3 大きなコーパスを共有しよう

田中 穂積 tanaka@cl.cs.titech.ac.jp
東京工業大学大学院情報理工学研究科
森口 稔 moriguch@cow.nara.sharp.co.jp
シャープ株式会社パソコン事業部
亀井 真一郎 s-kamei@cw.jp.nec.com
NEC情報通信メディア研究本部
加藤 安彦 kateaux@kokken.go.jp
国立国語研究所言語体系研究部

なぜコーパスの共有が大切なのか？

自然言語は自然現象である

日本語や英語といった人間の言葉「自然言語」は，人類の誕生とともに自然発生し，長い時間をかけて変化し現在の形になったものである．人間が意図して設計・構築した人工物ではないので，物理や化学の現象と同様の，一種の自然現象とみなすことができる．人間はごく簡単に言葉をあやつることができるため，言葉が自然現象であるという考え方には一般にはなじみが薄いかもしれない．しかし，言葉の使い方の規則は誰が決めたものでもなく，その規則をすべて完全に記述することは誰にもできない．

コーパスは自然観察の基本環境である

したがって，言葉の研究は，他の自然現象の場合と同じく，まず事実の収集・観察から始めなければならない．実際に言葉の使用例を大量に集め，言葉が使われている状況を広く分析的に観察し，個々の語の振る舞いを詳細・正確に記述し，その背後にある規則をモデル化することによって初めて，言葉を技術として取り扱うことができるようになる．これが「コーパス（corpus；書かれたり話されたりした文章を集めたもの）」や「言語資源（Language Resource；辞書や各種コーパス類，およびそれらを作成・利用するツール群を合わせた環境）」が言語研究にとって重要な理由である．

コーパスから言語的知見を得て自然言語処理技術を高度化させる方法についてはいくつかのアプローチがある．規則ベース（コーパスを分析して従来は扱えなかった言語現象を見出し，それを規則化して枠組みを拡張する），統計ベース（コーパスを分析してある現象の統計的傾向を求め，それを組み込む），用例ベース（コーパスに含まれる例文そのものを直接利用する），あるいはそれらのハイブリッド型と，そのアプローチは異なっても，実際の言語現象を大量に集めたコーパスが必要である点は共通している．

自然言語を扱う分野が広がっている

近年，言葉を処理対象とする技術が急速に拡大している．インターネットの発展やパソコンの普及によって，大量の電子テキストが流通するようになった．それを有効活用するために，情報検索，情報抽出，要約，テキスト・マイニング，アンケート分析，個人向け広告，情報フィルタリング，といった新技術が急速に発展してきている．また，インターネット上では英語をはじめとする世界中の言葉で情報が記述されているから，機械翻訳やクロス言語検索といった言語の壁を越える技術の重要性が再認識されるようになった．さらに，音声技術が高度化し，音声ワープロや，声で機械を操作する技術も実用になってきた．自然言語処理技術は，これら言葉に関係するすべての技術の共通基盤であり，コーパスは，その自然言語処理技術の高度化に必要不可欠な基本環境ということができる．

コーパスを作るのには大変な労力がかかる

コーパスから有意義な知見を得るにはコーパスは大量になければならないが，大規模コーパスの構築には膨大な労力がかかるという問題がある．日本語や英語の場合，日常的に数十万語が使用されるが，仮に特に重要で使用頻度の高い数万語の使われ方を調べるために，各語の使用例を百例ずつ集めたとしても数百万の例文が必要となる．実際には，経済記事なのか，日常会話なのかといった内容（ドメイン・分野）の違いによって，使われる語彙も表現も意味も大きく異なるので，それぞれに対して大量の例文が必要

となり，全体として非常に大規模なコーパスが必要となる．

近年，言語処理の基本技術，たとえば，文の単語への分割（形態素解析），文の構造の認定（構文解析）などの精度が飛躍的に向上したとはいえ，コーパスから言語知識を自動獲得できるまでには至っていない．したがって，元の生のコーパスにツールで単語区切りや文構造の情報を付与した後に，必ず人が目で見てチェックする必要がある．また，日本語と外国語にまたがる言語的知見を得るために，たとえば日本語と英語が対になったパラレルコーパスを作ろうとすれば，人の手による翻訳が必要となり，時間面・費用面で多大なコストがかかる．このような理由から，個別の機関において大規模なコーパスを開発するのには限界がある．

