論文 / 著書情報
Article / Book Information

論題(和文)	大規模知識資源の体系化と活用基盤構築
Title(English)	
著者(和文)	古井貞熙
Authors(English)	SADAOKI FURUI
出典(和文)	情報処理, Vol. 46, No. 5, pp. 488-496
Citation(English)	, Vol. 46, No. 5, pp. 488-496
発行日 / Pub. date	2005, 5
権利情報 / Copyright	

大規模知識資源の体系化と活用基盤構築

古井 貞熙　東京工業大学大学院情報理工学研究科計算工学専攻　furui@cs.titech.ac.jp

文理融合の旗印のもとに，人文系と情報系スタッフの融合によって，大規模知識資源を構築し，それを活用・流通するための基盤を構築することを目的に，研究教育拠点の構築を進めている．マルチメディアドキュメントを含む多様な知識資源を対象とし，大規模知識資源の標準的な体系化と構造化のための枠組みを構築する．多様な知識資源を相互に関連付け，体系化するための意味体系に関する研究とともに，体系化された大規模知識資源の構築を行う．これにより，誰でも容易に知識資源を構築し，体系的に活用することを可能にする基盤を確立することを目指している．さらに，これを通じて，境界領域あるいは領域横断型の新しい学問の開拓と，知識資源研究者の輩出を目指している．

知識資源の時代に向けて

21世紀は知識資源あるいはコンテンツの時代といわれており，あらゆる分野の研究・教育・生活において，大規模知識資源の構築と活用が不可欠になる．これまでの情報化社会の進展を振り返ってみると，1970年頃に，大型コンピュータを多数の利用者が順番に利用する時代（システム中心時代）が始まり，1980年頃からパーソナルコンピュータ（PC）が作られるようになって，1985年頃に，大型コンピュータよりも，小さなコンピュータを各個人が持つ時代（PC中心時代）が到来した．1990年頃からは，インターネットを中心とするネットワークの重要性が意識されるようになり，コンピュータは単独で使われるものではなく，ネットワークにつながることによって，新たな価値を生み出すようになった．2000年頃からは，コンピュータそのものよりも，ネットワークをいかに活用するかが，より重要な時代（ネットワーク中心時代）となった．このネットワーク中心時代の次にくるのが，知識資源中心時代である．ネットワークが知識資源を物理的に結びつけるのに対して，知識資源中心時代の課題は，知識資源をいかにして意味・概念のレベルで結びつけ，体系化するかである．

「知識」と関連のある言葉に，「情報」と「データ」があり，これらは，図-1に示すように，3階層をなしている．データを抽象化したものが情報，それをさらに一段高いレベルで抽象化したものが知識で，知識そのものがいろいろな形で抽象化され，体系化される．知識は，いわゆるコンテンツが観測される規則性や，一般的なルールの表現形態であり，それを用いて推論を行い新たな情報を創出したり，情報に対する解釈を与えたり，問題解決をする源となるものである[1]．したがって，知識は，特定の話題に関する情報が包括的に集積されたもので，情報よりも重い存在であるということができる．知識をいかに体系化するかによって，我々が外界から得ることができる情報が，まったく変わってくる．言い換えると，情報をどうやって活かすかは，我々個人あるいは社会の中に，知識がどのように体系化されているかによって決まる．知識資源とは，大量の知識を利用できる形で蓄積したもので，メタ知識，すなわち知識に関する知識を含むという意味で，コンテンツよりも広い概念を指す．

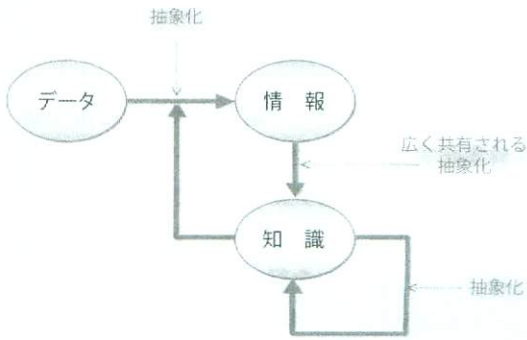

図-1　データ・情報・知識の関係

所属など	氏名など	専門分野
情報理工・計算工学	拠点リーダー：古井 貞熙	音声・マルチメディア
学術国際情報センター	サブリーダー：馬越 庸恭	言語教材
社会理工・人間行動システム	中川 正宣，赤間 啓之	文献資源
社会理工・価値システム	往住 彰文，山室 恭子	意味，感性，歴史資料
学術国際情報センター	横田 治夫，松岡　聡，望月 祐洋	データベース，グリッドコンピューティング，遠隔教育
留学生センター	仁科 喜久子	留学生日本語教育，第二言語習得
情報理工・計算工学	田中 穂積，中嶋 正之，佐伯 元司，徳田 雄洋，徳永 健伸，米崎 直樹，篠田 浩一，佐藤 泰介，亀井 宏行	言語，映像，ソフトウェア，ウェブ，論理，文化財，音声，データマイニング
精密工学研究所	奥村 学	言語要約
COE 特任教授（専任）	佐藤 大和	音声，言語
COE 特任教授（客員）	河合 直樹（NHK 放送技術研究所），前川 喜久雄（国立国語研究所）	マルチメディア，音声，言語

