

TR-I-0109

係り受け関係を用いた文節候補選択手法

Selecting the Phrase Candidates
Using Dependency Relationship

柿ヶ原康二

森元 暹

Koji KAKIGAHARA Tsuyoshi MORIMOTO

1989.9

概要

文節間の係り受け関係を用いた文節候補選択処理について報告する。音声認識部の認識結果である文節ラティスに対して、係り受け情報を用いて優先度の高い文節候補を順次選択することにより文節認識率の向上を図る。従来から提案されている手法は最適文節候補を効率良く求めるものであるが、限定されたタスクにおいては係り受けの曖昧性が低く、多くの無駄な探索を行う。そこで、係り受け関係に表層表現、正規表現、素性表現の3つのレベルを設けて優先順位を付けることにより、優先度の高い文節候補群を選択することが可能となった。

ATR 自動翻訳電話研究所

ATR Interpreting Telephony Research Laboratories

©ATR 自動翻訳電話研究所 1989

©ATR Interpreting Telephony Research Laboratories

目次

1. はじめに	1
2. 対象とする対話文	2
3. 音声認識結果	2
4. 係り受けによる文節候補の選択処理	2
4.1 係り受け関係	2
4.2 係り受けの条件	3
4.3 係り受けデータ	3
4.4 係り受けの整合度	7
4.5 係り受けのレベル	7
4.6 部分文節ラティス	9
5. 文節候補選択処理手順	11
5.1 係り受け関係のサーチ	11
5.2 文節候補選択処理手順	12
6. 文節候補選択処理実験結果	14
6.1 サンプル会話	14
6.2 文節候補選択処理結果	15
7. 係り受けデータ量	22
7.1 係り受けデータ量と絞り込み結果の関係	22
8. おわりに	27
参考文献	28
<付録 文節候補の絞り込み結果>	29

1. はじめに

電話による、異なる言語間のコミュニケーションをはかる自動翻訳電話システムを実現するには、音声認識、機械翻訳、音声合成の3つの技術の確立が必要である。特に、音声認識には認識誤りがあり、認識率を上げるためには言語情報の効果的な利用が必要である。

音声認識結果が文節ラティスになっている場合に係り受け関係を用いて最適な文節候補を選択する手法がいくつか提案されている[1]-[4]。これらは最適文節候補を効率良く求めるものであり、係り受けの曖昧性が高い場合に有効である。しかしながら限定されたタスクにおいては係り受けの曖昧性は低く、無駄な探索を行ってしまう。また、係り受け関係だけではその文の持つ深い意味などを抽出することは困難であり、そのためにはなんらかのパーサが必要になる。しかし、文節ラティスをそのままパーサの入力とするとその組み合わせ数が膨大なものとなり、満足のいく処理時間が得られない。すなわち、比較的簡単な手法で文節候補を選択し、それをパーサの入力とすることが現実的である。また、文節候補の選択においてその優先度の高いものから順次求めることにより、パーサとの並列処理が可能になる。

本稿では、限定されたタスクでの対話文音声認識結果の文節ラティスから、文節間の係り受け関係を用いて優先度の高い文節から順次候補を選択する手法について報告する。

2. 対象とする対話文

タスクは『国際会議の参加に関する問い合わせ』を想定しており、事務局と質問者との対話を設定している。それを構成している発話は大部分が質問文とその応答文になっている。

3. 音声認識結果

音声認識結果の出力形態は文節ラティスを想定している。例を図1に示す。図1の例は「登録用紙で手続をして下さい」という発話文に対する音声認識結果の文節ラティスである。それぞれの文節の認識結果の候補が第5候補まで出力されている。それぞれの候補の下の数値は認識の確からしさを表すスコアであり、この例では値の大きいものほど正解の確率が高い。

この文節ラティスに対して文節間の係り受け関係を用いて文節候補の選択処理を行う。

発話文	登録用紙で	手続を	して下さい
文節ラティス	登録用紙へ 0.965184 登録として 0.009604 登録用紙は 0.009604 登録用紙で 0.009604 登録用紙から 0.009604	手続も 0.713338 手続を 0.187402 手続の 0.063391 手続と 0.025924 手続への 0.009944	して下さい 0.907620 して下さい 0.062346 7月に 0.010011 7月で 0.010011 7月へ 0.010011

図1 文節ラティスの例

4. 係り受けによる文節候補の選択処理

4.1 係り受け関係

「登録の用紙を至急そちらにお送りします」に対する文節間の係り受け関係を図2に示す。この文には「登録の→用紙」、「用紙を→お送りします」、「至急→お送りします」「そちらに→お送りします」の4つの係り受け関係が存在する。このように文節間の意味的なつながりを考慮すると、図1の文節ラティス

の文節候補の中でどの文節候補が実際に発話されたものであるかを特定することが可能である。すなわち、図1の文節ラティスにおいて「登録用紙へ→する」、「登録として→する」、「登録用紙で→する」、「登録用紙から→する」、「手続も→する」、「手続を→する」等の係り受け関係が考えられるが、これらの係り受け関係の中から「登録用紙で→する」と「手続を→する」の2つの係り受け関係が最も優先度が高いものとして選び出すことできれば発話文を特定できることになる。

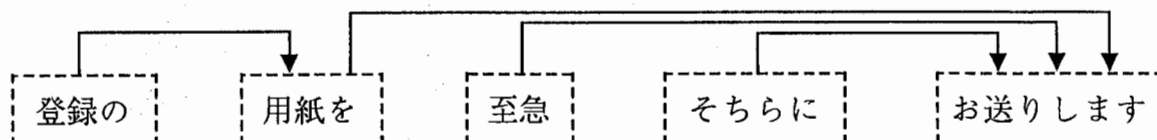


図2 係り受け関係の例

4.2 係り受けの条件

係り受け関係を用いて文節ラティスから正解の可能性の高い文節候補を選択する際の係り受けの条件としてつぎの2つを考える。

- [K1] 係りの唯一性:文末の文節以外の文節は必ずそれより後ろの文節の1つにかかる。
- [K2] 係り受けの非交差性:2つの文節の係り受け関係は他の2つの文節の係り受け関係と交差しない。

4.3 係り受けデータ

係り受けによる文節候補選択を行うための辞書データとして、上記タスクのもとで収集している対話文から係り受け関係を抽出している。現在、44142語の対話文から9029個の係り受けデータを作成している[5]。

図3に係り受け関係データの一部を示す。図中、分類1は係り受けの構文関係、分類2は係り受けの意味関係を表している[6]。頻度はその係り受け関係のデータ中の出現回数を記述するための項である。分類1、分類2は現在の文節候補選択処理には使用していない。

表層表現は対話文データの表現そのままを記述したものである。正規表現はタスク内で同等の内容を示すものを共通の単語で表現したものであり、活用語の場合はその終止形を用いる。意味素性は個々の単語を意味素性で表現したものであり、用いた意味素性を表1に示す。

分類1	分類2	頻度	表層表現	正規表現	意味素性
01	PRD	000	何-でしょうか : でしょうか	: 何 : です	: div/div ; : pre/des
01	RNG	000	事-で : 聞き	: 事 : 聞く	: abs/log ; : pre/kik
01	CON	000	聞き : たいのですが	: 聞く : たい	: pre/kik ; : pre/tai
02	RNG	000	場-の : 事で	: 場 : 事	: loc/ins ; : abs/log
01	SPF	000	駅-からは : 行けば	: 駅 : 行く	: loc/ins ; : pre/yuk
04	CND	000	行け-ば : 宜しいですか	: 行く : 宜しい	: pre/yuk ; : adj/yrs
04	MAN	000	どう : 行けば	: どう : 行く	: div/div ; : pre/yuk
01	TOO	000	地下鉄-で : 行って	: 地下鉄 : 行く	: con/tra ; : pre/yuk
01	SPF	000	駅-から : 行って	: 駅 : 行く	: loc/ins ; : pre/yuk
01	SPT	000	駅-まで : 行って	: 駅 : 行く	: loc/ins ; : pre/yuk
01	SPF	000	そこ-から : 乗り	: 其処 : 乗る	: loc/pos ; : pre/nru
01	OBJ	000	番-か : 乗り	: 番 : 乗る	: abs/sig ; : pre/nru
01	OBJ	000	タクシー-に : 乗り	: タクシー : 乗る	: con/tra ; : pre/nru
02	SEL	000	バス-の : 番か	: バス : 番	: con/tra ; : abs/sig
01	OBJ	000	時間-が : かかりますか	: 時間 : 掛る	: tim/tim ; : pre/kkr
01	PRD	000	分-位です : です	: 分 : です	: tim/tim ; : pre/des
01	TOO	000	バス-で : 行くよりも	: バス : 行く	: con/tra ; : pre/yuk

