大語彙日本語連続音声認識研究基盤の整備 —学習・評価テキストコーパスの作成—

伊藤克亘（電総研）伊藤彰則（山形大）宇津呂武仁（奈良先端大）河原達也（京都大）
小林哲則（早稲田大）清水徹（KDD）田本真詞（NTT）荒井和博（NTT）
峯松信明（豊橋技科大）山本幹雄（筑波大）竹沢寿幸（ATR）武田一哉（名大）
松岡達雄（NTT）鹿野清宏（奈良先端大）

大語彙連続音声認識研究の推進のためには、標準（ベースライン）となるコーパス（音声、テキスト）やソフトウェア（言語モデル・音響モデル・認識プログラム）が必要であり、著者らはその基盤整備を進めている。本稿では、システム評価のためのテキストコーパスについて設計方法とその諸元を述べる。

Common Platform of Japanese Large Vocabulary Continuous Speech Recognition Research — Developement of text corpus —

Katunobu Itou(ETL), Akinori Ito(Yamagata Univ.),
Takehito Utsuro(NAIST), Tatsuya Kawahara(Kyoto Univ.),
Tetsunori Kobayashi(Waseda Univ.), Toru Shimizu(KDD),
Masafumi Tamoto, Kazuhiro Arai(NTT), Nobuaki Minematsu(TUT),
Mikio Yamamoto(Univ. of Tsukuba), Toshiyuki Takezawa(ATR),
Kazuya Takeda(Nagoya Univ.), Tatsuo Matsuoka(NTT),
Kiyohiro Shikano(NAIST)

For Japanese large vocabulary continuous speech recognition (LVCSR) research, we are developing standard baseline software repsitory that includes language models, acoustic models and recognition engines. In this report, design and specification of the text corpus are described.

1 はじめに

数万単語を越える語彙を対象とする大語彙連続音声認識は、音声認識技術の応用分野の拡大には極めて重要な課題である。わが国では、要素技術は高水準にあるが、その評価が孤立単語認識による音声制御装置や電話音声応答装置など応用システムに依存して行なわれており、要素技術間の相互比較が極めて困難な状況にある。

このような背景の中で、1995年11月に、情報処理学会音声言語情報処理研究会に「大語彙連続音声認識研究用データベースに関するWG」が発足した。このWGでは、大語彙連続音声認識に含まれる様々な要素技術の性能を迅速かつ厳密に評価しうる基盤を整備することを目的としている。評価基盤を整備することにより、以下の3点の効果が期待される。

1. 要素技術間の相互比較が容易になり、先端的な研究領域での技術進歩を加速する。
2. 要素技術の共通化により、音声認識技術を応用した製品等の開発の効率化を可能にする。
3. 音声認識応用システムの問題点を要素技術に還元することが容易になる。

評価基盤として整備すべきものには、共通の音声コーパス（音韻モデル学習用・認識システム評価用）とその読み上げ用テキスト、共通のテキストコーパス（言語モデル学習用）、標準的な認識用言語モデル、標準的な音韻モデル、標準的なツール（認識エンジン・形態素解析ツール・言語モデル作成ツール）などがあげられる。本稿では、このうち、読み上げ用テキストと共通のテキストコーパスの整備・構築について述べる。

2 テキストコーパスの構築

2.1 対象文書データの選定

大語彙連続音声認識研究用のテキストデータの構築には、評価用音声データとしての読み上げ文を抽出するデータと、その読み上げ文の抽出に必要な言語モデルを学習するデータが必要である。つまり、大規模なデータのうち、一部から評価用の読み上げ文を抽出するが、他の部分でその抽出に必要な言語モデルを学習するのである。

そのような対象データとして、文体や内容がある程度統一され、かつ多量に電子化されているため、新聞記事を選定した。新聞記事は書き言葉であるなどの問題点があり、音声認識の対象として必ずしも適していない面もある。しかし、上記の条件を満たす現実的な選択として米国の DARPA[1] や欧州の SQALE[2] でも対象とされており、また、日本でも同様の選択がなされていた [3]。新聞記事データのうち、標準的に利用可能な研究基盤とする点を考慮して、利用許諾権の観点から最も適切であった毎日新聞の記事を選んだ。

また、詳しくは後述するが、音声認識の要素技術の多様な評価には、テキストコーパスは形態素解析されていることが必須である。だが、WG 活動当初 (95 年末) では、標準的に利用できる高精度の日本語形態素解析システムが存在しなかった。そのため、毎日新聞 CD-ROM 1991 年版から 1994 年版までの形態素解析情報を収録したテキストコーパスとして、RWC テキストデータベース (RWC-DB-TEXT-95-1)[4] を利用することにした。

実際の統計モデルの適用を考えると、過去のデータを利用して新たなデータの認識を行なうという形態が自然である。よって、RWC テキストデータベースのうち、頻度リストや言語モデル学習用には、91 年 1 月から 94 年 9 月までの 45ヶ月分の東京版記事を利用し、評価用には 94 年 10 月から 94 年 12 月までの 3 ヶ月分の東京版記事を利用した。

2.2 読み上げ文コーパスの構築

読み上げ文コーパス構築の手順は、次のようになる。(最初の 4 段階は言語モデルの学習と共通の手順である)

1. テキストの形態素解析
2. 読み上げに適さない文や表現の削除
3. 形態素の頻度リストの作成
4. 文の複雑さを計算するための言語モデルの作成
5. 読み上げ文の仮セットの構築、検査
6. 読み上げ文の本セットの構築、検査

以上の手順について、以下で詳細に説明していく。

2.2.1 形態素解析

形態素解析が必要であるのは、音声認識の要素技術の多様な評価が可能なように、評価用読み上げ文の抽出には、i) 語彙の規模 ii) 文長 iii) 文の複雑さ、の 3 つのパラメータを考慮しなければならないためである。

但し、標準的なコーパスを作成するという視点で形態素解析を考えると、実際には次の二つの問題がある。i) 現状で、広く共通に利用できる形態素解析結果は、95% 程度の精度であり、解析誤りが相当数ある。ii) 異なる形態素解析システムを適用した場合には、形態素の頻度リストの内容も変わる可能性がある。

この i) ii) の問題点が、文選別のパラメータに影響を与える可能性がある。これに対して、i) の解析誤りについては、最終的には人間が検査して、読み上げ文からは排除する方針を採用した。ii) については、形態素解析システムの違いが統計的言語モデルに与える影響に関する知見はほとんどなく、複数の形態素解析システムを用いて作業することは、作業量の観点で不可能である。よって、今回は RWC テキストデータベースの単独の形態素解析システムの結果を、そのまま利用することにした。

2.2.2 読み上げには適さない文や表現の排除

新聞記事データに現われる読み上げに適さない表現には、(1) 記事や段落自体が文章でないもの、(2) 文章中に部分的に出現する読み上げには適さない表現、がある。

記事や段落自体が文章でないものとしては、人事情報、株式市況などの表、俳句／川柳／和歌の欄、料理欄の材料一覧、スポーツの結果、アンケート結果などがある。これらの段落の多くに共通な特徴として、段落全体にひとつも句点「。」が含まれないことがあげられる。そこで、これらの段落を機械的に判別する手法として、句点のない段落を排除する方法で対処した。

文章中に部分的に出現する読み上げに適さない表現は、単語以外の記号や空白記号を使った字下げなどの用法と密接に関わる。これらの記号が構成する表現には、その表現がないと文が成立しないものと、その表現がなくても文が成立する表現がある。後者は、「拮（きっ）抗」の「(きっ)」のように、自然な読み上げを妨げる表現がある。この場合、読み上げを対象とするデータからは「(きっ)」の部分を削除することが望ましい。

括弧表現に関しては、「引用句」「強調表現」「リストのラベルの代名詞表現」「リストのラベル」は削

除せず、「段落などの見出し」「補充要素」は削除することにした。これらの表現がどの括弧に対応しているかは、新聞によって変化すると考えられるが、該当するデータにおける用法を調査した結果と、自動判別手法 [5] を元に、削除を行なった [6]。

