Mining the Web

Goal: Given two language names, find parallel texts.

- Hervé Saint-Amand's master's thesis (Saarbrücken).
  - Train language identification on Wikipedia.
  - Search for pages in English containing the word *česky*.
- Bitextor: Esplà-Gomis and Forcada (2010)
- PANACEA tools (`http://myexperiment.elda.org/workflows/7`)
- Students' project ParaSite: proof of concept, fixes needed.

Quasi-comparable sources (incl. Wikipedia):

- Texts on the same topic but written independently.
- Can hope to find parallel sentences but no longer segments.
- BUCC workshops 2008–2020: `https://comparable.limsi.fr/bucc2020/`
- "Lightly supervised training" (Schwenk, 2008) = basis of unsupervised MT.