明海大学大学院応用言語学研究科紀要
応用言語学研究 No.4 抜刷
(2002年3月)

英語学習者コーパス研究 (1):
コーパスの種類とデータ形式

投野 由紀夫

Graduate School of Applied Linguistics
Meikai University

英語学習者コーパス研究（1）：
コーパスの種類とデータ形式

投野 由紀夫

1. はじめに

「学習者コーパス」という用語はこの10年ほどで定着してきたものであるが、英語では(computerized) learner corpus （複数形は corpora） または learner's (learners') corpus という言い方をする。簡潔に言うと、学習者コーパスとは「ある言語を外国語として学ぶ学習者が話したり書いたりしたものを大量に収集・電子化し、コーパスとして整理したもの」である。

外国語学習者の output を記述・調査するという発想自体は別に目新しいものではない。いわゆる「よくやる間違い(common errors)」を収集した本は昔から存在してきた（e.g. George 1972)。しかし、これらの大部分の資料が教師の教授経験から来るものであったり、一部の作文などのエラーを整理したものなど極めて断片的な資料をもとに作られたものであったのに対し、学習者コーパスは大量の学習者データの整備と自然言語処理の技術によって科学的な言語習得データの記述を行うことが出来るという点で新しい可能性を秘めている。

1960年代後半から徐々に応用言語学の分野で第2言語学習者の発話・作文データを記録・分析するということが行われてきた。その主要な観点は、70年代前半までが中間言語（学習者の持っている目標言語に至る中間的文法体系）のエラー分析（cf. Dulay and Burt 1973; Burt 1975; Burt and Kiparsky 1972)、70年代後半から80年代前半が習得段階の記述研究（Clahsen 1980, 1984; Clahsen, Meisel and Pienemann 1983)、80年代以降が教室内での教師－生徒間のインタラクションやインプット、アウトプットの関係を記述する研究（cf. Chaudron 1988）などに発展していった。

しかし、これらの研究の大部分が、学習者データの処理に関して現在の学習者コーパス研究の観点から見ると不備が多かった。その最大の理由は、データの再利用、共有化という観点が欠けている点であった。現代の学習者コーパスが過去の学習者データと異なる最も大きな特徴は、学習者データをコンピューターが処理加工しやすい形で保存し、多角的な言語分析を行えるように整備している、という点にある。これは過去20年ほどのコーパス言語学や自然言語処理の分野の発展に負うところが大きい。

学習者データを大量に収集し、レベル別、技能別、抽出タスク別などのインデックスをつけて管理し、コーパス言語学、自然言語処理のノウハウを応用して統語構造や語彙使用についての使用頻度や例文検索、結びつきのパタン、誤用例等々の情報を容易に取り出せるようにデータの整備をしたもの。これが学習者コーパスである。

本論ではこの学習者コーパスの最新動向をレポートするが、特に主要なコーパス・プロジェクトの紹介 (2.)、コーパス・データのデータ形式 (3.) を中心に紹介したい。

2. 学習者コーパスの整備状況

2.1. 学習者コーパスの分類基準

世界の学習者コーパス・プロジェクトを整理する際に、以下のような分類評価基準を設けることが有益である。

2.1.1. コーパスの規模

学習者コーパス開発は歴史のあるものでもまだ10年程度と緒に付いたばかりで、一部の商用コーパスを除きそれほど大規模なものはまだ開発されていない。コーパス言語学的観点からあまり規模が小さいものは抽出する情報の一般性に欠けてしまう。学習者データの採取は非常に困難を伴う作業であり、一般の大型コーパスに比してサイズは小さくなり勝ちであるが、多くのプロジェクトが目指す目標として1単位グループのコーパスサイズが最低20万語、理想的な規模としては100万語程度を標準としている。

2.1.2. 被験者の学習段階

コーパス・データを採取する際に、学習者の学習段階の区分は重要な要素である。学年などで分類することも考えられるが、世界の他の学習者コーパスと比較する際には外部基準である学校歴では不十分である。その学習者の客観的な英語力の指標が別途参照できることが必要だ。現在の世界における学習者コーパスのプロジェクトでは、学習段階に関する表示があるものも、その妥当性に問題があるものが多い。そのためにICLE, HKUSTなどのプロジェクトでは特定の英語力レベルが統一された等質集団を用いてデータを採取するという方法をとっている。

2.1.3. 母語の相違

母語の背景が異なる英語学習者のデータを扱う場合は注意が必要だ。多くの第2言語習得研究の先行研究が母語の干渉の影響を指摘しており、母語の異なる被験者グループを一律同列に扱うことはできないので、良質な学習者データはサブグループとして母語に関する情報を詳細に明記する必要がある。研究テーマによって、母語が異なる学習者データは比較して興味深いデータが得られる場合もあれば、単一グループのみを観察しないと、きちんとした結果が導かれてこない場合もある。

2.1.4. タスクの種類

どのようなタスクによって発話（または作文）データを収集したかも重要な情報である。タスクが明解に記述されていれば、そのコーパスと同等のデータを第三者が比較データとして新たに作ることができ、そのコーパスに付加して規模を拡大したり比較参照したりすることが可能だ。このタスクの情報は、課題の内容、指示の仕方、作業時間、実際に使ったプリントなどさまざまな情報を含む物で詳細であればあるほどよいと言える。

2.1.5. デザイン

学習者コーパスには、分析を有効に行うために目標言語の本国人データを基準にする方法(IL vs TL)[1]と、異なるレベルの学習者データを比較する方法(IL vs IL)がある。それ以外に、母語のデータと比較する方法(IL vs L1)、3者を比較する方法(IL vs TL vs L1)などの可能性がある。それぞれの学習者コーパス・プロジェクトがどのようなコーパス比較の方法を念頭に置いて作成されたかも重要な分類の基準になる。

2.2. 世界の主要な学習者コーパス

特に注目を集めているいくつかの学習者コーパス・プロジェクトの概略を以下にまとめる。詳細は表1の主要コーパス・リストを参照されたい。

2.2.1. 海外の主要学習者コーパス

海外には例外的に早い時期から学習者データをコーパス化した例がいくつか見られる(例えばFærchのPIF Corpus, Perdueらのthe ESF Databaseなど)が、ここではそれらは扱わずに、比較的この10年ほどでコーパス言語学の影響を受けて構築が行われつつある学習者コーパスのプロジェクトを概観する。

■ International Corpus of Learner English (ICLE)

(URL: http://www.fltr.ucl.ac.be/FLTR /GERM/ETAN/CECL/abs.html)

1990年から始まったInternational Corpus of English (ICE)という世界18地域の英語変種コーパス構築プロジェクトの傘下で、英語の変種(variety)の1つとして英語学習者コーパスを作ろうというものである。現在世界の学習者コーパスの中でも、サンプリング方法の科学的なこと、規模と整備状況から見ても最も本格的なものだ。Louvan大学(ベルギー)のSylvianne Grangerが中心になり、多くの関連する研究発表を行っており(Granger 1998)、学習者コーパスの理論化という面でも世界をリードしているプロジェクトだと言えよう。

現在は15の異なる母語の上級学習者の論説文からなる自由英作文データを各々20万語ずつ集め、それに品詞タグ、エラータグを施す作業を進めており、予定が多少遅れているが、2002年春ごろを目処にICAMEよりデータが公開される予定。

■ HKUST Learner Corpus

香港科学技術大学のJohn Miltonが中心になって集めている中国人大学受験生、大学生の英作文コーパス。電子メールで作文課題を提出させたものと、大学入試の共通英作文問題のtranscriptionが中心で、1999年12月現在全体の規模が1000万語ほどあり、単一の学習者グループのコーパスとしては最大規模のものである。その一部分(約100万語)は品詞タグ付きで20万語がマニュアルによるエラータグの処理を施している。Miltonはこの学習者コーパスのエラーデータをもとに学習困難点をtutorial形式で自学自習用grammar tutorにコンコーダンサーを併用した、AUTOWORDと題するソフトを97年3月のアジア辞書

[1] これを中間言語(interlanguage: 以下IL)対目標言語(target language: 以下TL)の対立なので、IL vs TLと呼ぶ。以下同様。

学会で発表している。その後、コンコーダンサー機能を充実させて、困難語リストをもとに発見学習をさせるようなパッケージにまとめたWordPilotというシェアウェアも開発している（http://home.ust.hk/~autolang/）。

■ Longman Learners' Corpus (LLC)

Longmanが所有する商用の学習者コーパス。約1000万語の規模は世界最大級。80年代初期からデータを収集しており、LDOCEという英英辞典の第2版作成時に、Michael Rundellを中心に整備された。規模はデータの採取方法はかなり不統一で、文法問題や自社のcoursewareのモニター資料、Cambridge英検のessayデータなどとソースは雑多である。これらを世界70カ国以上から集めたデータベースをMichael Rundellがcorpus managerとして全体の統括をした。

ファイルはアルファベットごとの国別拡張子がついたテキスト・データで、ヘッダ部分には(1)データ入力日、(2)データ採取地、(3)被験者の母語、(4)言語グループ、(5)学習者の英語力レベル（8段階）、(6)タスク情報（15分類）などの情報がコード化されている。これをLongman社が所有しているCorpusbench（Textware社製）というOS/2版のコンコーダンサーを利用することでヘッダ情報を複雑に組み合わせ、統計を駆使した検索が可能になる[2]。

品詞・エラータグなどは元ファイルにはついていない。特に日本、中国、フランス、ドイツの学習者データは100万語規模で、タスクは非常に雑多であるが、比較的上級学習者の傾向をつかむには適しているといえよう。このデータをもとにしてLongman Dictionary of Common Errors (1997)やLongman Essential Activator (1996)などのコラムが執筆された。商用で本格的に出版物に学習者コーパスを利用しているのはロングマン社が最初である。簡単な解説であるが、web pageが参照可能。(http://www.longman-elt.com/dictionaries/corpus/lclearn.html)。

■ Cambridge Learners Corpus (CLC)

Longmanよりも遅く90年代になってから収集されたが、ケンブリッジ大学はUCLES (the University of Cambridge Local Examination Syndicate)というtesting serviceの機関を有しているので、その試験で使われた課題作文などを利用して、非常に短期間に1000万語規模のデータを収集し、現在もデータを逐次追加中。Cambridge International Dictionary of English (1995)という英英辞典作成のため、大規模にコーパス整備を行い、Longman社から経験ある編集者などを招いて、ノウハウを急速に身につけた。CLCは社内の辞書・語学教材編集専用のデータで一般には公開されていない。2000年末現在、250万語程度に人手によるエラータグ付与がされており、それらを元にした統計情報などのコンコーダンサーでの抽出も非常に精密に出来るようになっている。

2.2.2. 日本の主要学習者コーパス

■ Corpus of Japanese Learners of English（http://www.lb.u-tokai.ac.jp/lcorpus/）

[2] 辞書部編集長のDella Summers氏によると、ロングマンは現在、社内の辞書編集者用のコーパス利用環境を刷新しており、Windows仕様のweb-basedなコーパス・辞書編集統合環境を開発中という話である。

JACET'96で呼びかけがあり、ハイパーメディア研究会のメンバーが中心で発足した日本人英語学習者コーパス作成のプロジェクト。中・高・大と連携したデータの広範囲な採取、WWW上での共有化、エラータグ開発、音声コーパスへの取り組みなど全国規模で呼びかけようとした。実際にはまだ時期尚早であったのか、あまり期待したほど広範囲のデータ収集が行えず、プロジェクト・メンバー中心で収集が行われ、100万語規模に近いデータは集まったが、その主要なものは大学生のエッセイでかなりタスクのデザインなどにもばらつきがあった。

1997年度からは東海大学の朝尾幸次郎氏が代表となって科研のプロジェクトが3年計画で進行した。2000年に報告書が出版されている。日本初の学習者コーパス関連のメーリングリストもある（上記Web siteを参照)。データは公開の予定であるが、その後、整備のための十分な手当てがなされていない状況である。

