言語資源の共有と再利用」
シンポジウム
1999年2月1日,
電子情報通信学会「言語とコミュニケーション
研究会」

言語資源コンソーシアム設立に向けて

田中 穂積
東京工業大学大学院情報理工学研究科

1999年2月1日

1 はじめに

多くの自然言語処理システムは，現実の文を対象としなくてはならない段階に至っている．現実の文には，言語学者ですら十分に検討したことのないさまざまな言語現象が存在する．たとえば話し言葉に関する言語現象は，言語学者によってもまだ十分研究されていない．多くの言語学者の自己内省に頼るトップダウン的なアプローチによる研究では網羅性に限界があり，まだ多くの洩れがあるのである．

最近，多量の音声データを収集し，そこに含まれる統計的な性質を抽出することにより，音声認識システムの開発に大きな進展が見られたことは良く知られている．一方，電子化された日本語の文書は，日本語ワープロの普及とメモリー価格の低下，ネットワークの広域化とともに，近年その量が急激に増大している．こうした背景から，音声の分野と同様に自然言語の分野でも，電子化された大量の例文（コーパス）を集め，それを用いた自然言語処理の研究が活発になってきた．このような自然言語処理の研究は，多くの言語学者のように仮説から出発するのではなく,具体的な文の分析から研究を始めるという意味で，ボトムアップ的な研究のアプローチをとっているといえる．その主なものを以下にあげる．

1. 言語資源に含まれるさまざまな言語現象を網羅的に調べ分析する研究，
2. 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，
3. 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．

電子化された大量の例文は，それに加工を施すことにより音声や自然言語処理にとってより一層有用な言語資源となる．日本語でいえば例文にわかち書きがしてあること，わかち書きした結果にさまざまな詳細さのレベルで統語的なタグ（品詞など）や意味コードが振られていること，文全体に正しい統語構造や依存構造が付与してあること，文章にディスコースの情報が付加してあることが望ましい．これらの中には自動的に付加可能な情報もあるが，統語構造，意味コード，ディスコース情報は，残念ながら自動的に付加可能であるとはいえない．手作業に頼らざるをえないのである．手作業は，加工すべき文の量が増えるにつれて大量の労力を必要とするので，個々の研究者が個別に作成するのでは自ずと限界がある．多くの研究者は，このような言語資源を大量に作成したいと思っていても，時間と労力の問題で思うにまかせない，というのが現状である．さらに，研究課題が地味であり予算獲得が難しいだけでなく，時間がかかり論文が書きにくいという問題もある．

こうした状況を打開するために，われわれは協力し合い，言語資源の共有化をはかる必要がある．共有化を通じて，言語資源の量の拡大も可能になる．わが国では言語資源のベースになる例文を集めようとすると，著作権や知的所有権の問題をクリアにしなければならない．この問題に対しても，個々の言語資源使用者が個別に対応していたのでは効率が悪い．言語資源の提供者，利用者をめぐる様々な問題を議論する場がわが国でも必要になってくる．

欧米では公的な資金をベースにした言語資源コンソーシアムがすでに存在し活動している．わが国と異なり，言語資源はこれからの情報技術のインフラストラクチャとして重要であるとの認識に立っているから

だろう．わが国でも，自然言語処理技術だけでなく，音声学・言語学のベースとなる言語資源インフラを充実させるための組織だった活動の場が必要である．我々は現在，日本電子工業振興協会の中に言語資源コンソーシアム（GSK）設立準備会を設けて議論を重ねてきた．準備会での議論の内容については第 3 章で説明する．なお，設立趣意書（案）は最後に付録として示してあるので一読して御批判をいただければ幸いに思う．

GSK 設立の別の目的を述べておきたい．量的にはそれほどでもなくても，個々の研究者の汗の結晶ともいえる言語資源がある．ところがそれが死蔵されていることがある．このような死蔵された言語資源を発掘することもGSK 設立の大きな目的である．わが国では，プロジェクトが終り予算がなくなると，せっかく構築した貴重な言語資源が散逸してしまう．これらもGSK できちんと管理したいと思っている．

