ParCorOEv2
An open access annotated parallel corpus Old English-English
2021

MANUAL 3: Tokenisation
Version 1.0

Martín Arista, Javier, Sara Domínguez Barragán, Laura García Fernández,
Esaúl Ruíz Narbona, Roberto Torre Alonso & Raquel Vea Escarza (comp.)

Nerthus Project
Universidad de La Rioja
www.nerthusproject.com.

This manual describes and illustrates the corpus processing of ParCorv2, includining the tokenisation, concordance and triangulation of the source language and target language texts.

Tokenisation, concordance and triangulation

Tokenisation: text and section

The corpus has four types of units: texts, sections, fragments and tokens.
Each text is divided into sections (books, chapters, homilies, years, etc.).
Each text is identified by means of a sequence of four or five letters and, in some cases, one digit:

BOET, ÆHOM1

Each token is identified by means of an alpha-numerical sequence of the form:

Text code.Section number.Fragment number.Word number.

Section, fragment and word numbers consists of three digits and are separated by dots:

MART.190.008.009.

If the edition does not divide the text into any kind of part, all the fragments are assigned the section number 1.
The fragments in chronicles are numbered after the year in the chronicle entry:

ASCA.YEAR912.001.024.

If there is more than one entry to the chronicle per year, the fragment number corresponding to the year is followed by a lower case consecutive letter, as in 1066a, 1066b, 1066c, etc.
The index, if it is written in Anglo-Saxon, is rendered as 0. The first section number is 1, irrespectively of the existence of a prefix or an index.
The preface, if it is written in Anglo-Saxon, is rendered as 00. The first section number is 1, irrespectively of the existence of a prefix or an index.
The fragmentation criterion (division into sections) must be specified in the field Notes on tokenisation.

Tokenisation: fragment and token

A fragment is a period (a meaningful segment between two full stops).
As a general rule, fragments are between one and three lines long (in the original text).
The maximal length of fragments is 75 words. Exceptionally, a fragment may be longer if there are no punctuation marks that delimit a period with less than 75 words.
Only exceptionally (to avoid a period longer than three lines) can a period end with a colon or a semicolon.

Only exceptionally (to avoid periods shorter than one line) can a fragment consist of more than one period).

Coding

A new file is created for each token.
Tokens are numbered consecutively.
The alignment of the corpus as well as the parallel corpus layout crucially depend on the accuracy or token numbers.
Fragments and tokens are identified by source and by target.
The source includes the source text and the translation.
Coding by source (text and translation) is the information on the page of the text and the translation.
Coding by source is rendered in the following form:

> Editor/translator surname (year: page).

If the translation is split between two non-consecutive pages, the reference has the following form:

> Editor/translator surname (year: page, page).

Coding by target is an alpha-numerical sequence of the form:

> Text code.Text section number.Fragment number.Word number.

Illustration of corpus processing

Source text

[Required edition changes in bold]
Godes gelaðung wurðað **þ**isne dæg ðam mæran apostole Paule to wurðmynte**,** forðam ðe he is gecweden ealra ðeoda l**á**reow: **þ**urh soðfæste lare wæs ðeah**-**hwæðere his martyrd**ó**m samod mid ðam eadigan Petre gefremmed**.** H**é** wæs fram cildh**á**de on ðære ealdan **ǽ** getogen, and mid micelre gecnyrdnysse on ðære begriwen wæs. Æfter Cristes ðrowunge, ðaða se soða geleafa aspr**á**ng **þ**urh ðæra apostola bodunge**,** ða ehte he cristenra manna **þ**urh his nytennysse**,** and sette on cwearterne**,** and eac wæs on geðafunge æt ðæs forman cyðeres Stephanes slege: nis ðeah**-**hwæðere be him geræd, **þ**æt h**é** handlinga ænigne man acwealde**.**

Source translation

The church of God celebrates this day in honour of the great Apostle Paul, for he is called the teacher of all nations: though his martyrdom, for true doctrine, was accomplished with the blessed Peter's. He had from childhood been bred up in the old law, and by great diligence was therein deeply imbued. After Christ's passion, when the true faith had sprung up through the preaching of the apostles, he persecuted Christian men through his ignorance, and set them in prison, and was also consenting to the slaying of the first martyr Stephen: it is not, however, read of him that he killed any man with his own hands.