■ JEFLL (Japanese EFL Learners) Corpus

東京学芸大学で10年間にわたって行われた科学研究費補助金研究による英作文プロジェクトで蓄積されたデータを母体とし、その後、投野をプロジェクトの中心として、中・高のボランティアの先生方を加え、さらにデータ増殖を大幅に行って現在の形態に至っている。中学2年から高校3年まで同一のトピックで自由英作文をさせたデータを電子化している。現在、ファイル管理、品詞タグ、エラータグなどに関して整備中で、現在の規模は学習者コーパス部分のみで約50万語（作文45万語、会話5万語)。特に、インプットの影響を調べるために中学・高校英語教科書コーパス、また母語の特徴分析のために同一トピックを日本語で書かせた作文コーパスおよび研究用の一般日本語コーパスを比較コーパスとして同時に整備しているのが特徴的。詳細は投野のweb page参照(http://leo.meikai.ac.jp/~tono/index.html)。

■ SST (Standard Speaking Test) Corpus

投野がアルク所有の英語インタビュー・テストの音声データをコーパス化することを提案し、独立行政法人通信総合研究所および認可法人通信・放送機構のプロジェクトの1部として進めている日本人学習者の英会話コーパス構築プロジェクト。Standard Speaking Testは米国のACTFL OPIという面接テストをもとにして日本人英語学習者向けに開発されたもので、15分間の面接により英語力を10段階で診断するテストである。その15分間の英会話中には大きく分けると3種類の発話誘因タスクが用意されており、可能な限り学習者の発話を多面的に引き出すことを目的としている。このコーパスはそのデータを100万語書き起こして話し言葉学習者コーパスとして公開することを目的としている。

通信・放送機構では、このコーパス作成を「適合型コミュニケーション技術」の研究の一貫として位置づけている。これは、機械が人間の言語を理解する際に、言い誤りや不完全な文などを文意を類推して判断できるような技術であり、英語学習者のエラーを含んだ発話はそれらの基礎資料として貴重だというわけである。このプロジェクトは現在3年計画の2年目で、すでに500人の発話約125時間分の書き起こしデータが完成しており、またそれに付随する書き起こしツール（TagEditor）やエラータグ専用ツールの開発なども同時に行われており、非常に意欲的なプロジェクトである。

表１：世界の主要な学習者コーパス・プロジェクト

プロジェクト	被験者/タスク コーパスサイズ	タグ付与 利用可能性	対象分析	参考文献
ヨーロッパ:				
International Corpus of Learner English (ICLE)	- 大学 EFL 3/4 年 - 15 ヶ国 - 英作文 - 300 万語	- エラータグ - 品詞タグ - 2001 年	- IL – IL (different L1s) - TL – IL	Granger (1993; 1994; 1996; 1998) Virtanen (1998a,b) de Haan (1997, 1998, 1999)
Longman Learners' Corpus (LLC)	- 全レベル - 英作文 - 1000 万語	- 品詞タグ - 商用利用可	- IL – IL	Gillard and Gadsby (1998)
Polish-English Language Corpus Research and Applications (PELCRA)	- 全レベル - 英作文/会話 - ポーランド人英語学習者	- 品詞タグ - 利用不可	- IL – IL (発達的データ) - L1 – IL - TL – IL	Uzar (1997) Mason & Uzar (2000) Leńko-Szymańska (2000a, b) Lewandowska-Tomaszczyk, Leńko-Szymańska, and McEnery (2000)
The ISLE Corpus of non-native spoken English	- 20 分スピーチ - ドイツ&イタリア英語学習者	- 書記 - 音韻・強勢 - ELRA より配布	- TL – IL	http://nats-www.informatik.uni-hamburg.de/~isle/speech.html
JPU (Janus Pannonius University) Corpus	- 大学 EFL - 英作文 - 約 40 万語	- タグなし - 近く公開予定	- IL – IL (発達的データ)	József (1999)
Cambridge Learners Corpus (CLC)	- 全レベル - 1000 万語 - 英作文	- 品詞タグ - エラータグ (2.5M) - 利用不可	- IL – IL	http://uk.cambridge.org/elt/reference/clc.htm
Indianapolis Business Learner Corpus (IBLC)	- 米大学ビジネス専攻学生 - ビジネス文書	- タグなし - 利用不可	- IL – IL (異なる母語比較)	Connor & Precht (1998)
アジア				
JEFLL Corpus (Japan)	- 全レベル EFL - 英作文/会話 - 50 万語	- 品詞タグ - エラータグ (部分的) - 公開予定	- IL – IL (発達的データ) - L1 – IL - TL – IL	Tono (1996; 1998; 2000a, b; 2002) Tono and Aoki (1998) Tono et al. (2001)
SST Corpus	- 10 レベル EFL - 英会話 - 100 万語	- 品詞タグ - エラータグ - 公開予定		

Corpus of English by Japanese Learners	- 全レベル EFL - 英作文 - 100 万語	- タグなし - エラータグ (部分的) - 公開予定	- IL – IL (発達的データ)	Asao (1998)
Japanese/ English Translation corpus	- 中高 EFL - 和文英訳データ	- タグなし - Web から取得可	- TL – IL	http://home.hiroshima-u.ac.jp/d052121/eigo1.html
TELEC Student Corpus	- 香港英語学習者 - 入試作文課題 - 300 万語	- タグなし - 共同研究のみ利用可	- TL – IL	Allan (1998)
PolyU Corpus	- 大学院生 - 論文草稿など - 28 万 2000 語	- タグなし	- TL – IL	Farmer and Mead (1998)
NTOU Corpus	- EFL - 53,000	- タグなし	- TL – IL - IL – IL	Chen (1998)
A parallel corpus of Japanese learners of English	- 英作文 - Native の添削文とのパラレルデータ	- データベース形式	- TL – IL - IL – L1	Mark (1998a, b)
MET Corpus	- 中国人英語学習者（中学生） - 英作文 - 15 万語	- タグなし	- TL – IL	Anping (1998)
HKUST Corpus of Learner English (HKUST)	- 中国人大学生 EFL - 1000 万語 - 英作文 & 入試エッセイ	- 品詞タグ (1M) - エラータグ (10 万語)	- IL – IL	Flowerdew (1996) Flowerdew (1997) Milton (1998; 2001) Milton and Tsang (1993) Milton and Chowdhury (1994) Milton and Freeman (1996) Milton and Hyland (1999)

