inf, sem}$ の場合:

$$N_x = \begin{cases} H_{\text{in}_x} & (H_{\text{eg}_x} = E_x) \\ E_x & (H_{\text{eg}_x} \neq E_x) \end{cases} \quad (\text{eq. 18})$$

3.4 優先条件

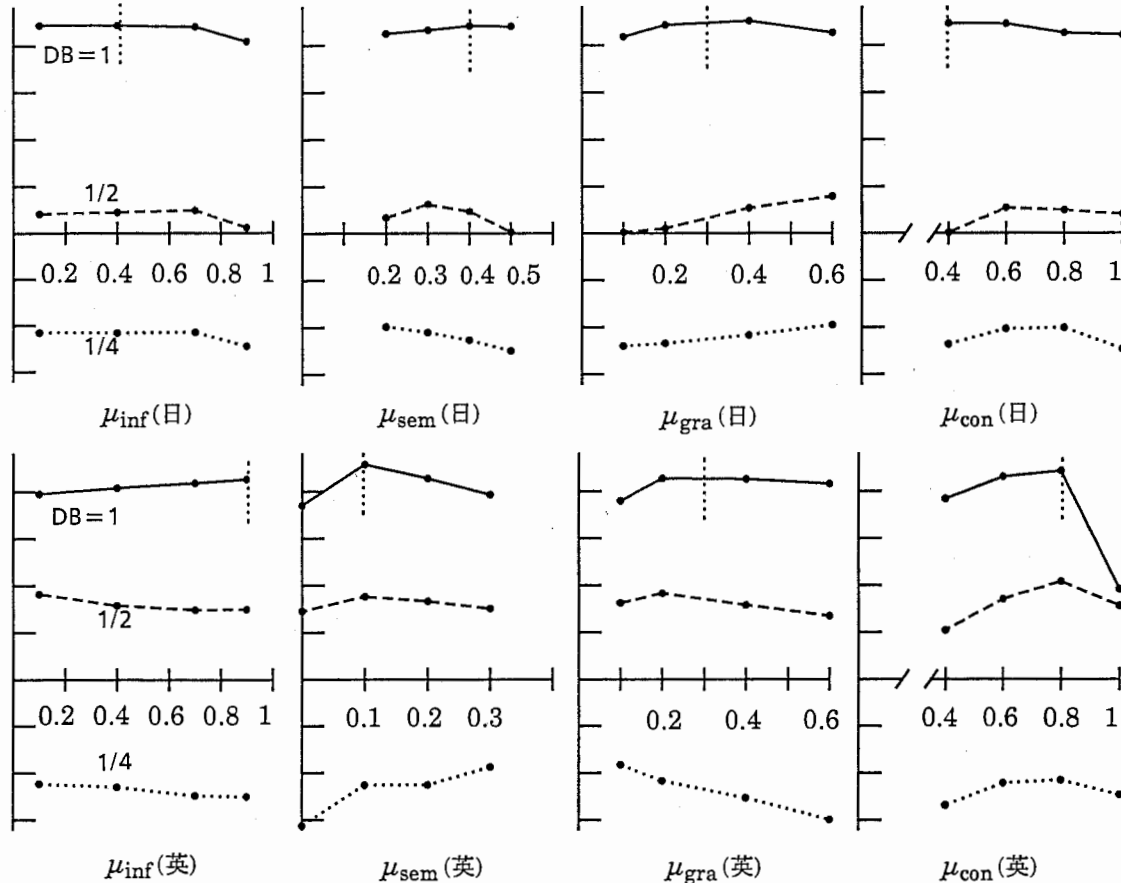
新ノードとなる要素列の決定は、(1)統合した類似度の高いもの、もし複数の同点があれば(2)要素数が多いもの、もし同数であれば(3)日本語は左端寄り、英語は右端寄り、の順に優先する。

注11 eq. 8 の機能類似度 ϕ_{fun} は、品詞分類の類似度 $\phi_{\text{pos}}(\sigma_{\text{pos}})$ と活用の類似度 $\phi_{\text{inf}}(\sigma_{\text{inf}})$ の積となる。

注12 活用の制約を受けないSヘッドについては、活用の一致は無視する($\mu_{\text{inf}} = 0$)。

3.5 類似度係数の決定

この実験では、用例からランダムに選んだ日英被験文各80メッセージについて、品詞の類似度係数($\mu_{pos}=1$)以外の係数値を組合せ(注14)被験文の部分木データは学習させず、構文解析実験を行った。図11に、各係数ごとの成功得点(注13)の合計のグラフを示す。この結果から、表5の係数値を採用する。



<凡例> 上段: 日本語, 下段: 英語 (DBサイズは全体を1)
縦軸: 得点合計 (目盛: 300点), 横軸: 各係数値

図11 類似度係数決定実験結果

表5 採用した類似度係数値

言語	μ_{pos}	μ_{inf}	μ_{sem}	μ_{gra}	μ_{con}
日本語	1	0.4	0.4	0.3	0.4
英語	1	0.9	0.1	0.3	0.8

注13 構造と属性の正解に2点、構造のみ正解に1点、誤った構造に-1点、解析失敗を-2点とした。

注14 係数 ($4 \times 4 \times 4 \times 4 = 256$), 被験文 (80), 言語 (2), 学習データサイズ (3) を組合せ、計 122,880 回行った。

3.6 クローズテスト

準備した用例からランダムに被験文100メッセージを選び、このデータを含めたまま学習データの量を半減させて構文解析を行った。図12はこの実験を3回行って得た結果の平均を表すグラフである。また、図13に、同じ実験で文脈類似度を無視した場合 ($\mu_{con}=0$ と設定) の結果を示す。この結果からわかるように、文脈類似度を利用することにより、用例追加による副作用を抑え、96%以上の精度を保持している。

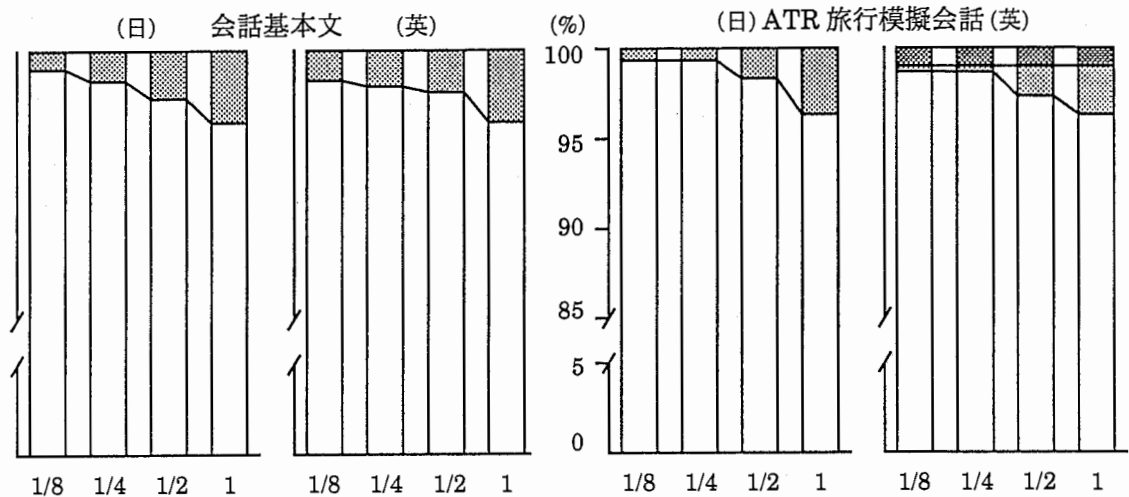
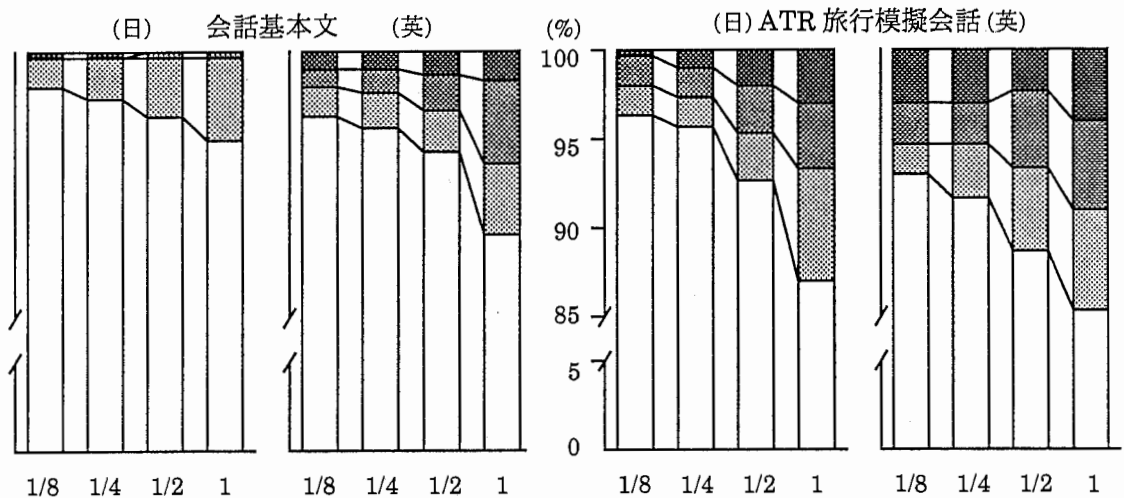


