

# INTERFACE OF OLD ENGLISH DICTIONARIES

Sorting out headword  
spelling and format  
differences

SELIM 32

Javier Martín Arista  
Universidad de La Rioja

PID2020-119200GB-100, funded by MCIN  
/ AEI / 10.13039/501100011033



# What is an interface?

- A **relational lexical database**.
- Devised for **comparing** the information provided by dictionaries of Old English. The focus is on **headwords and inflectional forms**, but lexical class, morphological categories, meaning definition, etc. can also be compared.
- The general method is based on **normalisation**, mainly aimed to format.
- The interface also allows for comparison with **corpora**, from which morphological tags, lexical categories and textual frequencies are imported.

# Talk plan



Method and sources



Building the Interface



Exploiting the Interface



Conclusions and further research



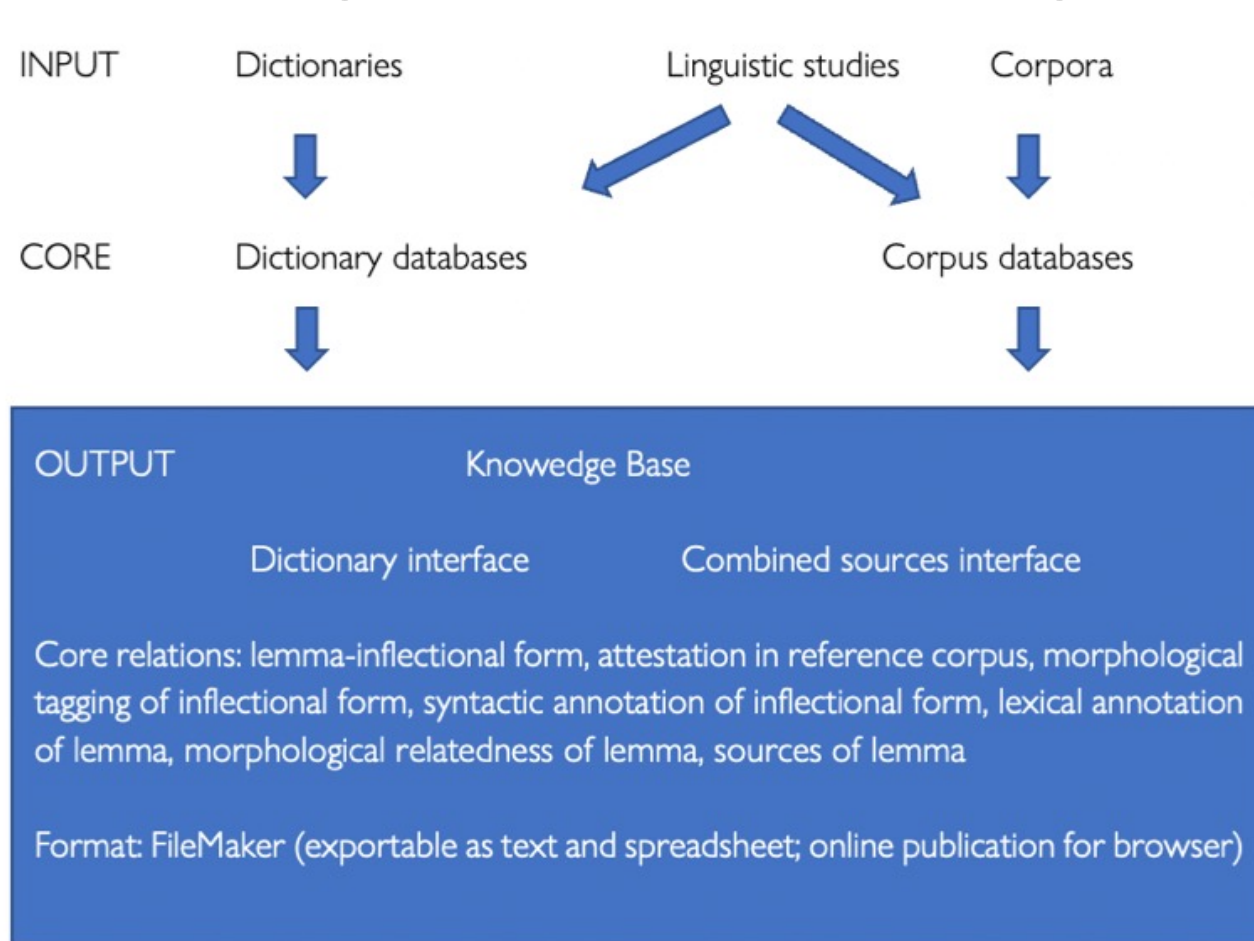
# METHOD AND SOURCES



# The concept of knowledge base

- A relational lexical **database**
- Implemented in database software
- Database software is used to gather, analyse and disseminate the data
- It contains layouts, fields and values
- **Relations are stored as data**
- Elements and the relations holding among them are represented as knowledge graphs
- Output should be readable by humans and **interpretable by machines**