図3 係り受けデータの例

表1 意味素性一覧

名詞	抽象概念 (abstract,abs)	論理(logic,log); 事、理論、方法 状態(state,sta); 様子、場合 言語(language,lan); 日本語、何語 学術(learning,lea); 分野、専門 意向(intention,int); 興味、希望 価値(value,val); 必要 記号(sign,sig); 名前 労役(labor,lab); 用件
	具象物 (concrete,con)	文書(document,doc); 要約、原稿 交通(transport,tra); バス、地下鉄 物品(article,art); 写真 貨財(money,mon); お金、カード 任意(various,var); これ、どれ
	人間 (human,hum)	人間(human,hum); 発表者、学生 任意(various,var); だれ、どなた
	場所 (location,loc)	施設(institution,ins); ホテル、会場 地域(region,reg); 京都、海外 位置(position,pos); 間、方、近く 任意(various,var); どこ
	時間 (time,tim)	時間(time,tim); 間、後、今、時間 任意(various,var); 何時、いつ
	数 (number,num)	数量(amount,amo); 字数、何名 単位(unit,uni); 円、ドル 費用(cost,cos); 登録費、参加料 価格(price,pri); 無料、いくら
	行為 (action,act)	予定 act/yti、制限 act/sgn、 参加 act/san、発表 act/hap (総数42)
	任意 (diverse,div)	ほか、なに
動詞 (predicate,pre)	申し込む pre/msk、来る pre/kur、 書く pre/kak、持つ pre/mot (総数48)	
形容詞 (adjective,adj)	近い adj/tik、多い adj/ooi、 遠い adj/toi (総数16)	
副詞 (adverb,adv)	既に adv/adv、ちょうど adv/adv、 もう adv/adv (総数41)	

この係り受けデータに対して係り受け関係のある文節間の距離を示したものが表2である。表2からわかるように62.4%の文節がその隣りの文節に係っており、文節ラティスの文節候補の係り受け関係を調べる場合、文節間の距離が係り受けの優先度に大きく影響している。

表2 係り受け関係にある文節間の距離

距離	比率(%)
1	62.4
2	17.2
3	10.3
4以上	10.1

係り受けデータに対して正規表現での出現頻度を示したものが表3である。係り受け関係が対話文データに一回だけ出現するものが約1割存在し、1回と2回出現する係り受け関係で全体の2割5分を占めている。

表3 係り受けデータの内訳(正規表現)

等しい係り受けの数	1	2	3	4	5	6	7	8	9	10
係り受けの数	4189	632	278	120	58	43	31	21	11	11

11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
4	9	4	6	4	4	1	2	3	1	3	2	0	1	0

26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
0	1	0	2	0	1	0	1	0	1	0	1	0	0	0

	45		52	53		71
	1		1	1		1

4.4 係り受けの整合度

文節ラティスの文節候補間の係り受け関係を求める際に、どの文節間の係り受け関係の発話された可能性が高いかを知る必要がある。言い換えると、多数存在する係り受け関係の優先順位を決める必要がある。そのためには係り受けの整合度 $P(x,y)$ (x :係り元の文節候補、 y :係り先の文節候補)を調べるのであるが、係り受けの出現頻度 $F(x,y)$ 、係り受けの距離 $D(x,y)$ 、各文節の認識スコア $S(y)$ を考慮する。すなわち整合度は次式により求める。

$$P(x,y) = F(x,y) - W1 * D(x,y) + W2 * S(y) \quad (W \text{は定数}) \quad (1)$$

係り受けの出現頻度は係り受けデータより求める。係り受けデータ中に出現する頻度の高い係り受け関係は発話される可能性が高い。係り受けの距離については表2に示されているように近くの文節に係る可能性が高く、整合度の式において係り受けの距離は負の要因として働く。認識スコアは音声認識の認識の確からしさを示すものであり、当然認識スコアの高いものは発話された可能性が高い。

整合度の式の定数($W1, W2$)の値を設定することは難しい問題であるが、現在はそれぞれの項の位が揃うように設定している。

4.5 係り受けのレベル

文節ラティス内の文節に対して係り受け関係を調べる際に、表4に示す係り受け関係のレベルを用いる。表層表現、正規表現、意味素性表現は先に説明したように係り受けデータに記述されているものである。例を表5に示す。

係り受け関係をサーチする場合、係り受け関係のレベルの順位の高いものから優先的に係り受け関係を付ける。例えばある文節を係り元の文節として係り先の文節をサーチする場合、係り受けのレベルの順位1からサーチを開始する。係り受け関係が付けられない場合は順位2でサーチする。係り受け関係が成立するまで順位を下げ、成立すれば係り先の文節のサーチは終了する。

同じ係り受け関係のレベルで多数係り受け関係が成立する時は、先に説明した係り受けの整合度の最も高いものに係り受け関係を付ける。

表4 係り受け関係のレベル

レベル	係り元文節	係り先文節	順位
表層表現	表層表現	正規表現	1
	表層表現	だ文表現	2
正規表現1	正規表現+付属語	正規表現	3
	正規表現+付属語	だ文表現	4
素性表現1	素性表現+付属語	素性表現	5
	素性表現+付属語	だ文表現	6
正規表現2	正規表現	正規表現	7
	正規表現	だ文表現	8
素性表現2	表層表現	素性表現	9
	表層表現	だ文表現	10

表5 係り受け関係のレベルの例

レベル	表現	表現
表層表現	受付に	登録用紙を
正規表現+付属語	事務局+に	用紙+を
素性表現+付属語	human+に	con/doc+を
正規表現	事務局	用紙
素性表現	human	con/doc

4.6 部分文節ラティス

文節ラティス内の文節に対して係り受け関係を調べる際に、係り受け関係の付けられていない部分文節ラティスが存在する。これは図4に示されるように係り受け関係のある文節には含まれている。このとき部分文節ラティスの右端の文節候補は、係り受けの条件を考慮すると、部分文節ラティスに右接する文節が部分文節ラティスに左接する文節の係り先の文節に到達するまでの係り受け関係のある文節 $j_h(1 \leq h \leq m)$ のいずれかに係る。図4では文節候補群 k は文節 j_1 、文節 j_2 、文節 j_3 のいずれかに係る。とくに文節 j_h の系列の文節数が1つの場合には図5に示されるように部分文節ラティスの右端の文節候補は文節 j_1 のみに係る。

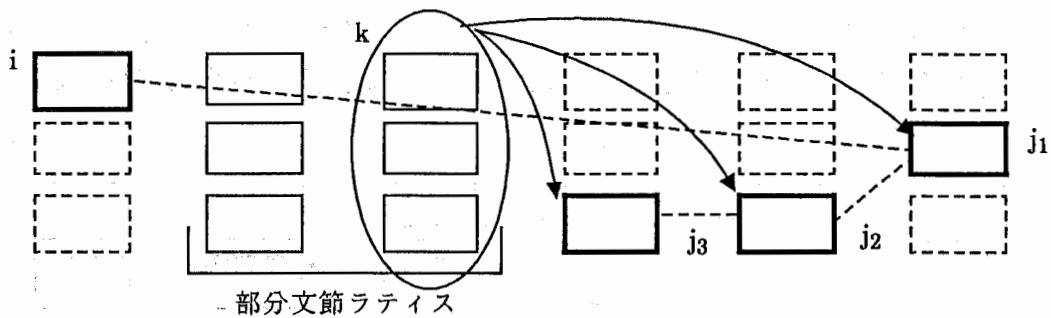


図4 部分文節ラティス

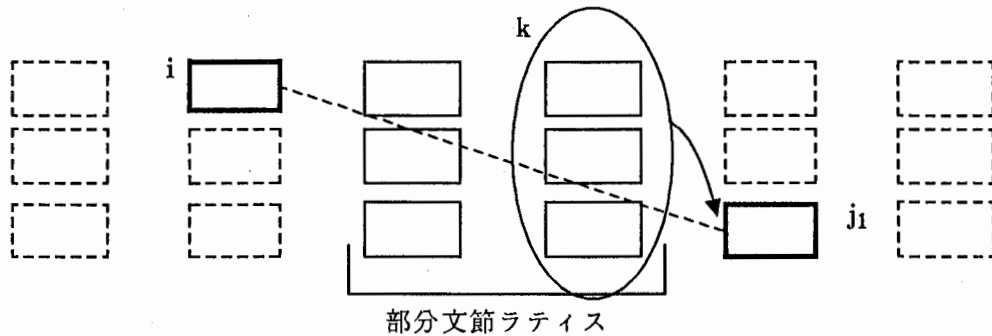


図5 部分文節ラティス

文頭を含む部分文節ラティスに対しても同様である。(図6)

