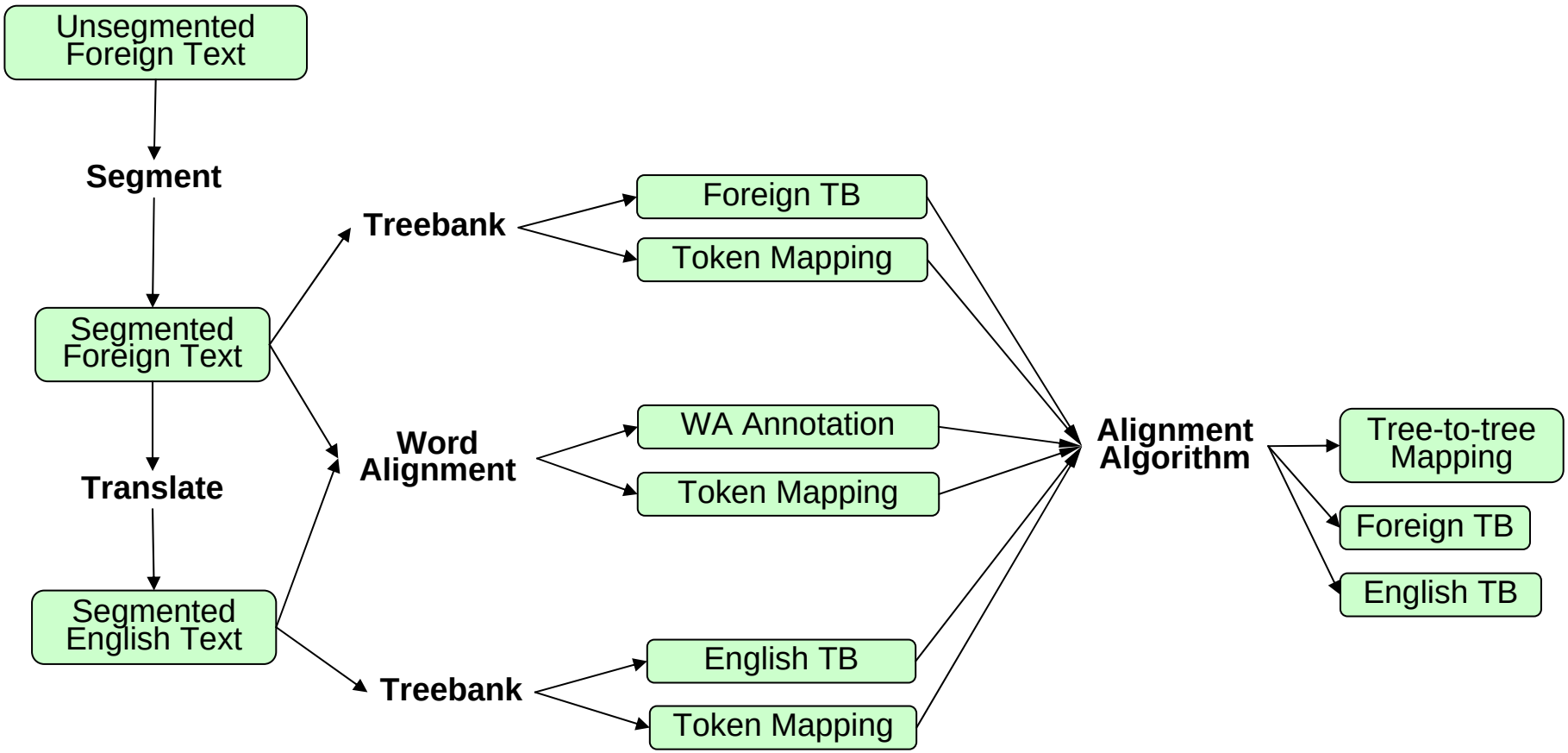
Parallel Aligned Treebanks at LDC: Update on Current Efforts

Ann Bies and Haejoong Lee
Linguistic Data Consortium
{bies,haejoong}@ldc.upenn.edu