# The Knowledge Base of Old English (KBOE)



# Entities and relations in the KBOE

scearpa	IS AN INFLECTIONAL FORM OF	scearp
scearp	BELONGS TO THE LEXICAL CATEGORY	ADJ
scearpa	IS THE ACCUSATIVE OF	scearp
scearpa	HAS A TEXTUAL FREQUENCY OF	12
scearpe	IS A DERIVATIVE OF	scearp
sharp	IS A MEANING DEFINITION OF	scearp
Durkin (2014: 191)	IS A RELEVANT REFERENCE TO	scearp
scearpere	CAN BE FOUND IN DOEC TOKEN	[Alc (Warn 35) 003800 (90)]
scearpere	CAN BE FOUND IN YCOE TOKEN	(ID coalcuin,Alc_[Warn_35]:90.65)

---

# Scope and sources

## Scope

- An Anglo-Saxon Dictionary* (Bosworth and Toller 1973, hereafter BT)
- The student's Dictionary of Anglo-Saxon* (Sweet 1976, henceforth SWT)
- A Concise Anglo-Saxon Dictionary* (Hall 1996, hereafter CHM)
- The Dictionary of Old English in Electronic Form A-I* (Healey et al. 2018, henceforth DOE).

## Textual Sources

- The York-Helsinki Parsed Corpus of Old English Poetry* (Pintzuk and Plug 2001)
- The York-Toronto-Helsinki Parsed Corpus of Old English Prose* (Taylor et al. 2003)
- The Dictionary of Old English Corpus* (Healey et al. 2004).

## Digitised dictionaries

The lexicographical sources of this study include the text file of BT (which excludes the 1921 *Supplement*) that can be found at <https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3532#>; the text file of SWT that can be downloaded from [https://archive.org/stream/studentsdictiona00sweerich/studentsdictiona00sweerich\\_djvu.txt](https://archive.org/stream/studentsdictiona00sweerich/studentsdictiona00sweerich_djvu.txt); the text file of CHM available from <https://www.gutenberg.org/ebooks/31543>; and the *The Dictionary of Old English A-H* on CD-ROM, supplemented by *The Dictionary of Old English: A to I online* that can be accessed at <https://tapor.library.utoronto.ca/doe/>.

\*I would like to thank Professor Ondřej Tichý for sharing 12,000 corrected entries to the online version of BT in progress, which have been added to the approximately 20,000 entries previously available at <http://hdl.handle.net/11234/1-3532>.



# Data from lexicographical sources

	CHM	SWT	BT	DOE
A	397	288	1,645	1,534
Æ	437	261	454	630
B	2,453	1,692	2,617	2,268
C	1,387	980	1,296	1,411
D	830	578	817	921
E	1,016	587	1,272	1,488
F	2,919	2,009	3,015	3,010
G	1,460	2,259	4,583	1,322
H	2,961	2,319	2,471	2,962
I	370	245	343	435
A-I total	14,230	11,218	18,513	15,981
Dictionary total	36,742	25,922	36,104	15,981

# REVIEW OF DICTIONARIES OF OLD ENGLISH



# The entry for *calan* ‘to be cold’ in CHM, SW, BT and the DOE

CHM. **calan** 6 to grow cool or cold.

SW. **calan** 2 be cold intr.: *hine cælp, him cælp*.

BT. **CALAN**, ic cāle, ðú calest, cælst, he caleþ, cælp, pl. calap; p. cól, pl. cólón; pp. calen; v. intrans. To be or become cool or cold; algere, frigesce :- *Ðonne him cælp, he cépp him hlywþe* when he is cold, he betakes himself to shelter, Hexam. 20; Norm. 28, 22. *Hwæðer ða wélgan ne ne cale* do the rich never become cold? Bt. 26, 2; Fox 92, 34. [Wyc. kele, koole: Orm. kelenn: Plat. kölen: O. Sax. kólón: O. Frs. kela: Dut. koelen: Ger. kühlen: M. H. Ger. kuolén UNCERTAIN to become cold: O. H. Ger. kuoljan: Dan. koele: Swed. koela: Icel. kala; p. kól; pp. kalit algere: Lat. gelare.] DER. a-calan, of-: calian: célan, a-, ge-: céle, cýle, ftér-; -gicel, -wyr: célnes, ge-: céling: célung, ge-: cól, -nes: cólían, a-: ceald, cald, æl-, brim-, eal-, hrím-, ís-, morgen-, ofer-, sin-, snáw-, wæl-, winter-: caldu, sin-: cald-heort: cealdian, a-: cǫlian: cǫélan, ge-.

DOE. **calan**

Vb., st. 6

Att. sp.: calan || cæleð, cælp, cælð; cælid (ErfGI) | caelith (EpGI), kaelið (CorpGI) || cale (Bo ms B) || col || cole || calendan

11 occ.

1. of human beings: to be or become cold

...

1.a. with dative or accusative of person affected

ÆHex 528: *ðonne him hingrað, he yt grædilice, eft þonne him þyrst, he drincð gif he hæfð, þonne him cælð, he cepð him hlywðe* (cf. ÆCreat 282 *him cæleð*).

...

2. glossing *infrigidare* ‘to make cold’

EpGI 421: *infridat caelith*.

...

Lat. equiv. in ms: *algere, infrigidare*

See also: a-, of-calan; cf. cyle, ceald adj, ceald noun, cōl; gecylled; cēlan, cealdian, cōlian.

