社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2005－NL－167　(8)
2005－SLP－ 56　(8)
2005／5／26

音声翻訳のための中国語対話コーパスの整備とその評価

胡　新輝†，劉 敏†‡，山本　博史†，菊井　玄一郎†

†ATR 音声言語コミュニケーション研究所
〒619-0288 京都府相楽郡精華町光台２－２－２
‡現在，北京 NEC 集成電路設計有限公司
中国，北京市海淀区知春路２７号大運村北京集成電路設計園７F
Email† {xinhui. hu, hirofumi. yamamoto, genichiro. kikui}@atr. jp ‡ liumin@bjnec.com.cn

あらまし　現在，自由対話の音声翻訳システムにおける音声認識及び翻訳では，統計的言語モデルが広く使われている．統計言語モデルでは，信頼できる統計量を得るために，大規模で高品質な学習用コーパスが必要である．
本報告では，日中音声対話翻訳システムのための，中国語対話文のコーパスの整備（セグメンテーション及び品詞付与）の方法及びその特徴を述べる．更に，そのコーパスを用いて構築した言語モデルのパープレキシティ及び連続音声認識による性能評価結果を報告する．

Development and Evaluation of Chinese Conversational Corpus for a Speech-to-Speech Translation System

Xinhui Hu†,　Min Liu†‡, Hirofumi Yamamoto†, Genichiro Kikui†

†ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
‡ Currently,　NEC IC Design Beijing Co., LTD.
Floor 7, Quantum Plaza, Zhichunlu 27#, Haidian District, Beijing, 100084, P.R.C.
Email† {xinhui. hu, hirofumi. yamamoto, genichiro. kikui}@atr. jp ‡ liumin@bjnec.com.cn

Abstract： In speech-to-speech translation systems, statistical language models are widely used in both speech recognition and translation. In statistical language models, a large amount of training data is required to calculate reliable statistics. Performance of language model depends heavily on its quantity and quality. Therefore, development of a training corpus is one of the most important issues for speech-to-speech translation systems.
In this paper, we report our conversational Chinese morphological corpora (with segmentation and part-of-speech tags), which are mainly used in our Japanese-Chinese speech translation system. Besides describing their statistical characteristics, we will also report the evaluation results of language models that are trained on these corpora.

1. まえがき

近年，音声言語処理において，統計的な手法が有効であることが確認されており，また，広く用いられている．しかしながら，信頼できる統計量を得るためには大規模な言語コーパスが必要とされる．我々の音声翻訳システムでは音声認識，翻訳双方で統計ベースの手法を採用している．そのため，日英音声翻訳システムにおいては，日本語及び英語の対話コーパスの構築は非常に大きな位置を占めており，これは日中音声翻訳システムにおいても同様と考えられる．このコーパス整備のためには，日中対訳テキストの整備と，それに対する形態素解析，すなわち，セグメンテーション及び品詞付与が必要とされる．本稿では，この中国語テキストの特徴と，それに対する形態素解析方法について紹介すると共に，このコー

パスを用いて構築した言語モデルに対し，パープレキシティおよび，連続音声認識で評価した結果を報告する．

まず，第２章では対象である中国語旅行会話文の内容を簡単に紹介した後，中国語形態素の定義を，他の既存の中国語形態素定義と比較しながら紹介する．第３章では形態素解析済みコーパスの作成手順を説明する．第４章では、対訳元の日本語版コーパス，そして他の中国語コーパスそれぞれとの特徴の比較を行う．第５章では，作成されたコーパスを用いて構築した言語モデルの性能をパープレキシティ，音声認識の二つで評価した結果を述べる．第６章ではむすびと共に，現状における問題及び今後の計画について述べる．

2. コーパスと形態素解析定義

2.1 対象データ

頑健な音声翻訳システムを構築するためには，多彩な言語表現を含むコーパスが必要である．ＡＴＲではこの目的のために複数の対話コーパスを構築している．表１に，それら，ＳＬＤＢ，ＢＴＥＣ，ＭＡＤという三つのコーパスの概要[1]を示す．