言語資源共有機構（GSK）の設立

コーパスは自然言語処理研究の基盤であるが，前章で述べたように，個々の機関で大規模に開発するのは困難である．そこで，コーパス類を集積し流通させることによって，我が国の言語関連分野の研究を推進する組織として「言語資源共有機構（略称GSK）」が設立された．その役割としては大きく次の2点がある．

(1) 言語資源の保持者と利用者の間の基本合意の確立

一般に，一定量以上の質の良い言語データ（生のコーパス）は多くの場合，自然言語処理の研究機関（たとえば大学や企業）とは異なる出版社や新聞社などで開発される．そのため，言語データを利用したい者は，個別にデータ保持者と著作権や価格を交渉しなければならず，多大な労力がかかる．また言語データを自然言語処理に利用するという使い方は今まで想定していなかった利用形態であるため，データ保持者側にも言語データの提供に躊躇や戸惑いがみられる．また言語資源提供・利用のための一般的ルールも確立していない．このような状況が結果的に我が国の音声・自然言語処理技術の研究・発展の著しい阻害要因となっている．そこで言語資源の保有者，利用者の双方が納得できる提供・利用の仕組みの確立を目指す．

(2) 言語資源の流通促進

GSKは，言語データ，自然言語処理用ツール，シソーラスなど，自然言語にかかわる種々の知識，すなわち言語資源一般に関して，その所在と入手方法などの情報を一括管理し，言語資源の保有者と利用者の仲介をすることで，言語資源の流通・共有化の促進を目指す．

GSKが仲介することで，言語資源の保有者，利用者の双方にメリットが生じる．言語資源の保有者にとっては，自分の保有している言語資源を新しい用途に供することができ，新たな需要を広く喚起して利益につなげることができる．利用者にとっては，必要な言語資源の所在が容易に分かり，それを繁雑な交渉を経ずに安価に利用できる点が最大のメリットである．また言語資源は，いったん作られた後，適切な維持管理が行われず死蔵されてしまう危険があるが，GSKによってそのような言語データの有効活用を図ることができる．このようにして言語資源を流通させ活用することで言語処理技術の高度化が進めば，情報化社会に氾濫する言語情報を有効活用する新しいサービスが実現される可能性がある．これにより新たな市場が開けて，情報化社会の健全な発展への貢献が期待される．

欧米では，次章で述べるように，言語資源の重要性が早くから強く認識されており，すでに公的支援をベースにした会員制コンソーシアムが設立されている．我が国でも，早期にGSKの活動を軌道に乗せ，音声・自然言語処理関連技術の発展を一層加速する必要がある．

海外の動向

欧州における動向——

欧州には，英語を中心として多数のコーパスが存在し，その流通基盤も整って，研究や教育に活用されている．本節では，日本に先んじてコーパスへ取り組んできた欧州の現状について概観する．

英国の大規模コーパスと辞書編纂

英国にはBritish National Corpus (BNC) とBank of English (BoE) という2つの大規模コーパスが存在する．

BNCは，英国政府からの援助を得て，オックスフォード大学が中心となり，ランカスター大学，British Library，主要辞書出版社3社の協力のもとに構築された．データの採集に際しては，辞書出版社のLongmanが話し言葉を，オックスフォード大学出版局が書き言葉を主に担当した．現在1億語からなるBNCは，Oxford University Computing Servicesで保守管理され，主として自然言語処理研究者や辞書編集者が利用している．たとえばオックスフォード大学出版局では，Oxford English Dictionary (OED) にかかわる40名以上の辞書編集者がこのBNC

名称	Australian Corpus of English	Bank of English	British National Corpus	Brown University Standard Corpus of Present-Day Edited American English	Corpus of Professional, Spoken American-English	French Polyphone Database(SpeechDAT(M))
略称	ACE	BoE	BNC	Brown Corpus	CPSA	FRESCO
管理・作成機関	Macquarie Univ.	COBUILD	Oxford University Computing Services	Brown University	Athelstan	Philips, SPEX
入手先	同上	同上	同上	ICAME, OTA	同上	ELRA
Written/Spoken	written	mixed	mixed (w90, s10)	written	spoken	spoken
Language	Australian English	English	British English	American English	American English	French
語数	100万	3億	1億	100万	200万	35,000 utterances
対象年代	1986-	1990-	1975-	1961	1994-1998	不明
備考	LOBのオーストラリア版	—	OUP, Longman, Chambers Hallap, UCREL, British Council	世界最初の英語電子コーパス	49米ドル	—