3. コーパスの論理的構造

1.1. コーパス・フォーマットの種類

書き起こしは会話データ（または手書き作文データ）の精度の問題だったが、もう１つ忘れてならないのはコーパスデータをコンピューターに理解させる際の表記方法の問題だ。これを一般にテキストの論理的構造を付与するという意味で markup という。これに対して、テキストへの言語的情報を付加する部分を annotation という[3]。この部分だけでもコー

[3] Edwards (1995:20) 参照

パスの専門書が1冊書けてしまうほどであるが、ここではごく初歩的なフォーマットの方法を解説する。

コーパスデータを扱うと目的に応じていろいろな markup の形式があることがわかる。代表的なものに (1) Fixed Format References、(2) COCOA 形式、(3) CHAT 形式、(4) SGML/XML 形式がある。サンプルを見ながら大まかな特徴を解説する。

3.1.1. Fixed Format References

この形式はテキストの各行の左端にテキスト情報と行番号が来るという形式。テキストそのものの書誌情報は別ファイルに記されている。1960 年～ 70 年代にかけて Brown/LOB Corpus[4] やその他の電子テキストに広範に利用された。現在でも、日本の TXTANA, 海外の WordSmith などの主要コンコーダンサーにはこの形式に対応した設定を出来るようになっているものが多い。一時代前ではあるが、Brown/LOB という二大コーパスの標準書式だったので今でも影響力がある。しかしこの方式は、最近編纂されているコーパスではあまり使われていない (図 1 参照)。

```
R01 0010 It was among these that Hinkle identified a photograph of Barco!
R01 0020 For it seems that Barco, fancying himself a ladies' man (and why not,
R01 0030 after seven marriages?), had listed himself for Mormon Beard roles
R01 0040 at the instigation of his fourth murder victim who had said: "With
R01 0050 your beard, dear, you ought to be in movies"!   Mills
```

図 1 ： Fixed Format References (Brown Corpus からの引用)

3.1.2. COCOA format

80 年代に OUP から発売されていた Micro-OCP というコンコーダンサーはじめ広範囲に利用されていた形式。最初のアルファベットが変数名 (T = title; A = act; S = scene; C = character など) でそれに続いて変数の値が文字列で記入され、全体を <> で囲むというようになっている。これらの変数名は自分で任意につけることが出来、柔軟性が高かった。この頃からテキスト部分とヘッダ部分を区別して記述するようになった。ただし、後の SGML などと比べると、テキスト自体が構造化はされておらず、変数の定義ファイルなどもないので電子文書としては過渡的なフォーマットだったといえよう (図 2 参照)。

```
<T Merchant of Venice>
<A 1>
<S 1>
((Enter Antonio, Salerio, and Solanio))
<C Antonio>
In sooth, I know not why I am so sad.
It wearies me, you say it wearies you,
But how I caught it, found it, or came by it,
What stuff 'tis made of, whereof it is born,
I am to learn;
```

図 2 ： COCOA 形式 (Merchant of Venice の冒頭)

[4] 百万語のバランスを考えた米語・英語コーパス

3.1.3. CHAT format

幼児の言語習得のデータベースとして有名なCHILDESのフォーマット形式。現在、言語習得関係の標準フォーマットの1つ。サンプルを見ていただくとわかるが、開始は @Begin, 終了は @End でそれぞれ改行する。@Begin の直後に @Participants という変数行で発話中に出てくる speaker を定義して、発話部分（speaker tier）はそこで定義した3文字の speaker code に * をつけて発話の先頭に置くという決まりがある。記号と発話を区切るコロンの後は必ずタブ区切りにする。% で始まる tier はいろいろな注釈行になる。

```
@Begin
@Participants: CHI Adam Target_Child, MOT
Mother, URS Ursula_Bellugi, Investigator,
DIA Diaperman Adult
@ID: brown.ad.adam20.0300=CHI
@Sex of CHI: Male
@Age of CHI: 3;0.10
@Date: 15-JUL-1963
@Time Duration: 11:00-12:00
@Situation: Adam has a cough and runny nose
and had it for a week. Only Mrs Smith,
Ursula and baby Paul are present. Paul's
crib has been moved into the living room
with tape recorder
*CHI:  why dis got holes?
%act:  looking at holes in Ursula's pad
*URS:  so you can put it in a notebook # if
you like.
*CHI:  0.
%act:  falls from bike
*URS:  what happened?
*CHI:  I fall # broke my head.
%spa:  $RES
*URS:  you didn't.
*CHI:  tell me story.
*CHI:  tell me story.
*URS:  shall we look at these first?
%act:  gives Adam bag of toys
```

図3：CHAT形式で書かれたRoger BrownのAdam corpus

CHILDESはCHAT形式をサポートする大量のツール群（CLANという）を開発しているので、特に対話データの分析に適している。また談話分析などの標識も数多く用意されており、会話分析などの音声データとの融合も積極的に行われている。

ただし、テキストの構造化という点では、XML/SGMLのような柔軟性と複雑な構造の表記が出来ない。またテキスト中に言語的情報を付加する方法も dependent tier という注釈行によるところが大きいので品詞情報などは別ファイルでデータを二重に持つなどの工夫が必要になる（図3参照）。

3.1.4. SGML/XML format

世界的に電子テキストの標準となっているフォーマット。これはコーパスデータ専用のフォーマットというわけではなく、広く電子テキストを共有化するための一般的な規格である。大量の文書の構造化やそのチェックをするためのツール群がほとんど無償で公開されているので、書式に慣れてしまえば非常に柔軟性が高い電子テキストを構築できる。80年代はSGML (Standard Generalized Markup Language) が主流であったが、最近はWWWの技術が進歩して、SGMLよりもハイパーテキストを柔軟に扱えるXMLという書式の方が優勢になってきている。ちなみに1億語のイギリス英語コーパスであるBritish National CorpusはSGML形式であるし、現在主要辞書出版社の持っている電子データのかなりのものがSGML形式になっている。