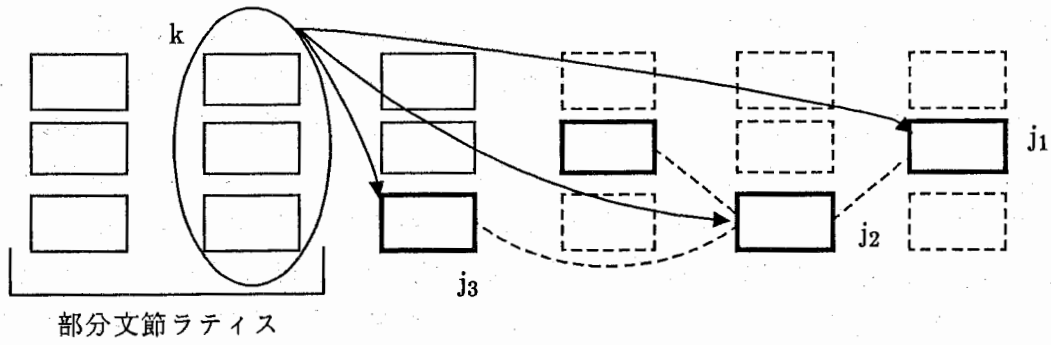


図6 部分文節ラティス

文末を含む部分文節ラティスにおいては図7に示されるようにそのラティス内のいずれかの文節に文末を含む部分文節ラティスに左接する文節に係る。

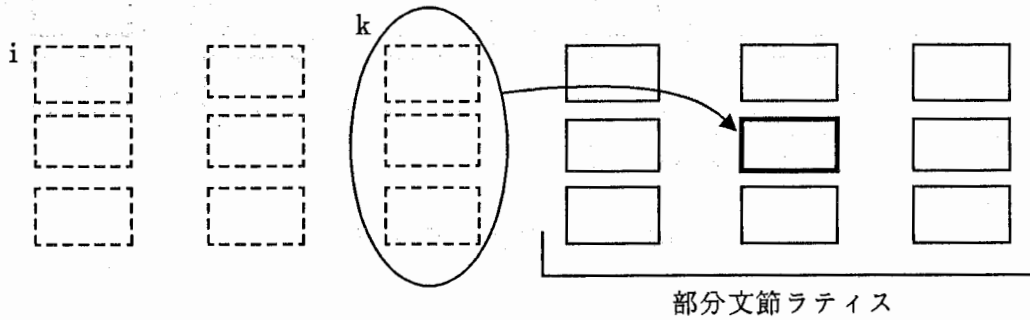


図7 部分文節ラティス

5. 文節候補選択処理手順

5.1 係り受け関係のサーチ

ある文節と他の文節との係り受け関係をサーチする場合、次の2種類のサーチを考える。

- (1) 右方向サーチ：ある文節*i*に右接する部分文節ラティスに対して、文節*i*と係り受け関係があり、優先度の最も高い文節*j*をサーチする。このとき係り受けのレベルの高いものから順にサーチし、なければ係り受けのレベルを下げてサーチする。同じ係り受けのレベルで2つ以上の係り受け関係が存在する場合には、それぞれの係り受け関係の整合度を求め、整合度の最も高いものを求める係り受け関係とする。1つの文節の文節候補群に自立語の部分の同じ文節が複数存在する場合には、そのグループに係るものとする。

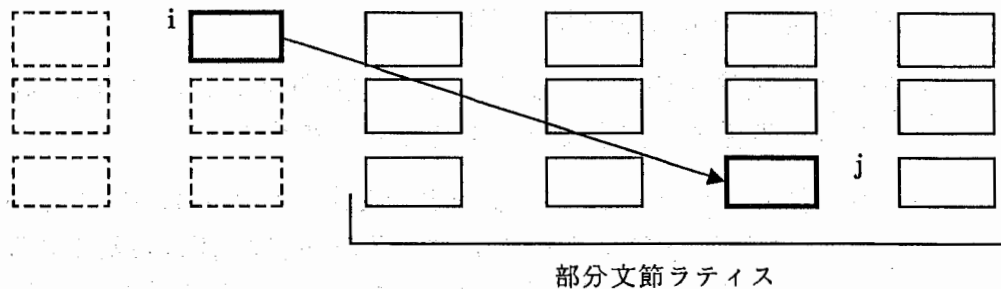


図8 部分文節ラティス

- (2) 左方向サーチ：ある文節*j*に左接する部分文節ラティスに対して、その右端の文節候補の中で文節 $j_h (1 \leq h \leq m)$ と係り受け関係があり、優先度の最も高い文節*k*をサーチする。右方向サーチと同様に、係り受けのレベルの高いものから順にサーチし、なければ係り受けのレベルを下げてサーチする。同じ係り受けのレベルで2つ以上の係り受け関係が存在する場合には、それぞれの係り受け関係の整合度を求め、整合度の最も高いものを求める係り受け関係とする。

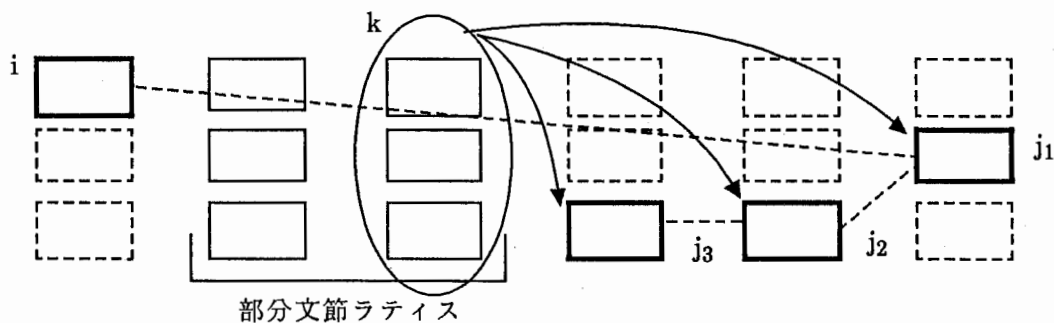


図9 左方向サーチ

5.2 文節候補選択処理手順

上で説明した部分文節ラティスの特性とサーチの方法を用いて文節候補を選択する手順を図10に示す。この手法は優先度の高い文節から順に文節候補群を選択するものであり、認識スコアが高く明らかに正解である文節候補が存在する場合や、文脈情報などから出現する確率の高い文節候補が存在する場合に、その文節から係り受けのサーチを開始する。まず右方向サーチにより文末の文節まで係り受けの整合度の高い文節系列を求め、次に文末寄りの部分文節ラティスに対して左方向サーチにより係り受けの整合度の高い文節をサーチし、部分文節ラティスがなくなるまで繰り返す。最適候補は求まらないが比較的はやい段階で正解の文節群が選択される。

上で説明した係り受けサーチの方法を用いて文節候補を選択する手順を以下に説明する。

- (step1) 文節ラティスからサーチのスタートの文節を選択する。これは認識スコアの最も高いものから順に選択する。
- (step2) スタートの文節を係り元の文節として右方向サーチにより文末の文節まで係り受け関係の成立する文節系列をサーチする。
- (step3) 係り受け関係の付けられていない部分文節ラティスに対して、文末に最も近いものから左方向サーチにより係り受け関係の成立する文節をサーチし、係り受け関係の付けられていない部分文節ラティスがなくなるまで繰り返す。