表-1　本COEの構成

世界中で，これまで種々の知識が電子化され蓄積されてきたが，それらは個人や，図書館，研究機関，研究分野に分散しており，統一された概念のもとで開発されていないため，管理・拡張・活用が容易でない．たとえば，自然言語処理や音声言語処理の基礎となる辞書は，多数の研究機関やプロジェクトで異なるものが用いられており，それらを対応付けるのは容易でない．Web検索でも，単語の表記のゆれを始めとして，同じ概念が異なる表記で行われていることが，大きな障害となっている[3]．マルチメディア情報に関しても，それらが多くの利用者の間で共有されるためには，何らかの共通の構造にモデル化され，蓄積されていることが必要である．これまでにすでに多くの知識ベースシステム（知識表現言語で記述された知識表現の集まりと推論エンジンの組合せ）が作られているが，それらを統合して，大規模な知識ベースシステムを構築することは，知識の共有と再利用という観点からも，非常に重要である．そのためには，知識の記述方法がそれぞれの知識で異なっている場合の統合をいかに行うか，などの課題の解決が必要である．以上のように，多様な知識資源を真に関連付けるためには，意味概念のレベルまで踏み込んだ高度な構造化が必要になるが，その実現には，関連する多様な学術を背景とした研究を推進することが不可欠である．

大規模知識資源COEの研究拠点形成実施計画

このような背景から，「大規模知識資源の体系化と活用基盤構築」のCOE研究拠点の形成が計画され，平成15年度に開始された[2]．大規模知識資源を構築し活用できるようにするためには，知識自体の集積のほかに，さまざまな基本技術の創造が必要である．本COEでは，表-1に示すように，東京工業大学大学院情報理工学研究科計算工学専攻，社会理工学研究科人間行動システム専攻，価値システム専攻，および学術国際情報センターを中心とする20名の事業推進担当者を核に組織を構成し，COE特任教授（専任および客員），助手，ポスドク研究者，博士課程学生を加えて，人文社会系・理工系を融合（文理融合）した多様な学際的研究を行う．具体的な知識資源として，話し言葉，書き言葉，マルチメディアコンテンツ，遠隔教育教材，マルチメディア教材，古典文献，歴史文書，文化財知識などを対象とし，新しい学問の開拓と，知識資源研究者の育成を目指す．並行して，拠点形成のための基礎となる大規模計算基盤および大規模情報蓄積・検索・流通基盤の整備を行う．

本COEの拠点形成の目的を，図-2に示す．上で述べた多様な知識資源を対象とし，大規模知識の標準的な体系化と構造化のための枠組みを構築する．多様な知識資源を相互に関連付け，体系化するための意味体系に関する研究とともに，体系化された大規模知識資源の構築を行う．これにより，誰でも容易に知識資源を構築し，体系的に活用することを可能にする基盤を確立することを目指す．同時に，開かれた拠点として知識資源を蓄積・活用し，新しい知見を創造することで，既存学問のより一層の発展と，境界領域あるいは領域横断型の新しい学問の開拓を目指す．本COEから得られる新たな知見は，音声・言語学，メディア学，教育学，歴史学，文献学，考古学などの，新たな発展を促すと期待される．本

拠点が対象とする知識資源には，古典文献，歴史文書（史料），古典美術など，我が国の伝統，文化，言語に直接関連するものが多くあり，本拠点での研究により，それらの知識が大量に蓄積できるとともに，知識資源の体系化により，従来の個別知識のレベルでは不可能であった新しい種々の事実の発見が期待できる．

遠隔教育や第二言語の自律学習のための教材と利用技術の構築も，本COEにおける拠点形成の目的の1つである．また，音声認識を用いたコンピュータとのインタフェースシステムを始めとする，種々の情報処理技術への利用も目的としている．講演録や講義録，テレビ放送への字幕やインデックスの自動付与など，種々の技術への寄与が期待される．

大規模知識資源の構築と，高度な活用を可能とするためには，大規模計算，マルチメディア処理，データベースシステム，知識ベースシステム，XML，データマイニング，情報検索，可視化，セキュリティ，モバイル，ネットワークなどIT技術の支援が不可欠で，それがまたIT関連学問分野の発展に寄与すると期待される．

