TR-I-0326

日本語形態素解析法の評価

Evaluation of Japanese Morphological Analysis for ATR Dialogue Database

浦谷 則好　　松尾 秀彦*　　高橋 誠*

Noriyoshi Uratani　Hidehiko Matsuo　Makoto Takahashi

1993.3

概 要

ＡＴＲ自動翻訳電話研究所では言語データベースの構築を進めてきた。言語データベースのデータに付与する形態素情報は格・係り受け関係属性や日英対応の基礎ともなっているので重要である。日本語形態素情報付与の作業は計算機による形態素解析結果を人手によって全数チェックするという半自動的な方法で行っている。計算機による形態素解析の精度や速度はその後の作業負担に大きな影響を及ぼす。そこで、定量的側面および定性的側面から形態素解析の精度や速度の評価を実施した。本報告ではこの評価結果をその後の人手による作業効率もあわせて報告する。

ＡＴＲ自動翻訳電話研究所

ATR Interpreting Telephony Research Laboratories

---

*株式会社　東洋情報システム

目　次

付 録

1. 概要

本報告書では、今後の研究に有益な知見を残すことを目的として、定量的側面，定性的側面から、ＡＴＲ対話データベース (以下、ADD)作成の日本語形態素情報作成工程で利用されている日本語形態素解析システムの評価実験について報告する。

ＡＴＲ自動翻訳電話研究所では、自然言語特に話し言葉を対象とした研究を行うために、電話あるいはキーボードを介した模擬会話により対話収集を進めてきた。収集した対話を文字化した対話テキストは、日本語形態素情報，格・係り受け情報，英語による対訳対応情報などを付与して、ADD に格納されている。

本報告書が扱う日本語形態素解析システム（以下、本システム）は、ADD の作成において当所が定めた基準（テクニカルレポートTR-I-0077「形態素情報利用解説書（兼作業マニュアル）」：篠崎、水野、小倉、吉本 共著参照）に基づいて日本語形態素情報を付与する工程で利用されている。

まず、第２章で本システムの概要を述べ、解析手法の特徴に触れる。次に第３章では、本報告の主題であるシステムの評価実験を行う。ここでは、本システムの解析性能，解析精度，および辞書構成の評価、ならびに本システムを利用した作業効率の評価等、様々な角度から評価を加える。

2. 日本語形態素システムの概要

本システムは、対話データベース作成における日本語形態素作成工程で利用されている日本語形態素データ作成環境であり、入力テキストの形態素分割を行なうメインモジュールの他、サブモジュールにおいて解析誤りの修正，文節切り，作成データの検証などをサポートしている。

本章では、2.1おいてデータ作成のフローにより本システムの全体的な構成を概説し、2-2 では形態素解析で採用している手法を説明する。

2-1 形態素情報

形態素解析で分割された単語に付与される情報は、表記データ,かな表記,標準表記,品詞,活用型,活用形,音便の７種類である。各情報の意味と実際の解析例を以下に示す。

表2-1. 形態素解析で付与される情報

情報の種類	説　　明
表記データ	テキストに現われた形 (以下、表記)。
かな表記	単語の読み。基本的にひらがな表記 (以下、かな)。
標準表記	活用語の場合、終止形。表記に揺れがある場合、活用・非活用を
	問わず、†典拠とする辞書見出しの表記を用いる(以下、標準)。
品詞	普通名詞、本動詞などのカテゴリー。
活用型	五段活用、下一段活用など活用型分類。
活用形	終止形、連体形などの活用形分類。
音便	促音便、撥音便、ウ音便、イ音便などの音便情報。

表2-2 形態素解析の例

入力文：「お世話になっております。」

表記	かな	標準	品詞	活用型	活用形	音便
お	お	御	接頭辞	ーー	ーー	
世話	せわ	世話	普通名詞	ーー	ーー	
に	に	に	格助詞	ーー	ーー	
なっ	なっ	だ	本動詞	五段	連用形	つ音便
て	て	て	接続助詞	ーー	ーー	
おり	おり	おる	補助動詞	五段	連用形	
ます	ます	ます	助動詞	ーー	終止形	
。			記号	ーー	ーー	

† ADDにおける日本語形態素の標準表記は、三省堂「新明解第二版」を典拠としている。

2-2 システム構成

(1) 基本フロー

形態素データ作成は、図2-2-1 に示すような流れで行われる。データ作成の基本的な流れは、†形態素解析 → 自動修正 → 手修正 である。日本語のような膠着語の形態素解析は技術的に非常に難しく、このように人手を介在させた半自動的な作成方法を採用せざるをえない。

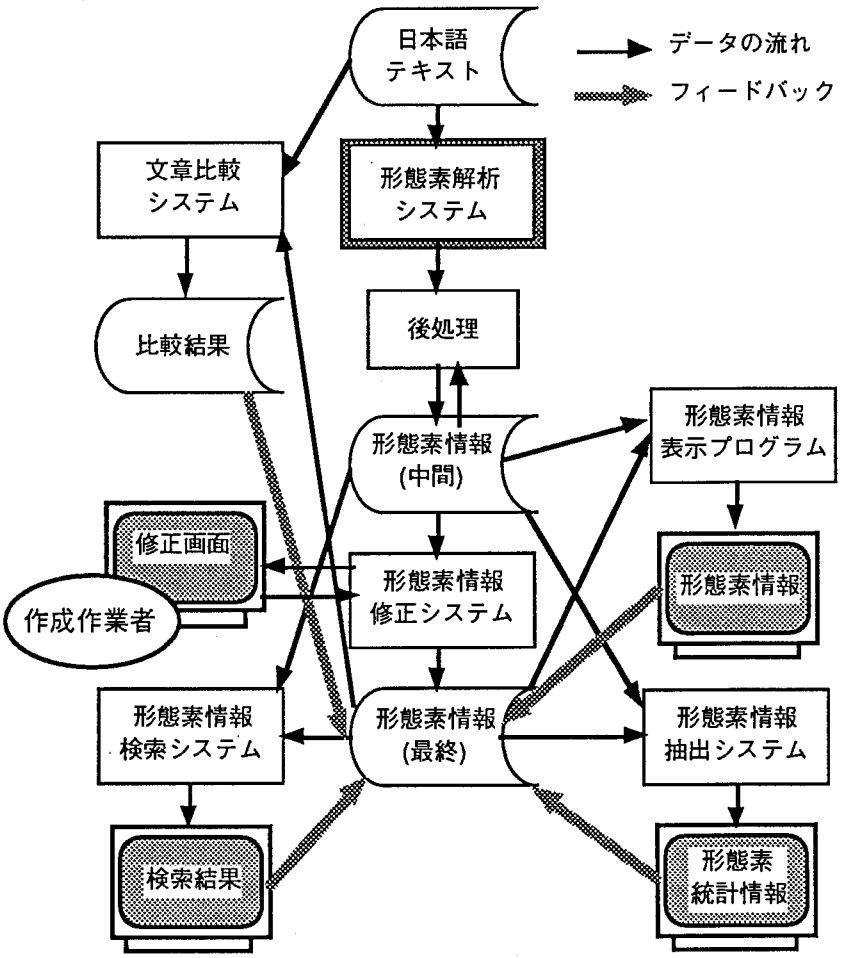


図2-2-1. 形態素情報作成の流れ

† ここでは、日本語形態素解析システムに焦点を当てているため、文節切り以降の説明は省いている。文節切り以降のデータ作成の説明は、3-C. 作業評価の節に譲る。

(2) 周辺システムとの関連

形態素解析システムとそれに関連する周辺システムの関係を次頁図2-2-2に示し、データの流れに沿って各システムの説明をする。

形態素解析システム(ANA)

日本語形態素解析システムは、日本語テキストを入力とし、形態素解析処理用の辞書や連接テーブル，その他の統計情報を参照しながら、2-1 で示したような形態素情報を出力する。

後処理システム(RULE REPLACE)

形態素解析の出力は、後処理システムにかけられ、形態素解析の誤りとその修正方法をパタン化した書換え規則を参照して、ある程度の自動修正が施される。

形態素情報抽出システム(EXTRACT)・日本語辞書編集プログラム(EDKDIC)

形態素情報抽出システムを使って、人手により修正された最終的な形態素情報から、日本語辞書のもとになる統計情報（異なり単語数）ファイルを作成する。次に、この統計情報ファイルを入力として日本語辞書編集プログラムにより、日本語自立語辞書(KDIC)を自動作成する。日本語辞書編集プログラムは、自立語辞書の他，付属語辞書，不規則変化辞書を編集することができる。

形態素解析用辞書構築ツール(SDICGEN)

形態素解析用辞書構築ツールは、日本語自立語辞書，付属語辞書，不規則変化辞書およびイディオム辞書の内容をマージして、形態素解析用辞書(SDIC)を構築する。

その他のツール

イディオム管理システム	イディオム辞書を作成・編集する。
連接テーブル修正ツール	連接テーブルのメンテナンスを行う。

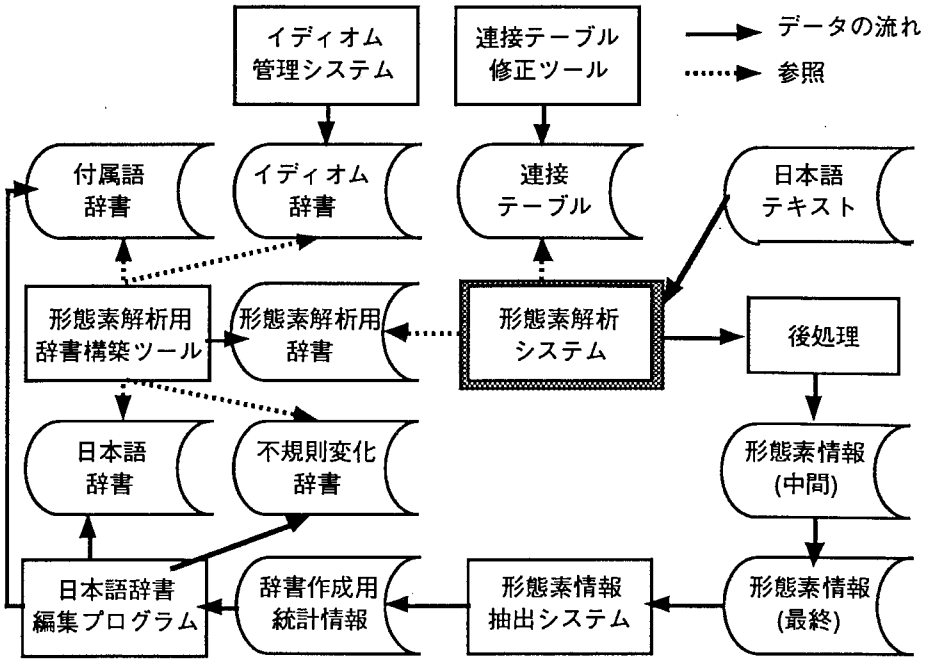


図2-2-2. 日本語形態素解析システムと周辺システム

2-3 形態素解析手法

本システムでは形態素解析の手法として、各種統計情報をパラメータとした最尤度探索を行なう、最尤度法を採用している。

2-3-1 特徴

基本的には、文節内での連接をチェックしながら、深さ優先で探索を行ない、第一解を出力とする。最尤度法の特徴は、探索経路の分岐点において、単語長と品詞優先度，品詞バイグラム，頻度などの4種類のパラメータにより単語の尤度計算を行ない、分岐点の単語候補を尤度順に並べ替えて評価する点にある。これにより、最も尤度の高い単語の組み合わせが第一解として得られる。

その他の特徴として、†字種切りをもとにした未知語処理機能を持つ。

以降、最尤度法に焦点をあてて説明する。

2-3-2 最尤度法

(1)尤度評価関数で使用されるパラメータ

単語長　　L

表記データの文字長。このパラメータだけを用いると、最長一致法による解析となる。

品詞優先度　　H

基本的には品詞の相対度数である。

但し、普通名詞，感動詞など一部の自立語については、相対度数に補正を加えたものを品詞優先度としている。これは、付属語には単語頻度が付与されていないので、自立語の品詞優先度を下げることによって調整をとるためである。

普通名詞は特に相対度数が高く、他の品詞の解析に悪影響を及ぼす傾向が見られたため、相対度数を下げている。現在設定している値は経験値であり、正当性を保証するものはない。

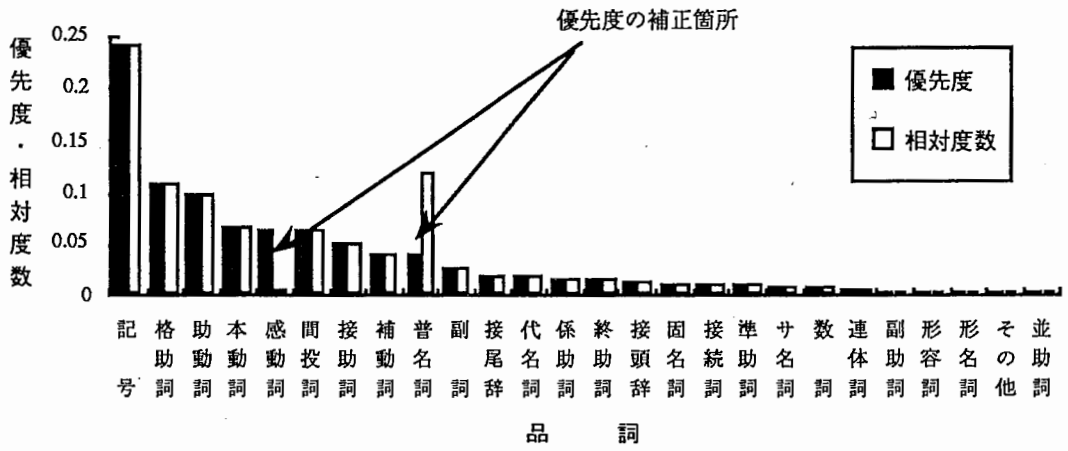


図2-3-1 品詞の相対度数と優先度

† 漢字，カタカナあるいは数字から成る同字種の列を普通名詞として処理する。

また、感動詞と間投詞は字面上同じで、”［］(間投詞の記号)”に囲まれているかどうかにより前者か後者かが決まるので、両者の相対度数を合わせている。

イディオムの優先度は 1 に設定している。

品詞バイグラム B

検査中の単語と前の単語、それぞれの品詞が連接する頻度。

単語頻度　　F

辞書作成用に使用したコーパスにおける単語の頻度。

付属語やイディオムの頻度は1 に設定している。

(2) パラメータの正規化

尤度評価を行う際、4つのパラメータのオーダを合わせる必要がある。本システムでは、0から1の間にパラメータを正規化している。

単純な方法として最大値で割る方法が考えられるが、単語長を例にとった場合、最大単語長で割ると、値の小さな（この場合単語長）ものが出現頻度としては高い割りに、不当に悪く評価されてしまう(図2-2-2 を参照)。バイグラムや単語頻度に対しても同じことが言える。

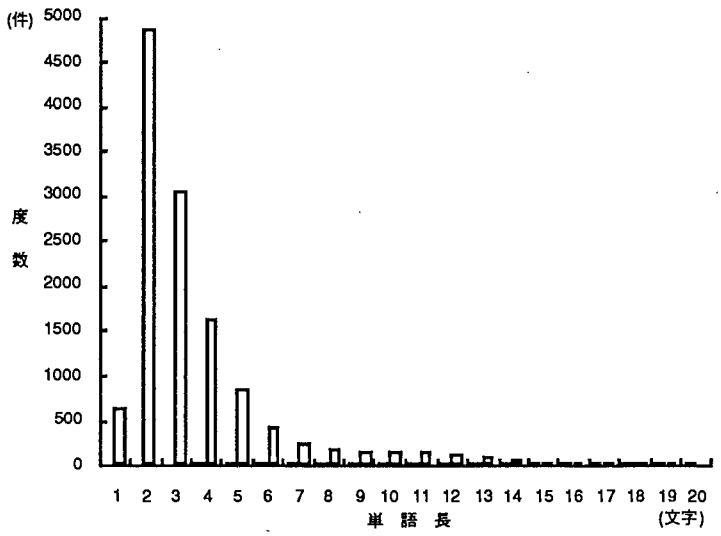


図2-3-2. 単語長の度数ヒストグラム

これを避けるため、本システムではしきい値を設け、パラメータをしきい値で割り，しきい値を越えた場合は1とする、という正規化方法をとっている。

但し、品詞優先度はもとから０～１の数値であるためそのままの数値を使用する。本評価実験での各パラメータのしきい値設定を下表に示す。

表2-3-1. パラメータの正規化のためのしきい値

パラメータ	しきい値	値の範囲
単語長(L)	13	$0,1/13 \leqq L' \leqq 1$
品詞優先度(H)	なし	$0 < H' < 1$
品詞バイグラム(B)	300	$0,1/300 \leqq B' \leqq 1$
単語頻度(F)	50	$0,1/50 \leqq F' \leqq 1$

