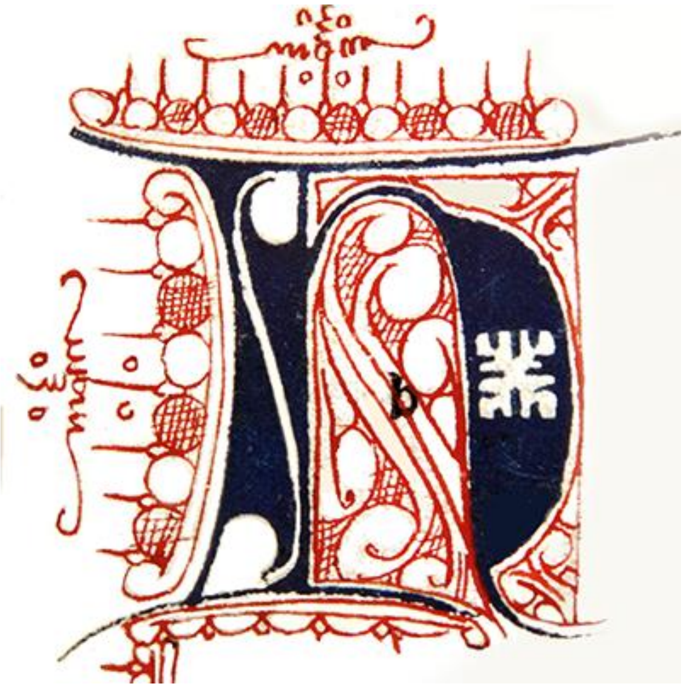
User Guide
Corpus Oudfries / Old Frisian Corpus

Corpus compiled and annotated by

Rita van de Poel, Leiden University

g.i.van.de.poel@hum.leidenuniv.nl



1 Introduction

1.1 Overview

This manual describes the corpus exploitation environment for the Corpus Oudfries (Old Frisian). The frontend application was developed by the Dutch Language Institute in Leiden (IvdNT). The application can be found at http://corpora.ato.ivdnt.org/corpus-frontend/OFR/search.
The application is a web-based frontend for the BlackLab search engine for corpora with token-based annotation.

The Corpus Oudfries/Old Frisian contains a large sample of the Old Frisian language from ca. 1200-1550, which has been tokenized, lemmatized and PoS-tagged by Rita van de Poel as part of her PhD research. Named entities (persons, locations) have also been identified and tagged. The corpus can be searched on three linguistic levels: words (as occurring in the text witnesses), lemmata (word forms linked to their dictionary headwords) and/or part-of-speech. The corpus has also been enriched with metadata (for instance, dialect, region, date) made accessible through filtering and grouping options.

1.2 DAH

The corpus Oudfries is founded on the "Database Altfriesisches Handwörterbuch" (DAH), which, in its turn, is the basis for the Old Frisian-Modern German dictionary: *Altfriesisches Handwörterbuch* (Heidelberg, 2008), edited by D. Hofmann and A. Popkema. The DAH is a MS Access database produced during the compilation of the dictionary in cooperation with CAU Kiel and the Fryske Akademy in Leeuwarden. The DAH has been made available by the creators for the purpose of this research. The lemmata in the corpus Oudfries are identical to the headwords in the DAH.

1.3 Corpus Size

At this point in time the corpus Old Frisian contains 235,462 tokens, 177 text witnesses from 11 manuscripts (114 distinct texts). The entire corpus of Old Frisian is estimated to be approximately 1,000,000 tokens.

1.4 Late Medieval Frisia

In the late Middle Ages the Frisian lands in the northern coastal regions in the Netherlands and Germany were only nominally ruled by the counts from Holland and Saxony. In practice the Frisians managed to ward off feudal overlords from exercising their power. The Frisian lands developed their own legal system in which the community of free men was responsible for judicial decisions. The legal system was based upon Germanic customary laws and relied very much on vengeance culture. Codifying their legal customs was not only a means of preserving their customs and rulings but also an important instrument in promoting their rights to self-governance. The vernacular language in which these customary laws were written down is Old Frisian, a West-Germanic language closely related to Old English. The textual witnesses in Old Frisian that remain today were written down between 1200 and 1550 and the texts that have been preserved are mostly legal in nature, with only very few samples of different genres.

The majority of Old Frisian texts are older than the manuscripts in which they can be found and the versions of the textual witnesses can differ markedly from each other. The largest part of the corpus of Old Frisian can be found in about 16 manuscripts (some of them being transcripts of original manuscripts). Most of the manuscripts have been transcribed and published and I have used these editions as the basis for my corpus.

If you are inexperienced in medieval Frisia and its language I highly recommend Rolf Bremmer's textbook, *An Introduction to Old Frisian. History, Grammar, Reader, Glossary* (Amsterdam, Philadelphia, 2009).

1.5 Acknowledgments

This website was developed by the Instituut voor de Nederlandse Taal, IvdNT, in Leiden. The corpus search engine is powered by BlackLab, an open source corpus retrieval engine allowing fast and complex searches on large volumes of annotated text. Special thanks go out to Jesse de Does at IvdNT for his invaluable assistance.

The corpus has been annotated in Cobalt, an application also developed by the IvdNT in which the complete lemmata list from the DAH has been pre-loaded. I would like to express my gratitude towards Anne Popkema for providing me with the DAH.

Finally, I want to thank Eduard Drenth of the Fryske Akademy for writing several scripts enabling me to transform the texts.

2 Corpus Design: Known Issues and Considerations

2.1 Named Entities

The PoS-tags PPRON (personal pronoun) and TOPO (toponym) have been used to distinguish word forms that indicate persons and locations. These lemmata (except the ones indicating countries) were not included in the DAH so new lemmata have been introduced in the annotation process.

2.2 First Lemma in Dictionary Entry

There are some consequences to using the dictionary/DAH as the corpus framework and these are important to keep in mind when using this search tool.

The corpus ONLY uses the first lemmata listed for an entry in the dictionary, which is very often the East Old Frisian. Other boldfaced words are not considered separate lemmata in the corpus, but variants that can be analysed in other ways. Hence, the dictionary entry ”**afte, āfte**^{WL}**, āft**^{WL}**, ēft-**^{WL}**, echt**^{WL} n.” yielded one lemma in the corpus (the first word) while the other boldfaced words are variants.

2.3 Homonyms

Homonyms were separate entries in the dictionary and hence also separate lemmata in the corpus. For instance, Old Frisian “*afte*” can either be a noun, an adjective or an adverb. In the dictionary the separate entries are distinguished with a superscripted numeral, while in the corpus these are represented with an underscore followed by a numeral AFTER the lemma (^{1}**afte** becomes “afte_1” and ^{2}**afte** becomes “afte_2”). This is an important design feature to keep in mind when searching on lemma level! You could use wildcards, like for instance “*afte**” to obtain results for all lemmata starting with the string “*afte*”.

2.4 Variants

All words in the corpus texts have been lemmatised, i.e. linked to a dictionary headword, here referred to as lemma, but in corpus linguistics also known as type. In practice this means that all morphological and orthographical variants have been connected to its headword or lemma. This feature is a major advantage in analysing medieval corpora (esp. on a semantic level) that are notorious for its complex and diverse orthography (through the eyes of a modern reader). In this way there is no need to take into consideration all variants of a lemma. For example, a relatively simple word like child; OFris “*kind”,* has a total of 36 variants in the Old Frisian corpus: *kynden, kind, kind, kijnden, kijnd, kyndes, kindes, kinder, kyndis, kyndena, kindis, kindan, kint, kinda, kinde, kinden, kiinden, kindem, kindar, kindum, kijnde, kijndena, kijndes, kijndis, kynde, kijndt, kijnt, kin-der, kindt, khinda, kiinderen, kijndem, kyn, kyndt, kindera, kynt.*

2.5 Multiwords

Spelling variants with a hyphen (see “*kin-der*” in the previous section) indicates the merging of a so-called multiword. In the manuscript “*kin der*” was the original spelling. However, a space cannot easily be used in corpus analyses (this would indicate token division), so it was substituted by a hyphen.

Another problematic issue while annotating the texts was the fact that Old Frisian texts contain many instances of word forms that have been merged into one word either because of scribal error, orthographical variation or cliticization. For example, “*hine”* (*hi* + *ne*), ”*hit*” (*hi* + *hit*), “*thetter*” (*thet* + *ther*), “*meyma*” (*mey* + *ma*) and “*hwersama*” (*hwer* + *sa* + *ma*). In the annotation tool that was used there was no way to split these tokens into several lemmata, so either a new lemma was introduced (i.e. “*hwersama*” -> “*hwersama*”) or the word form was linked to one of the lemmata (i.e. “*meyma*” -> “*muga*”), thus effectively losing one lemma.

2.6 PoS-tag Simplification

Rita van de Poel’s PhD research is mainly focused on lexical semantic patterns in the Old Frisian language and less on purely morphological and syntactical patterns. In the first version of the annotated corpus Oudfries therefore, some less strict tagging procedures have been used in order to save effort without compromising the possibilities of the PhD research.

You will often see homonyms (with different PoS-tags) being combined into the same lemma.

For example:

- Three dictionary entries for “*afte*” (noun, adjective and adverb) are reflected by just two separate lemmata in the corpus: noun (“*afte_1*”) and adjective (“*afte_2*”). The adverbial meaning was combined with the adjective and hence tagged with the “*afte_2*” lemma.
- The PoS-tag “X” is used as an indication that lemmata with different PoS-tags have been simplified into one lemma. For instance, the dictionary has two headwords “*sâ*”: conjunction and adverb. All occurring word forms have been combined to one lemma “*sâ*”, with the PoS-tag “X”.

2.7 Determiners and Pronouns

All morphological indications of case, gender and number have been standardized.

Determiners: The definite article has been standardized to the masculine, first person form: “*thî*”. For the indefinite article “*ên*” has been used. Please note, that the tagging does not differentiate between the use of “*ên*” as determiner or numeral.

Demonstrative forms have also been simplified to the masculine nominative case.

Personal pronouns have been lemmatized as the first person lemma form, retaining gender and number, but also, tagging the genitive forms separately. Hence, *hî*, *hiu*, *hit*, *hia* and *sîn*, *hire*, *sîn*, *hiara*.

Resolving (some of) the above-mentioned suboptimal tagging and inconsistencies in the annotation process will probably be taken care of in a new version of the corpus. Please help us improve the data and report any inconsistencies or errors you may come across: g.i.van.de.poel@hum.leidenuniv.nl

3 Metadata

3.1 List of PoS-tags

Each lemma has been provided with a PoS-tag: Part of speech label (syntactical category).