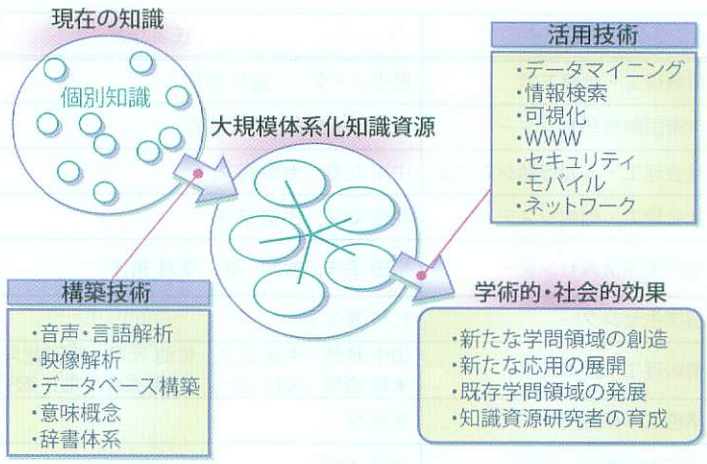

図-2　本COEの目的

図-3　本COEの研究課題の体系

3つの知識資源グループによる拠点形成

本拠点形成が対象とする，大規模知識資源の体系化と活用基盤構築のための具体的研究課題の相互関連を，図-3に示す．大規模知識資源の蓄積，計算基盤，およびネットワークのインフラストラクチャを構築し，それを基盤として，オントロジーを始めとする知識資源の体系化のための基本的な原理やアルゴリズムを研究する．それに基づいて，音声，言語，映像，Webに代表される，素材としての知識資源の構築法を研究し，実際に種々の知識資源を構築する．それらをベースとして，多様なアプリケーションを対象とした，知識資源の活用法について研究し，実際に種々の応用システムを構築する．さらに，これらの多様な知識資源を用いた，文理融合による種々の分析を行う．

拠点形成を組織的に進めるため，言語・文献知識資源グループ，教育・学習知識資源グループ，音・映像・感覚知識資源グループの，3つの知識資源グループを構成して，活動を進めている．表-1に示した各担当者は，必ず1つあるいは関連する複数のグループに属している．以下に，各グループの活動の概要を紹介する．

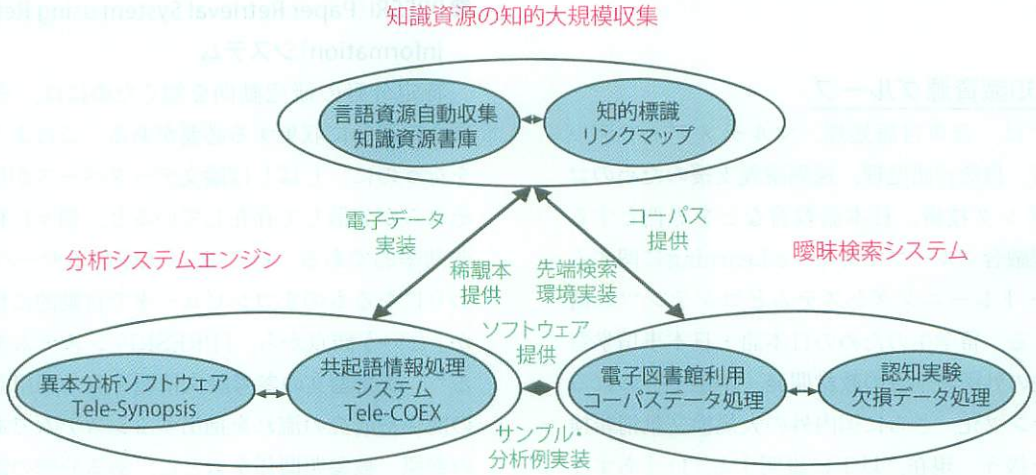

図-4　言語・文献知識資源グループの3サブグループの連携

言語・文献知識資源グループ

本グループは，いわば書き言葉で表現される知識資源を対象とするグループである．図-4に示す，(1) 知識資源の知的大規模収集，(2) 分析システムエンジン，(3) 曖昧検索システムの，3サブグループ間の連携のもと，文理融合により活動を進めている．辞書の作成，さまざまな言語資料の体系化，比喩理解のモデル化など，人文的フィールドにおいてこれまで難しかった新しい知識資源的研究分野を開拓することを目指している．

■知識資源の知的大規模収集

言語知識資源と文献知識資源の収集・構築を，インターネット上で知的，自動的，大規模に行い，しかも資源の質の管理まで行うことを目的とする．例文資源自動収集プロジェクトでは，インターネットを経由して世界中のサイトにエージェント（ロボットあるいはクローラーなどとも呼ばれている）を放ち，自然言語（日本語など）の典型的例文を，数十万例収集している．これに文法構造や意味などのメタデータを付加して知識ベース化すると，機械翻訳や，人間における外国語学習支援システムにとって非常に強力な知識資源となる．一方，人類が生みだした知識の凝縮物ともいえる書物の知識資源化についても，インターネットを経由して世界中の入力者が分散入力し，かつ分散評価も行う仕組みを検討している．世界規模のディジタル図書館を構築するための，要素技術を目指している．

■分析システムエンジン

人文科学に確率的言語モデルを導入し，自然言語処理と人文科学の融合によって，計算人文科学という新しい領域を拓くことを目標とする．特に，単語の共起頻度情報のデータマイニングにより，テキスト言語の異本比較・潜在的文脈解析などを行うため，必要な条件を満たす計量言語学ソフトウェアを構築している．種々のプログラミング言語を利用して，テキストの共起語をさまざまなオプションつきでカウントし分析するシステム「Tele-COEX」の公開を目指している．また言語対象の客観的な比較研究のため，異本テキストを並行フレーム中に共観提示しつつ，それらの間の差異性・類似性を計算して，著者同定や相互引用による系統発生関係の推定を行うためのソフトウェア「Tele-Synopsis」もあわせて開発中である．これらの分析システムエンジンにより，新約聖書のマルコ，ルカ，ヨハネの3つの福音書が書かれた起源に関する分析や，新聞見出し，フランス語語法などの計量言語学的分析で成果を上げつつある．