## Entry structure: headword, category, morphology and occurrence

	CHM	SW	BT	DOE
headword entry	+	+	+	+
lexical category	±	+	+	+
morphological class	+	+	+	+
inflectional paradigm	±	±	+	-
attested forms	-	-	-	+
weak forms	-	-	-	+
doubtful form	+	+	+	+
hypothetical form	+	+	-	+
poetical form	+	+	±	+
alternative spelling in manuscripts	-	-	±	±
textual frequency	-	-	-	+

## Entry structure: syntax, derivational morphology and phraseology

	CHM	SW	BT	DOE
grammatical information (transitivity)	-	+	±	+
grammatical information (case government)	±	±	±	+
grammatical information (other: passive, reflexive, absolute, correlative, etc.)	-	±	-	+
morphological base	+	+	±	-
morphological derivatives	-	+	+	+
phraseology	±	-	±	+

## Entry structure: definition, citations, cross-references and links to other sources

	CHM	SW	BT	DOE
meaning definition	+	+	+	+
citations	-	±	+	+
vowel length in citations	-	+	+	-
textual sources of citations	+	±	+	+
translation of citations	-	±	±	-
cross- references to alternative spellings	+	+	+	-
cross-reference to infinitive	±	±	±	-
cross-reference to nominative	±	±	±	-
conferre	±	±	±	±
references or links to other lexicographical sources	±	-	±	+
references to grammars of Old English	-	-	±	-

# Headword spelling

	<b>Bosworth</b>	<b>BT</b>	<b>SW</b>	<b>CHM</b>	<b>DOE B-D</b>
'to beckon'	bicnian/ bycnan	beacnian	biecnan	biecnan	bicnan
'to bend'	bigan	bugan	biegan	biegan	bigan
'arch, vault'	bigels	bigels	*biegels	bigels	bigels
'to encourage'	byldan	byldan	bieldan	bieldan	byldan
'boldness'	byldo	byld(o)	bieldo	bieldo	byldu
'trumpet'	byme	byme	bieme	bieme	byme
'birch'	birce	birce	*bierce	beorc	byrc
'to shine'	---	beorhtan	bierhtan	bierhtan	beorhtian
'brightness'	bierhte	beorhtu	bierhto	bierhtu	beorhtu
'to burn'	byrnan	beorman	biernan	biernan	byrnan
'mallet'	bytl	bytl	bietel	bietl	bytel
'concubine'	cyfes	cyfes	ciefes	cifes	cifes
'to call'	cygan	cigan	ciegan	ciegan	cigan
'cold'	cyle/cile	cyle	ciele	ciele	cyle
'merchant'	cypa	cypa	ciepa	ciepa	cypa
'onion'	cipe	cipe	ciepe	cipe	cipe
'shout'	cym	cirm	*cierm	cirm	cirm
'to turn'	cirran	cyrran	cierran	cierran	cyrran
'fastidious'	cis	cies	*cies	cis	cies
'cheese'	cyse	cyse	*ciese	cyse	cyse
'coffin'	cist	cyst	*ciest	cist	cist
'kettle'	cytel/cetel	cytel	*cietel	citel	cytel
'to kill'	dydan	dydan	*diedan	dydan	dydan
'to dip'	dufan	dufan	*diefan	dyfan	dyfan
'hidden'	diogol	digol	diegle	diegol	digol
'to dip'	dyppan	dyppan	*diepan	dyppan	dyppan
'to conceal'	dyrnan	dyrnan	diernan	diernan	dyrnan

(Ellis 1993: 8-9)



# BUILDING THE INTERFACE



# Overview

## Tasks

- Gathering and representing derivational paradigms in SW
- The normalisation of entry formats
- The normalisation of BT and DOE inflections

## Output

- The layout By\_headword\_A-I
- The layout By\_headword\_L-Y
- The layout By\_inflectional\_form
- The layout By\_POS\_tag\_YCOE

# Architecture

The IOED is a **relational lexical database** that links a given headword or inflection to its correlates in the other dictionaries filed in the database.

The IOED has a lemmatised component and an unlemmatised one. The **lemmatised component** copes with differences of headword format across dictionaries, while the **unlemmatised component** deals with divergent orthography in inflectional forms. In this sense, the makings of the relational database rely on both token analysis and type analysis because dictionary words (types) are considered along with textual attestations and frequencies (tokens).

The input to the IOED is comprised of the **dictionaries** as well as the **corpora** of Old English, including the DOEC, from which inflectional forms as well as textual frequencies have been extracted; as well as the YCOE, on which the IOED draws for inflectional forms and morphological tags.

# Software and tabulated data

The relational database has been implemented in **Claris FileMaker Pro** (version 19.4.2.204) and populated with files imported from **Microsoft Excel** or FileMaker (Update option).

Some dictionary files have required post-edition of the available versions with **TextEdit**, particularly CHM and SVV. The textual sources have undergone two types of procedure.

The DOEC has been gathered in a single .txt file, concorded with **AntConc** and registered in the database. A total of 200,659 fragments have been obtained, which correspond to a concordance of 3,016,937 lines and an index of 187,114 types (one form per attestation in the DOEC). The index is provided with textual frequencies (number of tokens of each type).

The POS files of the YCOE have been merged in one .txt file, from which inflectional forms and morphological tags have been extracted with **BBEdit**. This has turned out a total of 82,773 inflectional forms (tokens), disambiguated by means of the morphological tag, which includes lexical category.