(3) 尤度計算を行なう評価関数P

尤度は評価関数Pにより計算される。評価関数Pは、単語長，品詞優先度，品詞バイグラム，頻度、４つのパラーメータを持ち、これらを正規化した値（それぞれ、L', H', B', F' とする）に重み付けをして次のように計算される。

$$P(L,H,B,F) = w_1L' + w_2H' + w_3B' + w_4F';$$

重み付けの設定は現在、$w_1 = 7$, $w_2 = w_3 = w_4 = 1$となっている。

3. 評価実験

◆評価項目

A. 形態素解析プログラムの能力評価

解析の速度，形態素分割精度，形態素付与情報精度

B. 形態素辞書構成の評価

辞書ヒット率(付与されている尤度の妥当性)，分野依存率

C. 形態素情報修正作業の評価

作業内容，作業効率

◆評価の対象となる工程

上記の3つの評価項目が対象とする工程を以下の図に示す。

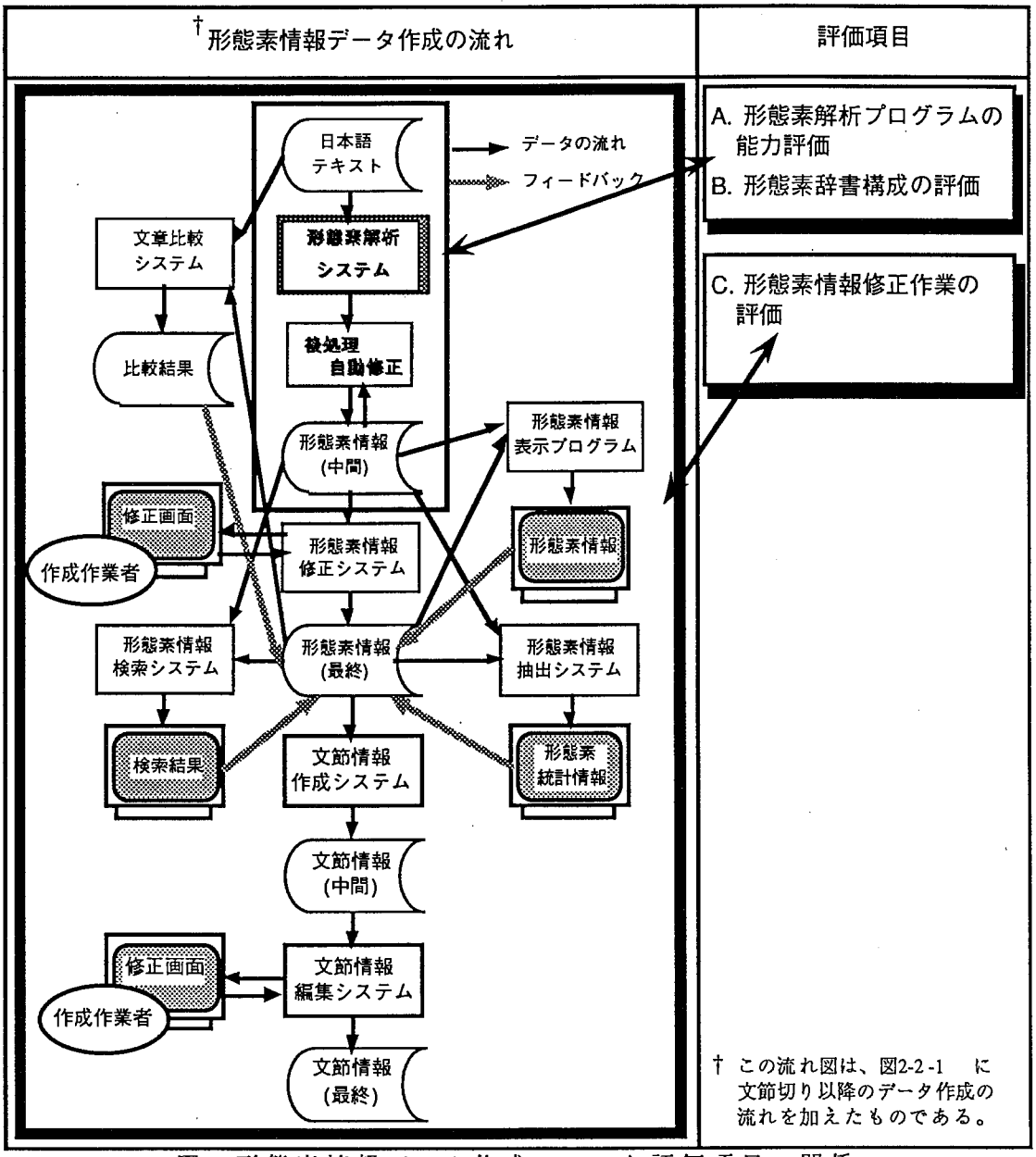


図3 形態素情報データ作成フローと評価項目の関係

◆評価用データ†

評価実験で使用するデータとして、以下のデータを準備する。

○テキスト

2タスク×2メディアの4種類の分野別に分類したテキストを使用する。

大きさは、それぞれ約3，０００語。

表3-1．評価実験用テキストの分類

タスク	メディア	本報告書での呼び方
国際会議	電話	テキストKt
国際会議	キーボード	テキストKk
旅行	電話	テキストRt
旅行	キーボード	テキストRk

○辞書

2タスク×2メディアの4種類の分野の会話テキストを学習データとして、4種類の分野別辞書を作成する。加えて、4種類の分野別辞書をマージして1つの辞書（本報告書では、”総合辞書”と呼ぶことにする）にしたもの、さらに形態素情報データ作成作業用に使用している辞書（”データ作成用大規模辞書”と呼ぶことにする）の2種類を加える。

表3-2．評価実験用辞書の分類

タスク	メディア	本報告書での呼び方
国際会議	電話	分野別辞書Kt
国際会議	キーボード	分野別辞書Kk
旅行	電話	分野別辞書Rt
旅行	キーボード	分野別辞書Rk
国際会議+旅行	電話+キーボード	総合辞書
国際会議+旅行	電話+キーボード	データ作成用大規模辞書

†本評価実験で使用した評価用データ，および学習用データは、添付資料 A-1 で示している。

○テキストと辞書の組み合わせ

実験では、表3-3 に示す9通りの組み合わせでテキストと辞書を使用する。

本報告書中で、実験に使用したテキストと辞書の組み合わせパタンを指す際は、表. 3-3 中に記したパタン名で呼ぶことにする。

表3-3. 評価実験時のテキスト・辞書の組み合わせ

パタン名	テキスト		辞書	
	テキスト分類	単語数	辞書分類	見出し語数
Ktt	テキストKt	3,110	分野別辞書Kt	2,747
Kta	↓	3,110	総合辞書	7,904
Kkk	テキストKk	3,051	分野別辞書Kk	4,763
Kka	↓	3,051	総合辞書	7,904
Rtt	テキストRt	3,454	分野別辞書Rt	4,139
Rta	↓	3,454	総合辞書	7,904
Rkk	テキストRk	3,206	分野別辞書Rk	5,131
Rka	↓	3,206	総合辞書	7,904
RkSD_ALL	↓	3,206	データ作成用大規模辞書	19,665

3-A 形態素解析プログラムの能力評価

3-A-1 解析の速度(文字/s)

(1) 内容

単位時間あたりの平均処理文字数を、テキストや辞書のタスク，メディアの組み合わせパタン別に計測する。計測は、自動修正を行なう場合と自動修正を行なわない場合に分けて行なった。

なお、計測はVAX8800 Ultrix 上で動作する vaxlisp にて行なった。

参考データとして、SUN版での計測結果を示したが、約６～９倍の実行速度であった。

(2) 実験結果

実験の結果を表3-A-2に示す。

１秒あたりの処理文字数は７～１０文字で、平均約９文字であった。１時間あたりに直すと、約 32,400 文字である。

自動修正に要する時間(T2)は、純粋な解析時間(T1)の１０倍以上も掛かった。T1とT2はオーダーが異なるため、自動修正を含めた処理時間(T3)をもとにした処理文字数の比較は行わなかった。

表3-A-1 解析速度実験結果

パタン	文字/T1	単語/T1	文/T1	T1(秒)	T2(秒)	T3(秒)	文字数	単語数	文数	辞書見出し数	ルール数
Ktt	10.30	5.43	0.38	573.05	6740.56	7313.61	5,904	3,110	220	2,747	71
Kta	9.04	4.76	0.34	652.86	7009.81	7662.67	5,904	3,110	220	7,904	71
Kkk	9.50	5.01	0.39	608.56	6782.17	7390.73	5,782	3,051	235	4,763	71
Kka	8.44	4.45	0.34	685.39	6992.50	7677.89	5,782	3,051	235	7,904	71
Rtt	10.02	5.48	0.33	630.68	7793.00	8423.68	6,321	3,454	205	4,139	71
Rta	8.76	4.79	0.28	721.36	7668.08	8389.44	6,321	3,454	205	7,904	71
Rkk	9.61	5.18	0.36	619.48	6935.01	7554.49	5,951	3,206	224	5,131	71
Rka	8.26	4.45	0.31	720.40	7074.89	7795.29	5,951	3,206	224	7,904	71
RkSD_ALL	7.13	3.84	0.27	835.16	6945.07	7780.23	5,951	3,206	224	19,665	71

T1	自動修正を含まない解析時間 (cpu time)。単位は、秒。
T2	自動修正に要する時間。単位は、秒。
T3	T1 + T2。単位は、秒。
文字/T1, 単語/T1, 文/T1	それぞれ、1秒あたりの解析文字数, 単語数, 文数。単位は、個／秒。
ルール数	後処理システムによる自動修正に使用する書き換えルールの数

参考データ　　表.SpracStaion2版での解析速度

	文字/T1	単語/T1	文/T1	T1(秒)	文字数	単語数	文数	辞書見出し数
SUN版	69.96	28.49	10.72	106.35	7,440	3,030	1,140	2,402

(3) 考察

実験結果より、1文字あたりの解析時間は辞書見出し数に対して見かけ上ほぼ比例している(図3-A-1a)。しかし、辞書見出し数の増加が直接の原因となって解析時間が長くなるわけではないので、この関係をもう少し検討してみたい。

解析時間を決定する主な要因として、ア.探索空間をたどる時間的コスト，イ.辞書検索の時間的コスト，ウ. 辞書索引に要する空間的コストなどが挙げられるが、辞書見出し数が増えることにより探索空間，辞書検索空間ともに大きくなるので、辞書見出し数はア，イ，ウの3者すべてに影響する。このようにいくつかの要因が重なり、解析時間は辞書見出し数に対して見かけ上比例して延びていると言える。

次に自動修正については、処理時間はテキストの延べ単語数に比例して増加している。これは順次探索によるルール検索を行なっているためである。自動修正の性能向上は単に検索速度にかかっていると言える。

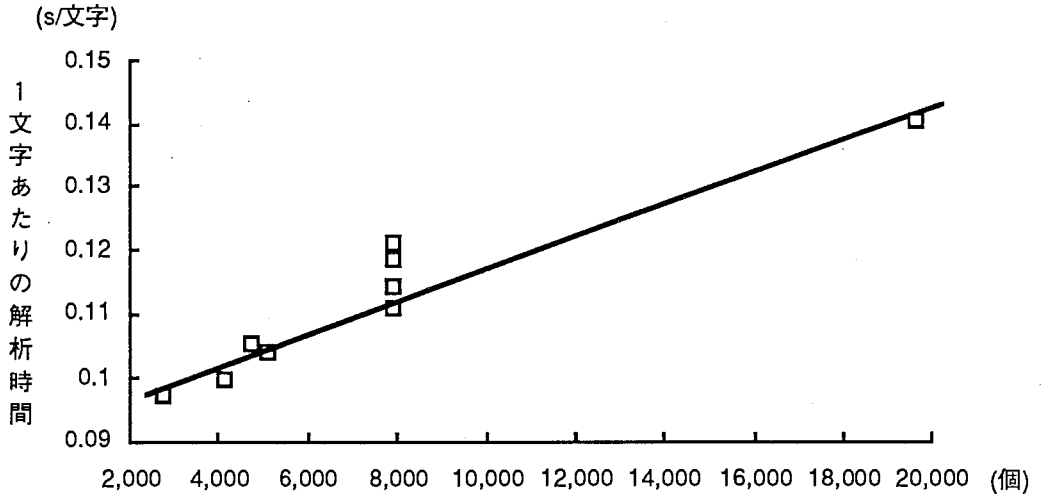


図3-A-1a. 辞書見出しと解析時間の相関関係

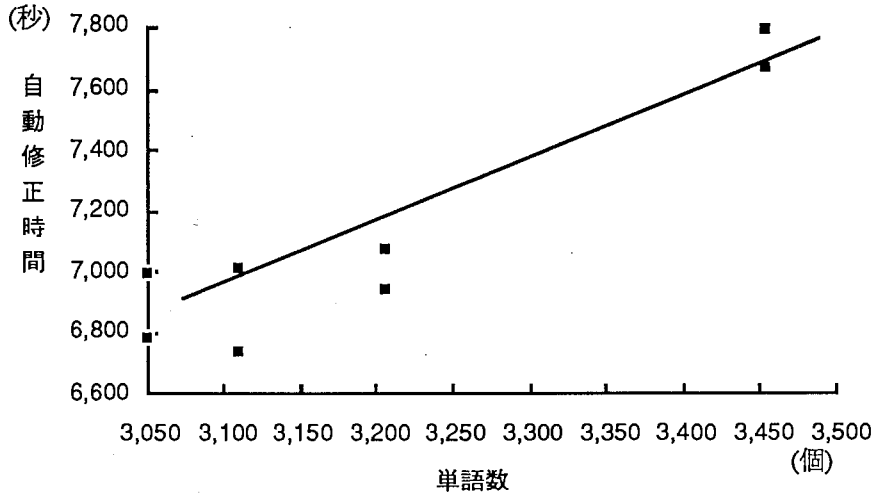


図3-A-1b. 単語数と自動修正時間の相関関係

3-A-2 形態素分割の精度と誤りの傾向

(1) 内容

3-A-1 の解析結果に対して、形態素分割精度の調査を行なう。属性情報の誤りは、この検査では無視する。

分割誤り数の計上方法

下の例によって示されるように、正しい分割をした場合の単語数を基準にして、計上する。

例１．短く切った場合

○解析結果

表記	かな	標準	品詞	活用型	活用形
ん	ん	ん	準助詞	ー ー	ー ー
で	で	で	格助詞	ー ー	ー ー

○正しい分割

表記	かな	標準	品詞	活用型	活用形
んで	んで	んで	接助詞	ー ー	ー ー

○誤り数

１単語あたり、分割誤り１個

例２．長く切った場合

○解析結果

表記	かな	標準	品詞	活用型	活用形
そうです	そうです	そうです	助動詞	ー ー	終止

○正しい分割

表記	かな	標準	品詞	活用型	活用形
そう	そう	そう	副詞	ー ー	ー ー
です	です	です	助動詞	ー ー	終止

○誤り数

２単語あたり、分割誤り２個

(2) 実験結果

実験の結果を表3-A-2a に示す。

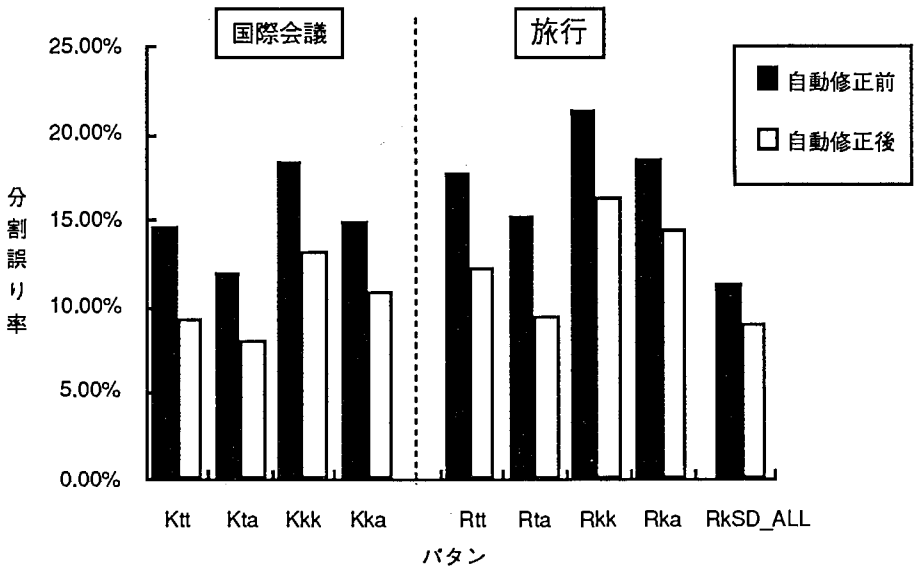
分野別辞書を使用した場合に比べと総合辞書を使用した場合の解析では、解析の精度は、２．５％～３．５％向上している。

また、後処理システムで自動修正を行うことにより、約５％の分割誤りが解消された。

表3-A-2a. 形態素分割精度実験結果

パタン	単語数 (個)	辞書見出し数 (個)	分割誤り率	
			自動修正前	自動修正後
Ktt	3,110	2,747	14.63%	9.26%
Kta	3,110	7,904	11.93%	8.10%
Kkk	3,051	4,763	18.42%	13.18%
Kka	3,051	7,904	14.85%	10.98%
Rtt	3,454	4,139	17.72%	12.33%
Rta	3,454	7,904	15.34%	9.55%
Rkk	3,206	5,131	21.37%	16.28%
Rka	3,206	7,904	18.50%	14.47%
RkSD_ALL	3,206	19,665	11.29%	8.92%

図3-A-2a. 形態素分割精度実験結果



(3) 考察

a. 定量的側面からの考察

総合辞書を使用した解析で見出し数を固定した場合、分割精度の順位は変わらない。テキストに出現する語の分散の度合いによるものと考えられ、次のような予測が成り立つ。

すなわち、分散の度合いが大きい程、未知語が増えて分割精度は低下する、ということである。

これを確認するために、各テキストの異なり単語数を調べてみると、テキストの異なり単語数と誤り率の関係は、表3-A-2b. の様になった。これを相関図であらわしたものが図3-A-2b. である。

表3-A-2b 異なり単語数と誤り率の関係

パタン	異なり単語数	分割誤り率
Ktt	433	14.63%
Kkk	486	17.72%
Rtt	556	18.42%
Rkk	648	21.37%

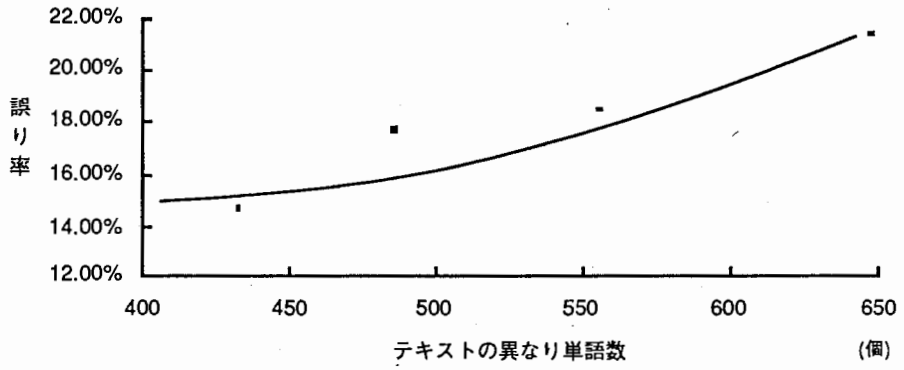


図3-A-2b 異なり単語数と誤り率の関係

b. 定性的側面からの考察 — 誤りの傾向

この考察は、3-A-3. (3) b. でまとめて行なう。

3-A-3 形態素付与の精度と誤りの傾向

(1) 内容

3-A-1 で行なった解析出力に対して、付与されている属性情報の精度の調査を行ない、a. 定量的側面からの考察，b. 定性的側面からの考察（誤りの傾向分析）を行なう。

誤り数の計上方法

属性情報の誤り数は、ア. 単語単位で何個誤りがあるか，イ. 属性情報の項目単位（表記，かな，標準表記，品詞，活用型，活用形の6項目）で何個誤りがあるか、の2通りでカウントする。こうして計上された数を本報告ではそれぞれ、ア. 単語誤り数，イ. 項目誤り数と呼ぶ。

分割誤りの単語は、例3. に示すように、全ての項目が誤りであると見なす。下の例では、袋文字で誤り箇所、下の行に括弧付きで正解示す。

例1.

○解析結果

表記	かな	標準	品詞	活用型	活用形
か	か	か	副助詞	ーー	ーー
			(格助詞)		

○誤り数

ア. 単語誤り数　1

イ. 項目誤り数　1

例2.

○解析結果

表記	かな	標準	品詞	活用型	活用
で	で	だ	助動詞	ーー	連用形
		(で)	(格助詞)	ーー	(ーー)

○誤り数

ア. 単語誤り数　1

イ. 項目誤り数　3

例3.