表１．音声対話翻訳のためのデータベース

名称	収集方法	目的	発話文数	ドメイン
SLDB	２言語模擬対話（通訳）	音声翻訳	１６Ｋ	ホテル対話限定
MAD	２言語模擬対話（ＭＴ）	音声翻訳	１１Ｋ	旅行会話一般
BTEC	テキスト	翻訳	６８０Ｋ	旅行会話一般

ＳＬＤＢ（Spoken Language Database）は異言語話者に人間の通訳を介して対話させ，対話者及び通訳の発話を収録したものである．

ＢＴＥＣ（Basic Travel Expressions Corpus）は旅行の様々な場面で用いられると考えられる表現を書き出して対訳を付与したものである．

MAD（Machine Translation Aided Dialogue）はＡＴＲの音声翻訳システムを介して日英の話者に対話を行わせ，その発話を収集したものである．

上記のデータは，元々日英対訳コーパスとして構築されたが，現在，それらに中国語対訳を追加することによって，日中の対訳コーパスとしても用いることができるようになっている．今回，我々が対象とするのは，この中国語対訳である．

2.2 他の中国語コーパス

現在，利用可能な形態素解析済み中国語コーパスが幾つかある．代表的なものとして，北京大学計算語言研究所と富士通が１９９９年から２００２年まで，中国の最大新聞紙である「人民日報」の１９９８年を対象にしたコーパス（略 PKU）と，台湾中央研究院語言学研究所のバランスドコーパス（略 SINICA）が挙げられる[1]．前者で用いられている「人民日報」は，文章の構造及び文法，用語等に対して厳密なチェックが行われており，人民日報社はこの中の半年分，およそ７５０万語のデータを研究のために条件付きで公開している．後者は，初めての品詞タグ付きの中国語コーパスで，特徴としては，トピック，文体に対するカバレッジの大きさと，品詞の定義が詳細であることであり，コーパスの規模も５００万語に至っている．代表的なトピックは，政治，科学，社会，文学などである．また，文体としては，報道，評論，小説，ドラマの脚本，演説文，及び会話などが含まれている．しかしながら，会話文は全体(749,886 文)の１０％（75,017 文）でしかなく，ラジオ放送やテレビ番組のインタビューで収集した文章が多く採用されている．この二つのコーパスの統計的特徴は，後で我々のコーパスと比較する時に説明する．

2.3 形態素の定義

形態素解析の整備作業は，厳密かつ明確な仕様が不可欠である．この仕様は，作業者のガイドラインになり，コーパスの品質にとって大切である．ここで，我々が，既存の幾つか仕様書を参考にして，自由対話のテキストの特徴と音声認識及び機械翻訳の応用の目的を考慮しながら，中国語形態素解析の仕様書を作成した．以下，この仕様の作成について，説明する．

2.3.1 形態素の定義の基本方針

（１）既存の形態素定義との互換性

ＰＫＵコーパスにおける形態素定義は，１９９０年代前期発表された中国国家標準 GB13715“信息処理用現代漢語分詞規範”基本としている．ここでは，形態素を“単独で使える最小言語単位”と定義している．中国語単語，複合語，フレーズ（短語）の間に明確な境界が存在しないという理由から，“分詞単位”（つまり，セグメンテーション単位）という概念を，中国語の情報処理の応用に確定的な語義及び文法機能を持つ単位として定義するとされている．より具体的な定義は，《現代漢語語料庫加工規範―詞語切分与詞性標注，2001》[2]というセグメンテーションと品詞タグ付け仕様書に規定されており，これに従ったコーパスとなっている．さらに，これを拡張した規定が２００１年に中国教育部言語応用研究所から≪信息処理用現代漢語詞類及詞性表記集規範》[4]として発表されており，多数の言語学者の研究成果を吸収し，当時既存のコーパスに使われていた品詞セットを参考した上で，現代中国語の文章の品詞分類及び表記符号集を推薦し，各種の品詞セットを統一することを狙いとしている．この規定は，中国国家重大基本研究プロジェクト（９７３）で検討されたものであり，中国科学院，清華大学，