Abbreviation	Part of speech
NOUN	Noun
VRB	Verb
ADJ	Adjective
ADV	Adverb
AUX	Auxiliary verb
CONJ	Conjunction
SCONJ	Subordinating conjunction
CCONJ	Coordinating conjunction
DET	Determiner
ADP	Adposition
PRON	Pronoun
NUM	Numeral
INT	Interjection
PPRON	Proper noun
TOPO	Topography
X	Undefined or a combination of lemmata

3.2 List of Manuscrips

MS sigil	Title	Region	Dialect	From	To
R1	First Rüstringen Manuscript	Rüstringen	East Old Frisian	1250	1300
H1	First Hunsingo Manuscript	Hunsingo	East Old Frisian	1300	1350
H2	Second Hunsingo Manuscript	Hunsingo	East Old Frisian	1300	1350
R2	Second Rüstringen Manuscript	Rüstringen	East Old Frisian	1327	1327
B2	Second Brokmer Manuscript	Brokmerland	East Old Frisian	1345	1345
E1	First Emsingo Manuscript	Emsingo	East Old Frisian	1375	1400
F	Fivelgo Manuscript	Fivelgo	East Old Frisian	1427	1450
Bas	Baseler Codex	West Lauwers	West Old Frisian	1440	1475
E3	Third Emsingo Manuscript	Emsingo	East Old Frisian	1440	1460
E2	Second Emsingo Manuscript	Emsingo	East Old Frisian	1450	1475
Ro	Codex Roorda	West Lauwers	West Old Frisian	1490	1504
J	Jus Municipale Frisonum	West Lauwers	West Old Frisian	1520	1540

3.3 List of Text Witnesses

Text sigil	Text	MS sigil	Edition	From	To	Category
Adm	Adam's Creation	E1	OTR 4: V 21			biblical
Afr	Afrethe	R2	OTR 8: VI			regional law
Asg	Asega Law	F	OTR 12: XIX	1200	1250	oaths/formulas
Aug	Gestation of a Foetus	E1	OTR 4: V 19			compensation
		E3	OTR 10: I 199			compensation
		F	OTR 12: XII 1-2			compensation
BAg	General Register of Compensations	E1	OTR 4: VI	1000	1100	compensation
		H2	OTR 6: XI	1000	1100	compensation
		R1	OTR 11: V, XIV 3-8	1000	1100	compensation
BasI	Wedding Speeches I	Bas	Buma1957	1440	1450	sermon
BasII	Wedding Speeches II	Bas	Buma1957	1440	1450	sermon
BasIII	Wedding Speeches III	Bas	Buma1957	1440	1450	sermon
BEm	Emsingo Book of Compensations	E1	OTR 4: VII, X 5-8	1250	1400	compensation
		E2	OTR 7: III	1250	1400	compensation
		E3	OTR 10: I 1-198, 200-287	1250	1400	compensation
BEmRP	Compensation Register of Emsingo	E3	OTR 10: II 12-23	1250	1400	compensation
BFDg	Compensation Register Ferwerdera- and Dongeradeel	J	J XXIII			compensation
BFi	Compensation Register Fivelgo	F	OTR 12: 4, 10-39	1086	1165	compensation
BGr	Compensation Register South West	J	J XXIV			compensation
BHm	Compensation Register Hemmen	J	J XXV			compensation
Bhua	Compensation Register Hunsingo a	H2	OTR 6: VII 1-105			compensation
Bhub	Compensation Register Hunsingo b	H2	OTR 6: VII 112-168			compensation
Bhuc	Compensation Register Hunsingo c	H2	OTR 6: IX			compensation
Bir	Interregional Compensations Register	J	J XXVII	1200	1280	compensation
BKJ	Kampa Jeldric's Compensations Register	F	OTR 12: XI			compensation
BLw	Compensation Register Leeuwarderadeel	J	J XXIX			compensation
Bod	Ten Commandments	H2	OTR 6: XIII			biblical
		Ro	Jagersma I			biblical
BPr	Compensations for Killing a Man of the Church	E1	OTR 4: V 20	1150	1250	compensation
		H2	OTR 6: XII 2	1150	1250	compensation
BPRa	Compensations for Killing a Man of the Church	R1	OTR 11: XII			compensation

Text sigil	Text	MS sigil	Edition	From	To	Category
BPRb	Compensations for Killing a Man of the Church	R1	OTR 11: XVIII			compensation
BrBa	Brocmonna Bref a	B2	OTR 5: 1-77, 212-220	1200	1300	regional law
BrbB	Compensation Register Brokmerland	B2	OTR 5: 183-211	1200	1300	compensation
Bro	The Riddle of the Three Brothers	E1	OTR 4: V 15-17			land law
		E3	OTR 10: III			land law
		H2	OTR 6: V 2-4			land law
		Ro	Jagersma III 50: 46-48			land law
BRu	Compensation Register Rustringen	R1	OTR 11: VI			compensation
		R2	OTR 8: I			compensation
Bsk	Bishop's Treaty	E2	OTR 7: VIII	1276	1276	land law
BW5D	Compensation Register Wonseradeel + Vijf Delen	J	J XXVIII			compensation
BWb	Compensation Register Wymbritseradeel	J	J XXI 1-117			compensation
Bwda	Compensation Register Fan Walddedum I	J	J XXI 118-120			compensation
BWdb	Compensation Register Fan Walddedum II	J	J XXVI			compensation
CrCr	Minor Old Frisian Chronicle	J	J XXX	1248	1464	historiographical
Dad	On werjeld	F	OTR 12: XIV			compensation
Dem	Emsinger Dooms 1312	E2	OTR 7: VI	1312	1312	regional law
		E3	OTR 10: IV	1312	1312	regional law
Dom	Eight Dooms	F	OTR 12: VII	1200	1276	land law
		J	J X	1200	1276	land law
EAt	Oath Attha	J	J XXXVIII			oaths/formulas
EEh	Oath circular judge	J	J XXXVII			oaths/formulas
EFia	The Fia-eth; The Property Oath	E1	OTR 4: I			oaths/formulas
		E2	OTR 7: I, II, A1			oaths/formulas
EFo	Oath Church guardian	J	J XXXIX			oaths/formulas
EFre	Peace Oath	J	J XLI			oaths/formulas
Egmg	Oath judge South West	J	J XXXVI			oaths/formulas
EGmw	Oath judge Wymbritseradeel	J	J XXXV			oaths/formulas
ELed	Oath feud leader	J	J XL			oaths/formulas
Epla	Epilogue to the 17 Statutes	E1	OTR 4: III 19			pan-frisian
		F	OTR 12: III 23			pan-frisian
		J	J VI 18			pan-frisian
		R1	OTR 11: IV 25			pan-frisian
Eplb	Epilogue to the 24 Landlaws	E1	OTR 4: V 6			pan-frisian

Text sigil	Text	MS sigil	Edition	From	To	Category
		H2	OTR 6: III 26			pan-frisian
ETo	Oath witnesses	J	J XLII			oaths/formulas
FdB	Court session	J	J XII			oaths/formulas
FdBW	Peace judge Wymbritseradeel	J	J XLIIII			oaths/formulas
FnWa	Fana en Widukin	J	J XVII 40-41			regional law
Fri	Poem on Frisian Freedom (Fon alra Fresena fridome)	H2	OTR 6: XIV	1200	1250	historiographical
Fro	Marital property	F	OTR 12: XV			land law
		H2	OTR 6: VII 106-111			land law
FtT	Fifteen Signs for Doomsday	R1	OTR 11: XI			biblical
GoF	Crimes against God	J	J XVI 18			canon law
GoH	Permission to Break into a Church	R1	OTR 11: XIII			canon law
Hav	Three Capital Crimes	E1	OTR 4: V 18	1312	1312	land law
	Three Capital Crimes	H2	OTR 6: V 1	1312	1312	pan-frisian
HRt	Haet is riucht? What is law?	F	OTR 12: I			pan-frisian
		J	J II			pan-frisian
		Ro	Jagersma I 1-7			pan-frisian
JF	Jurisprudentia Frisica	Ro	Jagersma III			canon law
K17	Seventeen Statutes	E1	OTR 4: III 1-17	1000	1100	pan-frisian
		F	OTR 12: III 1-17	1000	1100	pan-frisian
		H2	OTR 6: II	1000	1100	pan-frisian
		J	J VI	1000	1100	pan-frisian
		R1	OTR 11: III	1000	1100	pan-frisian
Kap	Fon Kap: civil law	F	OTR 12: XIII			regional law
KaR	Legend of Charlemagne and Redbad	J	J IV			historiographical
Kei	Five Keys to Wisdom	H2	OTR 6: IV			biblical
KFiO	Interregional statutes Fivelgo and Oldambt	F	OTR 12: XXI			land law
KHu	Statutes of Hunsingo 1252	F	OTR 12: XXII	1252	1252	land law
		H2	OTR 6: XXIII	1252	1252	land law
KKm	Commentary on the Seventeen Statutes	R2	OTR 8: II			pan-frisian
Krua	Statutes of Rustringen	R1	OTR 11: VIII	1100	1200	regional law
Krub	Additional Statutes of Rustringen	R1	OTR 11: IX	1175	1225	regional law
L24	Twenty-four Landlaws	E1	OTR 4: IV, V 11-13, VIII 17-20, X 1-4	1086	1200	pan-frisian
		F	OTR 12: IV	1086	1200	pan-frisian
		H2	OTR 6: III 1-24	1086	1200	pan-frisian

Text sigil	Text	MS sigil	Edition	From	To	Category
		J	J VIII	1086	1200	pan-frisian
		R1	OTR 11: IV 1-24	1086	1200	pan-frisian
Lad	Oath of innocence	J	J VXII 33-39			oaths/formulas
LaFi	Fon Lawum, Inheritance	F	OTR 12: XX			regional law
		Ro	Jagersma III 50:35-40, 50:42-44			regional law
LaK	Fon Lawum, three Statutes	F	OTR 12: XVI			regional law
		Ro	Jagersma III 41			regional law
Lav	Inheritance rules from Emsingo	E1	OTR 4: VIII 1-5			regional law
		E2	OTR 7: IV			regional law
		H2	OTR 6: VII 169-172			regional law
LKm	Commentary on the Twentyfour Landlaws	R2	OTR 8: III			pan-frisian
Lon	Instructions for judges	R2	OTR 8: VIII			procedures
Mgn	Magnus Legend	F	OTR 12: V	1200	1300	pan-frisian
		J	J V	1200	1300	pan-frisian
Mor	Homocide	E1	OTR 4: VIII 35-37			land law
		E2	OTR 7: V			land law
		E3	OTR 10: II 1-11			land law
		F	OTR 12: XII 3, 5-8			land law
		H2	OTR 6: III 25			land law
		R1	OTR 11: XIV 1-2, 9A-B			land law
Mrk	Coins Rustringen	R1	OTR 11: XVI			regional law
		R2	OTR 8: IV			regional law
MrWl	Market Law	J	J XIV	1165	1240	regional law
Ned	Lawful Impediments	E1	OTR 4: V 7-10			procedures
		E2	OTR 7: VII			procedures
Pap	Papena Ponten: Statutes Wymbritseradeel, 1404	J	J XXXIV	1404	1404	regional law
Pay	West Frisian coins	J	J XXII			land law
Pen	Coining privilege	F	OTR 12: XVII 76			regional law
PnB	Book of Debts	B2	OTR 5: 78-182	1200	1250	regional law
		E2	OTR 7: IX	1200	1250	regional law
		E3	OTR 10: V	1200	1250	regional law
Prla	Prologue (a) to the 17 Statutes and 24 Landlaws	E1	OTR 4: II	1220	1250	pan-frisian
		F	OTR 12: II	1220	1250	pan-frisian

Text sigil	Text	MS sigil	Edition	From	To	Category
		H1	OTR 6: I	1220	1250	pan-frisian
		R1	OTR 11: I	1220	1250	pan-frisian
Prlb	Prologue (b) to the 17 Statutes and 24 Landlaws	J	J VII	1220	1250	pan-frisian
		R1	OTR 11: II	1220	1250	pan-frisian
RFi	Regional laws from Fivelgo	F	OTR 12: XVII 1-71	1086	1165	land law
RgJ	Register to Jus Municipale Frisonum	J	J I			misc.
Rom	How the Frisian conquered Rome	J	J XIX			historiographical
Rrua	This is also Frisian law a	R1	OTR 11: X			regional law
Rrub	This is also Frisian law b	R1	OTR 11: XV			regional law
Rruc	This is also Frisian law c	R2	OTR 8: V			regional law
Rrud	Collection of Regional Laws	R2	OTR 8: VII	1275	1327	regional law
Rud	Book of Rudolf	J	J XVIII	1220	1300	land law
SBw	Canon Law Bolsward	J	J XXXIII	1377	1450	canon law
Sib	On Killing a Relative	R1	OTR 11: XVII			canon law
Ska	Skakraf: robbery	J	J XVI 1-17			land law
SkRa	Elder Skeltenariucht	J	J III 1-50, 53-81			land law
		Ro	Jagersma II			land law
SkRb	Younger Skeltenariucht	J	J III 51-52, XIII	1165	1280	land law
Spr	How to Accuse a Thief	H2	OTR 6: XII 1			oaths/formulas
SRu	Synod Laws from Rustringen	R1	OTR 11: XX	1200	1230	canon law
SWl	General Western Frisia Synod Laws	F	OTR 12: VI, VIII	1000	1050	canon law
		J	J IX	1000	1050	canon law
SwS	Swarte Sweng: Unforgivable Crimes	J	J XV	1250	1350	land law
Thf	Orphan girl can save a thief	F	OTR 12: IX			land law
Ups	Statutes of the Upstalbam	E1	OTR 4: III 18			pan-frisian
Urk	Superior Statutes	E1	OTR 4: IX	1165	1230	pan-frisian
		F	OTR 12: XVIII	1165	1230	pan-frisian
		H2	OTR 6: VIII	1165	1230	pan-frisian
Urt	Peace for Homocide	J	J XX	1165	1230	oaths/formulas
W5D	Statutes of the Vijf Delen	J	J XVII 1-32	1200	1400	regional law
We16	Exceptions to the Seventeenth Statute	E1	OTR 4: VIII 6-10	1165	1230	pan-frisian
		F	OTR 12: XVII 72-75	1165	1230	pan-frisian
		H2	OTR 6: VI	1165	1230	pan-frisian
We17	Exceptions to the Sixteenth Statute	E1	OTR 4: V 1-3, V 14, VIII 6-10, VIII 12-16	1100	1200	pan-frisian