図12 クローズテスト結果



<凡例> 縦軸: 結果の割合(%) 横軸: 学習データ量(全体を1)
 □ 構造・属性正解 ▨ 構造正解 ▩ 誤り ■ 失敗

図13 クローズテスト ($\mu_{con}=0$) 結果

3.7 オープンテスト

クローズテストと同様の手順(100メッセージ3セット)でオープンテストを行った。ただし、被験文の部分木データは学習データに含めていない。図14にその結果を示す。

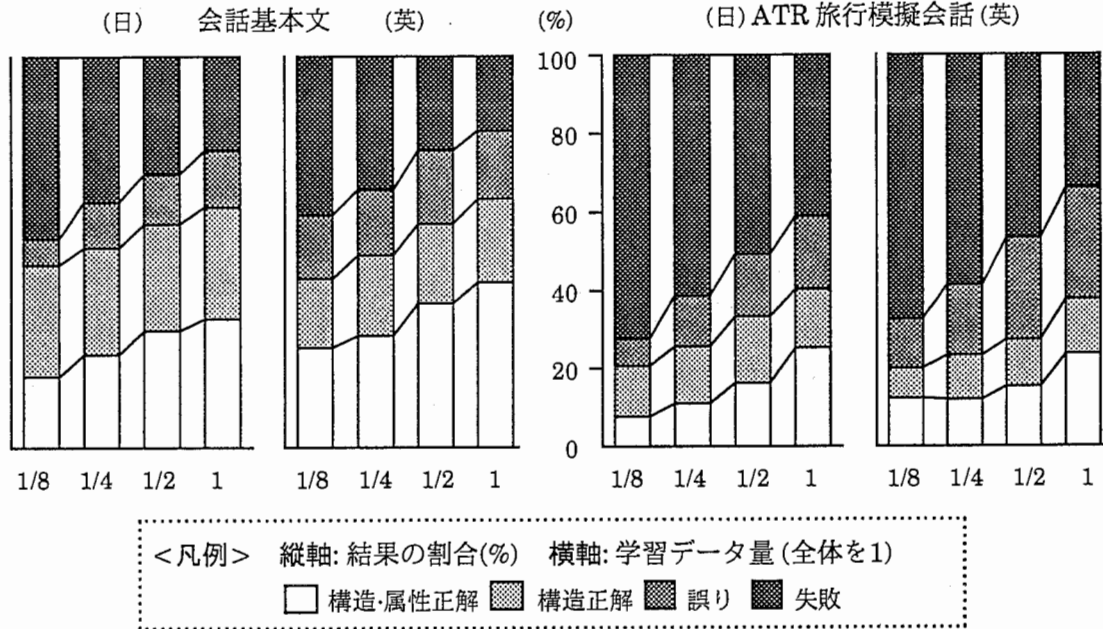


図14 オープンテスト結果

図15は、パース対象メッセージの構成部分木のうち、用例との類似度(用例中の全部分木と比較した場合の最大値)が最低となる部分木の類似度(クリティカル類似度と呼ぶ)とパース結果を表すグラフである(横軸の斜体数値はその値未満を表す)。このグラフから、クリティカル類似度に比例してパース精度が向上しており、類似度設定の妥当性を示している。

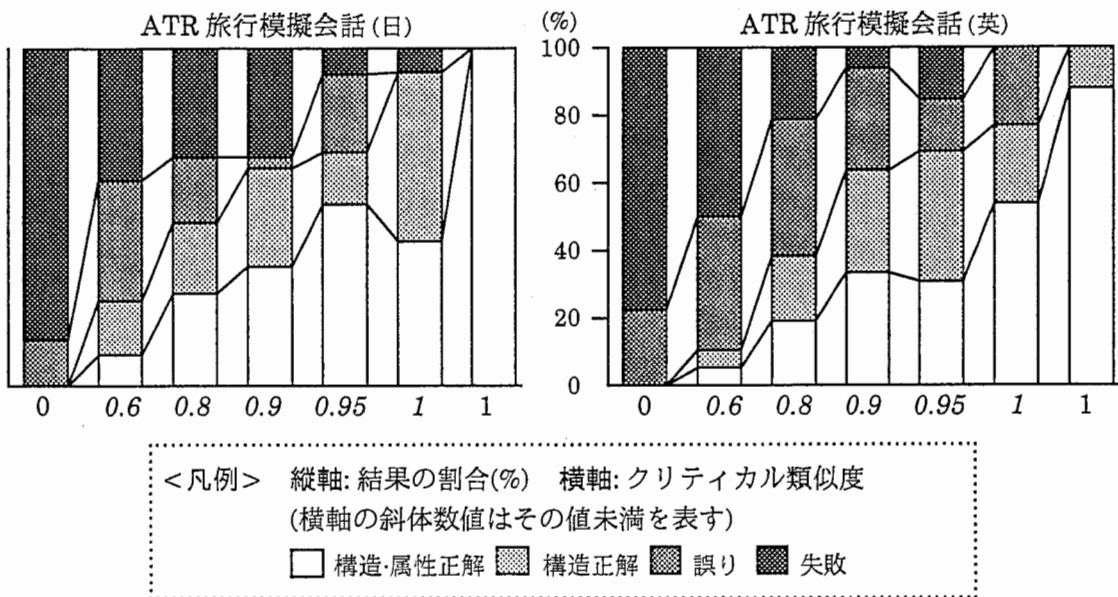


図15 クリティカル類似度とオープンテスト結果

第4章

結果の考察とまとめ

今回の実験では、メッセージ隣接用例を用いた属性の曖昧性解消は行っていない(注15)ため、模範解答と属性が異なり、「構造正解」と判定されたものが多くなっている。しかしながら、クローズテストでは、文脈類似度を利用することにより、用例追加による副作用を抑えて、96%以上の精度を保持している。また、オープンテストでは、用例の増加に比例して精度が向上しており、本手法の有効性を示している。

本報告書では、隣接語句の文脈を含む部分木用例との多角的な類似度計算に基づくボトムアップな構文解析手法の概要と、いくつかの実験結果を示した。本手法は、構文・意味・用法的制約を機能・意味・形態および局所文脈の類似度に置き換えることにより、制御を容易にし、かつ処理の頑強さを向上させている。特に、局所文脈の利用により、用例知識の追加による副作用を抑え、インクリメンタルな性能向上が可能であることを示している。

今後の課題としては、(1)さらにデータを増やして適用性を確認すること、(2)言語モデルを詳細化すること、(3)類似度計算対象を拡張すること(メッセージ間、場面や状況の取り込み)、があげられる。

注15 例えば、「そうですね」という表現の発話行為タイプ(反応的応答か判断要求か)の判定に直前メッセージ情報を使用していない。

謝辞

本研究の機会を与えて下さいました山崎泰弘社長、ご指導いただいた飯田仁室長、古瀬蔵主任、また、数々の貴重なご意見をいただいた研究室の연구원諸氏、本研究のデータ作成および実験に緻密な作業をいただいた増井淳子さん、皆さまに心から感謝いたします。