[1]中国本土と台湾では語彙，発音，文法には基本的に違いはない．しかしながら，現代用語や外来語の一部に違いが見られる．また，文字セットも異なる．

北京大学などで採用されている．

一方，SINICA では１９９０年代中期，規定された《中文資訊処理分詞規範》[3]に基づいており，その中では”信，雅，達“という三つの原則が提唱されている．

我々は自分の仕様書を作成する際に，上述の規範を元にし，音声対話の特徴を考慮しながら，分詞単位及び品詞セットを設計した．例えば，数字列を一文字毎に切り，一分詞とすることにより，データスパース問題を対応できるようにする．傾向動詞を普通動詞から分離して，単独の品詞とし，自由対話中に良く表れる傾向動詞の特徴をよりはっきりと表現できるようにする．同様に，会話特に旅行対話中に，頻繁に用いられる，問い合わせと希望を表す能願動詞(vw)，系動詞“是”（v1），および，所有関係を表す動詞“有”（v2）もそれぞれに一つ品詞として取り扱う．

（２） 将来の拡張性

ＡＴＲのコーパスは，現時点では音声認識及び機械翻訳を目的として構築されているが，今後，適用範囲を拡大する際に，単語の切り方や品詞セットの設定などが変更し易いことが望ましい．したがって，品詞の種類数が過大にならないように，４０種類程度に設定した．これによって，品詞の分類が簡単で，しかも将来拡張し易くなることが期待できる．

（３）他言語のコーパスとの関連性

我々が現在処理するこのコーパスの元言語は英語あるいは日本語であり，これらの言語の形態素解析作業が先行しており，音声認識と機械翻訳システムで高い性能が得られている．日本語及び英語における規定を参考にすることは有用である．例えば，日本語のデータと同じように，固有名詞の場合に，人名，組織名，飲食物名及び地名を示すサブカテゴリを設けた．更に人名では，中国，日本と欧米における姓名の順序の違いを考慮して「日姓」，「日名」，「中姓」，「中名」「欧米姓」，「欧米名」の六つのサブカテゴリを設けた．

2.3.2 形態素規定方法

形態素規定方法はセグメンテーション規定と品詞規定の二つの部分に分けて行う．

セグメンテーション規定では，文字列を如何に切断し，単語に分解する際に基準を定める．基本的に，２，３文字までを単語としてセグメンテーションを行う．４文字以上の単語は原則として，構文に従って分解する．ここでは，PKU コーパスの規定と異なる点のみを挙げる：

（１） 数字は，一文字毎に分ける．例えば，“一/ 百/ 二/ 十/ 八/，三/ 分之/ 一/，第/ 五/”．

（２） 頻繁に使われる挨拶等のフレーズは，一つの単位にする．例えば“您好/”，“再見/”，“好的/”．

（３） 複合詞の動詞，名詞に対して，構文構造（並列構造，偏正構造，動賓構造，主賓構造など）によって，結合するか，切断するかという基準を決める．

（４） 構造のみではなく意味も考慮する．ＰＫＵでは構造（修飾関係があるかどうか）のみで分割，非分割を決めており，例えば”想要”は常に二語に分割される．これに対し，われわれは構造の他に意味も考慮した分割を行っているため，”想要”はコンテキストによって分割されるか否かが変る。

例文：

想/ 要/ 一/ 个/ 苹果/ ./（リンゴが一つ**ほしい**.）

想要/ 住/ 你们/ 饭店/ ./（そっちのホテルに泊まり**たい**.）

品詞規定は，付属表１に示す品詞セットによって，セグメンテーション単位に対する品詞を確定する基準を決める．ここでは，セグメンテーションの規定と同じように，ＰＫＵと異なる点のみを挙げる．

（１）系動詞“是”を他の動詞と区別し，(v1)にする．

（２）“有”を単独の品詞（v2）とする．

（３）傾向動詞を動詞から分類し，別の品詞（vt）にする．

（４）可能，希望，願望などの意思を表す能願動詞，例えば，“能，想，要，应該”を，別の品詞(vw)にする．

（５）単純の数字(m)以外に，”数+量“のような構造の数量或いは程度を表す単語は，新しい品詞（ma）とする，例えば，”一些，一点“．また，”个把，左右，好些“のような概数を新しい品詞(mb)にする．