Text sigil	Text	MS sigil	Edition	From	To	Category
		F	OTR 12: III 18-22	1100	1200	pan-frisian
		H2	OTR 6: X	1100	1200	pan-frisian
		J	J XI	1100	1200	pan-frisian
		R1	OTR 11: VII	1100	1200	pan-frisian
Wif	Causes of Miscarriage	E1	OTR 4: VIII 11			regional law
		F	OTR 12: XII 9			regional law
		R1	OTR 11: XIV 9C			regional law
Wit	Oathswearing relics	J	J XIII 1E			oaths/formulas
WUps	Upstalboom Statutes, 1323	Ro	Jagersma IV	1323	1323	pan-frisian
Wwe	Road obstruction	R1	OTR 11: XIV 10			regional law
ZFr	Desacration of a Church	R1	OTR 11: XIX			regional law

3.4 List of Editions

Listed below are the editions I used as source texts for my corpus. All editions, except J, are diplomatic.

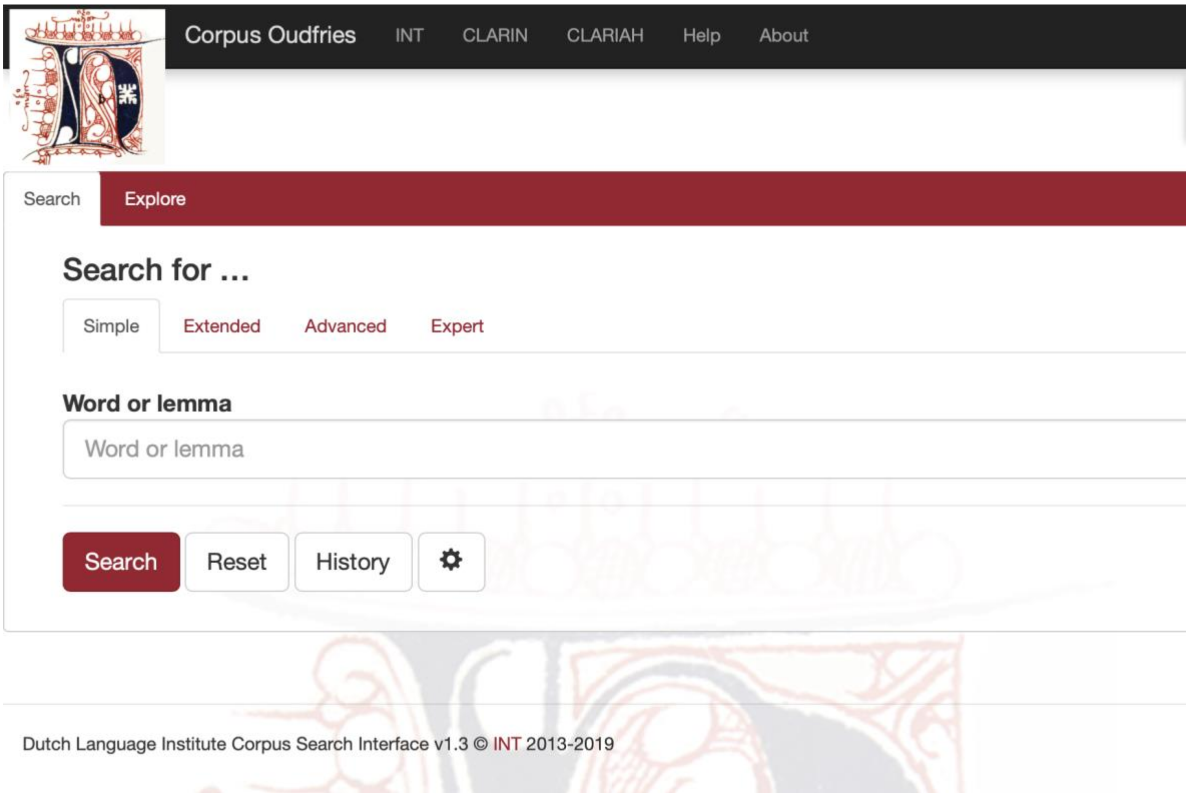
Abbreviation	Editor & Title
Buma1957	Buma, Wybren Jan, ed. Aldfryske houlikstaspraken (Assen, 1957).
J	Buma, Wybren Jan, Wilhelm Ebel and Martina Tragter-Schubert ed., Westerlauwerssches Recht I-II, Jus Municipale Frisonum, Altfriesische Rechtsquellen 6 (Göttingen, 1977).
Jagersma	Jagersma, Bram, Unpublished Diplomatic Edition of Codex Roorda, 2003
OTR 4	Sipma, P., ed. De eerste Emsinger codex, Oudfriesche taal- en rechtsbronnen 4 (Den Haag, 1943).
OTR 5	Buma, Wybren Jan, ed. Die Brokmer Rechtshandschriften, Oudfriesche taal- en rechtsbronnen 5 (Den Haag, 1949).
OTR 6	Hoekstra, J., De eerste en de tweede Hunsinger Codex, Oudfriesche taal- en rechtsbronnen 6 (1950).
OTR 7	Fokkema, K., ed. De tweede Emsinger codex, Oudfriesche taal- en rechtsbronnen 7 (Den Haag, 1953).
OTR 8	Buma, Wybren Jan, ed. Het tweede Rüstringer handschrift, Oudfriesche taal- en rechtsbronnen 8 (Den Haag, 1954).
OTR 10	Fokkema, K., ed. De derde Emsinger codex, Oudfriesche taal- en rechtsbronnen 10 (Den Haag, 1959).
OTR 11	Buma, Wybren Jan, ed. De eerste Riustringer codex, Oudfriesche taal- en rechtsbronnen 11 (The Hague, 1961).

OTR 12	Sjölin, B., ed. Die Fivelgoer Handschrift, Oudfriesche taal- en rechtsbronnen 12 (Den Haag, 1970).

4 Basics

4.1 Getting Started

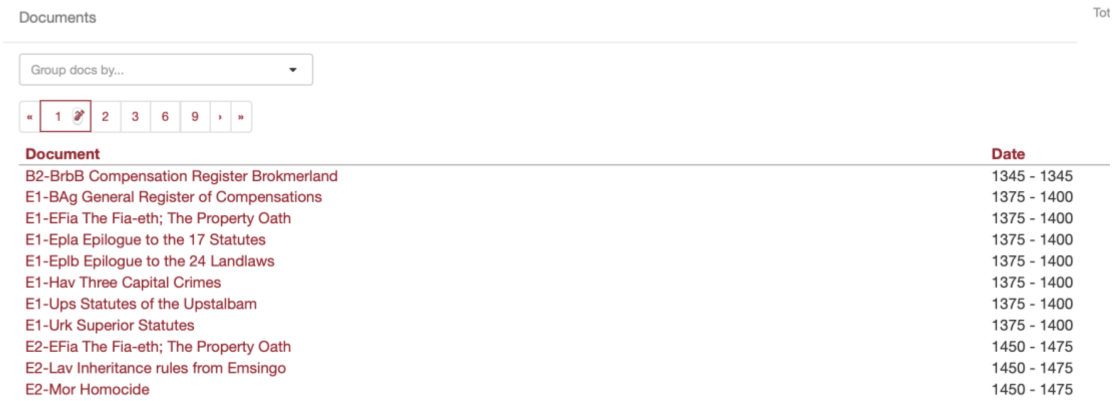
In the following chapters the different features of the corpus search application OudFries/Old Frisian will be clarified.



Dutch Language Institute Corpus Search Interface v1.3 © INT 2013-2019

As the homepage for the corpus shows (tabs in red) the application has two main functions: “Search” (with four options) and “Explore” (with three options) and these will be discussed in the next two chapters. The default opening screen is the “Smple search”. Furthermore, the Viewing, Filtering and Grouping options will be illustrated. First, some basic operations of this website will be explained.

The default opening screen is the Simple Search form. When you click on Search without entering a word form the application will display all documents/text witnesses in the corpus.



Document	Date
B2-BrbB Compensation Register Brokmerland	1345 - 1345
E1-BAg General Register of Compensations	1375 - 1400
E1-EFia The Fia-eth; The Property Oath	1375 - 1400
E1-Epla Epilogue to the 17 Statutes	1375 - 1400
E1-Eplb Epilogue to the 24 Landlaws	1375 - 1400
E1-Hav Three Capital Crimes	1375 - 1400
E1-Ups Statutes of the Upstalbam	1375 - 1400
E1-Urk Superior Statutes	1375 - 1400
E2-EFia The Fia-eth; The Property Oath	1450 - 1475
E2-Lav Inheritance rules from Emsingo	1450 - 1475
E2-Mor Homocide	1450 - 1475

Click on any of the hits/documents to open a new tab in your browser with more elaborate document information. See the chapter “Information on Documents/Text Witnesses”.

4.1.1 Reset

You can start a new search by pressing the “Reset” button. By doing so, both the search query and the hits found will be cleared. Your search history, however, will remain unchanged.

4.1.2 History

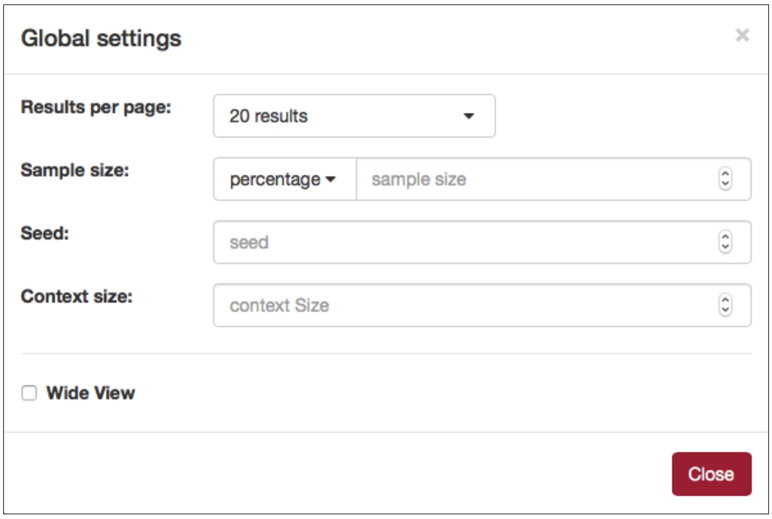
The “History” button will display your query history. Per search query there are several possibilities (as shown in the screenshot below): you can again perform the search (“Search”), you can copy the results as a link (“Copy as link”), you can download the results as a file (“Download as file”) or you can delete your search (“Delete”).

Every search query has its own URL. If you copy this URL from your browser, you can send it to someone else who can import this link via “Import from a link”. It offers that person the possibility to run the search on their own computer.

4.1.3 Global Settings

The “Global settings” dialogue, activated by pressing the wheel button ⚙ , allows you to configure five settings: “Results per page”, “Sample size”, “Seed”, “Context size” and “Wide View”.