また、括弧による「段落などの見出し」に準ずる表現として、注視記号 (○●★など) と空白などの字下げ表現を併用した表現がある。これらの表現についても、削除した。

これらの処理の効果を示すため、元のデータの例と処理後 (削除後) のデータの例を図 1、図 2 に示す。

```
|　|◇|高齢|化|社会|を|よく|する|女性|の|会
|講演|会|　|１２|日|（|火|）|午後|１|時半
|―|４|時|、|東京|都|千代田|区|神田|駿河|台
|３|の|９|、|三井|海上|火災保険|本社|ビル|１
|階|大|会議室|（|ＪＲ|御茶|ノ|水|駅|、
|地下鉄|新|御茶|ノ|水|、|淡路|町|、|小川|町
|駅|下車|）|。|スウェーデン|研究|で|同国
|から|北極|星|勲章|を|受け|た|岡沢|憲
|芙早大教授|が|「|ほんとう|の|『|生活|大国
|』|スウェーデン|事情|」|を|テーマ|に|講演
|する|。|問い合わせ|は|同|会|（|０|３|・
|３３５６|・|３５６４|）|へ|。|
```

図 1 元のデータ

```
|１２|日|午後|１|時半|―|４|時|、|東京|都
|千代田|区|神田|駿河|台|３|の|９|、|三井
|海上|火災保険|本社|ビル|１|階|大|会議室
|。|スウェーデン|研究|で|同国|から|北極|星
|勲章|を|受け|た|岡沢|憲|芙早大教授|が|「
|ほんとう|の|『|生活|大国|』|スウェーデン
|事情|」|を|テーマ|に|講演|する|。
|問い合わせ|は|同|会|へ|。|
```

図 2 削除後のデータ

この処理を行なった後のデータ量を 表 1 に示す。

表 1 コーパスのデータ量

	91/1～94/9	94/10～94/12
文数	2,371,932	194,372
段落数	1,438,311	138,590
記事数	281,818	21,186
形態素数	65,347,098	4,935,600
形態素 (種類)	290,939	97,409

2.2.3 形態素の頻度リストの作成

RWC の形態素解析結果としては、形態素の表記・原形・品詞名が付与されている。日本語の形態素の計量基準としては、以下の方法が考えられる。

1. 表層表現が同じなら同一と見なす
2. 原形 (標準形) が同じなら同一と見なす
3. 読みが同じなら同一と見なす

言語モデルの研究・評価を行なうためには、より詳細なレベルで単語が区別されていることが望ましい。しかし、現状では、テキストコーパスに読みを正確に自動付与することは困難であるため、表層・原形・品詞のいずれかが異なるものは、異なる単語として扱うこととした。この計量基準で、頻度リストを構築した。被覆率を表 2 に示す。

表 2 被覆率

語彙規模	被覆率 (%)
5000 語 (中語彙)	85.8
8129 語	90.0
20047 語 (大語彙)	95.7
27634 語	97.0

読み上げ文の選択用基準として語彙の規模を考慮することになっている。DARPA や SQALE では、5000 語と 20000 語という 2 段階の基準を設けている。しかし、日本語の場合、英語などの語という単位と形態素という単位には違いがあるという問題がある。そこで、英語の 5000 語、20000 語の被覆率がそれぞれ、約 90%、97% である点に着目し、日本語の場合には被覆率を 90%、97% となる語数で語彙を設定する基準が提案されている [3]。しかし、日本語の場合、形態素解析システムによって単位が異なるという問題もある。この場合に被覆率がどうなるか、などの知見はまだ得られていない段階であるため、他言語との比較を厳密に議論することは困難である。また、システム開発の困難さは、被覆率よりも登録語数に関わってくる。これらの観点から、本 WG では、5000 語、20000 語の 2 段階の語彙の規模を設定することにした。(実際には頻度が同数の語が多数あったため、20000 語は 20047 語で設定した。)

2.2.4 言語モデルの作成

読み上げ文を選択する単位にも、記事にするか、段落にするか、文にするかという選択があり、DARPA などでは段落単位としている。評価用データの母集団が十分にない点から、パラメータを制御する分に

ついては、文単位で選択することにした。ただし、段落単位、記事単位の文も補助的なデータとして用意することとする。

文単位に選択するためには、文という単位を定義する必要がある。一般には、句点から句点が文であると考えられているが、新聞では、「。」以外にも句点と同じ働きをする記号がある。また書き言葉では、引用句が多用されるため、文の定義は引用句の範囲の認定と密接な関わりがある。

毎日新聞では、「。」の他に、「?」「!」「。」の後に空白があるものが句点相当の表現として使われている。また記事の種類によって、「▼」や「▲」が句点および読点相当の表現として使われることがある。これらを総称して、句点相当表現と呼ぶことにする。引用句は「」を使って表現される(『』は「」内で「」と同じ働きをする)。