スタートの文節を次に認識スコアの高いものに変えて(step1)から(step3)を繰り返す。

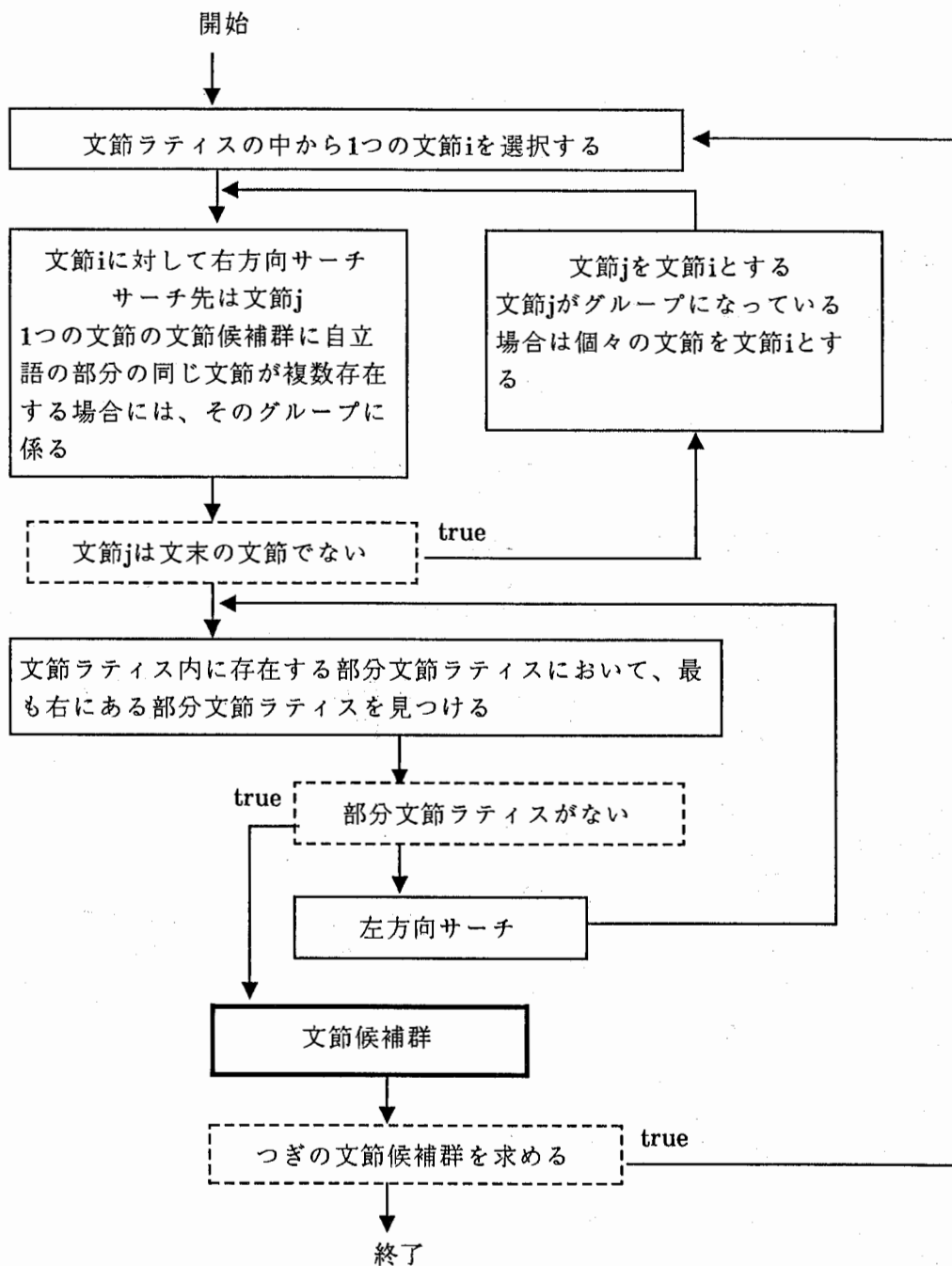


図10 文節候補選択処理手順

6. 文節候補選択処理実験結果

6.1 サンプル会話

実際にサンプル文に対して文節候補選択を行った。以下に実験に用いたサンプル会話を示す(a)。

サンプル会話AB

- (2) そちらは 会議事務局ですか。
1 2
- (5) 会議に 申し込みたいのですが。
3 4
- (6) 登録用紙は 既に お持ちでしょうか。
5 6 7
- (10) それでは 登録用紙を お送り致します。
8 9 10
- (11) ご住所と お名前を お願いします。
11 12 13
- (12) 住所は 大阪市 北区 茶屋町 二十三です。
14 15 16 17 18
- (13) 名前は 鈴木真弓です。
19 20
- (15) 登録用紙は 至急 送らせていただきます。
21 22 23
- (16) 分からない 点が ございましたら いつでも お聞き下さい。
24 25 26 27 28
- (18) それでは 失礼します。
29 30
- (21) こちらは 会議事務局です。
31 32
- (22) 会議に 参加したいのですが。
33 34
- (23) どうすれば よろしいですか。
35 36
- (24) 先ず 登録用紙で 手続きを していただかなくては なりません。
37 38 39 40 41
- (25) もう 登録用紙は お持ちでしょうか。
42 43 44
- (27) 用紙を 送ってください。

(a) 実際のサンプル会話には1文節の文が存在するが、係り受けによる文節候補選択処理の対象とならないのでここでは抜いている。

	45	46		
(28)	では	ご住所と	お名前を	お願いします。
	47	48	49	50
(30)	名前は	清水太郎です。		
	51	52		
(32)	参加料は	要るのでしょうか。		
	53	54		

6.2 文節候補選択処理結果

表6は文節ラティスから文節候補を選択する際に、係り受けレベルのどのレベルまでレベルを下げれば正解の文節が選択されるかを示すものである。係り元の文節が表層表現で係り先の文節が正規表現の場合には正解の文節が選択できるのは3文しかないが、係り受けのレベルを下げることにより正解の文節候補の選択できる文の数が増えている。このことは文節候補の表現を表層表現から正規表現、意味素性表現というように抽象度を上げて係り受け関係をサーチすることにより、係り受けデータの項目数が少なくても係り受け関係をサーチできることを示している。

表7は係り受けレベルを固定した場合に正解の文節候補が選択される順位を示したものである。音声認識結果の文節ラティスでの正解の文節の順位と、すべての係り受けレベルを許した場合の順位、正規表現と意味素性レベルを許した場合の順位、意味素性レベルのみの順位、付属語の制約を付けた場合の順位を示している。文節ラティスにおいて第1位の文節認識率が90.7%、第2位までが98.1%である。表7の場合は音声認識結果が良いので文節ラティスでの正解の順位が高く、文節候補選択による順位の向上はあまり見られない。係り受けレベルの制限に対しては、文節番号26、38において意味素性のみを用いて文節候補選択をするよりも表層表現レベル、正規表現レベルを用いた方が正解の文節の順位が高いことが分かる。このことは表層表現レベルや正規表現レベルで係り受け関係を付けられるものは、意味素性レベルで係り受け関係を付けられるものより実際に発話された可能性が高いことを示しており、係り受け関係のレベルを設けていることの有効性が示されている。付属語考慮の欄は係り受け関係をサーチする際に係り元の文節の付属語が等しい場合のみ係り受け関係を付けるようにして処理した結果である。-記号は付属語を考慮すると係り受け関係が付けられない文であることを示している。

表8は音声認識結果の文節ラティスに対して、その文節候補の順番を逆にし、認識率を下げて同様の文節候補選択処理を行った結果である。文節ラティスにおいて第1位の認識率が2.0%、第2位までが9.3%、第3位までが24.1%である。表中の△記号は文末の文節に対応しており、係り受け関係を調べる際に係り先の文節は自立語部分のみを考慮しており文末表現の区別が不可能であるため順位のカウン

トからは除外している。表7の結果では文節ラティスでの正解の文節の順位が高く、文節候補選択の効果がそれほど見られなかったが、表8ではその効果が顕著である。特に付属語を考慮した場合には正解の文節候補を選択できる可能性は高くなる。しかしながら付属語の制約を設けると係り受け関係を付けられない文が多くなり、これは係り受けデータの充実が必要である。

図11は係り受けのレベルごとの文節候補選択の可能な文の数の割合と、その時の正解の文節の平均順位を示している。係り受け関係のレベルを下げると係り受け関係の成立する組合せが増え、文節候補選択の可能な文の数が増えているが、正解の文節の順位は下がっている。すなわち、係り受け関係の高いレベルで文節候補選択処理を行える文は、選択された文節が正解である可能性が高く、低いレベルで選択処理を行うと、正解の文節を選択する可能性は低くなる。

図12は係り受け関係のレベルを制限した場合の正解の文節の平均順位の変化を示している。Aは文節ラティスの文節候補の順位が認識結果のままのもの(表7)、Bは文節候補の順位を逆にしたもの(表8)に対して文節候補選択処理を行った結果である。ABともに第5位までの文節認識率は100%である。表7、表8で見たようにAの場合は認識率が高いので文節候補選択の効果はほとんど見られないが、Bのように認識率が低い場合では正解の文節の順位の上昇が顕著である。

表9は音声認識結果の文節ラティスに対してその文節候補の順番を逆にしたものについて、文節候補選択処理による文節数の削減のようすを付属語を考慮しない場合と付属語を考慮した文節とに分けて示したものである。付属語を考慮した文節に対しては平均文節数が1.26と非常に小さい値になっており付属語の制約が係り受けの優先度に強く影響していることが分かる。

表10は以上の結果を認識率の変化として示したものである。係り受け関係を用いた文節候補の選択処理は音声認識率の向上に役立ち、特に認識率の低い場合に有効である。

表6 レベルごとの係り受け関係による文候補生成

文番号	2	5	6	10	11	12	13	15	16	18	21	22	23	24	25	27	28	30	32
表層→正規		○										○				○			
表層→です	○						○				○							○	
正規+付→正規		○	○	○	○							○			○	○			○
正規+付→です	○						○				○							○	
素性+付→素性		○	○	○	○							○			○	○	○		○
素性+付→です	○						○				○							○	
正規→正規	○	○	○	○	○			○	○	○	○	○	○	○	○	○			○
正規→です	○						○				○	○						○	
素性→素性	○	○	○	○	○		○	○	○	○	○	○	○	○	○	○	○	○	○
素性→です	○					○	○				○	○			○			○	○
表層→正規,です	○	○					○				○	○				○		○	
正+付→正規,です	○	○	○	○	○		○				○	○				○		○	○
素+付→素性,です	○	○	○	○	○		○				○	○				○	○	○	○
正規→正規,です	○	○	○	○	○		○	○	○	○	○	○	○	○	○	○		○	○
素性→素性,です	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○
付属語を考慮	○	○	○	○	○		○				○	○			○	○	○	○	○