○解析結果

表記	かな	標準	品詞	活用型	活用
ん	ん	ん	準助詞	ーー	ーー
で	で	で	格助詞	ーー	ーー
(んで	んで	んで	接助詞	ーー	ーー)

○誤り数

ア. 単語誤り数　1

イ. 項目誤り数　6

(2) 実験結果

実験の結果を表3-A-3 に示す。

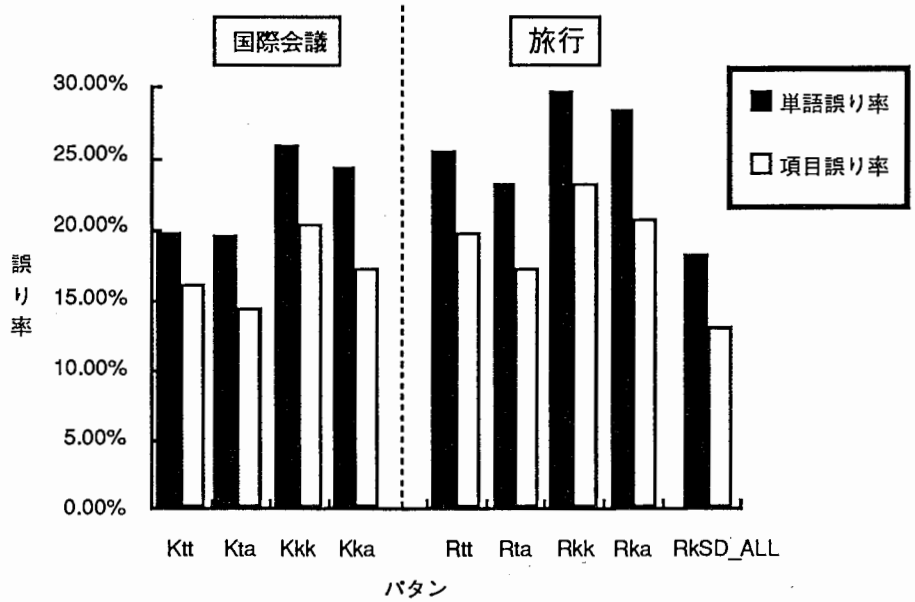
分野別辞書を使用した場合に比べと総合辞書を使用した場合の解析では、解析の精度は、項目誤り率で約１．８～３．１%、単語誤り率で約０．４～２．３%向上した。

また、後処理システムで自動修正を行うことにより、項目誤りも単語誤りも５%前後が解消された。

表3-A-3 形態素付与精度実験結果

パタン	単語数 (個)	辞書見出し数 (個)	単語誤り率		項目誤り率	
			自動修正前	自動修正後	自動修正前	自動修正後
Ktt	3,110	2,747	19.74%	14.82%	15.99%	10.79%
Kta	3,110	7,904	19.36%	13.57%	14.23%	9.71%
Kkk	3,051	4,763	25.83%	21.93%	20.26%	15.21%
Kka	3,051	7,904	24.25%	20.16%	17.17%	13.20%
Rtt	3,454	4,139	25.42%	19.80%	19.68%	14.06%
Rta	3,454	7,904	23.13%	16.21%	17.28%	11.13%
Rkk	3,206	5,131	29.69%	25.95%	23.31%	18.56%
Rka	3,206	7,904	28.42%	24.20%	20.71%	16.73%
RkSD_ALL	3,206	19,665	18.28%	15.13%	13.01%	10.54%

図3-A-3 形態素付与精度実験結果（自動修正前のみ）



(3) 考察

a. 定量的側面からの考察

解析の精度が、辞書の内容とテキストに出現する語彙に依存するのは直感的に明らかであるが、実験 3-A-2，および3-A-3 の結果によく現われている。

b. 定性的側面からの考察 － 誤りの傾向

以下に、解析誤りをその原因により分類した。

［1］未知語

ア. 平仮名表記の単語 ・・・ 解析不能

<説明>

本システムの未知語処理は、漢字または片仮名の同字種かつ１個以上の連続を普通名詞と認定するか数字の１個以上の連続を数詞と認定するかである。平仮名の未知語がある場合、解析不能となり、それまで解析された単語を含め、文中のすべての単語が区切り誤り、単語誤りとなる。

<例>

データ　　「１４日の２泊ということでよろしゅうございますか。」

解析結果　　解析不能

原因　　”よろしゅう”が未知語

イ. 漢字，片仮名，数字 ・・・ 分割は正しくても”読み”が誤りになる

<説明>

上で説明したような未知語処理を行なった場合、認定された語の読みは特定できない。なんらか方法で読みを振ることはできるが本システムでは対応していないので、変わりに表記データと同じ文字を読みとして振っている。このため、未知語処理された漢字，片仮名，数字の列は、例え分割が正しくても、読みが必ず誤りになる。

<例>

データ　　「妻を同行したい」

解析結果

	妻	妻	妻	普通名詞	－－　－－

正解

	妻	つま	妻	普通名詞	－－　－－

原因　　”妻”が未知語処理された。

ウ. 辞書にある別の語（未知語の部分文字列）を選ぶ

<説明>

正しい語が辞書にない場合、その部分文字列を辞書から見つけてくる場合がある。この場合、必ず分割誤りとなる。無論項目誤りでもあるし、単語誤りでもある。

<例>

データ　「お急がせして・・・」

解析結果

急が	いそが	急ぐ	本動詞
せ	せ	す	助動詞

正解

急がせ	いそがせ	急がす	本動詞

原因

”急がす”が未知語で、”急ぐ”＋”する”という部分文字列の解析が成功した。

［2］辞書中の不良データ

<説明>

解析用の辞書辞は、すでに作成された形態素情報ファイルを加工して作成される。形態素情報ファイルに人為的な誤りが存在すれば、その誤りは辞書にも直接反映する。

<例>

データ　「ありがとうございました」

解析結果

ございまし	ございます	ございます	補助動詞	五段	連用

正解

ございまし	ございます	ございます	補助動詞	特殊	連用

原因

辞書の元になる形態素情報ファイルに入力ミスがあったため。

［3］連接テーブルの不備によるもの

<説明>

連接テーブルが、「連接しない語の連接を許している場合」と「連接するの語の連接を禁止している場合」がある。前者の不備による悪影響は多く潜在していると思われるが他のヒューリスティクスによって回避することができる。後者による被害は、特に頻度の高い連接の場合には相当大きくなる。

<例>

データ　　「承知しました」

解析結果

し	し	する	補助動詞	サ変	連用
ま	ま	ま	副詞	ーー	ーー
し	し	為る	本動詞	サ変	連用

正解

し	し	する	補助動詞	サ変	連用
まし	まし	ます	助動詞	ーー	連用

原因

補助動詞「する」と助動詞「ます」の連接が禁止されているため。

［4］固定的な原因で尤度が低くなるもの

ア. 長さ(L)

<説明>

s1,s2という２つの文字列が並んでおり、s1を一つの単語w1として分割したいとする。このときもし、２つの文字列s1s2を結んだような単語w2があっとすると、w1の長さ(L)はw2より短くなる（L(w1)<L(w2)）。

<例>

データ　　「それで結構」

解析結果

それで	それで	それで	接続詞	ーー	ーー

正解

それ	それ	それ	代名詞	－－	－－
で	で	で	格助詞	－－	－－

原因

L("それで") > L("それ")

イ. 品詞優先度(H)

<説明>

品詞優先度(H)は、表記が同じであるときの品詞の選択に影響する。

問題となるのは、

H("本動詞") > H("補助動詞")

H("助動詞") > H("助詞"類)

H("格助詞")>H("接続助詞")>H("終助詞")>H("準体助詞") >H("副助詞") >H("並列助詞")

などである。

<例>

データ　「事務局でございます」

解析結果

で	で	で	格助詞	－－	－－

正解

で	で	だ	助動詞	－－	連用形

原因

H("格助詞") > H("助動詞")

ウ.. 単語頻度(F)

<説明>

表記が同じであるときの品詞の選択に影響する。

問題となるのは同じ表記の語の場合、

F(”サ変名詞”の単語) > F(”普通名詞”の単語) (多くの例で)

など。

しかし、ＡＤＤではサ変名詞は「スル」などの補助動詞の前に現われる時だけであるので品詞バイグラムが有効に働けばこの誤りは回避できるし、自動修正の処理でも回復可能と思われる。(サ変名詞をこう定義しているのはＡＤＤの問題点かも知れない)

<例>

データ　「郵送の方でお願いします」

解析結果

郵送　ゆうそう　郵送　サ変名詞　－－　－－

正解

郵送　ゆうそう　郵送　普通名詞　－－　－－

原因

F("サ変名詞「郵送」") > F("普通名詞「郵送」")

[5] 活用形の解析をしていないことによるもの

<説明>

本システムは、活用語の活用で派生した表記をすべて辞書の見出しとして持っており、活用形の解析は単に辞書見出しとのマッチングにより行なっている。そのため、未然形と連用形，連体形と終止形など表記が同じ場合に活用形を誤る。連接テーブルである程度この誤りを防いでいるが、すべての場合に対処できていない。

現在のインプリメントでは、活用形解析が連接テーブルの内容に依存しているため、その意味で言えば連接テーブルの不備が原因ということになるかも知れない。

<例>

データ　　「テクニカルビジットというのがあった」

解析結果

いう　　いう　　言う　　本動詞　五段　終止

正解

いう　　いう　　言う　　本動詞　五段　連体

原因

終止形と連体形が同じ表記をしているため。連接テーブルの不備が原因ともとれる。

［6］テキストの特殊な仕様によるもの

<説明>

会話をテキスト化するうえでの仕様として、間投詞，言い淀みはそれぞれ" [] ", " () "で囲むことになっている。逆に言うと、" [] ", " () "で囲まれた文字列は、それぞれ必ず間投詞，その他(言い淀みに対する品詞付け)としなければならない。しかし、本システムでは一般的なテキストの入力を対象としており、間投詞，言い淀みのための特殊なルーチンを持たせていないため、" [] ", " () "で囲まれた文字列を通常どおりに最尤度法で形態素解析する。このため、間投詞が感動詞として解析されたり、言い淀みが普通に解析されたりする。

<例>

データ　　「［あっ］」

解析結果

あっ　　あっ　　有る　　本動詞　五段　連用　（つ音）

正解

あっ　　あっ　　あっ　　間投詞　－－　－－

原因

間投詞のための特別な処理をしていないため。

［7］前の語の誤りの影響によるもの

<説明>

前の単語の属性情報の付与を誤ったために、連接テーブルやバイグラムの影響で次の単語の選択を誤ってしまうことがある（付与誤り）。

また、第一単語の分割を誤って短く切った場合、次の単語(第二単語)の分割は必ず誤りとなる。長く切ってしまった場合も、次の単語(第二単語)は分割誤りとしている（分割誤り）。

<例>

データ　「急行します」

解析結果

急行	きゅうこう	急行	普通名詞	ーー	ーー
し	し	為る	本動詞	サ変	連用

正解

急行	きゅうこう	急行	サ変名詞	ーー	ーー
し	し	する	補助動詞	サ変	連用

原因

前の単語「急行」の品詞を誤って普通名詞としたため、
バイグラムが以下のようになった。

B("普通名詞", "本動詞") > B("サ変名詞","補助動詞")

3-B 形態素辞書構成の評価

3-B-1 辞書ヒット率(付与されている尤度の妥当性)

(1) 内容

◆現状の尤度計算 ・・・ 各パラメータに対する重み付け

形態素解析過程で辞書引きされた語は、評価関数Pによって尤度計算され最も尤度の高いものから順に、次の単語の解析を進めていく。

評価関数Pは、単語長，品詞優先度，品詞バイグラム，頻度、4つのパラーメータを持ち、それらを正規化した値（それぞれ、L', H', B', F'とする）に重み付けをして次のように計算されている。（パラメータ、正規化方法等の説明は、2.3.2を参照のこと）

$$P(L,H,B,F) = w_1L' + w_2H' + w_3B' + w_4F';$$

現在の設定は、$w_1 = 7$, $w_2 = w_3 = w_4 = 1$となっているが、この妥当性を検討するため、重み(w1~4)の値を変化させてその影響を考察する。

具体的には、尤度計算時に使用される重みを変えて解析を行ない、その解析の分割精度および属性情報精度を算出する。

◆使用するテキストおよび辞書

使用する入力テキストと辞書の組み合わせパタンは、Kta, Rkaの2通りとする。

パタン名	テキスト	辞書
Kta	テキストKt	総合辞書
Rka	テキストRk	総合辞書

◆調査を試みた重み付けのパターン

重みの組み合わせパターンは、表3-B-1に示す8通りを検討した。6), 7),8) については 1) ～ 5) の結果を考慮し、効果が期待できる値を順次設定した。

表3-B-1. 実験に使用する各種パラメータの重み付けパタン

パタン	w1	w2	w3	w4	特　徴
1)	7	1	1	1	通常使用されているパタン
2)	0	1	1	1	長さなし
3)	0.5	1	0	0	長さ＋品詞重視
4)	0.5	0	1	0	長さ＋バイグラム重視
5)	0.5	0	0	1	長さ＋単語頻度重視
6)	1	0	1	1	品詞なし
7)	5	0	3	2	長さ重視＋品詞なし
8)	8	0	1	1	長さ重視＋品詞なし

(2) 実験結果

実験の結果を表3-B-2 に示す。

Kta, Rka どちらの場合も、通常の形態素解析で使用されているパタン１の精度が最も高かった。バイグラム(B)を重視したパタン４が分割精度の上位を占めている。

長さを用いないパタン２は、付与精度(分割誤りや項目誤り)が最も悪かった。また、長さと品詞優先度のみを用いたパタン３は、総じてパタン２に次いで精度が低く、付与精度が特に悪かった。

パタン６～８では、パタン１～５の結果より有効と思われる重み付けを試行した。これらの重み付けは、以下の推測に従っている。但し、結果的には、パタン１を凌ぐには到らなかった。

［推測１］品詞優先度(H)は使用しない方がよい。　パタン６,７,８

［推測２］バイグラム(B)は重み付けを大きくしたほうがよい。　パタン７

［推測３］長さ(L)の重み付けは大きい程よい。　パタン８

表3-B-2 辞書ヒット率実験結果

パタン	条件					分割誤り		単語誤り		項目誤り		単語数(個)	辞書見出し数(個)
	各パラメータの重み				テキスト辞書								
	L	H	B	F		誤り率	順位	誤り率	順位	誤り率	順位		
1	0.7	0.1	0.1	0.1	Kta	11.93%	2	19.36%	1	14.23%	1	3,110	7,904
2	0	0.5	0.5	0.5	↓	16.75%	8	23.44%	6	18.89%	8	3,110	↓
3	0.5	1	0	0	↓	12.38%	4	27.78%	8	17.12%	7	3,110	↓
4	0.5	0	1	0	↓	11.51%	1	26.46%	7	15.68%	5	3,110	↓
5	0.5	0	0	1	↓	12.89%	7	22.32%	5	15.94%	6	3,110	↓
6	0.5	0	0.5	0.5	↓	12.57%	6	21.51%	4	15.27%	4	3,110	↓
7	0.5	0	0.3	0.2	↓	12.38%	4	21.45%	3	15.11%	3	3,110	↓
8	0.8	0	0.1	0.1	↓	11.93%	2	20.90%	2	14.62%	2	3,110	↓
1	0.7	0.1	0.1	0.1	Rka	18.50%	1	28.42%	1	20.71%	1	3,206	7,904
2	0	0.5	0.5	0.5	↓	21.90%	8	32.00%	2	24.11%	6	3,206	↓
3	0.5	1	0	0	↓	19.84%	6	38.83%	8	24.71%	7	3,206	↓
4	0.5	0	1	0	↓	18.53%	2	37.40%	7	23.73%	5	3,206	↓
5	0.5	0	0	1	↓	20.74%	7	35.53%	6	25.09%	8	3,206	↓
6	0.5	0	0.5	0.5	↓	19.53%	5	33.28%	5	23.42%	4	3,206	↓
7	0.5	0	0.3	0.2	↓	18.90%	4	32.91%	4	22.86%	3	3,206	↓
8	0.8	0	0.1	0.1	↓	18.68%	3	32.53%	3	22.60%	2	3,206	↓

(3) 考察

パタン１、パタン４が分割精度の上位を占めることから、長さ(L)に次いでバイグラム(B)が有力であるとの推測を立て(前頁［推測２］)パタン７を試行したが、推測に反して分割精度は劣化した。このことから、どのパラメータの重み付けを大きくすればよいかを、誤り率だけを見て判断することは出来ないことが分かる。

そこで、実際のデータに対する各重み付けパタンの解析結果を比較して、各パラメータが正解不正解にどのように影響しているかを調べた。表3-B-3 は、その結果を表にしたものである。

表3-B-3 パラメータの重み付けと解析誤りの分析（Ktaの解析結果から）

	データ	正解	解析誤りのパタン	重み付けパタン								パラメータ値比較			
				1	2	3	4	5	6	7	8	L	H	B	F
[1]	（〜の）分	普通名詞（ぶん）	接尾辞（ぶん）	○	○	○	○	×	○	○	○	＝	＞	＞	＜
[2]	でき（ましたら）本動詞		補助動詞	○	○	○	○	×	○	○	○	＝	＞	＝	＜
[3]	（普通名詞）に	格助詞	助動詞「だ」	○	○	○	×	×	×	×	×	＝	＞	網目	網目
[4]	（一緒）に	格助詞	助動詞「だ」	○	○	○	×	×	×	×	×	＝	＞	網目	網目
[5]	（て）おります	補助動詞	本動詞	○	○	×	○	○	○	○	○	＝	＜	＞	＞
[6]	いただく	補助動詞	本動詞	○	○	×	○	○	○	○	○	＝	＜	＞	＞
[7]	承知（いたしました）	サ変名詞	普通名詞	○	○	×	×	○	○	○	○	＝	＜	＝	＞
[8]	郵送	普通名詞	サ変名詞	×	×	○	○	×	×	×	×	＝	＞	＞	＜
[9]	（普通名詞）中	接尾辞	普通名詞「じゅう」	○	○	×	×	○	○	○	○	＝	＜	網目	＞
[10]	ああ	間投詞	間投詞「あ」＋間投詞「あ	○	×	○	○	○	○	○	○	＞	＝	＝	網目
[11]	詳しい	形容詞	形容詞語幹＋本動詞（居る）	○	×	○	○	○	○	○	○	＞	＝	＝	＝
[12]	どうも	副詞	副詞「どう」＋係助詞「も」	○	×	○	○	○	○	○	○	＞	＝	＝	＞
[13]	それで	接続詞	代名詞「それ」＋格助詞	○	×	○	○	○	○	○	○	＞	＜	網目	網目
[14]	あさって	普通名詞	あ＋さ＋って	○	×	○	○	○	○	○	○	＞	＞	網目	＜
[15]	シングルルーム	普通名詞	普通名詞「シングル」＋普通名詞「ルーム」	○	×	○	○	○	×	○	○	＞	＝	＝	＜
[16]	今度	普通名詞	普通名詞「今」＋接尾辞「度」	○	×	○	○	×	×	×	○	＞	＝	＝	＜
[17]	（記号）申し訳ない	形容詞	本動詞「申す」＋普通名詞「訳」＋形容詞「無い」	○	×	×	×	○	×	×	○	＞	＜	＜	＜
[18]	それ＋で	代名詞「それ」＋格助詞	接続詞	×	○	×	×	×	×	×	×	＜	＞	＝	網目
[19]	いつも	副詞	代名詞「いつ」＋係助詞	×	×	○	○	×	×	×	○	＞	＞	＝	＜
[20]	前回	普通名詞	普通名詞「前」＋接尾辞「回」	×	×	○	○	×	×	×	×	＞	＝	＝	＜
[21]	（〜です）とか	接続助詞	接続助詞「と」＋接尾辞「か」	×	×	×	×	○	×	×	×	＞	＜	＜	網目

”正解”，”解析誤りのパタン”欄
それぞれ、正解および各パタンでの誤り方を示す。

”重み付けパタン”欄
各パタンが正解であったかどうかを次の記号で示す。

○（大）：　正解
×（大）：　誤り。誤り方は”解析誤りのパタン”欄
○×（小）：　使用しているパラメータによる比較結果が等しく、事実上（局所的に）最長一致法のアルゴリズムになっていることを示す。

”パラメータ値比較”欄
単語選択の際、正解単語と誤り単語について各パラメータの価を比較した結果を次の記号で示す。

＝　等しい場合。このときパラメータは解析に対して何の効果も及ぼさない。
＞　正解単語の方の価が高い場合で、このパラメータは正解に貢献している。
＜　誤り単語の方の価が高い場合で、このパラメータは悪影響を及ぼしている。
網目　バイグラム(B)では、両者ともしきい値を越えたため等しくなったことを
単語頻度(F)では、両者とも付属語で頻度がデフォルト値（＝１）のため等しくなったことを、それぞれ示す。

表中の各パラメータの価比較と各重み付けパタンでの正解・不正解の関係を分析し、どのパラメータが結果に影響しているかを以下データ毎に示す。

付与誤りのデータ[1]～[9]