3.2. XML による論理的構造の定義

ここでは最も新しい XML によるテキスト構造の定義の概略を説明する。XML は "eXtensive Markup Language" の略で、なぜこの形式を採用するかというと、(1) SGML のように構造化テキストを扱える、(2) HTML のようにハイパーテキストを扱える、(3) 将来的に web ベースのツール類やデータベースなどと連携する際に威力を発揮する、といった点が挙げられる。今å„最も利用される可能性の高いフォーマットだといえよう。現在、Internet Explorer の最新バージョン（5.5 以上）では XHTML という形式が採用されていて、これは XML でフォーマットされた内容を XSL と CSS という 2 種類のスタイルシートを通してブラウザに表示させる仕組みになっている。一般の電子テキストの世界でも基礎的な構造部分のフォーマットはどんどん XML になってきている。CES(Corpus Encoding Standard) という大規模コーパスの標準フォーマットを決めようというプロジェクトでも、昨年辺りから SGML から XML へとフォーマットを移行して、ハイパー文書的にコーパスを扱うという試みが研究され始めている[5]。

3.2.1. XML による学習者データの記述

それでは具体的な例として生徒の英作文データをフォーマットしてみよう。まず重要なことは、作文データとして管理したい内容を出来るだけ詳細なリストにすることである。その際に作文データの「テキスト外情報 (extra-textual information)」と「テキスト情報 (textual information)」を大きく分けて考えるとよい（図 4 参照)。テキスト外情報というのは、例えば「学校」「学年」「クラス」「氏名」「実施時期」「課題内容」「制限時間」「辞書使用」「添削」などなどのデータ管理上必要な情報である。これはどれだけデータを一般的に共有するかにもよるので、必要な項目を各自で選定すればいい。

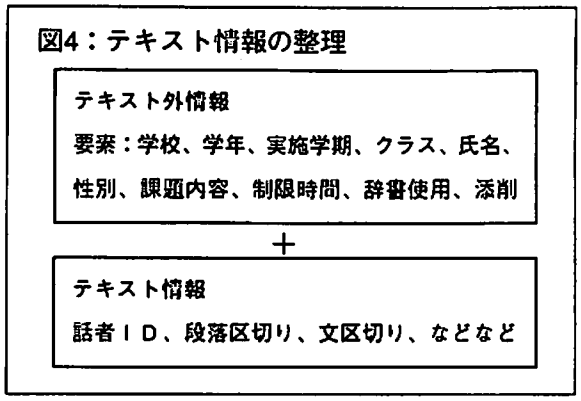


図4：テキスト情報の整理

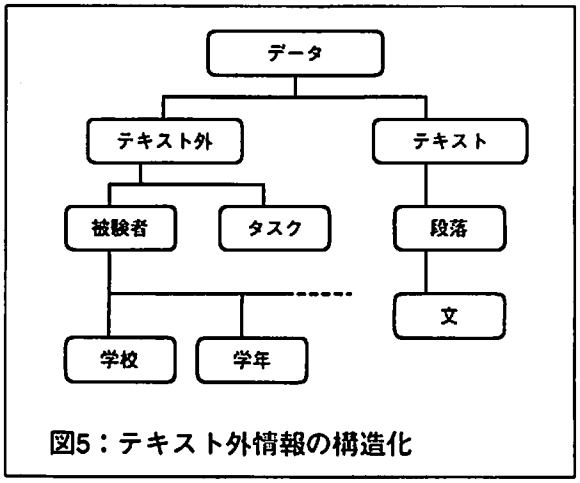


図5：テキスト外情報の構造化

テキスト情報は実際に書かれた作文や録音した発話データの中身に関する情報になる。発話データならば、話者 ID，作文ならば段落区切りなどの情報がまず考えられる。文区切りに関しても途中で切れたりして境界がはっきりしないものが多いので、あった方がよいだろう。また prosodic な情報や、error tag などの実際のデータに付ける annotation （注釈）部分については後述する。

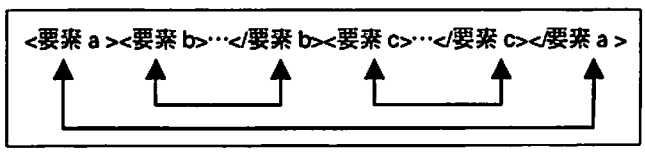
次にこれらの中で階層構造を持たせた方がいいものを特定する（図 5 参照)。例えば、テ

[5] CES の web page 参照(http://www.cs.vassar.edu/CES/)

キスト外情報とテキスト情報は並列の関係、テキスト外情報のうち「学校」「学年」「実施学期」「クラス」「氏名」「性別」に関するものは被験者情報、「課題内容」「制限時間」「辞書使用」「添削有無」などは課題情報として分類する。テキスト情報の方も話者 ID（作文の場合は不要）－段落－文といった階層を持たせる。次にこれを XML で記述してみる。

3.2.2. XML の記述方法

XML は通例、構造に関する定義を DTD (Document Type Definition) と呼ばれる部分に書き込んで参照し、それに対応する文書（XML instance という）をペアにして持つ。しかし、ここでは DTD の解説は紙数を要するので、DTD を使わずに XML instance のみを作成して利用しておこう。こうしておけば、今後、大量の XML ファイルを扱うようになった際に、DTD をしっかり作ってそれにそったチェックを再び行ったり出来るので、別に心配はない。DTD でチェックをしなくても XML の書き方を守っていれば well-formed XML instance ということができる。



