作文履歴をトレース可能な大規模こどもコーパスの構築

永田　亮† 河合　綾子†† 須田　幸次††† 掛川　淳一††† 森広浩一郎†††
† 甲南大学知能情報学部 †† 甲南大学理工学部 ††† 兵庫教育大学
E-mail: †rnagata@konan-u.ac.jp, ††ss564042@center.konan-u.ac.jp, †††{kakegawa,mori}@hyogo-u.ac.jp

1. はじめに

自然言語処理や言語学においてコーパスは重要な役割を果たすが，従来のコーパスは大人の文章を集めたものが中心で子供の文章を集めたコーパスは少ない．特に，著者らが知る限り，書き言葉を収集した大規模な子供のコーパスは存在しない．対象となる子供を集める困難さに加えて，文章の収集自体も難しいことがその理由と考えられる．更に，著作権も大きな問題である．対象者は未成年であり著作権の譲渡には，一人一人，親権者の承諾を得ることが必要になる．実際，現存する子供のコーパスは話し言葉を対象にしており，対象となる子供の数も少ない．特に，この問題は，日本語の子供コーパスで顕著である．例えば，Child Language Data Exchange System（CHILDES）[3], [4] の日本語サブコーパスである Hamasaki コーパス [2]，Ishii コーパス [3], [4]，Aki コーパス [7]，Ryo コーパス [6]，Tai コーパス [8]，Noji コーパス [3], [4] では，対象となる子供の数は 1 人である（表 1 に，従来の日本語の子供コーパスの概要を示す．英語コーパスについては，中條ら [1] に詳しい）．言語習得に関する研究や子供の文章を対象とした言語処理の研究には子供の言語データが重要となるが，一般性の高い知見や分析結果を得るためには，対象となる子供の数は可能な限り多いことが好ましい．

このような背景を受け，現在，我々は，子供の書き言葉を対象としたコーパス「こどもコーパス」の構築を進めている．小学生 100 人を対象として，2 年間，各児童週 1 回，言語データを収集することを目指している．現在までに，小学 5 年生 3 学級 86 人に対して，約 6ヵ月間，言語データを収集している（31548 形態素分の文章を収集）．表 1 に，「こどもコーパス」の概要を示す．

本研究では，コーパスを効率的に構築するために，図書ブログシステム [9] を利用する．須田ら [9] は， 同システムは図書という身近なものを話題にすることによって，児童は楽しみながら積極的に文章を書き込むと報告している．そのため，児童の言語データの効率的な収集が期待できる．また，ブログ上に蓄積されたデータは，既に電子化されているので児童の文章を電子化する必要がないというメリットもある．更に，ブログには，書き込みの日時やユーザ ID を管理する機能が備わっていることもメリットである．書き込みの日時とユーザ ID がわかれば，誰がいつどのような文章を書いたかを容易にトレースできる．加えて，同システムでは編集履歴も記録されるので，誰がいつどのように書き換えを行ったかという推敲履歴や書き直しの履歴もトレース可能となる．以上をまとめると，「こどもコーパス」には，(i) 書き言葉である，(ii) 対象とする児童の数が多い，(iii) トレース可能である，という 3 つの特徴があるといえる．

以下，2. で，コーパスの詳細について述べる．3. で，コーパス構築のためのガイドラインについて説明する．4. で，「こどもコーパス」の現状と課題について述べる．

2. こどもコーパス

2.1 言語データの収集方法

1. で述べたように，本研究では，児童の言語データの収集に図書ブログシステム [9] を利用する．以下，同システムを利用した言語データの収集方法を説明する．

まず，児童は自分の読みたい本を学校図書館で借りる．貸出の際に，本のバーコードを読み込むことで，ISBN が図書貸出管理システムに送られる．図書貸出管理システムは，図書ブログシステムへ，ISBN，本のタイトルなどの書誌情報を送る．図書ブログシステムは，送られて来た情報を基に，ブログのアイテムを自動的に作成する．本のタイトルがアイテムのタイトルとなる．

次に，児童は借りた本を読み，その本を推薦する「おすすめメッセージ」と呼ばれる文章をアイテム上に書き込む．この「おすすめメッセージ」が「こどもコーパス」の基本となる．必要があれば，児童は，書き込んだ「おすすめメッセージ」を修正し，再度，書き込むことができる．書き込みの度に，書き込み日時と書き込んだ内容がシステムに保存される．この書き込みと修正の繰り返しによりシステム上に言語データが蓄積される．基本的に週一回授業時間を設け，その中で書き込みをしてもらうこととしている．加えて，休み時間や放課後にも書き込みが行えるようシステムを開放している．なお，児童は，他の児童のブログの内容を検索，閲覧できる環境とした．