表-1 海外の主要コーパス（1）

を語義・用法の確認といった編集作業のために利用しており，さらに一般から「OEDにないことば」をみつけたら知らせてもらうような体制をとって辞書見出しとコーパスデータの拡充に努めている．

Bank of Englishは，3億3千万語（1998年分までのデータ）からなるコーパスで，バーミンガム大学内に拠点を持つ辞書出版社Collins COBUILDが作成し，現在もアップデートを続けている．BoEの当初の目的は英語を母語としない学習者のための辞書編纂であったが，英語の母語話者からも新語のチェックなどの目的での利用がある．また，コーパス内での語の出現頻度から判断して，高頻度語を除いた母語話者のための辞書も企画・出版されている．

コーパス流通機関

大規模なコーパスを構築する一方で，既存のコーパスの流通も言語研究には重要な要因である．欧州には，パリに拠点を置くEuropean Language Resource Association（ELRA）と，ノルウェーのベルゲンに拠点を置くInternational Computer Archive of Modern and Medieval English（ICAME）が存在する．

ELRAは，EUからの援助の下，1995年ルクセンブルクに設立された．その後パリに移り，現在，会員数は約100を数え，ELRAの財政の約50%が，コーパス流通による収入で支えられている．会員の多くは，欧州域の自然言語処理研究者で，多言語コーパスやスピーチコーパスなど，80を超えるさまざまなコーパスを利用している．ELRAの運営は，ニーズ調査，マーケティング，プロジェクトの管理，年4回のニューズレターの発行，広報活動などを含み，6人の専任スタッフが携わっている．また，関連領域の専門家集団がパネルという形で年2回のミーティングを開き，データの妥当性の検証，マニュアルの作成，法律に関する助言なども行っている．コーパス流通の際に焦点となる標準化については，ELRA自身は関知せずユーザからクレームがあったり，修正の必要がある場合にのみ何らかの介入をしている．

一方，ICAMEは，20ほどの英語コーパスを保有しており，その中には欧州のみならず，米国で作られた世界最初の英語コーパス，Brown Corpusや，オーストラリア，ニュージーランド，インドのコーパスも含まれている．利用はCD-ROMの購入という形をとる．

コーパスを取り巻く環境

歴史的・地理的要因により言語意識の高い欧州ではコーパスそのものの整備が進んでいるだけではなく，それを取り巻く環境も整っている．言語関係の研究者だけではなく，語学教師や翻訳者などの実務者もコーパスの利用に積極的であり，自然言語処理技術者と言語学者との協力関係も進んでいる．また公的機関からのバックアップも得やすく，出版社も著作権について協力的といえそうである．たとえば英国での大規模コーパス構築にあたっても著作権保有者に対価を支払うことはなかったし，ELRAでも著作権に関する法的問題は今のところ起こっていないそうである．

日本との関係

日本はコーパスの整備という観点で，欧州に大きく立ち遅れている．著作権問題も大らかな欧州の事情に比べると，意識し過ぎて「跼天蹐地（きょくてんせきち）」ともいうべき状況に陥っているようにも見える．もっとコーパス化ということに関し，新聞社，出版社，著作権者などに欧州並みの理解を得ておくことが必要である．欧州のコーパス関連機関は，日本のGSKの活動に期待しており，協力的な姿勢を示

名称	International Telecommunications Union	Kolhapur Corpus of Indian English	Lancaster/Oslo-Bergen Corpus	Multilingual Corpora for Cooperation
略称	CRATER, ITU	—	LOB	MLCC
管理・作成機関	Corpus Resources and Terminology Extraction Project (MLAP-9320)	Shivaji Univ.	Lancaster Univ.,University of Oslo, Norwegian Computing Centre for the Humanities	—
入手先	UCREL, ELRA	ICAME	ICAME, OTA	ELRA
Written/Spoken	written	written	written	written
Language	English/Spanish/French	Indian English	British English	major European Languages
語数	各100万	100万	100万	言語により異なる（最低100万）
対象年代	1980s	1978	1961-	1986-1994
備考	電気通信分野	Brown (米), LOB (英) のインド版	ブラウンコーパスとの対比を可能にするために規模，手法などを同じにして開発	6カ国語の経済誌と，9カ国語の Official Journal of the European Commission

