Document Alignment Attempted Many Times

Goal: Given bag of texts in two languages, find pairs.

- A project at this very seminar at FJFI: (Jahoda et al., 2007)
- A project at MFF: (Klempová et al., 2009)
  - Evaluation suggested that the first step is tricky: finding source URLs.
- Václav Novák (ÚFAL, ∼2009): aligning subtitles.
  - Proper minimum pairing algorithm.
  - Not generic enough: focus on named entities at the beg. and end only.
- ParaSite: probably good, re-evaluation would be useful.
  - Problem: Based on libraries with conflicting licenses (GPL 2.0 vs 3.0).
- Parallel **Paragraphs** from CommonCrawl (Kúdela et al., 2017)
  - Recall 63%, precision 94% when re-aligning shuffled CzEng.
  - 149TB of CommonCrawl ⇝ 115k en-cs sentpairs from 2k webdomains.
  - **Targetted re-crawl would be highly desirable (project suggestion).**
- `paracrawl.eu` large but noisy. Aligns documents, not paragraphs.