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

（H11年～H15年、H１３年度予算額：　2.0億円（2.0億円））
研究代表者（研究総括責任者）：古井貞熙（東京工業大学教授）
融合研究機関：国立国語研究所、通信総合研究所

研究の概要・目標

1．研究の目標

自発的な「話し言葉」の情報処理技術、特に、話し言葉からその意味・内容・話し手の意図などを自動的に抽出する技術の基盤を確立する。

3年後の目標
- １千万語規模の話し言葉コーパスのプロトタイプを開発
  （音声入力装置の基本となるデータベース）

5年後の目標
- 話し言葉情報処理システムのプロトタイプを開発
  （検索、要約、自動字幕付与、自動速記起こし等）

2．研究の内容

(1) 大規模な話し言葉コーパス、すなわち音声を文字化し言語情報を付与するとともに、パラ言語情報（音声の文字化によって欠落する情報）も付与したデータベースを構築する。このために、必要な情報の半自動付与技術を開発する。
(2) 上記コーパスを用いて、話し言葉に含まれる言語情報とパラ言語情報の体系化をはかるとともに、両情報を用いた「話し言葉工学」基礎技術の研究を行う。
(3) まとまった内容を持つ音声を自動的に認識し、要約情報を出力する話し言葉要約システムのプロトタイプを開発する。

3．新規性

(1) 欧米の言語と比較して、遥かに複雑で、世界で最も難解な言語の一つと言われている日本語を対象とし、書き言葉より困難な話し言葉の情報処理を目指すものである。
(2) 難解な日本語の話し言葉の情報処理を可能とするため、パラ言語情報を数値モデル化するなど、新しい情報処理技術を構築する。

諸外国の現状

1．現状

欧米では、話し言葉のコーパスの構築が、３～４年前から積極的に行なわれており、すでに大量のコーパスを利用した音声認識研究が活発に行なわれている。

特に米国では、国家予算を用いた研究が大規模に行なわれ、世界をリードしている。

2．我が国の水準

・小規模のコーパスが、複数の研究機関で個別に作られてきているが、これらを総合的に用いることができず、それができたとしても量的に少ない。
・音声認識技術は先進諸国に肩を並べているが、大量のコーパスがないため、研究の促進に問題を生じている。

研究進展・成果がもたらす利点

1．世界の水準との関係

(1) 話し言葉コーパスの量として、世界のトップレベルとなる。
(2) 諸外国のコーパスでは、言語情報のみが付与されているものがほとんどであるのでパラ言語情報が付与されたコーパスとしては世界に類を見ないものができる。
(3) 言語情報とパラ言語情報を組み合わせた話し言葉工学として、独自技術が開発できる。

2．波及効果

(1) インデックスの自動付与により、音声を含む大量のマルチメディア情報の検索が可能となる。
(2) 放送などへの字幕付与の自動化など、福祉技術の向上に寄与できる。
(3) 会議・講演などの速記や文字起こしの自動化の研究を促進する。
(4) 音声ワープロの高性能化の研究開発を促進する。

「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」

1．目的，意義，必要性

本研究は自発的な「話し言葉」の情報処理技術の基盤を確立することを目的とする．

「話し言葉」情報処理技術には，インデックス付与による既存音声データの有効活用，福祉技術への応用，速記や文字起こしの自動化などの波及効果が期待される．

本研究では，通信総合研究所の有する自然言語処理技術と，国立国語研究所の有する言語学的知見を，研究総括責任者の有する音声情報処理に関する知見のもとに統合して研究を推進する．