# Gathering and representing derivational paradigms in SW

SW format	Normalisation	Meaning
fāc(e)n	fācen	'deceit'
~gecwiss	fācengecwiss	'conspiracy'
~dæð	fācendæð	'sin'
~full	fācenfull	'deceitful'
~lēas	fācenlēas	'guileless'
~lic	fācenlic	'deceitful'
~lice	fācenlice	'deceitfully'
~searo	fācensearo	'treachery'
~stafas	fācenstafas	'treachery'
~geswipere	fācengeswipere	'treachery'
~tācen	fācentācen	'deceitful token'

# The normalisation of entry formats 1

Dictionary	Headword	Normalisation	Meaning
CH	āhlin-ian	āhlinian	'to deliver'
CH	bealo-	bealo-	'bale, harm'
CH	+bearded	gebearded	'bearded'
CH	±bēatan	bēatan(ge)	'to beat'
CH	-cæfian	-cæfian	'to ornament'
CH	ed-winde	edwinde	'wind'
CH	efensārgung	efensārgung	'sympathy'
CH	eftonfōnd?	eftonfōnd?	'receiver'
CH	-ēgan	-ēgan	'to fear'
CH	feaxclāð	feaxclāð	'cap'
CH	lēctūn-	lēctūn-	'kitchen-garden'
SW	ā•bāedan	ābāedan	'to force out'
SW	abbod~dōm	abboddōm	'abbacy'
SW	abbod~isse	abbodisse	'abbess'
SW	ā•broþennes	ābroðennes	'cowardice'
SW	an(d)-byrd an	andbyrdan	'to strive'
SW	and-fęng~nes	andfengnes	'receptacle'
SW	and-giēt	andgiet	'intellect'

# The normalisation of entry formats 2

Dictionary	Headword	Normalisation	Meaning
SW	an(d)-mitta	andmitta	'scales'
SW	and-wæ̃t	andwæ̃t	'moist'
SW	be-tȳn-ing	betȳning	'conclusion'
SW	-briht	-briht	'bright'
SW	gebyrg	gebyrg	'protection'
SW	(ge)bytl-ung	gebytlung	'building'
BT	ÁDL	ād̃l	'disease'
BT	æfen-þéning	æfenðéning	'evening service'
BT	a-þeódan	aðeódan	'to disjoin'
BT	brád-nes	brādnes	'broadness'
BT	friþ-lice	friðlice	'freely'
BT	FÓÐER	fóðer	'fodder'
BT	ge-þeóðnes	geðeóðnes	'joining'
BT	ge-béd-ræden	gebēdræden	'office of praying'
BT	-elfen	-elfen	'of elves'
DOE	a-ȳpan	aȳðan	'to devastate'
DOE	āl-ge-weorc	ālgeweorc	'material for producing fire'
DOE	āgustus-mōnaþ	āgustusmōnað	'August'

# The normalisation of BT and DOE inflections

BT_inflectional_form	BT_headword	Normalisation	DOE_inflectional_form	DOE_headword
bearwes	BEARO	bearwes	bearwes	bearu
bebaðod	be-baðian	bebaðod	bebaþod	be-baþian
bebeorgan	be-beorgan	bebeorgan	bebeorgan	be-beorgan
bebeád	be-beódan	bebead	bebead	be-bēodan
bebirigde	be-birgan	bebirigde	bebirigde	be-byrgan
bebohte	be-bycgean	bebohte	bebohte	be-bicgan
bebudon	be-beódan	bebudon	bebudon	be-bēodan
bebycgean	be-bycgean	bebycgean	bebycgean	be-bicgan
bebyreþ	be-beran	bebyreþ	bebyreþ	be-beran
bebyride	be-byrian	bebyride	bebyride	be-byrgan
bebyrigede	be-byrgan	bebyrigede	bebyrigede	be-byrgan
bebyrignysse	be-byrignys	bebyrignysse	bebyrignysse	be-byrgnes
bebúgan	be-búgan	bebugan	bebugan	be-būgan
bebýt	be-beódan	bebyt	bebyt	be-bēodan
beceorfan	be-ceorfan	beceorfan	beceorfan	be-ceorfan
beceáþian	be-ceáþian	beceapian	beceapian	be-cēapian
beclemme	be-clemman	beclemme	beclemme	be-clemman
beclungen	be-clingan	beclungen	beclungen	be-clingan
beclyped	be-clyþian	beclyped	beclyped	be-cýþan
beclýsdon	be-clýsan	beclysdon	beclysdon	be-clýsan