最後に，収集したデータを 3. で述べるガイドラインに従い整備し，コーパスとする．データ形式は XML 形式である．コーパスのサンプルを図 1 に示す．

2.2 収録情報

「こどもコーパス」は，児童の書いた文章以外に様々な情報を収録している．以下，収録情報と対応するタグについて

表 1： 従来コーパスと「こどもコーパス」の比較

コーパス	Hamasaki	Ishii	Aki	Ryo	Tai	Noji	こどもコーパス
話し/書き言葉	話し言葉	話し言葉	話し言葉	話し言葉	話し言葉	話し言葉	書き言葉
規模（形態素数）	約 24000	約 10 万	約 45000	約 23000	約 93000	約 15 万	約 31548*
人数	1 人	1 人	1 人	1 人	1 人	1 人	86 人†
年齢	2〜3 歳	2〜5 歳	1〜5	1〜4 歳	1〜3 歳	0〜7 歳	5〜6 年生
期間	1 年	3 年	約 2 年	約 2 年	約 2 年	7 年	約 6ヵ月‡
収集間隔（平均）	月 2，3 回	月 2，3 回	週 1 回	週 1 回	週 1 回	月 24 回	週 1 回
収集方法	会話の録音	会話の録音	会話の録音	会話の録音	会話の録音	会話の録音	ブログ

*最終的には，約 10 万形態素を目指している．

†100 人を予定．

‡2 年間を予定．

```
<BLG>
   <BLG_NAME>LIBLOG</BLG_NAME>
   <USR>
      <USR_ID>2030023</USR_ID>
      <USR_GRADE>5</USR_GRADE>
      <ITEM>
         <ITEM_ID>100</ITEM_ID>
         <TITLE>西遊記</TITLE>
         <AUTHOR>呉承恩</AUTHOR>
         <ISBN>9784001145496</ISBN>
         <NDC>913.6</NDC>
         <EDTN>
            <EDTN_NO>1</EDTN_NO>
            <DATE>2008-03-03-10:30:00</DATE>
            <TEXT>難しい本です。
            </TEXT>
         </EDTN>
         <EDTN>
            <EDTN_NO>2</EDTN_NO>
            <DATE>2008-03-03-10:37:00</DATE>
            <TEXT>難しい本です。
                  でも面白いよ。
                  石から生まれた悟空が大活躍するよ。
            </TEXT>
         </EDTN>
      </ITEM>
   </USR>
</BLG>
```

図 1： こどもコーパスのサンプル

詳細に説明する（具体例は，図 1 を参照のこと）．

【 ユーザタグ：<USR>】 児童 1 人分の言語データを表すタグである．この中に,「おすすめメッセージ」を含む全ての情報が含まれる．ユーザ情報としては，各ユーザを識別するユーザ ID（<USR_ID>）とブログシステム開始時の学年情報（<USR_GRADE>）が含まれる．

【アイテムタグ：<ITEM>】ブログの 1 アイテムに対応するデータである．したがって，児童が登録したアイテム数と同じ数のアイテムタグ含まれることになる．アイテムには，アイテムを識別するアイテム ID（<ITEM_ID>），本のタイトル（<TITLE>），著者（<AUTHOR>），ISBN（<ISBN>），十進分類（<NDC>），書き込み履歴（<EDTN>）が含まれる．

【書き込み履歴タグ：<EDTN>】「おすすめメッセージ」の書き込み履歴である．<EDIT_NO> タグは，何番目に書き込まれた（編集された）「おすすめメッセージ」かを表す．また，<DATE> タグは，書き込み（編集）日時を表す．この二つのタグ情報から，作文履歴をトレースすることが可能となる.「おすすめメッセージ」本文は，<TEXT> タグに含まれる．一文一行に，文分割した形式とした．

以上が「こどもコーパス」に収録されている情報である．「こどもコーパス」は，児童の言語データだけでなく，関連する様々な情報を提供することがわかる．これらの情報により，多様な分析が可能となる．例えば，書き込み時間と編集履歴から，児童はどのように文章の推敲や修正を行うかということが分析できる．また，本のタイトルや十進分類の情報が得られるので，読んだ本のジャンルが児童の語彙の使用に及ぼす影響の分析などにも利用できる．

