Preparing the final corpus

choice of corpus

- new corpus – existing resources
- domain – genre specs: news/general – newspapers
- size c. 3500 – 4000 vMWEs
- texts **in the original** rather than translated
- corpus preferably free from copyright issues
- whole texts or paragraphs
- **pre-select** sentences containing MWEs **only if negative** examples are also available

Corpus characteristics

- input format: TSV, one-token-per-line, see website
- size of individual files (unrestricted; if big: the annotators should be heavily involved)
- file naming convention:
  - <language-code>-<number>-<free-text>
  - example: EN-23-Herald-Tribune.txt
- existence of pre-annotations: allowed but beware of the bias (the system does NOT know better, review the whole text, not only the pre-annotated candidates)

some recommendations…

- file management - best practices (don't assign all files in advance)
- spreadsheet for managing annotations
  - encourage all annotators to register to the FLAT server
  - note their usernames in the spreadsheet
- double annotation of a corpus fragment for IAA calculus - how big should it be?