参考文献

- [1] M. Nagao, "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," in *Artificial and Human Intelligence*, eds. A. Elithorn and R. Banerji, North-Holland, pp. 173-180 (1984)
- [2] S. Sato, "Example-Based Machine Translation," Doctorial Thesis, Kyoto University (1991)
- [3] 隅田 ほか, "英語前置詞句係り先の用例主導あいまい性解消" 信学論 Vol. J77-D-II No.3 pp.557-565 (1994)
- [4] 古瀬 ほか, "経験的知識を活用する変換主導機械翻訳" 情処学論 Vol. 35 No. 3, pp. 414-425 (1994)
- [5] Y. Sobashima, O. Furuse, H. Iida, "A corpus-based local context analysis for spoken dialogues," *Speech Communication*, Vol 15, pp. 205-212 (1995)
- [6] A. S. Hornby, *Verb Patterns in "Oxford Advanced Learner's Dictionary of Current English,"* published by Kaitakusha in Japan (1974)
- [7] 花本金吾, F. J. Edamatsu, "実用英会話必携," 全面改定版, 旺文社 (1993)
- [8] 斎藤なが子, "よく使われる英検3級レベルの会話表現333," 旺文社 (1993)
- [9] O. Furuse, *et al*, "Bilingual corpus for speech translation, " *Proc. of AAAI -94 Workshop on "Integration of Natural Language and Speech Processing,"* pp. 84-89 (1994)

付録

A 属性分類とデータ仕様

A.1 属性分類

この節では、多元的類似度計算に基づく文脈依存のボトムアップパーサ (MAC-BP) で使用した各種属性の分類コード表を示す。

表A-1 日本語品詞分類

表A-2 日本語活用分類

表A-3 英語品詞分類

表A-4 英語活用分類

表A-5 語句の意味分類 (日英共通)

表A-6 発話行為タイプ (日英共通)

表A-1 (1) 日本語品詞分類

N00	名詞類	#
N10	名詞	#
N20	名詞句	# (不使用)
N30	代名詞	#
N31	特定代名詞	# 私/それ/あれ
N32	不定代名詞	# 誰/どれ(が)/いつ(が)/どこ(が)/何(が)
N33	連体形式名詞	# 方/はず/わけ/ところ/こと/とき/もの
N40	固有名詞	#
N50	単位	#
N51	数量単位	# 枚/名/日間/円
N52	位置単位	# 号/章/日
Q00	数詞類	#
Q10	数詞	#
Q11	基本数詞	# 〇~九
Q12	倍数数詞	# 十/百/千/万/億/兆
Q20	値	# 2メートル
A00	形容詞類	#
A10	形容詞	#
A20	形容動詞	#
A30	連体詞	#
A31	特定連体詞	# この
A32	不定連体詞	# どの
A40	A型活用助動詞句	#
A41	形容活用型	# ~たい
A42	形動活用型	# ~べき
V00	動詞類	#
V10	動詞	#
V11	が	#
V12	を	# 始める
V13	に	# 会う
V14	と	# 知る
V15	に・を/と	# 伝える
V16	を・に/と	# 定義する/変える
V17	に・が	# 分かる
V18	から	# 離れる
V19	の/~する	#
V20	動詞句	# (不使用)
V30	V型活用助動詞句	#
V31	動詞活用型	# ~始める/~ている

表A-1(2) 日本語品詞分類

D00	副詞類	#
D10	文副詞	#
D11	評価副詞	# 幸い
D12	強調副詞	# 特に
D13	接続副詞	# もし/実は
D14	付加副詞	#
D15	順番副詞	# まず
D20	対象副詞	#
D21	前置対象限定	# 約
D22	後置対象限定	# ぐらい
D30	事象副詞	# しばしば/きっと
D31	時副詞	# 昨日
D32	場所副詞	# ここで
D33	頻度副詞	# よく/全く
D34	否定副詞	#
D35	完遂副詞	# まだ/もう
D36	条件副詞	# その他/~場合
D40	属性副詞	#
D41	強調副詞	# かなり
D42	程度副詞	# どのくらい
D43	様態副詞	# どう
D50	疑問副詞	# 誰が/どこで/何を/いつ(=>代名詞)
D51	疑問副詞	# どこで/いつ
D52	疑問連体詞	# (不使用)
D53	疑問名詞	# (不使用)
D54	疑問属性副詞	# (不使用)
S00	文	#
S10	平叙文	# ~です/~ます
S20	疑問文	# ~か
S30	念押文	# ~ね/~よ
R00	間投表現	#
R10	応答表現	# はい
R20	挨拶語	# ありがとうございます
R30	働き掛け語	#
M00	メッセージ要素	# (不使用)
M10	前置メッセージ	# 反応語「あ」、思考語「そうですね」含む
M20	中心メッセージ	# メッセージの中心
M30	後置メッセージ	# 作用語「ね/よ」、「~様」含む
J00	接続要素	#
J10	思考語	#
J20	接続(助)詞	#
J21	接続助詞	#
J22	接続詞	#
J30	記号	#
J31	句点	#
J32	読点	#

表A-2 日本語活用分類表

#####

- # 名詞類 (N~)
- # 数詞類 (Q~)
- # 副詞類 (D~)
- # 間投表現 (R~)
- # メッセージ要素 (M~)
- # 接続要素 (J~)

#####

0 活用なし

#####

形容詞類 (A~)

#####

		# (形容詞)	(形容動詞)
0	語幹	# 美し	元気
1	ない形	# 美しく	元気で
2	た形	# 美しかつ	元気だつ
3	に形	#	元気に
4	終止形	# 美しい	元気だ
5	です形	# 美しい	元気
6	連体形	# 美しい	元気な
7	条件形	# 美しけれ	元気なら
8	仮想形	# 美しかろ	元気だろ

#####

動詞類 (V~)

文 (S~)

#####

		# (五段)	(一段)	(サ変)
0	語幹	# 書	食べ	研究
1	せる形	# 書か	食べさ	研究さ
2	れる形	# 書か	食べら	研究さ
3	ない形	# 書か	食べ	研究し
4	ず形	# 書か	食べ	研究せ
5	ます・たい形	# 書き	食べ	研究し
6	た/て形	# 書い	食べ	研究し
7	だ/で形	# 踏ん(だ)		
8	御~致す形	# 書き	食べ	研究
9	終止・連体形	# 書く	食べる	研究する
A	条件形	# 書け	食べれ	研究すれ
B	命令1形	# 書け	食べろ	研究しろ
C	命令2形	#	食べよ	研究せよ
D	仮想形	# 書こ	食べよ	研究しよ

表A-3 (1) 英語品詞分類

N00	名詞類	#
N10	名詞	#
N11	可算名詞	# girl
N12	可算名詞	# information
N20	名詞句	# a room registration
N30	代名詞	#
N31	主格代名詞	# I
N32	目的格代名詞	# me
N33	再帰代名詞	# myself
N34	指示代名詞1	# this/mine
N35	否定代名詞	# none
N40	固有名詞	# Kyoto
N50	單位	#
N51	數量單位	# dollar
N52	位置單位	# Chapter (one)/Room ~
Q00	數詞類	#
Q10	數詞	#
Q11	基本數詞	# zero ~ nineteen
Q12	十數數詞	# twenty/thirty/.../ninety
Q13	倍數數詞	# hundred/thousand/million
Q20	值	# two meters
A00	形容詞類	#
A10	連體形容詞類	#
A11	數量形容詞	# many
A12	形狀形容詞	# large
A13	形質形容詞	# good/young/atomic/Chinese
A14	狀態形容詞	# happy
A20	連用形容詞類	#
A21	數量形容詞	#
A22	形狀形容詞	# large
A23	形質形容詞	# good/young
A24	狀態形容詞	# happy
A30	限定詞	#
A31	選折詞	# all/both/such/some/any
A32	否定限定詞	# no
A33	冠詞	# a/the
A34	所有格代名詞	# your/my
A35	指示代名詞2	# this/these
V00	動詞類	#
V10	動詞	#
V11	自動詞1	# (LONGMAN-I) sing
V12	自動詞2	# (LONGMAN-L) become (old)
V13	他動詞1	# (LONGMAN-T) meet
V14	他動詞2	# (LONGMAN-D) give (me) (your name)
V15	他動詞3	# (LONGMAN-V,X) saw (her) (singing)