[1]（～の）分　　単語頻度(F)の重みが大きいと不正解(パタン５)。

[2]でき（ましたら）　　〃

[3]（普通名詞）に　　品優先度(H)を使用したものだけが正解(パタン１～３)。

[4]（一緒）に　　〃

[5]（て）おります　　品優先度(H)の重みが大きいものは不正解(パタン３)。

[6]いただく　　〃

[7]承知（いたしました）　品優先度(H)の重みが大きいものは不正解(パタン３)
同条件化では、First Hit になる(パタン４)．

[8]郵送　　単語頻度(F)の重みが大きいものは不正解(パタン３，４以外)

[9]（普通名詞）中　　[7] と同じ

分割誤りのデータ[10]～[21]

[10]ああ　　長さ(L)のみ有効。事実上最長一致法(パタン２以外)。

[11]詳しい　　〃

[12]どうも　　長さ(L)を使用していれば正解(パタン２以外)。

[13]それで　　長さ(L)の重みが大きければ正解。
品詞優先度(H)の悪影響は小(パタン３)。

[14]あさって　　長さ(L)の重みが大きいものは正解。
単語頻度(F)の悪影響は小(パタン５)。

[15]シングルルーム　　長さ(L)の重みが大きいものは正解。
単語頻度(F)の重みが大きいものは不正解(パタン6)

[16]今度　　〃

[17]（記号）申し訳ない　長さ(L)の重みが大きいものは正解。
単語頻度(F)の悪影響は小パタン５)。

[18]それ＋で　　長さ(L)を使ったものは不正解。
長さ(L)を使わない場合、First hit で動作し、パタン２では偶然的に正解している。

[19]いつも　　長さ(L)の重みに対する単語頻度(F)の重み比率が大きければ不正解（パタン１，２，５～７）。
パタン１(7対1)が不正解でパタン８(8対1)が正解であることから、境界はこれらの間にあることが言える。

[20]前回　　単語頻度(F)を使ったものは不正解(パタン３，４以外)。

[21]（～です）とか　　品詞優先度(H)またはバイグラム(B)を使用したものは不正解。
使用しなければ、同条件となりFirst Hit (パタン５)。

２１個のデータを分析した結果を見ると、単語頻度が悪影響を及ぼすケースが多いが、この悪影響は他のパラメータ特に長さ(L)の重みを大きくすることで解消できるということが分かる。このことを[19]のデータは良く示している。

品詞優先度(H)は、特に他のパラメータの価が同条件であるときに効力を発揮している([3], [4])。また、品詞優先度(H)だけなら誤りを選んでしまうような場合(Hが"＜"のとき)でも、他のパラメータによって打ち消されることの方が多い([5]～[7],[9],[13])。このことから、品詞優先度(H)は重みは小さくても良いから使用すべきだと言える。

次は、各パラメータの有効性に注目する。

パラメータが好影響を及ぼした場合，悪影響を及ぼした場合，および効果なしの場合の件数をそれぞれ比較すると、表3-B-4のようになる。

最も好影響を与えているパラメータは、長さ(L)，次いで品詞頻度(H)である。逆に悪影響を与えているパラメータは、単語頻度(F)，次いで品詞頻度(H)である。

品詞頻度(H)は、なんらかの効果を及ぼしている件数が１６件で最も多い。同じ品詞でない限り、Hは異なる価を取る（間投詞、感動詞は例外）のに対して、他のパラメータでは †同じ価を取る場合が多く、その場合無効果となってしまう。

特にバイグラム(B)，単語頻度(F)は、頻出する品詞連接や単語において、しきい値を越えてしまい、効力を成さないのは問題である。しきい値設定を見直す必要がある。

推測１～３を振り返ってみると、推測１，２とはまったく反対の結論になった。推測３は、表3-B-4を見るかぎり、否定は出来ない。

表3-B-4 各パラメータの効果の有無(２１データから)

パラメータ	好影響 (件)	悪影響 (件)	無効果 (件)
長さ(L)	10	1	10
品詞優先度(H)	8	8	5
バイグラム(B)	5	2	14
単語頻度(F)	5	9	7

† Lは分割誤りでないかぎり同じ価をとる。Bは、文頭である，同じ連接である，両者共しきい値を越えているなどの理由で、同じ価をとる。Fは、同じ単語である，両者共付属語のため頻度がデフォルト値に設定されている，両者共しきい値を越えているなどの理由で、同じ価をとる。両者共しきい値を越えているというケースが5～6件も見られることから、しきい値の設定に問題があると言わざるを得ない。その他の理由でパラメータが同じ価をとるのは、パラメータそのものの性質である。

以上の考察より、次のようなことが言える。

・長さの有効性

長さ(L)は最も強力なパラメータであり、他のパラメータよりも重み付けを大きくすべきである。

・品詞優先度の利用価値

品詞優先度(H)は意外に利用価値の高いパラメータであり、重みを小さくして使うことで、より利用価値を引き出すことができる。

・かな漢字混じりの文の形態素解析における単語頻度の重要度

単語頻度(F)は５分５分の成績であり、かな漢字混じりの文のように同表記異単語が少ない場合には効力を発しない。日本語かな漢字変換などであれば同音異義語が多く、単語頻度は非常に重要なパラメータとなる。

・バイグラムの正規化と使用上の問題

バイグラム(B)は利用価値はあるかも知れないが、有効に使われていない。これは、しきい値設定やバイグラムの使用法そのものに問題があるためである。しきい値を設定することにより、頻度の高い品詞連接パタンが正規化された際、等値になってしまう。また、バイグラムの使用法の問題は、２単語の品詞連接頻度を比較していく場合、第１単語の品詞の頻度で割っていない点である。このことにより、品詞の頻度はここでも効いてしまうことになる。おまけに、品詞優先度の調整（2-3-2参照）はこの処理によって打ち消されることになりかねず、非常に問題である。

・妥当な重み付け

パラメータの重み付けは、下に示す比ぐらいが適当であろう。

w1：w2：w3：w4 = 8：0.6：1：0.4 (w2 < w4)

・最尤探索に対する提言 ― 正規化の功罰とパラメータの独立性

各パラメータは評価関数によってマージされ、非常に不透明な使われ方をする。つまり、どのパラメータが何に効いているのかが分からなくなる。これは、しきい値によって強制的に数値のオーダーを揃える正規化演算のためである。

各パラメータ独自で評価を行ない、その結果の数値で並べ替えを行ないそれぞれの順位をなんらかの方法で総合すれば、オーダーを揃える必要はないうえ、パラメータの独立性を保つことができる。

尤度計算は、言語処理の様々なフェーズで行なわれるが、パラメータの独立性を保つことで最尤探索の制御は扱いやすいものとなり、より効果的な利用法が生み出されるであろう。

3-B-2 分野依存率

(1) 内容

分野により単語選択がどのように異なるかを調査する。

以下に挙げる4つの角度(イ,ロ,ハ,ニ)から辞書を比較し、a. 見出し語の頻度が著しく異なる語、b. 一方にのみ出現する語、の2点を指標として分野依存度をみる。

1) メディア依存度

イ. 「国際会議に関する問い合わせ」におけるメディア依存度

ロ. 「旅行代理店への問い合わせ」におけるメディア依存度

2) タスク依存度

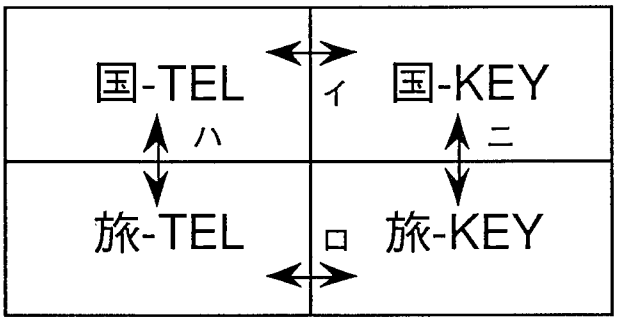
ハ. 「電話会話」におけるタスク依存度

ニ. 「キーボード会話」におけるタスク依存度

イ ～ ニの依存度は以下の図に示す比較パタンにより調査する。

但し、国-TELは、「国際会議に関する問い合わせ」をタスクとし、電話をメディアとする会話のテキスト文から抽出された辞書を示している。その他、国-KEY,旅-TEL,旅-KEYも同様に表している。

表3-B-5. 辞書比較のパタン



(2) 調査結果およびその考察

1) メディア依存度

イ. 「国際会議に関する問い合わせ」におけるメディア依存度

・・・ 電話会話(国際会議)辞書とキーボード(国際会議)辞書の比較

【調査結果】

a. 見出し語の頻度が著しく異なる語

電話会話(国際会議)辞書とキーボード(国際会議)辞書において、見出し語の相対度数が、著しく異なり，かつ見出し語の度数が比較的大きいものを抽出した。表3-B-? に示す。

表3-B-イ(a) 頻度が著しく異なる語 - 品詞別(国際会議)

x: 電話会話／y:キーボード会話

品詞	単語	度数x	相対度数x	度数y	相対度数y	Larger	相対度数比(大/小)
普通名詞	こと	136	1.03%	542	2.91%	Y	2.83
	会議	19	0.14%	182	0.98%	Y	7.00
	方	116	0.88%	2	0.01%	X	88.00
	講演	27	0.20%	120	0.64%	Y	3.20
	会場	4	0.03%	112	0.60%	Y	20.00
	先生	7	0.05%	76	0.41%	Y	8.20
	時間	9	0.07%	76	0.41%	Y	5.86
	ファックス	17	0.13%	66	0.35%	Y	2.69
	前	2	0.02%	60	0.32%	Y	16.00
	キャンセル	10	0.08%	60	0.32%	Y	4.00
代名詞	私	5	0.04%	133	0.71%	Y	17.75
	わたくし	56	0.42%	11	0.06%	X	7.00
	これ	16	0.12%	66	0.35%	Y	2.92
副詞	どう	15	0.11%	70	0.38%	Y	3.45
接続詞	では	46	0.35%	208	1.12%	Y	3.20
	で	87	0.66%	24	0.13%	X	5.08
	じゃあ	70	0.53%	24	0.13%	X	4.08
接尾辞	様	74	0.56%	28	0.15%	X	3.73
	さん	7	0.05%	85	0.46%	Y	9.20
	時	8	0.06%	79	0.42%	Y	7.00
	円	55	0.42%	17	0.09%	X	4.67
	回	12	0.09%	58	0.31%	Y	3.44
感動詞	はい	1718	13.01%	495	2.66%	X	4.89
	ええ	280	2.12%	108	0.58%	X	3.66
	あ	1	0.01%	84	0.45%	Y	45.00
連体詞	もしもし	10	0.08%	60	0.32%	Y	4.00
本動詞	有る	33	0.25%	137	0.74%	Y	2.96
	畏まる	64	0.48%	10	0.05%	X	9.60
	書く	1	0.01%	63	0.34%	Y	34.00
	待つ	1	0.01%	45	0.24%	Y	24.00
形容詞	良い	20	0.15%	148	0.79%	Y	5.27
	無い	21	0.16%	87	0.47%	Y	2.94
補助動詞	いただく	1	0.01%	109	0.59%	Y	59.00
	いただける	1	0.01%	68	0.37%	Y	37.00
	なる	12	0.09%	68	0.37%	Y	4.11
	みる	7	0.05%	35	0.19%	Y	3.80

x: 電話会話。度数xは、電話会話辞書の見出し語数示す。
相対度数xは、総単語数に対する度数xの比を示す。

y: キーボード会話。以下、上と同様。

相対度数比: 相対度数x，yのうち、小さい方の数に対する大きい方の数の比を示す。

Larger 相対度数x，yの大きい方を指す。

以降、表3-B-ロ(a)，3-B-ハ(a)，3-B-ニ(a) でも同様の形式を使用する。

b．一方にのみ出現する語

表3-B-イ(b)[TEL] 電話会話のみに出現する語
－ 品詞別5件以上(国際会議)

間投詞		普名詞		固名詞		感動詞	
あのー	695	論文	33	武蔵野電気 ｼｽﾃﾑ開発研究所	26	ごめんください	20
あっ	248	セカンド アナウンスメント	27	鈴木	22	おそれいります	17
ああ	160	ファーストアナウンスメント	27	三鷹市	17		
えー	129	会員	23	鈴木一郎	17	本動詞	
あの	118	アブストラクト	22	小金沢	15	決る	17
あ	55	締切り	19	東京工大	12	伺う	14
ちょっと	34	休み	11	井の頭	11	構う	10
えーと	25	さき	10	佐藤恵子	11	行う	9
そうですね	20	3月	8	中村啓子	9	持つ	7
はあ	18	ふう	8	アイシーシー	8	知る	7
ああー	12	現在	8	月刊コンピューター	8	受付ける	5
ま	11	１７日	7	武蔵野	8	入れる	5
そのー	10	テクニカルビジット	7	サンハイツ	7		
えーとですね	9	消印	7	井の頭	7	サ名詞	
え	8	郎	7	大岡山	7	同封	11
えーっと	8	応募	6	原田	6	コピー	7
もう	7	初め	6	はらだ	5	応募	6
えええとですね	6	審査	6	赤塚	5	副　詞	
はあはあ	6	振替	5	太郎	5	早速	9
まあ	6	中身	5	板橋区	5	いつも	7
えええと	5	土曜	5				
		理学	5			代名詞	
						あたくし	8

表3-B-イ(b)[KEY] キーボード会話のみに出現する語
－ 品詞別5件以上(国際会議)

補動詞		間	22	打合せ	7	請求	5
いただく	109	手配	22	下	6	宣伝	5
いただける	68	席	22	月日	6	対向	5
まいる	6	発	22	生年	6	病気	5
あう	6	テクニカルツアー	20	専門	6	払い戻し	5
		ページ	20	速達	6	方々	5
普名詞		側	20	大勢	6	理事	5
スライド	66	イー	18	まま	5	本動詞	
ビザ	58	ｵﾌｨｼｬﾙ･ｲﾝﾋﾞﾃｰｼｮﾝ	18	アール	5	書く	63
演者	42	パンフレット	16	ガイド	5	待つ	45
演題	36	外務	16	スケジュール	5	話す	30
議長	34	機会	16	スペル	5	振込む	21
時	34	件	16	バス	5	取れる	11
手紙	32	見学	16	ブース	5	会う	11
ＯＨＰ	30	後	16	プロジェクター	5	渡す	11
事	30	省	16	レディースプログラム	5	降りる	10
セッション	28	心配	16	開会	5	映す	9
応答	26	人	16	外国	5	支払う	9
今	26	担当	16	楽しみ	5	知れる	7
質疑	26	昼	16	帰り	5	関する	6
上	26	オペレーター	7	休憩	5	代る	6
ポスター	24	データショー	7	業者	5	置く	6
演台	24	マイク	7	指示	5	持つ	5
ほう	22	ミーティング	7	式	5	映る	5
スクリーン	22	機材	7	終了	5	座る	5
間	22	切符	7	振り込み	5	存じる	5

(次頁に続く)

形名詞		名古屋ヒルトン	8	頃	7	推薦	5
残念	28	グリーン教授	7	方	5	配布	5
急	10	ロバートソン	7				
自由	10	成田	7	形容詞			
だいじょうぶ	8	クーパー	6	細い	16		
十分	8	クリストファーソン教授	6	大きい	16		
個人的	6	ブラウン教授	6	相応しい	8		
大切	6	モスクワ	6	近い	6		
		仁科先生	6	広い	6		
感動詞		野原	6				
ああ	25	スペンサー	5	代名詞			
あの	22	フーバー教授	5	彼女	12		
えー	20	ペンシルバニア	5				
あのー	12	ホプキンス教授	5	副　詞			
ええと	8	マクガバン	5	つまり	8		
は	5	坂口	5	よく	8		
		山本理事	5	少々	7		
固名詞		接尾辞		直接	6		
田端	18	証	16	是非	5		
保刈	15	状	14				
岩本	10	分	12	サ名詞			
大阪	10	ヶ	10	登録	8		
ＪＴＢ	9	階	10	紹介	7		
ＡＴＲ	8	時間	10	セット	6		
ワルシャワ	8	さ	9	申請	6		
山本	8	メートル	9	案内	5		

【考察】

a. 見出し語の頻度が著しく異なる語より

補助動詞の頻度差に特徴的な違いが出ている。キーボード会話は書面語の表現を比較的多くとるため、発話者にもよるが、特に丁寧な言い方をする補助動詞がよく使われる。

キーボード会話の書面語的な特徴を表すその他の例としては、「では」と「じゃあ」が挙げられる。どちらも 0件ではなく、口語と書面語の間で揺れている様子が現われており、しかも電話会話では口語的な方に偏り，キーボード会話では書面語的な方に偏るという、特徴が出ている。

	TEL:件数	KEY:件数	比（大/小）
「では」	TEL:0.35	< KEY:1.12	3.2
「じゃあ」	TEL:0.53	> KEY:0.13	4.1

その他の特徴としては、一人称を電話会話では「わたくし」，キーボード会話では「私」ということが多い点があった。

b. 一方にのみ出現する語より

メディアの依存度が最も顕著に現われている品詞は間投詞である。電話会話では、相手の発話内容の理解と自分自身の発話内容の思索が並列的に行なわれるために、間投詞が出現する。

また、キーボード会話は電話会話に比べ語彙の種類が多い、これは普通名詞，固有名詞，形容詞，本動詞などの内容語に良く現われている。トピックの設定による影響とも考えられるが、今回の調査ではトピックの違いを考慮した実験データ選定をしていないので、上述以上のことは言えない。

ロ．「旅行代理店への問い合わせ」におけるメディア依存度
・・・　電話会話(旅行)辞書とキーボード(旅行)辞書の比較

【調査結果】

a. 見出し語の頻度が著しく異なる語

電話会話(旅行)辞書とキーボード(旅行)辞書において、見出し語の相対度数が、著しく異なり，かつ見出し語の度数が比較的大きいものを抽出した。表3-Bロ(a) に示す。

表3-B-ロ(a) 頻度が著しく異なる語 - 品詞別(旅行)

x: 電話会話／y:キーボード会話

品詞	単語	度数x	相対度数x	度数y	相対度数y	Larger	相対度数比(大/小)
普通名詞	方	152	1.06%	5	0.03%	X	35.33
	ツアー	4	0.03%	81	0.53%	Y	17.67
	もの	56	0.39%	1	0.01%	X	39.00
	いま	51	0.35%	14	0.09%	X	3.89
	形	43	0.30%	9	0.06%	X	5.00
	だいたい	34	0.24%	10	0.07%	X	3.43
	明日	1	0.01%	35	0.23%	Y	23.00
	申込	2	0.01%	34	0.22%	Y	22.00
	ほか	13	0.09%	32	0.21%	Y	2.33
代名詞	私	12	0.08%	94	0.62%	Y	7.75
	どちら	11	0.08%	33	0.22%	Y	2.75
	そちら	7	0.05%	29	0.19%	Y	3.80
	わたくし	24	0.17%	2	0.01%	X	17.00
サ変名詞	失礼	23	0.16%	57	0.37%	Y	2.31
形容名詞	同じ	10	0.07%	25	0.16%	Y	2.29
副詞	たとえば	31	0.22%	4	0.03%	X	7.33
	いかが	10	0.07%	30	0.20%	Y	2.86
	いくら	11	0.08%	30	0.20%	Y	2.50
接尾辞	さん	24	0.17%	73	0.48%	Y	2.82
	日	56	0.39%	1	0.01%	X	39.00
	人	36	0.25%	2	0.01%	X	25.00
感動詞	はい	950	6.60%	373	2.45%	X	2.69
	ええ	450	3.13%	52	0.34%	X	9.21
	うん	56	0.39%	1	0.01%	X	39.00
本動詞	為る	1	0.01%	168	1.10%	Y	110.00
	分る	27	0.19%	71	0.47%	Y	2.47
	願う	64	0.44%	1	0.01%	X	44.00
	送る	16	0.11%	52	0.34%	Y	3.09
	伺う	38	0.26%	2	0.01%	X	26.00
	持つ	13	0.09%	33	0.22%	Y	2.44
	致す	15	0.10%	32	0.21%	Y	2.10
補助動詞	いる	37	0.26%	108	0.71%	Y	2.73
	おく	12	0.08%	30	0.20%	Y	2.50