（６）“的，地，得”などのような構造助詞を新たな品詞(de)に入れる．

（７）中国人人名の姓を(nppx)に，名を（nppm）に，日本人の姓（nppxj），名を（nppmj）に，欧米人の姓を（nppxw），名を（nppmw）に定める

（８）料理名，飲食物名は(npfd)にする．

3. コーパスの作成手順

図１は，コーパス整備作業の全体の流れを示す．ここでは，この図について説明する．

A- 学習用コーパスのデータに基づいて，セグメンテーションと品詞タグ付けのためのモデル，すなわち，解析用言語モデルを訓練する．ただし，最初は訓練データが存在しないため，PKU と SINICA のデータから抽出したもので代用する．

B- 処理対象テキストを解析用言語モデルに通して，セグメンテーション及び品詞タグ付けのデータが得られる．

C- 上述のタグ付けのデータに対して，人手で形態素解析仕様書を基準にして修正作業を行う． 人手でチェックしたデータに対し評価を行い，性能が不十分であれば再チェックを行う．

D- 性能が十分であればコーパスを更新する．

E- チェック作業をする際に，固有名詞に関しては，日本語対訳データにおける位置及び品詞タイプの情報を利用して，作業効率を向上させる．

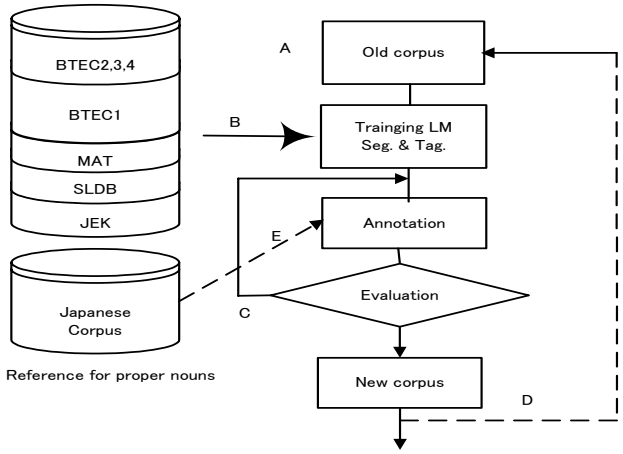

図１．コーパス整備作業の流れ

4. コーパスの統計的特徴

表２は，日本語のコーパス（ATR-J と記す），今回の対象である対訳中国語データ（ATR－C と記す）と，PKU 及び SINICA コーパスの統計的特徴を示している．図２は，ATR－C における単語の長さの分布を示す．

表２．各コーパスの統計データ

	発話数/文数	平均単語数	最大単語数	最大単語長	単語数	辞書数
ATR-C	207,982	7.2	116	13	1,496,374	23,724
ATR-J	220,199	8.83	57	27	1,943,873	25,255
PKU	290,193	25.1	920	30	7,284,877	163,700
PKU-short[1]	797,216	9.14	920	30	7,284,877	163,700
SINICA	749,886	7.76	105	44	5,819,922	153,400
SINICA-spoken	75,017	7.32	31	18	549,162	17,107

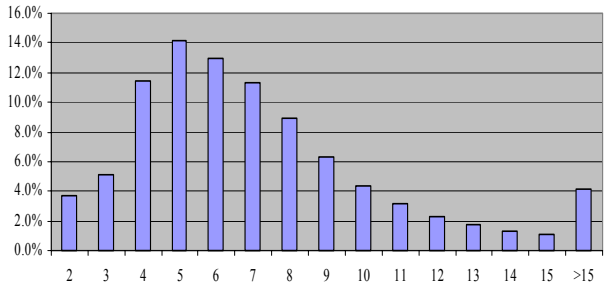

図２．単語の長さの分布（単位：文字）

図３に，各品詞の分布を示す．比較のために，SINICA と PKU の統計データを示してある[2]．図４，５に辞書（のべ語数）とコーパスの中の多品詞語と単一品詞語の比率を示す．