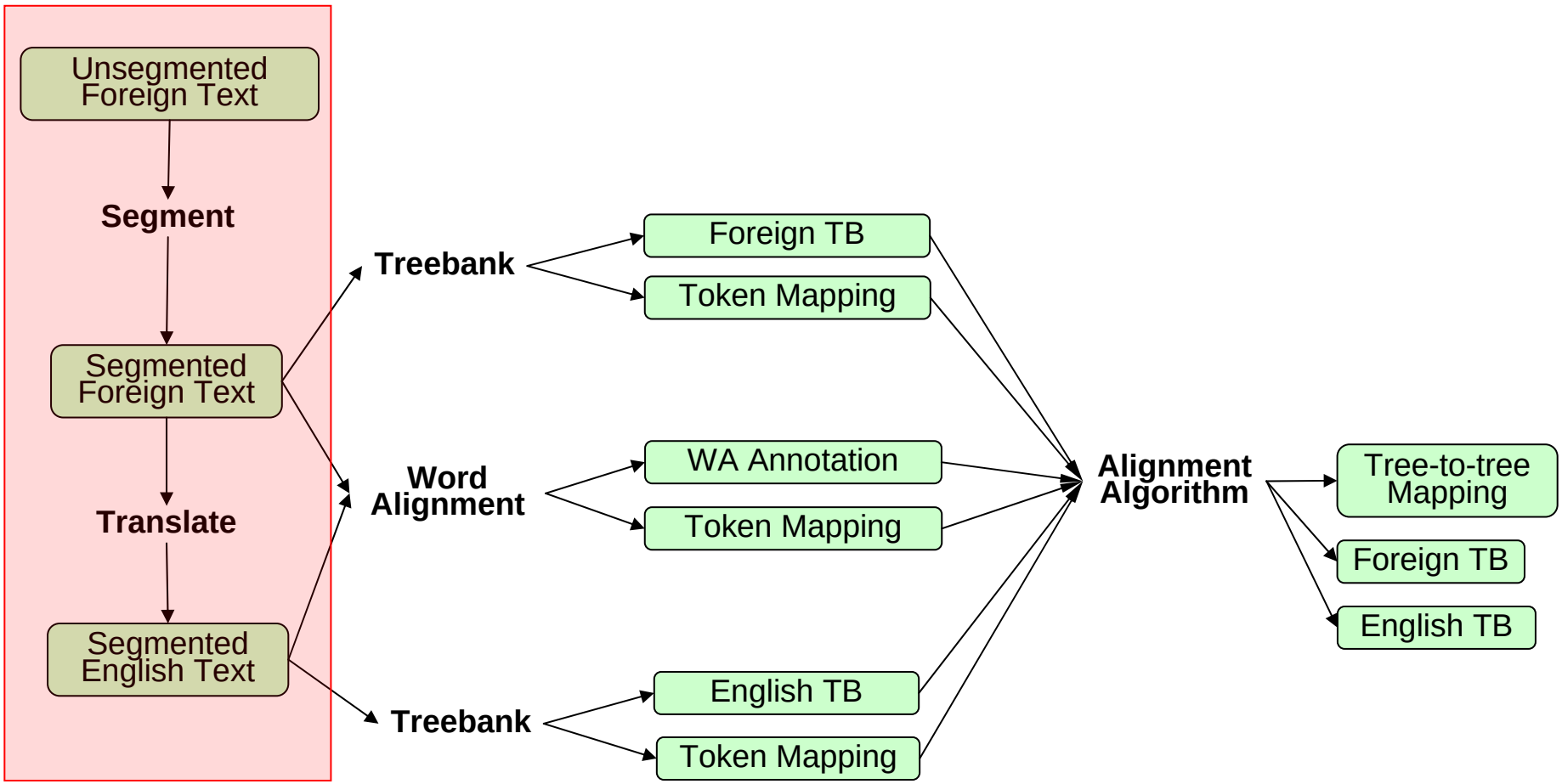
Goal and Challenges

- Goal: Sentence-aligned parallel Treebanks
- Challenges
  - Desire to correct sentence/SU boundaries during TB annotation
  - Different data priorities/schedules for WA vs. TB
- Proposal
  - If downstream annotation can change SU boundaries, need to use WA annotation as intermediary to get token-aligned parallel TB alignment → proposal following
  - If SU boundaries are inviolable for downstream annotation, then SUs will already be parallel