表-2　海外の主要コーパス（2）

てくれているので，それに応えるべく我々も努力すべきときが来ているといえよう．

米国LDCの動向 ——

本節では，米国で言語資源の収集と頒布の中心となっているLDC (Linguistic Data Consortium) に焦点を当てて報告する．

LDCとは

LDCは，1992年Defense Advanced Research Projects Agency (DARPA) およびNational Science Foundation (NSF) の援助を受けて設立された，ペンシルベニア大学の非営利団体である．現在はLDC会員からの会費とコーパス頒布収入による独立採算形態をとっている．専任の職員は20名ほどで，データの作成や頒布の繁忙期にはパートタイムで20～30名ほどの作業者を雇い入れるようである．研究・技術開発・教育といった方面から，数千万語に及ぶテキストや数千人の発話データ，数万語からなる辞書など，大量データに対する要望が高まってきたにもかかわらず，それをこなせるほどの人材，設備，財源を持った組織がなかったことがLDC設立に至った直接の動機であるという．そうした設立の動機に基づき，言語資源データの構築と公開，データ収集の推進と保管を使命とし，言語に関連する教育や研究，実践的な練習，技術開発に積極的に利用できるデータの収集・頒布を行っている組織がLDCという組織である．

LDCの基本方針

- 研究者の必要とするデータを公開
- 万人へのデータ提供
- 言語資源共有化の促進

というのが，LDCの基本方針である．誰もが会員となることができ，非会員であっても対価さえ支払えばほとんどのデータベース利用が可能である．ただ，コーパスについては，非会員でも入手可能なデータはあるものの，多くは会員限定となっているようである．会員は，会員となった年次に公開・頒布された言語資源を1部ずつ無償提供される．また，言語資源共有化を促進する観点から，データ提供者とデータ利用者間における知的所有権に関する仲立ちをしたり，知的所有権に関してのアドバイスも業務として行ったりしている．公開にあたっては，より利用価値の高い公開とするための標準の規定とツールの開発を同時に行っている．

会員と公開・頒布の現状

現在までのところ，会員数は約300で，公開した言語資源の数は，発話コーパス，テキストコーパス，コーパスから得られた語のリスト（レキシコン）をあわせて200種類近い．700近い非会員への頒布も含めて9,000件以上の頒布を行っている．

アジアの動向 ——

ここでは，中国と韓国におけるコーパス事情について述べる．中国については，北京大学におけるコーパスに基づいたデータベース作成の現状を，また，韓国については国立国語研究院におけるコーパスを利用した辞書作成の現状に触れる．

北京大学

中国全体では北京大学，精華大学をはじめ，さまざまな大学や研究機関がコーパス（語料庫）作成を行ってきている．ここではそうしたコーパス作成の進め方として北京大学に例をとる．

北京大学では人民日報を対象とした「人民日報コーパス」を作成している．コアにあるのは人民日報

1年分，約2,600万字のデータで，そこに新たな人民日報データが付加されつつある状況である．このテキストデータをもとに検討を加えた文法的・意味的な情報をタグとして付与する作業が行われている．

このプロジェクトには，北京大学計算語言学研究所（計算言語学）を中心に，北京大学中文系（中国語学），上海師範大学，南京師範大学，烟台師範学院，北京語言文化大学が加わり，総勢50名ほどの研究者が参加している．

作業の進め方としては，北京大学計算語言学研究所が提供する形態素解析ツールによってデータ加工し，その加工されたデータのチェックを協力体制にある他大学が行って，最終的にまた北京大学でチェックを行って解析済データとするとのことである．付与されるタグは，品詞が約40種類，語素と呼ばれる文法範疇と意味範疇にまたがる情報成分が7,223種類ある．これらは共に1年分のデータを加工しながら検討されて改良を加えられてきたものである．

人民日報の著作権については研究利用に目的を限定し，対価を支払って許諾を得ているとのことである．中国国内でも以前に比べるとはるかに著作権に関する意識が高まってきているようである．