表A-3 (2) 英語品詞分類

V20	動詞句	# saw her singing/given a book
V30	HAVE動詞句	# have been given a book
V31	HAVE動詞	# => テンプレート記述
B00	BE動詞句	# been given a book
B10	BE動詞	# => テンプレート記述
X00	助動詞句	# can be seen there
X10	助動詞	# => テンプレート記述
D00	副詞類	#
D10	文副詞類	# ■文全体を形容する
D11	評価副詞	# happily
D12	強調副詞	# particulary
D13	接続副詞	# however/also/too
D14	付加副詞	# reportedly
D15	順番副詞	# firstly
D20	対象副詞類	# ■名詞句を形容する
D21	前置対象限定	# about
D22	後置対象限定	# .. alone/or less
D30	事象副詞類	# ■事象を形容する
D31	時副詞	# yesterday
D32	場所副詞	# here
D33	頻度副詞	# often/sometimes/all
D34	否定副詞	# seldom/never
D35	完遂副詞	# yet/already/almost
D36	条件副詞	# unless/otherwise/altogether
D37	様態副詞	# (study) hard/well
D38	内容規定副詞	# (go) further/(I will be) off
D40	属性副詞類	# ■形容詞・副詞を修飾する
D41	強調副詞	# very/far (apart)
D42	程度副詞	# how (old)
D43	その他	#
D50	疑問副詞類	# ■疑問文の文頭に置かれる
D51	疑問副詞	# where/how
D52	疑問連体詞	# whose/which/how
D53	疑問名詞	# what/who/whom
D54	疑問属性副詞	# how (many people)
S00	文	#
S10	完全文	# ■名詞類と動詞句 or 助動詞句
S11	平叙文	# I would like to make a reservation
S12	疑問文	# could you make a reservation
S13	感嘆文	# What a nice house the man has!
S20	骨格文	# ■代名詞と助動詞 or DO
S21	平叙文	# it is/he would
S22	疑問文	# is it/does she

表A-3 (3) 英語品詞分類

S30	省略文	# ■動詞 or 助動詞なしの疑問文・感嘆文
S31	疑問文	# for how many?, what nights?
S32	感嘆文	# What a nice house!/how nice!
S40	要素抜け文	# ■動詞支配要素が抜けた文
S41	平叙文	# what ("you said")
S42	疑問文	# which camera ("do you like")?
R00	間投表現	#
R10	応答語	# oh/yes/okay/hello/very good
R20	挨拶語	# thank you/you are welcome/not at all
R30	働き掛け語	# hello/sir/hey
M00	メッセージ要素	#
M10	前置メッセージ	# yes/okay/well
M20	中心メッセージ	#
M30	後置メッセージ	#
J00	接続要素	#
J10	思考語	# uhm
J20	接続詞	# though
J30	記号	#
J31	句点	# .?!
J32	読点	# ,

表A-4 (1) 英語活用分類

#####		
# 名詞類 (N~)		
#####		
1	単数形	# man
2	複数形	# men
#####		
# 形容詞類 (A~)		
# 副詞類 (D~)		
#####		
1	原級	# big
2	比較級	# bigger
3	最上級	# biggest
#####		
# 数詞 (Q~)		
#####		
1	基数形	# hundred
2	序数形	# hundredth

表A-4 (2) 英語活用分類

```
#####
# 動詞類 (V~)
#####
1  原形                # play
2  三単現形            # plays
3  過去形              # played
4  過去分詞形          # played
5  現在分詞形          # playing

#####
# BE動詞・句 (B~)
#####
1  原形                # be
2  一人称単数形        # am
3  二人称・複数形      # are
4  三人称単数形        # is
5  過去(単数)形        # was
6  過去(複数)形        # were
7  過去分詞形          # been
8  現在分詞形          # being

#####
# 助動詞・句 (X~)
#####
1  現在形              # can/will
2  過去形              # could/would

#####
# 文(S~)
#####
1  原形(不定形)
2  現在形
3  過去形

#####
# 間投語 (R~)
# メッセージ要素 (M~)
# 接続要素 (J~)
#####
0  表出形
```

表A-5(1) 語句の意味分類(日英共通)

100	人(HUMAN)	#
110	個人	# 神,人,観光客,国民,証人,投票人
111	個人名	# キリスト,鈴木
112	性識別	# 女,婦人,紳士
113	職業識別	# こじき,タレント,科学者,大工
114	地位・関係識別	# あなた,おじ,お母さん,花嫁,友,社長,知事
115	形状識別	# 巨人,白人,黒人
116	形質識別	# うそつき,悪魔,野蛮人,天才,左翼,民主主義者
117	状態識別	# アマチュア,患者,生徒,破産者,捕虜
118	年齢識別	# シニア,若者,中年,子供
119	出身識別	# アジア人,日本人,民族,国民,住民
120	組織	# 社会,組織
121	組織名	# アイスランド,国際連合,東京ホテル,京都大学
122	機能区分	# 支店,事務所,ホテル,研究所,メーカー
123	セクション	# チケット窓口,学部,厚生省
130	群衆	# 群れ,群衆,暴徒
200	物(OBJECT)	# お気に入り,ごみ,本体,媒介,偽物
210	施設	# ダム,港,大学,建造物
211	建物	# オフィス,ホワイトハウス,銀行,商店
212	部屋	# アパート,ツインルーム,一時預かり所,台所
213	道路・鉄道	# アーチ,街路,橋,鉄道,歩道
214	公園・広場	# コート,公園,庭園
215	乗物	# カヌー,ヨット,電車,航空機,急行,コーチ
220	家具・装置	# ハードウェア,装備,家財
221	設備	# 玄関,洗面所,エレベーター,プール
222	施設・設備部分	# 窓,柱,非常口,壁,改札口
223	家具	# いす,本棚,ベッド,席,カウンター
224	装置	# カメラ,ドライヤー,冷蔵庫,シャワー,テレビ
225	備品	# じゅうたん,寝具,クッション,カーテン
226	容器	# やかん,カップ,郵便受け,金庫,口座,箱
227	作用道具	# ナイフ,武器,鉛筆,小さじ,やすり,むち
228	部品・部分	# ねじ,取っ手,弁,棒,窓ガラス,タイル,ヒューズ
229	その他	# 模型,メダル,鏡,爆弾,墓
230	衣類・小間物	# レース,商品,必需品,贅沢品
231	衣服	# エプロン,替え,肌着,衣装,絹製品
232	装飾品	# ネクタイ,マスク,ポケット,宝石,帯,そで
233	帽子	# 帽子,ずきん,フード
234	靴	# サンダル,下駄,靴,スニーカー
235	小間物	# 財布,香水,包帯,口紅
236	荷物	# かばん,つつみ,手荷物,積荷,貨物

表A-5 (2) 語句の意味分類 (日英共通)

240	飲食物	# 産物,食糧
241	食物	# うどん,朝食,和食,野菜,鶏肉,チョコレート
242	飲物	# コーヒー,酒,ハチミツ,葉巻
243	薬品	# アスピリン,ビタミン,丸薬,毒,麻薬
250	作品・書類	# オペラ,ソフトウェア,ドラマ,一流品,音楽
251	読み物	# カタログ,百科辞典,案内書,小説,週刊誌
252	証明書・記録	# カード,荷物札,許可証,申込書
253	絵画・書	# イラスト,図,写真,版画,メニュー
260	金銭類	# ストック,基金,負債
261	現金	# つり銭,貨幣,金貨,紙幣,小銭
262	小切手・切符	# クレジットカード,乗車券,切手,宝くじ
270	物体・物質・材料	# 原料,混合物,材料
271	固体	# 金属,小麦粉,塩,脂肪,磁石,石炭,丸太
272	液体	# 雨,油,ニス,水,しずく,酢
273	気体・におい	# 酸素,ガス,臭い,風
274	火・光・熱	# エックス線,炎,日光,熱
275	音・声	# 音,つぶやき,歌,旋律,足音
276	色	# カラー,ピンク,黄色,陰
277	形	# 円筒,しみ,しわ,平行線,泡,縫い目,模様,縁
280	身体要素・部分	# 傷
281	身体部分	# あご,かかと,歯,耳
282	身体器官	# 胃,静脈,神経,盲腸
283	その他	# 癌,組織,皮膚,涙,血
290	生物	# 群れ,生き物
291	動物	# いけにえ,ウサギ,動物,猟犬
292	植物	# 木,草,花,イチゴ,稲,タマネギ
293	微生物	# ばい菌,バクテリア
2A0	景色・光景	# 景色,自然,壮観,名所
2A1	空・雲・星	# 空,雲,月,虹,北極星,彗星
2A2	陸・山・地	# ほら穴,岩,山,地球,地面
2A3	海・湖・川	# 海,小川,滝,湾
300	抽 (ABSTRACT)	# 何か,無,命
310	抽象素材	# 美,悪,知性,問題,起源
311	思考	# 意見,考え,空想,言い訳
312	心情	# あわれみ,自慢,確信,心配,熱意
313	意図・期待	# 意向,希望,望み,要求
314	情報(文字)	# 言葉,数字,暗号,アクセント,句
320	知的生産物	# 経過,実情
321	表現・約束	# 計画,演劇,警告,決まり,指導,結論,索引
322	理論・法則・制度	# 経済学,課程,マナー,時制
323	歴史・成果	# 影響,結果,文化,歩み,思い出
324	知的所有物	# 機知,経験,趣味,資格,安全,権利