# The layout By\_headword\_A-I

..! DOE_headword	DOE_headword_match_field	DOE_lexical_class	BT_headword	BT_lexical_class	CHM_headword	SW_headword
fild	fild	noun	fild	adjective	fild	fild
fild-cumb	fildcumb	noun	fild-cumb	noun	fildcumb	fild-cumb
finc	finc	noun	FINC	noun	finc	finc
findan	findan	verb	FINDAN	verb	findan	find an
finger	finger	noun	FINGER	noun	finger	finger
finn	finn	noun	Finn	noun	finn	finn
finol	finol	noun	FINOL	noun	finol	finol
finta	finta	noun	finta	noun	finta	finta
firen	firen	noun	FIREN	noun	firen	firen
firen-lust	firenlust	noun	firen-lust	noun	firenlust	firen-lust
fisc	fisc	noun	FISC	noun	fisc	fisc
fisc-cynn	fisc cynn	noun	fisc-cynn	noun	fisc cynn	fisc-cynn
fisc-hūs	fischūs	noun	fisc-hūs	noun	fischūs	fisc-hūs
fisc-mere	fisc mere	noun	fisc-mere	noun	fisc mere	fisc mere
fisc-pól	fisc pól	noun	fisc-pól	noun	fisc pól	fisc-pól
fiscere	fiscere	noun	fiscere	noun	fiscere	fisc-ere
fiscian	fiscian	verb	fiscian	verb	fiscian	fisc ian
fipelere	fiðelere	noun	fiðelere	noun	fiðelere	fipel-ere
fipelestre	fiðelestre	noun	fiðelestre	noun	fiðelestre	fipel-estre
fipere	fiðere	noun	fiðere	noun	fiðere	fiper e
flacor	flacor	adjective	flacor	adjective	flacor	flacor
fleardian	fleardian	verb	fleardian	verb	fleardian	fleard-ian
fleax	fleax	noun	FLEAX	noun	fleax	fleax
fleswian	fleswian	verb	fleswian	verb	fleswian	fleswian
flicce	flicce	noun	flicce	noun	flicce	flicce
flint	flint	noun	FLINT	noun	flint	flint
flint-græg	flintgræg	adjective	flint-græg	adjective	flintgræg	flint-græg
flocc	flocc	noun	FLOCC	noun	flocc	flocc
flocc-mælum	floccmælum	adverb	flocc-mælum	adverb	floccmælum	flocc-mælum
flot	flot	noun	flot	noun	flot	flot
flota	flota	noun	FLOTA	noun	flota	flota
floterian	floterian	verb	floterian	verb	floterian	floter ian
flotian	flotian	verb	flotian	verb	flotian	flotian
flugol	flugol	adjective	flugol	adjective	flugol	flugol
flyge	flyge	noun	flyge	noun	flyge	flyge
flyht	flyht	noun	flyht	noun	flyht	flyht



# The layout By\_headword\_L-Y

BT_lexical_class	BT_headword	BT_match_field	CHM_headword	SW_headword	+
verb	tó-hlídan	tōhlídan	tōhlídan	tō-hlídan	
noun	toh-líne	tohlíne	tohlíne	toh-líne	
verb	tó-hnescian	tōhnescian	tōhnescian	tō-hnescian	
noun	tó-hopa	tōhopa	tōhopa	tō-hop a	
noun	tó-hopung	tōhopung	tōhopung	tō-hop-ung	
verb	tó-hrēran	tōhrēran	tōhrēran	tō-hrēran	
noun	tohte	tohte	tohte	tohte	
verb	tó-hweorfan	tōhweorfan	tōhweorfan	tō-hweorfan	
noun	tó-hyht	tōhyht	tōhyht	tō-hyht	
noun	tól	tōl	tōl	tōl	
verb	tó-lætan	tōlætan	tōlætan	tō-læt an	
adverb	tolcendlice	tolcendlice	tolcendlice	tolc-endlice	
verb	tó-licgan	tōlicgan	tōlicgan	tō-licgan	
verb	tó-liðian	tōliðian	tōliðian	tō-liþian	
noun	toll	toll	toll	toll	
noun	toll-scír	tollscír	tollscír	toll-scír	
noun	toln	toln	toln	toln	
noun	tolnere	tolnere	tolnere	toln-ere	
adjective	tóm	tōm	tōm	tōm	
verb	tó-mearcian	tōmearcian	tōmearcian	tō-mearc ian	
preposition	tó-middes	tōmiddes	tōmiddes	tō-middes	
noun	tó-nama	tōnama	tōnama	tō-nam a	
verb	tó-nemnan	tōnemnan	tōnemnan	tō-nemnan	
verb	tonian	tonian	tonian	tonian	
noun	tord	tord	tord	tord	
noun	tord-wifel	tordwifel	tordwifel	tord-wifel	
verb	tó-rendan	tōrendan	tōrendan	tō-rendan	

# The layout By\_inflectional\_form

BT_inflectional_form	BT_headword	BT_inflections_match_field	DOE_inflectional_form	DOE_headword	+
antefn	antefn	antefn	antefn	antefn	
antimber	an-timber	antimber	antimber	an-timber	
antran	antre	antran	antran	antre	
anweorce	an-weorc	anweorce	anweorce	and-weorc	
anwlite	an-wlite	anwlite	anwlite	and-wlite	
anwlátan	an-wlæta	anwlatan	anwlatan	and-wlita	
anihst	a-nihst	anihst	anihst	a-nýhst	
anydde	a-nýdan	anydde	anydde	a-nýdan	
anyde	a-nýdan	anyde	anyde	a-nýdan	
apan	apa	apan	apan	apa	
aparod	a-parian	aparod	aparod	a-parian	
apinsode	a-pinsian	apinsode	apinsode	a-pinsian	
aplantod	a-plantian	aplantod	aplantod	a-plantian	
aplantode	a-plantian	aplantode	aplantode	a-plantian	
aplum	ÆPPEL	aplum	aplum	æppel	
apostatan	apostata	apostatan	apostatan	apostata	
apostol	apostol	apostol	apostol	apostol	
apostola	apostol	apostola	apostola	apostol	
apostolas	apostol	apostolas	apostolas	apostol	
apostolhád	apostol-hád	apostolhad	apostolhad	apostol-hád	
apostoli	apostol	apostoli	apostoli	apostol	
apostolum	apostol	apostolum	apostolum	apostol	
appla	ÆPPEL	appla	appla	æppel	
aprotanan	aprotane	aprotanan	aprotanan	aprotane	
apulder	apulder	apulder	apulder	apulder	
apuldorinde	apuldor-rind	apuldorinde	apuldorinde	apuldor-rind	
apuldre	apuldre	apuldre	apuldre	apulder	
apæce	a-pæcan	apæce	apæce	a-pæcan	
arc	arc	arc	arc	earc	
arce	arce-	arce	arce	earc	
arce-bisceop	arce-	arcebisceop	arcebisceop	arce-bisceop	
arcebisceoprice	arce-bisceop-ríce	arcebisceoprice	arcebisceoprice	arcebisceop-ríce	
arcebiscoprice	arce-bisceop-ríce	arcebiscoprice	arcebiscoprice	arcebisceop-ríce	
arcebysceop	arce-bisceop	arcebysceop	arcebysceop	arce-bisceop	
arcestóle	arce-stól	arcestole	arcestole	arce-stól	
archidiácones	arce-diácon	archidiacones	archidiacones	arce-diacon	
ardlice	ardlice	ardlice	ardlice	arodlice	