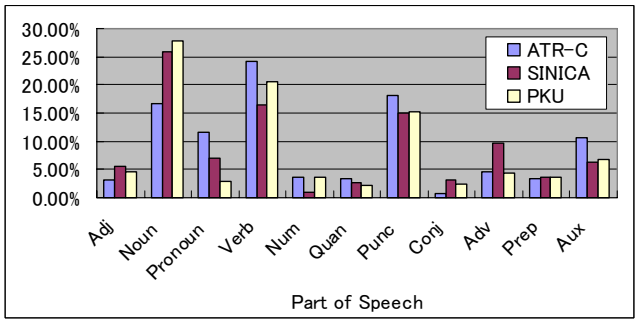

図３．コーパス中の主な品詞の分布

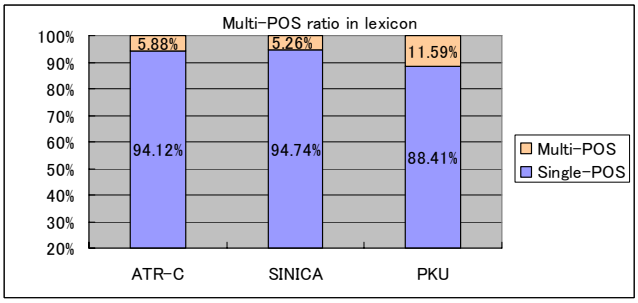

図４．辞書中の多品詞と単一品詞の比率

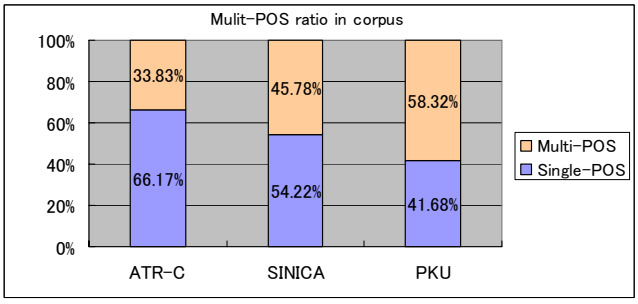

図５．コーパス中の多品 詞と単一品詞の比率

5. 言語モデルの評価

３．の手順に従って作成した四つのコーパス（SLB,MAD,BTEC1）を使って言語モデルを構築し，パープレキシティおよび音声認識によるモデル性能の評価実験を行った．評価対象のモデルは，山本らが提案したマルチクラス 2-gram 及びマルチクラス複合 2-gram というモデル[5]である．これらのモデルはすでに ATR の日英音声翻訳システムで用いられており，データスパースの問題に対して有効であることが確認されている．また，従来の単語 2-gram 及び単語 3-gram モデルとの比較も同時に行う．これらのモデルの評価実験の目的は，現在のコーパスの質がどの位か，また，日本語や英語などに比べてどのような特徴があるかを調べる事により，コーパスの整備と改善，及びモデルの選択などに役立たせることである．

[1] PKU_short は，コンマを文境界とする時のデータ
[2] 異なる品詞セットを使っているので、完全の品詞マッピングが出来ない，ここで主な品詞のみを提示している。

5.1 言語モデル

5.1.1 マルチクラス N-gram モデル

このモデルはクラスNグラムを基本にして，直前直後の単語の接続性を考慮しながら各単語に対して複数のクラスを割り当てたモデルである．各単語の生起確率の計算は次式で行う．

$$P(w_i \mid w_{i-1}) = P(w_i \mid C^t_{Wi}) P(C^t_{Wi} \mid C^f_{Wi-1}) \qquad (1)$$

ここで，w_{i-1} は単語または単語系列である．$C^f_{wi_{-1}}$ は先行単語 w_{i-1} が属するクラスであり，C^t_{wi} は後続単語 w_i が属するクラスである．右辺第一項はクラスから単語または単語系列が出現する確率，第二項は先行単語のクラス $C^t_{w_{i-1}}$ から後続単語のクラス $C^t_{w_i}$ への遷移確率である．これらのクラスは，自動クラスタリング手法に基づいて学習データから自動的に求められる．

5.1.2 マルチクラス複合 N-gram モデル

これは，マルチクラス N-gram モデルを拡張したものであり，頻度の高い連続単語文字列を単語 N-gram，頻度の低い単語をそのままクラス N-gram にし，両者を結合して，単語の予測を行う方式である．