表A-5 (3) 語句の意味分類 (日英共通)

325	言語名	# 日本語,英語
326	宗教名	# 仏教,キリスト教
327	技能・方式名	# こつ,航空便,遠近法,文体,輸送
328	産業名	# 農業,工業,漁業,ビジネス
330	関係定義語	#
331	名称・識別ラベル	# 名前,住所,色,年代,つづり
332	内容・要素	# 中身,趣意,本質,要因,詳細
333	構成・構造	# 体系,構成,外観,モデル,配列,模範,書式
334	種類	# 科目,種類,分野
335	目的・原因・条件	# きっかけ,目的,用途,理由,場合,なぜ
336	手段・方法	# 手段,対策,用法
340	能力・機能	# 理性,才能
341	力・能力	# 影響力,重力,洞察力,努力,権限,技量
342	機能	# 機能,働き,代用,役割,係り,サービス
343	可能性	# 可能性
350	水準・基準	# 海拔,単位,JIS,標準時
400	時 (TIME)	# 最盛期
410	時識別ラベル	# 時間,時期,時代,チェックアウトタイム,いつ
420	絶対時間	# スパン,幼時,老後
421	年	# 年,学年,歳
422	季節	# 春,夏,休み,収穫期,新年
423	月	# (三)[月]
424	曜	# 週,火曜
425	日	# 毎日,元日,明日,誕生日,クリスマス
426	朝・昼・夜	# 朝,前夜,昼間,今夜,夕暮れ
427	時	# アワー,正午
428	時間分割	# 半
429	分・秒	# 分,秒
430	相対時間	# 最初,最後,途中
431	前	# 過去,かつて,最近
432	同時	# 今,本日,まだ,もう
433	後	# 将来,明日,後,まもなく
440	頻度	# また,再び
441	最高頻度	# 常に,いつも
442	高頻度	# 普通,たいてい,通常
443	中頻度	# よく,度々
444	低頻度	# 時々
445	極少頻度	# たまに
446	0頻度	# 決して,全く

表A-5 (4) 語句の意味分類 (日英共通)

500	所 (LOCATION)	# 目標, 余地, 広がり, 社会, etc. (抽象的な場所)
510	所識別ラベル	# 位置, 場所, 地域, どこ, どちら
520	絶対位置	# 空欄, 目的地, 所在地
521	天・宇宙・地球	# 宇宙, 空, 惑星, 天国, 空中
522	地形	# がけ, アルプス, 海岸, 沼地, 運河, 川, 山脈, 地平線
523	大陸・海洋	# アジア, ヨーロッパ, 北極, 半球
524	国	# 日本, アメリカ, 民主主義国, 国家, 植民地
525	州・地方	# 山形県, カリフォルニア, ハワイ, フロンティア
526	都市・町・村	# ニューヨーク, 東京, 村
527	区画	# 空港, 基地, 行楽地, 農園, 中野区, 六本木, 道のり, 鉄道, 路線
528	建物	# 駅, 映画館, 住居, ホテル, 金閣寺
529	部屋・廊下	# 入り口, 売店, 廊下, バス停, 浴室, ホール
530	相対位置・方向	# うわべ, 側, 角
531	前後	# 前, 後ろ
532	左右	# 右, 左, わき
533	上下	# 上, 以下, 底, 下
534	東西南北	# 西, 南西, 北方
535	遠近	# 遠方, 近く, 向こう, 手元, 辺り
536	内外	# 外, 間, 内部, 中, まず
540	時間・位置的数	# 一, 二, 百
600	量 (QUANTITY)	#
610	数量ラベル	# どれくらい
611	人数	# 人口, 会員数, 人手
612	個数	# 数, 票, 部数
613	量	# エネルギー, シェア, 雨量, 温度, 成績, 能率
614	期間	# スパン, 一生, 期間, 年齢
615	金額	# チップ, ローン, 運賃, 手当, 収入, 利息, いくら
616	大きさ	# 長さ, 面積, 寸法, 規模, 角, いくつ
620	数量値	# 違い, 制限
621	人数	# シングル, ツイン, (二)[名]
622	個数	# 行, (四)[個]
623	量	# パーセンテージ, 割合, 量, (百)[メートル]
624	期間	# 時代, 年間, (三)[年]
625	金額	# 大金, (五)[ドル]
626	大きさ	# (三)[倍]
630	数量単位	#
631	人数単位	# 人, 名
632	個数単位	# 個, 点, セット
633	量単位	# フィート, 袋, 杯 (数を使った計量)
634	時間単位	# 年, 月, 週, 日, 時, 分, 秒, 学年, 世紀, 歳
635	金額	# ドル, 円
636	大きさ	# 号

表A-5 (5) 語句の意味分類(日英共通)

640	数	#
641	ゼロ	# 零,ゼロ
642	1	# 一
643	複数	# 二,百,万,億
644	極少数	# 微,極
645	少数	# 数
646	多数	# ほとんど,大勢,たくさん
647	不定数	# 何,幾,いくつ
650	絶対量	# 一団,塊,最大
651	0量	# ぜんぜん
652	極少量	# たった
653	少量	# 少量
654	大量	# たくさん,大容量
660	相対・比較数量	# 半分
661	同程度	# 程度,約,ぐらい,前後
662	比較小	# 以下
663	比較大	# 以上,もう,もっと
664	一部・一方・他方	# いずれ,分,それぞれ
665	両方・全部	# すべて,全員,全室,両方,双方,皆
700	態 (STATE)	# どうやって
710	形状記述	# 外観,美しさ
711	対人形状記述	# 笑い,表情,日焼け
712	対物形状記述	# 三角,坂,和室,傾き,小型
713	その他	#
720	形質記述(評価)	# くず,類似,多様性
721	対人形質記述	# 欲張り,几帳面,良心
722	対物形質記述	# 正確,中古,オートマチック
723	その他	# 普通,現実,ナンセンス
730	状態記述	# 圧力,安全性,危険,仕上げ,手があく,偶然
731	心理状態	# いらだたしさ,怒り,失望,人気,熱中,感心する
732	物理状態	# カーブ,ほてり,体格,洪水,高さ
733	社会状態	# 違反,不法,敗北,状況,緊張
734	その他	# 完全
740	存在	# 出席,生存,ある
750	関係	# 関係,リンク,比較,対応,仲
751	等価	# 一致,等価,似る,同じく
752	相違	# 違い,逆,裏,違う,他
753	所有・内含	# 結合,所有,付く
754	所属・構成	# 帰属
760	程度	# どのくらい
761	同程度	# ちょうど,そんなに
762	0程度	# ぜんぜん,とても,決して
763	極少程度	# あまり

表A-5 (6) 語句の意味分類 (日英共通)