3. コーパス構築のためのガイドライン

収集した言語データを整備してコーパスとするためには，整備のためのガイドラインが必要となる．本節では，我々がこれまでに策定したガイドラインについて説明する．

【基本方針】 基本方針として，収集した言語データ（「おすすめメッセージ」）は，可能な限りそのままの形でコーパスに収録することとした．そのため,「こどもコーパス」には，意味不明な文字列が含まれる場合もある（例：“jhshsxsainvtquoicab”）．例外として，個人名に対する処理，文分割処理，文字の処理がある．

【個人名の処理】 個人名の処理とは，ある個人が特定される名前などが「おすすめメッセージ」に含まれていた場合，個人が特定されない別の文字列（例：<NAME> 人名 </NAME> など）に置き換える処理のことである．固有表現タガーなどを利用した半自動の処理も検討したが，対象文章が児童の文章ということを考慮し，全て人手で作業することとした．

【文分割処理】 文分割処理とは,「おすすめメッセージ」中の文を同定し，一文一行形式に調える処理のことである．

文末に適宜改行を挿入するだけでなく，文の途中に含まれる余分な改行を削除することも含む．基本的には，文末記号（“。”，“？”，“！”など）で改行することとする．ただし，様々な例外処理をガイドラインとして策定した．例えば，文末記号の直後に顔文字（例：“犬を飼いたい人にとてもおすすめです。（＾0＾）”）がある場合，顔文字の終了までを一文とする．また，引用符内や括弧内に文末記号がある場合は改行しない．現在のところ，このような文分割に関するガイドラインが 22 項目存在する．

【文字の処理】 文字の処理とは，「おすすめメッセージ」中の “>” や”&” などの文字をエスケープする処理のことである．これは，「こどもコーパス」が XML 形式を採択しているためである．具体的には，XML でエスケープする必要がある全ての文字に対してエスケープを行う．

4. こどもコーパスの現状と課題

2008 年 6 月 9 日より，小学校 5 年生 3 クラス 86 人を対象として言語データの収集を開始した．現時点(注 1)で，861 の「おすすめメッセージ」（31548 形態素分の文章）を収集できている．そのうち，460 が編集履歴であった．

以上のデータから，1 つの「おすすめメッセージ」は，平均 36.6 形態素から成ることがわかる（形態素のカウントには茶筌 [5] を利用した）．また，1 人の児童は平均約 10 回の書き込みを行っていることがわかる．収集期間は，約 6ヵ月間であるが，8 月が夏休みであることを考慮すると，実質 5ヵ月間（約 20 週間）の収集期間となる．したがって，児童は，平均で 2 週間に一回のペースで書き込みを行っていることがわかる(注 2)．また，児童は，約半分の「おすすめメッセージ」について，何らかの編集を行っていることもわかる（ただし，編集時に「おすすめメッセージ」に何の修正も加えず，そのまま登録したものも含む）．

このように，「こどもコーパス」には既に多くの言語データが収録されており様々な応用が期待されるが，使用の際には注意しなければならない点がいくつかある．以下，この点について議論する．

第一に，データの偏りが挙げられる．「おすすめメッセージ」は，本の推薦文であるため，内容は本に関するものに偏っている．実際，“本”（頻度 65，58 位）や “話”（頻度 56，63 位）など本に関する単語が多く出現する傾向にある．また，推薦文であるため勧誘表現が多いことも予想される．そのため，「こどもコーパス」から得られた語句の頻度と他のコーパスから得られた語句の頻度とを単純に比較することは意味を持たない場合があるということに注意しなければならない．

第二に，入力方法の問題がある．児童たちは，キーボードもしくはソフトウエアキーボードを用いて，「おすすめメッセージ」を入力する．漢字入力は，コンピュータの漢字変換機能を利用する．したがって，児童たちは，自分では書けない漢字を「おすすめメッセージ」に使用している可能性が高い．このことは，「こどもコーパス」を利用して，漢字の習得に関する分析などを行う際には注意が必要なことを意味する．

最後に，ブログを利用して収集された言語データであることにも注意しなければならない．ブログ上の文章であるため，紙と鉛筆で書く通常の作文とは，語用や文体が異なる可能性がある．このことも，「こどもコーパス」を利用して何らかの分析を行う際に，念頭においておく必要がある．