表の見方は、表3-B-イ(a) の下を参照のこと。

b．一方にのみ出現する語

表3-Bロ(b)[TEL] 電話会話のみに出現する語
－ 品詞別５件以上(旅行代理店)

間投詞		えと	11	まま	8	金沢	6
あのー	472	えーとですね	10	年末	8	第３営業部	6
あの	392	そうですねー	10	夜行	8	藤枝	6
ああ	170	なんか	10	２月	7	オアフ島	5
えー	153	はあはあ	10	ブルートレイン	7	ロスアンゼルス	5
あっ	151	ん	9	会	7	河口湖	5
あ	69	うー	8	春	7	中村	5
えーと	69	っと	7	夫婦	7	東京駅	5
まあ	69	ふーん	7	忘年	7	熱海	5
あー	61	えっと	6	連休	7		
そうですね	57	ほう	6	シャトルバス	6	副　詞	
うん	53	はあはあはあ	5	車中	6	なるほど	8
その	51			流氷	6	そっ	6
ちょっと	44	本動詞		バラドール	5	おそらく	5
ああー	42	願う	64	フルムーン	5	とにかく	5
んー	38	伺う	38	冬	5	全然	5
ええ	35	入る	18	母	5		
ま	33	着く	14			サ名詞	
うーん	32	聞く	14	固名詞		承知	7
はあ	20	変る	9	ジェーティービー	14		
もう	20	組立てる	8	渋谷支店	11	形名詞	
え	17	振込む	7	ワイキキ	10	デラックス	6
こう	14	合う	5	近鉄	10		
そのー	14			マウイ島	9	形容詞	
はあー	14	普名詞		十兵衛	9	寒い	5
はい	13	コンドミニアム	22	マドリッド	7		
そう	12	１月	13	小金沢篤子	7		
ふんふん	12	１１月	12	ハワイ	6		
えっ	11	むこう	9	海外旅行	6		

表3-B-ロ(b)[KEY] キーボード会話のみに出現する語
— 品詞別5件以上(旅行代理店)

普名詞		１３日	5	白樺湖	11	形名詞	
ほう	58	１８日	5	ＪＴＢ	10	フリー	10
７月	36	２２日	5	北軽井沢	10	自由	5
社	26	２７日	5	ゴールドコースト	9		
件	24	テニス	5	シンガポール	9	本動詞	
２０日	21	テニスコート	5	安曇野	9	休む	9
先日	21	夏	5	広島	8	入る	8
保険	20	夏休み	5	ＮＷ	6	為す	7
今	19	現金	5	ラマダ	6	戴く	6
写真	17	事項	5	残波	6	寄る	6
宅	15	手段	5	柴田	6	過す	5
ビザ	14	直通	5	順子	6	見付ける	5
色々	12	病院	5	瀬戸大橋	6	乗る	5
コテージ	11	本日	5	穂高	6		
手数	11	迷惑	5	オーストラリア	5	副　詞	
ゴルフ	10	問い合わせ	5	タヒチ	5	ただいま	8
前後	10	来週	5	テッド	5	先ほど	7
近く	9			フランス	5	もしも	6
在宅	9	感動詞		ベルリン	5	全く	5
振込	9	あ	36	メルボルン	5		
他	9	ああ	26	奥の細道	5	形容詞	
存知	8	あの	17	熊野本宮	5	早い	8
変更	8	ええと	9	長谷川	5		
用紙	8	まあ	9			サ名詞	
トラベラーズチェ	7	さようなら	7	連体詞		到着	7
支払い	7	ごめんなさい	5	当	23	変更	6
度	7					周遊	5
補償	7	固名詞		接尾辞		同封	5
牧場	7	新宿支店	25	頃	18		
いっしょ	6	日本交通公社	18	カ	12	代名詞	
学生	6	橋上	17	中	12	あなた	6
座禅	6	ソウル	13	センチ	11	何	5
最終	6	ＪＲ	12	等	7		
民宿	6	与論島	12	付き	6	補動詞	
旅程	6	静内	11	語	5	なれる	5

【考察】

イ．「国際会議に関する問い合わせ」におけるメディア依存度で考察した内容のうち、次の点が「旅行問い合わせ」におけるメディア依存度でも言える。

・間投詞が電話会話に多い

・内容語の種類がキーボード会話に多い

・一人称の違い。（電話会話では「わたくし」、キーボード会話では「私」が多い）

但し、丁寧な言い方をする補助動詞の差が見られなかった。

2) タスク依存度

ハ. 「電話会話」におけるタスク依存度

・・・ 国際会議(電話会話)辞書と旅行(電話会話)辞書の比較

【抽出結果】

a. 見出し語の頻度が著しく異なる語

国際会議(電話会話)辞書と旅行(電話会話)辞書において、見出し語の相対度数が、著しく異なり，かつ見出し語の度数が比較的大きいものを抽出した。表3-B-ハ(a) に示す。

表3-B-ハ(a) 頻度が著しく異なる語 - 品詞別(電話会話)

x: 国際会議／y:旅行

品詞	単語	度数x	相対度数x	度数y	相対度数y	Larger	相対度数比(大/小)
普通名詞	方	116	0.88%	1	0.01%	X	88.00
	１５日	79	0.60%	2	0.01%	X	60.00
	あと	23	0.17%	85	0.59%	Y	3.47
	参加	54	0.41%	6	0.04%	X	10.25
	ホテル	10	0.08%	53	0.37%	Y	4.63
	今日	44	0.33%	3	0.02%	X	16.50
	形	10	0.08%	43	0.30%	Y	3.75
代名詞	こちら	133	1.01%	58	0.40%	X	2.53
	わたくし	56	0.42%	24	0.17%	X	2.47
	これ	16	0.12%	58	0.40%	Y	3.33
サ変名詞	失礼	75	0.57%	23	0.16%	X	3.56
副詞	よろしく	101	0.76%	21	0.15%	X	5.07
	少し	4	0.03%	34	0.24%	Y	8.00
接続詞	それで	70	0.53%	28	0.19%	X	2.79
	それでは	68	0.51%	3	0.02%	X	25.50
	じゃ	8	0.06%	45	0.31%	Y	5.17
接頭辞	ご	82	0.62%	201	1.40%	Y	2.26
接尾辞	書	50	0.38%	3	0.02%	X	19.00
	時	8	0.06%	37	0.26%	Y	4.33
間投詞	あの	118	0.89%	392	2.72%	Y	3.06
	まあ	6	0.05%	69	0.48%	Y	9.60
	あー	4	0.03%	61	0.42%	Y	14.00
	うん	1	0.01%	53	0.37%	Y	37.00
	その	4	0.03%	51	0.35%	Y	11.67
感動詞	はい	1718	13.01%	950	6.60%	X	1.97
	うん	3	0.02%	56	0.39%	Y	19.50
本動詞	願う	153	1.16%	64	0.44%	X	2.64
	送る	141	1.07%	16	0.11%	X	9.73
	為る	119	0.90%	1	0.01%	X	90.00
	分る	71	0.54%	27	0.19%	X	2.84
	申す	64	0.48%	21	0.15%	X	3.20
	畏まる	64	0.48%	18	0.13%	X	3.69
	待つ	1	0.01%	34	0.24%	Y	24.00
補助動詞	いたす	229	1.73%	88	0.61%	X	2.84
	いただく	1	0.01%	118	0.82%	Y	82.00
	いただける	1	0.01%	47	0.33%	Y	33.00
形容詞	良い	20	0.15%	77	0.53%	Y	3.53
	安い	5	0.04%	34	0.24%	Y	6.00

表の見方は、表3-B-イ(a) の下を参照のこと。

b．一方にのみ出現する語

表3-B-ハ(b)[K]「国際会議」のみに出現する語
－ 品詞別5件以上(電話会話)

普名詞		原稿	9	第一開発部	21	接尾辞	
登録	55	プログラム	8	三鷹市	17	部	16
発表	36	ポスターセッション	8	鈴木一郎	17	中	14
論文	33	委員	8	東京工業大学	15	語	7
１６日	32	１７日	7	斉藤先生	14		
セカンドアナウンスメント	27	テクニカルビジット	7	東京工大	12	形容詞	
ファーストアナウンスメント	27	控え	7	井の頭	11	早い	15
局	27	消印	7	佐藤恵子	11		
講演	27	先生	7	東京大学	10	サ名詞	
７月	25	郎	7	名古屋観光ホテル	10	同封	11
用紙	25	イベント	6	中村先生	9	拝見	8
会員	23	扱い	6	アイシーシー	8	コピー	7
アブストラクト	22	応募	6	月刊コンピュータ	8	発表	7
工学	20	期限	6	日下部	8	応募	6
会議	19	子	6	武蔵野	8	出席	6
聴講	19	初め	6	サンハイツ	7		
締切り	19	書面	6	井の頭	7	本動詞	
ファックス	17	招待	6	斉藤	7	構う	10
振込	15	審査	6	大岡山	7	繰返す	9
早期	15	送金	6	原田	6	行う	9
手数	14	欄	6	東工大	6	戴く	8
研究	13	つもり	5	目黒区	6	受付ける	5
１３日	12	カード	5	はらだ	5	遅れる	5
証明	12	国際	5	英	5		
所属	11	振替	5	赤塚	5	副　詞	
情報	11	中身	5	太郎	5	必ず	6
さき	10	理学	5	東大	5		
科	10			板橋区	5	問投詞	
学生	10	固名詞				ええとですね	6
大学	10	コンピュータ 国際会議事務局	48	感動詞		ええと	5
封筒	10	コンピュータ 国際会議	30	ごめんください	20		
コピー	9	武蔵野電気システム開発研究所	26	おそれいります	17	接頭辞	
ナンバー	9	佐藤	22			不	5
バンケット	9	鈴木	22				

表3-B-ハ(b)[R]「旅行代理店」のみに出現する語
－ 品詞別５件以上(電話会話)

補助詞		むこう	9	車中	6	ほう	6
いただく	118	ゴールデンウィー	9	席	6	はあはあはあ	5
いただける	47	飛行	9	泊	6	ふうん	5
まいる	13	まま	8	流氷	6		
		気候	8	とこ	5	本動詞	
普名詞		航空	8	パラドール	5	待つ	34
コース	75	年末	8	フルムーン	5	出掛ける	23
出発	55	平日	8	企画	5	取れる	22
旅行	43	夜行	8	歳	5	違う	19
パンフレット	39	２月	7	自由	5	書く	16
バス	32	２日	7	女性	5	知れる	14
客	29	ブルートレイン	7	待ち	5	組立てる	8
予約	24	以上	7	中	5	振込む	7
コンドミニアム	22	往復	7	昼	5	合う	5
現地	17	温泉	7	冬	5		
時	17	春	7	馬	5	副　詞	
パッケージツアー	15	程度	7	母	5	たとえば	31
１月	13	添乗	7	友達	5	なかなか	14
３日	13	発	7	旅	5	だいぶ	12
スキー	13	夫婦	7			ほとんど	10
代金	13	忘年	7	間投詞		多少	9
方面	12	連休	7	んー	38	かなり	7
時期	11	１１日	6	こう	14	さきほど	7
食事	11	１日	6	そう	12	結構	7
正月	11	オプショナルツア	6	ふんふん	12	そっ	6
利用	11	シャトルバス	6	えと	11	たぶん	6
交通	10	パック	6	ん	9	あまり	5
場所	10	感じ	6	うー	8	あんまり	5
２９日	9	帰り	6	ふーん	7	つまり	5
５月	9	行動	6	えっと	6		

表3-B-ハ(b)[R]「旅行代理店」のみに出現する語
－ 品詞別５件以上(電話会話)

つづき

サ名詞		高木	14	金沢	6	少い	5
案内	17	渋谷支店	11	第３営業部	6	新しい	5
利用	10	ワイキキ	10	鳥羽	6	短い	5
出発	9	近鉄	10	藤枝	6		
		交通公社	10	万座	6	形名詞	
接尾辞		マウイ島	9	オアフ島	5	デラックス	6
歳	16	十兵衛	9	スペイン	5	値段的	5
機	11	杉並区	8	ロスアンゼルス	5		
代	11	奈良	8	河口湖	5	接頭辞	
時間	8	バリ島	7	熱海	5	いく	6
さま	7	マドリッド	7				
ページ	5	ハワイ	6	形容詞		接続詞	
		伊勢	6	近い	12	だから	5
固名詞		伊豆	6	長い	7		
ヨーロッパ	15	沖縄	6	きつい	6	代名詞	
ジェーティービー	14	海外旅行	6	寒い	5	おたく	5

【考察】

b. 一方にのみ出現する語より

タスクの違いが顕著に現われているものは、固有名詞の種類であった。「旅行」の場合，地名（特に海外）が多く、「国際会議」の場合，大学名や人名が多い。この違いは、それぞれタスクの中心的な話題であるという理由による。これは話題の焦点の違いに関係なく成り立つことが予想される。

その他、各タスクにおいて話題の中心となるような名詞としては、それぞれのタスクに次のような単語が集中した。

「国際会議」

発表，論文，講演，工学，聴講，研究，コンピュータ会議事務局，

「旅行」

コース，旅行，パンフレット，コンドミニアム，現地，パッケージ

形容詞，副詞などの修飾語は、「旅行」の場合に非常に多く出現し、種類も電話に比べれば豊富だと言える。特に、”近い”，”遠い”，”寒い”などはタスクによる違いだろう。

a. 見出し語の頻度が著しく異なる語より

「旅行」では、間投詞が頻度，種類共に多い。「旅行」と「国際会議」ではタスクの広さが違う（辞書の規模からすると、国際会議：7,510語/旅行：9,270語）。このため「国際会議」に比べ「旅行」では、話者（特に模擬会話実験の被験者）の”考える”対象が多くなり、その影響で「旅行」では間投詞が増えたのではないだろうか†。

† 「旅行」と「国際会議」のタスクの性質を調べると、次のようなことが分かった。

・1会話に含まれる内容の豊富さ

「旅行」，「国際会議」それぞれ約6,000語に対する会話数を調べてみると、「旅行」6会話に対して「国際会議」19会話で、1会話あたりの単語数にすると、「旅行」は「国際会議」の3倍になる。この数は、単に会話の長さを示すだけでなく、1会話の内容の豊富さを表している。

・模擬会話の被験者の一般的傾向

被験者にとって「旅行」と「国際会議」では、「旅行」の方が卑近で想像しやすく、現実にも「旅行」の会話を経験している可能性も高い。このため、被験者にとっては「旅行」の方が考えることが多いし、また考えやすい。

これらを通して、話者が”考える”ことと間投詞の関係は、あくまで推測の域であるが、データから見ても全く関係がないとはいえない。

ニ．「キーボード会話」におけるタスク依存度

・・・　国際会議(キーボード会話)辞書と旅行(キーボード会話)辞書の比較

【抽出結果】

a．見出し語の頻度が著しく異なる語

国際会議(キーボード)辞書と旅行(キーボード)辞書において、見出し語の相対度数が、著しく異なり，かつ見出し語の度数が比較的大きいものを抽出した。表3-B-ニ(a) に示す。

表3-B-ニ(a) 頻度が著しく異なる語 - 品詞別(キーボード)

x: 国際会議／y:旅行

品詞	単語	度数x	相対度数x	度数y	相対度数y	Larger	相対度数比(大/小)
普通名詞	こと	136	1.03%	542	2.91%	Y	2.83
	会議	19	0.14%	182	0.98%	Y	7.00
	方	116	0.88%	2	0.01%	X	88.00
	講演	27	0.20%	120	0.64%	Y	3.20
	会場	4	0.03%	112	0.60%	Y	20.00
	先生	7	0.05%	76	0.41%	Y	8.20
	時間	9	0.07%	76	0.41%	Y	5.86
	ファックス	17	0.13%	66	0.35%	Y	2.69
	前	2	0.02%	60	0.32%	Y	16.00
	キャンセル	10	0.08%	60	0.32%	Y	4.00
代名詞	私	5	0.04%	133	0.71%	Y	17.75
	わたくし	56	0.42%	11	0.06%	X	7.00
	これ	16	0.12%	66	0.35%	Y	2.92
副詞	どう	15	0.11%	70	0.38%	Y	3.45
接続詞	では	46	0.35%	208	1.12%	Y	3.20
	で	87	0.66%	24	0.13%	X	5.08
	じゃあ	70	0.53%	24	0.13%	X	4.08
接尾辞	様	74	0.56%	28	0.15%	X	3.73
	さん	7	0.05%	85	0.46%	Y	9.20
	時	8	0.06%	79	0.42%	Y	7.00
	円	55	0.42%	17	0.09%	X	4.67
	回	12	0.09%	58	0.31%	Y	3.44
感動詞	はい	1718	13.01%	495	2.66%	X	4.89
	ええ	280	2.12%	108	0.58%	X	3.66
	あ	1	0.01%	84	0.45%	Y	45.00
連体詞	もしもし	10	0.08%	60	0.32%	Y	4.00
本動詞	有る	33	0.25%	137	0.74%	Y	2.96
	畏まる	64	0.48%	10	0.05%	X	9.60
	書く	1	0.01%	63	0.34%	Y	34.00
	待つ	1	0.01%	45	0.24%	Y	24.00
形容詞	良い	20	0.15%	148	0.79%	Y	5.27
	無い	21	0.16%	87	0.47%	Y	2.94
補助動詞	いただく	1	0.01%	109	0.59%	Y	59.00
	いただける	1	0.01%	68	0.37%	Y	37.00
	なる	12	0.09%	68	0.37%	Y	4.11
	みる	7	0.05%	35	0.19%	Y	3.80

表の見方は、表3-B-イ(a) の下を参照のこと。

b．一方にのみ出現する語

表3-B-ニ(b)[K]「国際会議」のみに出現する語
－品詞別5件以上(キーボード会話)

普名詞		テクニカルツアー	20	レディースプログラム	5	座る	5
会議	182	ポスターセッション	20	開会	5	受取る	5
講演	120	開催	20	休憩	5	存じる	5
会場	112	イー	18	業者	5		
登録	92	フィシャル・インビテーション	18	国際	5	固名詞	
スライド	66	発表	18	指示	5	コンピュータ国際会議事務局	55
１１月	52	外務	16	式	5	コンピュータ国際会議	30
演者	42	見学	16	終了	5	日下部	25
バンケット	40	省	16	振り込み	5	田端	18
プログラム	40	心配	16	宣伝	5	アンダーソン教授	16
委員	40	オペレーター	7	対向	5	保刈	15
プロシーディングス	36	データショー	7	返金	5	シミュレーション学会	13
演題	36	ナンバー	7	方々	5	トヨタ	10
控え	36	ミーティング	7	理事	5	岩本	10
議長	34	機材	7			ＡＴＲ	8
原稿	32	打合せ	7	本動詞		アメリカ	8
手紙	32	科	6	願う	151	ワルシャワ	8
招待	32	会長	6	話す	30	山本	8
ＯＨＰ	30	学会	6	出す	24	名古屋ヒルトン	8
広告	30	月日	6	振込む	21	グリーン教授	7
セッション	28	雑誌	6	見る	14	ロバートソン	7
応答	26	書面	6	着く	13	成田	7
会	26	生年	6	会う	11	クーパー	6
研究	26	専門	6	映す	9	クリストファーソン教授	6
質疑	26	速達	6	入る	7	ブラウン教授	6
ポスター	24	通訳	6	代る	6	モスクワ	6
演台	24	アール	5	置く	6	仁科先生	6
スクリーン	22	スペル	5	返す	6	中村	6
工学	22	ブース	5	持つ	5	東京工業大学	6
欄	22	プロジェクター	5	映る	5	東工大	6