# The layout By\_POS\_tag\_YCOE

inflectional_form_YCOE	lexical_category_YCOE	morphological_tag_YCOE	inflectional_form_DOEC	inflectional_form_BT	textual_frequency_DOEC
folcgefehtum	noun	N^D	folcgefehtum	folcgefehtum	3
folcum	noun	N^D	folcum	folcúm	167
foldan	noun	N^D	foldan	foldan	182
folgerum	noun	N^D	folgerum	folgerum	10
foimum	noun	N^D	foimum	foimum	39
forannihte	noun	N^D	forannihte	foran-nihte	3
fore	noun	N^D	forę	fore	1130
forengan	noun	N^D	forengan	fórengan	23
foresceawunge	noun	N^D	foresceawunge	foresceawunge	34
forespræce	noun	N^D	forespræce	fórespræce	13
foręægednysse	noun	N^D	foręægednysse	foręægednysse	17
forhæfdnesse	noun	N^D	forhæfdnesse	forhæfdnesse	74
forhæfdnysse	noun	N^D	forhæfdnysse	forhæfdnysse	7
forhæfednesse	noun	N^D	forhæfednesse	forhæfednesse	22
forhæfednysse	noun	N^D	forhæfednysse	forhæfednysse	66
forligere	noun	N^D	forligere	forligere	18
forlore	noun	N^D	forlore	forlore	8
forscapunge	noun	N^D	forscapunge	forscapunge	2
forsewennysse	noun	N^D	forsewennysse	forsewennysse	24
forwyrde	noun	N^D	forwyrde	forwyrde	150
fostre	noun	N^D	fostre	fóstre	19
fotadle	noun	N^D	fotadle	fótadle	38
fotscamele	noun	N^D	fotscamele	fótscamele	3
fotum	noun	N^D	fotum	fótum	468
frætewum	noun	N^D	frætewum	frætewum	3
frætwwm	noun	N^D	frætwwm	frætwwm	24
fræcdnesse	noun	N^D	fræcdnesse	fræcdnesse	9
fræcdnysse	noun	N^D	fræcdnysse	fræcdnysse	44
fræcdnyssum	noun	N^D	fræcdnyssum	fræcdnyssum	9
fræcnesse	noun	N^D	fræcnesse	fræcnesse	17
fræme	noun	N^D	fræme	fræme	21
frædome	noun	N^D	frædome	fræódome	22
fræolsdæge	noun	N^D	fræolsdæge	fræólsdæge	34
fræolsdagum	noun	N^D	fræolsdagum	fræólsdagum	27
fræolse	noun	N^D	fræolse	fræólse	32
fræonde	noun	N^D	fræonde	fræónde	16
fræondscipe	noun	N^D	fræondscipe	fræóndscipe	75
fræondscype	noun	N^D	fræondscype	fræóndscype	33
fræondum	noun	N^D	fræondum	fræóndum	87
fríþe	noun	N^D	fríþe	fríþe	17
frófre	noun	N^D	frófre	frófre	190
fruman	noun	N^D	fruman	fruman	272
frumslæpe	noun	N^D	frumslæpe	frumslæpe	3
frymþe	noun	N^D	frymþe	frymþe	47

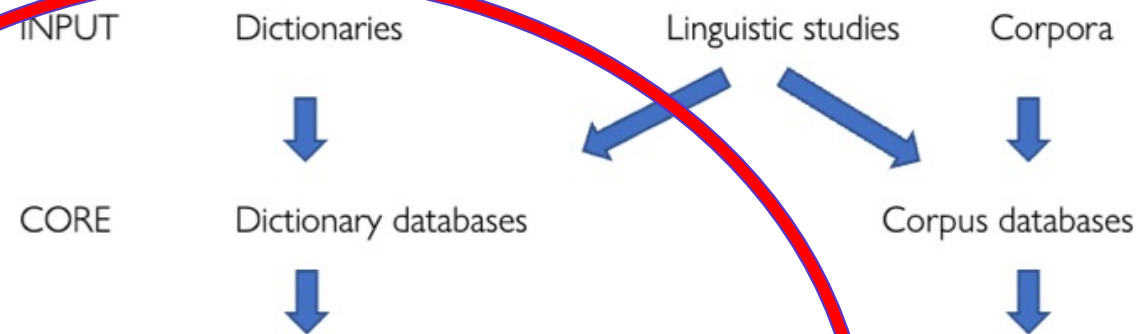
# TOWARD A KNOWLEDGE BASE



+

o

# KBOE: Relations as data



OUTPUT

Knowledge Base

Dictionary interface

Combined sources interface

Core relations: lemma-inflectional form, attestation in reference corpus, morphological tagging of inflectional form, syntactic annotation of inflectional form, lexical annotation of lemma, morphological relatedness of lemma, sources of lemma