2．研究概要

サブテーマ１として，パラ言語情報（音声を文字化することによって欠落する情報）など，話し言葉固有の特徴を利用した「話し言葉工学」基礎技術の研究をおこなう．

サブテーマ２として，大規模な話し言葉コーパスを構築する．またその構築のために必要な付加情報の半自動付与技術の開発をおこなう．

サブテーマ３として，まとまった内容をもつ音声を認識して要約情報を出力する話し言葉要約システムのプロトタイプを開発する．

５年後の目標は，

・「話し言葉工学」基礎技術の確立と言語学的知識の体系化
・大規模話し言葉コーパスの構築とその効率化技術の確立
・話し言葉要約システムの構築

である．

3．研究総括責任者　古井貞熙　　（東京工業大学 教授）

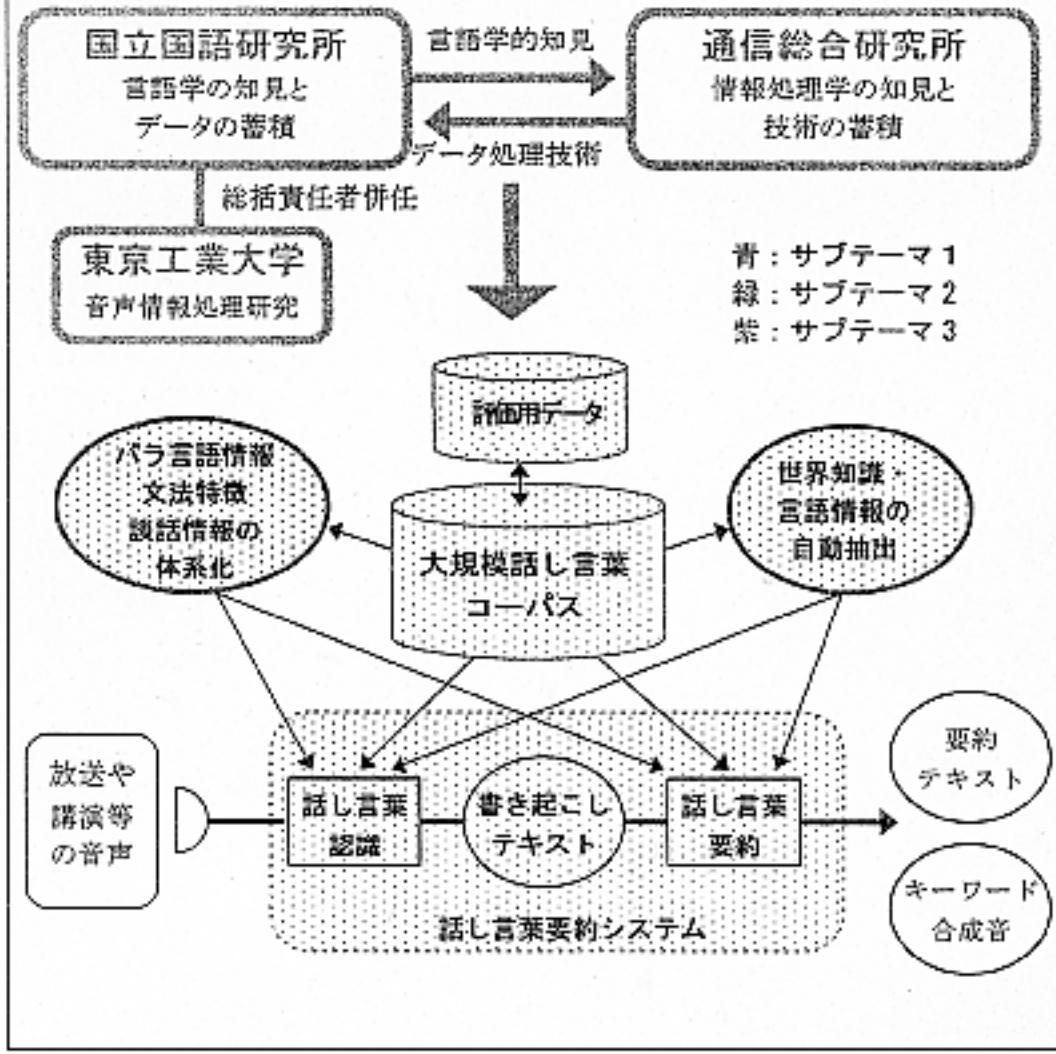
4．融合研究機関

総務省　独立行政法人　通信総合研究所
文化庁　独立行政法人　国立国語研究所

5．研究期間

平成 11 年度～平成 15 年度

（参考図）



融合研究の研究体制について

	通信総合研究所	国立国語研究所	融合の形態
サブテーマ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	研究内容 １　発話の流れの把握に関する研究 話し言葉を対象に書き起こしたテキスト情報や韻律情報を利用して，提案，賛否，話題提起といった発話者の意図を抽出する手法を開発する．また得られた意図を用いて，言語解析の精度を向上する． 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，話し言葉の分類に有益な属性を自動的に抽出する手法を検討する． ３　　パラ言語情報の数値モデル化の研究 大量の実データを統計処理し，個々の言語表現のもつパラ言語情報（例えば待遇表現）を数値化し，データ処理が可能なモデルを構築する． 融合のメリット 通信総合研究所においては，意図抽出，知識獲得についての工学的研究を行ってきた．国立国語研究所の持つ言語・パラ言語データの分類に関する知見を利用することにより，研究の基礎となる意図・知識の客観的分類が可能となる． また国語研究所が実社会で収集した大量の実データを利用することにより，心理実験によらないパラ言語情報の数値モデル化が可能になる．	研究内容 １　話し言葉の文法研究 音声の書き起こしテキストから国語辞典などには登録されていないような表現を抽出し，演劇および映画などにみられる書き言葉における会話表現と比較し，話し言葉の文法的特徴を解明する． ２　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． 融合のメリット 通信総合研究所の有するデータベースからの情報抽出技術を利用することにより，従来から蓄積されてきた言語情報・パラ言語情報に関する知見の妥当性を大量のデータを対象として検証できるところに国語研究所側のメリットがある． サブテーマ２で用意される話し言葉データベースをこの目的に利用する．	通信総合研究所（15 人） （3 人）↑↓（5 人） 国立国語研究所（9 人）