■曖昧検索システム

言語コーパスを用いた人間的な曖昧検索手法の開発を目的とする．曖昧検索とは，「雲のような犬」という曖昧な表現（ここでは比喩表現）から「白くてふわふわした犬」の画像イメージ等の関連情報を検索することができるようなシステムのことである．本研究では，実際に人間が行うであろう曖昧検索を行うことができるシステムの開発を目的とする．曖昧表現，たとえば「AのようなB」という表現は，「あのBはAのようだ」という形式の比喩表現（直喩）として考えることができる．そこで，本研究では実際に人間が行う比喩理解・生成過程のモデル化を通し，人間的な曖昧検索が可能なシステムの構築を行う．心理学実験を通じて大規模データを収集し，それらのデータを基に，人間が行う比喩の理解・生成過程を明らかにするとともに，人間的な比喩

の理解・生成過程を表現するモデルの構築を行うことを目指している．

教育・学習知識資源グループ

本グループは，音声言語処理，マルチメディアデータの統合技術，自然言語処理，遠隔講義支援のためのコンピューティング技術，日本語教育などを専門とする研究者の文理融合グループである．e-Learningに関する，Webベース・トレーニングシステムとコンテンツの研究を目的とする．留学生のための日本語・日本事情学習およびその他の外国語学習の教材開発，本学で行われる講義のコンテンツ化，さらに国内外の大規模文献情報検索の体系化を扱う．現在，以下に説明する(1)「あすなろ」，(2)「PRESRI」，および(3)「UPRISE」の3つのシステムを開発している．これらのシステムをさらに洗練された有機的な統合マルチメディアシステムとして完成させ，国内外の理工系学生に利用してもらうことを目標としている．これにより，国内外に向けて，本学の教育コンテンツを提供することができる．また，これらのシステムを多くの日本人学生および留学生に利用してもらうことで，国際的なリーダシップを持つ学生を輩出することを目指す．

■日本語読解学習支援システム「あすなろ」

日本語学習者のための，日本語の語の意味や文の構造が多言語によって提示できる，学習支援システムである[4]．現在すでに，Web上で公開中（http://hinoki.ryu.titech.ac.jp）で，学内外で利用されている．その目標は，次の通りである．

(a) 理工系留学生のために，Web上で学習可能な科学技術日本語読解学習支援を開発する．英語圏以外の学習者でも，母語による支援により文章理解できることを目指す．

(b) 細分化された専門分野別，学習者の日本語能力別の学習を可能にする．一斉授業で個々の学習者が満足できる専門読解を目指すことは難しいが，Web上では，個別に学習者に最適な内容を選択でき，学習レベルに合わせた時間配分も可能になる．

(c) 自然言語処理，日本語学，第二言語習得理論（外国語学習理論），教育工学などの学際的視点から，各分野に新しい知見を加える．

システムの主な機能は，学習者が入力した日本語の文章に対し，文章中の単語の訳と文ごとの構文構造を出力することである．その際，Web画面表示や辞書データベースをUNICODEで構成することにより，日本語，英語，マレー語，インドネシア語の他に，中国語，タイ語等の特殊な文字を含めた多言語表示ができる．将来は，日本語学習のみならず，他の言語学習も可能なシステムへの発展を計画している．

■PRESRI（Paper Retrieval System using Reference Information）システム

特定分野の研究動向を知るためには，その分野の論文を網羅的に収集する必要がある．このような文献調査を行うのに，しばしば論文データベースが用いられるが，それらが分散して存在していると，個々に検索するのは非効率的である．そこで，「人が作るサーベイ論文の代わりになるものをコンピュータで自動的に作り出せないか」という観点から，「PRESRI」システムを構築している[5]．研究論文の参考論文情報を有効利用して，論文間の関係や研究の流れを抽出するシステムである．論文間の参照・被参照関係をもとに，ある分野の論文集合を集めてきて表示するとともに，各論文のアブストラクトが読め，ある論文を参照する他の論文中で，その論文に関して記述している部分（参照個所）を読むことができる．参照の目的に関して，手がかり語を用いて，問題点指摘型（他の論文の理論や手法等の問題点を指摘する），論説根拠型（ある理論を提案する場合や仮定をする場合にその根拠となる論文）およびその他に，自動的に分類している．このように，複数の論文を統合するシステムは，高度な複数文要約システムと考えることもできる．

Web上に存在する日英論文データのみならず，国際会議のCD-ROM，種々の組織の図書館にある論文データベース，学会や出版社の論文データベースなどを統合し，参照関係をたどって効率的に関連論文を集める試みも行っている．インタフェースとして，論文間の参照関係をグラフで表示し，ユーザがグラフ上の論文アイコンにカーソルを重ねるとその論文の情報が，参照関係を示す矢印にカーソルを重ねると参照関係が提示できるシステムを実現している．図-5に，システムの構成図を示す．本システムは，http://www.presri.comから利用可能で，日英以外の言語で記述された論文も扱えるようにする拡張も検討されている．今後，多言語意味表示機能を導入することで，さらに広範な利用者に便宜をもたらすことが考えられる．

