世界の言語研究所（19）　言語資料コンソーシアム（Linguistic Data Consortium : LDC）　（アメリカ合衆国）

言語資料コンソーシアム(Linguistic Data Consortium: LDC)（アメリカ合衆国）

黒橋　禎夫（京都大学）

1. 概略

言語資料コンソーシアム（Linguistic Data Consortium: LDC, http://www.ldc.upenn.edu/）は，フィラデルフィアのペンシルバニア大学内にあり，言語関連の教育，研究，技術開発の支援を目的として，言語データの作成や配布を行っている機関である。筆者は昨年10月にLDCを訪問する機会にめぐまれた。

LDCは1992年に設立され，当初は大学内の1室を使って数人で活動していたということであるが，現在はビルの1フロアを占め，40名をこえる常勤職員とパートタイム50名ほどの体制で運営されている。この10数年間でまさに倍々の成長をとげてきた。

常勤職員には言語学出身の研究者が多いが，それぞれがさらにプログラミングまたはマネージメントのスキルをもって運営にあたっている。施設内には，電話での音声収録のための端末室，アノテーション用の端末室，フィラデルフィア地域の放送アーカイブ用の大規模ストレージ，音声収録用の防音室，リラックスした環境で会話を収録するための応接室など多彩な設備が整えられていた。

2. LDCの主な業務

LDCの主要な仕事は次の3つである。

(1) コーパスの作成と収集
(2) 出版・配布
(3) 会員関連業務

この中で中心的な仕事はコーパス作成である。特に，TIPSTERプロジェクト[1]，TIDESプロジェクト[2]など，政府系プロジェクトのための言語データ作成の占める割合が大きい。実際，LDCの収入の大半がこの部分ということである。

コーパス作成については企業からの依頼もあるが，その場合には，たとえばデータ作成後1年間だけその企業がデータを独占利用し，その後は一般に公開するというような方法で契約を行っている。ただし，現在は政府系プロジェクトのコーパス作成で手一杯であり，企業からの依頼を実際に受けつけることは非常に少ないという。また，潜在的に存在する世界のコーパスを収集，管理していく仕事も視野に入ってはいるものの，残念ながら現在はほとんど手を出す余裕がないとのことであった。

2005年の1年間にリリースされたコーパスは38個で，音声コーパス12個，テキストコーパス25個，辞書1個であった。また，これまでリリースされたコーパスの出荷数の上位10位は次のとおりである。

最初の数字が出荷数，次がLDCカタログ番号（発行年，コーパスの種類［音声コーパス：S，テキストコーパス：T，辞書：L］，年内ID)。

761 LDC93S1 TIMIT Acoustic-Phonetic Continuous Speech Corpus
（マサチューセッツ工科大学(MIT)，SRI International, Texas Instruments, Inc. (TI)によって構築された，英語に関する連続音声データ）

598 LDC96L14 CELEX2
（オランダで開発された，英語，オランダ語，ドイツ語の音韻情報，形態素情報，頻度情報などの辞書）

350 LDC93S10 TIDIGITS
（Texas Instruments, Inc.（TI）による，音声認識学習用の英語数字列音声データ）

329 LDC94T5 ECI Multilingual Text
（European Corpus Initiative, the Multilingual Corpus 1（ECI/MC1）による，多言語テキストコーパス。収録言語は，アルバニア語，ブルガリア語，中国語，チェコ語，デンマーク語，オランダ語，英語，エストニア語，フランス語，ゲール語，ドイツ語，ギリシャ語，イタリア語，日本語，ラテン語，マレー語，ノルウェー語，ポルトガル語，ロシア語，セルビア語，スペイン語，スウェーデン語，チベット語，トルコ語，ウズベク語，リトアニア語。）

285 LDC93S2 NTIMIT
（米国の電話会社NYNEXの電話回線による，音声認識学習用の英語の音声データ）

239 LDC93T3A TIPSTER Complete
（TIPSTERプロジェクトによる，英語のテストコレクション）

233 LDC94S16 YOHO Speaker Verification
（話者認識のための英語の会議録データ）

225 LDC99T42 Treebank-3
（Penn Treebankプロジェクトによる，英語のテキストデータ）

216 LDC2000S85 Santa Barbara Corpus of Spoken American English Part-I
（米国内の，様々な出身地・年齢・職業の人々の実際の会話を録音した英語音声データ）

202　LDC2001T02　Message Understanding Conference (MUC) 7
（情報抽出のための英語の訓練データとテストデータ）

たとえば，この中の Treebank-3 は，Penn Treebank とよばれる，同じくペンシルバニア大学の Mitch Marcus 教授らが主導した90年代前半のプロジェクトで，「ウォールストリート・ジャーナル」および「ブラウン・コーパス」の40,000万文に，形態素情報と構文情報を付与したものである。このデータの第1版は LDC の初期のコーパスとして1993年にリリースされたものであり，機械学習による形態素解析や構文解析という，いわゆるコーパスベースの自然言語処理の端緒となったデータである。

3. LDC のデータの入手

LDC のデータを入手するには，会員になる方法と，個別にコーパスを購入する方法がある。会員の場合，アカデミック会員は年会費2,000ドル，一般会員は年会費20,000ドルである。会員はその年に出版された全コーパスを入手することができる。

アメリカ，特にペンシルバニア大学では，言語学，音声学，計算言語学の関係が非常に緊密であり，連携しながら研究を進めている。LDC はそのような土壌から生まれたものであろう。その活発な活動がまたそれぞれの分野の活性化をうながしており，LDC はまさに言語リソースに関する世界のセンターとなっている。

筆者らのグループでも最近，中国語の構文解析を試したいということがあったが，LDC から出版されている Chinese Treebank（300ドル）を購入するとともに，このコーパスの情報を基に動作する構文解析システム nlparser（研究目的であれば web から無料でダウンロードできる；http://www.cs.brown.edu/~ec/）を利用することができた。このような環境は自然言語処理研究を強力に後押ししており，これは音声研究，コーパス言語学研究などにおいても同様であろう。

なお，おそまきながら，日本においても LDC に相当する（ことを目指した）組織として「言語資源協会」（GSK; http://www.gsk.or.jp/）が設立され，昨年度から徐々に活動をはじめている。

注

1　TIPSTER プロジェクト（1991-1998）：DARPA（Defense Advanced Research Projects Agency:「高等研究計画局」の略。米国防総省の研究・開発部門）や，NIST（National Institute of Standards and Technology:「（米国）標準技術局」の略。連邦政府の機関で工業技術の標準化を支援している）などの主導で，文書検索，情報抽出文書要約などテキスト処理技術の向上を目指したプロジェクト。TREC（Text Retrieval Conference:1992-），MUC（Message Understanding Conference:1990-1998）などのシステム評価型会議を開催した。

2　TIDES（Translingual Information Detection, Extraction, and Summarization）プロジェクト（1999-2005）：TIPSTER プロジェクトの後継。多言語で提供される膨大な情報へ効率良くアクセスすることを目指し，TIPSTER の分野に加え機械翻訳の研究開発も積極的に行った。

付　記

著者らの訪問を快く引き受けて下さったDirectorのMark Liberman教授とExective DirectorのChristopher Cieri博士に御礼申し上げる。