ここでは、引用句を「」で囲まれた範囲として定義し、地の文の句点相当表現から句点相当表現の間を文と定義した。この定義に基づき、前処理をおこなった記事データ(段落単位になっている)を文に分割する。そのように前処理したデータから、大語彙に含まれる20047語以外は未知語としたバックオフバイグラム言語モデルを作成した。言語モデルは、CMUツールキット[7]を利用し、バイグラムのカットオフは2として作成した。バイグラムの種類は2,402,695であった。

2.2.5 読み上げ文の抽出

抽出のための3つのパラメータi)語彙の規模ii)文長iii)文の複雑さ、について以下で順に説明する。

語彙の規模

語彙の規模は前述の2段階を基本にさらに再分類して5段階とする。5000語を中語彙、20000語を大語彙と呼ぶことにする。中語彙に含まれる形態素のみからなる文によるクラスMID、中語彙に含まれる形態素とそれ以外の1語から構成される文をMID+、大語彙に含まれる形態素のみからなるクラスLARGE、大語彙に含まれる形態素とそれ以外の1語から構成される文をLARGE+、大語彙に含まれる形態素とそれ以外の2語から構成される文をLARGE++とした。語彙に含まれない語を含むクラスは、未知語(Out of Vocaburary word)に関する評価を可能にするために用意している。

中語彙クラスの場合には、形態素解析の誤りの影響は余りない。しかし、大語彙クラスになると、そもそも、20000語の語彙にすら解析を誤っている形態素が含まれるような状況であるため、形態素解析の誤りを含む文が多い。今回の作業手順では、形態素解析の誤りは手作業で除くことにしているため、歩止りを悪くしないため、LARGE++に該当するクラスは、DARPAなどでは無制限としているところを、大語彙を外れる語は2語に制限した。また、たとえばMID+とLARGEの両方の条件を満たす場合もありえるが、そういう場合には、語彙クラスの小さいクラス(この場合は、MID+)にのみ含めるようにする。

文長

文長に関しては、2段階を用意することにした。読み上げのときの読みやすさを考慮して、句読点などの記号も含めて39形態素以下となるようにした。また、短すぎる文には、形態素解析誤りや前処理で排除しきれない表現が多いため、6形態素以上の文を対象とした。

文の複雑さ

文の複雑さについては、3段階区別することにした。文の複雑さとして、上記の言語モデルを利用し、文ごとにパープレキシティを計算した。

計算には、文頭と文末を示すシンボルを付与し、句点や引用句を示す括弧などの記号も全く取り除かずに計算している。また、句点と文末の間の遷移確率は取り除いていない。未知語については、CMUツールキットのデフォルトの計算方法に準じている。すなわち、大語彙に含まれない語が全て一つの語であるとして、出現確率や遷移確率を計算している。したがって、未知語を含むほどテストセットパープレキシティが高くなる傾向が生じている(大語彙に含まれない語は、全体の4%強を占めるため、まとめて扱うと非常に出現頻度の大きな語になってしまうことが原因であろう)。

文のテストセットパープレキシティの大小によって、小さいものから順に、L, M, Hのクラスを設けた。パープレキシティの大きな文は、形態素解析誤りしている文が多いと考えられるため、399以下の文だけを対象とした。

仮セットのパラメータの決定

この基準で分類すると、全部で30クラスに区別される。この30クラスについて、話者ごとに90文発話することとして、文の内訳を以下のように設定することにした。

実際の語彙クラスごとと全体の文長やパープレキシティの分布を表4に示す。

この内訳で文を選択することで、コーパス全体が新聞記事全体の統計的な性質から若干離れたものになる可能性がある。しかし、前述したように、本WGでは「新聞記事の音声認識」よりも「大語彙連続音

表 3 セット当りの文の内訳

	NORMAL			LONG		
語彙クラス	L	M	H	L	M	H
MID	2	6	2	1	3	1
MID+	2	6	2	1	3	1
LARGE	4	12	4	2	6	2
LARGE+	2	6	2	1	3	1
LARGE++	2	6	2	1	3	1