■UPRISE（Unified Presentation Slide Retrieval by Impression Search Engine）システム

インターネットで情報発信を行い，かつ大量の非定型データやマルチメディアデータを扱うアプリケーションの1つとして，e-Learningが注目されている．インターネットを介した教育コンテンツの配信の試みは，米国のMIT（マサチューセッツ工科大学）を始めとして，国内外の多くの大学などですでに多数試みられているが，これらで配信される教育コンテンツは，シラバスや，講義ビデオ，プレゼンテーション資料などの教育素材がほぼそのままの形で提供されるにとどまっている．

本COEでは，e-Learningシステムにおける教育コンテンツの柔軟な統合機能と高度な検索機能の実現を目指し，講義・講演のビデオ画像とその中で用いられているプレゼンテーション資料を，実際に使われている状況の情報を有効利用して統合的に蓄積することにより，的確なキーワード検索を可能にする「UPRISE」システムを構築している（図-6）[6]．外国語学習の講義などを遠隔地で受講する学生にとって，復習，自習に役立つツールである．

システムの具体的目標は，次の通りである．

(a) 教育コンテンツの統合：講義ビデオとプレゼンテーション資料等を，それぞれのコンテンツは無修正のままで，同期させて提示することで，緩い結合を実現する．コンテンツの修正や動的な統合も可能にする．

(b) 効率的なユーザインタフェース：教育コンテンツの緩い統合を前提に，その特徴を活かした検索機能を実現し，教育効果の高いユーザインタフェースを提供する．

本システムでは，講義を録画したビデオや，講義に使用したプレゼンテーション資料，さらに講義に関連する教科書や配布資料等を格納し，それらの多様な教育コンテンツをメタデータで緩く統合するアプローチをとっている．検索の対象は基本的にこのメタデータとなる．メタデータの記述にはXMLを用い，動画のどの時刻にスライドの切り替えが起こったかという同期情報や，スライドに含まれる文字列へのインデクス，キーワードに対する適合度などが含まれる．動画とスライドの同期情報には，録画時のクリック情報を使うほかに，動画とスライドのパターン認識等を利用する．動画とスライドの対応関係を保持することで，スライドのIDが変更されない限りは，スライドの事後の修正も可能である．このようにして構成されたメタデータにより，講義ビデオだけでは不可能であった統合的な検索機能が実現可能となる．実際のプレゼンテーションを用いた評価を行っており，今後は，音声情報やポインタ情報を統合していく予定である．