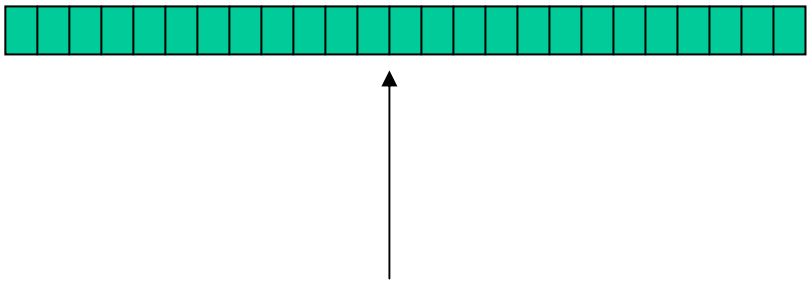
Parallel Aligned Treebanks: LDC Proposal



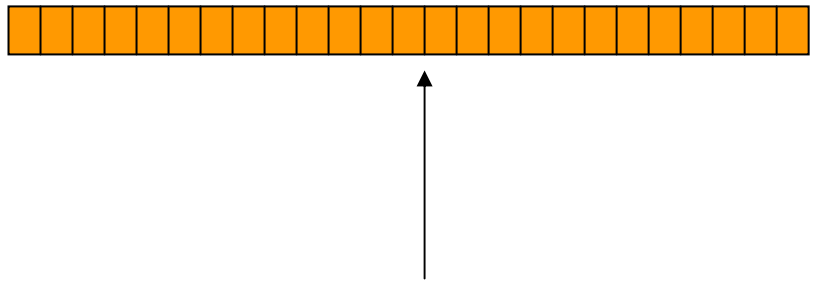
- Step 1: LDC provides sentence segmented source text to TB, WA teams
  - Source text is locked, immutable