サブテーマ２ 話し言葉コーパス構築とその効率化に関する研究	研究内容 話し言葉コーパス作成の効率化を図るために，話し言葉の音声データと書き起こしテキストの両者を入力して，音声データに話し言葉独自のラベルを付与する作業を自動化する方式を検討し，話し言葉コーパスのラベリングを半自動的に行うシステムを構築する． そして整備が遅れている日本語の話し言葉コーパスを充実させてサブテーマ１の研究を推進するために，１千万語規模の話し言葉コーパスを構築する． 融合のメリット 通信総研が有するデータベース構築の技術と国語研究所が有する音声・言語ラベリングに関する知見を融合することにより，大規模かつ高品質のデータベースを構築することが可能となる．双方が独立して同様のデータベースを構築するよりも，予算や期間を大幅に短縮することが可能となる．	研究内容 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視点から検討を加える． 話し言葉の文字への書き起こし規準を確定し作業用マニュアルを作成する． コーパスへのラベル付与に必要な音素・韻律・形態素・談話機能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． 半自動的に付与されるラベル情報の評価用データとして，100 万語程度を対象に手作業によるラベリングをおこなう． 融合のメリット 通信総研が有するデータベース構築の技術と国語研究所が有する音声・言語ラベリングに関する知見を融合することにより，大規模かつ高品質のデータベースを構築することが可能となる．双方が独立して同様のデータベースを構築するよりも，予算や期間を大幅に短縮することが可能となる．	通信総合研究所(6 人) (3 人) ↑↓ (2 人) 国立国語研究所(8 人)

サブテーマ３ 話し言葉要約システムの研究	研究内容 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構築する． 具体的にはニュースや講演など，まとまった内容をもった音声を主入力とし，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． 話し言葉からの情報受信という点に着目し，話し言葉の内容を効率的に相手に伝達する手法を検討する．具体的には要約文章の提示だけではなく，キーワードの提示，図表の提示，音声による指示などを融合して，効率的な情報伝達環境を構築する手法を検討する． 融合のメリット 高性能の音声認識・生成システムを情報伝達環境の入出力に用いることにより，実験用プロトタイプの開発が容易になる．	研究内容 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構築する． 具体的にはニュースや講演など，まとまった内容をもった音声を主入力とし，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． 多様性に富む話し言葉音声を正確に認識する技術を大規模話し言葉コーパスを利用して開発する． 融合のメリット 通信総合研究所の有するテキスト要約技術を利用することによって，サブテーマ１，２の成果を具体的なシステムに結実させることができる．	通信総合研究所(16人) (0人) ↑ ↓ (6人) 国立国語研究所(9人)

所要経費

（1）研究費の配分一覧（サブテーマ毎）

（単位：千円）

サブテーマ名	サブテーマリーダー	11年度予算	12年度予算	13年度予算
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	井佐原　均	７４，１５２	４４，２０５	４３，０００
2. 話し言葉コーパス構築とその効率化に関する研究	前川 喜久雄	１００，２９７	１２９，３６５	１３３，６１１
3. 話し言葉要約システムの研究	古井　貞熙	２６，５５５	２７，８０７	２６，０７６
合　計　額		２０１，００４	２０１，３７７	２０２，６８７