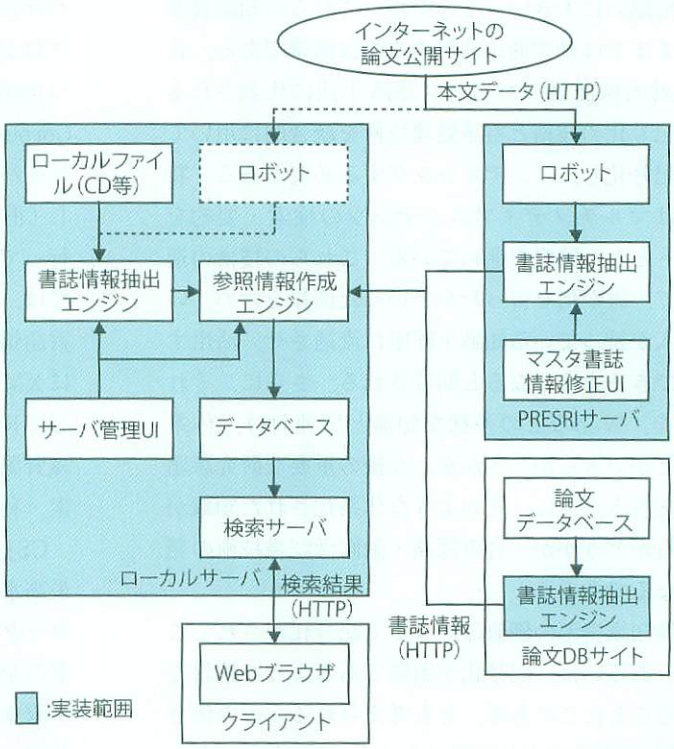

図-5　論文データベース統合システムの構成

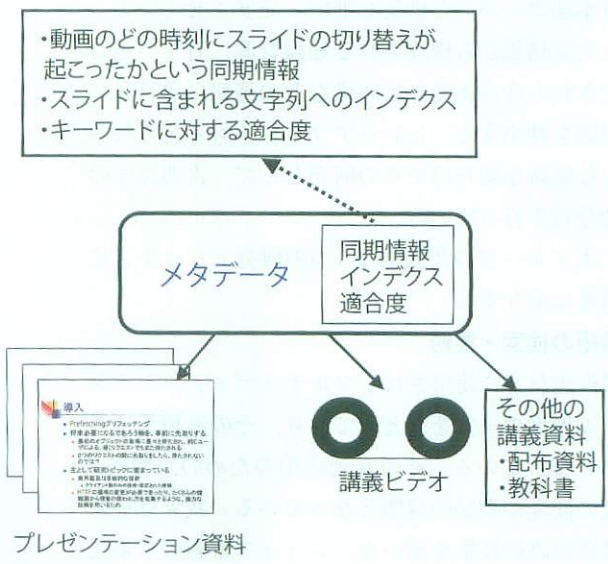

図-6　メタデータによる教育コンテンツ統合

音・映像・感覚知識資源グループ

本グループでは，テレビのニュース，学会や一般の講演，大学の講義，インタビューなど，音声，映像で表現された知識資源（マルチメディアコンテンツ・ドキュメント）の体系化とその応用を目指し研究を行う．これらの資源は，毎日大量に生まれているが，主に信号の形で記録されているため，その大きさは文字化された知識資

源と比べ桁違いに大きい．そのため，これらの知識資源は，このままでは効率的に利用するのは困難である．音声認識，動画像認識のパターン認識手法に代表されるように，記号化の技術と言語処理技術を統合的に用いてデータを記号化し，インデキシングする必要がある．特にここではマルチメディアコンテンツの検索・要約をターゲットとし，研究を進めている．これらの技術の進展によって，誰が何を言ったかといった情報を含め，いろいろな人が持っている知識を即座に流通させ，活用することができるようになると期待される．さらに，それらの知識を，Webなどの多様な知識と関連付け，体系化することができるかどうかが，今後の重要な研究課題となるであろう．逆に，そのような体系化された知識が利用できるかどうかが，音声認識・動画像認識技術の発展の鍵となるであろう．

音・映像知識資源の価値は，今まで記号化がされてこなかった，あるいは，記号化が困難であるような情報を含んでいることにこそある，とも考えられる．その例としては，話し言葉に含まれるアクセント，イントネーションなどの韻律情報が，話し手の意図，感情などを含んでいることがあげられる．そのような知識をどのように抽出し，体系化するかという目的意識のもと，後述する大規模な日本語コーパスの構築を継続・発展させている．さらに，文字情報から構成される知識資源に対し，そこから推定される音声あるいは映像などの情報に着目して，新たな知識を抽出する，というアプローチも考えられる．このような斬新な切り口からの研究として，古典文学の音声学的分析を行っている．

以下，本グループの代表的な3つの研究トピックスについて簡単に紹介する．

■放送番組の検索・要約

テレビ放送などに使用されるマルチメディアコンテンツは，日々増大の一歩をたどっており，その蓄積量は膨大なものとなっている．その有効活用のための，要約・検索手段の開発が喫緊の課題となっている．我々は従来から，音声認識の技術を用いた，ニュース番組の字幕化および要約の研究や，動画像の認識技術を用いた，日常生活における人間のさまざまな動作のモデル化と認識の研究を行ってきた．ここでは，このような技術を基盤として，音声情報と映像情報を共に用いた検索・要約の研究を行う．放送ビデオのインデキシングでは，音声情報と映像情報をいかに組み合わせるか，対象（野球，サッカーなど）の知識をいかに用いるかが鍵である．NHK放送技術研究所との共同研究を開始している．

■話し言葉コーパスの構築と活用

本COEに先立つ，国立国語研究所，通信総合研究所，および本学の共同プロジェクト（「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」プロジェクト）により，日本語の話し言葉750万語（のべ650時間）を格納した「日本語話し言葉コーパス（CSJ: Corpus of Spontaneous Japanese）」が構築されている[7]．コーパス全体について，セグメンテーション，書き起こし（正書法による基本形と発音形），形態素解析などを行っているが，全体の約7%（50万語）の「コア」については，人手による形態素解析のほか，韻律情報などパラ言語情報まで含めたコーパスとなっている．CSJはすでにXML化されているが，種々のユーザの利便性を考慮してRDB化を進めており，本COEで構築中の大規模知識資源蓄積システムに蓄積して，外部の研究者などが検索・利用できるようにする準備を進めている．

CSJは，話し言葉コーパスとしては質・量ともに世界最高水準のデータベースであるが，講演音声などのモノローグを主な対象としている．CSJを拡張して，話し言葉の全体像を適切に把握するためには，どのようなデータ収集手法が有効であるかを検討し，多様な音声を収集して，話し言葉データベースの体系化をはかる．このために，放送ニュースの音声のように書かれたテキストを読み上げた音声と，通常の話し言葉音声の音響的・言語的違いの分析を，国立国語研究所との共同研究として進めている．