764	小程度	# ちょっと,しばらく(数量は660付近を参照)
765	中程度	# ある程度,いくらか
766	大程度	# かなり,大変(数量は660付近を参照)
770	思考判断内容	# 好都合,たぶん,どうも,やっぱり
771	義務	#
772	必要	# 必要性,絶対に
773	可能・正当	# 可能性,無理,有効,できるだけ
800	事 (EVENT)	# ごたごた,事
810	開始	# 復活
811	(人)生誕	# 出生,誕生
812	(物)出現	# 出現,発見
813	(抽)有効になる	#
814	(所)出発	# 出発,離陸
815	(事)発生	# スタート,開始,突発
820	継続	# 継承,連続
821	(人)生活する	# 生活
822	(物)無変化	# 不変
823	(事)継続する	# 繰り返す,続発,相続
830	変化	# 過程,転換
831	(人)~になる	# 生長
832	(物)変わる	# 交換,変化
833	(抽)変質する	# 改善,衰え,低下
834	(所)移動する	# 移動,逃亡,避難,飛行,再帰
835	(時)過ぎる	# 経過
836	(量)増減する	# 増加,膨張,割増
837	(様)変化する	# 回復,復興
840	終了	# 最後,終わり
841	(人)死ぬ	# 死
842	(物)消滅	# 破壊
843	(抽)無効になる	#
844	(所)到着	# 一致,完成,接触,達成,到着
845	(事)終了	# ポーズ,中断,停止,閉店
850	感覚・識別現象	# 感動,興奮
851	感覚	# 嗅覚,感じ
852	知覚	# 知覚,理解
853	認識	# 承認,是認,認識
854	思考	# 思考
855	表意	# 笑い
860	生理現象	# うたた寝,くしゃみ,骨折,食欲,眠り
870	物理現象	# うめき,日光,洪水,汽笛,雪解け,故障
880	社会現象	# ききん,スキャンダル,戦争,犯罪,普及
890	催し物	# 大会,裁判,レセプション,祝祭,試験,結婚
900	行 (ACTION)	#
910	動作	# 清掃,食事

表A-5 (7) 語句の意味分類(日英共通)

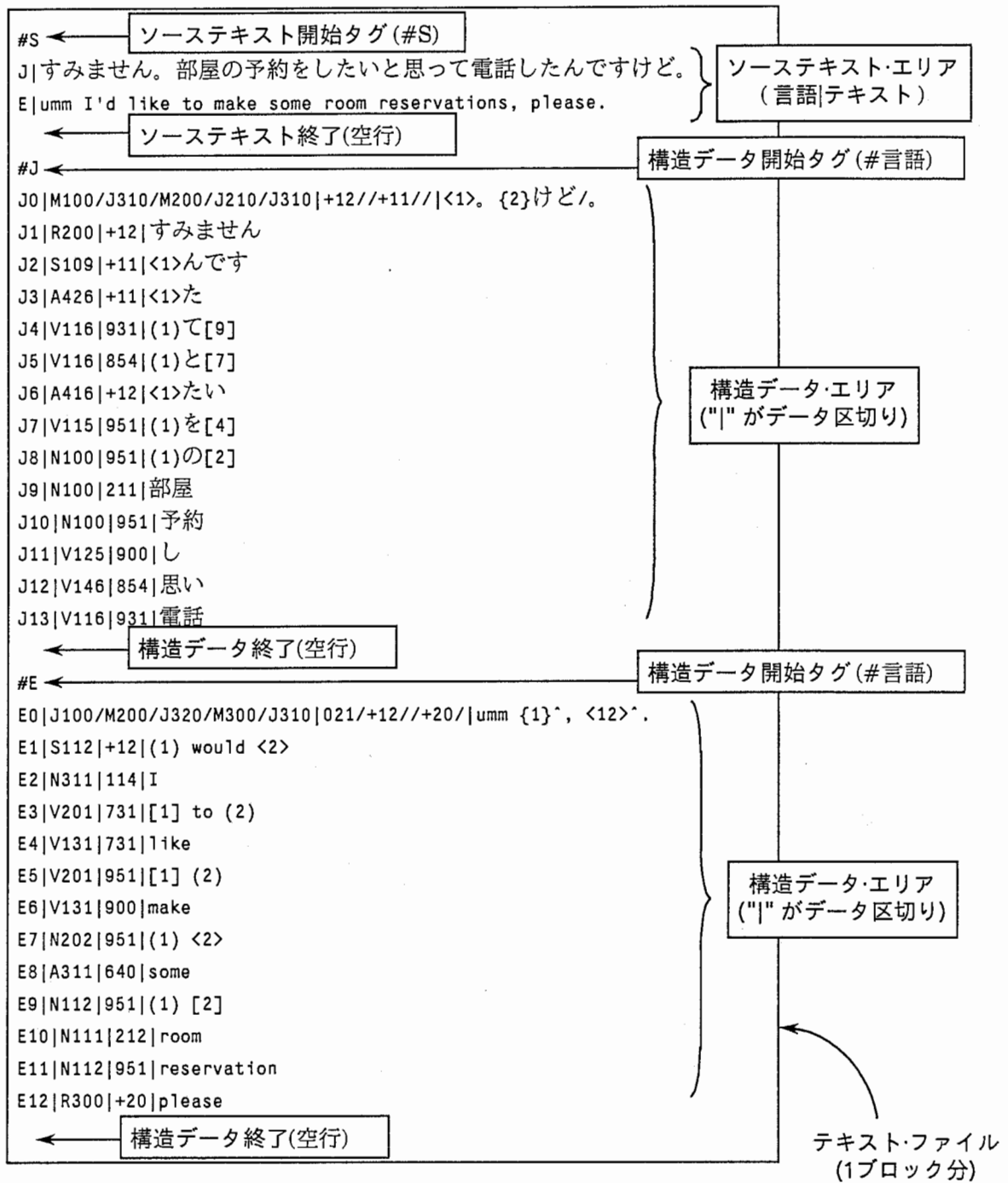
911	対物動作	# 切断,連結,手術,打撃,咀嚼,ノック
912	身体動作	# おしゃべり,まね,踊り,身ぶり,喫煙,話す
913	感覚動作	# のぞき見,凝視
920	移動(人)	# フライト,横断,行進,通学,パレード
921	出発	# 出発,出る
922	到着	# 出席,上陸,訪問
923	滞在	# 宿泊,駐車,避難,加入
930	移動(物・情報)	#
931	運搬・連絡	# コール,急送,伝達,郵便,輸出,送る
932	入手・収集	# 購入,収穫,導入,万引き,輸入,取る
933	譲渡・分配	# デモ,演説,小売り,宣言,発行,注意,教える
934	交換	# 会話,文通,交換する
935	保持	# 持ち込み,保持
936	放棄	# 降伏,処分
940	セッション(事)	#
941	開始	# 開始,着手,起こす
942	中断・終了	# 辞職,停止,終了,済ます
950	作用(物・情報)	# 治癒,手続き,準備,整理,用意
951	作成・記録	# 作曲,入力,描画,処方,登録する
952	変更・加工	# 拡張,割引,乗換,装飾,訂正
953	消去・解除	# キャンセル,破壊,チェックアウト
954	保管・維持	# 維持,運営,貯蔵,保存,管理
955	使用・操作	# 運転,操作,調節,表示,利用
960	作用(人・所・事)	# 関与,拒絶,激励,抗議,招待
961	実施・攻撃	# 一撃,強盗,遂行,処罰,逮捕,妨害,反乱
962	保護・援助	# 援助,防衛,防止,支持,弁護,世話
963	許可	# 委任,入会,認可
964	強制	# コントロール,制限,支配,規制,刑罰,限定
965	要求	# 願い,主張,抵抗,モーニングコール
966	応答	# 応答,反応,返事
969	受諾・服従	# 償い,是認
970	思考作用	# 謝罪,信仰,評論
971	調査・研究・搜索	# 検査,考察,実験,分析,勉強する
972	計算・推論	# 会計,勘定,推測,天気予報,理解
973	分類	# 割り当て,区別,分離
974	選択	# 雇用,選挙,投票,選択
975	決定・約束	# 解決,デート,不承知,予定,判決,宣言
980	機能・活動	# 業務,活動,用事
981	機能・労働	# 任務,営業,実行
982	生産・開発	# 印刷,建設,制作,発行
983	消費・浪費	# 消費,買い物
984	維持・運営	# 運営,統治,保護,推進,家事,監督
985	討論・交渉	# 会議,議論,競争,交際,交渉,試合,討議,闘争,非難
986	運動・娯楽・芸術	# ロック,ゴルフ,娯楽,趣味,レース
987	情報伝達	# 授業,講演,通信,通話,広告,出版