まずは XML 文書だということを宣言する。これには <?xml version="1.0" ?> というおまじないの言葉を文書の先頭につける。「XML の version1.0 を用いなさい」という命令文だと思えばいい。次からが構造を示すタグ（記号）が来る。XML は必ず文書の要素はタグで表すことに決まっていて、内容（content）を開始タグ（<>）で始めて、終了タグ（</>）で終る。そこで、コーパスデータ全体は大きくは、<data> ... </data> というくくりを設けてみよう。その中が 2 つに別れる。<head> 部分と <text> 部分だ。そして <head> 部分に大別して <subject> 関連の要素（ここではそれぞれ <school>, <year>, <term>, <class>, <sex>, <name> としてみた）、<task> 関連の要素（<p>, <s> など）をそれぞれ記述する。ここで記述上の注意点を記しておく。

```
<?xml version="1.0" encoding="Shift_JIS" ?>
<data>
<head>
<subject>
<school>Lancaster High School</school>
<year>9</year>
<term>summer</term>
<class>4</class>
<sex>f</sex>
<name>Yuki Tono</name>
</subject>
<task>
<topic>breakfast</topic>
<time>20 min</time>
<dicuse>no</dicuse>
<revision>no</revision>
</task>
</head>
<text>
<s>I like rice better than bread. </s>
...
</text>
</data>
```

図 6 ： XML 形式で書かれた英作文データの例

1. タグの中の要素名は小文字にする
   これは XML の約束事として決まっている。
2. 必ず終了タグをつける
   HTML では終了タグを省略する形式が頻繁に用いられるが、XML では終了タグを必ずつけたほうがよい。
3. タグが交差しない（上図参照）
   タグが交差すると、ill-formed 文

書になってしまう。必ず、タグは入れ子になるように配置する。

以上の書き方に注意して、書き起こしデータをXML形式でフォーマットしたものが図6のサンプルである。データそのものはわかりやすくするために割愛してあるが、大きく<head>部分の記述と<text>中に書き起こしたデータを置くという形式を理解していただけると思う。

3.2.3. DTDの利用

学習者データの量が増えてくると、マニュアルでタグを付けていたものに関してはどうしてもタグの挿入ミスや文書構造違反などが出てくる。この際にXMLは威力を発揮する。XMLではこの文書内の構造チェックをvalidationといって特に重要視しており、そのためのツール類もたくさん出回っているので、これを利用可能だ。

XMLでは文書構造の定義はDTD（文書型定義）によって行う。DTDには単にタグ付けで使用する要素や属性が定義されるだけでなく、文書の1つ1つの部分の全体に対する関係などの「意味付け」が行われる。文書構造がしっかりしていれば、XML文書としての利用価値がそれだけ高くなる。

前節でごく基本的な学習者コーパスのファイルのフォーマットを紹介したが、これに対応するDTDは図7のようになる。

```
<!ELEMENT data (head, text)                                    >
<!ELEMENT head (subject, task)                                 >
<!ELEMENT subject (school, year, term?, class?
sex? name?)  >
<!ELEMENT task (topic, time, dicuse, revision)                 >
<!ELEMENT text (s*)                                            >
<!ELEMENT school (#PCDATA)                                     >
<!ELEMENT year (#PCDATA)                                       >
<!ELEMENT term (#PCDATA)                                       >
<!ELEMENT class (#PCDATA)                                      >
<!ELEMENT sex (#PCDATA                                         >
<!ELEMENT name (#PCDATA)                                       >
<!ELEMENT topic (#PCDATA)                                      >
<!ELEMENT time (#PCDATA)                                       >
<!ELEMENT dicuse (#PCDATA)                                     >
<!ELEMENT revision (#PCDATA)                                   >
<!ELEMENT s (#PCDATA)                                          >
```

図7 ： DTDの例

個々の変数の値を属性として細かく定義することも可能だが、ここではDTDのイメージが分るように非常に簡略にしてある。これを外部サブセットとして1つのDTDファイルとして保存して、XMLフォーマットで書かれた学習者コーパス・データから参照させるように指定することも出来るし、これを各ファイルの先頭に置いて、内部サブセットとして指定することも可能だ。

図8はこのDTDを用いて、

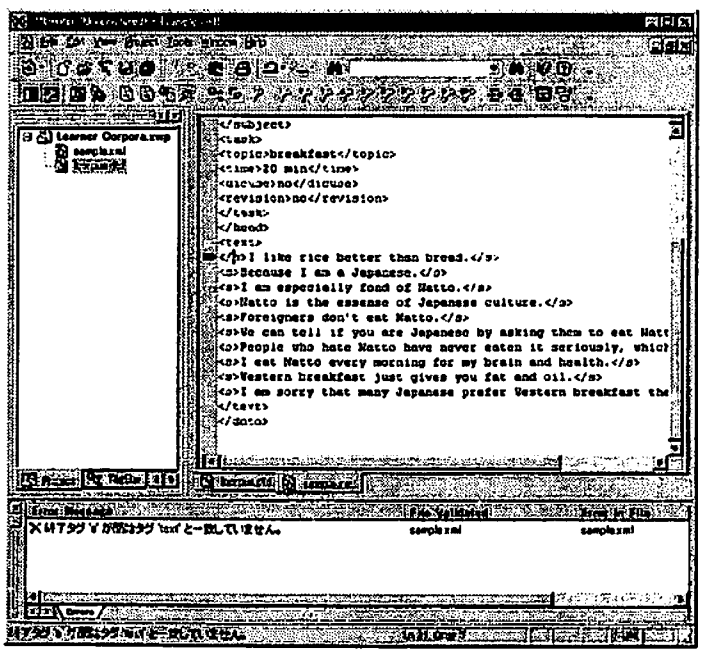


図8 Tag validationの様子

XMLwriterという商用のXMLエディタ兼パーザーで文書構造の検証をさせているところである。パーザーが構文チェックをして<text>中の、<s>の開始タグの誤りを指摘している。このようにDTDによってコーパスデータの持つさまざまな情報をきちんと定義してやることで、コーパスデータが「情報利用」という観点から意味を持ってくる。さらに、一般のXMLツールが自動的に大量のファイルの検証を行ってくれるので、特別に個別のプログラムなどを書かなくてもよい。こういった点で、XML文書の形式でコーパスをフォーマットするのは今後の方向性として非常に重要である。

4. おわりに

本論では学習者コーパス研究の概要のうち、コーパスの定義、世界の学習者コーパス構築の動向、そしてコーパスのフォーマットに関する解説を行った。以下の参考文献も、プロジェクト一覧表とともに参照されたい。学習者コーパスを語る際に、書き言葉データとしての自由英作文の収集、および話し言葉データとしての発話データの採取は非常に重要な項目である。ここでは紙数の関係で詳しい説明が出来ないが、別の機会にデータ収集方法を書き言葉・話し言葉の両面から詳しく論じてみたい。また、学習者コーパスの具体的な応用分野に関してもさまざまな研究が始まっている。先行研究の概要を別の機会にレポートしてみたい。

まだ新しいこの分野のより一層の発展のため、各分野から学習者コーパス研究への視点・要望をお寄せいただいて、日本人学習者の英語習得プロセスの客観的記述が組織的に可能になるように切に望んでいる。

参考文献

Allan, Q. G. (1998) The TELEC Student Copus: a recource for teacher development. In S. Granger and J. Hung (eds) (1998): 4−6.