表3-B-ニ(b)[K]「国際会議」のみに出現する語
－品詞別5件以上(キーボード会話)

つづき

野原	6	調整	8	方	5	補動詞	
スペンサー	5	登録	8			あう	6
フーバー教授	5	申請	6	代名詞			
ペンシルバニア	5	開催	5	彼女	12	副　詞	
ホプキンス教授	5	決定	5			とんでも	5
マクガバン	5	推薦	5	形名詞		なるほど	5
坂口	5	聴講	5	だいじょうぶ	8		
山本理事	5	配布	5	具体的	6	感動詞	
人工知能学会	5			個人的	6	は	5
		接尾辞		大切	6		
サ名詞		部	14			接続詞	
講演	17	メートル	9	形容詞		それに	5
出席	9	パーセント	8	相応しい	8		

表3-B-ニ(b)[R]「旅行代理店」のみに出現する語
－ 品詞別5件以上(キーボード会話)

普名詞		デラックス	8	フリータイム	5	橋上	17
旅行	80	温泉	8	夏	5	ソウル	13
金	30	空き	8	夏休み	5	ＪＲ	12
社	26	自由	8	手段	5	与論島	12
予算	24	周遊	8	直通	5	バリ島	11
パック	23	６日	7	毎日	5	静内	11
現地	22	トラベラーズチェック	7			白樺湖	11
保険	20	パッケージ	7	本動詞		北海道	11
便	16	行動	7	持つ	33	ヨーロッパ	10
宅	15	主人	7	決る	13	札幌	10
旅館	15	船	7	調べる	13	北軽井沢	10
代金	13	追加	7	込む	12	ゴールドコースト	9
パスポート	12	補償	7	回る	10	シンガポール	9
色々	12	牧場	7	違う	9	安曇野	9
部屋	12	シーズン	6	入る	8	沖縄	9
３月	11	海	6	為す	7	新宿	9
コテージ	11	研修	6	寄る	6	羽田	8
プラン	11	座禅	6	捜す	6	広島	8
休み	11	昼食	6	拘る	5	野沢	7
カウンター	10	添乗	6	見る	5	NW	6
ゴルフ	10	平日	6	過す	5	イタリア	6
以上	10	民宿	6	知る	5	ラマダ	6
前後	10	旅	6			残波	6
土曜	10	旅程	6	固名詞		順子	6
方面	10	２１日	5	交通公社	25	瀬戸大橋	6
旅券	10	２７日	5	新宿支店	25	穂高	6
４月	9	９日	5	高木	21	あかねホテル	5
あたり	9	オプショナルツアー	5	神戸	19	スペイン	5
近く	9	テニス	5	京都	18	タヒチ	5
コート	8	テニスコート	5	日本交通公社	18	テッド	5

表3-B-ニ(b)[R]「旅行代理店」のみに出現する語
－ 品詞別5件以上(キーボード会話)
つづき

ナオミ・スズキ	5	返送	8	共	5	全く	5
ベルリン	5	変更	6	歳	5		
メルボルン	5	周遊	5			形容詞	
奥の細道	5	同封	5	感動詞		長い	6
熊野本宮	5			まあ	9	短い	5
長谷川	5	形名詞		あら	8		
奈良	5	フリー	10	さようなら	7	接頭辞	
						相	5
連体詞		接尾辞		副　詞			
当	23	代	9	あいにく	9	補動詞	
		等	7	あまり	9	なれる	5
サ名詞		室	6	いつも	6		
観光	13	付き	6	もしも	6		

【考察】

ハ.「電話会話」におけるタスク依存度で考察した内容のうち、次の点が「キーボード会話」におけるタスク依存度でも言える。

・話題のタスク依存性 － 「旅行」で地名，「国際会議」で人名・大学名が多いこと

但し、形容詞，副詞などの修飾語に、特徴的な差はなかった。

その他の特徴としては、「国際会議」では”こと”，”かた”，”もの”などの形式名詞が非常に多く「旅行」との差もはっきりしている点がある。

3-C. 形態素情報修正作業の評価

ここでは、3章の冒頭で示した図3－1の形態素情報データ作成の流れに沿い、3-C-1においてデータ作成作業の内容を説明したうえで、3-C-2において作業効率を定量的に評価する。

なお、本節で言う形態素情報データ作成は、文節情報データ作成工程を含む。文節情報データ作成工程の説明では、分節区切りつき形態素情報データのことを文節情報データと呼んでいる。

3-C-1 作業内容

説明中、データ作成作業者のことを作業者、対話データベース作成管理者のことを管理者と記す。

(1) 日本語テキストの受取

管理者から形態素データのもとになる日本語テキストのファイル名の指定を受ける。作業者はリモート端末のディスプレイ上で確認し、さらに印刷して検査を行う。

派生作業

●仕様検査

日本語テキストが会話の文字化仕様に合っているかどうかを検査する。もし誤りがあれば管理者に報告する。

この検査は、管理者があらかじめ行なっているものであるが、検査漏れに備えて作業者の方でも再度行なっている。データ作成はテキストに忠実に行なわれるため、万一作成作業に入ってからテキストに誤りがあると分かった場合には、作業に大きな支障を来たすことになるからである。

検査の内容

・規定外の発話ラベルが使われていないか
・ローマ字による数字の読みが入っているか
・余分な空白は無いか
・全角の空白は無いか
・文が句点「。」で終わっているか
・間投詞、言い淀み、相槌などのための特殊記号が正しく使用されているか
・明らかに誤字・脱字と分かる文字は無いか

など

●形態素数見積り

日本語テキストのバイト数から形態素数の見積りを立て、指定を受けたファイルのうちどのファイルを処理するかを決定する。決定後、管理者に見積書を提出し、承認を受ける。

データ作成完了後に、規定の形態素数を下回らないよう、1割程度多目に見積るようにしている。

(2) 形態素解析（後処理を含む）

日本語テキストを形態素解析システムおよび後処理システムで処理し、形態素情報データ（中間）を作成する。

派生作業

●バッチファイルの作成

3万語分のテキストを形態素解析システムおよび後処理システムで処理すると、30時間以上の時間を要する(3-A-1 で計測した時間は cpu time であるが、real time では約2倍の時間がかかる）。このため、3万語を2～3回に分け、夜間にバッチで処理している。作業者はこの準備作業としてバッチファイルを作成する。

(3) 形態素情報データ修正

形態素情報修正システムを使用して、(2) の解析誤りを修正する。この作業は複数の作業者により、複数回行なわれる。すなわち、一つの日本語テキストの形態素情報データを、相異なる作業者が2回以上†検査・修正する。

† ここで言う検査は人為的な検査であり、(4) で説明する機械的な検査と区別する。

(4) 検査（ツールを利用した機械的な検査）

形態素情報データ修正(複数回) が一通り終わった後、形態素情報検索システム，形態素情報抽出システム，および文章比較システムの3種類のシステムを利用して、形態素情報データの検査を行なう。これにより、入力ミス，検査漏れなどの人為的ミスを防ぐ。もし、誤りが発見された場合は、形態素情報修正システムで修正を行

なう。

3種類のシステムによる検査は以下のようにして行なわれる。

・形態素情報抽出システムによる検査

形態素情報抽出システムは本来、辞書のもとになるデータを作成するためのものであるが、検査にも流用されている。

このシステムは、形態素情報データを品詞毎に分類し、異なり単語毎の件数を出力する。作業者は、この異なり単語毎の件数を手掛かりとし、件数の少ないもの（特に1件しかないもの）の中から誤った品詞に分類されている単語や、誤った分割の仕方をしている単語を探し出す。

・形態素情報検索システムによる検査

形態素情報抽出システムを利用して発見した誤り単語の存在場所（テキストファイル名，ファイル内の位置）を特定する。

これは誤った形態素を形態素データの中から検索し、そのファイル名とその形態素を含む前後の形態素を出力することにより特定するできる。

・文章比較システムによる検査

文章比較システムは、作成された形態素情報データの表記データと日本語テキストを比較して、修正作業中に誤った変更をしていないかを検査する。

派生作業

●品詞判定基準の追補

品詞の判定は『形態素データ作成マニュアル』（以下、マニュアル）に準拠している。しかし実際のデータ作成作業においては、マニュアルだけでは判定基準が定まらない場合がある。この場合、判定基準を検討しATR担当者の了解を得て、新規事項としてマニュアルの判定基準に加える。

品詞判定基準の追補内容は、半期に一度提出する品質管理報告書のなかで、「新規事項」という項目で報告する。

●形態素分割基準の追補

品詞判定基準の追補と同様に行なわれる。

(5) 辞書登録

新しく作成された形態素情報データを加工し、解析用辞書に登録する。

この作業は、形態素情報抽出システム(EXTRACT)を用いて形態素情報データから抽出した辞書情報を、日本語辞書編集プログラム(EDKDIC)により既存の日本語辞書(KDIC)に登録し、その後解析用辞書構築ツール(SDICGEN)により解析用辞書(SDIC)を作成するという、3つのステップを踏む。

3つのステップを踏むのは非常に非効率であるが、これは、解析用辞書構築ツール(SDICGEN)が辞書検索の高速化のために、日本語辞書(KDIC)とその他不規則変化辞書辞書等を統合して1つの解析辞書にするための加工ツールとして開発されたという経緯に由来する。

(6) 文節情報データ作成

文節情報作成システムを使用し、形態素情報データをもとに文節区切りの入った文節情報データを作成する。

(7) 文節情報データ修正

文節情報修正システムを使用して、(6) の出力を検査し文節区切りの誤りを修正する。

この時点で形態素情報データの誤りが発見された場合は、形態素情報修正システムで修正したあと、(6) からやり直す。

●文節認定基準の追補

品詞判定基準の追補と同様に行なわれる。

(8) 管理情報の作成

対話データベース作成管理のための管理情報を作成する。

作成する管理情報を以下に挙げる。

・管理票

・納品書

・進行状況表

(9) 納品（毎月分）

(1) ～ (4) で作成した形態素情報データは、計算機上の所定のディレクトリに移すことにより納品される。

また、(6) ～ (7) で作成した文節区切りつき形態素情報データは、MS-DOS形式に変換しフロッピーディスクに格納し、管理者に手渡すことにより納品される。

派生作業

●データのバックアップ作業

月々の納品時に、6カ月前に納品した形態素解析データ作成のために作った中間ファイルをテープにダウンロードし、ディスク上の中間ファイルを消去する。

(10) 品質管理

6カ月に一度、過去6ヶ月間に作成した形態素情報データを対象に抜取検査を行ない、「品質管理報告書」としてまとめる。これは、全対象データから1％を無作為に抽出したものを、作業者が検査するという方法を採っている。もし抽出されたデータの中から1％以上の不良データが発見されれば、管理者から6ヶ月分のデータの差し戻しを受けるという規定になっている。

(11) 納品（半期分）

6カ月に一度、過去6ヶ月間に作成した形態素情報データに対する報告を行なう。内容は、(10)で示した品質管理の他、マニュアルの追補内容を「新規事項」という項目を設けて報告する。

納品物件には通常以下のものが提出される。

・「品質管理報告書」

・同報告書の電子化ファイル

派生作業

●データのバックアップ作業

過去6ヶ月間に作成した形態素情報データをテープにダウンロードする。

3-C-2 作業効率

(1) 調査内容

形態素情報データ作成作業のうち、修正作業（形態素情報データ修正および文節情報データ修正）の効率を調査する。

この調査は、９２年度下期に作成した形態素データ 15,000 語のうち、10,000語分に対する作業を対象として行なったものである。

表3-C-1 調査対象の日本語テキスト

番号	会話ＩＤ	ファイル名
1	T0607	tel-722-02
2	T0608	tel-722-03
3	T0609	tel-722-04
4	T0610	tel-722-05
5	T0611	tel-722-06
6	T0612	tel-722-07
7	T0613	tel-722-08
8	S8631	sdb-282-10

(2) 調査結果

形態素情報データ修正および文節情報データ修正に要した時間を、それぞれ、表3-C-2、表3-C-3 に示す。

表3-C-2 形態素情報データ修正時間

番号	形態素数	単語誤り	誤り率	修正1(分)	修正2(分)	検査(分)	修正合計(分)	形態素/h	修正/h
8	1,125	140	12.44%	269	94	56	419	161.10	20.05
2	1,132	167	14.75%	288	63	50	401	169.38	24.99
6	1,178	150	12.73%	288	138	69	495	142.79	18.18
3	1,275	142	11.14%	281	56	38	375	204.00	22.72
5	1,397	205	14.67%	294	144	81	519	161.50	23.70
1	1,477	217	14.69%	313	113	44	470	188.55	27.70
4	1,546	210	13.58%	338	181	100	619	149.85	20.36
7	1,717	282	16.42%	431	281	100	812	126.87	20.84
計	10,847	1,513	13.95%	2,502	1,070	538	4,110	1,304.04	178.53

表の見方
表は１行が１会話分のテキストファイルに対応している。
番号　　表3-C-1 で示したテキストファイルに対応
単語誤り　　3-A-3で示した単語誤りの計上方法による。
誤り率　　単語誤り÷形態素数
修正１　　第一回目の修正に要する時間
修正２　　第二回目の修正に要する時間
検査　　修正作業終了後の検査に要する時間
修正合計　　修正１＋修正２＋検査
形態素／ｈ　　１時間あたりの形態素作成個数
修正／ｈ　　１時間あたりの修正形態素数

表3-C-3 文節情報データ修正時間

番号	文節数	修正(分)	文節/時間
1	505	69	439.13
2	384	100	230.40
3	390	56	417.86
4	517	138	224.78
5	442	131	202.44
6	418	119	210.76
7	559	163	205.77
8	492	113	261.24
計	3,707	889	2,192.38

(3) 考察

計測結果より、形態素データ約１０，０００形態素分の作成にかかる時間は、６８．５時間(=4,110分)。文節データ作成時間を合わせると、約８３．１時間である。

ここで計測した作業時間は形態素情報作成における修正作業のの最も基本的な部分である。

3-C-1での作業内容で示すと、(3)形態素情報データ修正，(4)検査，(6)文節データ修正の３項目が、ここで時間計測した作業にあたる。

時間計測の対象外の作業には、

(1)テキストの受取

＋仕様検査

＋形態素数見積り

(2)形態素解析

＋バッチファイルの作成

(3)形態素修正のうち派生作業

品詞判定基準，形態素分割基準の追補

(5)辞書登録

(6)文節情報データ作成

(7)文節情報データ修正のうち派生作業

文節認定基準の追補

(8)管理情報の作成

(9)納品（１カ月分）

＋データバックアップ作業

(10)品質管理

(11)納品（半期分）

＋データバックアップ作業

がある。（"＋"は派生作業を示す。(3),(4)では派生作業のみなので"＋"を付けていない）

4.まとめ

本報告書では、まず第2章の日本語形態素解析システムの概要において、システムの現状について把握し、第3章の評価実験において、A.解析の性能面とB.辞書の構成面およびC.データ作成作業面の3つの側面からシステムの評価を行なった。

報告の目的の中心である、第3章の評価実験での成果をもう一度ここで確認しておきたい。

1. 最尤度法による形態素解析は、尤度計算に使用するパラメータの重み付け，設定法により、精度向上の可能性があることが分かった。

2. 実験結果より、妥当と思われるパラメータの重み付けを推定することが出来た。

3. 本システムで使用している最尤度法の問題点を指摘し、パラメータの独立性を保った尤度計算についての提言を行ない、形態素解析以外でも有効に利用できるという点に言及した。

謝辞

本研究の機会を与えてくださった樽松社長に感謝します。日頃、ご指導をいただいている森元室長に感謝します。

付　録

付録A.テストに使用したデータ（会話ＩＤによる表示）

１．テキストＫｔ（国際会議／電話会話）

４０８	４１３
４０９	４１４
４１０	４１５
４１１	４１６
４１２	４１７

２．テキストＲｔ（旅行／電話会話）

５００
５０１

３．テキストＫｋ（国際会議／キーボード会話）

３６１９	３６２４
３６２０	３６２５
３６２１	３６２６
３６２２	３６２７
３６２３	

４．テキストＲｋ（旅行／キーボード会話）

３３９５
３３９６
３３９７
３３９８

付録B．品詞優先度

品詞	優先度	相対度数
記号	0.241198	0.241198
格助詞	0.109083	0.109083
助動詞	0.099572	0.099572
本動詞	0.067218	0.067218
感動詞	0.064793	0.007465
間投詞	0.064793	0.064793
接続助詞	0.053183	0.053183
補助動詞	0.040455	0.040455
普通名詞	0.040089	0.118113
副詞	0.027606	0.027606
接尾辞	0.018679	0.018679
代名詞	0.017733	0.017733
係助詞	0.016942	0.016942
終助詞	0.016512	0.016512
接頭辞	0.014345	0.014345
固有名詞	0.012677	0.012677
接続詞	0.012023	0.012023
準体助詞	0.011834	0.011834
サ変名詞	0.009804	0.009804
数詞	0.009133	0.009133
連体詞	0.007448	0.007448
副助詞	0.006450	0.006450
形容詞	0.005573	0.005573
形容名詞	0.005452	0.005452
その他	0.004111	0.004111
並列助詞	0.002597	0.002597
自立語イディオム	1.000000	1.000000
付属語イディオム	1.000000	1.000000

