理科教科書を例にしたOpenCHJ構築の流れ

髙橋 雄太（明治大学／国立国語研究所）

OpenCHJワークショップ
2024年3月10日(金)

- OpenCHJのモデルケースとして明治大学大学院田中牧郎研究室で作成した、明治初期理科教科書3点のコーパスについて説明する。

- CHJの「明治・大正編」収録のコーパスと同じ形式で構築したコーパス
- 明治初期の理科教科書『物理階梯』『小学化学書』『初学人身窮理』の3点（約10万語）
- 今回は「中納言」（CHJ）に組み込むことを前提としたコーパス構築について解説

明治大学側で作業

①電子テキストの入力（業者外注）

②テキストの変換、解析用テキスト作成

③XMLタグの付与

④形態素解析、「大納言」へインポート

⑤形態論情報の整備

⑥コーパス情報、書誌情報の整備

⑦各種情報の統合、納品

⑧「中納言」での公開

⑨解説書の執筆、公開

国語研に依頼

コーパス構築のフロー

【作業内容】電子テキストを用意する

- 専門業者に入力外注する
- OCRで書籍から読み込む
- 再配布可能なオープンリソースの電子テキストを利用する

【理科教科書の場合】

- 『日本教科書大系 近代編』のテキストをもとに専門業者にテキスト入力外注
- 14～15万字（概算）
- 旧字体テキスト入力は1字2.0円～（+税）
- 外注期間：約3ヵ月
- 作業者：1名

①電子テキストの入力（入力外注）

【作業内容】解析用にテキストを加工する

- 解析に適したテキストに加工する
- サンプル（記事、章節項…）単位にファイル分割

【理科教科書の場合】

- 漢字片仮名交じりテキストを漢字平仮名交じりテキストに変換
- 濁点落ちのテキストに濁点付与ツール「AYTC」（岡2012）を利用
- 原本（和本）の表記との照合・修正作業（同音異字の修正等）
- 課ごとにデータを分割してXML形式のファイルを作成
- 作業時期：約4ヵ月　作業者：4名

②テキストの変換/解析用テキスト作成 》

【作業内容】XMLタグを付与する

- 文境界タグ、ルビタグなど
- 「中納言」検索に必須のタグと任意付与のタグがある

【理科教科書の場合】

- 必須タグ：article（サンプル）、s（文境界）
- 任意タグ（中納言表示アリ）：quotation（引用）、ruby（右ルビ）、pb（ページ番号）、odoriji（踊り字）、vMark（濁点落ち）等
- 任意タグ（中納言表示ナシ）：warigaki（割書き）等
- 作業期間：約4ヵ月　作業者：1名

【作業内容】形態素解析とDBインポート

- XMLデータのセットを小木曽智信氏に共有し、形態素解析を依頼
- 解析前にXMLエディタなどでwell-formed（タグのエラーチェック）の確認することが必須
- 中村壮範氏にDB（「大納言」）インポートを依頼

【理科教科書の場合】

- 作業期間：数日
- 作業者：2名（小木曽智信氏、中村壮範氏）

④形態素解析/DBインポート》

【作業内容】「大納言」での形態論情報の修正

- 誤解析の修正
- 新規要素のUniDicへの登録

【理科教科書の場合】

- 基本的に1名で整備し、もう1名がダブルチェックとして確認作業を行った
- **1時間あたり1000短単位前後**のスピードで整備
- 未知語の登録・処理が最も困難（「延展」「寒慄」「越素」「カヒール（キャピタル）」）
- 作業期間：約5ヵ月　作業者：整備1名　＋　確認1名

⑤形態論情報の整備 》

【作業内容】「中納言」の表示に必要な情報整備

- 形態論情報以外の検索結果の列の表示項目
  - コーパス情報（サンプルID等）
  - 書誌情報（底本、章名、成立年、出版社など）
  - 著者情報（著者名、生年、性別、NDLリンク）
  - 画像リンク（URLリスト）
- EXCELベースでまとめて中村丈範氏に提出