表 4 文長とパープレキシティの分布

語彙クラス	文長		パープレキシティ	
	平均	分散	平均	分散
MID	12.3	8.4	75.9	81.8
MID+	15.2	9.3	83.2	82.0
LARGE	26.4	14.0	119.3	99.1
LARGE+	26.8	14.6	106.0	82.9
LARGE++	28.6	15.7	96.5	74.4
全体	26.2	16.5	101.0	84.5

声認識」を主目的している。したがって、当面の研究レベルに合わせて、中語彙のパープレキシティが小さめの文を比較的多く含むように文を選択している。

この文データは、単独の文から構成されるが、単独の文より広い範囲での言語現象にも対応できるように、段落を単位とするデータも用意することにする。(この段落のデータと文データが重複する場合はある) この段落データを話者一人当り 3 段落分割り当てることにした。したがって、話者一人当り合計約 100 文程度のセットが作成されることになる。

また、さらに広い範囲での言語現象にも対応できるように、上記のセットとは別に、35 文以上からなる記事を単位とするデータも用意した。このセットでも、1 名当り 100 文程度になるように、1 名当り 3 記事を割り当てることにし、全部で 5 セット構築することを目標とした。

実際に文を分類するためには 表 3 に示した割合で分類できるように、文長と複雑さの値を設定する必要がある。ここで留意しないといけないのが、語彙クラスによって文長と複雑さの分布がかなり異なることである。具体的には、小さい語彙クラスでは、パープレキシティは小さく文長は短くなる。逆に大きい語彙クラスでは、パープレキシティは大きく、文長は長くなる。したがって、同じ境界値で分類することは困難なので、中語彙と大語彙で違う境界値を設定することにした。設定された境界値を以下に示す。

段落データに関しては、表 3 のいずれかのクラスの文だけからなる 4 以上 14 以下の文からなる段落のみを対象とすることにした。このパラメータに基いて、94 年 10 月から 12 月のデータを分類した。194372 文中、131507 文が、いずれかのクラスに分類された。それぞれのクラスに分類された文の数を表 7 に示す。人間による検査によって捨てられる文の数も考慮にいれ、各クラス 2 倍のマージンをとって一セット当り 270 文 + 9 段落で、122 セット作成した。

表 5 文長に関する境界値

語彙クラス	NORMAL	LONG
MID 群	6 ～ 19	20 ～ 39
LARGE 群	6 ～ 29	30 ～ 39

表 6 複雑さに関する境界値

語彙クラス	範囲 (単位:パープレキシティ)
MID 群	$0 \leq L < 40 \leq M < 85 \leq H < 400$
LARGE 群	$0 \leq L < 70 \leq M < 130 \leq H < 400$

表 7 クラス別文数

	NORMAL		
語彙クラス	L	M	H
MID	2504	3726	2190
MID+	4142	7476	5673
LARGE	7790	10481	7815
LARGE+	11639	11338	7030
LARGE++	8583	6354	3400
	LONG		
語彙クラス	L	M	H
MID	413	1239	458
MID+	927	3956	1876
LARGE	2400	3607	1540
LARGE+	2892	4075	1725
LARGE++	2087	2896	1275

2.2.6 読みの付与・文の検査

このように作成された 122 セットに、形態素と読みの文脈独立的な対応を取った読み付与辞書ファイルを利用して機械的に読みを付与して、仮セットとした。仮セットの読みを人間によって検査し、誤っているものは正しい読みに修正した。

読みに関しては、括弧や句読点などの記号は読まないこととした。ただし、記号類でも、「%」(パーセント) のように読んだ方が自然な記号については読むこととした。

また、この際、形態素解析が明らかに誤っている表現を含む文・段落や、前処理で排除しきれない表現や、内容的に読み上げ作業者に精神的な苦痛を与えるよう

な表現を含む文などは排除した。これを各セットあたり 4 人の検査者で検査し、最終的に、90 文+3(4) 段落の読み上げ文セットを 150 セット (13500 文+451 段落) と、読み上げセット補充用のコーパス 11120 文 + 160 段落を構築した。

2.3 読み上げ文セットの諸元

構築された読み上げ文セットの文当りの平均音韻数は、以下の通りである。

表 8 文当りの音韻数

	NORMAL		LONG	
分類	平均	分散	平均	分散
MID 群	40.4	14.8	88.3	21.8
LARGE 群	63.0	24.5	113.1	22.2