言語資源の共有を目指した機関の設立については，まだ議論されていないようだが，いくつかのコーパスを中心として，大学・機関のグループが存在し，そのグループを単位にデータやツール類の共有化を図っている様子である．

韓国

韓国では過去数年間に少なくとも7つのコーパスが作られているが，その中で最大のものは，現在も進行中の「21世紀世宗プロジェクト」によるコーパスである．このプロジェクトは，1998年に国立国語研究院を中心として10年計画で始まったプロジェクトで，高麗大学，延世大学，蔚山大学，全州大学，ソウル大学などの協力によって構築を進めている．最終的には，10億語からなるコーパスの構築を目指しており，2000年5月時点までで，1億3,000万語のコーパスができ上がっている．ちなみにコーパスベースの辞書，「南北韓国総合国語大辞典」も8年間で100億ウォンという予算によって国立国語研究院で編纂されている．

また，同じ1998年に，ターミノロジーに関する情報収集，標準化と，理論研究，応用研究，ターミノロジーのための特定分野での大規模コーパス構築をミッションとしてKORTERM（Korea Terminology Research Center for Language and Knowledge Engineering）も設立されて活発に活動している．

著作権については，数年前までそれほど厳しい取り決めはなかったようで，書籍・新聞などから収集されたデータのコーパス化や，そのコーパスを利用して作成した辞典類の刊行などが比較的自由に行われていた．その結果，コーパスの潜在的需要，必要性の認識が広く社会に浸透し，同時に著作権に対する意識も大きく変化を見せた．現在，コーパス公開時にいかにして著作権問題をクリアするか，日本同様コーパス関係者，新聞・出版関係者の間で盛んに議論がなされているところである．

コーパスの応用分野

文科系分野におけるコーパスの利用

文科系分野の現状

国内の文科系分野でコーパス利用の可能性や，量データを利用した研究手法についての議論が盛になってきたのは1990年代も半ば近くなってからある．その実際の利用ということであればさらに下てここ数年間の話である．

コーパス利用に限っていえば，英語学など外国学において利用されている現状がある．ネットを介て豊富な英語データ，殊に海外のさまざまな分野コーパスやテキストアーカイブを利用できることら，条件的には英語学が他の文科系分野よりも一先んじている感があり，研究目的・教育目的での用が盛んになってきている．また，若干の補助記号必要であるけれども基本的にはアルファベットでの書法が確立されている，といった言語を対象とす語学分野においても，ネットを介しての研究目的教育目的での利用が増加しつつあるところである．

振り返って日本語についていえば，国内のあちちにボランティアによって成り立っているテキスアーカイブなどもあるが，積極的にこうしたデータ利用した研究はまだ稀なようである．理科系分野でそうした利用法が多く見られるが，新聞のCD-RC版をコーパスとして使った研究を目にすることがい．あるいは，CD-ROM化された国語辞典データ利用し，各見出し語に付与されているアクセント報などを利用した研究や，電子ブック版の文学作を利用するなどした小規模データによる語の研究あるが，本格的な利用はまだまだこれからである．

国語学分野におけるバックグラウンドと課題

国語を対象とする学問分野においては，古く江時代から書籍・文書などに対して「索引」を作成

ることが行われてきた．テキストクリティークから始めて，その言語資料中で用いられている主要な語や人名・地名，あるいはすべての語について資料中での所在を明らかにしていくのが「索引」の作成作業である．これは作成する者のライフワークとなるほどに時間のかかる作業であったが，大量のテキストがコーパスとして電子化され，語の単位で簡便に検索することが可能となれば，こうした伝統的背景を持つ国語学の分野でのコーパス利用は大いに高まる可能性がある．ただ，古い文献を扱う場合の障害となるのがJISで扱っていない文字・表記の問題である．国立国語研究所では，明治から昭和初期にかけて読まれた雑誌「太陽」を対象とした「太陽」コーパスの作成が今現在も進行中であるが，こうした文字・表記の問題に対処する方法として，包摂規準を設け，JIS第1・第2水準の文字の範囲でこれらの文字を包摂して処理する方法を試行している．しかし，包摂規準によっても補いきれない文字に対してどのように処理するのが最も適切であるか，まだ明確な方法論を持つまでには至っていない．JISで扱われていない漢字・文字が言語資料中に出現した場合の処理方法，処理規則の確立が国語学分野におけるコーパス利用の鍵となるだろう．