- *Results per page*: you can choose whether you want 20, 50, 100 or 200 results to be shown;
- *Sample size:* selecting a value here will instruct the search engine to return a random sample drawn from the complete result set. The sample size can be limited by a percentage of the total number of search results (percentage) the number of results displayed (count);
- *Seed:* a “random seed” is a number used to initialize a so-called pseudo-random number generator. Keeping the same seed will ensure that two samples drawn from the same result set are identical. A new seed will most likely result in a different sample.
- *Context size*: by enter a number to determine the number of words “Before hit” and “After hit”;
- *Wide View:* the default setting is “wide view”; you can change to small view by unticking the checkbox.



Global settings

Results per page: 20 results

Sample size: percentage | sample size

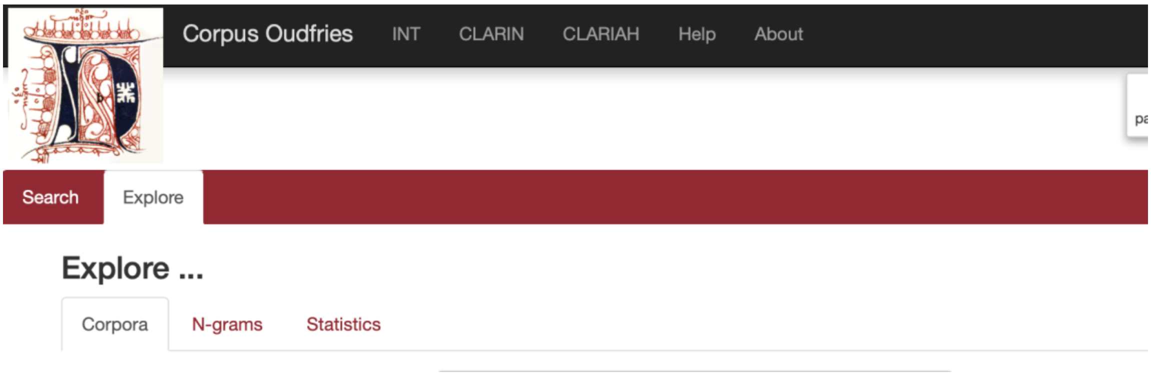
Seed: seed

Context size: context Size

□ Wide View

5 Exploring the Corpus

The “Explore” options help you to gain insight into the corpus and its characteristics and can be useful before starting your analyses of the language. The Explore tab has three subdivisions: “Corpora”, “N-grams” and “Statistics”.



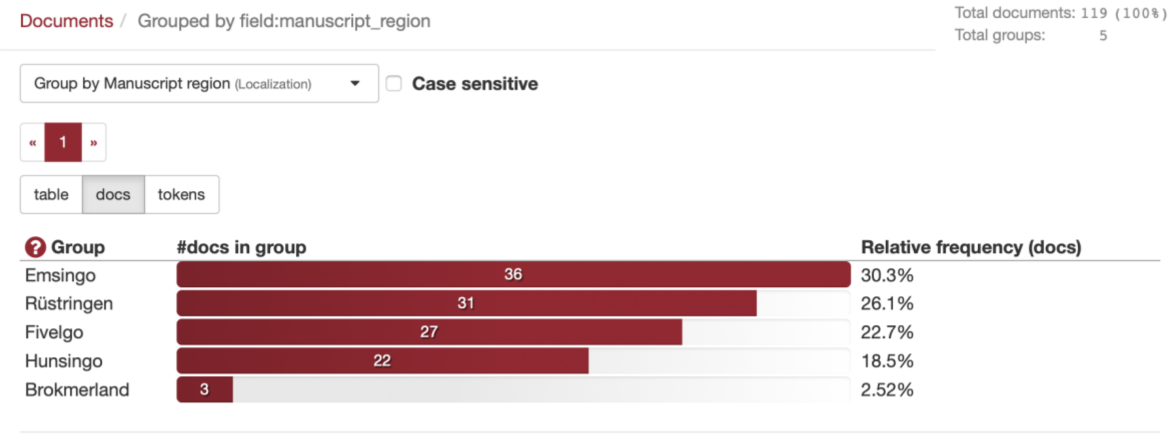
5.1 Corpora

This tab allows you to investigate the corpus composition. It consists of a small form to specify a grouping criterion, a metadata search form and a grouped result display.

For example, suppose we want to obtain information about the manuscripts provenances within the East Old Frisian subcorpus/dialect group.

- ○ In the “Group documents by metadata” drop-down menu, choose “Group by Manuscript region”
- ○ In “Show groups as”, select “docs”
- ○ In the metadata search form (“Filter search by”), select from the Localization tab: Manuscript dialect = East Old Frisian.
- ○ Click “Search”

You will see these results:



Documents / Grouped by field:manuscript_region

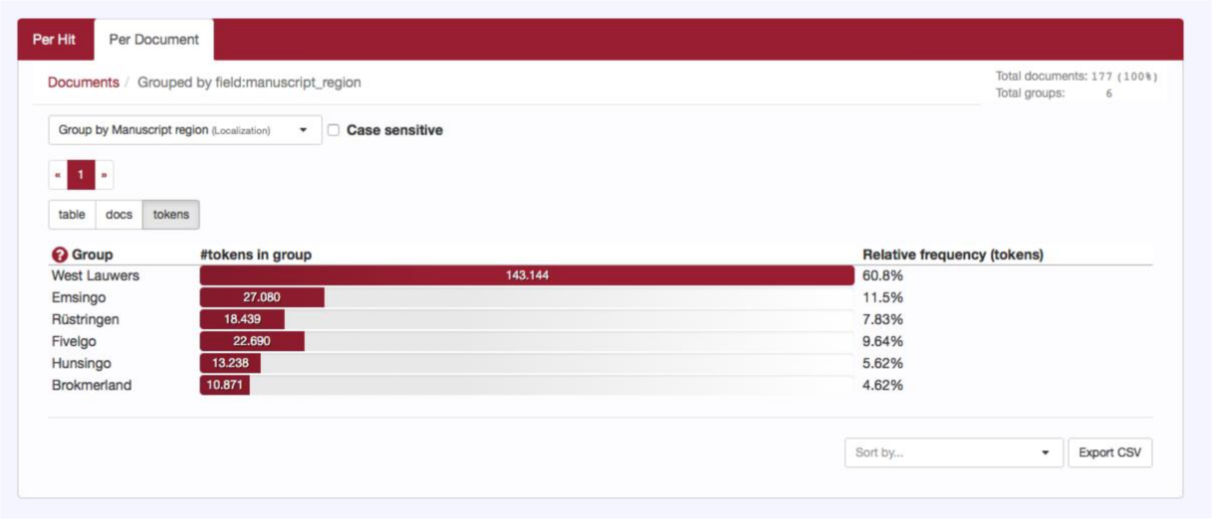
Total documents: 119 (100%)
Total groups: 5

Group by Manuscript region (Localization)

☐ Case sensitive

Group	#docs in group	Relative frequency (docs)
Emsingo	36	30.3%
Rüstringen	31	26.1%
Fivelgo	27	22.7%
Hunsingo	22	18.5%
Brokmerland	3	2.52%

Another example: Group the results by Manuscript regio and you will notice that about 67% of the corpus consists of words from the West Lauwers region, while 33% is from east of the Lauwers river.



Documents / Grouped by field:manuscript_region

Total documents: 177 (100%)
Total groups: 6

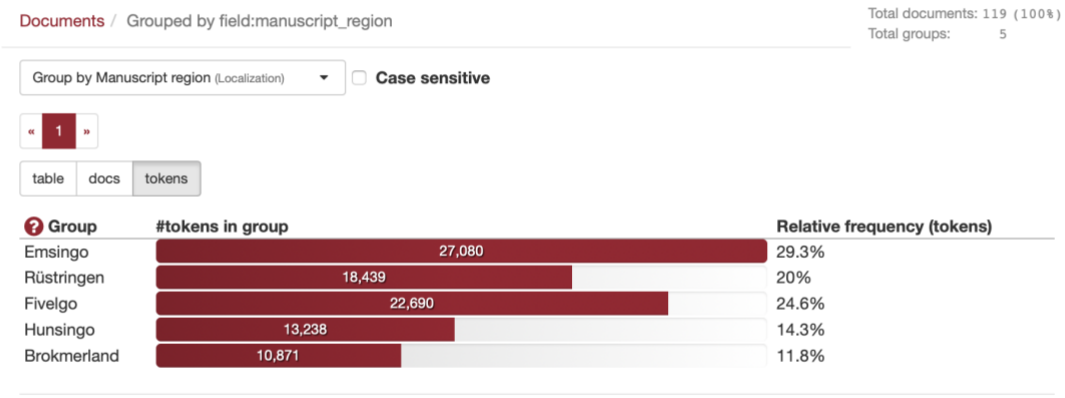
Group by Manuscript region (Localization)

□ Case sensitive

Group	#tokens in group	Relative frequency (tokens)
West Lauwers	143.144	60.8%
Emsingo	27.080	11.5%
Rüstringen	18.439	7.83%
Fivelgo	22.690	9.64%
Hunsingo	13.238	5.62%
Brokmerland	10.871	4.62%

5.2 Token View

The token view will also represent the infromation in a vertical bar chart in which the bars displays the number of tokens each filter variable (in this case: region) contains.



Documents / Grouped by field:manuscript_region

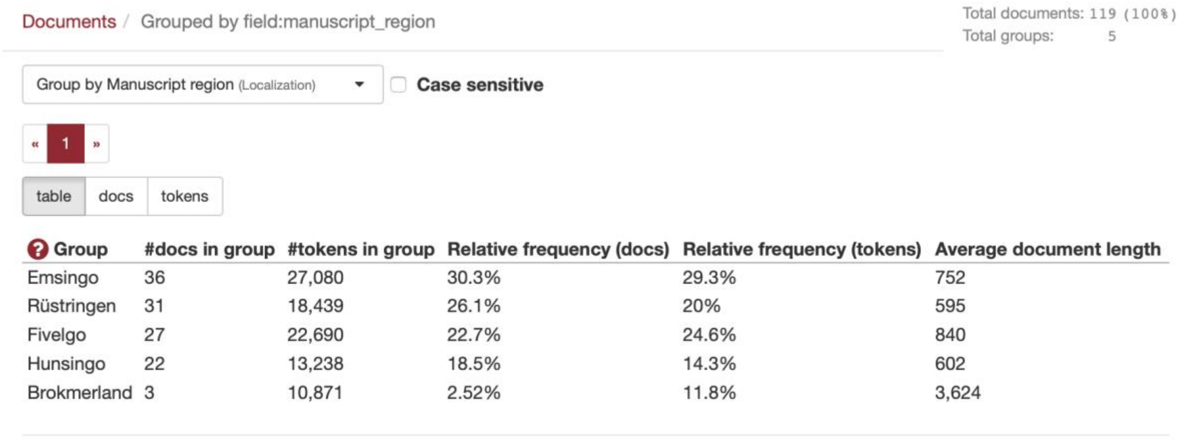
Total documents: 119 (100%)
Total groups: 5

Group by Manuscript region (Localization)

□ Case sensitive

Group	#tokens in group	Relative frequency (tokens)
Emsingo	27,080	29.3%
Rüstringen	18,439	20%
Fivelgo	22,690	24.6%
Hunsingo	13,238	14.3%
Brokmerland	10,871	11.8%

5.3 Table View



Documents / Grouped by field:manuscript_region

Total documents: 119 (100%)
Total groups: 5

Group by Manuscript region (Localization)

□ Case sensitive

Group	#docs in group	#tokens in group	Relative frequency (docs)	Relative frequency (tokens)	Average document length
Emsingo	36	27,080	30.3%	29.3%	752
Rüstringen	31	18,439	26.1%	20%	595
Fivelgo	27	22,690	22.7%	24.6%	840
Hunsingo	22	13,238	18.5%	14.3%	602
Brokmerland	3	10,871	2.52%	11.8%	3,624

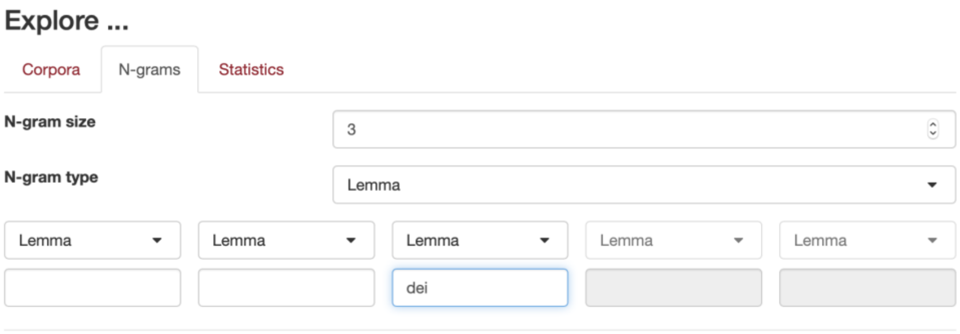
5.4 N-grams

An ***n*-gram** is a contiguous sequence of a particular number (*n*) of items from a given sample of text. This option will list the frequency of different N-grams in a (sub)corpus.

Options:

- "N-gram size": the length of the sequence (number from 1 to 5). Bigram and trigrams are most often used in corpus liguistics.
- "N-gram-type": choose for sequences of word form, lemma, part of speech tag.
- Restrict to N-grams with some slots already specified.
- Subcorpus selection by metadata filtering

Example: Give all trigrams with the lemma "*dei*" in the third position.



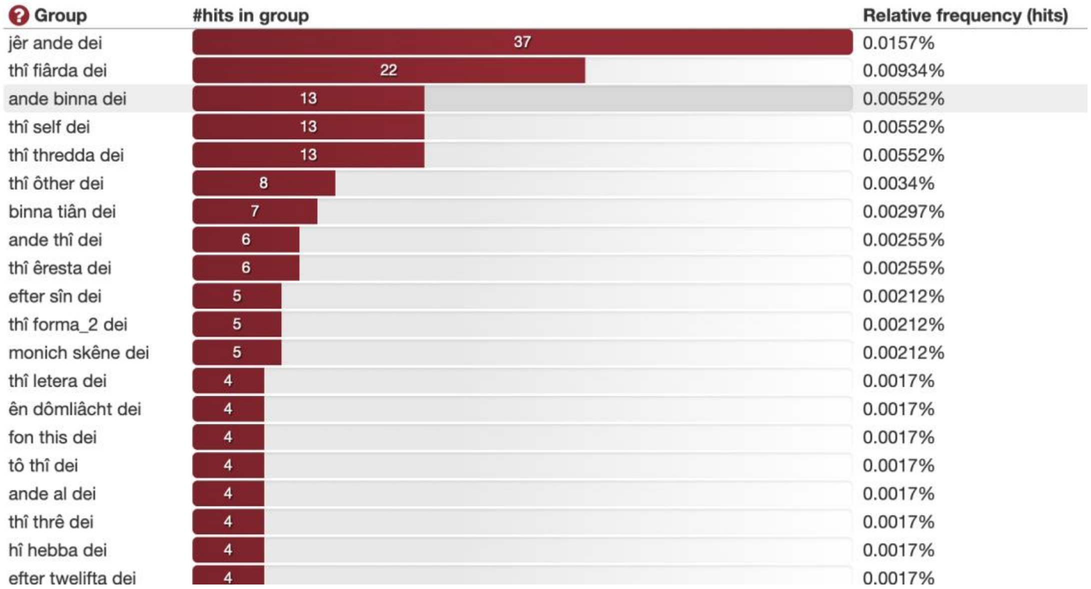
Explore ...

N-gram size: 3

N-gram type: Lemma

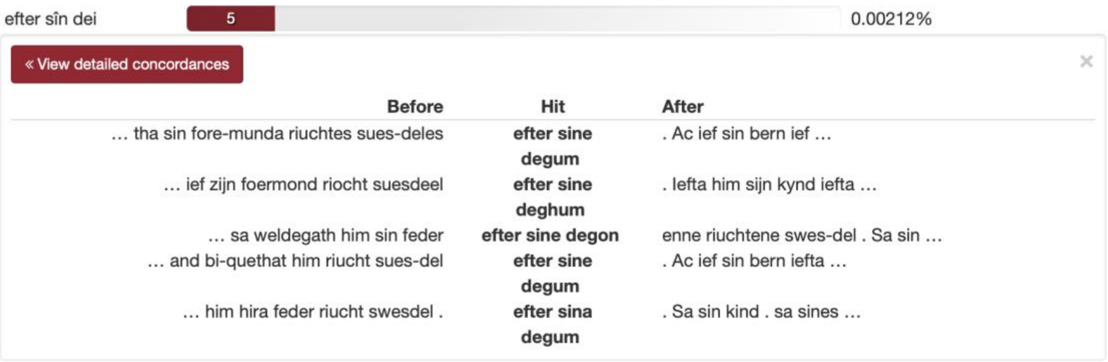
Lemma	Lemma	Lemma	Lemma	Lemma
		dei		

Results:



Group	#hits in group	Relative frequency (hits)
jêr ande dei	37	0.0157%
thî fiârda dei	22	0.00934%
ande binna dei	13	0.00552%
thî self dei	13	0.00552%
thî thredda dei	13	0.00552%
thî ôther dei	8	0.0034%
binna tiân dei	7	0.00297%
ande thî dei	6	0.00255%
thî êresta dei	6	0.00255%
efter sîn dei	5	0.00212%
thî forma_2 dei	5	0.00212%
monich skêne dei	5	0.00212%
thî letera dei	4	0.0017%
ên dômliâcht dei	4	0.0017%
fon this dei	4	0.0017%
tô thî dei	4	0.0017%
ande al dei	4	0.0017%
thî thrê dei	4	0.0017%
hî hebba dei	4	0.0017%
efter twelifta dei	4	0.0017%

These results can be further analysed on the level of an indivdual hit by clicking on one the bars. Clicking on the bar for "*efter sîn dei*", literally "after his days" the occurrences of this phrase in the corpus are being presented.

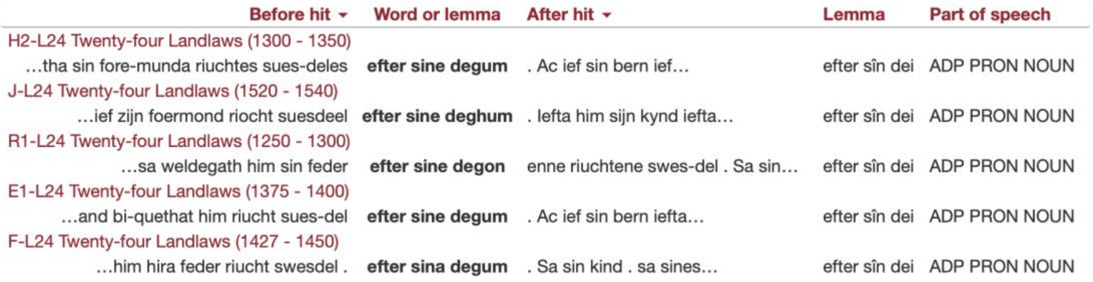


efter sîn dei 5 0.00212%

« View detailed concordances

Before	Hit	After
… tha sin fore-munda riuchtes sues-deles	efter sine degum	. Ac ief sin bern ief …
… ief zijn foermond riocht suesdeel	efter sine deghum	. Iefta him sijn kynd iefta …
… sa weldegath him sin feder	efter sine degon	enne riuchtene swes-del . Sa sin …
… and bi-quethat him riucht sues-del	efter sine degum	. Ac ief sin bern iefta …
… him hira feder riucht swesdel .	efter sina degum	. Sa sin kind . sa sines …

When clicking on "View detailed concordances" you will see the specific hits for "*efter sîn dei*" (after his death) in the regular "View results" layout. This specific phrase only occurs in five renditions (text witnesses) of the same text in the corpus.



Before hit	Word or lemma	After hit	Lemma	Part of speech
H2-L24 Twenty-four Landlaws (1300 - 1350)				
…tha sin fore-munda riuchtes sues-deles	efter sine degum	. Ac ief sin bern ief…	efter sîn dei	ADP PRON NOUN
J-L24 Twenty-four Landlaws (1520 - 1540)				
…ief zijn foermond riocht suesdeel	efter sine deghum	. Iefta him sijn kynd iefta…	efter sîn dei	ADP PRON NOUN
R1-L24 Twenty-four Landlaws (1250 - 1300)				
…sa weldegath him sin feder	efter sine degon	enne riuchtene swes-del . Sa sin…	efter sîn dei	ADP PRON NOUN
E1-L24 Twenty-four Landlaws (1375 - 1400)				
…and bi-quethat him riucht sues-del	efter sine degum	. Ac ief sin bern iefta…	efter sîn dei	ADP PRON NOUN
F-L24 Twenty-four Landlaws (1427 - 1450)				
…him hira feder riucht swesdel .	efter sina degum	. Sa sin kind . sa sines…	efter sîn dei	ADP PRON NOUN

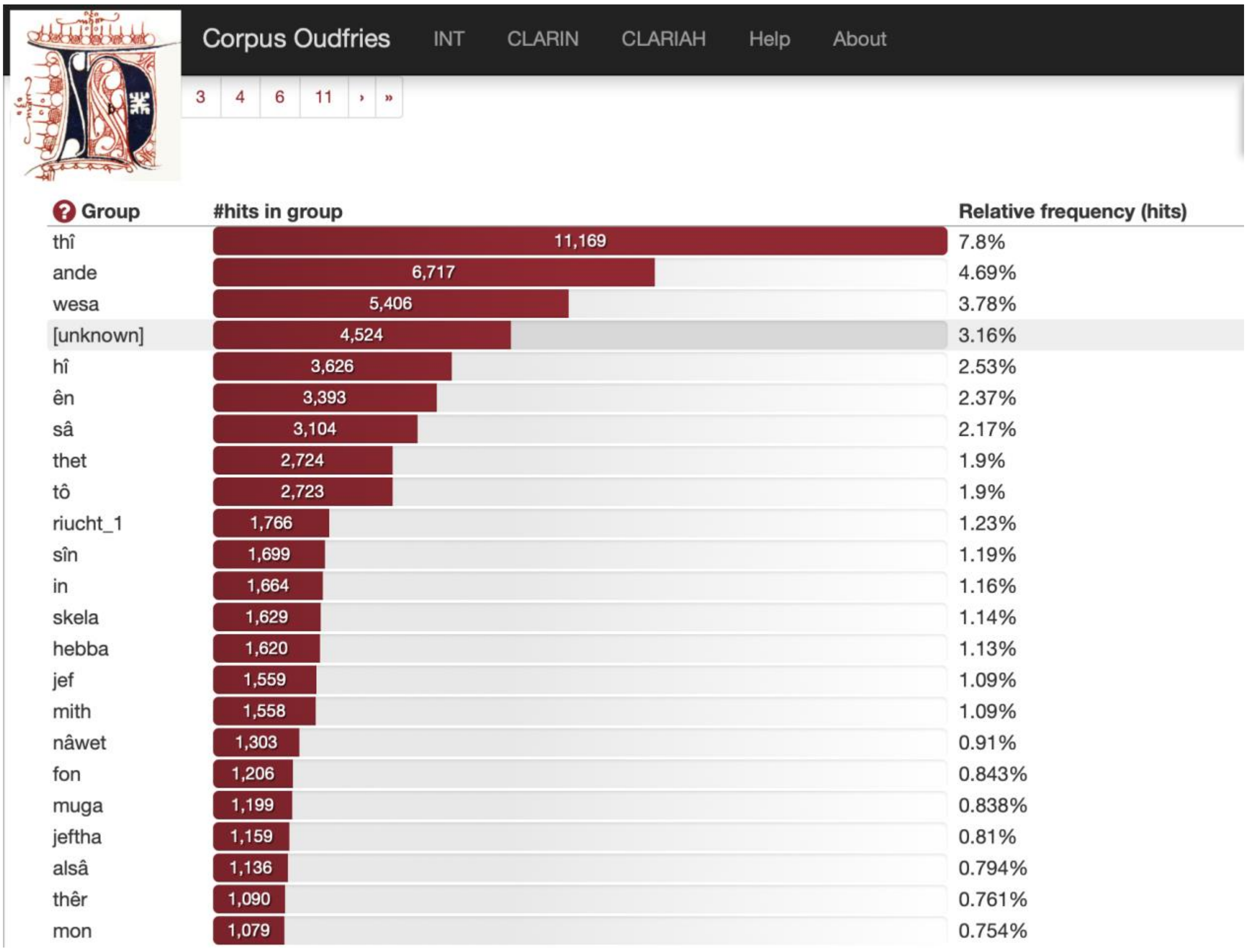
5.5 Statistics (Frequency Lists)

Here you can produce frequency lists for the entire corpus or for a subcorpus.