Asao, K. (1998) Corpus of English by Japanese learners. In S. Granger and J. Hung (eds) (1998): 10−13.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *Longman Grammar of Spoken and Written English.* London: Longman.

Burt, M. (1975) Error Analysis in the adult EFL classroom. *TESOL Quarterly* 9 (1), 53−64.

Burt, M. and C. Kiparsky (1972) *The Gooficon: a Repair Manual for English.* Rowley, Mass.: Newbury House.

Chaudron, C. (1988) *Second Language Classroom: Research on Teaching and Learning.* Cambridge: Cambridge University Press.

Chen, H-J.H. (1998) Underuse, overuse and misuse in Taiwanese EFL learner corpus. In S. Granger and J. Hung (eds) (1998): 25−28.

Chi, Amy, K. Wong Pui-yui and M.W. Chau-ping (1994) Collocational problems amongst ESL learners: a corpus-based study. In Flowerdew, L. and A. K.K. Tong (eds.) *Entering Text.* Language Centre. The Hong Kong University of Science and Technology.

Clahsen, H. (1980) Psycholinguistic aspects of L2 acquisition. In S. Felix (ed.) *Second Language Development: Trends and Issues.* Tübingen: Gunter Narr.

Clahsen, H. (1984) The acquisition of German word order: a test case for cognitive approaches to L2 development. In Andersen, R.W. (ed.), *Second Languages: a Cross-linguistic Perspectives.* Rowley, MA: Newbury House: 219–42.

Clahsen, H., J. Meisel and M. Pienemann (1983) *Deutsch als Zweitsprache: der Spracherwerb ausländischer Arbeiter.* Tübingen: Gunter Narr.

Connor, U. and K. Precht (1998) Business English: learner data from Belgium and the U.S. In S. Granger and J. Hung (eds) (1998): 29–33.

David, M. (1998) The pedagogical implication of a learner corpus. In S. Granger and J. Hung (eds) (1998): 87–88.

Dulay, H. and M. Burt (1973) Should we teach children syntax? *Language Learning* 23: 37–53.

Edwards, Jane (1995) Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In Leech, Geoffrey N., et. al. (eds.), *Spoken English on Computer: transcription,mark-up and application,* London: Longman, pp.19–34.

Farmer, R. and K. Mead (1998) The language of citations: an analysis via computer learner corpus. In S. Granger and J. Hung (eds) (1998): 34–37.

Flowerdew, J. (1996) Concordancing in language learning. In M. Pennington (eds.) The Power of CALL. Houston, TX: Athelstan, 97–113.

Flowerdew, L. (1997) Interpersonal strategies: investigating interlanguage corpora. RELC Journal 28 (1): 72–88.

Flowerdew, L. (1998) Concordancing on an expert and learner corpus in ESP. CÆLL Journal 8 (3): 3–7.

French, F. (1949) *Common Errors in English.* London: Oxford University Press.

George, H.V. (1972) *Common Errors in Language Learning: Insights from English.* Rowley, Mass.: Newbury House.

Gillard, P. and A. Gadsby (1998) Using a learners' corpus in compiling ELT dictionaries. In Granger, S. (ed.) *Learner English on Computer.* London and New York: Addison Wesley Longman: 159–171.

Granger, S. (1993) The International Corpus of Learner English. In Aarts, J., P. de Haan and N. Oostdijk (eds.) *English Language Corpora: Design, Analysis and Exploitation.* 57–69. Amsterdam: Rodopi.

Granger, S. (1994) The learner corpus: a revolution in applied linguistics. *English Today* 39 (10/3): 25–9.

Granger, S. (1996) From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora. In Aijmer, K., B. Altenberg and M. Johansson (eds.) *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies, Lund* 4–5 March 1994. Lund: Lund University Press: 37–51.

Granger, S. (ed.) (1998a) *Learner English on Computer.* London: Addison Wesley Longman.

Granger, S. (1998b) A bird's eye view of computer learner corpus research. In S. Granger and J.

Hung (eds) (1998): 45–48.

Granger, S. (1998c) Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. Cowie (ed.) *Phraseology: theory, analysis and applications.* Oxford: Oxford University Press, 145–160.

Granger, S. (1999) Use of Tenses by Advanced EFL Learners: Evidence from an Error-tagged Computer Corpus. In Hasselgard, H. and S. Oksefjell (eds) *Out of Corpora - Studies in Honour of Stig Johansson.* Amsterdam: Rodopi, 191-202.

de Haan, P. (1997) An experiment in English learner data analysis. In Aarts, J., de Mönnink, I. and Wekker, H. (eds) *Studies in English Language and Teaching.* Amsterdam: Rodopi, 215–229.

de Haan, P. (1998) How native-like are advanced learners of English? In Renouf, A. (ed.) *Explorations in Corpus Linguistics.* Amsterdam: Rodopi, 55–65.

de Haan, P. (1999) English writing by Dutch-speaking students. In Hasselg_rd, H. and Oksefjell, S. (eds) *Out of Corpora.* Amsterdam: Rodopi, 203–212.

Housen, A. (1998) An analysis of grammatical form-function mapping in L2 data using the CHILDES system. In S. Granger and J. Hung (eds.) (1998): 59–62.

József, H. (1998) *Advanced Writing in English as a Foreign Language: A Corpus-based Study of Processes and Products.* Unpublished PhD dissertation. Janus Pannonius University, Pécs, Hungary.