Character stream is pre-defined and locked

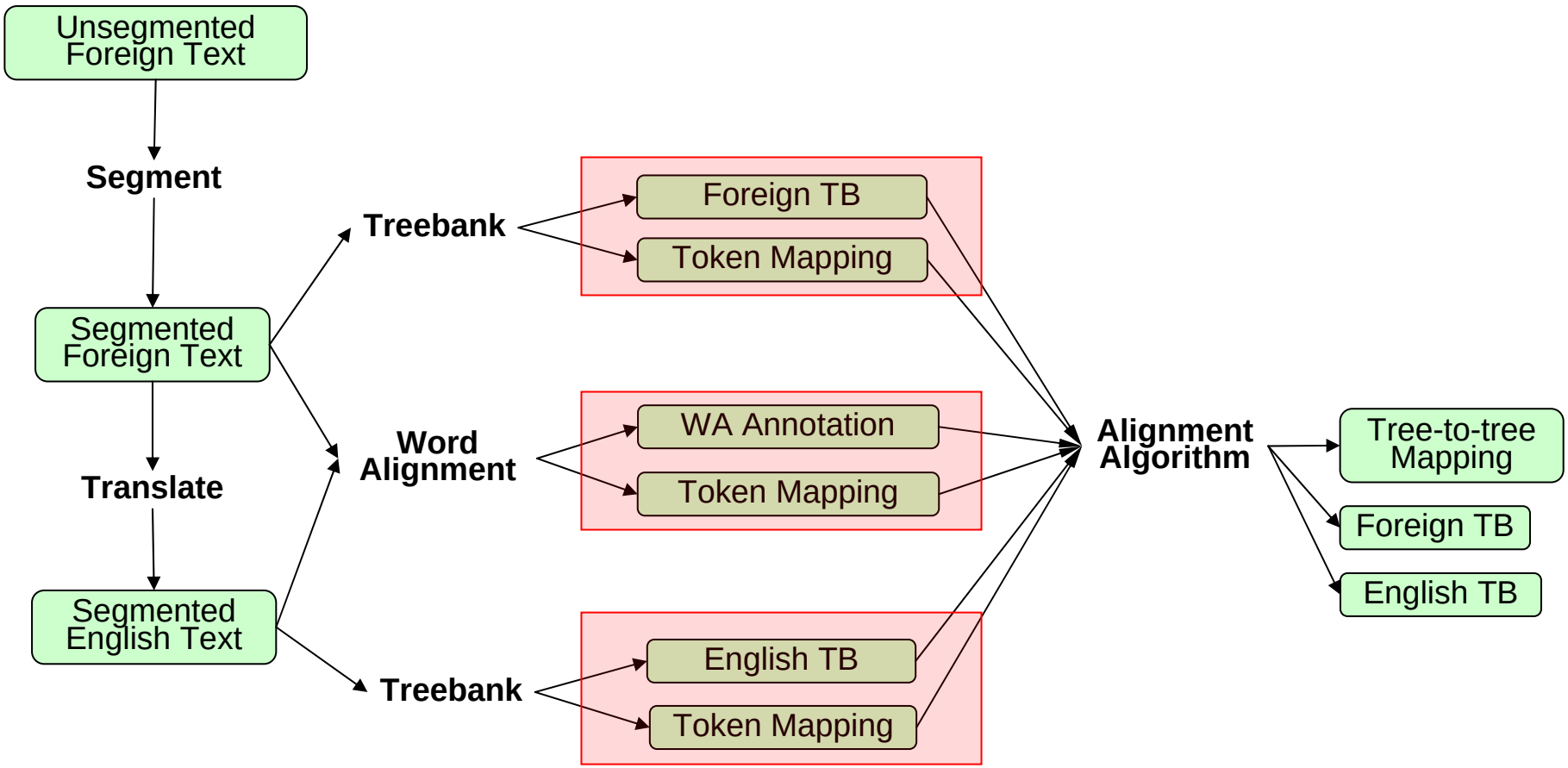


English Source Text
As Character Stream

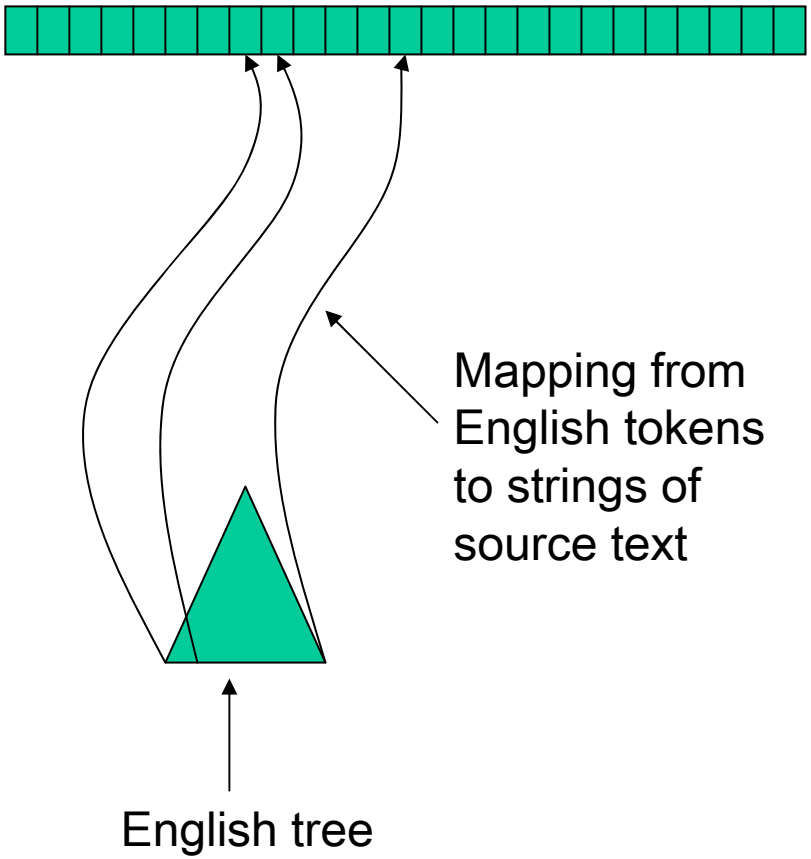


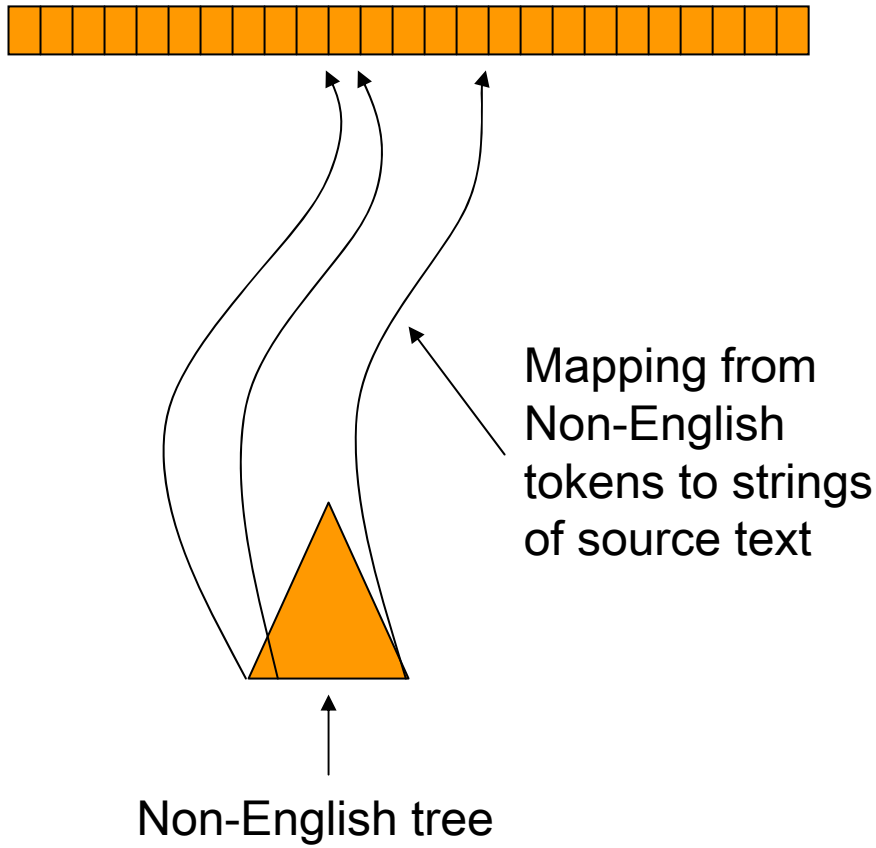
Non-English Source Text
As Character Stream

- Step 2: Treebank, WA teams produce annotation based on LDC source text
  - Provide mapping table from TB, WA tokens to strings in source text