表7 各文節に対する正解候補の選択順位

文節番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
文節ラティス	1	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	4	1
全レベル	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	2	2
正規、素性	1	1	1	1	1	2	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	2	2
素性	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1	1	2	2
付属語考慮	1	1	1	1	1	1	1	1	2	1	1	1	1	-	-	-	-	-	1	1	-	-	-	-	-	-	-	-	-	-	-
文節番号	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54							
文節ラティス	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1							
全レベル	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1							
正規、素性	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1							
素性	1	1	1	1	1	2	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1							
付属語考慮	1	1	1	1	-	-	-	-	-	-	-	1	1	1	1	1	1	1	1	1	1	1	1	1							

表8 各文節に対する正解候補の選択順位 (認識候補を逆に並べ変えたもの)

文節番号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	
文節ラティス	5	2	5	5	5	2	5	5	3	5	2	5	5	5	3	5	5	4	5	5	3	5	5	5	5	5	5	5	3	2	5
全レベル	1	△	2	△	3	1	△	2	1	△	2	5	△	1	2	5	5	△	2	△	3	1	△	3	1	2	2	△	2	△	
正規、素性	1	△	2	△	3	1	△	3	1	△	2	5	△	1	2	5	5	△	2	△	3	1	△	3	1	2	4	△	2	△	
素性	5	△	2	△	3	1	△	3	3	△	2	5	△	1	2	5	5	△	3	△	3	1	△	3	2	2	4	△	2	△	
付属語考慮	1	△	1	△	1	1	△	1	1	△	1	3	△	-	-	-	-	-	1	△	-	-	-	-	-	-	-	-	-	-	
文節番号	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54							
文節ラティス	5	4	5	3	5	3	4	5	5	1	5	5	5	5	3	5	5	3	5	5	5	4	5	5							
全レベル	1	△	2	△	1	△	2	5	1	1	△	1	2	△	1	△	2	1	5	△	2	△	4	△							
正規、素性	1	△	2	△	1	△	2	5	1	1	△	1	2	△	1	△	2	1	5	△	2	△	4	△							
素性	1	△	2	△	1	△	2	5	5	1	△	1	3	△	2	△	2	1	5	△	3	△	5	△							
付属語考慮	1	△	2	△	-	-	-	-	-	-	-	1	1	△	1	△	2	1	2	△	1	△	1	△							

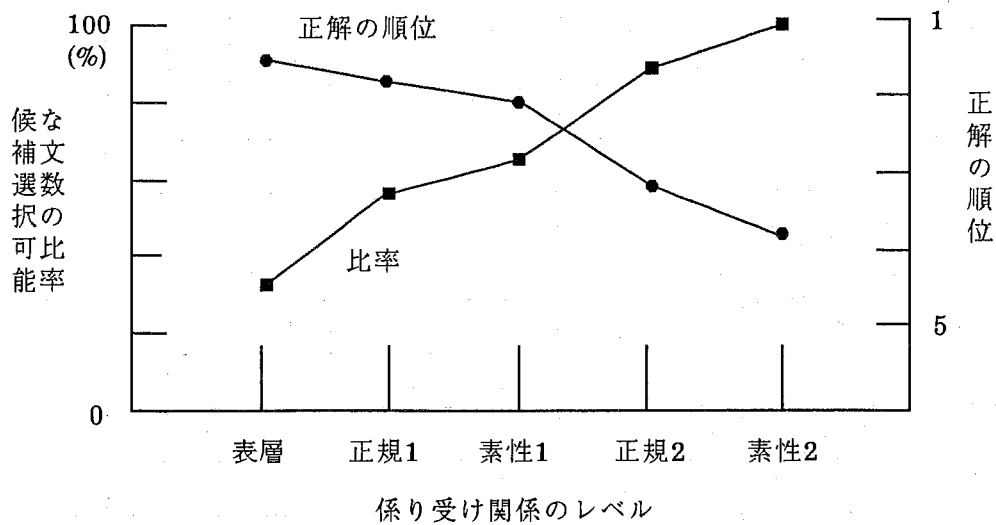


図11 係り受けレベルと絞り込みの関係

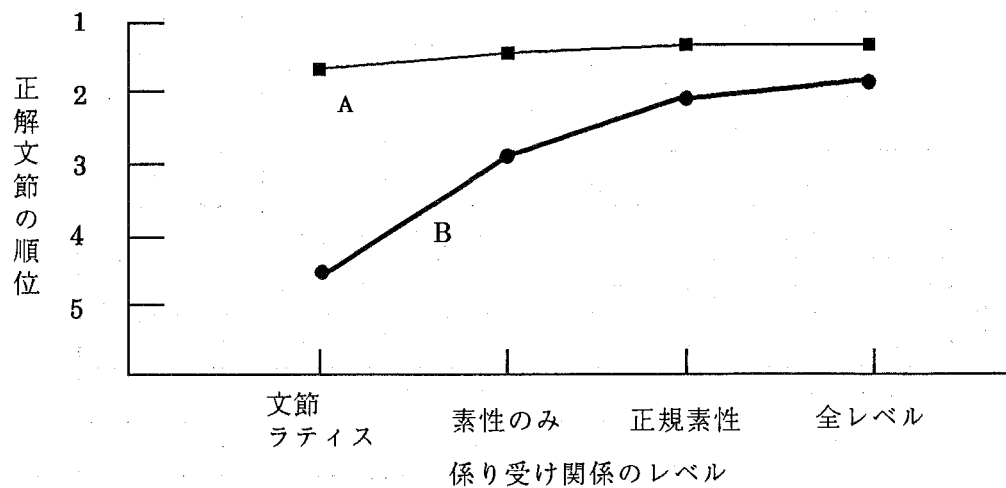


図12 係り受けレベルと平均文節数の関係

表9 平均文節数 (認識候補を逆に並べ変えたもの)

	文末以外の文節	付属語を考慮した文節のみ
文節ラティス	4.31	4.37
全レベル	2.20	2.11
正規、素性	2.26	2.11
素性	2.74	2.74
付属語考慮	-	1.26

表10 文節候補選択処理による認識率の変化

	A	B
1位	88.9 / 90.7	37.1 / 2.0
2位	100 / 98.1	74.3 / 9.3
3位	100 / 98.1	82.9 / 24.1
4位	100 / 100	85.7 / 31.5
5位	100 / 100	100 / 100

選択処理(%) / 文節ラティス(%)

7. 係り受けデータ量

音声認識結果の文節ラティスに対して、係り受け関係を用いて候補の絞り込みを行う際に、どれくらいの係り受けデータがあれば十分な絞り込みが行えるかを調べる必要がある。現在、44142形態素の会話データから抽出されたのべ9029項目の係り受けデータを用いて文節候補選択処理を行っている。

7.1 係り受けデータ量と絞り込み結果の関係

係り受けのデータ量を変化させた場合に文節候補数がどのように変化するかを示したのが図13である。横軸に係り受け関係ののべ数を、縦軸に文節候補選択処理を行った後の文節候補数を示している。この時の文節候補数は、各々の文節に対して正解の文節候補が出現するまでの候補数の平均を表している。

図14は文節ラティスの文節候補の順位を逆にして認識率を下げた場合の文節候補数がどのように変化するかを示したものである。

図15は係り受けのデータ量を変化させた場合に、絞り込めない文節の数がどのように変化するかを示したものである。すなわち、どこにも係らない文節候補群の数を示している。

図16は絞り込み結果の最大の文節候補数の変化を示している。この場合、絞り込めない文節が1つでも存在すれば最大文節候補数は5になる(文節ラティスの文節候補数は5である)。

図13、図14から、係り受けデータの数が3000くらいから文節候補数はそれほど減少していないのが分かる。また図15から、絞り込めない文節数も係り受けデータが3000からは1つだけであり、係り受けデータが8000からは絞り込めない文節はなくなる。これらのことから係り受けデータは3000程度で十分な絞り込み結果が得られていると考えられる。

しかしながら、この絞り込み処理を音声言語統合化システムに組み込む場合には、1つでも絞り込めない文節が存在すると、文節候補数を5にしなければならなくなり実用性はなくなってしまう(係り受けデータ数が9000の場合には最大文節候補数は2であり、絞り込んだ後の文節数は高々2である)。言い換えるならば、絞り込めない文節が発生しないだけの係り受けデータが必要であり、現状では9000項目以上の係り受けデータが必要である。