話し言葉音声には，言い直し，言い誤り，間投詞など，冗長な表現が多く含まれているほか，それを音声認識した結果には，認識誤りがどうしても含まれてしまう．一方，音声ドキュメントの検索などにおいては，音声をそのまま録音したり書き起こしたりするだけでなく，要約したテキストや，要約音声で提示することが求められることが多い．音声認識の結果として重要なのは，あくまでも話し手の意図を抽出することであって，音声がそのままどういう文字で表されるかは，2次的な問題ということもできる．このような背景から，これまでに，音声認識結果に対して，重要文抽出と単語抽出の組合せによって，自動的に冗長な部分や認識誤りを除き，重要な部分だけを抽出することにより，音声を要約する方法を提案して，その有効性を確認している[7),8)]．図-7に，要約システムの構成を示す．要約性能を定量的に評価する方法も提案している．現在，日本語だけでなく英語音声の話し言葉コーパスを構築しながら，要約技術のより一層の性能向上をはかる共同研究を，京都大学や欧米の大学と始めている．

■古典文学の音声学的分析

日本文化の貴重な財産として，数多くの古典文学が残されており，国文学資料館などでディジタル文化資源として一般に提供されつつある．これら古典文学については，これまで主に書き言葉（文字文学）として研究がな

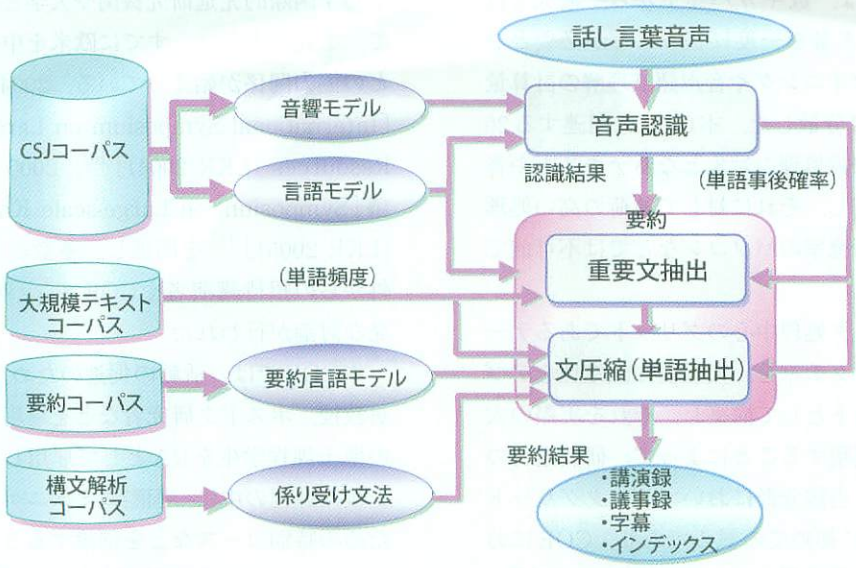

図-7　音声要約システムの構成

されてきた．しかし，古典文学のなかには，そもそも演じたり，物語ることを目的として作られ，結果として文字化されて残っているものがある．また，語ることを目的とした作品であっても，文字文化の浸透によって，話し言葉に影響を及ぼしていることもあると考えられる．本COEでは，文字言語として継承されている古典文学について，書かれた文字の背後にある音韻や韻律など音声的側面の特徴に関して統計的研究を行い，日本語における話し言葉と書き言葉の性質とその変遷を，体系的に研究しようとしている．比較対照の例として，まず「源氏物語」，「平家物語」などをとりあげ，相互の特徴と違いについて分析を行っている．すでに，日本の古来の詩歌のリズムの階層構造がほぼ黄金比になっていることなど，興味深い結果が得られている．本COEの目的の1つである文理融合の典型となる研究を目指している．

大規模知識資源の蓄積および計算基盤の構築

大規模知識資源蓄積システム

本COEでは人文社会学系・理工学系の研究者が連携して知識資源を構築していくことを目指しており，そのためには多種多様の大量の知識素材を蓄積し，それらを整理しながら利用して研究を進めていくことが肝要である．幸い近年の情報技術の発達により，テキスト，図表，画像，音声，動画など知識資源の元となる多くの素材が電子化され，情報環境を利用してそれらを蓄積して効率的に研究を進めることが可能となってきている．しかし，そのような電子化された多種多様で大量の素材を有効利用するためには，それらを統一的に蓄積するとともに，それらに対する高度な検索機能を提供することが必要となる．

本COEでは，そのような知識資源構築のための研究基盤として，大規模知識資源蓄積システム(KS：Knowledge Store)[9]の開発を行っている．KSでは，さまざまな利用形態や多様な素材の蓄積，さらに格納手法に関する試行錯誤等を可能とするため，柔軟性・拡張性に重点を置き，共通する基本的な蓄積・検索機能を提供し，利用分野ごとに外部システムを用意する方針をとっている．その内部では，各素材を統一的に扱い高度な検索を可能とするため，素材に関する情報であるメタデータを用い，さらに柔軟なメタデータ定義を可能としている．また，外部システムとの間はネットワーク環境での利用を想定して標準化されているWebサービスインタフェースを提供する．利用者は外部システムを通し，あるいは直接WebインタフェースでKSを利用することができる．