（2）年度毎予算額推移（機関毎）

（単位：千円）

年度／機関	11年度予算	12年度予算	13年度予算	合　計
１．独立行政法人国立国語研究所	１１１，１０２	１１１，３９８	１１５，９６４	３３８，４６４
２．独立行政法人通信総合研究所	８９，９０２	８９，９７９	８６，７２３	２６６，６０４
合　計	２０１，００４	２０１，３７７	２０２，６８７	６０５，０６８

研究成果の概要

研究期間前半最大の目標は、『日本語話し言葉コーパス』と名づけた研究用コーパスの構築におかれていたので、まずサブテーマ２に触れる。

● サブテーマ２

コーパス構築作業は順調に進捗しており、2001 年 8 月の時点で、目標とする 700 万語の話し言葉音声のうち 85%を収録し 50%の書き起こしを終了した。2001 年 12 月には音声収録を、2002 年前半には書き起こしを終了する予定である。またコアを含む 88 万語に対しては既に手作業での形態論的解析を終了した。現在、サブテーマ１では、この解析結果を学習データとして、形態素解析プログラムの開発を進めており、今年度中には試験的に自動解析を開始する予定である。

コア（50 万語）を対象としたパラ言語情報に関する情報付与のうち、分節音ラベルは現在までに約 20%のラベリングが終了しており、2002 年 12 月終了を目標として作業を進めている。イントネーションは、分節音に比べてラベリング体系の整備が困難であったが、2001 年 8 月にラベリング仕様を確定することができた。今後は 2003 年 3 月終了を目標に作業を進める。

2001 年 8 月には、『日本語話し言葉コーパス』のデータの一部（86 時間相当）を、試用を希望する外部研究者に対してモニター公開した。これに対して、工学・人文科学・社会科学にわたる広い領域の研究者 130 名以上からの申込みがあり、『日本語話し言葉コーパス』への強い期待の存在が判明した。

● サブテーマ１

書き言葉を対象として開発されてきた従来の自然言語処理技術を、必ずしも書き言葉の文法に従うとは限らない話し言葉に適用するための基礎研究を行った。最大エントロピー法による話し言葉書き起こしテキストの形態素解析手法の研究、話し言葉テキストの文相当単位への分割手法の研究、自動分割された文から重要文を抽出する手法等の研究である。これらに加えて、形態素解析されたコーパスに含まれる解析誤りを自動的に発見するための手法の研究も行った。

これらのうち、形態素解析手法と解析誤りの発見手法はサブテーマ２に、重要文抽出手法はサブテーマ３においてそれぞれ利用されている。

● サブテーマ３

話し言葉を対象とした音声認識技術および要約技術の研究を進めた。音声認識実験の結果、『日本語話し言葉コーパス』から構成した音響モデル・言語モデルを使用した場合、朗読音声に基づくモデルを用いた場合に比べて認識率が大幅に（平均 20%から 70%へ）向上することが判明し、コーパスの有用性が確認された。この実験に利用したデータ量はコーパス全体の 1/6 程度であるので、今後コーパスの拡張とともに認識精度の一層の向上が期待される。

しかしコーパスの拡張だけで実用的な認識率（例えば 95%以上）を達成することは困難であり、認識アルゴリズム上の問題を発見し解決することも必要である。そのために、発話速度の変動に対応可能な HMM（隠れマルコフモデル）構築や、文の区切りの決定と音声認識を同時におかなう効率的な文仮説探索法について研究をおこなった。この方面の研究は、以下に述べる要約技術とともにプロジェクト後半の重要な課題である。

要約技術に関しては、音声認識結果から、話題を担う単語の重要度、認識結果の信頼度、単語連鎖の

言語尤度、単語間の係り受け確率などのスコアを計算し、それらに基づいて要約文を自動作成する手法を提案した。また実際に話し言葉音声の認識結果を用いて、種々の要約率の要約文を作成する実験をおこなった結果、原音声の意味を保存した要約文を得る見通しが得られた。

研究成果公表等の状況<課題全体>

【研究成果発表等】

	原著論文による発表	左記以外の誌上発表	口頭発表	合　計
国内	２（３）件	３（０）件	３９（７）件	４４（１０）件
国外	０（１）件	０（０）件	１７（５）件	１７（６）件
合計	２（４）件	３（０）件	５６（１２）件	６１（１６）件

（注：既発表論文について記載し、投稿中の論文については括弧書で記載のこと）

【特許出願等】

なし

【受賞等】

１．電子情報通信学会フェロー，古井 貞熙，平成１３年９月１９日

２．言語処理学会第６回年次大会優秀発表賞，内元清貴・関根聡・井佐原均，「最大エントロピーモデルに基づく形態素解析と辞書による影響」２０００年６月

（計 ２件）

【主要雑誌への研究成果発表】

最大エントロピーモデルに基づく形態素解析---未知語の問題の解決策---，内元 清貴・関根 聡・井佐原 均，自然言語処理（言語処理学会誌），pp. 127-141，2001.