表A-6 発話行為タイプ(日英共通)

-00	Reaction (反応)	#
	-10 Response (反応的応答)	# 「あ」
	-11 Surprise (驚き)	#
	-12 Confirmation (確認)	# 「そうですか」
	-13 Understanding (理解)	# 「分かりました」
	-20 Reply (陳述的応答)	# 「はい」
	-21 Judgement (判断)	# 「そうです」
	-22 Information (情報)	# 「京都観光ホテルです」
	-23 Promise (約束)	# 「お送り致します」
000	Thinking (思考)	#
	010 Murmuring (独言)	#
	011 Understanding (理解)	# 「そうか」
	012 Judgement (判断)	# 「しかたないな」
	013 Information (情報)	# 「3日かー」
	020 Thinking-words (思考語)	#
	021 Content (内容思考語)	# 内容思考語「うーん」
	022 Express (表現思考語)	# 表現思考語「あの一」
	023 Word (語思考語)	# 単語思考語「えーと」
+00	Action (作用)	#
	+10 Informing (提供)	#
	+11 Event (事象)	# 過去・現在・未来
	+12 Feeling (心情)	# 陳謝危惧・希望要望・挨拶
	+13 Judgement (判断)	# 理解受諾・正誤肯否・決意選択 # 義務必要・可能許可
	+20 Asking (要求)	#
	+21 Judgement (判断)	# Y/N疑問・確認
	+22 Information (情報)	# 選択疑問・提供依頼
	+23 Action (行為)	# 行為依頼・禁止・許可

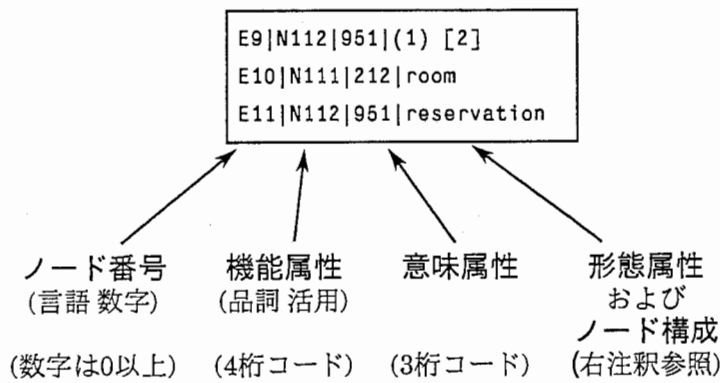
A.2 用例ソースファイルの仕様

図A-1 に用例ソースファイルの例を示す。



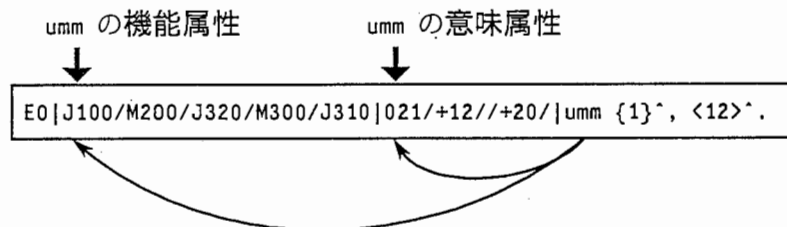
図A-1 用例ソースファイル

用例ソースファイルは、メッセージ単位に対応する「ブロック」から構成される。また、1つのブロックは"#S"で開始するソーステキストエリア、"#言語 (J: 日本語,E:英語)"で開始する言語ごとの構造データ・エリアから構成される。いずれのエリアも空行を用いて範囲を表す。各エリアは、いくつかの行(ここではレコードと呼ぶ)からなる。各行のレコードは、区切り文字"|"によりフィールドに分けられる。図17に示すように、構造データ・エリアの各レコードは、1つの部分木(構造)または単語(要素)を表し、各レコードのフィールドは、構造または要素の属性を表す。また、図18に示すように、トップノードの各要素の属性は、区切り文字"/"(スラッシュ)を用いて列挙する。



第4フィールド(形態属性 およびノード構成)中の番号は構成要素の相対位置を表す。例えば"room reservations"のノード(E9)はE10 ノードとE11 ノードから構成される。それぞれの相対位置は+1,+2であり、役割はサブとSヘッドである。そのため、E9 ノードのノード構成は"(1) [2]"と記述する。

図A-2 レコード(ノード・データ)の構成



メッセージノードの各要素
(J100, 021, "umm")
(M200, +12, main-message)
(J320, NIL, ".")
(M300, +20, post-message)
(J310, NIL, ".")
(機能,意味,形態)

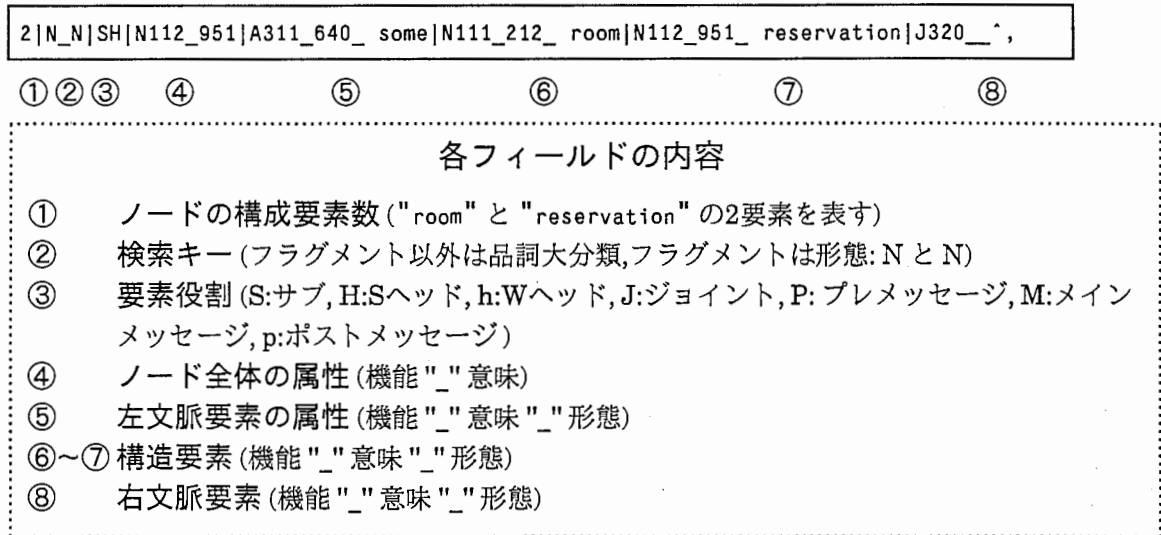
図A-3 トップノード・レコードの構成

第4フィールドで記述する形態属性は、語の標準形を用いる。ただし、日本語の場合、サ変動詞、形容詞、形容動詞は語幹、動詞は「ます」に接続する連用形を採用する(サ変「する」は語幹がないため「し」を標準形とする)。例えば「～たいと思って～」の「思

う」は「思い」が標準形となる。このような標準形の採用により、他品詞への派生語(例えば「思い」という名詞)とのリンクが可能となる。また、英語の標準形では、空白以外の非アルファベット文字(例えば",")とアルファベット文字との間には"."を挿入する。このことにより、英語では空白または"."を語の区切り文字に使用し、語の先頭はこのいずれかの区切り文字で開始する(例「room」「'11」(willの短縮形)。

