Bigger and better is best: Construction and evaluation of the COW14 web corpora

Previous releases of our "Corpora from the Web" (COW12) introduced only incremental improvements over first-generation web corpora (WaCky). However, the latest releases, COW14 (Dutch, English, French, German, Swedish, Spanish), represent a major step forward. In the corpus creation tool chain, we implemented numerous measures to improve data quality, which we are in the process of evaluating. Also, the COW14 corpora are larger than the COW12 corpora by a factor of roughly 2, and they are no longer based on web crawls in single top-level domains. In this talk, we provide an overview of some of the COW14 technology and the COW14 data.

We cover the following topics:

- improved filtering (deduplication, bad documents, etc.)
- non-destructive cleaning
- noise removal (e.g., removal of hard-coded hyphenation)
- improved and extended linguistic annotation
- automatically generated meta data (e.g., geolocalization)
- corpus composition (text-types and registers)
- evaluation of filtering methods using distributional semantics
- COW distribution model and interfaces

http://corporafromtheweb.org/
https://webcorpora.org/