各種言語資源を組み合わせて使う場合に，言語資源相互間に整合性がなければならない．たとえば辞書という言語資源のもつ品詞体系は，文法という言語資源で使う品詞体系と整合していなければならない．いくつかの辞書を統合し大きな辞書を作る場合にも整合性が必要になる．これは言語資源の構築に際して標準化が重要な課題になることを意味している．GSK はこれらの問題を議論する場を提供することになるだろう．

2　言語資源

これまでは，言語資源という用語を，電子化された（タグ付き）例文集という意味に使ってきた．しかしGSK が対象とする言語資源という用語はそれより広く解釈している．言語資源には，書き言葉だけでなく，話し言葉も含まれることはすでに述べた．音声による対話の研究では，話し言葉の言語資源が必要になる．話し言葉は書き言葉とは異なる性質をもっているので，話し言葉の研究が最近重要になってきている．その他に言語資源には，辞書や概念体系，種々の自然言語処理ツールも含まれる．これを以下にまとめて示す．

言語資源の型

1. コーパス
   (a) 音声言語（音声データ，音声を書き起こしたもの,...）
   (b) 文字言語（小説，新聞記事，論文,...）
      i. 単一言語（モノリンガルコーパス）
      ii. 多言語（マルチリンガルコーパス；パラレルコーパス）
2. 音韻辞書
3. 単語辞書
4. 文法体系
5. 意味・概念体系（シソーラス，オントロジー,...）
6. 百科辞典的知識（常識）
7. ソフトウェアツール（形態素・構文解析システム，音声認識システム，文生成システム，音声合成システム，音響分析システム，言語資源構築システム，統計情報抽出ソフト，視覚化ソフト，用例検索システム,...）

コーパスを，未加工と加工済みのものに分けることもできる．

1. コーパス
   (a) 未加工コーパス
      - 音声データ
      - プレインテキスト
   (b) 加工コーパス
      - KWIC

- 音素記号列
- 形態素・単語列（わかち書き済み）
- 品詞タグ・意味・談話コード付き
- 構造付き（統語・係受け関係など）

3 GSK設立の三段階（案）

GSKの設立には三段階を考えている．資金がなくても有志の協力を得て運営可能な体制からコンソーシアムを立ちあげ，その後資金の手当を考慮した本格的なGSKコンソーシアムの運営につなげたいと考えている．最終的にはGSKの活動をアジア諸国にも広げたい．以下に箇条書したものは，日本電子技術振興協会の自然言語処理技術委員会の中のGSK設置準備会で議論した試案である．

第一段階：

目的 - 無償の既存言語資源の発見と散逸防止
- 既存の（無償，有償の）言語資源のカタログリストの作成と配布

設置場所 - 日本電子工業振興協会

組織 (a) GSK諮問委員会
- 全体統括
- 第二，第三段階構想の検討
- 発起人名簿の検討
- 任意法人としての会則検討
- 財源（会費，課金方法など）の検討
- 言語資源設置場所の確保
- 海外組織との連携
- 言語資源提供者ヒアリングに移行
- ホームページ内容作成
- 公開シンポジウム企画

(b) GSK事務局
- ホームページ (http://www.jeida.or.jp/corpus) の管理
- カタログリスト（配布先アドレス・ポインタ付き）の作成
- 言語資源の登録呼びかけ
- 使用者からの提言・バグレポート収集
- ニュース発行

(c) 言語資源調査研究WG
- 言語資源登録フォーマット作成
- 既存の言語資源の調査発掘
- 言語資源開発プロジェクトの立案
- タグ付き・構造付きコーパスの形式標準化
- 言語資源自動獲得技術の調査研究
- 言語資源利用技術の調査研究
- 使用者からの提言・バグ報告の検討と分析

(d) 言語資源提供者 WG

- 有償言語資源の調査
- 利用に際しての条件・知的所有権問題

スケジュール １９９８年度末に準備終了，４月上旬に設立集会，活動期間１年

第二段階 :

目的 第一段階の目的の他に，

- GSK 設立趣意書に沿うコンソーシアム設立
- 既存言語資源の保守改良
- 言語資源の開発新プロジェクトの実施準備
- 言語資源の仲介業務（CD ー ROM 化とその配布など）