Step 1: Edition

Godes gelaðung wurðað ðisne dæg ðam mæran apostole Paule to wurðmynte, forðam ðe he is gecweden ealra ðeoda lareow: ðurh soðfæste lare wæs ðeahhwæðere his martyrdom samod mid ðam eadigan Petre gefremmed. He wæs fram cildhade on ðære ealdan æ getogen, and mid micelre gecnyrdnysse on ðære begriwen wæs. Æfter Cristes ðrowunge, ðaða se soða geleafa asprang þurh ðæra apostola bodunge, ða ehte he cristenra manna ðurh his nytennysse, and sette on cwearterne, and eac wæs on geðafunge æt ðæs forman cyðeres Stephanes slege: nis ðeahhwæðere be him geræd, ðæt he handlinga ænigne man acwealde.

Step 2: Fragmentation

Godes gelaðung wurðað ðisne dæg ðam mæran apostole Paule to wurðmynte, forðam ðe he is gecweden ealra ðeoda lareow:

ðurh soðfæste lare wæs ðeahhwæðere his martyrdom samod mid ðam eadigan Petre gefremmed.

He wæs fram cildhade on ðære ealdan æ getogen, and mid micelre gecnyrdnysse on ðære begriwen wæs.

Æfter Cristes ðrowunge, ðaða se soða geleafa asprang þurh ðæra apostola bodunge, ða ehte he cristenra manna ðurh his nytennysse, and sette on cwearterne, and eac wæs on geðafunge æt ðæs forman cyðeres Stephanes slege:

nis ðeahhwæðere be him geræd, ðæt he handlinga ænigne man acwealde.

Step 3: Translation

The church of God celebrates this day in honour of the great Apostle Paul, for he is called the teacher of all nations:

though his martyrdom, for true doctrine, was accomplished with the blessed Peter's.

He had from childhood been bred up in the old law, and by great diligence was therein deeply imbued.

After Christ's passion, when the true faith had sprung up through the preaching of the apostles, he persecuted Christian men through his ignorance, and set them in prison, and was also consenting to the slaying of the first martyr Stephen:

it is not, however, read of him that he killed any man with his own hands.

Step 4: Concordance (of fragment 5)

[Punctuation is kept; if the punctuation mark immediately follows the ConcTerm, it is the first character in the postfield]

Prefield	**ConcTerm**	**Postfield**
	nis	ðeahhwæðere be him geræd, ðæt he handlinga ænigne man acwealde.
nis	ðeahhwæðere	be him geræd, ðæt he handlinga ænigne man acwealde.
nis ðeahhwæðere	be	him geræd, ðæt he handlinga ænigne man acwealde.
nis ðeahhwæðere be	him	geræd, ðæt he handlinga ænigne man acwealde.
nis ðeahhwæðere be him	geræd	, ðæt he handlinga ænigne man acwealde.
nis ðeahhwæðere be him geræd,	ðæt	he handlinga ænigne man acwealde.
nis ðeahhwæðere be him geræd, ðæt	he	handlinga ænigne man acwealde.
nis ðeahhwæðere be him geræd, ðæt he	handlinga	ænigne man acwealde.
nis ðeahhwæðere be him geræd, ðæt he handlinga	ænigne	man acwealde.
nis ðeahhwæðere be him geræd, ðæt he handlinga ænigne	man	acwealde.
nis ðeahhwæðere be him geræd, ðæt he handlinga ænigne man	acwealde	.

Step 5: Tokenisation (of fragment 3)

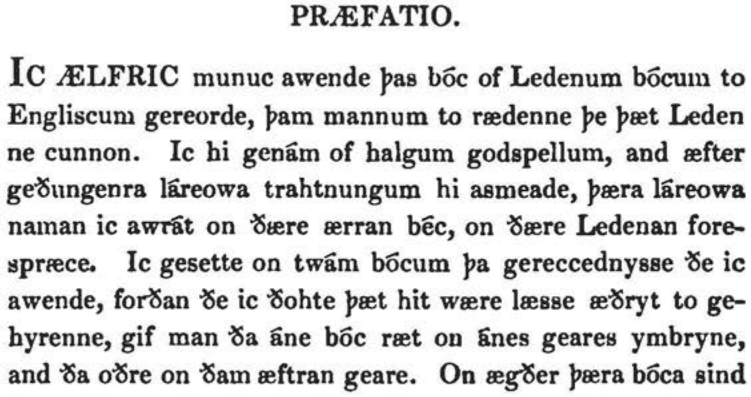
Text code: ÆHOM1
Text section: 00 (preface)
Fragment: 3
Words: 1-