大規模計算基盤システム

本COEでは，さまざまな形式の自然言語，画像，音声に関する信号処理，知識処理，およびデータマイニングを行う．また，従来型の数値処理や，Webに対する

自動情報収集なども行われる．セマンティックWebのデータマイニングでは，数ギガバイトから，最大では数テラバイトのデータ量を一度に扱うことが必要となり，また，自然言語マイニングや音声認識理解の計算量は莫大となる．学内に分散した，本COEに関連する20の研究室が，有効に知識処理の対象となるテキストや音声データを作成・共有し，それに対して負荷の高い処理を行うのは，従来の研究室のパソコンなどでは不可能である．

そこで，我々はデータ処理中心のグリッドであるデータグリッドのアーキテクチャを，大規模知識処理をアプリケーションターゲットとして構築し，それを実際の大規模計算基盤として実現することによって，研究遂行の礎とすることとした．当該分野においてデータグリッド技術の適用は世界的に初めての試みであり，COEにおける研究活動での実運用を通じて，高信頼性・冗長性・可容性・高いアクセシビリティの実現のための技術的要件の洗い出しと実現，ならびに学内キャンパスグリッドを中心とした他の情報インフラとの連携を実現することとした．すでに，データグリッドサーバ群のアーキテクチャの設計を行い，調達・設置を行った．今後，実際のさまざまな知識処理を行い，かつデータグリッドとしての研究開発を行いながら，さらにハードウェア・ソフトウェアとも拡充していく予定である．

夢の実現へ

本COEは，壮大な夢のもとに計画され，拠点形成が始まった．文理融合という旗印のもとに，人文系スタッフと情報系スタッフの融合によって，大規模知識資源を構築し，それを活用流通するための，新しいアルゴリズムや体系化の研究と教育が始まっている．この拠点形成により，次のような成果が期待されている．

(1) マルチメディアドキュメントを含む多様な知識の体系化，遠隔教育や第二言語の自律学習のための教材と利用技術の構築，教科書・古典文献・文化財・歴史資料などの知識資源化による相互関連の分析や体系化．
(2) 研究活動における知的創造活動の強化．
(3) 多様性に富みかつ効率的な教育とともに，境界領域あるいは領域横断型の新しい学問の創造．
(4) 種々の知識処理技術の発展．
(5) 文科系・理工系を融合した幅広い学際的知識，高い学力，教養および論理的思考力，さらに幅広い国際性を備えた新しいタイプの知識資源研究者の輩出．
(6) 大規模知識資源を対象とした研究センターの設立．

これらの拠点形成の実を上げるためには，国内のみならず国際的先進研究機関や大学と協調することが必要で，上述のように，すでに欧米を中心とする多数の機関との協力関係が始まっている．2004年3月には国際会議「International Symposium on Large-scale Knowledge Resources (LKR 2004)」[10]，2005年3月には国内会議「Symposium on Large-scale Knowledge Resources (LKR 2005)」[11]を開催し，多数の参加者を得て，国内外からの招待講演者やCOEメンバの講演を中心に，活発な討論が行われた．

本COEでは，活動の促進のため，COE特任教授，客員教授，ポスドク研究者などを雇用しているほか，多数の博士課程学生をRAとして雇用している．関係する博士後期課程の中に，国際コミュニケーション能力を養うための特別コースなどを創設するとともに，博士課程学生の自主的な活動の基盤としての「博士フォーラム」を立ち上げている．すでに多数の学生が，成果の発表，討論，共同研究などのために，海外に派遣されている．

本COEがねらう大規模知識資源の体系化と活用基盤構築は，次世代Webとも関連して，すでに世界的なレベルでのホットな研究課題となりつつあり，諸外国と協力しかつ切磋琢磨しながら，研究と教育の実を上げていくことが求められている．

参考文献

1) 西尾他：情報の共有と統合，岩波講座マルチメディア情報学，7，岩波書店（1999）．
2) 古井：21世紀COEプログラム「大規模知識資源の体系化と活用基盤構築」，電子情報通信学会技術報告，SP2003-162，pp.47-52（2003）．
3) 萩野他：特集 セマンティックWeb，情報処理，Vol.43，No.7，pp.707-750（July 2002）．
4) 阿辺川他：多言語表示日本語読解学習支援システム「あすなろ」の開発，第3回「日本語教育とコンピュータ」国際会議Castel/J 2002論文集，pp.71-74（2002）．
5) 難波他：Web上のデータを中心とした複数論文データベースの統合，未踏ソフトウェア創造事業報告書（2003）．
6) Yokota, H. et al.: UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine, IEICE Trans. on Information and Systems, Vol.E87-D, No.2, pp.397-406（2004）．
7) 古井：話し言葉の音声認識と自動要約，第3回話し言葉の科学と工学ワークショップ講演予稿集，pp.1-6（2004）．
8) 堀他：単語抽出による音声要約文生成法とその評価，電子情報通信学会論文誌，J85-D-II，2，pp.200-209（2002）．
9) Yokota, H.: An Information Storage System for Large-scale Knowledge Resources, Proc. International Symposium on Large-scale Knowledge Resources (LKR 2004), Tokyo, pp.87-90（2004）．
10) Proceedings of International Symposium on Large-scale Knowledge Resources (LKR 2004), Tokyo（2004）．
11) Proceedings of Symposium on Large-scale Knowledge Resources (LKR 2005), Tokyo（2005）．

（平成17年1月10日受付）