組織 (a) GSK 運営委員会

- 全体統括
- 第三段階構想の検討
- 資金・財務の検討
- アジア太平洋地域をカバーする GSK の設立準備

(b) 事務局

(c) 言語資源調査研究 WG

- 第一段階の項目の他に，新言語資源の開発
- 既存言語資源の保守改良作業

(d) 言語資源提供者 WG

- 第一段階の他に，利用条件の明確化，新たな法律問題への対処法

スケジュール １９９９年度末に設立

第三段階 :

目的 アジア太平洋地域をカバーする GSK の設立

4 言語資源の自動構築技術・標準化

第１章では，言語資源の構築には大量の人手と時間がかかることを指摘した．これは音声・自然言語処理技術が未成熟な技術であることによる．しかし，たとえば日本語の形態素解析の精度は近年向上が目覚しく，加工コーパスとして，全文検索法による KWIC の作成だけでなく，形態素タグつきコーパスの自動構築は可能になってきている．言語資源の自動構築技術は，知識獲得技術の一つであり，人工知能やデータマイニングの分野の魅力的な研究課題であるといえる．筆者が現在考えている言語資源の自動構築のスキームを図１に示す．図中の編み目を付けた部分は，フリーのソフトとして公開されている[1]．音声・自然言語処理技術者は，言語資源の利用者と供給者の共に役割を果たすことが期待されている．

繰り返し述べるが，人手であろうと自動的にであろうと，構築した言語資源を多くの人が共有して利用するためには，データの形式，タグセットの標準化が必要になる．これも GSK で真剣に検討すべき課題であろう．