Format: FileMaker (exportable as text and spreadsheet; online publication for browser)

# Edge lists: the lemmatised component

clifr-ung\_SW>>>clifrung\_CHM

+flāeschamod\_CHM>>>ge-flāeschamod\_BT

andsacian\_CHM>>>and-sac-ian\_SW

ymb-þreodung\_BT>>>ymbðreodung\_CHM

fæderen-bróðor\_BT>>>fæderenbróðor\_CHM>>>fæd(e)ren-brōþor\_SW

gebregd-stafas\_BT>>>+bregdstafas\_CHM>>>gebregd-stafas\_SW

inwit-gecynd\_BT>>>inwitgecynd\_CHM>>>in-wit-gecynd\_SW

be-smītan\_DOE>>>be-smítan\_BT>>>besmītan\_CHM>>>be-smītan\_SW

ēpel-turf\_DOE>>>éðel-turf\_BT>>>ēðelturf\_CHM>>>ēpel-turf\_SW

fīc\_DOE>>>FÍC\_BT>>>fīc\_CHM>>>fīc\_SW

# Edge lists: the unlemmatised component

abeþecige\_VBPS\_abeþecian\_BT>>>a-bedecian\_DOE

abroden\_VBN\_abredan\_BT>>>a-bregdan\_DOE

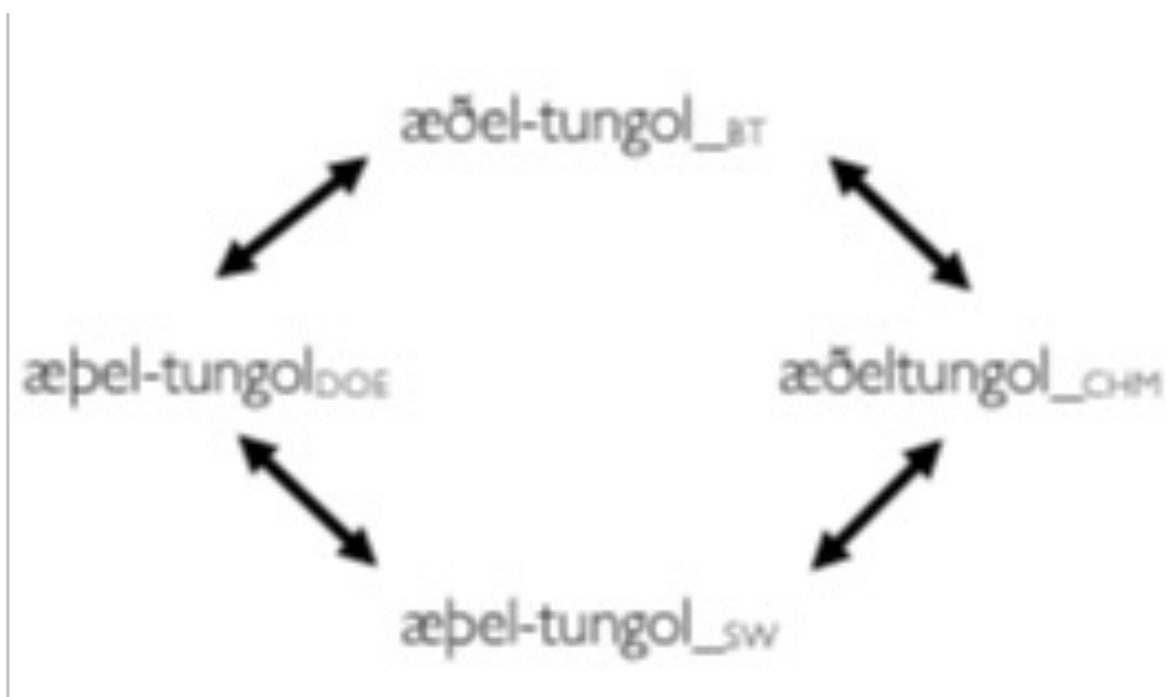
adylegode\_VBD\_adylegian\_BT>>>a-dīligian\_DOE

agæþ\_VBPI\_agæþ\_BT>>>a-gān\_DOE

ahsan\_N^A\_ahse\_BT>>>axe\_DOE

bedryda\_ADJ^N\_bedryda\_BT>>>bedd-rida\_DOE

# Knowledge graphs from edge lists 1



Closed undirected graph (headword to headword relation)



# Knowledge graphs from edge lists 2



Open directed graph (headword to inflectional form relation)

+  
○ ●

# EXPLOITING THE INTERFACE



# Dictionary review 1

The **DOE** has **83,642 inflectional forms**, **12,646 of which coincide with BT**. BT has 20,123 inflectional forms (mainly in in the segment A-I but also in the partial xml-annotated file of the L-Y segment kindly shared by Ondrej Tychy).