付録C. バイグラムデータ

普名詞	格助詞	10691
記　号	間投詞	6595
間投詞	記　号	6589
記　号	記　号	6536
格助詞	本動詞	5825
記　号	普名詞	5551
格助詞	普名詞	5145
助動詞	接助詞	4605
本動詞	助動詞	4603
補動詞	助動詞	4456
助動詞	記　号	4254
接助詞	記　号	3888
接助詞	補動詞	3145
普名詞	普名詞	2758
助動詞	助動詞	2735
記　号	副　詞	2664
本動詞	接助詞	2568
終助詞	記　号	2341
助動詞	終助詞	2261
記　号	代名詞	2200
サ名詞	補動詞	1975
格助詞	記　号	1684
接頭辞	本動詞	1511
普名詞	係助詞	1483
記　号	接続詞	1450
準助詞	助動詞	1443
代名詞	格助詞	1437
本動詞	補動詞	1432
普名詞	助動詞	1410
接尾辞	格助詞	1358
補動詞	接助詞	1307
記　号	固名詞	1280
固名詞	格助詞	1268
普名詞	記　号	1229
記　号	感動詞	1136
記　号	本動詞	1132
本動詞	普名詞	1100
数　詞	接尾辞	1061
助動詞	普名詞	1053
格助詞	接頭辞	1034
係助詞	記　号	972
普名詞	接尾辞	961
形名詞	助動詞	952
接続詞	記　号	950
感動詞	記　号	940
接頭辞	普名詞	932
助動詞	準助詞	908
記　号	接頭辞	907
連体詞	普名詞	887
記　号	連体詞	782
格助詞	サ名詞	777
助動詞	格助詞	765
副　詞	本動詞	739
副　詞	記　号	732
格助詞	係助詞	639
補動詞	記　号	588
副　詞	普名詞	574
係助詞	本動詞	551
記　号	数　詞	547
補動詞	普名詞	518
本動詞	準助詞	479
感動詞	補動詞	443
助動詞	補動詞	440
係助詞	普名詞	439
普名詞	本動詞	427
格助詞	固名詞	406
接頭辞	サ名詞	401
格助詞	副　詞	396
記　号	形名詞	377
記　号	サ名詞	376
副　詞	接頭辞	365
その他	記　号	364
代名詞	係助詞	354
副　詞	助動詞	350
係助詞	副　詞	344
補動詞	準助詞	330
記　号	その他	326
格助詞	助動詞	324
普名詞	副助詞	322
代名詞	接尾辞	305
格助詞	数　詞	301
副助詞	格助詞	289
本動詞	格助詞	284
固名詞	助動詞	277
接尾辞	記　号	274
記　号	格助詞	269
接尾辞	普名詞	266
代名詞	副助詞	264
固名詞	接尾辞	264
接続詞	普名詞	256
固名詞	記　号	251
形容詞	普名詞	251
接助詞	係助詞	242
接助詞	普名詞	237
格助詞	格助詞	232
普名詞	並助詞	222
終助詞	格助詞	222
補動詞	格助詞	216
代名詞	記　号	216
接尾辞	助動詞	212
格助詞	形容詞	212
接尾辞	接尾辞	211
準助詞	係助詞	211
数　詞	記　号	206
助動詞	本動詞	206
接尾辞	副助詞	199
形容詞	助動詞	197
格助詞	代名詞	190
普名詞	接頭辞	187
記　号	形容詞	186
副　詞	感動詞	179
形容詞	準助詞	170
接助詞	助動詞	167
格助詞	形名詞	167
副　詞	数　詞	164
副　詞	副　詞	162
普名詞	数　詞	151
本動詞	記　号	147
副　詞	形容詞	147
係助詞	助動詞	146
接助詞	本動詞	143
接助詞	形容詞	141
接助詞	副　詞	140
副助詞	記　号	135
係助詞	形名詞	135
接尾辞	係助詞	133
代名詞	固名詞	130
係助詞	接頭辞	130
並助詞	普名詞	127
副　詞	代名詞	126
記　号	助動詞	126
助動詞	係助詞	123
係助詞	代名詞	123
副　詞	格助詞	120
係助詞	固名詞	119
副助詞	本動詞	118
副助詞	普名詞	117
準助詞	終助詞	115
普名詞	副　詞	114
助動詞	接頭辞	114
並助詞	記　号	113
係助詞	数　詞	113
形容詞	格助詞	111
係助詞	形容詞	109
副　詞	サ名詞	108
接頭辞	接尾辞	108
補動詞	終助詞	106
準助詞	格助詞	101
接助詞	接頭辞	99
固名詞	固名詞	92
格助詞	連体詞	92
代名詞	普名詞	90
数　詞	格助詞	90
接尾辞	本動詞	89
連体詞	助動詞	85
接助詞	格助詞	85
本動詞	終助詞	84
接助詞	代名詞	82
係助詞	サ名詞	82
副　詞	副助詞	78
接続詞	副　詞	77
代名詞	助動詞	75
助動詞	固名詞	75
本動詞	副助詞	72
接続詞	代名詞	72
普名詞	終助詞	71
接助詞	形名詞	71
副助詞	接頭辞	70
形容詞	本動詞	70
副助詞	助動詞	69
副　詞	形名詞	68
数　詞	普名詞	67
普名詞	代名詞	66
固名詞	普名詞	66
数　詞	助動詞	64
終助詞	副　詞	63
連体詞	記　号	62
副　詞	連体詞	62
普名詞	サ名詞	62
接尾辞	接頭辞	62
普名詞	形容詞	61
副助詞	係助詞	59
普名詞	接続詞	59
格助詞	補動詞	59
助動詞	形容詞	57
終助詞	終助詞	57
接続詞	助動詞	55
接助詞	連体詞	54
係助詞	連体詞	53
副助詞	形容詞	52
接助詞	サ名詞	52
形容詞	接助詞	51
記　号	補動詞	50
本動詞	接尾辞	48
接頭辞	形名詞	47
副　詞	固名詞	43
普名詞	固名詞	42
普名詞	感動詞	42
接頭辞	形容詞	42
固名詞	数　詞	42
形容詞	終助詞	41
普名詞	接助詞	40
接頭辞	数　詞	40
形容詞	記　号	39
接尾辞	数　詞	38
助動詞	サ名詞	37
形容詞	接頭辞	37
固名詞	係助詞	36
係助詞	補動詞	35
助動詞	形名詞	33
終助詞	係助詞	33
接尾辞	並助詞	32
接続詞	連体詞	32
接助詞	終助詞	32
連体詞	固名詞	31
副助詞	サ名詞	29
副　詞	係助詞	29
接続詞	接頭辞	29
連体詞	数　詞	28
形容詞	接尾辞	28
副助詞	形名詞	27
普名詞	形名詞	27
接尾辞	副　詞	27
接尾辞	サ名詞	26
接続詞	固名詞	26
接助詞	接助詞	26
接助詞	数　詞	26
助動詞	並助詞	26
普名詞	連体詞	25
接助詞	感動詞	25

格助詞	副助詞	25
補動詞	固名詞	24
副助詞	副　詞	24
準助詞	接助詞	24
終助詞	本動詞	23
感動詞	接助詞	23
助動詞	副　詞	22
格助詞	感動詞	22
本動詞	接頭辞	21
助動詞	代名詞	21
終助詞	普名詞	21
固名詞	並助詞	20
並助詞	固名詞	18
接尾辞	固名詞	17
接続詞	数　詞	17
記　号	副助詞	17
接助詞	固名詞	16
助動詞	連体詞	16
副助詞	副助詞	14
並助詞	本動詞	13
並助詞	接頭辞	13
副助詞	代名詞	13
助動詞	副助詞	13
記　号	接助詞	13
連体詞	接頭辞	12
補動詞	補動詞	12
並助詞	接続詞	12
並助詞	助動詞	12
形名詞	普名詞	12
感動詞	助動詞	12
サ名詞	記　号	12
連体詞	副助詞	11
本動詞	数　詞	11
本動詞	その他	11
接頭辞	記　号	11
接続詞	サ名詞	11
助動詞	感動詞	11
終助詞	数　詞	11
固名詞	本動詞	11
形容詞	サ名詞	11
補動詞	副助詞	10
副　詞	接助詞	10
代名詞	本動詞	10
代名詞	副　詞	10
接尾辞	接続詞	10
接尾辞	終助詞	10
接続詞	本動詞	10
終助詞	代名詞	10
記　号	接尾辞	10
本動詞	本動詞	9
補動詞	並助詞	9
接尾辞	代名詞	9
接続詞	形名詞	9
記　号	係助詞	9
格助詞	その他	9
補動詞	接頭辞	8
並助詞	格助詞	8
普名詞	準助詞	8
代名詞	終助詞	8
代名詞	サ名詞	8
接続詞	接助詞	8
助動詞	数　詞	8
固名詞	副助詞	8
間投詞	間投詞	8
格助詞	終助詞	8
連体詞	本動詞	7
普名詞	その他	7
接尾辞	連体詞	7
接頭辞	固名詞	7
数　詞	固名詞	7
終助詞	並助詞	7
終助詞	接頭辞	7
固名詞	接続詞	7
形名詞	記　号	7
間投詞	助動詞	7
並助詞	副　詞	6
並助詞	数　詞	6
副　詞	終助詞	6
代名詞	並助詞	6
代名詞	接頭辞	6
接尾辞	準助詞	6
接助詞	並助詞	6
数　詞	接頭辞	6
形容詞	係助詞	6
本動詞	並助詞	5
本動詞	固名詞	5
補動詞	その他	5
副助詞	連体詞	5
副助詞	数　詞	5
副助詞	固名詞	5
接頭辞	その他	5
接続詞	形容詞	5
接続詞	感動詞	5
準助詞	記　号	5
終助詞	接続詞	5
その他	助動詞	5
連体詞	形容詞	4
連体詞	格助詞	4
連体詞	サ名詞	4
本動詞	連体詞	4
本動詞	副　詞	4
本動詞	代名詞	4
補動詞	感動詞	4
並助詞	代名詞	4
普名詞	補動詞	4
代名詞	代名詞	4
接尾辞	接助詞	4
接頭辞	副　詞	4
数　詞	接続詞	4
助動詞	その他	4
終助詞	固名詞	4
形容詞	補動詞	4
記　号	並助詞	4
感動詞	副　詞	4
感動詞	感動詞	4
その他	補動詞	4
連体詞	形名詞	3
本動詞	形容詞	3
本動詞	感動詞	3
補動詞	代名詞	3
並助詞	連体詞	3
並助詞	係助詞	3
副助詞	接続詞	3
副助詞	終助詞	3
副助詞	感動詞	3
代名詞	数　詞	3
代名詞	形容詞	3
代名詞	形名詞	3
接尾辞	形名詞	3
接助詞	副助詞	3
接助詞	接続詞	3
接助詞	その他	3
終助詞	助動詞	3
固名詞	接頭辞	3
形容詞	副　詞	3
係助詞	接続詞	3
係助詞	終助詞	3
感動詞	代名詞	3
感動詞	固名詞	3
格助詞	接続詞	3
その他	本動詞	3
連体詞	代名詞	2
本動詞	形名詞	2
本動詞	係助詞	2
本動詞	サ名詞	2
補動詞	連体詞	2
補動詞	副　詞	2
並助詞	形名詞	2
並助詞	サ名詞	2
副助詞	接助詞	2
副　詞	並助詞	2
副　詞	その他	2
接尾辞	形容詞	2
接尾辞	感動詞	2
接頭辞	接頭辞	2
接続詞	接続詞	2
数　詞	副　詞	2
数　詞	数　詞	2
数　詞	サ名詞	2
準助詞	本動詞	2
準助詞	並助詞	2
固名詞	その他	2
形容詞	代名詞	2
形容詞	形容詞	2
形容詞	形名詞	2
形名詞	副助詞	2
形名詞	接助詞	2
形名詞	格助詞	2
係助詞	その他	2
感動詞	普名詞	2
感動詞	終助詞	2
格助詞	並助詞	2
格助詞	準助詞	2
サ名詞	助動詞	2
サ名詞	格助詞	2
連体詞	その他	1
補動詞	本動詞	1
補動詞	接尾辞	1
補動詞	数　詞	1
補動詞	形名詞	1
補動詞	サ名詞	1
並助詞	形容詞	1
副　詞	補動詞	1
副　詞	接尾辞	1
副　詞	接続詞	1
普名詞	間投詞	1
代名詞	連体詞	1
代名詞	接続詞	1
接尾辞	その他	1
接頭辞	補動詞	1
接頭辞	助動詞	1
数　詞	連体詞	1
数　詞	本動詞	1
数　詞	副助詞	1
助動詞	接続詞	1
準助詞	準助詞	1
準助詞	サ名詞	1
終助詞	副助詞	1
終助詞	形容詞	1
終助詞	感動詞	1
固名詞	副　詞	1
固名詞	接助詞	1
固名詞	終助詞	1
形容詞	連体詞	1
形容詞	副助詞	1
形容詞	その他	1
係助詞	準助詞	1
係助詞	感動詞	1
係助詞	格助詞	1
記　号	終助詞	1
感動詞	本動詞	1
感動詞	サ名詞	1
格助詞	接助詞	1
サ名詞	本動詞	1
サ名詞	その他	1
その他	副助詞	1
その他	普名詞	1
その他	接助詞	1
その他	終助詞	1

付録D.解析用辞書の例

("１０月" "１０月" "じゅうがつ" (4) 10 (22 . 4))
("１１月" "１１月" "じゅういちがつ" (4) 25 (22 . 4))
("１２月" "１２月" "じゅうにがつ" (4) 2 (22 . 4))
("１２日" "１２日" "じゅうににち" (4) 1 (22 . 4))
("１３日" "１３日" "じゅうさんにち" (4) 12 (22 . 4))
("１４日" "１４日" "じゅうよっか" (4) 26 (22 . 4))
("１５日" "１５日" "じゅうごにち" (4) 79 (22 . 4))
("１６日" "１６日" "じゅうろくにち" (4) 32 (22 . 4))
("１７日" "１７日" "じゅうしちにち" (4) 7 (22 . 4))
("１７日" "１７日" "じゅうななにち" (4) 11 (22 . 4))
("１８日" "１８日" "じゅうはちにち" (4) 3 (22 . 4))
("１９日" "１９日" "じゅうくにち" (4) 3 (22 . 4))
("２０日" "２０日" "はつか" (4) 2 (22 . 4))
("３０日" "３０日" "さんじゅうにち" (4) 5 (22 . 4))
("３１日" "３１日" "さんじゅういちにち" (4) 1 (22 . 4))

("３月" "３月" "さんがつ" (4) 8 (22 . 4))
("４月" "４月" "しがつ" (4) 3 (22 . 4))
("７月" "７月" "しちがつ" (4) 25 (22 . 4))
("８月" "８月" "はちがつ" (4) 2 (22 . 4))
("９月" "９月" "くがつ" (4) 31 (22 . 4))
("あ" "あ" "あ" (11) 1 (426 . 164))
("あ" "あ" "あ" (33) 55 (201 . 19))
("あー" "あー" "あー" (33) 4 (201 . 19))
("ああ" "ああ" "ああ" (33) 160 (201 . 19))
("ああー" "ああー" "ああー" (33) 12 (201 . 19))
("あい" "合う" "あい" (32 1 1) 1 (166 . 18))
("あいだ" "間" "あいだ" (4) 1 (22 . 4))
("あう" "合う" "あう" (32 2 1) 1 (167 . 18))
("あう" "合う" "あう" (32 3 1) 1 (168 . 18))
("あえ" "合う" "あえ" (32 4 1) 1 (169 . 18))
("あえ" "合う" "あえ" (32 5 1) 1 (170 . 18))
("あお" "合う" "あお" (32 0 1) 1 (165 . 18))
("あたくし" "私" "あたくし" (6) 8 (24 . 6))
("あたり" "辺" "あたり" (4) 1 (22 . 4))
("あっ" "あっ" "あっ" (11) 3 (426 . 164))
("あっ" "あっ" "あっ" (33) 248 (201 . 19))
("あっ" "ある" "あっ" (19 1 1 1) 11 (38 . 13))
("あっ" "合う" "あっ" (32 1 1 1) 1 (166 . 18))
("あっ" "有る" "あっ" (32 1 1 1) 33 (166 . 18))
("あと" "後" "あと" (4) 23 (22 . 4))
("あの" "あの" "あの" (33) 118 (201 . 19))
("あのー" "あのー" "あのー" (33) 695 (201 . 19))
("あのーね" "あのーね" "あのーね" (33) 1 (201 . 19))
("あのですね" "あのですね" "あのですね" (33) 1 (201 . 19))
("あら" "ある" "あら" (19 0 1) 11 (37 . 13))
("あら" "有る" "あら" (32 0 1) 33 (165 . 18))
("あらーっ" "あらーっ" "あらーっ" (33) 1 (201 . 19))
("あらかじめ" "予" "あらかじめ" (8) 1 (26 . 8))
("あり" "ある" "あり" (19 1 1) 11 (38 . 13))
("あり" "有る" "あり" (32 1 1) 33 (166 . 18))
("ありがた" "有難い" "ありがた" (1 6) 4 (7 . 1))
("ありがたい" "有難い" "ありがたい" (1 2) 4 (3 . 1))
("ありがたい" "有難い" "ありがたい" (1 3) 4 (4 . 1))
("ありがたかっ" "有難い" "ありがたかっ" (1 1) 4 (2 . 1))
("ありがたかろ" "有難い" "ありがたかろ" (1 0) 4 (1 . 1))
("ありがたく" "有難い" "ありがたく" (1 1) 4 (2 . 1))
("ありがたけれ" "有難い" "ありがたけれ" (1 4) 4 (5 . 1))
("ありがとう" "ありがとう" "ありがとう" (11) 70 (426 . 164))
("ある" "ある" "ある" (19 2 1) 11 (39 . 13))
("ある" "ある" "ある" (19 3 1) 11 (40 . 13))
("ある" "有る" "ある" (32 2 1) 33 (167 . 18))
("ある" "有る" "ある" (32 3 1) 33 (168 . 18))
("あれ" "ある" "あれ" (19 4 1) 11 (41 . 13))
("あれ" "ある" "あれ" (19 5 1) 11 (42 . 13))
("あれ" "有る" "あれ" (32 4 1) 33 (169 . 18))
("あれ" "有る" "あれ" (32 5 1) 33 (170 . 18))
("あれですよね" "あれですよね" "あれですよね" (33) 1 (201 . 19))
("あろ" "ある" "あろ" (19 0 1) 11 (37 . 13))
("あろ" "有る" "あろ" (32 0 1) 33 (165 . 18))
("あわ" "合う" "あわ" (32 0 1) 1 (165 . 18))
("い" "い" "い" (33) 1 (201 . 19))
("い" "いる" "い" (19 0 0) 57 (31 . 13))
("い" "いる" "い" (19 1 0) 57 (32 . 13))
("い" "居る" "い" (32 0 0) 5 (159 . 18))
("い" "居る" "い" (32 1 0) 5 (160 . 18))
("い" "良い" "い" (1 6) 20 (7 . 1))
("いー" "いー" "いー" (33) 1 (201 . 19))
("いい" "言う" "いい" (32 1 1) 167 (166 . 18))
("いい" "良い" "いい" (1 2) 20 (3 . 1))
("いい" "良い" "いい" (1 3) 20 (4 . 1))
("いいえ" "いいえ" "いいえ" (11) 13 (426 . 164))
("いう" "言う" "いう" (32 2 1) 167 (167 . 18))
("いう" "言う" "いう" (32 3 1) 167 (168 . 18))
("いえ" "いえ" "いえ" (11) 2 (426 . 164))
("いえ" "言う" "いえ" (32 4 1) 167 (169 . 18))
("いえ" "言う" "いえ" (32 5 1) 167 (170 . 18))
("いえいえ" "いえいえ" "いえいえ" (11) 1 (426 . 164))
("いお" "言う" "いお" (32 0 1) 167 (165 . 18))
("いか" "行く" "いか" (32 0 1) 7 (165 . 18))
("いかが" "如何" "いかが" (8) 3 (26 . 8))
("いかっ" "良い" "いかっ" (1 1) 20 (2 . 1))
("いかろ" "良い" "いかろ" (1 0) 20 (1 . 1))
("いき" "行く" "いき" (32 1 1) 7 (166 . 18))
("いく" "行く" "いく" (32 2 1) 7 (167 . 18))
("いく" "行く" "いく" (32 3 1) 7 (168 . 18))
("いく" "良い" "いく" (1 1) 20 (2 . 1))
("いくら" "幾等" "いくら" (8) 11 (26 . 8))
("いけ" "行く" "いけ" (32 4 1) 7 (169 . 18))
("いけ" "行く" "いけ" (32 5 1) 7 (170 . 18))
("いけ" "行ける" "いけ" (32 0 3) 2 (177 . 18))
("いけ" "行ける" "いけ" (32 1 3) 2 (178 . 18))
("いけよ" "行ける" "いけよ" (32 5 3) 2 (182 . 18))
("いける" "行ける" "いける" (32 2 3) 2 (179 . 18))
("いける" "行ける" "いける" (32 3 3) 2 (180 . 18))
("いけれ" "行ける" "いけれ" (32 4 3) 2 (181 . 18))
("いけれ" "良い" "いけれ" (1 4) 20 (5 . 1))
("いけろ" "行ける" "いけろ" (32 5 3) 2 (182 . 18))
("いこ" "行く" "いこ" (32 0 1) 7 (165 . 18))
("いたさ" "いたす" "いたさ" (19 0 1) 229 (37 . 13))
("いたさ" "致す" "いたさ" (32 0 1) 15 (165 . 18))
("いたし" "いたす" "いたし" (19 1 1) 229 (38 . 13))
("いたし" "致す" "いたし" (32 1 1) 15 (166 . 18))
("いたす" "いたす" "いたす" (19 2 1) 229 (39 . 13))
("いたす" "いたす" "いたす" (19 3 1) 229 (40 . 13))
("いたす" "致す" "いたす" (32 2 1) 15 (167 . 18))
("いたす" "致す" "いたす" (32 3 1) 15 (168 . 18))