Options:

- "Frequency list type": choose for lists of word form, lemma or part of speech tag.
- Subcorpus selection by metadata filtering

The frequency list for the entire corpus displays many function words (determiners, conjunctions, auxiliary verbs) that are always more frequent than other words. The most striking high frequency words in this screenshot are: "*riucht_1*" and *"mon"*.

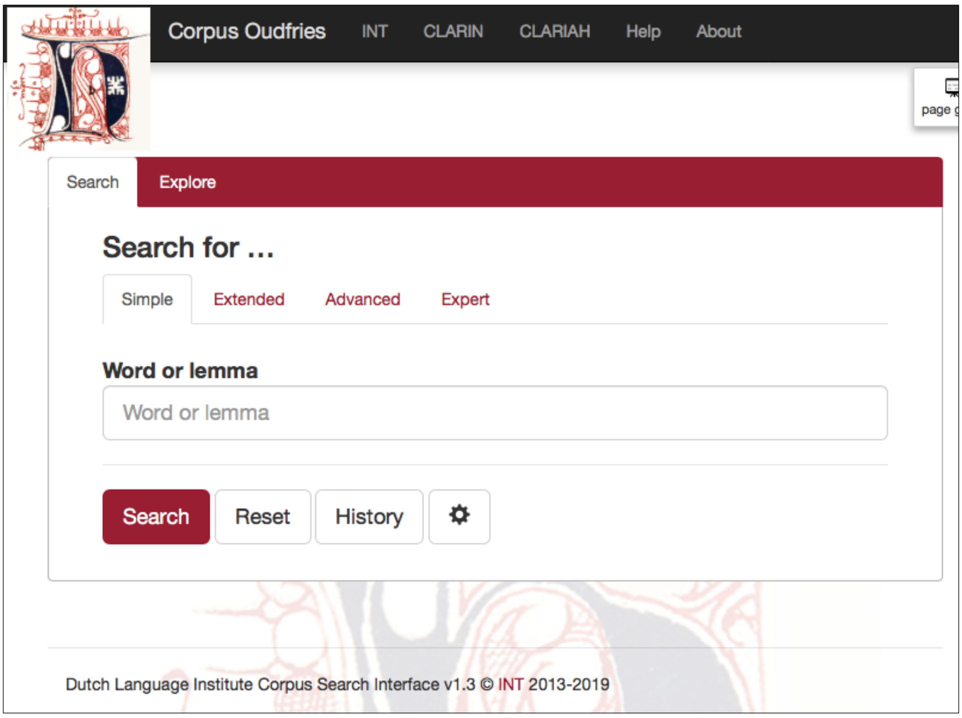


Group	#hits in group	Relative frequency (hits)
thî	11,169	7.8%
ande	6,717	4.69%
wesa	5,406	3.78%
[unknown]	4,524	3.16%
hî	3,626	2.53%
ên	3,393	2.37%
sâ	3,104	2.17%
thet	2,724	1.9%
tô	2,723	1.9%
riucht_1	1,766	1.23%
sîn	1,699	1.19%
in	1,664	1.16%
skela	1,629	1.14%
hebba	1,620	1.13%
jef	1,559	1.09%
mith	1,558	1.09%
nâwet	1,303	0.91%
fon	1,206	0.843%
muga	1,199	0.838%
jeftha	1,159	0.81%
alsâ	1,136	0.794%
thêr	1,090	0.761%
mon	1,079	0.754%

6 Searching the Corpus

The website has four Search options (four tabs in the window), ranging from elementary to elaborate: Simple, Extended, Advanced and Expert.

6.1 Simple Search



Dutch Language Institute Corpus Search Interface v1.3 © INT 2013-2019

The “Simple search” allows you to quickly search for specific lemmata and/or word forms. It is also possible to enter a phrase: *“fri fresa”*or *”fri fresa is”.* Note that in “Simple search” the patterns will be matched case-insensitively: *”fri fresa”* will deliver the same results as *”fri Fresa”* or *”FRI FRESA”.* If you wish to search case-sensitively, you will have to use “Extended search”.

6.1.1 Wildcards

A wildcard is a symbol used to replace or represent one or more characters. The following two wildcards are supported:

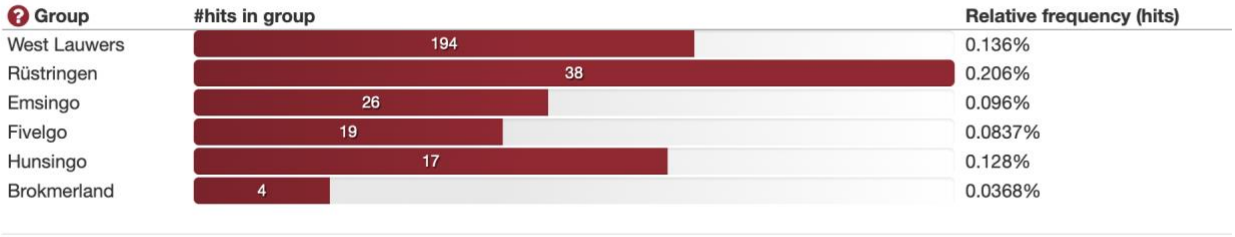
* The asterisk matches any character zero or more times. Therefore, a*b matches all values (words or lemmata) that start with an *a* and end with a *s*, e.g. *””, ”als”, “augustus”* and *“alderlas”.*

? The question mark matches a single character once. Therefore, *a?s* matches *only* three-letter word forms starting with “a” and ending with “s”. This wildcard can be used more than once, in several positions in a search string. Thus *“a???s”* matches *“alles”, “aynes”* and *“aftes”*.

Note that searching with wildcards is limited to “Simple search” and “Extended search”. In “Advanced search” and “Expert search” you can use so-called *regular expressions (Regex)* instead of wildcards.

6.1.2 Relative Frequency

Example: Search for lemma=”*frî*” and Group by Region will give these results:



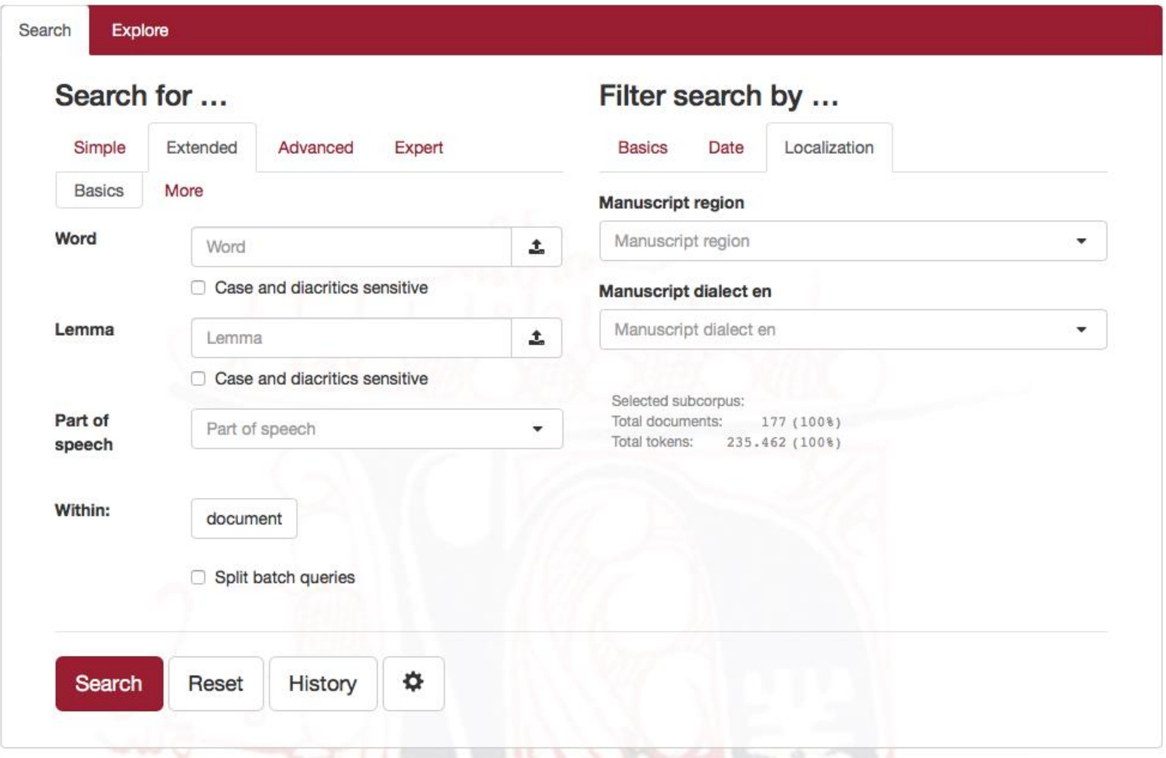
Group	#hits in group	Relative frequency (hits)
West Lauwers	194	0.136%
Rüstringen	38	0.206%
Emsingo	26	0.096%
Fivelgo	19	0.0837%
Hunsingo	17	0.128%
Brokmerland	4	0.0368%

The region West Lauwers has 194 instances of the lemma “*frî*”, while the region Rüstringen merely has 38 instances (also known as “raw frequencies” in corpus linguistics). However, the bar in the chart for Rüstringen is considerably longer than the West Lauwers bar. The bars represent the relative frequency; which takes into account the number of words in the texts from the regions. The chance of the lemma “*frî*” being found in a large text is obviously higher than in a small text. Use the “Explore” function to delve into the corpus design and metadata.

6.1.3 Duplicates

An important issue with the results from a “Simple Search” can be the duplicate hits when the word form or lemma searched for is both found as lemmata or as word forms in the texts. For instance, searching for “fri” will give several duplicate result because it is both a lemma (dictionary entry) and a word form.
Therefore, I recommend using the Extended Search.

6.2 Extended Search



Selected subcorpus:
Total documents: 177 (100%)
Total tokens: 235.462 (100%)

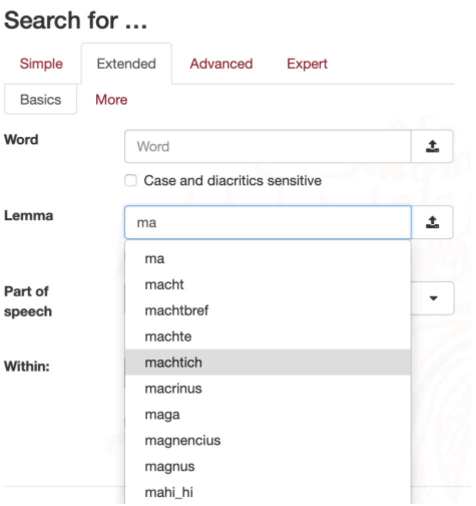
6.2.1 Basics Tab

The three linguistic attributes you can search for in queries are:

- “Word” (word form),
- “Lemma” (dictionary entry)
- “Part of Speech” (syntactical category)

All supported attributes are shown in the “Basics” search form (see the above screenshot). The “More” tab will display another search form, but these options are not implemented in the corpus at this time.

In the search fields “Word”, “Lemma” and/or “Part of Speech” enter the value of the attributes you are looking for. Then press Enter or click the “Search” button below to execute the search and view the results. You will see that while typing a query the application will display a dropdown list with possible lemmata.