9000項目以上にデータを増やした場合に絞り込み結果がどのように変化するかは現状では判断不可能であり(データが増えると絞り込み結果が悪くなる可能性がある)、今後データを増やして考察する必要がある。

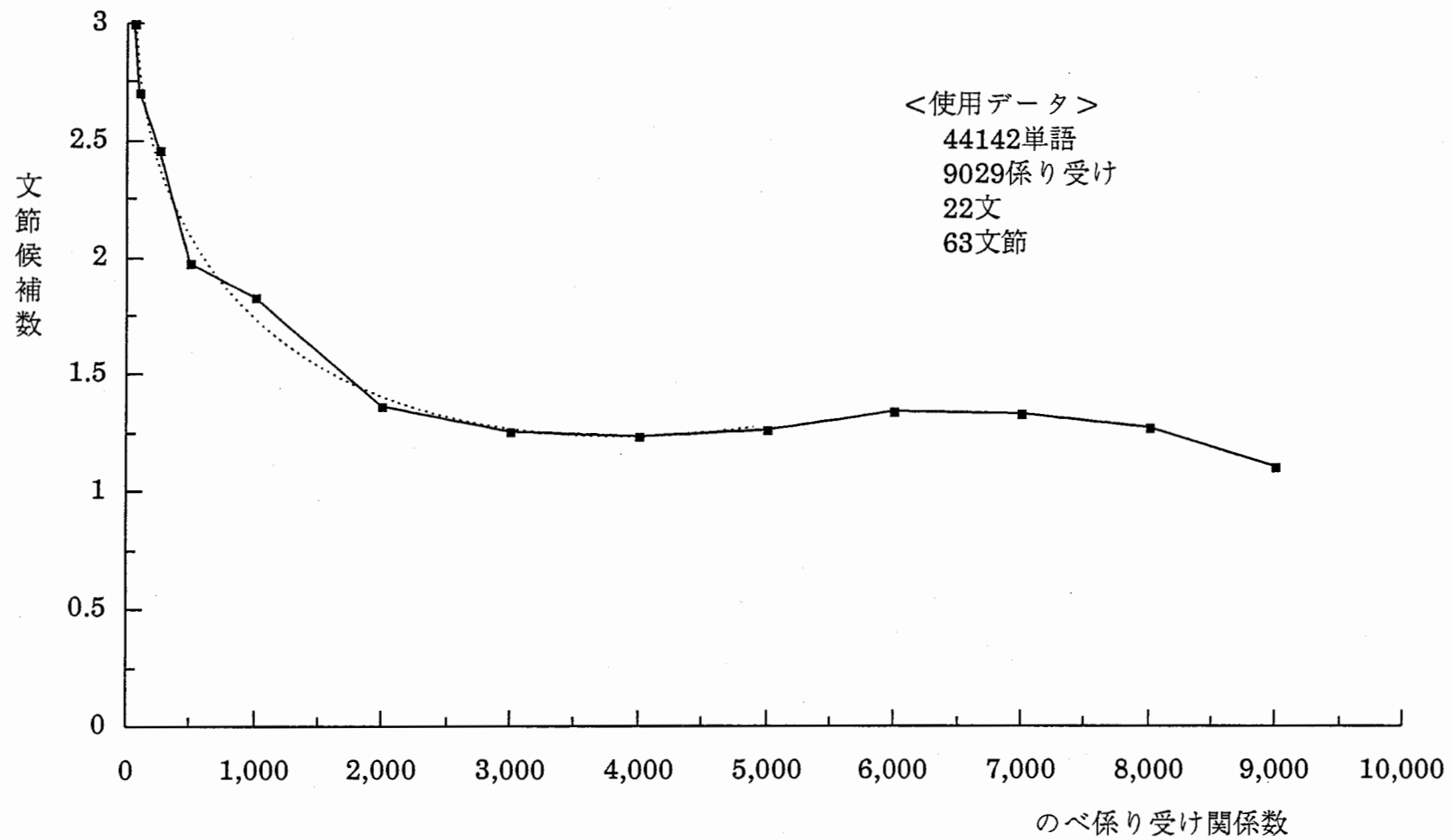


図13 係り受けの数と文節候補数の関係

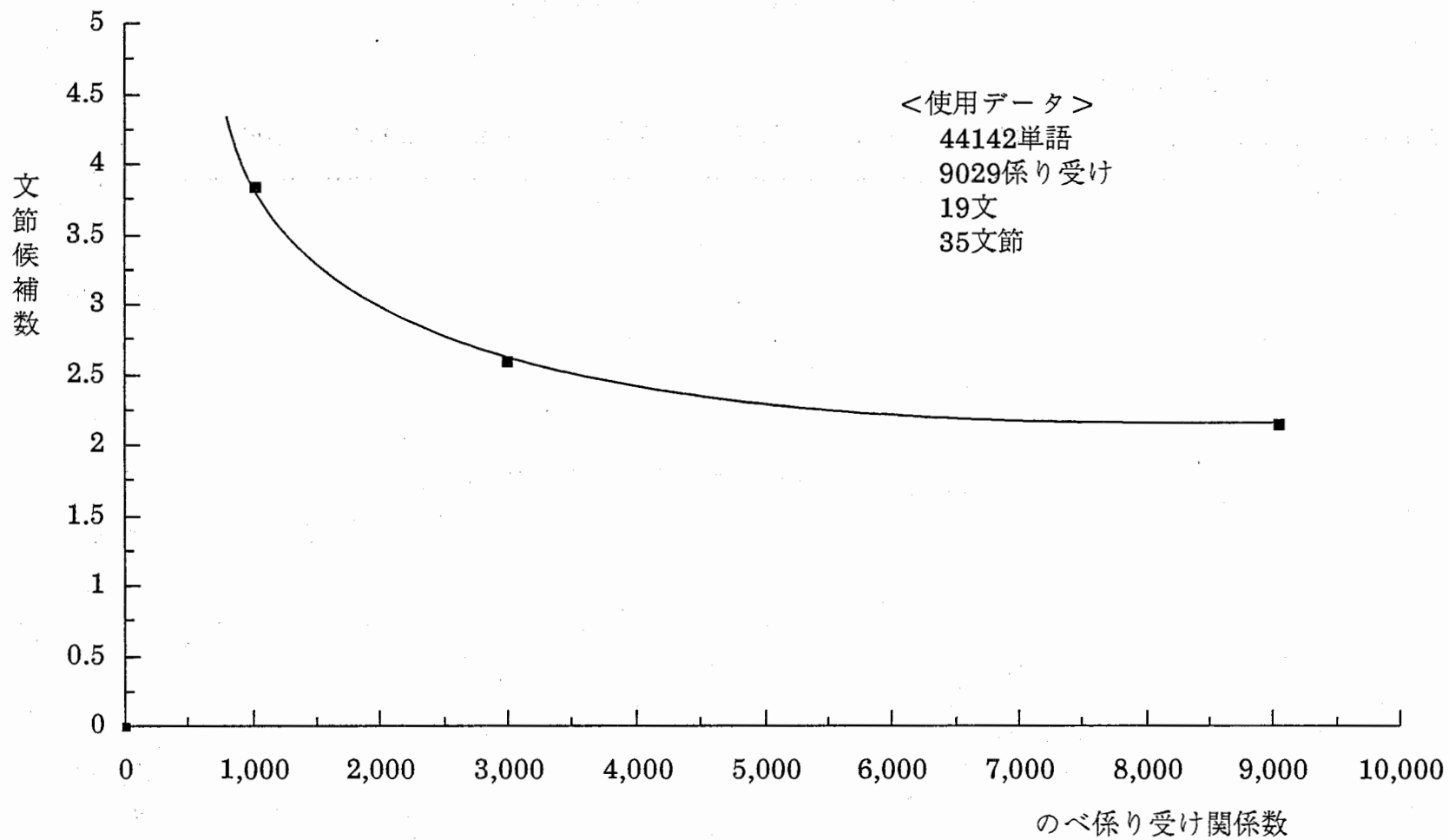


図14 係り受けの数と文節候補数の関係 (認識候補を逆に並べたもの)