5.2 モデルのパープレキシティによる評価

モデルの評価は，テスト用テキスト文のパープレキシティにより行う．ここで，モデル訓練用データは，SLDB,MAD,BTEC1 の学習セットで，合計１７万文，１４９万語である．テスト用データは，BTEC1 から抽出した学習データに対しオープンな１０２４文である．

図６．は単語 2-gram と単語 3-gram を含む各モデルのパープレキシティを表している．図に示されるように，クラスの数の増加に伴って，マルチクラス 2-gram と複合 2-gram のパープレキシティは減少する．また，複合 2-gram はマルチクラス 2-gram よりパープレキシティが小さい．これは，日本語のモデルと同じであるが，マルチクラスモデルと単語 2-gram，複合モデルと単語 3-gram の交差点が存在していない事が日本語のモデルと異なっている．

単純にパープレキシティの大小でモデルを選定するのであれば，単語 3-gram，マルチクラス複合 2-gram，単語 2-gram，マルチクラス 2-gram の順となる．しかしながら，モデルの良し悪しは，他の様々な要因を考慮しなければならない．例えば，音声認識の単語の正解率，モデルのサイズなども重要な要素である．低いパープレキシティは必ずしも音声認識における認識率に直結するとは限らない．そのため，次節では，各種のモデルを使って，引き継ぎ連続単語認識による評価実験を行う[1]．

[1] 現時点の条件により、単語 3-gram モデルを除く．

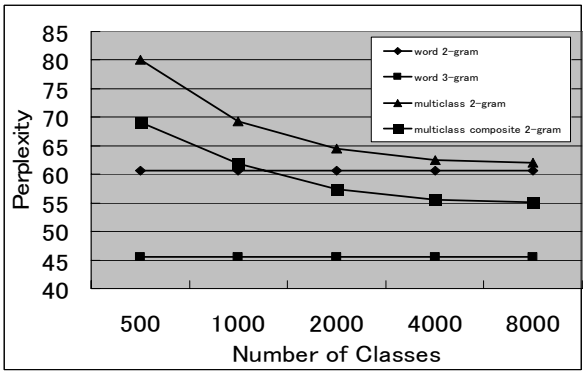

図６．各言語モデルの Perplexity

5.3 連続音声認識実験

テスト用音声データは，話者６４人（男女それぞれ３２人）で発話した，上述のモデル評価時に使用した同じ１０２４文のテキスト文である．

言語モデルは以下の種類で試みる－単語 2-gram，マルチクラス 2-gram（クラス数＝2000），マルチクラス複合 2-gram（クラス数=2000）

音響モデル－ATRPTHHKU という中国語音声データベース[6]によって訓練された，性別依存自動選択の HMNET モデルである．この音響モデルの構成は，表３に示す．

表３．音響モデルの学習条件

パラメータ＝MFCC(12)，ΔMFCC(12),ΔE(1)
フレーム周期 ＝10ms，フレーム長 20ms
状態数=1200，混合＝５

図７に，各種のモデルを使った音声認識の結果を示す．図に示されるように，マルチクラス複合 2-gram が，全体的に良い性能を示しており，これは日本語と英語と同じ傾向である．但し，マルチクラス 2-gram のパープレキシティは単語 2-gram より高いが，認識はマルチクラス 2-gram の方が高い．この理由はマルチクラス 2-gram は認識の難しい低頻度語に対しでロバストなモデルであるため，認識率では有利に働いたためと考えられる．

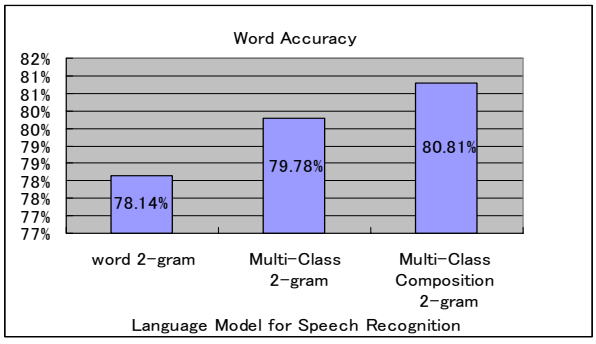

図７．連続音声認識結果

評価基準は，次の式で示す単語認識精度の計算を行った．

$$WordAccuracy = \frac{W - D - I - S}{W} \times 100 \quad (2)$$

(W: 正解単語数，D:削除誤り数，I : 挿入誤り数，S : 置換誤り数)