バランストコーパス構築の必要性

漢字を用いた人名・地名を除けば，この30年間の現代語を対象としたときに前述のような問題はほとんど考慮しなくてよい．しかし，だからといって新聞データや文学作品データのみではやはり偏りがあり，本格的な言語研究や辞典編集，日本語教育といった目的に供するには困難な点がある．英語データの研究・教育目的での盛んな利用を可能にしているのは，バランストコーパスをはじめとする多彩かつ豊富なデータの存在があると考える．コーパスに対する潜在的な需要は理科系よりもむしろ文科系の研究分野の方があるにもかかわらず，日本語データの利用がいまひとつ積極的ではない背景にはジャンルの偏りが少なからず影響していると考えられる．この状況を打開するためにも，さまざまなジャンルの言語資料を網羅した，大規模な日本語の「バランストコーパス」構築が待たれるところである．

実務におけるコーパスの利用 ——

研究や教育だけでなく実務においてもコーパスは2つの点で利用価値がある．1つは，日本語，外国語を問わず，文章を書く際のお手本としての利用である．もう1つは，用語の普及度を調査するための道具としてのコーパスである．

「コーパス」という用語が人口に膾炙（かいしゃ）するずっと以前から，過去に書かれたテキストを，これから書くテキストのお手本とすることは行われてきた．特に，外国語で学術論文を書くような場合，章立てや特有の言い回しなどは，参考書や一般の辞書を調べるより，過去に掲載された論文から抽出した方が，正確かつ迅速である．事実，化学の分野では，すでに1970年に，『化学英語の活用辞典』（化学同人）という形で発行されている（斎藤他，1998，p.233）．英語のコーパスを使えば「非母語話者にとっても，ハンディキャップなしに英語を研究できる」（前掲書，p.4）が，研究のみならず，実務として英語を書く場合でも同様であるといえよう．

お手本としてのコーパスからさらに発展したのが，過去のテキストをそのまま対訳で保存する「トランスレーション・メモリー」である．たとえば，取扱説明書を翻訳した場合，これを対訳で保存しておき，バージョンアップしたときには，変更部分だけを翻訳し直す．通常のコーパスの概念とはかなり離れるが，プロの翻訳者にとっては便利なツールとして普及している．

もう1つ，コーパスの実務での利用方法として，用語の普及度の調査が考えられる．そういった調査が必要となるのが，パソコンなどの取扱説明書の執筆とユーザインタフェースの設計である．

初心者向けの取扱説明書で，概念や操作方法を説明するとき，使おうとする用語がどの程度世の中に浸透しているかを考慮しなければならない．ところが，現時点での用語選択は，テクニカルライターの経験に頼るか，せいぜいが辞書の見出しを調べる程度である．しかも，辞書の見出しも，普及度という意味では必ずしも信頼がおけない．たとえば，部品や周辺機器がその主たる製品といっしょに箱に入っていることを「同梱」と呼ぶが，この語は『新明解国語辞典・第五版』（1998）には収録されていない．かといって，では「同梱」は使うべきでないかどうか判断に迷うところである．また，辞書に見出しとして載っているからといって，その語が一般に普及しているとも限らない．「ドライブ」という語は，『新明解』には「駆動装置」という意味で記載されているが，パソコンの初心者向けの取扱説明書では説明しないわけにはいかない．

同じことはユーザインタフェースについてもいえる．ユーザインタフェースというと，グラフィクスや音声認識などが目立つ分野だが，ユーザ操作の基本となるのは，画面上のテキストである．ソフトウェアのユーザビリティを向上させるには，画面に表示される各機能や部分の名称が分かりやすくなければならない．たとえば，「表」を意味する「テーブル」が果たしてどれほど一般的に使用されているか，などはイ

ンタフェースとして機能に名前を与える前に，調べるべきだろう．

しかし，お手本としてのコーパスが，かなり前から存在していたのに対し，コーパスを用語の普及度調査に使うという意識は，実務者には，まだほとんどない．これは，お手本の場合とは異なり，利用そのものを実務者が思い付かないためと，ある語の普及度を知りたいと思ってもそれを調べるためのコーパスが日本語には存在しないためだと思われる．