Search for ...

Word: Word

☐ Case and diacritics sensitive

Lemma: ma

- ma
- macht
- machtbref
- machte
- machtich
- macrinus
- maga
- magnencius
- magnus
- mahi_hi

Part of speech

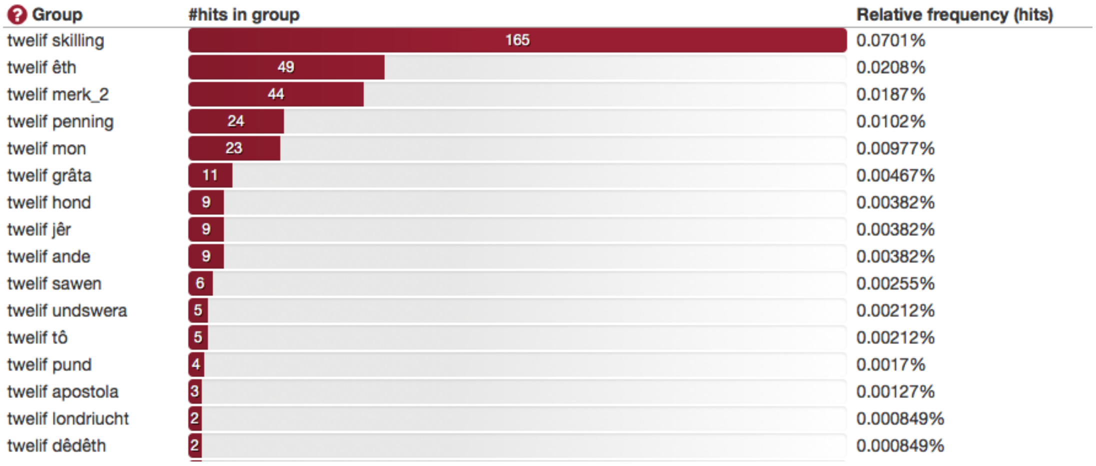
Within:

The default setting for search is case and diacritics insensitive. For example, *“sin”* will result in all occurrences of both *“sin”* (sense, mind, meaning) and *“sîn”* (his). By ticking the box “Case and diacritics sensitive” and typing in “*sîn*” you will only find hits in which *“sîn”* appears.

In the “Extended search” wildcards are supported. (See for a short explanation of wildcards “Simple Search”.)

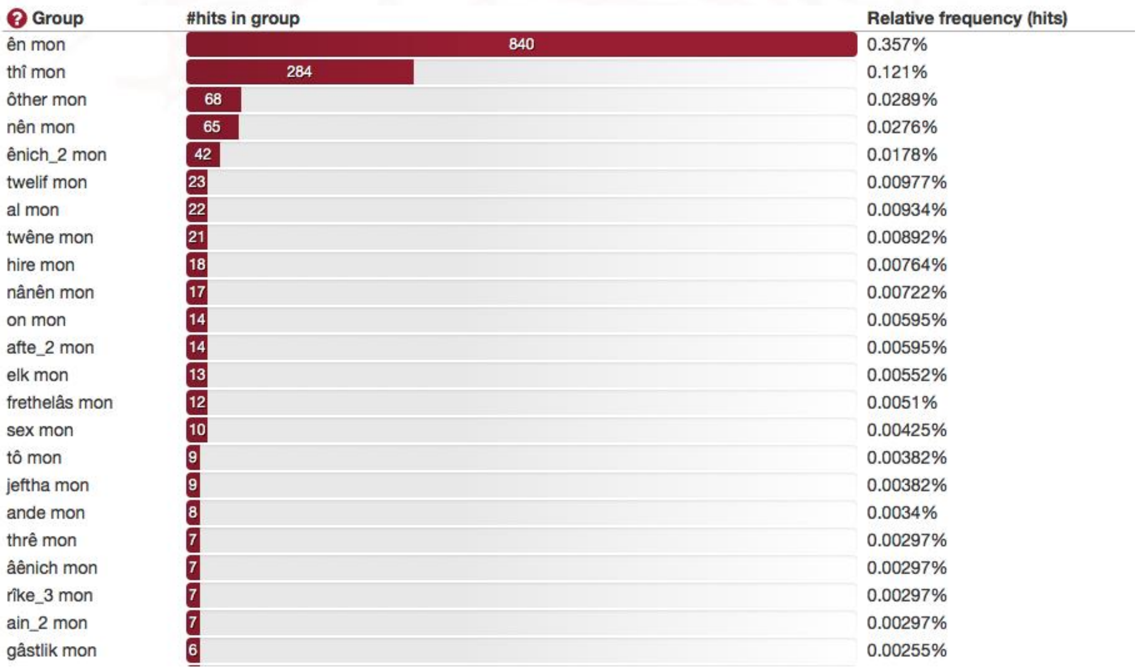
Please note (as also explained previously) that there is an important difference between the search fields “Word” and “Lemma”. For example, entering the value *“feder”* in “Word” will only provide you with occurrences of that exact string of characters which will almost always give less results than when you enter *“feder”* in the search field “Lemma”. You will, besides the word *“feder”*, also find all word forms that are linked to that lemma, such as the plural form *“faders”* and the genitive form *“faders”.*

For the search fields “Word” and “Lemma” it is also possible to search for a series of tokens by entering multiple values, including wildcards, separated by a space, e.g. *“twelif skilling,”* or “*twelif *”.*
Values at the same position in different fields are grouped together as a single token, meaning that all values in the first position of each field are grouped to match a single token.
The results of the search for all lemmata following “*twelif*” (grouped by lemma) are displayed below.



? Group	#hits in group	Relative frequency (hits)
twelif skilling	165	0.0701%
twelif êth	49	0.0208%
twelif merk_2	44	0.0187%
twelif penning	24	0.0102%
twelif mon	23	0.00977%
twelif grâta	11	0.00467%
twelif hond	9	0.00382%
twelif jêr	9	0.00382%
twelif ande	9	0.00382%
twelif sawen	6	0.00255%
twelif undswera	5	0.00212%
twelif tô	5	0.00212%
twelif pund	4	0.0017%
twelif apostola	3	0.00127%
twelif londriucht	2	0.000849%
twelif dêdêth	2	0.000849%

The wildcard can also be used before a lemma: “* *mon*”, giving all lemmata occurring one position before “*mon*”.



Group	#hits in group	Relative frequency (hits)
ên mon	840	0.357%
thî mon	284	0.121%
ôther mon	68	0.0289%
nên mon	65	0.0276%
ênich_2 mon	42	0.0178%
twelif mon	23	0.00977%
al mon	22	0.00934%
twêne mon	21	0.00892%
hire mon	18	0.00764%
nânên mon	17	0.00722%
on mon	14	0.00595%
afte_2 mon	14	0.00595%
elk mon	13	0.00552%
frethelâs mon	12	0.0051%
sex mon	10	0.00425%
tô mon	9	0.00382%
jeftha mon	9	0.00382%
ande mon	8	0.0034%
thrê mon	7	0.00297%
âênich mon	7	0.00297%
rîke_3 mon	7	0.00297%
ain_2 mon	7	0.00297%
gâstlik mon	6	0.00255%

A multi-lemma example: searching for *”oxa|ku”* in the Lemma field should find all occurrences of either “*ku*” or “*oxa*”.

6.2.2 Upload List of Values

At the right side of the search fields “Word” and “Lemma” there is an option to upload a file with a list of multiple values; these values must all be separated by a white space or a “|”. Note that this function only works for Plain text files (*.txt). Every word in the uploaded file will be added to the list of values to search for. To remove the word list simply press the “Reset” button.

6.2.3 Part of Speech Dialog Box

Clicking on the arrow next to the search field “Part of Speech” provides you with the Part of speech options list. See also the Table in this document with all the PoS-tags used in this corpus.

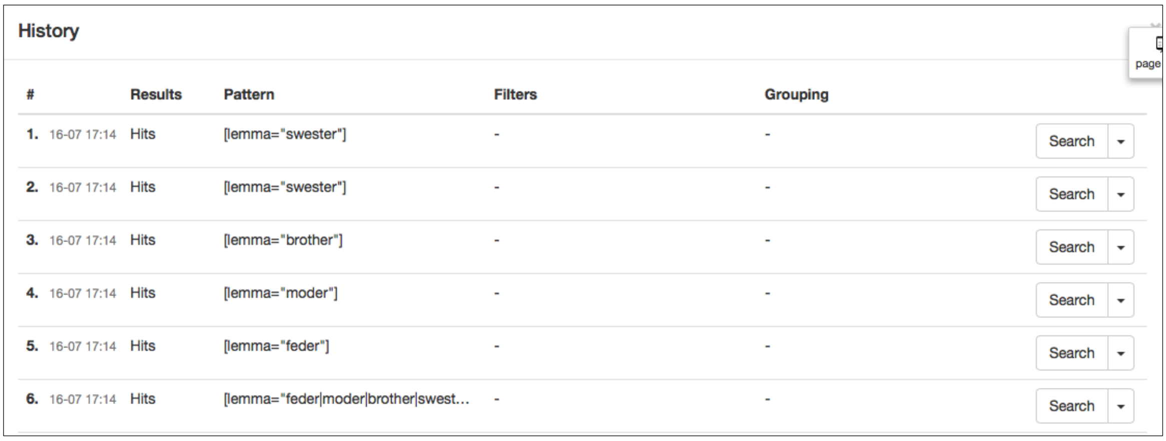
6.2.4 Within

Below the Part of Speech dialog box the option “Within” is displayed. This feature is not implemented in the corpus Oudfries.

6.2.5 Batch Splitting

Ticking the option “Split batch queries” will split a query containing a certain number of attributes, separated by the character | or the Boolean operator OR, in a set of smaller queries equal to the number of chosen attributes and store these separate queries in the History list. Thus every value can be executed as a separate query. The metadata filters filled in by “Filter search by” (see below) will apply to all sub queries. After running the query, only the hits for the *last* value will be shown in the results. To obtain the results for the other values you have to go to “History”, select the value (i.e. pattern) you are looking for and press “Search” on the right hand side of the history panel.

An example will clarify this. Searching for *“feder*|moder|brother|swester*”* in “Lemma” will show all hits for *“feder”*, *“moder”*, *“brother”* and *“swester”*. When the option “Split batch queries” is ticked, the same query will show only hits of *“swester”* (being the last value). In “History” you will see the attributes separately in individual queries which can now also be searched for individually.



History

#	Results	Pattern	Filters	Grouping	
1. 16-07 17:14	Hits	[lemma="swester"]	-	-	Search ▾
2. 16-07 17:14	Hits	[lemma="swester"]	-	-	Search ▾
3. 16-07 17:14	Hits	[lemma="brother"]	-	-	Search ▾
4. 16-07 17:14	Hits	[lemma="moder"]	-	-	Search ▾
5. 16-07 17:14	Hits	[lemma="feder"]	-	-	Search ▾
6. 16-07 17:14	Hits	[lemma="feder|moder|brother|swest...	-	-	Search ▾

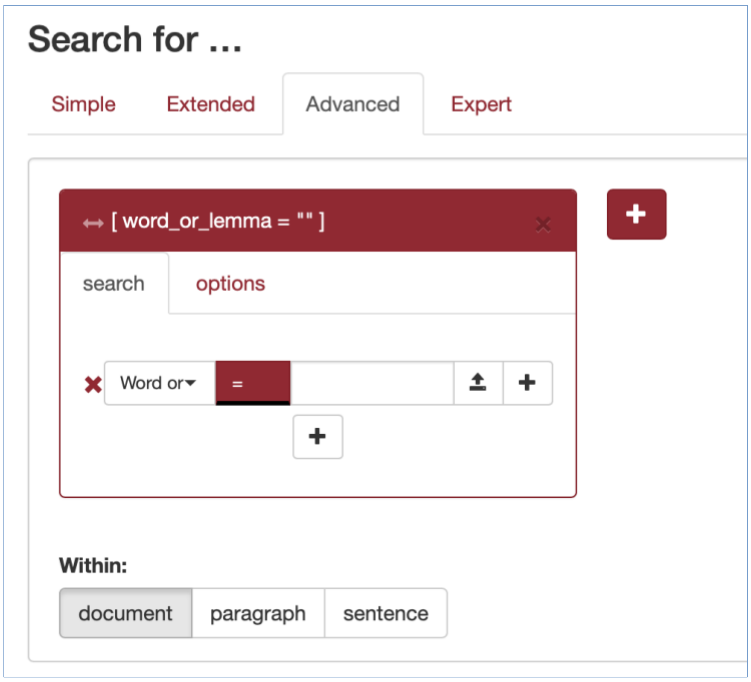
6.3 Advanced Search

6.3.1 The Query Builder

You can use the query builder to create complex queries with multiple components without having to write CQL (see “Expert Search”). Any time a query is created in the query builder, it is also copied to the Advanced Search CQL editor, in which you can further edit the query by hand.

