世界の言語研究所（20）　特定非営利活動法人　言語資源協会（GSK）　（日本）

特定非営利活動法人 言語資源協会(GSK)（日本）

橋田 浩一（産業技術総合研究所・GSK副会長） 田中 穂積（中京大学・GSK会長）

1. 設立の目的と経緯

言語資源協会（GSK; http://www.gsk.or.jp/）は，平成15年6月に設立された特定非営利活動法人である。言語資源の流通を促進することにより，言語資源を必要とする音声・自然言語処理分野の学術・研究・産業の発展，さらには，言語学分野の研究の推進に貢献することを目的とし，大学や公的研究所や企業に所属する言語処理技術や言語学の研究者を中心とする委員会によって運営されている。

言語資源流通促進の仕組みは，米国ではLDC（Linguistic Data Consortium; http://www.ldc.upenn.edu/）[1]，欧州ではELRA（European Language Resources Association; http://www.elra.info）という組織において運用されている。これらの組織では，さまざまな機関で開発された言語資源を集積し，希望する者にそれらを配布する仲介業務等が行われている。これにより言語の研究者や開発者は，必要な言語資源を簡単な手続きで入手し利用することが可能になっている。ところが，日本国内にはこのような組織がこれまでに無かった。そのため，研究者は個別に言語資源を探しその使用について所有者と交渉するなど，研究以外の事務処理等に多くの労力と時間を割かざるを得ない状況にあった。そこで，そのコストを大幅に軽減するために言語資源協会が設立された。

2. 言語資源とは何か

音声・自然言語処理分野の学術・研究・産業の発展に欠かせない基礎データとしての音声・言語データおよび関連するソフトウェアツールを「言語資源」と位置づけている。特に最近の「コーパス（言語データ集）に基づく音声・自然言語処理」の潮流にみられるように，大規模な実データ・コーパスを利用した確率・統計的手法が成果をあげ，言語資源の必要性はますます高まっている。一方，言語研究においても，近年「コーパス言語学」と呼ばれる研究が目立つようになってきており，研究利用可能な言語データの整備を求める声が高まっている。しかしながら，一般に音声・言語データは，音声・自然言語処理や言語研究を行う機関とは業種を異にする，新聞社，出版社，テレビ局などで開発されたものが多く，また，本来そのような研究目的で開発されたものではない。よって，入手が非常に困難である，利用範囲が限定されている，利用にはかなりの加工が必要である，などといった問題が多いのが現状である。言語資源協会では，そういった問題に取り組み，これまでに新聞社，出版社，テレビ局などで開発されてきたようなものか

ら，新たに各所で開発されるものまでを含め，あらゆる「言語資源」を利用しやすい形で提供することを目指している。

以下に，現時点において「言語資源」として想定しているものと，「言語資源」を利用する研究の例を示す。研究の例は，これまでの限られた言語資源を活用しながら実際に行われてきているものでもあり，今後，さらなる言語資源の活用により発展が見込まれる。

■ 言語資源

○テキスト・音声・映像等のコーパス（言語データ集）
○レキシコン（辞書）
○ターミノロジー（分野別用語集）
○言語処理ソフトウェアツール等

■ 言語資源を利用する研究

(1) 言語資源に含まれる様々な言語現象を網羅的に調べ分析する言語学的研究

(2) 言語資源から様々な言語知識を獲得する技術の研究

○辞書
○文法
○概念間の関係，オントロジー
○翻訳用辞書

(3) 言語資源から統計的な情報を抽出して，それを音声・自然言語処理技術に生かそうとする研究（統計ベース／コーパスベースの音声・自然言語処理）

(4) 各種言語処理ツールを開発するための研究

○形態素・構文解析ツール 例：茶筌，JUMAN, KNP, MSLR
○音声認識ツール 例：Julius，Julian
○文生成ツール
○音声合成・分析ツール
○構文情報抽出ツール
○加工（品詞付き・構造付き）コーパス構築支援ツール
○用例検索ツール

(5) 自然言語処理応用システムを開発する研究

○テキスト検索，分類，要約
○機械翻訳

○音声認識
○対話

(6) 自然言語処理技術の評価用例文として言語資源を利用する研究

3. 言語資源の流通の効果

言語資源協会は，言語資源の保有者と利用者の双方にとって，次に示すような意義・メリットのある言語資源の流通の仕組み・サービスの提供を促進している。

■ 言語資源保有者にとって

○公開により，新しい用途，新たな需要も喚起され，従来想定していた以上の利用に供することができ，言語資源の迅速な改良が期待できる。
○公開に必要な加工を代行してもらえる。
○契約・配布業務を言語資源協会が代行することにより，煩雑な契約手続きの必要がなくなる。
○著作権や知的所有権等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が利用されることにより，不正使用や権利侵害が防止される。
○開発後の保守・管理が保証される。
○言語資源の価値が高まることで新たな言語資源の開発が促進される。