[1] http://tanaka-www.cs.titech.ac.jp/pub/mslr/index.html

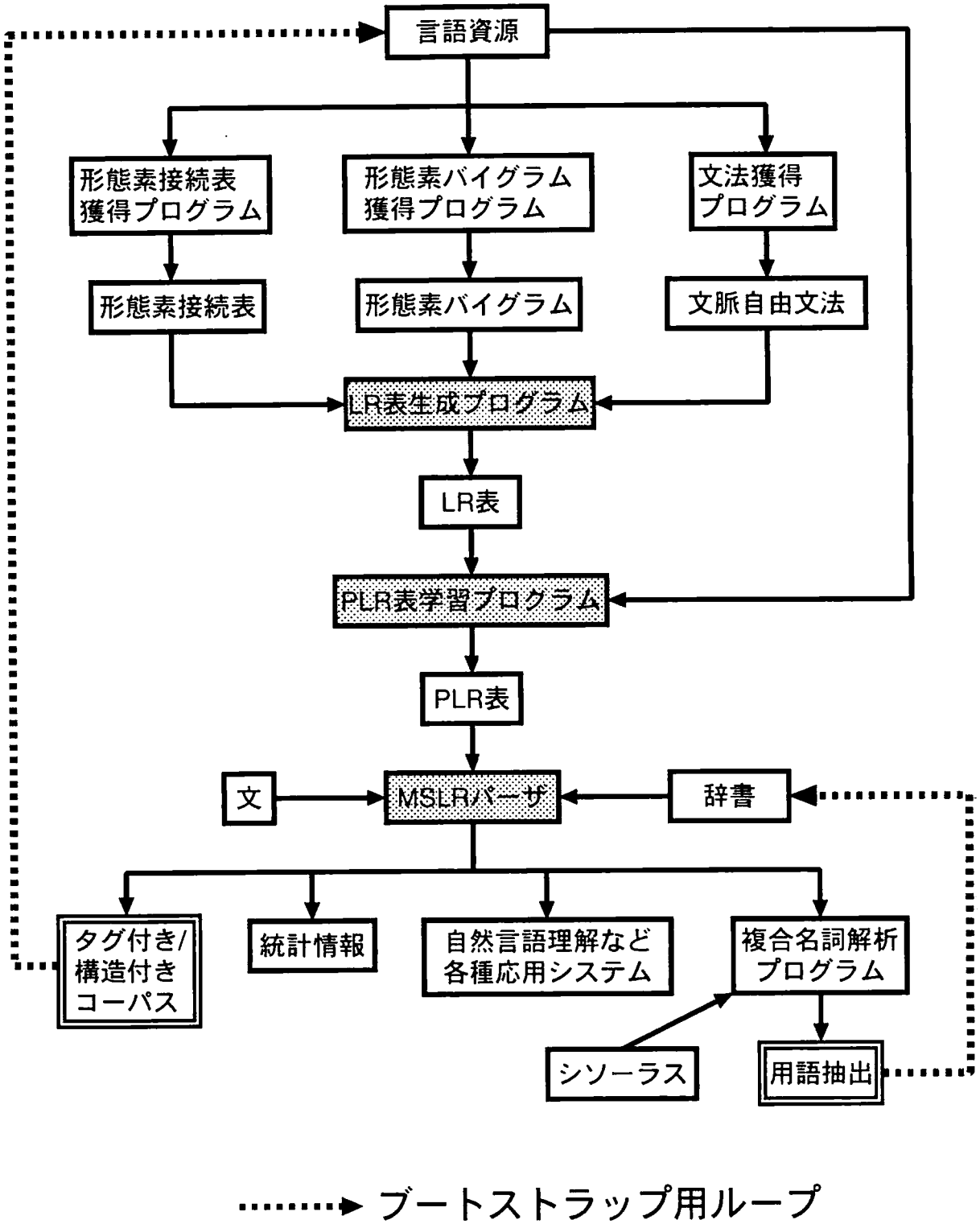

図 1: 言語資源の自動構築

5 海外の言語資源の例

海外では言語資源コンソーシアムが設立され，さまざまな言語資源が流通している．英国では辞書の作成用として始まった言語資源構築の長い歴史がある．参考のために，以下に海外の代表的な言語資源のURLを，亀井氏の文献から引用しておく[2]．

- Linguistic Data Consortium(LDC)
  http://www.ldc.upenn.edu/ldc/sites/index.html
- European Language Resource Association(ERLA)
  http://www.icp.grenet.fr/ERLA/home.html
- Corpus Linguistics
  http://www.ruf.rice.edu/~barlow/corpus.html
- British National Corpus
  http://info.ox.ac.uk:80/bnc/
- Oxford Text Archive
  http://firth.natcorp.ox.ac.uk/ota/public/index.shtml
- International Computer Archive of Modern Medieval English
  http://www.hd.uid.no/icame.html
- COBUILD
  http://titania.cobuild.collins.co.uk/

なお，国内の言語資源関連のURLとして，

- 奈良先端科学技術大学院大学 松本裕治研究室
  http://cactus.aist-nara.ac.jp/lab/resorce/resorce.html

がある．

6 おわりに

昨年の5月末から6月はじめにかけてスペインのグラナダで，言語資源に関する国際ワークショップ (First International Conference on Language Resource and Evaluation) が開催された．この会議は当初小規模の会議として企画されたと聞いていたが，実際には424編の論文発表があり，自然言語と音声の言語資源に関心がある全世界の研究者が一同に会した大きな会議として盛況の内に閉会した．このことは，言語資源に対する世界的な関心の高さを示しているといえよう．この会議の終りに，米国と欧州が共同で言語資源を構築する計画が米国のホワイトハウスから提案され，資金提供者（米国：NSF, 言語技術とその応用に関する欧州コミッション）と研究者の間で議論された．共同研究が開始されたと聞く．このように，情報技術のインフラストラクチャとして，国際的な規模での言語資源の構築が，欧米の言語を中心にして進められようとしている．アジアの諸言語を視野においたGSKを早急に設立する必要があろう．

最後に，言語資源は音声・自然言語処理技術の研究に役立つだけではない．言語を研究テーマとする学術，特に言語学，哲学，音声学の発展にも貢献することになろう．現代語を中心にした言語資源だけでなく，歴史的な多数の古文書を言語資源化しておけば，古文の研究にも役立つであろう．GSKがわが国で成功するかしないかは，音声・自然言語処理技術者だけでなく，人文系の学術研究者の協力も必要になる．

[2]亀井真一郎:自然言語処理技術とコーパス研究，電子工業月報，9，pp.19-23，1998.