「『日本語話し言葉コーパス』における書き起こしの方法とその基準について」，小磯 花絵・土屋 菜穂子・間淵 洋子・斎藤 美紀・籠宮 隆之・菊池 英明・前川 喜久雄，日本語科学，9 号，pp. 43-58，2001.

（計 ２件）

開放的融合研究に向けた研究体制の概要＜課題全体＞

【研究総括責任者の指導状況】

全体として、音声認識・自然言語処理・言語学の三領域間にまたがる研究を共通の目標に向かって収斂させるために必要な指導をおこなった。サブテーマ２におけるコーパス構築については、各研究領域で要求される仕様の差異に配慮しつつ、プロジェクト全体の目標が達成されるように指導をおこなった。これと並行して、大学等から非常勤研究員を招く、公開ワークショップを開催するなどの方法によって、国立国語研究所、通信総合研究所以外の研究者との連係をとることに努めた。

【サブテーマ間の連携状況】

プロジェクト前半における最重要課題である『日本語話し言葉コーパス』の構築作業（サブテーマ２）には国立国語研究所が主要な役割を果たしたが、データの蓄積が進んでからは、コーパスに含まれる解析誤りの発見に通信総合研究所が協力した。また、現在は、国語研究所において作成された形態素解析結果を学習データとして、通信総合研究所が自動形態素解析プログラムの開発に取りくんでいる。2001年度後半からは、両機関の密接な協力のもと、コア以外の書き起こしテキスト（約620万語）に対する自動形態素解析とその誤り修正を実施する段階に入る。

サブテーマ１において通信総合研究所が開発したテキスト分割技術は、国語研究所が今後実施するコアのラベリング作業におけるラベル初期値の決定に応用する。

同じく通信総合研究所が開発した重要文抽出システムは、サブテーマ３の話し言葉要約システムに応用する。

【開放的融合研究に向けた取り組み状況】

原則として毎月１回、延べ22回の融合研究研究会（全員参加）を開催して意見交換の場とし、サブテーマ間および研究機関間の有機的な協力体制の実現に努めた。

特定のテーマ（形態素解析、エラー検出、談話構造ラベリング）については、国語研究所および通信総合研究所のメンバー少数からなるワーキンググループを編成して問題の解決にあたった。

2001年３月には、プロジェクトで作成したデータの一部（２時間分）を共通データセットとして希望者に公開し、それを様々な角度から分析して発表してもらう話し言葉工学ワークショップを公開開催した。さらに2001年８月にはコーパスの一部(86時間分)を希望者に対してモニター公開した。これによって、コーパスに含まれる誤りやコーパスの利用可能性等について、広い範囲からの意見を反映させる予定である。

【融合研究推進委員会の支援状況】

融合研究推進委員会は年１回の開催を原則としているが、それ以外にも必要に応じて緊密な連絡をとりあっている。過去２年間には、通信総合研究所井佐原グループの京阪奈への移動、両研究所の独立行政法人への移行、さらに国語研究所の大幅な組織改変など、国語研究所、通信総合研究所ともに研究環境が著しく変動したが、融合研究の実施に大きな影響を及ぼすことはなかった。これは融合研究推進委員会のサポートの賜物であった。

評価結果＜課題全体＞

1．進捗状況について

(1) 目標の達成度について

ほぼ予定どおり、順調に目標を達成している。一部については予定よりも進んでいる。

(2) 研究全体の進捗状況について

プロジェクト全体の基盤となる『日本語話し言葉コーパス』の構築は当初予定を上回る進捗を見せている。音声認識および言語処理関係の研究も順調に進展している。プロジェクト後半では、当初の予定どおり、サブテーマ１および３に研究の重点を移行してゆくことになるが、その際、特に以下の点が重要となる。