■ 言語資源利用者にとって

○研究開発に効果的な言語資源の発見が容易になる。
○価値ある言語資源を無償あるいは安価に利用できる。
○利用しやすいように加工された言語資源が入手できる。
○契約・配布業務を言語資源協会が代行することにより，資源の保有者と直接個別の交渉をすることなく，簡単な手続きで言語資源を利用することができる。
○開発後の保守・管理が保証されることにより，より質の高い資源が利用できる。
○価値ある言語資源によって，新たな研究開発が促進される。

4. 現在の活動

現在行っている主な活動は次のとおりである。

○言語資源の集積・配布およびそのための標準的規約の策定
○言語資源に関する調査・研究
○言語資源の標準化
○言語資源の解析・加工
○言語資源関連技術等の普及・啓発

○言語資源に関連する無体財産権の活用推進
○言語資源に関する国際連携

また，現時点で取り扱っているまたは取り扱う予定の言語資源は次のものを含む。(詳細はweb上に公開しているカタログの各説明書を参照されたい。)

〈コーパス〉

◆ 毎日新聞GDAコーパス2004
毎日新聞1994年版に含まれる記事のうち約3,000件に形態素，統語構造，語義等に関するGDAタグを付与し，人手で修正したもの。原テキストデータを含まないので，毎日新聞1994年版CD-ROMが別途必要。

◆ 講談社和英辞典コーパス2003
講談社和英辞典の各項目をXMLで構造化し，和英対訳の用例データに句単位の対応関係に関するGDAタグを人手で付与したもの。

◆ 電総研道案内対話音声コーパス1998
機械と人間との間の道案内についての音声対話をWizard of OZ法によって記録したもの。発話の番の交換・うなずき・割り込み・割り込みへの適切な対応などを分析できるように設計されている。40名の話者による197対話のデータを含み，全部で1300分以上に及ぶ。

◆ 岩波国語辞典コーパス2004
岩波国語辞典第5版の本体である約5万6千項目のデータにGDAタグ等のXMLタグを付与したもの。形態素，統語構造，岩波国語辞典自身に基づく語義などのタグを含む。

◆ EDRコーパス改良版2001
EDRコーパスのうち約6,000文を東工大で公開されている文法に基づき形態素と統語構造について機械的にタグ付けし，これを人手によって修正したもの。

◆ WSJコーパス詳細化版2003
Penn TreeBank Wall Street Journalコーパスのタグを GDA化し，照応・共参照に関するタグを人手で付与したもの。原テキストデータを含まないので，Penn Treebank Release 2のCD-ROMが別途必要。

◆ (財)CICCで開発したアジア諸国言語の電子化辞書。

〈辞書〉

◆ EDR共起辞書改良版2001
EDR日本語共起辞書のうち主要な用言に関する係り受け関係約40,000件について，関係子を人手で修正したもの。語義の記述が正しい概念識別子になっていないものについても修正してある。

〈ツール〉

◆ XML オーサリングツール2004

GDA等に基づくXMLタグをテキストデータに付与する作業を支援するソフトウェア。自然言語のデータへのアノテーションを容易にするため，畳み込んだエレメントの中身のテキストデータも表示する。GDA の DTD に基づく整式チェックや検索等の機能を含む。

5. 今後の予定

言語資源協会は，その活動を活発化させ本来の機能を発揮させるべく活動を進めている最中である。中でも言語資源のカタログの充実を重要課題として取り組んでいる。現在入手が困難になっている『計算機用日本語基本辞書 IPAL（動詞・形容詞・名詞)』(情報処理振興事業協会(現：情報処理推進機構)) の辞書データ及び PDF 化したマニュアルの公開を準備中である。

また，将来は，日本国内の言語資源に限定せず，アジア地域に活動を拡張することにより，音声・自然言語処理技術，言語研究に関する国際貢献を目指している。

最後に，言語資源協会は会員制度によって活動が支えられている。会費等の収入は，言語資源の保守，発掘，開発等に及ぶさまざまな活動に用いられる。言語に関連する研究，教育その他の事業に携わる人々のコミュニティを形成し発展させる場として学会を補う役割を果たすことが，言語資源協会には期待される。多くの研究者，研究組織の御支援，御協力をお願いする。

注

1　前号（19号）の本欄でも紹介。「言語資料コンソーシアム（アメリカ合衆国)」(黒橋禎夫)