6. むすび

本稿では，ATR で行っている中国語対話コーパスの整備に関して，形態素定義とそれに基づくセグメンテーション及び品詞タグ付与の手順を説明した．形態素解析済みのコーパスを使って単語の長さ，品詞分布，単一／多品詞の比率などの統計量を分析することにより，中国語コーパスと，対訳元の日本語のコーパス，そして，他の種類の中国語コーパスとの比較を行い，対話文の統計的特徴を調べた．これらの情報を利用して，言語モデルの構築に役立たせると考えられる．また，このコーパスを用いた言語モデルのパープレキシティ及び音声認識による評価実験を行った．中国語の言語モデルは，日本語，英語のモデルと同様にマルチクラス複合 N-gram モデルが最も良い性能を示した．一方，残された問題として，日本語のモデルに比べて，中国語の言語モデルのパープレキシティが高く，音声認識の単語認識率もまた低い（13-14%位の差が存在している）ことがあげられる．原因は，音響モデルの訓練データの不足以外、言語モデルの学習データも，日本語に比べて，まだ不十分である（日本語の学習データは，BTEC3，４まで，合わせて 500 万語以上，辞書のサイズは 2.5 万語）と考えられている．また，翻訳時における誤りや揺れ（特に固有名詞の訳に問題が多い）によるコーパスの質の劣化も原因に一つにあげられる．今後，コーパス量を増加させるとともに，コーパスの揺れ等の問題を改善し，学習データの質と量を確保することによって，音声認識及び翻訳性能の向上を目指す．

謝辞

本研究は，総務省からの研究委託「携帯電話等を用いた多言語の自動翻訳システム」により実施したものである．

参考文献

[1] 菊井玄一郎，竹澤寿幸，山本誠一，“対話翻訳のための音声言語コーパスの現状”,日本音響学会２００４年春季研究発表会講演論文集，Vol.1, pp.55-56, March, 2004

[2] 俞士汶，段慧明，朱学鋒，孫斌，“北京大学現代漢語語料庫基本加工規範”，http://icl.pku.edu.cn/

[3] 中央研究院資訊科学研究所，中文詞知識庫小組，《中央研究院平衡語料庫的内容与説明》，http://www.sinica.edu.tw

[4] 中国教育部言語応用研究所，“信息処理用現代漢語詞類及詞性標記集規範（請求意見稿)”，２００２年

[5] Hirofumi Yamamoto, Shuntaro Isogai, Yoshinori Sagisaka, “Mutli-class composite N-gram language model”, Speech Communication, 2003, Vol.41, pp369-379.

[6] Jin-Song Zhang, Mitsunori Mizumachi, Frank K.Soong and Satoshi Nakamura, “An Introduction to ATRPTH: A Phonetically Rich Sentence Set based Chinese Putonghua Speech Database Developed by ATR Spoken Language Translation Research laboratories”，日本音響学会２００３年秋季研究発表会講演論文集，3-Q-21, pp167-168，Sep. 2003.

付属表１，中国語形態素の品詞セット（４１個）

品詞タグ		説明	品詞タグ		説明
a		形容詞		nppxj	日本人人名の姓
b		区別詞		nppmj	日本人人名の名
c		連結詞		nppxw	欧米式人名の姓
d		副詞		nppmw	欧米式人名の名
de		構造助詞，“的”など		npl	地名
e		感動詞		npo	組織名
g		語素字		npfd	飲食物名
h		語頭詞		ns	場所詞
i		熟語		nt	時間詞
j		略語		nx	非漢字の符号
k		結尾詞	p		前置詞
m	m	数詞	q		量詞
	ma	数量定詞	r		代名詞
	mb	概数詞	u		助詞
n	n	普通名詞	v		動詞
	nd	方位詞		v1	系動詞“是，系，像是“
	np	固有名詞		v2	関係動詞“有”
	npp	人名		vt	傾向動詞
	nppx	中国人人名の姓		vw	能願動詞
	nppm	中国人人名の名	w		句読点
			y		語気助詞