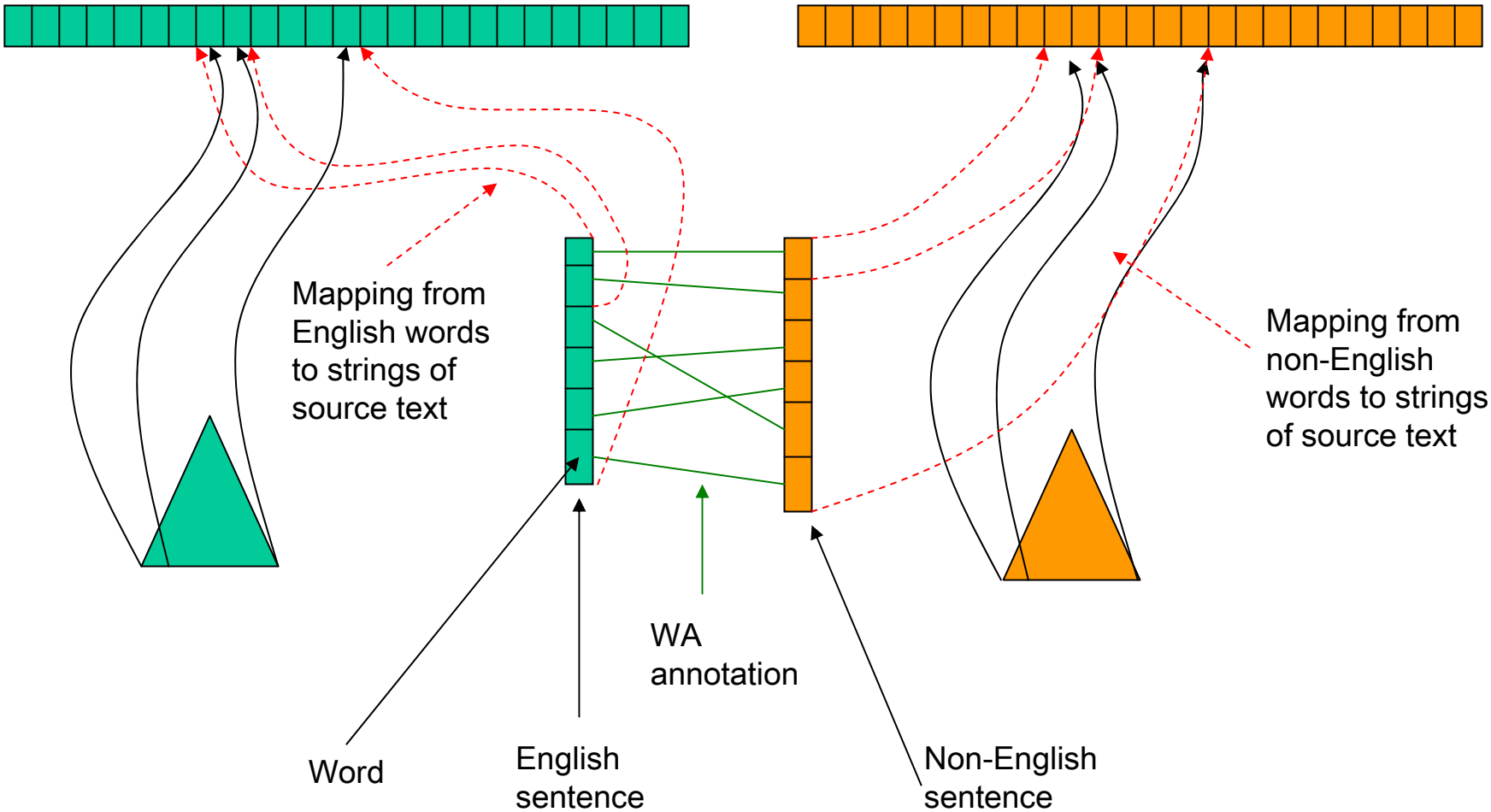


Depiction of TB Mapping Tables



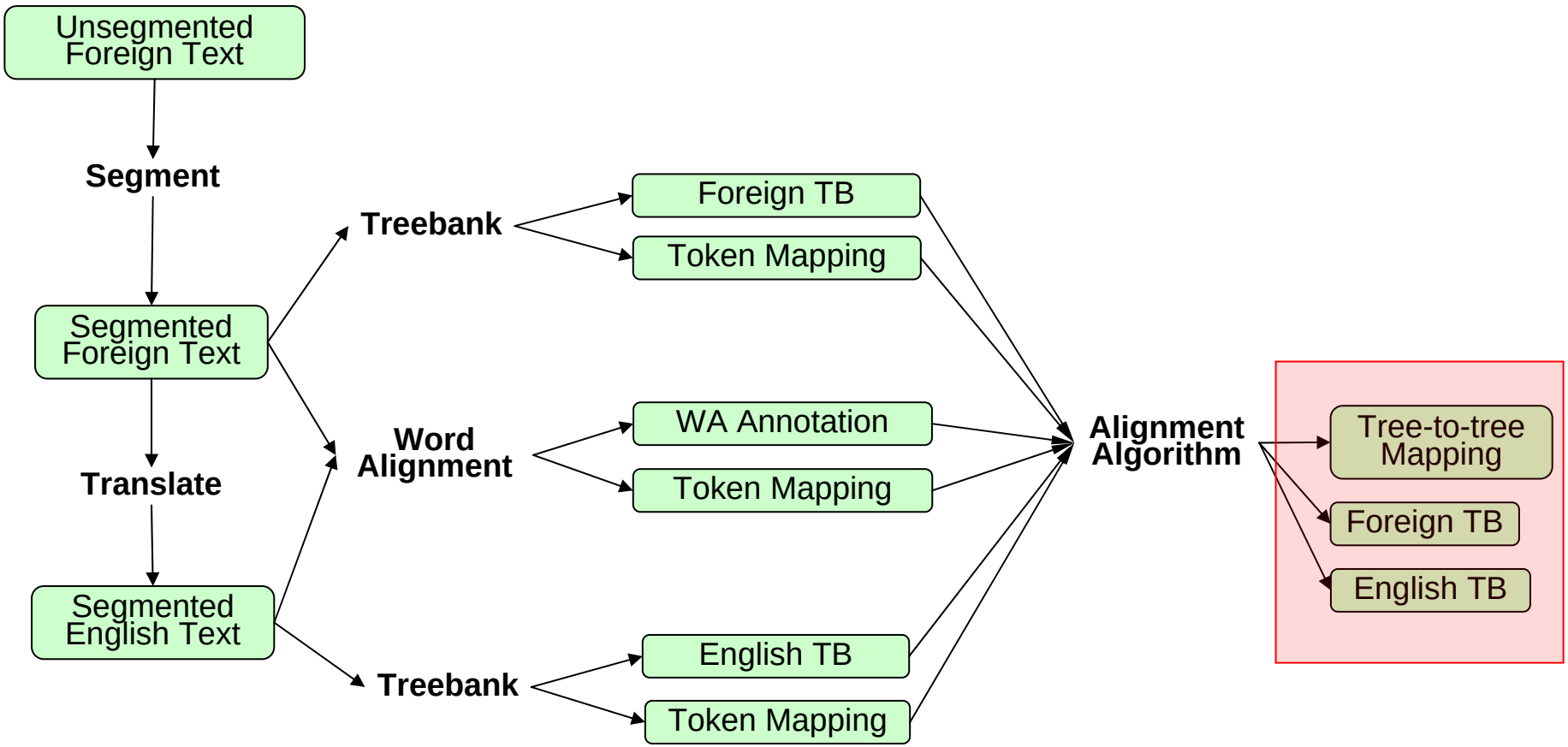


Depiction of WA Mapping Tables

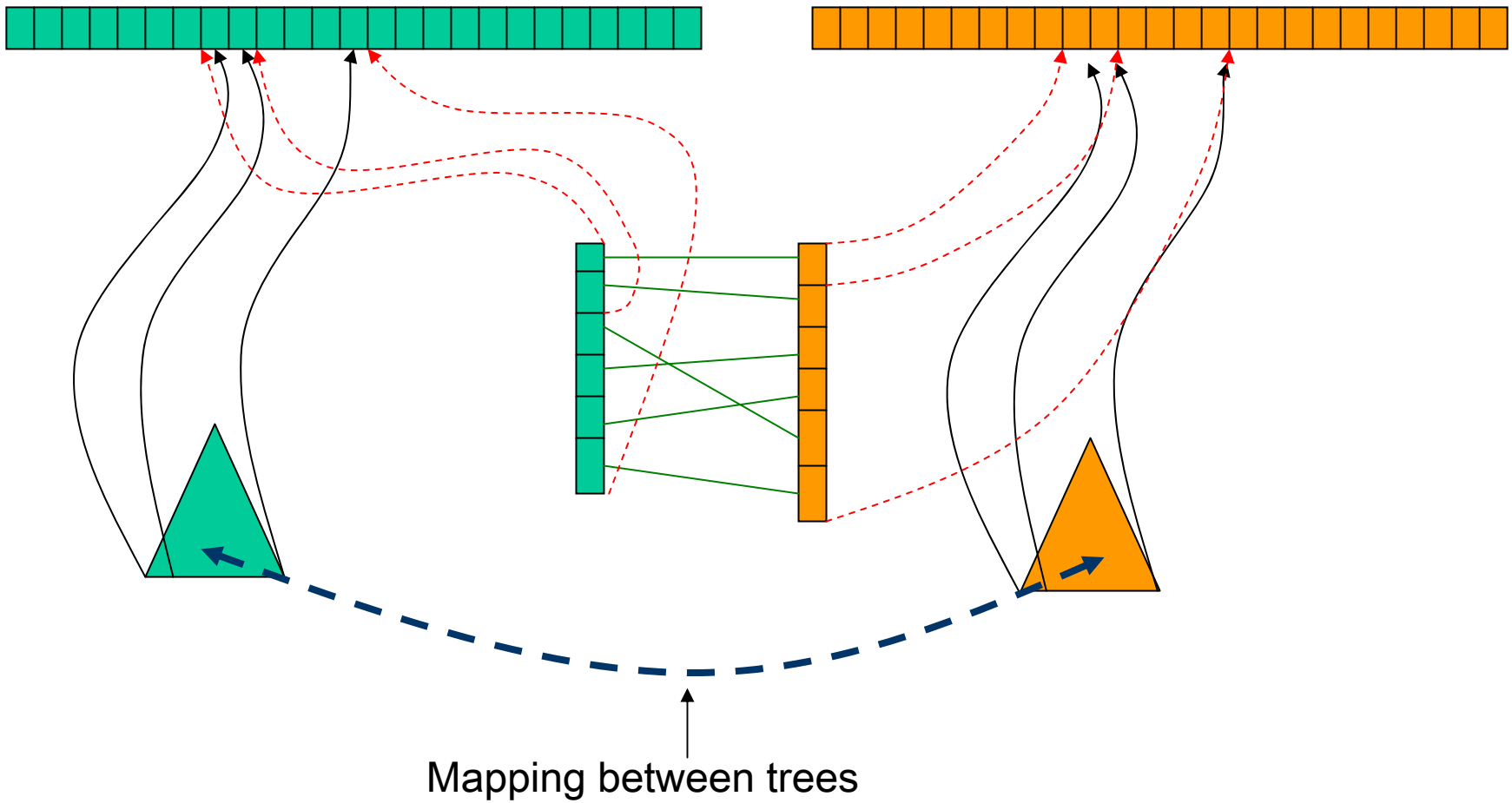


Step 3: LDC runs alignment algorithm

- Incorporate mapping tables provided by TB and WA teams to create mapping between English and non-English trees



Depiction of Tree-to-Tree Mapping



Progress & Milestones

- May 2008: Discussion begins in BAC committee
- Jun-Aug 2008: LDC develops, tests proposal
- Aug-Sep 2008: BBN, LDC discuss and refine proposal
- Nov 2008: Final proposal circulated to BAC group
- Dec 2008: Sample package circulated to BAC group
- Current: Waiting for feedback from BAC
- Pending BAC buy-in, plan will be implemented for all future releases
  - Full implementation requires both TB and WA annotation on same data
  - Requires TB, WA teams to use LDC source text and produce mappings
- Past data sets can be addressed as needed, over time
- Need additional site input to synchronize data set priorities for WA, TB teams