【理科教科書の場合】

- CHJ「明治・大正編」と同様の情報を実装
- 作業期間：約1ヵ月　作業者：1名

⑥コーパス情報/書誌情報の整備》

【作業内容】コーパスの納品・公開

- ⑤の形態論情報と⑥の各種情報の統合→納品データの作成・納品（中村氏）
- 「中納言」での公開
- 公開に際してかかる費用は通時プロジェクト持ち

【理科教科書コーパス】

- 作業期間：2ヶ月程度（1月から3月の間に随時進行）
- 作業者：1名（中村氏）＋ 各種相談（小木曽氏）

⑦各種情報の統合/納品
⑧「中納言」での公開 ❯

【作業内容】コーパスの仕様書（解説文書）の執筆

- 収録資料の簡易解説（資料的価値など）
- テキストの作成（②、③）の方針の説明
- 形態論情報（⑤）、各種情報（⑥）の説明

【理科教科書の場合】

- CHJのWebページから公開
- https://clrd.ninjal.ac.jp/chj/doc/abstract-ScienceTextbook-202303.pdf
- 作業期間：2週間程度
- 作業者：執筆1名、確認・修正1名

⑨解説書の執筆/公開 ›

- **作業者1名**（髙橋）が**週5時間程度**で継続的に作業をした場合、**10万語規模**の形態論情報付きコーパスについて、**2年**で**入力～公開の行程を完遂**できる概算

作業工程	作業期間（概算）
①電子テキストの入力（外注）	3ヵ月程度
②解析用テキスト作成	4ヵ月程度
③XMLタグの付与	4ヵ月程度
④形態素解析/DBインポート	数日～2週間
⑤形態論情報の整備	5ヵ月程度
⑥各種情報の整備	1ヵ月程度
⑦各種情報の統合/納品	1ヵ月半程度
⑧「中納言」での公開	半月程度
⑨解説書の執筆/公開	2週間程度

実質「中の人」が作成
- 作業レクチャー不要
- 作業スピードが速い
- 国語研との連絡・連携がスムーズ

外部ではじめて構築する場合、同規模同基準であっても、同じペースで構築できるとは限らない

構築フローのまとめ》

- 理科教科書は「完全形」で公開したが、いくつかの行程をオミットすることができる

作業工程	理科教科書	簡易化の例
②解析用テキスト作成	原本準拠で テキスト校正	全集テキストなどをそのまま採用
③XMLタグの付与	必須タグ+任意タグ	必須タグのみ付与 ruby、quotationなどを省略
⑤形態論情報の整備	全編コア （人手修正済）	非コアデータ（全編or一部） 数十万語以上規模の場合推奨
⑥各種情報の整備	全情報付与 画像リンク実装	画像リンクを実装しない 一部情報（出版社等）をオミット
⑨解説書の執筆/公開	執筆・公開済み	省略

コーパス構築の簡素化》

- 岡照晃（2012）「近代文語論説文を対象とした濁点の自動付与アプリケーション」NLP若手の会 第7回シンポジウムhttps://www2.ninjal.ac.jp/past-events/2009_2021/event/specialists/project-meeting/files/JCLWorkshop_no2_papers/JCLWorkshop2012_2_37.pdf
- 髙橋雄太・田中牧郎（2023）「『日本語歴史コーパス 明治・大正編Ⅱ教科書』明治初期理科教科書コーパス（短単位データVer.1.0）概説書」https://clrd.ninjal.ac.jp/chj/doc/abstract-ScienceTextbook-202303.pdf
- 田中牧郎・髙橋雄太（2023）「明治初期理科教科書コーパスの構築と活用―『物理階梯』『小学化学書』『初学人身窮理』を対象として―」『国際日本学研究』第15巻1号

参考文献

本発表は、国立国語研究所共同研究プロジェクト「開かれた共同構築環境による通時コーパスの拡張」（プロジェクトリーダー：小木曽智信）による成果の一部です。

付記