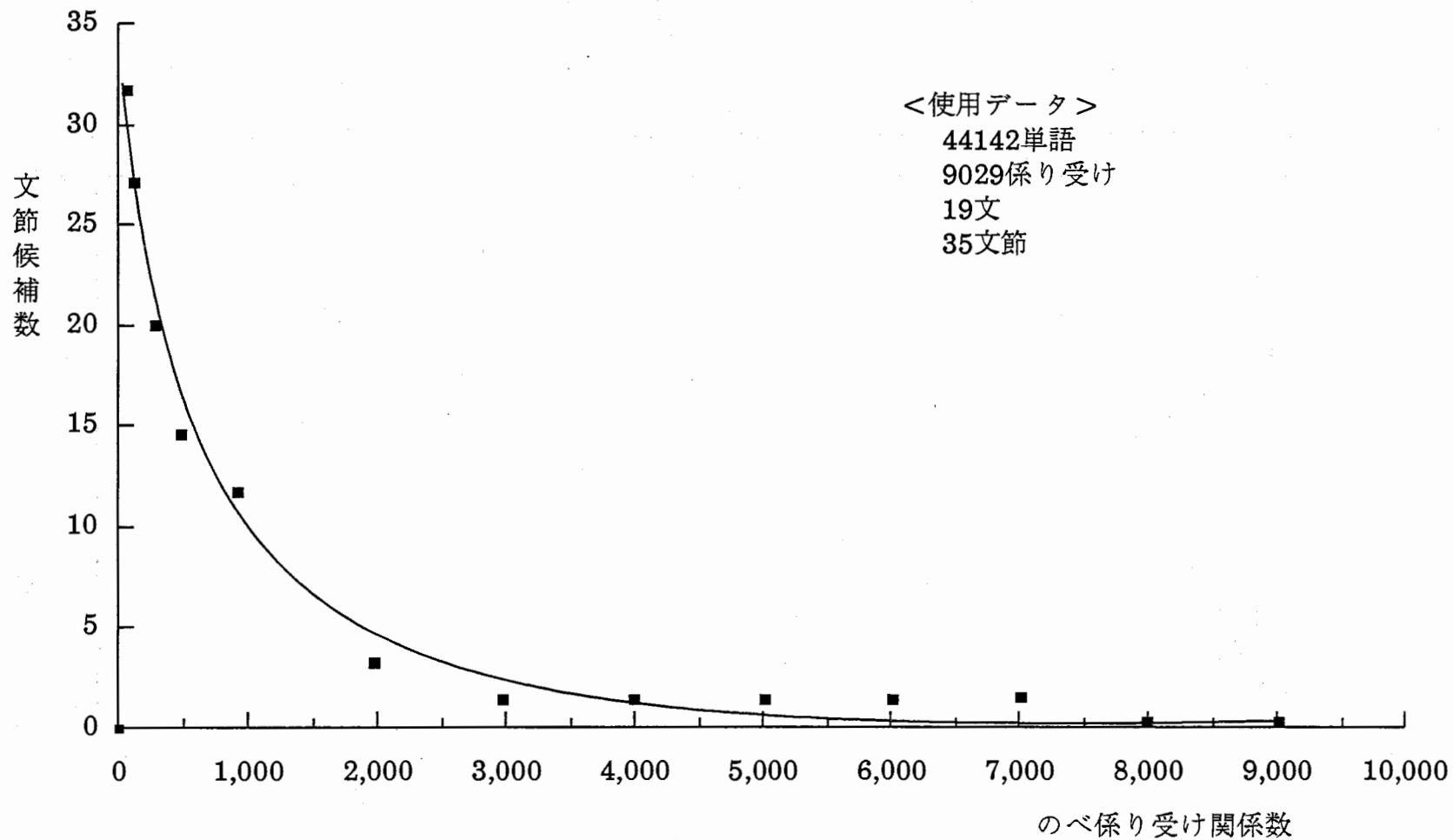


図15 係り受けの数と絞り込めない文節数の関係

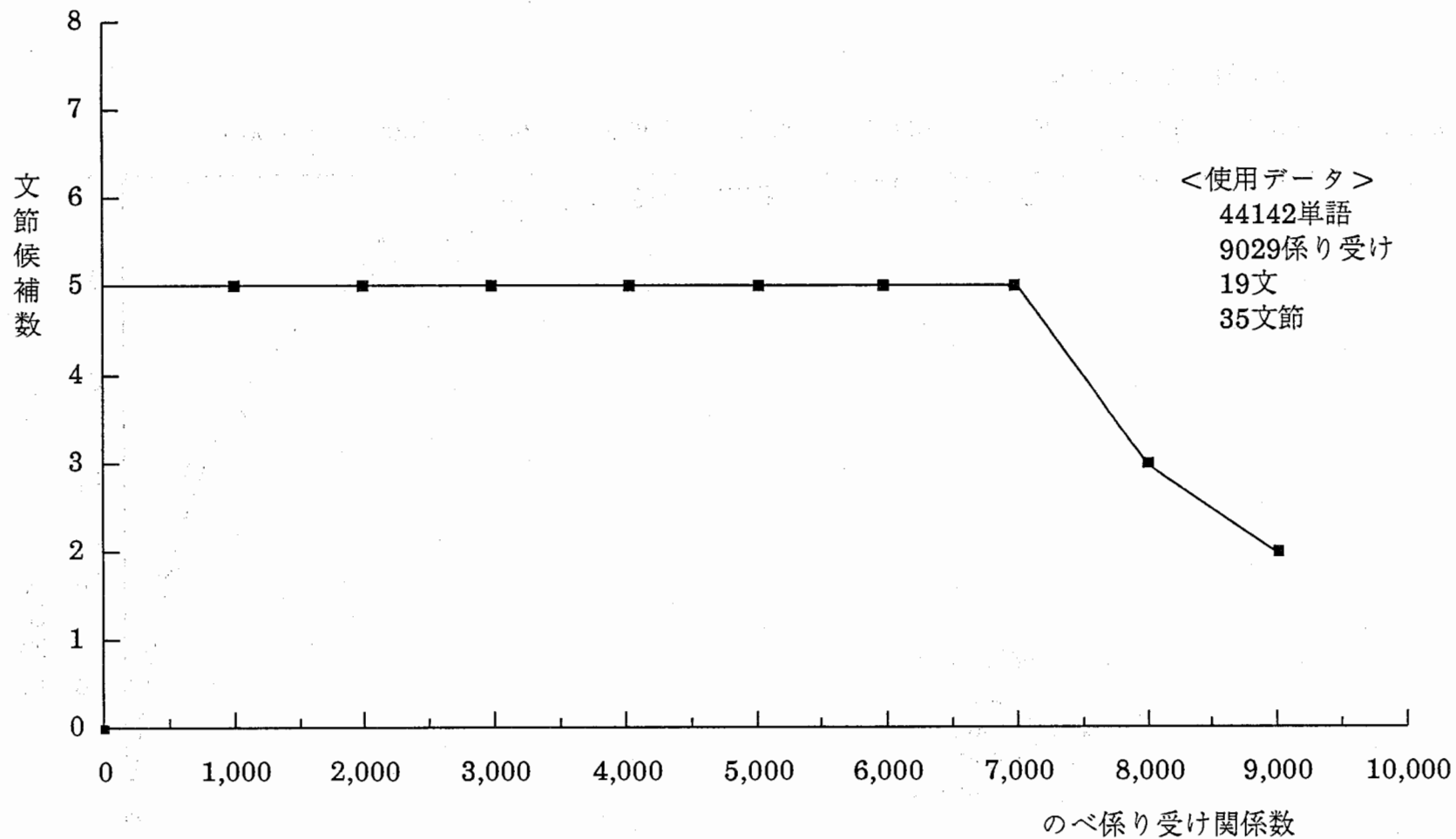


図16 係り受けの数と最大文節候補数の関係

8. おわりに

音声認識結果の文節ラティスに対して、文節間の係り受け関係を用いて文節候補を選択する手法について報告した。タスクが限定されている場合に文節間の係り受け関係は候補選択に有効であることが実験により実証された。しかしながら、サンプル会話は比較的簡単なものであり、その数も少ないので今後はさらに多くの会話について検討する必要がある。また、整合度の計算式の評価や文末表現の取扱い、意味素性の付与など、多くの検討課題が残されている。今後はこれらに対して検討を加えて行きたい。なお付録にサンプル会話に対する文節候補の絞り込み結果を示す。

参考文献

- [1] 尾関：「多文節間の係り受け整合度に基づき最適文節列を選択する多段決定アルゴリズム」信学論J71-D(1988-04)
- [2] 松永他：「係り受け規則により候補を限定した文節音声の認識」信学技法SP87-29(1987-06)
- [3] 中川他：「音節標準パターンと逆時間向き係り受け解析法を用いた日本語文音声の認識」信学論J70-D(1987-12)
- [4] 好田：「文節ラティス上で最適文節列を選択するためのアルゴリズム」信学技法SP86-72(1986-12)
- [5] 柿ヶ原、泉：「文節候補選択処理のための係り受けデータ」ATRテクニカルレポート TR-I-0111 (1989)
- [6] 井上、小倉、森元：「言語データベース用格・係り受け意味体系」ATRテクニカルレポート TR-I-0029 (1988)
- [7] 柿ヶ原、森元：「係り受け関係を用いた文節候補選択についての一考察」情報処理学会第38回全国大会 (1989)
- [8] 柿ヶ原、森元：「SL-TRANSにおける文節候補の削減」情報処理学会第39回全国大会 (1989)
- [9] 柿ヶ原、森元：「文節間の係り受け関係を用いた文節候補選択手法」音響学会秋季研究発表会 (1989)
- [10] 柿ヶ原、桶谷、泉：「SL-TRANSにおける文節候補選択処理、問合せ処理」ATRテクニカルレポート TR-I-0110 (1989)
- [11] 荻野：「文法と意味I」朝倉書店 (1983)
- [12] 南：「文法と意味II」朝倉書店 (1985)
- [13] 大野：「類語国語辞典」角川書店 (1985)