Kaszubski, P. (1998) Enhancing a writing textbook: a national perspective. In Granger, S. (ed.) *Learner English on Computer.* London and New York: Addison Wesley Longman: 172–185.

Lam, P and J.Hung (1998) The use of multi-word verbs in advanced Chinese ESL learners. In S. Granger and J. Hung (eds) (1998): 80–82.

Leech, G. (1998) Learner corpora: what they are and what can be done with them. Preface to S. Granger (ed.) *Learner English on Computer.* xiv-xx. London and Harlow: Addison Wesley Longman.

Leńko-Szymańska, A. (2000a) Passive and active vocabulary knowledge in advanced learners of English. In Lewandowska-Tomaszczyk, B. and J.P. Melia (eds), 287–302.

Leńko-Szymańska, A. (2000b) How to trace the growth in learner's active vocabulary. A corpus-based study. Paper presented at the 4th International Conference on Teaching and Language Corpora. Graz, 19–23 July 2000.

Lewandowska-Tomaszczyk, B. and J.P. Melia (eds.) (2000) *PALC' 99: Practical Applications in Language Corpora.* Frankfurt: Peter Lang.

Lewandowska-Tomaszczyk, B., T. Leńko-Szymańska, and A. McEnery (2000) Lexical problem areas in the PELCRA learner corpus of English. In Lewandowska- Tomaszczyk, B. and J.P.Melia (2000): 303–312.

Mark, K. (1998a) A parallel learner corpus approach to English curriculum development at a Japanese university. In S. Granger and J. Hung (eds) (1998): 89–90.

Mark, K. (1998b) The Significance of Learner Corpus Data in Relation to the Problems of

Language Teaching. *Bulletin of General Education* 312: 77–90. Meiji University.

Mason, O. and R. Uzar (2000) NLP meets TEFL: Tracing the zero article. In Lewandowska-Tomaszczyk, B. and J.P. Melia (2000): 105–116.

Milton, J. (1998) WORDPILOT: enabling learners to navigate lexical universes. In S. Granger and J. Hung (eds) (1998): 97–98.

Milton, J. (2001) *Describing and overcoming environmental limitations on the interlanguage of Hong Kong Chinese learners of English: a computational and corpus-based methodology.* Unpublished PhD thesis. Lancaster University.

Milton, J. and E. Tsang (1993) A corpus-based study of logical connectors in EFL students' writing. In R. Pemberton & E. Tsang (eds.) *Studies in Lexis.* Language Centre, The Hong Kong University of Science and Technology, 215–246.

Milton, J. and N. Chowdhury (1994) Tagging the interlanguage of Chinese learners of English. In Flowerdew, L. and A. K. K. Tong (eds.) *Entering Text.* Language Centre, The Hong Kong University of Science and Technology, 127–143.

Milton, J. and R. Freeman (1996) Lexical variation in the writing of Chinese learners of English. In C.E. Percy, C.F. Meyer and I. Lancashire (eds.) *Synchronic Corpus Linguistics. Papers from the sixteenth International Conference on English Language Research on Computerized Corpora.* Amsterdam: Rodopi, 121–131.

Milton, J. and K. Hyland (1999) Assertions in students' academic essays: a comparison of L1 and L2 writers. In R. Berry, B. Asker, K. Hyland and M. Lam (eds.) *Language Analysis, Description and Pedagogy.* Hong Kong: HKUST, 147–161.

Perdue, C. (ed.) (1984) *Second Language Acquisition by Adult Immigrants.* A Field Manual. Rowley, Mass.: Newbury House.

Perdue, C. (ed.) (1993) *Adult language acquisition: cross-linguistic perspectives.* 2 vols. Cambridge: Cambridge University Press.

Tono, Y. (1996) Using learner corpora for L2 lexicography. *LEXIKOS* 6. Stellenbosch: Universiteit van Stellenbosch: 116–132.

Tono, Y. (1998) A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In *TALC (Teaching and Language Corpora) 98 — Conference Proceedings,* Keble College Oxford, 24–27 July 1998: 183–187.

Tono, Y. (2000a) A computer learner corpus-based analysis of the acquisition order of English grammatical morphemes. In Burnard, L. and T. McEnery (eds.) *Rethinking Language Pedagogy from a Corpus Perspective.* Frankfurt: Peter Lang. 123–132.

Tono, Y. (2000b) A corpus-based analysis of interlanguage development: Analysing POS tag sequences of EFL learner corpora. In Lewandowska-Tomaszczyk, B. and J.P. Melia (2000): 323–343.

Tono, Y. (2002) *The Role of Learner Corpora in SLA Research and Foreign Language Teaching: the Multiple Comparison Approach.* Unpublished Ph.D. Dissertation. Lancaster University.

Tono, Y. and M. Aoki (1998) Developing the optimal learning list of irregular verbs based on the native and learner corpora. In S. Granger and J. Hung (eds) (1998): 113–118.

Tono, Y., K. Kaneko, H. Isahara, T. Saiga, E. Izumi, M. Narita and E. Kaneko (2001) The Standard Speaking Test Corpus: a 1-million-word spoken learner corpus of Japanese learners of English and its implications for L2 lexicography. *ASIALEX 2001* Proceedings (The Second ASIALEX International Congress, August 8–10, 2001), pp. 257–262. Center for Linguistic Informatics Development, Yonsei University, Korea.

Turton, N.D. and J.B. Heaton (1997) *Longman Dictionary of Common Errors.* Harlow: Longman.

Uzar, R. (1997) Was PELE a linguist? In Lewandowska-Tomaszczyk, B. & P. J. Melia (eds.) *PALC '97 (Practical Applications in Language Corpora),* Lódz, Poland 10–14 April 1997.

Virtanen, T. (1998a) Direct questions in argumentative student writing. In S. Granger, S. (ed.) (1998): 94–118.

Virtanen, T. (1998b) Argumentative uses of the progressive in NS and NNS student compositions: notes on clause status and grounding. In S. Granger and J. Hung (eds) (1998): 119–120.