セット全体に出現した形態素は、読みまで考慮して異なるものを数えると 20668 種類であった。中語彙のうち出現したものは、4940 種類であり、大語彙では 15256 種類であった。中語彙のうち欠けているものは、解析誤りが 1/3 程度で、後は記号や固有名詞をはじめとする名詞であった。各セットごとに出現する形態素の種類は、平均 903 で、セット当りの平均形態素数は 2114 である。

3 音声コーパス

以上の手順で構築した読み上げ文セットをもとに、日本音響学会 音声データベース調査委員会 (主査:板橋秀一) において、読み上げ音声を収録し、CD-ROM が作成された [8][1]。実際の収録時には、再度、上記の検査を行なった。また、実際に読み上げられた文の読みに関しては、予め付与していた読みと違う読みで発声された場合には、実際の読みに修正した。

4 今後の活動

2.2 でも述べたように、言語モデルの学習のためには、形態素解析や不要部分を削除するなどのテキスト処理が必要となる。本 WG で整備したテキストコーパスでは、これに対応する形態素解析システムは提供されていない。

そこで、より柔軟な研究基盤とするためには、これらの処理のための標準ツールが必要である。そこで、1997 年より情報処理振興事業協会 (IPA) 「独創的情報技術育成事業」の支援を得て、著者ら (代表者:鹿野) を中心にその整備を進めている。具体的には、共通的な形態素解析ツールとして利用可能な ChaSen[9][2]。を基盤に、本 WG で作成したコーパスや前述の音声コーパスを利用するための周辺的なツール (不要部分削除用ツールなど)、自動読み付与プログラムなどの整備を計画している。

謝辞

本研究は、情報処理学会音声言語情報処理研究会の「大語彙連続音声認識研究のためのデータベース整備 WG」での活動の成果である。また、本研究は、情報処理振興事業協会「独創的情報技術育成事業」の一環として行なわれている。関係各位のご支援に感謝いたします。読み上げ文セットの構築には、日本音響学会音声データベース調査委員会の方々はじめ多くの方々に協力して頂きました。心から感謝いたします。

参考文献

[1] D. B. Paul and J. M. Baker. The design for the wall street journal-based csr corpus. In *Proc. DARPA Speech & Natural Language Workshop*, pp. 357–361, 1992.

[2] H. J. M. Steeneken and V. Leeuwen. Multilingual assessment of speaker independent large vocabulary speech recognition system: the SQALE-project. In *EUROSPEECH*, pp. 1271–1274, 1995.

[3] 大附克年, 森岳至, 松岡達雄, 古井貞熙, 白井克彦. 新聞記事を用いた大語彙連続音声認識の検討. 信学会技術報告, Vol. SP95, No. 90, pp. 63–68, 1995.

[4] 井佐原均, 元吉文男, 徳永健伸, 橋本美奈子, 荻野紫穂, 豊浦潤, 岡隆一. RWC における品詞情報付きテキストデータベースの作成. 言語処理学会第 1 回年次大会, pp. 181–184, 1995.

[5] 荻野紫穂. リストのラベルとして使われる丸括弧とリストの範囲. 計量国語学, Vol. 19, No. 4, 1994.

[6] 伊藤克亘, 松岡達雄, 竹沢寿幸, 武田一哉, 鹿野清宏. 大語彙連続音声認識研究のためのテキストデータ処理. 日本音響学会講演論文集, pp. 105–106, September 1996.

[7] Ronald Rosenfeld. The CMU Statistical Language Modeling Toolkit and its use in the 1994 ARPA CSR Evaluation. In *Proc. ARPA Spoken Language Systems Technology Workshop*, pp. 47–50, January 1995.

[8] 板橋秀一, 山本幹雄, 竹沢寿幸, 小林哲則. 日本音響学会新聞記事読み上げ音声コーパスの構築. 日本音響学会講演論文集, pp. 187–188, September 1997.

[9] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明. 日本語形態素解析システム『茶筌』 version 1.0 使用説 明書. Information Science Technincal Report NAIST-IS-TR97007, NAIST, 1997.

[1] 詳細なコーパスの概要や申し込み方法などは、以下の URL を参照のこと。http://www.milab.is.tsukuba.ac.jp/jnas/

[2] 詳細は、以下の URL を参照のこと。
http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html