<付録 文節候補の絞り込み結果>

- (1) もしもし。
 (2) そちらは 会議事務局ですか。

<文節ラティス>

そちらは 会議事務局ですか
 いつからは 会議事務局です
 そちらから
 そちらとは
 そちらを

<絞り込み結果>

そちらは 会議事務局ですか
 いつからは

- (3) はい。
 (4) そうです。
 (5) 会議に 申し込みたいのですが。

<文節ラティス>

会議に 申し込みたいのですが
 八に 申し込みたいのですか
 八人 申し上げたいのですが
 会議 乗り継ぎたいのですが
 八人に 申し上げたいのですか

<絞り込み結果>

会議に 申し込みたいのですが
 申し込みたいのですか

- (6) 登録用紙は 既に お持ちでしょうか。

<文節ラティス>

登録用紙は 既に お持ちでしょうか
 登録用紙が いつに 思いましようか
 登録した 千に お持ちましようか
 登録用紙とは 九でしょうか
 登録したら お持ちしては

<絞り込み結果>

登録用紙は いつに お持ちでしょうか
 既に

- (7) いいえ。
- (8) まだです。
- (9) 分かりました。
- (10) それでは 登録用紙を お送り致します。

<文節ラテイス>

それでは	登録用紙も	お送り致します
三では	登録用紙を	お送りします
すれば	登録用紙と	お聞き致します
会議	登録用紙とも	お持ち致します
八人に	登録費としても	お送りしています

<絞り込み結果>

それでは	登録用紙も	お送り致します
	登録用紙を	

- (11) ご住所と お名前を お願いします。

<文節ラテイス>

ご住所と	お名前を	お願いします
ご住所とも	お名前も	お願いしますか
	お名前と	お願いしない
	お名前とも	お願いしましょう
	お名前は	お願いしたい

<絞り込み結果>

ご住所と	お名前を	お願いします
	お名前も	

- (12) 住所は 大阪市 北区 茶屋町 二十三です。

<文節ラテイス>

住所は	希望させる	北区	茶屋町	二十三です
十七は	大阪市	北区を	茶屋町に	二十三ですか
十四は	大阪市に	北区と	茶屋町と	二十三ですね
十四が	希望される	北区で	茶屋町の	二十三でしょう
十七が	送らせる	北区が	茶屋町が	

<絞り込み結果>

住所は	大阪市	北区	茶屋町	二十三です
				二十三ですか

(13) 名前は 鈴木真弓です。

<文節ラテイス>

名前は 鈴木真弓です
 名前が 鈴木真弓ですか
 七は 鈴木真弓ですね
 なれば 鈴木真弓でしょう
 七が 鈴木真弓として

<絞り込み結果>

名前は 鈴木真弓です
 鈴木真弓ですか

(14) 分かりました。

(15) 登録用紙は 至急 送らせていただきます。

<文節ラテイス>

登録用紙が 至急 送らせていただきます
 登録用紙は 七十 送らせて致します
 登録用紙から 七十を 含ませていただきます
 登録しました 七十五 含ませて致します
 登録用紙とは 七十も 送らせられます

<絞り込み結果>

登録用紙が 至急 送らせていただきます
 登録用紙は

(16) 分からない 点が ございましたら いつでも お聞き下さい。

<文節ラテイス>

分からない 点が ございましたら いつでも お聞き下さい
 送らない 意味が ございましたら 一でも お送り下さい
 分かりたい 入れるが 言われますか いつも お聞きしてある
 分からない 四が 行かれますか いつにも
 在りたい 言えば 五千ですか いつの

<絞り込み結果>

分からない 点が ございましたら いつでも お聞き下さい
 送らない

(17) 有難うございます。

(18) それでは 失礼します。

<文節ラテイス>

すれば	失礼します
三では	失礼しますか
いつでは	失礼しない
それでは	失礼したい
すれが	されます

<絞り込み結果>

すれば	されます
それでは	失礼します

(19) どうも 失礼致します。

<文節ラテイス>

五の	失礼致します
五も	して致します
どうも	失礼します
五を	失礼しています
五のも	失礼して致します

<絞り込み結果>

どうも	失礼致します
	失礼します

(20) もしもし。

(21) こちらは 会議事務局です。

<文節ラテイス>

こちらは	会議事務局です
九からは	会議事務局ですか
九から	会議事務局ですね
六が	会議事務局でしょう
送れば	

<絞り込み結果>

こちらは	会議事務局です
	会議事務局ですか

(22) 会議に 参加したいのですが。

<文節ラテイス>

会議に 参加したいのですが
 会議 参加したいのですか
 八に 参加されますか
 八人
 八

<絞り込み結果>

会議に 参加したいのですが
 参加したいのですか

(23) どうすれば よろしいですか。

<文節ラテイス>

どうすれば 用紙ですか
 五千が よろしいですか
 五千の 用意しますか
 五千も 用意しました
 五千を 用紙です

<絞り込み結果>

どうすれば 用紙ですか
 よろしいですか

(24) 先ず 登録用紙で 手続を していただかなくては なりません。

<文節ラテイス>

先ず 登録用紙で 手続を していただかなくては なりません
 まられる 登録費として 手続も 在りませんが
 七を 登録用紙が 手続は 七千が
 七で 登録として 手続が なりません
 登録用紙に 手続から 名前が

<絞り込み結果>

先ず 登録用紙で 手続を していただかなくては なりません
 なられる

(25) もう 登録用紙は お持ちでしょうか。

<文節ラテイス>

もう	登録用紙は	お持ちでしょうか
五を	登録用紙が	お持ちしては
五も	登録した	お持ちしたら
どうも	登録したら	思いましようか
二を	登録用紙とは	お持ちした

<絞り込み結果>

もう	登録用紙は	お持ちでしょうか
		お持ちしては

(26) まだです。

(27) 用紙を 送ってください。

<文節ラテイス>

用紙を	送って下さい
用紙と	送って下さる
四十を	お聞き下さい
	お送り下さい
	送っていただく

<絞り込み結果>

用紙を	送って下さい
用紙と	

(28) では ご住所と お名前を お願いします。

<文節ラテイス>

では	ご住所と	お名前を	お願いします
五は	五十七と	お名前も	お願いしますか
二は	五十七は	お名前と	お願いしない
七		お名前とも	お願いしたい
七を		お名前は	お願いしましょう

<絞り込み結果>

では	ご住所と	お名前を	お願いします
五は	五十七と		

(29) 住所は 大阪市 北区 茶屋町 一の 二です。

<文節ラテイス>

住所は	希望させる	東区	徳井町	一の	二です
十七は	希望される	東区を	徳井町に	一も	二ですか
十四とは		東区も	送ります	一を	二ですな
十四は		東区で	送りたい	いつの	二十
十七が		東区が		一にも	二でしょう

<絞り込み結果>

絞り込み不可能

(30) 名前は 清水太郎です。

<文節ラテイス>

名前は	清水太郎です
名前が	清水太郎ですか
七は	清水太郎ですね
七には	清水太郎でしょう
名前とは	

<絞り込み結果>

名前は	清水太郎です
	清水太郎ですか

(31) 分かりました。

(32) 参加料は 要るのでしょうか。

<文節ラテイス>

参加料は	要るのでしょうか
参加料が	入れるのでしょうか
参加のは	四でしょうか
参加とは	言うのでしょうか
参加料を	入れましょうか

<絞り込み結果>

参加料は	要るのでしょうか
参加料が	

(33) はい。

(34) 登録費として お一人 三万 五千円が 必要です。

<文節ラティス>

登録費として	お一人	三万	五千円が	必要です
登録として	お二人	三万を	五千が	必要ですか
登録しました	お一人に	三万に	五千円は	七十六です
登録したい	お二人に	三万も	五十四が	四十六です
お持ちして	送りたい	三万の	五千には	七十六ですか

<絞り込み結果>

登録費として	お一人	三万	五千円が	必要です
				四十六です

(35) そうですか。

(36) どうも 有難うございました。

<文節ラティス>

五の	有難うございました
五も	有難うございます
どうも	
五のも	
五人を	

<絞り込み結果>

どうも	有難うございました
	有難うございます

(37) 失礼致します。