こどもコーパスマニュアル Ver. 0.0.2

永田 亮
甲南大学知能情報学部

平成 22 年 5 月 2 日

1　概要

こどもコーパスは，児童が書いた図書を推薦する文章から成るコーパスである．教育研究活動に限り自由に利用可能である（ただし，こどもコーパスを利用して得た成果を発表する際には，文献 [1] を引用し出典を明記すること）．表 1 にこどもコーパスの概要と特徴を示す．こどもコーパスの詳細については文献 [1] を参照されたい．

表 1:「こどもコーパス」の概要と特徴

形体	書き言葉
人数	66 人
文書数	1068
年齢	10〜11 歳
期間	8ヵ月
平均収集間隔	約週 1 回
収集方法	ブログ
その他の特徴	トレース可能

2　収録ファイル

こどもコーパスには次のファイルが収録されている：

- kk_corpus.xml: こどもコーパス本体
- guideline.pdf: コーパス構築に関するガイドライン
- manual.pdf: マニュアル（本ファイル）

3　データ形式

こどもコーパスのデータ形式は XML 形式である．こどもコーパスに収録されている情報と対応する XML タグについて説明する（詳細は，文献 [1] を参照のこと.）．

【 ユーザタグ：<USR>】子供 1 人分の言語データを表すタグである．子供の文章を含む全ての情報がこのタグの間に含まれる．子供に関する情報としては，各子供を識別するユーザ 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，子供が登録したアイテム数と同じ数のアイテムタグが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>），タイトル（<TITLE>），図書の著者（<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>】文章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）文章かを表す．また，<DATE> タグは，書き込み（編集）日時（秒まで）を表す．この二つのタグ情報から，いつ何を書き込んだかがわかる．子供の文章自体は，<TEXT> タグに含まれる．一文一行に，文分割した形式とした．

参考文献

[1] 永田 亮，河合綾子，須田幸次，掛川淳一，森広浩一郎, “作文履歴をトレース可能な子供コーパスの構築,” 自然言語処理, Vol.17, No.2, pp.51–65, Apr. 2010.

Acknowledgments

言語データの収集にあたり，多大な協力をいただいた小学校関係者の皆様に感謝いたします．本研究に対して貴重な助言をいただいた（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いたします．著作権に関する情報を提供していただいた甲南大学フロンティア推進機構のスタッフの方々に感謝いたします．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．