用語の普及度を調べるためには，コーパス自体が社会の動きに合わせて絶えずアップデートされている必要がある．このようなコーパスはモニタコーパスと呼ばれるが，新聞や一般雑誌を中心として日本語の現状を反映し，実務者が手軽に使用できる大規模なモニタコーパスの編纂が待たれる．

情報処理技術におけるコーパスの利用 ——

自然言語を処理対象とする技術分野は，前述した通り，非常に広範囲にわたっている．それらのすべてでコーパスが利用されているが，ここでは機械翻訳を例にとって，コーパスの活用方法の概略を述べる．

機械翻訳で最も基本となる言語知識は，2つの言語の間の語彙・表現の対応を記述した対訳辞書という形で保持されている．また，翻訳のプロセスは，原文の認定を行う「解析部」と，訳文の出力を行う「生成部」とに大別される．機械翻訳の品質を向上させるには，対訳辞書，解析部，生成部のすべての個所でコーパスから抽出した言語知識を活用する必要がある．

我々が日常的に触れる言葉の数は，日々増大している．技術の進展，社会生活の変化，諸外国との交流の活発化などがその原因である．機械翻訳システムを有効活用するには，最新の語彙とその訳語を迅速に辞書に追加する必要がある．そのためには，大量のコーパスから最新の対訳辞書エントリを抽出する技術が必要である．

また，日本語や英語などの文の構造の把握にもコーパスの利用技術が必要である．文の構造を認定するには，文法規則を用いるのが基本であるが，文法規則は完全に書き尽くすことができないため，現実には人間にとって1つの解釈しかしない場合にも，たくさんの解釈の可能性が生じてしまう．その中から最も妥当な解釈を選択するには，文法規則では書ききれない，実際のコーパスに基づいた傾向の利用が不可欠である．

さらに，1つの言葉には複数の意味があるから，その中から，その文で使われている意味を正しく選択する必要があるが，その際にもコーパスの利用が必要である．言葉の意味の問題は，規則化しづらい．そこで実際のコーパスから言葉の使われ方と意味の関係に関する傾向を抽出して，意味の選択に利用する必要がある．

このように，言葉を対象とした処理を実現するには，まず対象である言葉の分析が必要であり，そのためには，実際に言葉が使用された姿であるコーパスが必要不可欠である．

コーパスは高度情報化社会の社会インフラである

自然言語の研究基盤である大量のコーパスを利用可能とするために「言語資源共有機構（GSK）」の活動が開始された．将来的には，煩雑な契約手続きに煩わされることのないよう契約・配布業務をGSKが代行することを目指している．また，著作権などの権利関係の扱いを明確に規定した契約のもとにデータを利用することとし，不正使用や権利侵害を防止するのもGSKの目的とする役割である．

GSKには，対象を日本国内の言語資源に限定せず，将来的にはアジア地域に拡張することにより，欧州・アメリカ・アジアの3大コンソーシアムの一翼を担い，自然言語処理技術，言語研究の国際貢献にもつながることが期待される．

言語資源の流通は，音声・自然言語処理に関連する幅広い研究を促進し，それによって新たに音声・言語産業（Speech and Language Industry）が創出されることが期待されている．そこには，狭い意味の音声・自然言語処理の分野だけでなく，言葉に関係する広範囲の技術が包含されている．この意味で，大規模コーパスは，高度情報化社会を支える社会インフラであるということができる．

参考文献

1) (社) 日本電子工業振興協会: 言語資源共有機構 (GSK) 活動報告, 『自然言語処理システムに関する調査報告書』所収 (2000).
2) 斎藤俊雄, 中村純作, 赤野一郎 編: 英語コーパス言語学・基礎と実践, 研究社出版 (1998).

ホームページアドレス

1) 言語資源共有機構 (GSK): http://tanaka-www.cs.titech.ac.jp/gsk/
2) British National Corpus (BNC): http://info.ox.ac.uk/bnc/index.html
3) Bank of English (BoE): http://titania.cobuild.collins.co.uk/index.html
4) European Language Resource Association (ELRA): http://www.icp.grenet.fr/ELRA/home.html
5) International Computer Archive of Modern and Medieval English (ICAME): http://www.hd.uib.no/icame.html
6) Linguistic Data Consortium (LDC): http://www.ldc.upenn.edu
7) Korea Terminology Research Center for Language and Knowledge Engineering (KORTERM): http://korterm.org/

(平成12年5月23日受付)