7 謝辞

本稿をまとめるにあたり，GSK 設立準備会のメンバー，特に初期の段階でご尽力いただいた EDR の酒井専務，三吉氏（現シャープ），東大の辻井教授，筑波大の板橋教授，奈良先端大の松本教授，電総研の伊藤氏，通総研の井佐原氏，NHK の江原氏，富士通の森本氏，松井氏，NEC の亀井氏，東芝の木村氏，三菱の鈴木氏，松下の安川氏，電子協の樋口氏，宮川氏に感謝する．

付録：言語資源コンソーシアム設立趣意書（案)

1 言語資源コンソーシアム設立の趣旨

音声・自然言語処理の研究開発において，音声データ，レキシコン，テキストコーパス，ターミノロジー，各種ツール等の言語資源の重要性はいうまでもない．特に最近の「コーパスに基づく音声・自然言語処理」の潮流にみられるように，大規模な実データを対象とした確率・統計的手法が成果をあげている．

知識情報処理分野の基礎データとしての音声・言語データがコンピュータ産業の発展にとって重要であるにもかかわらず，大規模な音声・言語データの構築は膨大な労力・費用・年月を要するものであり，各個別の研究サイトにおいて開発するのは困難である．データの利用を希望する研究サイトは，やむなく他所で開発されたデータを利用せざるを得ないのが現状である．一般に大規模な言語データは，音声・自然言語処理の研究開発を行なう機関とは業種を異にする出版社や新聞社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．

そのため，言語データを利用したいユーザは，個別にデータ保有者と著作権交渉や価格交渉をすることを余儀なくされ，膨大な労力を必要とする．一方，データ保有者においても，従来想定していなかった利用形態であるため，データを提供することへの躊躇や戸惑いがみられる．またデータ提供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わが国の音声・自然言語処理研究の発展の著しい障害となっている．

従って，データ保有者，データ利用者の双方が納得できる形でのデータ提供，データ利用の仕組みを確立することは，言語資源の流通を促進し，ひいては，わが国の音声・自然言語処理の研究を促進し，言語産業 (Language Industry) の発展に貢献することになるため，そのような仕組みを確立することが急務である．それはまた，音声・自然言語処理の分野だけでなく，広く言語学の分野の研究の発展にも貢献することになる．

一方欧米ではそのような仕組みの必要性は早くから認識されており，米国では LDC（Lingustic Data Consortium)，欧州では ERLA(European Language Resources Association)，という共に公的支援を受けた会員制コンソーシアムが設立され，各所で開発された音声・言語資源を集積し，それらの利用を希望するユーザに配布するという仲介業務（データ保有者に代わって利用料を徴収し，一定のマージンを取る）を行なっている．これによりユーザは，簡単な手続きで必要な言語資源を入手し利用することが可能になっている．わが国においても LDC，ELRA のような，言語資源の集積・配布を行なう組織の確立が望まれる．

言語資源コンソーシアム構想は，以上のような背景に基づき，音声・自然言語処理の研究開発に不可欠な言語資源の流通を促進することにより，わが国のこの分野の学術・学問の研究の推進に貢献する公的機関を設立しようとするものである．また対象を日本国内のものに限定せず，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・アジアの三大コンソーシアムの一翼を担い，自然言語処理技術の国際貢献にもつながることが期待される．

2 言語資源コンソーシアムの意義

言語資源コンソーシアムは，データ保有者及び利用者双方に以下のようなメリットがあるので，言語資源の流通が促進される．

2.1 データ保有者にとって

- 従来想定していなかった新しい用途に供することにより，新たな需要を喚起し，利益にもつなげることができる
- 契約・配布業務をコンソーシアムが代行するので，煩雑な契約手続きに手を煩わせることがない
- 著作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ので，不正使用や権利侵害のおそれがない
- 死蔵データの有効活用をはかることができる

2.2 利用者にとって

- 契約・配布業務をコンソーシアムが代行するので，データ保有者と直接個別の交渉をすることなく，簡単な手続きでデータを利用することができる
- 会員になることにより，言語資源を安価にりようできる可能性がある

１９９９年４月？日

言語資源コンソーシアム 設立発起人（組織名五十音順・敬称略）