A.3 部分木用例データの仕様

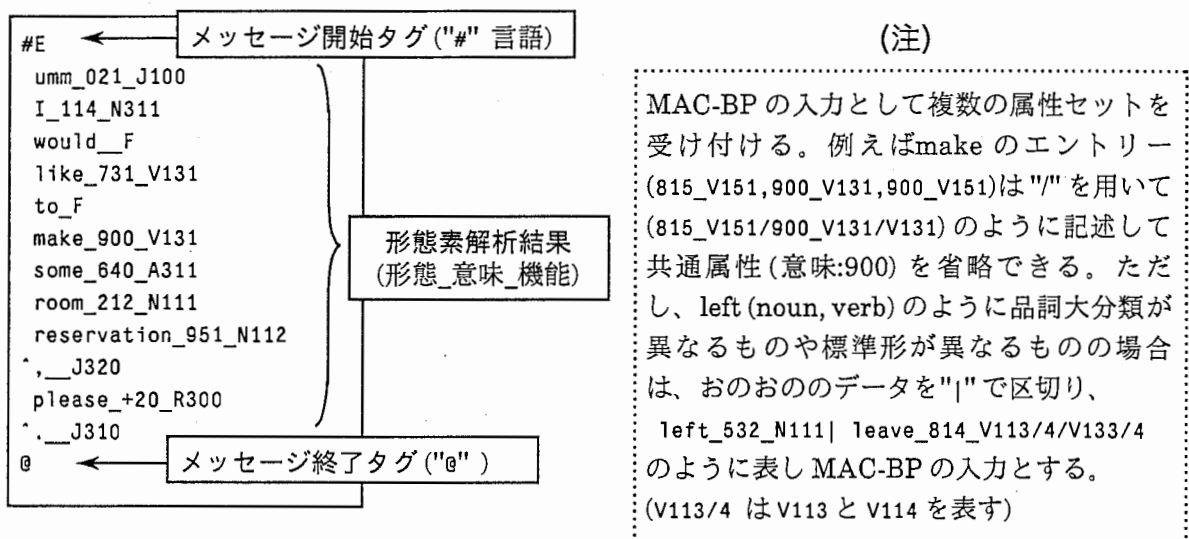
用例ソースファイルから部分木データを抽出し、構文解析の用例データとして使用する。図19は、「room + reservations → room reservation」の句構成を表す左右文脈付き部分木データの例である。ここでは、「some」が左文脈要素、「.」が右文脈要素となっており、「room」と「reservations」が結合した後生成される「room reservation」は「N112_951」という属性を持つことを表している（"_"はアイテムの区切り文字）。



図A-4 部分木データの構成

A.3 パーサ入力データの仕様

パーサ(MAC-BP)へは、図20のような形態素解析結果を入力する。



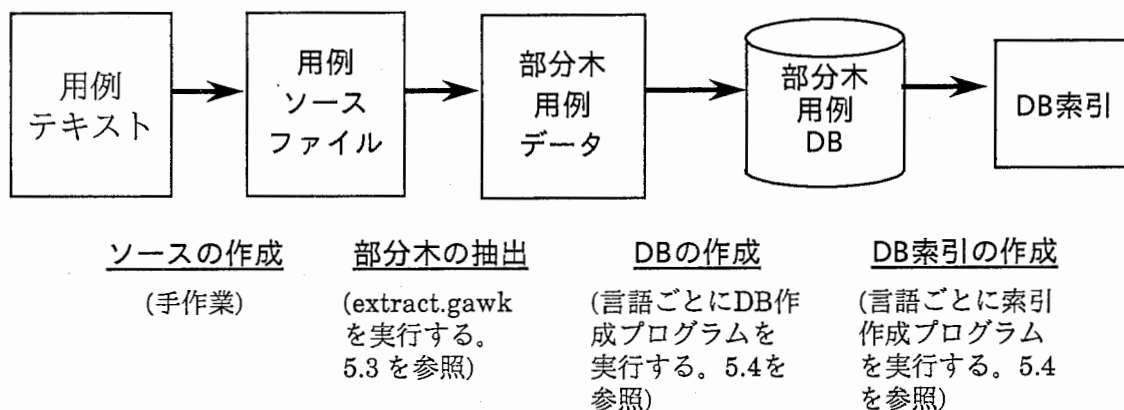
図A-5 部分木データの構成

付録

B 処理の手順とパーサの操作方法

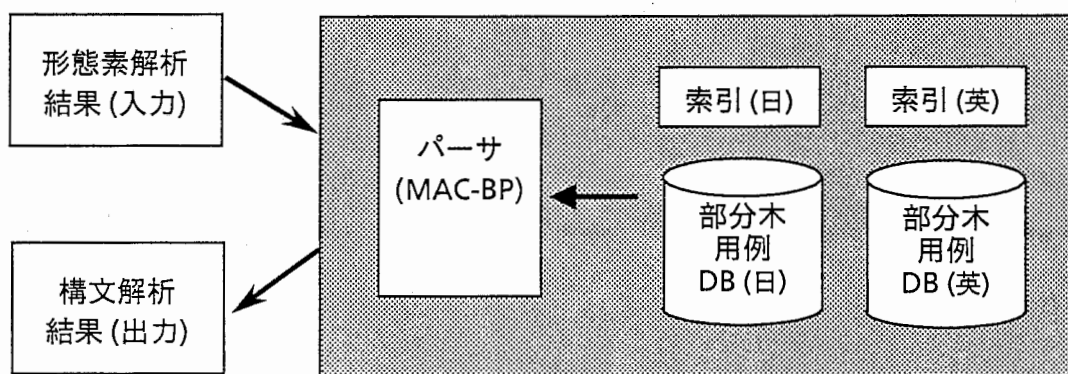
B.1 処理の概要

パーサ (MAC-BP) を実行するためには、用例ソースファイルを作成し、部分木用例データを抽出した後、部分木用例データベースを作成して構文解析のための知識ベースとする必要がある。図B-1にこの処理の流れを示す。



図B-1 パーサ利用のための準備

パーサは、このようにして作成された日英各言語ごとの部分木用例データベースおよびそれぞれの索引を利用して多元的類似度計算を実行し、ボトムアップに構文解析を行う (図B-2参照)。



図B-2 パーサの実行環境

B.2 用例ソースファイルの作成

用例ソースファイルは、エディタを用いて「付録A2」の仕様で作成する。ファイル名は、一般に「*fname.tag*」(*fname* は任意)のように拡張子として「*tag*」を使用する。

B.3 部分木用例データの抽出

用例ソースファイル (*fname.tag*) を入力として以下のコマンドを実行し、部分木用例データ (*fname.dat*) を抽出する。

```
% gawk -f extract.gawk fname.tag > fname.dat (日英部分木用例データ)
```

B.4 部分木用例データベースの作成

部分木用例データを入力として以下のコマンドを実行し、部分木用例データベース (*DB.sj*, *DB.se*) を作成する。

```
% gawk -f jsyndic.gawk fname.dat | sort | uniq > DB.sj  
                                     (日本語部分木用例データベース)
```

```
% gawk -f esyndic.gawk fname.dat | sort | uniq > DB.se  
                                     (英語部分木用例データベース)
```

さらに、以下のコマンドを実行し、各言語ごとの部分木用例データベースの索引 (*DB.sjz*, *DB.sez*) を作成する。

```
% gawk -f index.gawk DB.sj > DB.sjz (日本語部分木用例データベース索引)
```

```
% gawk -f index.gawk DB.se > DB.sez (英語部分木用例データベース索引)
```

B.5 構文解析の実行

パーサ (MAC-BP) を実行する場合、カレント・ディレクトリに、上記 (言語ごとの) 部分木用例データベース、およびそれぞれの索引 (4ファイル) が存在する必要がある。MAC-BP はフィルタであり、入力は標準入力、出力は標準出力に対して行われる。以下に示す例は、「*fname.mor*」という形態素解析結果のファイルを入力として画面に結果を表示するものである。

```
% macbp < fname.mor
```

B.6 構文解析の実行結果

MAC-BP を実行すると、途中経過(ノード作成の模倣を行った用例)と解析結果を出力する。図B-3は、MAC-BP 出力の例である。



図B-3 MAC-BP の出力

B.7 構文木表示ツール

MAC-BP の実行結果は、構文木表示ツール(`tree.awk`)を使用して構文木を表示する。
例えば構文解析結果ファイル `fname.syn` の構文木を画面に出力する場合、次のように入力する。

```
% awk -f tree.awk fname.syn
```

形態素解析結果ファイル `fname.mor` の構文木を画面に出力する場合、次のように入力する。

```
% macbp < fname.mor | awk -f tree.awk
```

図B-3 の構文解析結果を構文木表示させた例を、図B-4 に示す。

```
#E
{}^.*010_/+12/
|- () would <>_S112_+11
   |- I_N311_114
   |- [] to ()_V201_731
      |- like_V131_731
      |- [] ()_V201_951
         |- make_V131_900
         |- a <>_N201_951
            |- () []_N111_951
               |- room_N111_212
               |- reservation_N111_951
```

図B-4 構文木表示の例