ÆHOM1.00.003.001	He
ÆHOM1.00.003.002	wæs
ÆHOM1.00.003.003	fram
ÆHOM1.00.003.004	cildhade
ÆHOM1.00.003.005	on
ÆHOM1.00.003.006	ðære
ÆHOM1.00.003.007	ealdan
ÆHOM1.00.003.008	æ
ÆHOM1.00.003.009	getogen

ÆHOM1.00.003.010	and
ÆHOM1.00.003.011	mid
ÆHOM1.00.003.012	micelre
ÆHOM1.00.003.013	gecnyrdnysse
ÆHOM1.00.003.014	on
ÆHOM1.00.003.015	ðære
ÆHOM1.00.003.016	begriwen
ÆHOM1.00.003.017	wæs

Step 6: Triangulation

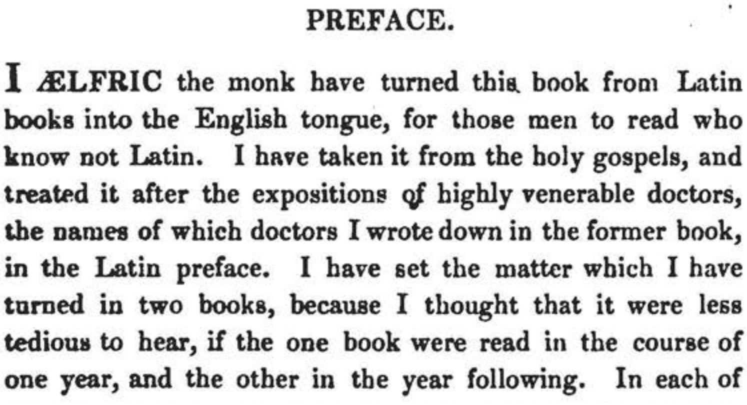
Source text (Thorpe 1846: 2)



PRÆFATIO.

IC ÆLFRIC munuc awende þas bōc of Ledenum bōcum to Engliscum gereorde, þam mannum to rædenne þe þæt Leden ne cunnon. Ic hi genām of halgum godspellum, and æfter geðungenra lārewa trahtnungum hi asmeade, þæra lārewa naman ic awrāt on ðære ærran bēc, on ðære Ledenan foresprǽce. Ic gesette on twām bōcum þa gereccednysse ðe ic awende, forðan ðe ic ðohte þæt hit wære læsse æðryt to gehyrenne, gif man ða āne bōc ræt on ānes geares ymbryne, and ða oðre on ðam æftran geare. On ægðer þæra bōca sind

Source translation (Thorpe 1846: 3)



PREFACE.

I ÆLFRIC the monk have turned this book from Latin books into the English tongue, for those men to read who know not Latin. I have taken it from the holy gospels, and treated it after the expositions of highly venerable doctors, the names of which doctors I wrote down in the former book, in the Latin preface. I have set the matter which I have turned in two books, because I thought that it were less tedious to hear, if the one book were read in the course of one year, and the other in the year following. In each of

ParCor Parallel Text

Edited fragment: IC ÆLFRIC munuc awende ðas boc of Ledenum bocum to Engliscum gereorde, ðam mannum to rædenne ðe ðæt Leden ne cunon.

ConcTerm: IC
Text name: Ælfric´s Catholic Homilies I
Text code: ÆHOM1
Source text reference: Thorpe (1846: 2)
Source translation reference: Thorpe (1846: 3)

PacCorOE number: ÆHOM1.00.003.001

ConcTerm: ÆLFRIC
Text name: Ælfric´s Catholic Homilies I
Text code: ÆHOM1
Source text reference: Thorpe (1846: 2)
Source translation reference: Thorpe (1846: 3)
PacCorOE number: ÆHOM1.00.003.002

ConcTerm: munuc
Text name: Ælfric´s Catholic Homilies I
Text code: ÆHOM1
Source text reference: Thorpe (1846: 2)
Source translation reference: Thorpe (1846: 3)
PacCorOE number: ÆHOM1.00.003.003