話し言葉の言語的・パラ言語的構造の解明に基づく『話し言葉工学』の構築

融合研究機関：㈱国立国語研究所
㈱通信総合研究所
（研究総括責任者：古井　貞熙 ）

I　研究の全体計画

1. 研究の趣旨

融合研究の背景：従来の音声言語情報処理研究や自然言語処理研究では，書き言葉の文法に正確に従った言語を対象としてきている。つまり，文法にも発音にも破綻のない文章，ないしはそれを朗読した音声である。しかし，我々が日常用いている自然な話し言葉（自発音声）には，言い誤りや言いよどみが頻出し，文法的にも破格な構造が多い。また，イントネーションのように文字には書き起こすことのできない情報（パラ言語情報）が話し手の意図の伝達に重要な役割を果たしている点も，日常の話し言葉の重要な特徴である。音声認識技術を始めとする音声言語処理技術を真に実用化するためには，自発音声の処理技術，すなわち『話し言葉工学』を開拓しなければならない。

融合研究の必要性：現在の音声言語ないし自然言語処理技術は大量データからの統計的学習に基礎を置いている。そのため，自然な話し言葉を大量に収集し，研究用に編纂したデータベースである「話し言葉コーパス」を構築しなければならない。そのためには，話し言葉を特徴づける様々な言語・音声現象に対する知見が必要不可欠であり，言語学や音声学など，いわゆる人文科学系の言語研究者との共同研究が要請される。本研究では総括責任者の指導のもとに国立国語研究所の有する人文科学的知見と通信総合研究所の有する情報処理技術を融合させることによって『話し言葉工学』の基盤を支える「話し言葉コーパス」の構築を進め，同時に話し言葉工学の要素技術の研究を進める。

融合研究の概要及び目標：研究期間の前半３年間における最大の目標は『日本語話し言葉コーパス』の構築である（サブテーマ２.）。このコーパスには，様々な度合いの自発性を有する共通語音声700万語分（時間にして650時間程度）を収録する他，音声を組織的な手法で精密に書き起こした転記テキスト及び転記テキストを単語に分割して品詞を与えた形態素解析情報を格納する。さらににコーパスの一部，50万語に対しては，イントネーションのラベルなど，パラ言語情報の研究に利用するための様々な研究用情報を付与する。研究期間全体を通じては，このコーパスを利用して『話し言葉工学』の要素技術を開拓する（サブテーマ１.）と同時に，要素技術の実用化をめざした「話し言葉要約システム」のプロトタイプシステムを構築する（サブテーマ３.）。これは，話し言葉を音声認識し，要約した内容をテキストその他の形で出力するシステムである。

2. 開放的融合研究の概要

【１】 研究の概要

1.「話し言葉固有の特徴を利用した「話し言葉工学」基礎技術の研究」

(1) 発話の流れの把握に関する研究

サブテーマ３.の要素技術として，要約処理に必要となる技術を開発する。具体的には，文切り出し技術，重要文抽出技術，文書分割技術の開発を行う。また，要約研究用コーパス（重要文抽出結果）の作成と公開を行う。

(2) 言語情報・世界知識の自動獲得に関する研究

言語の処理に必要となる言語情報・世界知識を大量の言語データから自動的に抽出する手法を開発する。また，サブテーマ２.を効率化する技術として，コーパスの誤り検出技術を開発する。

(3) パラ言語情報の音声特徴の研究

発話の意図や話し手の心的態度といった「パラ言語情報」が，実際の音声コミュニケーションのなかでどのように伝達されているかを，サブテーマ２.で構築する話し言葉コーパスにもとづいて解明する。

(4) パラ言語情報の数値モデル化の研究

言語モデルおよび社会モデルを用いて，個々の発話の待遇値を決定し，敬語表現の生成・理解のモデル化をおこなうとともに，音声による丁寧さの表現についても研究をおこなう。

(5) 話し言葉の文法研究

『日本語話し言葉コーパス』を解析して，話し言葉を特徴づける言語特徴を探索する。それを書き言葉のデータと比較して，話し言葉と話し言葉の文体上の相違点を明らかにする。

(6) 話し言葉自動解析システムの研究

サブテーマ２.で作成する『日本語話し言葉コーパス』に自動で形態素情報を付与するための解析システムを開発する。また，係り受け情報を得るための構文解析システムを開発する。

2.「話し言葉コーパス構築とその効率化に関する研究」

コーパスに収録すべき話し言葉のサンプルの言語学的属性について社会言語学的な観点から検討を加える。

話し言葉の文字への書き起こし基準を確定し，作業用マニュアルを作成する。

コーパスへの情報付与に必要な音素・韻律・形態素・談

話構造などに関するラベル体系を考案し，作業用マニュアルを作成する。

話し言葉の形態素解析を自動化するソフトウェアを開発する。

音声認識技術を応用して話し言葉の音素ラベリング作業の半自動化を図る。

以上の研究に立脚して，音声と形態素解析済み書き起こしテキストを含む700万語規模の話し言葉コーパスを作成する。また，コーパスの一部，50万語程度を対象として，音素・韻律ラベリングを施す。

3.「話し言葉要約システムの研究」

サブテーマ1.，2.の成果を応用して，話し言葉の要約システムを開発する。具体的にはニュースや学会講演など，まとまった内容をもった音声を入力とし，テキストが存在する場合はそれを副入力として，内容の要約テキストを出力するシステムのプロトタイプを構築する。

【2】 融合への取り組みの概要

1. 研究総括責任者の指導性

総括責任者は従来同様，研究全体の円滑な進行を旨とした指導を行うが，その際，特に『日本語話し言葉コーパス』を利用した種々の研究成果の発表を念頭において，指導にあたる。これと並行して海外への研究成果の普及にも努める。特に平成15年4月に話し言葉音声認識に関する公開国際ワークショップを日本で開催するための準備を進める。

2. サブテーマ間の連携

サブテーマ2.で構築を進めている『日本語話し言葉コーパス』のうち，利用可能となったデータは，これまでも随時サブテーマ1.，3.で利用してきているが，平成14，15年度には，種々の研究テーマの実施順序を念頭において，コーパスへの情報付与作業を精密に計画し実施してゆく必要がある。

特に書き起こしテキストの形態素解析作業については，通信総合研究所における解析ソフトウェアの開発，自動解析結果の国語研究所による検討，東京工業大学および京都大学におけるコーパス全体を用いた言語モデルの構築という三つの作業を遅滞なく実施するために，各作業の進行表を作成している。

3. 融合への取り組み

融合研究開始以来，毎月開催してきている融合研究会に加え，特に集中的に対処すべき問題については，通信総研，国語研双方の研究員からなるワーキンググループ（WG）を設置して問題の解決にあたってきている。現在活動中のWGには，自動形態素解析WGと談話構造ラベルWGがあり，これらは平成14年度にも活動を継続する。サブテーマ3の音声認識関係の研究では，京都大学工学部との共同研究が実績をあげてきており，今年度もこれを継続する。その他，国内外の関連する研究機関との間には，非常勤研究員制度ないし外部評価委員会を利用した情報交換を実施してきている。

さらに平成13年度に実施した『日本語話し言葉コーパス』モニター版の公開が，広い範囲の研究者に好評をもって迎えられたので，平成14年度にも形態素情報のモニター公開を実施する。

研究資金は，通信総合研究所・国立国語研究所で独立に運用している。データ作成作業の外注等に関しては，両研究所間で事前に密接な計画をたてて実施にあたっているので，これまでに問題が生じたことはない。

4. 融合研究推進委員会における支援の取り組み

平成13年度には国立試験研究所の独立行政法人化が実施され，これに伴って，両融合研究機関の組織変更，研究実施場所の移動等が生じたが，融合研究推進委員会の支援により，研究に遅滞が生じることはなかった。平成14年度以降，推進委員会において検討すべき重要事項には，融合研究終了後に『日本語話し言葉コーパス』をどのように管理するかという問題がある。この問題は，『日本語話し言葉コーパス』だけに局限された問題ではなく，広く日本の音声・言語研究資源全体の問題と重ね合わせて検討すべき問題である。

3. 年次計画

研　究　項　目	11 年度	12 年度	13 年度	14 年度	15 年度
1.「話し言葉固有の特徴を利用した「話し言葉工学」基礎技術の研究					
(1) 発話の流れの把握に関する研究					
・把握すべき情報の決定と有効な属性の抽出	←	→			
・解析アルゴリズムの検討			←	→	
・工学的実現と検討				←	→
(2) 言語情報・世界知識の自動獲得に関する研究					
・言語情報の自動獲得に関する研究	←	→	→	→	
・対象とする世界知識の確定	←	→			
・世界知識の自動獲得に関する研究			←	→	→
(3) パラ言語情報の音声特徴の研究					
・パラ言語情報・談話情報の体系化	←	→	→		
・パラ言語情報の音声特徴抽出		←	→		
・音声特徴によるパラ言語情報の分類				←	→
(4) パラ言語情報の数値モデル化の研究	←	→	→	→	
(5) 話し言葉の文法研究					
・話し言葉コーパスからの特徴抽出			←	→	→
・書き言葉との比較				←	→
(6) 話し言葉自動解析システムの研究					
・形態素自動解析システムの研究			←	→	→
・係り受け情報解析システムの研究				←	→
2.「話し言葉コーパス構築とその効率化に関する研究」					
・ラベル体系整備およびマニュアル作成	←	→			
・ラベリング半自動化の研究	←	→	→		
・評価用コーパス作成	←	→	→	→	→
・大規模コーパス作成		←	→	→	→
3.「話し言葉要約システムの研究」					
・システムの仕様検討	←	→	→		
・システムの概念設計		←	→	→	
・システムの構築			←	→	→
・システムの評価					← →
所　要　経　費　（合　計）	201 百万円	201 百万円	203 百万円	207 百万円	