If we consider the segment A-Y and, consequently, exclude the DOE from the assessment, the comparison throws a total of **10,386 headwords shared by CHM, SW and BT**. This represents a relevant part of CHM and BT (28.2% in CHM and 28.7% in BT) and an even more substantial portion of SW (40%).

# Dictionary review 2

Taken **in chronologically-arranged pairs** based on the date of publication of the first edition of each dictionary:

17,377 headwords of SW have a counterpart in CHM (**68.4%**);

15,501 headwords of BT have a counterpart in CHM (**42.9%**);

12,548 headwords of BT have a counterpart in SW (**48.4%**).

In the segment A-I of the DOE, which comprises 15,981 headword entries:

4,272 headwords of the DOE coincide with CHM (which has 14,230 entries for A-I). This amounts to **30% of coincidence**.

3,386 headwords of the DOE coincide with SW (which has 11,218 entries for A-I). This amounts to **30.1% of coincidence**.

4,127 headwords of the DOE coincide with BT (which has 18,513 headword entries for A-I). This amounts to **22.29% of coincidence**.

# Dictionary review 3

All in all, the degree of coincidence in the letters A-I of the dictionaries is not very high. A maximum of 30%, between the DOE, on the one hand, and between CHM and SW, on the other hand. The lesson that can be learned from this is that **headword spelling makes a difference. The inventory of headword entries of the dictionaries under analysis show at least 70% divergence.** This is unparalleled in dictionaries of natural languages. Headword spelling differences, however, seem unlikely to account for all the divergence considered here. Further research should explain the role played by the accuracy of lemmatisation, the status of encyclopaedic entries and the introduction of ghost forms.

# Cross-dictionary and cross-corpus data searches: What is the counterpart of headword X in dictionary Y?

CHM headword	SW headword	BT headword	DOE headword
alordrenc	alor-dręnc	alor-drenc	alor-drenc
æteglan	æt-ęglan	æt-eglan	æt-eglan
befeallan	be-feall an	be-feallan	be-feallan
croh	croh	croh	crög
drencan	dręncan	DRENCAN	drencan
egeðe	ęgeþ e	egeðe	ecgan
formyrðrian	for-myrþr(i)an	for-myrþrian	for-myrþrian
+biddan	gebidd an	ge-biddan	ge-biddan
hwer	hwęr	hwer	hwær
innoð	innop	INNOP	innop

# Cross-dictionary and cross-corpus data searches: What is the headword spelling of the inflectional form X of dictionary Y in dictionary Z?

DOE inflectional form	BT inflectional form	DOE headword	BT headword
agane	agáne	a-gān	AGAN
æftersange	æfter-sange	æfter-sang	æfter-sang
befæstest	befæstest	be-fæstan	be-fæstan
cwarterne	cwarterne	cweartern	cweart-ern
drignesse	drignesse	drȳgnes	dryg-nes
eðelstaðolas	éðelstaðolas	ēpel-staþol	eðel-staðol
ferigend	ferigend	feriend	feriend
geandwyrðan	geandwyrðan	ge·and-wyrðan	ge-andwyrðan
hreðe	hréðe	hrape	HREDDAN
ingeþonc	ingeþonc	in-ge·þanc	in-geþanc

Cross-dictionary and cross-corpus data searches: What textual evidence for inflectional form of dictionary X is provided by dictionary Y?

YCOE	YCOE	BT	DOEC	DOEC
inflectional	morphological	inflectional	inflectional	textual
form	tag	form	form	frequency
gebræd	VBN^N	gebræd	gebræd	28
geceosan	VBPS	geceósan	geceosan	28
adla	N^G	ádla	adla	28
adrencte	VBN^N	adrencte	adrencte	28
fictreow	N^N	fictreów	fictreow	28
forwurde	VBD	forwurde	forwurde	28
fyrena	VBI	fyrena	fyrena	28
gebræd	VBI	gebræd	gebræd	28
gehælde	N	gehælde	gehælde	280
awrat	VBDI	awrát	awrat	282



# CONCLUSIONS AND FURTHER RESEARCH



+

o

# Some conclusions

Previous work has not assessed dictionaries as to headword spelling and inflectional forms. The conclusion can be drawn in this area that **the DOE makes a noticeable quantitative difference with respect to the dictionaries published at the turn of the 20th. century, which show a remarkable degree of convergence as to the selection of headword entries.** On the qualitative side, the DOE constitutes a major breakthrough in the field of Old English lexicography, which can be attributed to the introduction of **computational corpus compilation** and analysis into historical lexicography.

However, hypertext is the new print. It relies on 19th century lexicographical principles that maximise alphabetical semasiological arrangement with cross-references to related entries. It changes the means from paper to computer and, as a result, remarkably increases the searchability, but it does not modify the core design of the dictionary. The hypertext does not allow for extensive multivariable searches, for instance. Data cannot be easily published or share. The conclusion in this respect is that the hypertext format should be replaced with online databases that can be read by humans and interpreted by computers. The **knowledge base format**, in which graphs relate entities to relations and treat relations as data, enhances applicability to NLP and AI

# Further research

- What is the mirror of the Interface?
- Why do some headwords and inflections not coincide?
- Can the normalisation be improved so that more coincidences are found?



Thank you for your attention!!!

[www.nerthusproject.com](http://www.nerthusproject.com)