- 話し言葉書き起こしテキストを対象とした自動形態素解析技術の開発
- 話し言葉音声認識率の一層の向上
- 誤りを含む音声認識結果の言語処理技術の開発
- 以上を実現するために、音声情報処理技術と自然言語処理技術の一層の融合
- 話し言葉処理技術の定式化

2．目標設定について

(1) 当初の目標設定が適切であったか否かについて

適切であったと判断される。

(2) 最終目標の変更の必要の有無について

目標変更の必要性は認められない。

3．研究成果について

(1) 研究成果の科学的価値について

本研究は、話し言葉という新しい研究領域を開拓しつつある。また世界的にも類例のない大規模な研究用データベースを開発しつつあり、研究終了時の公開を予定している。また次節でも指摘するように、本プロジェクトの影響下に海外でも類似の大型プロジェクトが開始されたりもしている。これらの点から判断して、本プロジェクトの科学的価値は総じて高いと判断される。

(2) 研究成果の波及効果について

本プロジェクトからの影響によって米国が音声要約研究を国家プロジェクトのひとつに取り挙げることになった他、英国でも類似のプロジェクトが開始される等、国際的な波及効果を生んでいる。

コーパス構築技術に関しては、国内の複数研究機関に加えて、米国・韓国・台湾の研究機関からも作業マニュアルの公開要請があり、広く注目されている。

(3) 研究成果の情報発信について

種々の国際会議で研究成果を発表する他、米国での国家プロジェクトの会議、種々の学会論文誌（日本音響学会、日本音声学会、言語処理学会等）などで成果を発表することによって、前項に述べたような国際的な関心を呼び起こしている。その他、ホームページによるプロジェクトの紹介も行っている。

コーパス構築のために準備されたマニュアル類は貴重な財産であるので、広く公開することが望まれる。外部評価委員会では、可能であれば英語版を作成してほしい旨の意見も出た。

これに関連して、プロジェクト終了後のコーパスの管理維持について、今からその方法を真剣に考えておくべきであることが、複数の外部評価委員から指摘された。

これまでにもプロジェクト主催のワークショップを２回開催してきているが、プロジェクト後半においても、このような意見交換の場を公開で設け、日本における話し言葉研究をリードしてゆくことが期待される。

４．研究体制について

(1) 研究総括責任者の指導性について

本プロジェクトにおいて研究総括責任者が積極的に指導性を発揮する必要があった問題には、１）コーパスの設計の基本方針と、２）音声研究と自然言語処理研究の円滑な交流の実現のふたつがあった。１）については、様々な専門分野からの要求を取捨選択して実現可能な設計を実施するために、時宜に応じた指導を行ったと判断できる。２）については、プロジェクトの後半においては一層強いリーダーシップを発揮してゆくことが期待される。

(2) サブテーマ間の連携状況について

現在もサブテーマ１と２、サブテーマ２と３の間には、コーパス開発を軸とした有効な連携が認められる。具体的には、サブテーマ２において構築したコーパスのサブテーマ１・３における利用、サブテーマ１の自動形態素解析技術とコーパス修正技術のサブテーマ２での利用、サブテーマ３の音声認識用音響モデルのコーパスに対する分節音ラベリングにおける利用等の点である。

今後、プロジェクト後半においては、科学的研究と話し言葉要約プロトタイプシステムを新たな軸として、より緊密な連携関係を実現することが望まれる。特に東京工業大学を中心とする音声情報処理研究チームと、通信総合研究所を中心とする自然言語処理チームとの成果を、話し言葉要約プロトタイプシステムの構築にむけて有機的に統合してゆくことが必要である。

さらに、現在コーパスに付与されている種々の付加情報、特に韻律ラベル（パラ言語情報）の利用法について、話し言葉要約プロトタイプシステムにおける活用を念頭において再検討することも必要であり、そのためには国語研究所とそれ以外の研究チームの間で、密接な議論が要請される。

(3) 開放的融合研究に向けた取り組み（積極的な融合が図られているか）

これまでの体制はコーパスの構築を軸とした共同研究体制であり、それなりによく機能してきたと認められる。今後は、従来とは異なる目的指向的な融合的研究協力関係を実現する必要がある。具体的には、話し言葉の科学的解明及びプロトタイプシステムの開発を目的とした融合である。

(4) 融合研究推進委員会の支援

プロジェクト開始後に始まった国立試験研究機関の独立行政法人化にともなう研究環境の激変のなかで、遅滞なく研究を推進できており、良好な支援がおこなわれてきたものと認められる。