Lexicography: Automating Parallel Alignment for PRC Press Conferences

CMSC435 Software Engineering

Team Members: Noah Cauchi, Tianhui Hou, Vincent Huang, John Ng, Jack Spoleti

A. James Clark School of Engineering

Context

Echtralex provides consulting services relating to linguistics such as parallel text compilation, dictionary services, and lexicography. For parallel text compilation, Echtralex publishes an online archive of parallel translations of People's Republic of China (PRC) government press releases, presenting a side-by-side **Chinese** and **English** transcript so readers can compare the two versions sentence by sentence.

Problem Definition

The current process is manual, and inefficient: Chinese speaking staff copy transcripts from ministry websites and paste them into Microsoft Word, where Chinese and English sentences are manually matched one row at a time. Manually aligning a single press release takes **20-35 minutes**; with releases from multiple ministries, the process can **exceed an hour**.

Our tool addresses the problem with the following features:

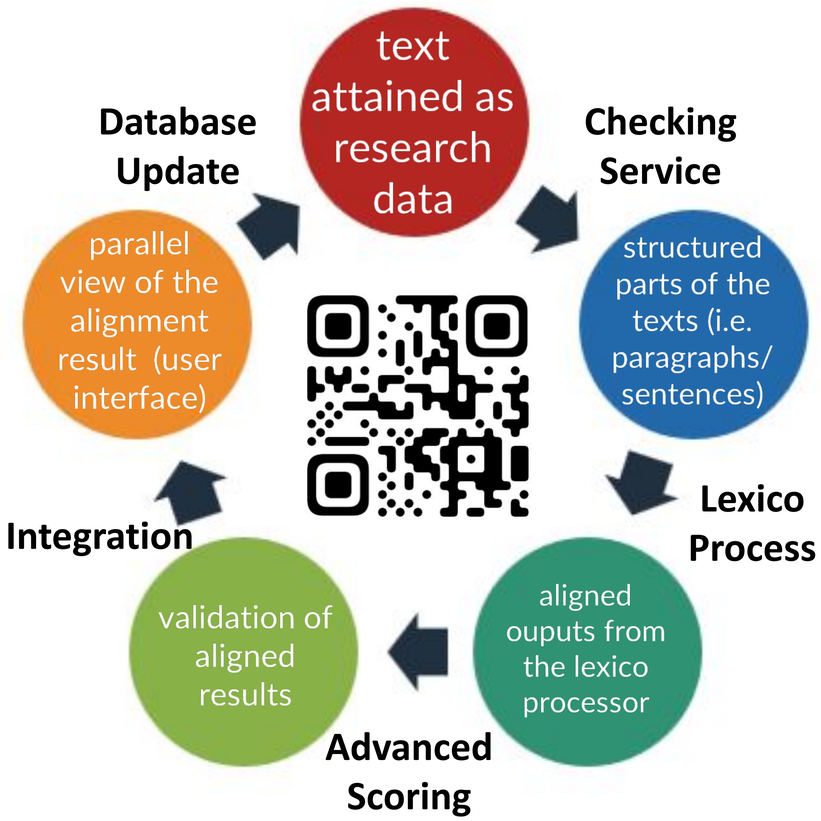
- ❏ Detect new press releases automatically
- ❏ Extract transcripts from ministry web pages
- ❏ Align Chinese and English translations at the sentence level
- ❏ Generate the final HTML page for publishing
- ❏ Ensure the system scales

请问中方对此有何评论?	What's China's comment?
【郭嘉昆】：我们昨天已阐明中方的严正立场。	Guo Jiakun: We've made China's serious position clear yesterday.
日方涉靖国神社一系列消极动向公然挑衅国际正义，粗暴践踏人类良知。	Japan's negative moves related to the Yasukuni war shrine constitute a blatant provocation against international justice and an affront to human conscience.
中方对上述恶行十分愤慨、严厉谴责。	China is strongly indignant and severely condemns it.

Alignment Example

Final Design: Rethinking Text

We treat text as a structured object rather than unstructured content. By modeling bilingual transcripts as analyzable data, we support automated extraction, sentence-level alignment, and transformation into publishable formats. This perspective underpins the end-to-end pipeline shown below.



System Overview

1. Use the **Checking Service** to scrape for new press release URLS.
2. URLs enter the **Lexico Process**, where we extract the transcript and use algorithms to match the parallel sentences.
3. **Advanced Scoring** provides a variety of property based tests to validate the aligned output.
4. **Integration** connects the backend processing with user-facing interfaces.
5. **Updates the database** and gets ready for a new pipeline cycle.

Design Calculations and Analysis

This section lays out the key design choice that we employed in each part of the system pipeline (as the diagram on the left):

Database Checking:

Design a rich SQL schema to capture detailed text features. This supports deeper processing and more flexible analysis downstream.

Lexico Process and Validation:

We treat alignment as a scoring task, combining semantic similarity (**LaBSE model**) with other customized rules to evaluate sentence matching.

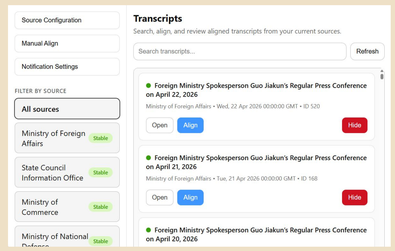
Integration:

Integrate backend logic and frontend interfaces to a full-stack pipeline. Ensures data moves smoothly from processing to user interaction.

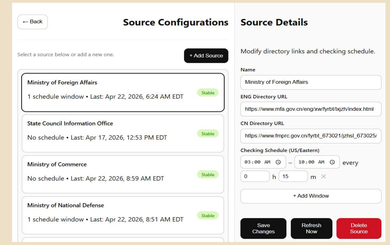
Database Update:

Test for reproducibility and robustness during source and transcript updates.

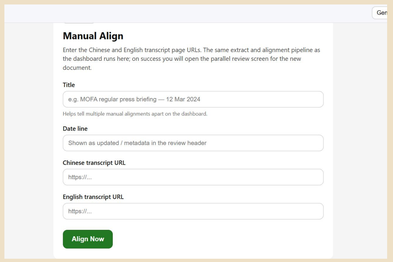
User Interfaces and Test Results



Dashboard



Source Configuration



Manual Align Page

Parallel Transcript Review

Accuracy: 96.3 % of sentences; Efficiency boost: 3536%