「こどもコーパス」には，上述の 3 点の注意しなければならない点があるものの，様々な応用可能性があると著者らは考えている．例えば，年齢と語彙数の関係を推定する重要な資料になると考えられる．また，児童間の語彙の伝搬に関する知見も得られるのではないかと期待している．児童は，他の児童のブログを検索，閲覧できる環境で，各自の「おすすめメッセージ」を書き込む．したがって，他の児童のブログから影響を受けることは容易に推測できる．例えば，他の児童の「おすすめメッセージ」中の単語や表現を利用して，自分の「おすすめメッセージ」を作成することなどが予想される．「こどもコーパス」には，書き込みおよび編集履歴が記録されているため，ある程度，語彙の伝搬の情報を得ることができる．

現在，より詳細な情報として，検索と閲覧に関する情報もコーパスに収録することを検討している．図書ブログシステムには，児童が，どのようなキーワードで検索を行い，検索された「おすすめメッセージ」のうちどれを閲覧したかという情報も記録している．将来的には，この検索に関する情報もコーパスに含めたいと考えている．更に，形態素情報をコーパスに付与することも計画している．そのためには，児童が書いた誤りを含む文章に対応できるよう，既存の形態素に関するガイドラインを拡張する必要がある．形態素情報が付与されたコーパスがあれば，子供の書いた文章専用の形態素解析が開発できる．子供の書いた文章専用の形態素解析は，更に詳細な，子供の文章の分析に繋がると期待できる．

5. おわりに

本稿では，我々が構築を進めている「こどもコーパス」について述べた．現在のところ，小学 5 年生 86 人 31548 形態素分の文章を収集している．今後は，言語データの収集を続けると共に，検索に関する情報の収録や形態素情報の付与などを検討していく予定である．また，近い将来，「こどもコーパス」を教育研究目的に限り公開する予定である．

（注 1）：2008 年 12 月 22 日現在．

（注 2）：基本的に，週一回の授業で収集を行っているが，自然学校や音楽会などで授業がなくなることもあり，平均すると週一回を下まわるペースとなっている．

謝　　辞

言語データの収集にあたり，多大な協力をいただいた神戸市立南落合小学校の米満芳人校長先生と諸先生方に感謝致します．著作権について情報提供と助言をいただいた甲南大学フロンティア研究推進機構のスタッフの方々に感謝致します．また，本研究に対して様々なアイデアをいただいた（株）ホンダ・リサーチ・インスティチュート・ジャパンの船越孝太郎氏に感謝致します．本研究の一部は，（株）ホンダ・リサーチ・インスティチュート・ジャパンからの助成金により実施した．

参 考 文 献

[1] 中條清美，内山将夫，中村隆宏，山崎淳史，“子供話し言葉コーパスの特徴抽出に関する研究，” 日本大学生産工学部研究報告 B, no.39, pp.65–78, 2006.

[2] N. Hamasaki, “The timing shift of two-year-olds' responses to caretakers' yes/no questions,” Studies in Language Sciences (2), pp.193–206, 2002.

[3] B. MacWhinney, The Childes Project: Tools for Analyzing Talk, Volume I: Transcription format and Programs, Lawrence Erlbaum, 2000.

[4] B. MacWhinney, The Childes Project: Tools for Analyzing Talk, Volume II: The Database, Lawrence Erlbaum, 2000.

[5] 松本裕治, “形態素解析システム「茶筌」,” 情報処理, vol.41, no.11, pp.1208–1214, 2000.

[6] S. Miyata, “「パパワ？」—子どもの「ワ」を含む質問について—,” 愛知淑徳短期大学研究紀要, vol.31, pp.151–155, 1992.

[7] S. Miyata, “アキ・コーパス —日本語を獲得する男児の 1 歳 5ヵ月から 3 歳までの縦断観察による発話データ集—,” 愛知淑徳短期大学研究紀要, vol.34, pp.183–191, 1995.

[8] S. Miyata, “The Tai corpus: Longitudinal speech data of a Japanese boy aged 1;5.20–3;1,” Bulletin of Shukutoku Junior College, vol.39, pp.77–85, 2000.

[9] 須田幸次，永田亮，掛川淳一，森広浩一郎, “児童が共同構築するブログにおける検索が情報発信能力に及ぼす効果,” 日本教育工学会研究報告集, pp.11–16, 2007.