Old English Corpora

P. S. Langeslag

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

DOEC: *Dictionary of Old English* Corpus

Definition

At least one copy of every surviving text in Old English

Format

XML with minimal markup (per-text header; sentences, language, source page/line references)

3060 "texts"

Table 1: DOEC statistics for the 2009 release

A	Poetry	177,480 OE words	6%
B	Prose	2,128,781 OE words	70%
C	Glosses	699,606 OE words	23%
D	Glossaries	26,598 OE words	.88%
E	Runes	346 OE words	.01%
F	Inscriptions	331 OE words	.01%
	Total	**3,033,142 OE words**	
	Incl. foreign words	**3,791,645 words**	

DOEC and Offshoots

Corpus	Word Count	Metadata
DOEC	3,033,142 (OE)	Header, sentences, language, page/line ref
Helsinki (OE subset)	413,250	Header, page/line reference
York (poetry)	71,490	Syntax, POS
Brooklyn (prose)	106,210	Syntax, POS, morphology (incl. lemmata)
YCOE (prose)	1,500,000	Syntax, POS

Non-DOEC

Corpus	Size	Metadata
CLASP	60,000 lines (incl. Latin)	Syntax, metrics, lemmata?
EnHiGLa (OE subset)	5,960 clauses	Syntax

ECHOE: Electronic Corpus of Anonymous Homilies in Old English

Definition

All copies of all prose homilies and saints' lives not by Ælfric of Eynsham

Format

XML with markup for words, sentences, (paragraphs), proper nouns and adjectives, numerals, language, biblical source references, content cross-references, scribal hands, ink colour, script, initial size, MS folio and line, emendations, scribal modifications, etc.

Statistics

Metric	Status	Notes
MS items drafted	335 (97%)	10 further, damaged items outstanding
Total OE words	537,724	547,047 including foreign words
Average item length	1605 words	1633 including foreign words

Our Copy of ECHOE

- Plaintext
- Stripped of foreign words
- Stripped of punctuation
- Lowercase
- Word and sentence division retained through spacing and newlines
- One manuscript item per file

Bibliography

Healey, Antonette diPaolo, ed. “Dictionary of Old English Web Corpus.” Toronto: Dictionary of Old English Project, 2009. https://tapor.library.utoronto.ca/doecorpus/.

Orchard, Andy, ed. “A Consolidated Library of Anglo-Saxon Poetry.” Accessed May 3, 2022. https://web.archive.org/web/20220320085850/https://clasp.ell.ox.ac.uk/.

Pęzik, Piotr, Anna Cichosz, Jerzy Gaszewski, and Maciej Grabski, eds. “EnHiGLa.” Accessed May 2, 2022. http://pelcra.pl/enhigla/.

Pintzuk, Susan, Ann Taylor, Anthony Warner, Leendert Plug, and Frank Beths, eds. “York–Helsinki Parsed Corpus of Old English Poetry.” Accessed May 2, 2022. https://www-users.york.ac.uk/~lang18/pcorpus.html.

Rissanen, Matti et al., eds. “Helsinki Corpus of English Texts.” University of Helsinki, 2011. https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/.

Rudolf, Winfried et al. “Electronic Corpus of Anonymous Homilies in Old English.” Accessed May 2, 2022. https://echoe.uni-goettingen.de.

Taylor, Ann, Anthony Warner, Susan Pintzuk, and Frank Beths, eds. “The Toronto-Helsinki-Parsed Corpus of Old English Prose,” 2003. https://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm.