Brooklyn-Korpus

- eigentlich: *Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English*
- erstmals 2000 publiziert
- 16 Prosatexte aus dem Altenglisch-Teil des (diachronen) Helsinki-Korpus
- (einigermassen) repräsentativ bezüglich Perioden und Textsorten
- total 106000 Wörter, 19000 Wortformen und 7000 Lemmata
- mehr Info hier

Reichhaltig annotiert

- auf Stufe Text
  - Periode
  - Textsorte
  - bibliographische Quellenangabe
- auf Stufe s-unit
  - syntaktische Annotation
  - Glosse (Modern English)
  - Referenz (Seiten- und Zeilennummer)
- auf Stufe w-unit
  - morphologische Annotation (Kasus/Numerus/Genus bzw. Person/Numerus/Modus/Tempus)
  - Part of Speech (Tagset mit 111 Tags)
  - Lemma

Konvertierung zu TEI

- ursprüngliches Korpus liegt in einem proprietären Format vor
- individuelle Texte mittels eines Perl-Skripts zu TEI konvertiert (mit minimaler manueller Nachbearbeitung :-)
- Corpus Header File von Hand angefertigt
- als Unicode bzw. UTF-8 kodiert (Spezialzeichen wie z.B. ę)