Ⅱ　平成14年度における実施体制

研究総括責任者：古井　貞熙（東京工業大学（教授），㈱国立国語研究所に併任）

研　究　項　目	担　当　機　関	研究担当者
1.「話し言葉固有の特徴を利用した「話し言葉工学」基礎技術の研究		
(1)　発話の流れの把握に関する研究	㈱通信総合研究所	○井佐原　均 村田真樹 内元清貴 矢野博之 神崎亮子
(2)　言語情報・世界知識の自動獲得に関する研究	㈱通信総合研究所	馬　青 内山将夫 野畑　周
(3)　パラ言語情報の音声特徴の研究	㈱国立国語研究所	前川喜久雄 小磯花絵 籠宮隆之 菊池英明 米山聖子
	㈱通信総合研究所	山田　篤 小作浩美
(4)　パラ言語情報の数値モデル化の研究	㈱通信総合研究所	白土　保 太田公子
	㈱国立国語研究所	前川喜久雄 籠宮隆之
(5)　話し言葉の文法研究	㈱通信総合研究所	井佐原　均 内元清貴 山田　篤
	㈱国立国語研究所	小椋秀樹 山口昌也 木村睦子 西川賢哉
2.「話し言葉コーパス構築とその効率化に関する研究」	㈱国立国語研究所	○前川喜久雄 小磯花絵 小椋秀樹 山口昌也 籠宮隆之 菊池英明 斉藤美紀 西川賢哉 米山聖子 木村睦子
	千葉大学（㈱国立国語研究所非常勤研究員）	伝　康晴
	京都橘女子大学（㈱国立国語研究所非常勤研究員）	宮島達夫
	大阪大学（㈱国立国語研究所非常勤研究員）	石井正彦

研究項目	担当機関	研究担当者
3.「話し言葉要約システムの研究」	㈱通信総合研究所	内元清貴 村田真樹 山田篤 馬青
	東京工業大学（㈱国立国語研究所併任）	○古井貞熙
	㈱国立国語研究所	前川喜久雄 菊池英明
	京都大学（㈱国立国語研究所非常勤研究員）	河原達也
	日本電信電話㈱サイバースペース研究所（㈱国立国語研究所非常勤研究員）	南康浩
	㈱国際電気通信基礎技術研究所音声言語コミュニケーション研究所（㈱国立国語研究所非常勤研究員）	田中英輝
	㈱通信総合研究所	井佐原均 小作浩美 山田篤

（注：○はサブテーマ責任者）

Ⅲ　融合研究評価委員会・融合研究推進委員会

(1)　融合研究評価委員会

委員	所属
○長尾真	京都大学　総長
板橋秀一	筑波大学　電子情報工学系　教授
白井克彦	早稲田大学　副総長
辻井潤一	東京大学　大学院理学系研究科　教授
土屋俊	千葉大学　文学部　教授
B. Juang	アメリカ　Avaya研究所　研究部長
J. Flanagan	アメリカ　ラトガース大学　副学長
J. Hirschberg	アメリカ　AT&T研究所　研究室長
M. Beckman	アメリカ　オハイオ州立大学言語学科　教授
R. Grishman	アメリカ　ニューヨーク大学計算機科学科　教授

（注：○は研究評価委員長）

(2)　融合研究推進委員会

委員	所属
○飯田尚志	㈱通信総合研究所　理事長
相澤正夫	㈱国立国語研究所　研究開発部門長
井佐原均	㈱通信総合研究所　自然言語グループリーダー
甲斐睦郎	㈱国立国語研究所　所長
木村直	㈱国立国語研究所　理事
酒井保良	㈱通信総合研究所　理事
田中宏	㈱通信総合研究所　総務部長
中山治人	㈱通信総合研究所　けいはんな情報通信融合研究　センター長
福地一	㈱通信総合研究所　情報通信部門長
前川喜久雄	㈱国立国語研究所　研究開発部門第二領域長

（注：○は研究推進委員長）