付録E．日本語テキストの例

電話会話（国際会議）

[会話408]

担当者：（はい、こちら）はい、こちら、コンピュータ国際会議事務局でございます。
申込者：東大の斉藤弘美<saitouhiromi>です。
担当者：［あ］斉藤先生でいらっしゃいますね。
　　いつもお世話になっております。
申込者：はい、どうも。
　　［えーと］今度のアイシーシーの会期中ですね、｜はい｜ ホテルに部屋をとっていただきたいんですけども。
担当者：はい。
申込者：［そうー］名古屋観光ホテルがいいかな。
担当者：［あ］はい。
申込者：［えー］それでですね、１５日講演を終えたら、東京に帰ることにしましたので、｜はい｜ １３日と１４日<juuyokka>の２泊<nihaku>でいいと思います。
担当者：はい、承知いたしました。
　　それでは、ご確認させていただきます。
　　名古屋観光ホテルに、１１月１３日、１４日<juuyokka>の２泊<nihaku>ということでよろしゅうございますか。
申込者：はい、それで結構。
担当者：［あっ］（そうで）［そうですね、あのー］シングルルームでよろしいんでしょうか。
申込者：いいですよ。
担当者：はい、承知しました。
　　早速ご手配いたします。
申込者：はい、よろしく。

[会話409]

担当者：はい、こちら、コンピュータ国際会議事務局でございます。
申込者：東大の斉藤弘美<saitouhiromi>です。
担当者：はい、いつもお世話になっております。
申込者：どうも、こんにちは。
　　［えー］セカンドアナウンスメントが届きました。
担当者：［あ］はい。
申込者：ありがとうございました。
　　それで、ちょっとお伺いしたいことがあるんですが。
担当者：はい。
申込者：前回、ユーシーエルエーのときですね、｜はい｜ 観光ツアーとか、テクニカルビジットというのがあったんですけども、｜［あっ］はい｜ 今回はどうなんでしょう。
担当者：［あっ］今回もですね、名古屋市の市内観光ですとか、｜ええ｜ お茶とお花の体験ツアーなど、レディースプログラムとして用意してございます。
申込者：［ふん、ふん、ふん］。
担当者：［あー］それからですね、テクニカルビジットといたしましては、トヨタ自動車の組立工場と｜［ほう］｜ エーティーアール自動翻訳電話研究所に行く予定になっております。
　　まだ［あのー］詳しいことが決まっておりませんのですが。
申込者：［ああ］そうですか。
担当者：いかがでございましょうか。
申込者：はい、はい、［えー］それでですね、そのときお世話になった［そのー］バベル研究所のアンダーソン教授夫妻と｜［あっ］はい｜ 旧交を温めたいと思いますので、｜はい｜ ［えー］私も妻を同行したいんですが。
担当者：［ああ］そうでございますか。
申込者：［えー］妻の宿泊とか、交通も一緒に手配をお願いします。
担当者：［あっ］わかりました。
　　そうしましたら、奥様の分も同じホテルで。
　　［あのー］｜ええ｜ 今度、部屋ですが、シングルルームをおとりしてあったと思うんですが、ツインにお替えいたしますか。
申込者：そうですね、できましたら、お願いします。
担当者：はい、わかりました。
　　そうしましたら、早速手配いたします。
　　それでは、お電話ありがとうございました。
　　失礼いたします。
申込者：はい。

[会話410]

担当者：はい、コンピュータ国際会議事務局です。
申込者：東大の斉藤弘美<saitouhiromi>と申します。
担当者：［あっ］東大の斉藤先生ですね、｜はい｜ お世話になっております。
申込者：はい、はい、どうも。
　　［えー］原稿と｜はい｜ アブストラクトがですね、｜はい｜ 締切りはあさって、｜１５日になっております｜ ９月１５日ですよね。
担当者：はい。
申込者：［えー、ちょっと］まにあいそうもないんですよ。
担当者：［あっ］そうですか。
申込者：それでー、｜はい｜ ［まあ］遅れますので、よろしくお願いします。
担当者：それではですね、［あのー］斉藤先生、原稿は［あのー］１５日すぎてもよろしいんですけれども、｜はい｜ できれば、アブストラクトだけでも、至急こちらの方にお送り願えませんでしょうか。
　　［あのー］｜そう｜ お急がせして、まことに｜うん｜ 申し訳ないんですけれども、｜はい［えー］｜ 同時通訳との兼ね合いもございますので、｜うん｜ アブストラクトだけでも、できたら、｜はい｜ 送っていただきましたら｜はい、はい｜ 助かるんですけれども。
申込者：なんとかやってみます。
担当者：［あ］そうですか。
申込者：はい。
担当者：それで、まことに申し訳ないんですけれども、今日は、もう１３日で、締切りまであと２日<futsuka>しかございませんので、できれば［あのー］郵送よりも、ファックスでこちらの方に送っていただければ助かるんですけれども。
申込者：そうですね。
担当者：はい。
申込者：はい、はい、そうします。
担当者：はい、こちらのファックスナンバーはご存じでいらっしゃいますね。
申込者：はい。
担当者：はい、では、おそれいりますが、そちらの方に、よろしくお願いいたします。
　　ほかにはございませんでしょうか。
申込者：ええ、それだけです。
担当者：はい、すいません、お急がせしますが、よろしくお願いいたします。
申込者：はい。
担当者：はい、失礼します。

キーボード会話（国際会議）

[会話361]

通訳者：もしもし。
事務局：はい、こちらは第１３回コンピュータ国際会議事務局です。
通訳者：フレミングと申しますが、どなたか印刷の係の方をお願いします。
事務局：印刷とおっしゃいますと、議事録の関係でしょうか。
通訳者：訂正してもらいたいことがあるんです。
　　会議のプログラムの私の名前がミスプリントなんです。
事務局：あ、それは大変失礼いたしました。
　　私から係の者に伝えますので、恐れ入りますがフレミング様の、所属と、お名前を、お願いします。
通訳者：マサチューセッツ工科大学、言語学教授のメアリー・フレミングです。
事務局：マサチューセッツ工科大学、言語学のメアリー・フレミング教授でいらっしゃいますね。
　　お名前のスペルアウトを、お願いします。
通訳者：はい、メアリーはエム・エー・アール・ワイ。
　　フレミングはエフ・エル・イー・エム・アイ・エヌ・ジーです。
事務局：はい、承知しました。
　　これから印刷するものにつきましては、正しいお名前に訂正いたします。
　　プログラムについては、訂正表を、はさむことにいたします。
通訳者：ええ、そうしていただければ結構です。
事務局：はい、お知らせいただきありがとうございました。
　　では、会議でお目にかかるのをたのしみにしております。
通訳者：私もです。
　　どうもありがとうございました。
　　失礼します。
事務局：はい。
　　失礼します。

[会話362]

通訳者：もしもし、フーバーです。
　　成田から電話してるんですが、日下部さんをお願いします。
事務局：はい、私、日下部ですが、フーバーさん何かお困りでしょうか。
通訳者：ええ、大変なんです。
　　荷物がなくなってしまって、航空会社の人が探してくれてはいるんですが。
事務局：えっ、そうですか。
　　こちらから航空会社に問い合わせて、荷物が見つかり次第ホテルに届けさせますので、とりあえずホテルでお待ちいただけますか。
　　どちらの航空会社の何便にお乗りでしたか。
通訳者：ちょっと待ってください。
　　私はこれからすぐ飛行機で名古屋に向かうんです。
　　今晩は名古屋に泊まるんですよ、身の回りのものは全部荷物の中に入っているし。
事務局：ええ、お困りのことと思いますが、とにかく荷物がいまどこにあるのか、調べてもらいます。
　　そのうえで、間に合えば名古屋行きの飛行機に積んでもらいます。
　　もし、間に合わないときは、名古屋のホテルまで届けさせます。
通訳者：そうしていただければ助かります。
　　ここに英語のわかる女性がいらっしゃるので彼女に事情を説明していただけますか。
　　阿蘇さんとおっしゃる方です。
　　よろしくお願いしますね。
事務局：はい。
　　承知しました。

[会話363]

通訳者：もしもし。
　　第１３回コンピュータ国際会議事務局ですか。
事務局：はい、第１３回コンピュータ国際会議事務局です。
通訳者：会議の招待講演者なんですが、私の旅費についてそちらで負担していただけると書面に書いてあったので。
　　これには観光旅行も含まれるんでしょうか。
　　東京と京都に行こうと考えているんです。
事務局：旅費というのは、会議に参加するための費用という意味でして、観光旅行の場合は、含まれません。
　　切符の手配はいたしますので、会議の当日にホスピタリティデスクにお立ち寄りください。
通訳者：わかりました。
　　ではホスピタリティーデスクに行ってみます。
　　会議の前の旅行についてはどなたがアレンジしてくださるんでしょうか。
事務局：今度の会議ではＪＴＢという旅行会社が、担当していますので、そこの電話番号と担当者の名前をファックスでお送りします。
通訳者：ＪＴＢですか、聞いたことがあります。
　　ファックスしていただかなくても今電話番号と担当の人の名前を教えてください。
事務局：はい。
　　１１２－３３４４、担当者は、中村さんです。
通訳者：どうもありがとうございました。
　　このほうがはやいので。
事務局：それでは、失礼します。
通訳者：失礼します。

付録F.形態素情報データ作成作業で扱われるデータ

次ページより、以下のデータを示す。

F-1. 日本語テキスト（3-C-1.(1)日本語テキストの受取）
F-2. 形態素情報データ[後処理前]（3-C-1.(2)形態素解析）
F-3. 形態素情報データ[修正後]（3-C-1.(3)形態素情報データ修正）
F-4. 形態素情報抽出システム出力ファイル（3-C-1.(4)検査, 3-C-1.(5)辞書登録）
F-5. 形態素情報検索システム検索結果（3-C-1.(4)検査）
F-6. 文章比較システムの出力（3-C-1.(4)検査）
F-7. 文節情報データ（3-C-1.(6)文節情報データ作成,(7)文節情報データ修正）

F-1. 日本語テキスト（3-C-1.(1)日本語テキストの受取）

FILE NAME: tel-772-16

t-772-16

担当者：はい、コンピュータ国際会議事務局でございます。
申込者：東大の斉藤と申します。
担当者：はい、［あっ］東大の斉藤先生ですね、〔はい、ええ〕お世話になっております。
申込者：ええ、［えー］１５日の午後２時４０分から講演することになっているんですが、〔はい〕実は用事が長引きましてね、〔はい〕今夜中に東京を発てないんですよ。
担当者：はい。
申込者：ええ、それで、〔ええ〕明日<asu>朝一番でそちらへ行きたいと思っているんですが、〔はい〕で、一番早く会場に着ける方法をですね、〔ええ〕教えていただきたいと思いましてお電話しました。
担当者：［ああ］そうですか。
申込者：ええ。
担当者：［そうですね］それでは［ま、あのー］東京駅から新幹線で名古屋まで来られまして、〔ええ〕あと（名古屋からは）名古屋駅からはタクシーで１０分ほどですので、〔ええ〕タクシーに乗られて、〔ええ〕国際会議場と言っていただければわかると思います。
申込者：［ああ］そうですか。
担当者：はい。
申込者：国際会議場だけでわかりますか。
担当者：はい、国際会議場でわかります。
申込者：［はあはあ］。
［えー］名古屋駅からタクシーで約１０分ですか。
担当者：１０分です、はい。
申込者：［ああ］そうですか。
担当者：ですから、ま、午前中に東京をお発ちになれば、〔ええ〕午後からの講演には大丈夫だと思います。
申込者：［ああ］そうですね。
担当者：はい。
申込者：わかりました。
ありがとうございました。
担当者：はい、それでは［あのー］〔ええ〕こちらの方でもお待ちしておりますので。
申込者：はい。
担当者：はい（よろし）。
申込者：それじゃあどうも〔はい〕お世話様でした。
担当者：はい、よろしくお願いいたします。
どうもわざわざありがとうございました。
申込者：はい、失礼します。
担当者：失礼いたします。

F-2. 形態素情報データ[後処理前] (3-C-1.(2)形態素解析)

FILE : ¥0_tel-772-16-01

文番号 1

担当者：			記　号	－－	－－	
はい	はい	はい	感動詞	－－	－－	
、			記　号	－－	－－	

文番号 2

コンピュータ	コンピュータ	コンピュータ	普名詞	－－	－－	
国際会議事務局	こくさいかいぎ	国際会議事務局	固名詞	－－	－－	
で	で	だ	助動詞	－－	連用	
ございます	ございます	ございます	補動詞	特殊	終止	
。			記　号	－－	－－	

文番号 3

申込者：			記　号	－－	－－	
東大	とうだい	東大	固名詞	－－	－－	
の	の	の	格助詞	－－	－－	
斉藤	斉藤	斉藤	普名詞	－－	－－	
と	と	と	格助詞	－－	－－	
申し	もうし	申す	本動詞	五段	連用	
ます	ます	ます	助動詞	－－	終止	
。			記　号	－－	－－	

文番号 4

担当者：			記　号	－－	－－	
はい	はい	はい	感動詞	－－	－－	
、			記　号	－－	－－	

文番号 5

[			記　号	－－	－－	
あっ	あっ	在る	本動詞	五段	連用	つ音
]			記　号	－－	－－	
東大	とうだい	東大	固名詞	－－	－－	
の	の	の	格助詞	－－	－－	
斉藤先生	斉藤先生	斉藤先生	普名詞	－－	－－	
です	です	です	助動詞	－－	終止	
ね	ね	ね	終助詞	－－	－－	
、			記　号	－－	－－	

F-3. 形態素情報データ[修正後] (3-C-1.(3)形態素情報データ修正)

FILE : ED_tel-772-16-01

文番号 1

担当者：			記　号	－－	－－
はい	はい	はい	感動詞	－－	－－
、			記　号	－－	－－

文番号 2

コンピュータ国	コンピュータこ	コンピュータ国	固名詞	－－	－－
で	で	で	格助詞	－－	－－
ございます	ございます	御座います	本動詞	特殊	終止
。			記　号	－－	－－

文番号 3

申込者：			記　号	－－	－－
東大	とうだい	東大	固名詞	－－	－－
の	の	の	格助詞	－－	－－
斉藤	さいとう	斉藤	固名詞	－－	－－
と	と	と	格助詞	－－	－－
申し	もうし	申す	本動詞	五段	連用
ます	ます	ます	助動詞	－－	終止
。			記　号	－－	－－

文番号 4

担当者：			記　号	－－	－－
はい	はい	はい	感動詞	－－	－－
、			記　号	－－	－－

文番号 5

「			記　号	－－	－－
あっ	あっ	あっ	間投詞	－－	－－
」			記　号	－－	－－
東大	とうだい	東大	固名詞	－－	－－
の	の	の	格助詞	－－	－－
斉藤先生	さいとうせんせ	斉藤先生	固名詞	－－	－－
です	です	です	助動詞	－－	終止
ね	ね	ね	終助詞	－－	－－
、			記　号	－－	－－

F-4. 形態素情報抽出システム出力ファイル（3-C-1.(4)検査，3-C-1.(5)辞書登録）

Jan 19 13:08 1993 FL_tel-772-16 Page 1

[記号]

。				30
、				24
担当者：				14
申込者：				13
｛				12
｝				12
［				11
］				11
（				2
）				2

[形容詞]

早い	早い	はやい		1

[形容動詞]

大丈夫	大丈夫	だいじょうぶ	ダ活	1

[普通名詞]

タクシー	タクシー	タクシー		3
世話	世話	せわ		2
午後	午後	ごご		2
１５日	１５日	じゅうごにち		1
こと	事	こと		1
用事	用事	ようじ		1
今夜	今夜	こんや		1
明日	明日	あす		1
朝	朝	あさ		1
一番	一番	いちばん		1
会場	会場	かいじょう		1
方法	方法	ほうほう		1
新幹線	新幹線	しんかんせん		1
あと	後	あと		1
午前	午前	ごぜん		1
講演	講演	こうえん		1
方	方	ほう		1

[サ変名詞]

失礼	失礼	しつれい		2
講演	講演	こうえん		1
電話	電話	でんわ		1

[代名詞]

そちら	其方	そちら		1
こちら	此方	こちら		1

[数詞]

１０	１０	じゅっ		3
２	２	に		1
４０	４０	よんじゅっ		1

[副詞]

そう	そう	そう		4
どうも	どうも	どうも		2

F-5. 形態素情報検索システム検索結果（3-C-1.(4)検査）

extract ED_tel-772-16

ええ、それで、〔ええ〕明日朝一番でそちらへ行きたいと思っているんです
格助詞 —— —— で

思っているんですが、〔はい〕で、一番早く会場に着ける方法をですね
接続詞 —— —— で

］それでは［ま、あのー］東京駅から新幹線で名古屋まで来られまして、〔ええ〕
格助詞 —— —— で

あと（名古屋からは）名古屋駅からはタクシーで１０分ほどですので、〔ええ〕タクシー
格助詞 —— —— で

そうですか。担当者：はい。申込者：国際会議場だけでわかりますか。担当者：はい、国際会議場でわかり
格助詞 —— —— で

だけでわかりますか。担当者：はい、国際会議場でわかります。申込者：［はあはあ］。［えー
格助詞 —— —— で

［はあはあ］。［えー］名古屋駅からタクシーで約１０分ですか。担当者：１０分です
格助詞 —— —— で

F-6. 文章比較システムの出力 (3-C-1.(4)検査)

< 1993 年 1 月 19 日 13 時 3 分 >
User Name : tdpmaker
形態素ディレクトリー : /data3/MORPH/edwork/tel
形態素ファイル : ED_tel-772-16-01
日本語ディレクトリー : /data2/CONV/TELEPHONE/77-9106-K
日本語ファイル : tel-772-16
形態素データ比較出力 : /data3/MORPH/gc/tel/CT_tel-772-16-01

F-7. 文節情報データ（3-C-1.(6)文節情報データ作成,(7)文節情報データ修正）

ファイル名: T0539.DEC
会話ＩＤ : 539

発話 : 10 担当者:

文番号: 100 はい、コンピュータ国際会議事務局でございます。

100 はい 、
感動詞 記 号
100 200

200 コンピュータ国際会議事務局 で ございます 。
固名詞 助動詞 補動詞 記 号
300 400 500 600

発話 : 20 申込者:

文番号: 200 東大の斉藤と申します。

300 東大 の
固名詞 格助詞
700 800

400 斉藤 と
固名詞 格助詞
900 1000

500 申し ます 。
本動詞 助動詞 記 号
1100 1200 1300

発話 : 30 担当者:

文番号: 300 はい、〔あっ〕東大の斉藤先生ですね、{はい、ええ} お世話になっております。

600 はい 、
感動詞 記 号
1400 1500

700 〔 あっ 〕
記 号 間投詞 記 号
1600 1700 1800

800 東大 の
固名詞 格助詞
1900 2000

900 斉藤先生 です ね 、 { はい 、 ええ }
固名詞 助動詞 終助詞 記 号 記 号 感動詞 記 号 感動詞 記 号
2100 2200 2300 2400 2500 2600 2700 2800 2900