6.3.2 Token Boxes

Each query component is represented by a building block, a so-called “token box”. A token is a term used in corpus linguistics to indicate an individual occurrrence of a linguistic unit (word). When multiple tokens are used, they are matched in order from left to right and in that way searching for a phrase that is made up of several words/tokens can be made possible.



A token box in the querybuilder has two tabs:

- “Search”: The set of attributes a token in the corpus must have to be matched by the query. Create attributes by clicking the + button within this token. Then enter a value that the attribute must have for the token to be found. It it also possible to define specific combinations of attributes, or specify lists of possible attributes, and more.
- “Options”: specify the repetition pattern and contextual properties, such as whether the token occurs at the start, or the end of a sentence.

The CQL query (see for information on CQL the following chapters) generated to match this token (the “*token query*”) in the corpus is displayed in the top bar of the box, to help you understand what is happening internally.

6.3.3 Token Attributes

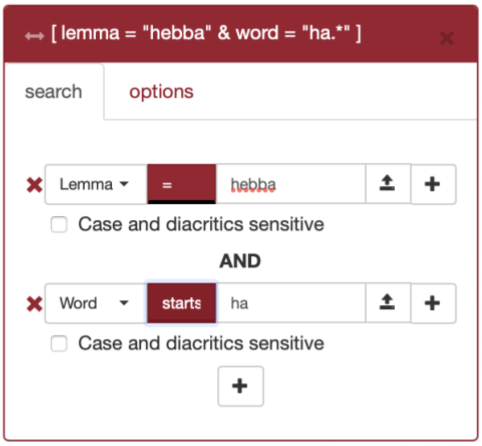
Specifying token attributes is similar to the “Extended search” form. Select the attribute a token should have, and enter the value that the attribute must have for the token to be matched. Attributes in the query builder are interpreted as *regular expressions*. Note that this is different from the extended search, where token patterns use wildcards.

6.3.4 Adding Attributes

A token box also allows you to combine several attributes and to specify repetition options.

Using the "+" button, new query components can be added. Two options exist: *AND* and *OR*.

The *AND* option creates a new attribute restriction that a token must match in addition to the ones which were already there. As an example: suppose we want to find all occurrences of the verb "*hebba*" (to have) with the spelling startng with "ha-". First, fill in the attribute "Lemma" with value "*hebba*", then click +, choose *AND*, and choose the value "ha" for word. And change the red "=" sign to "starts with".



[lemma = "hebba" & word = "ha.*"]

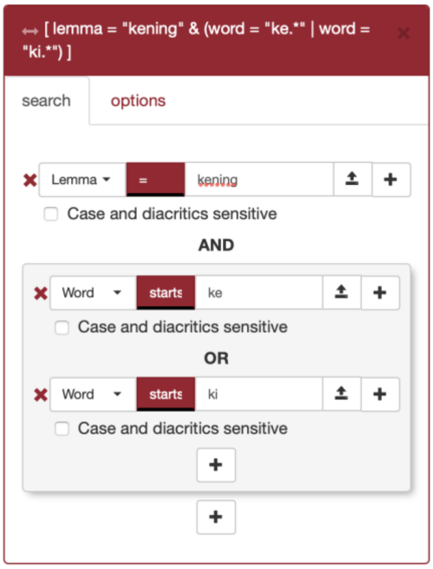
□ Case and diacritics sensitive

□ Case and diacritics sensitive

Creating a new attribute using *OR* will broaden the search results.

6.3.5 Creating Subclauses

The +-sign on the right of an attribute and the one below it, is that the +-sign on the right keeps the newly added attribute "within a subclause", comparable to a mathematical formula which includes parentheses.



[lemma = "kening" & (word = "ke.*" | word = "ki.*")]

Lemma = kening

□ Case and diacritics sensitive

AND

Word starts ke

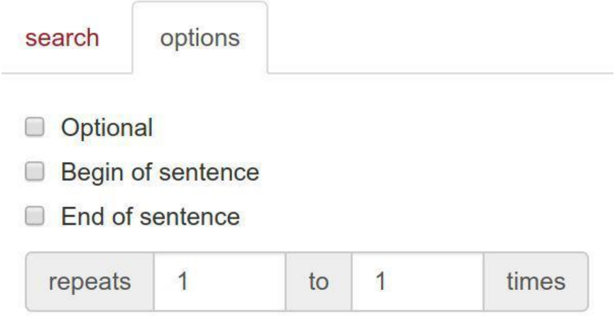
□ Case and diacritics sensitive

OR

Word starts ki

□ Case and diacritics sensitive

6.3.6 Specifying Repetition



□ Optional

□ Begin of sentence

□ End of sentence

repeats 1 to 1 times

Begin/end of sentence feature will not operate reliably in the corpus Oudfries because the medieval punctuation differs from modern methods and the punctuation in the corpus texts has not been annotated properly.

The Repeat options can be used to specify multiple token boxes with the same properties so as to avoid having to make multiple token boxes with the same attributes.

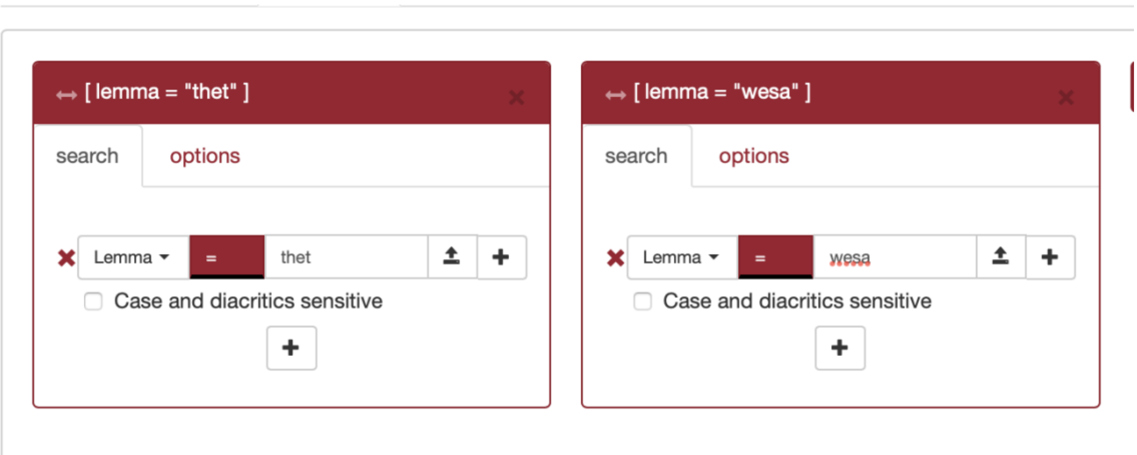
6.3.7 Managing Sequences of Token Boxes

Create new tokens by clicking the + button on the right of the rightmost box.

Rearrange a token by clicking and dragging the little arrow handle in the top-left corner.

Delete a token by clicking the x in the top-right corner.

The token boxes in the screenshot below represent a query for the phrase “thet wesa” in all its orthographical variants (“thet is”, “dat is”, “Thet weron”, etc.).



6.3.8 Uploading value lists in the query builder

It iss also possible to upload a list of values. To do so, click the upload button that appears when you hover the mouse over the text input for this attribute and select a text file. Tokens will then be matched for any of the values from the file.

Uploaded files must be plain text files (.txt), with each value on its own line.

After uploading a file, the text can be edited by clicking the yellow button that has appeared in place of the text field. Editing the text is temporary and will not modify your original file.s

To remove an uploaded file and go back to typing a value, click the upload button a second time, then press cancel instead of selecting a file.

6.4 Expert Search

CQL queries (Corpus Query Language; see https://www.sketchengine.eu/documentation/corpus-querying/) are expressions formulated with the help of a few sequence operators and brackets from basic blocks enclosed by square brackets, in each of which one or more token attributes are specified (these correspond to the token boxes in the query builder; “Advanced search”).

The CQL editor allows you to type your own CQL query, or edit a query further after creating it in the querybuilder. Queries generated by the query builder will be pasted here automatically. Note that this will overwrite any queries you type here yourself!

In some cases, when the query is relatively simple, it can also be imported into the query builder using the “*Copy to query builder”* button. A message will be displayed next to the button if the query could not be parsed.

Some examples:

- Simple: [word="bank"], [word="bank"], [word!="bank"]. Here a=b or a!= b should be understood as “attribute a matches the regular expression b” respectively “attribute a does not match the regular expression b”, e.g. [lemma=".*bank"] matches all lemmata ending with “bank”.
- abbreviated notation for word form queries: In most corpora, just "bank" will be a shortcut for [word="bank"]
- Combination of attributes (combining operators are &, |, !), e.g. [word="rust" & pos_head !="n"] or equivalently [word="rust" & ! pos_head ="n"]
- Simple sequence: [pos_head="adj"][lemma="bank"]
- Repetition operators: [pos_head="adj"]{3} matches a sequence of 3 adjectives,
- [pos_head="adj"]{3,5} matches a sequence of 3 to 5 adjectives, [pos_head="ADJ"]{3,} matches a sequence of 3 or more adjectives
- The empty [] matches any token, e.g. [pos_head="adj"][]{4}[pos_head="adj"] matches two adjectives with 4 arbitrary tokens in between
- Querying with tag positions using e.g. <s> (start of sentence), </s> (end of sentence), <s/> (whole sentence) or <s> ... </s>
- Example: <s> [pos_head="ww"][word="je|jij"].
- <s> []{4} </s>, matching 4-token sentences (in fact this includes punctuation tokens, so <s> [pos_head!="let"]{4} </s> gives nicer results)
- Operators |, & and parentheses and the repetition operators can be used to build complex sequence queries. Example: "brave" "hond" | "stoute" "kat", or even ("happy" "dog" | "sad" cat")+, matching any sequence of happy dogs and/or sad cats. Note that, while most queries up to this point could also have been constructed with the query builder, we really need the power of CQL from here on.

This short list does not cover all CQL features. For more detailed information on how to write CQL, please consult the short CQL manual in the last chapter, which contains further pointers.

6.4.1 Gap Filling

Use this option to upload a TSV file (tab separated) with terms to complete a query with marked gaps functioning as variables. For instance, given a query:

```
[lemma="@@"][pos="ADJ.*"][lemma="@@"]
```

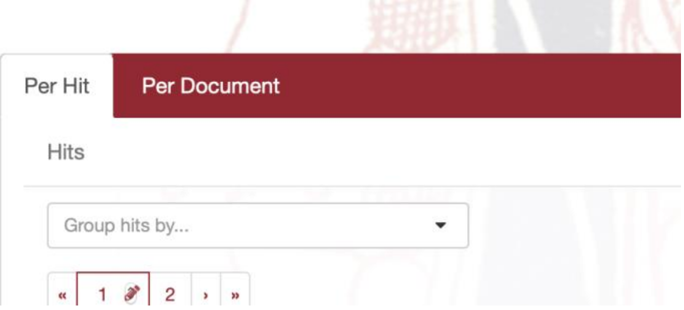
you supply a list with two tab-separated columns of terms. The terms in the first column will be entered at the position of the first gap (@@) and the words in the second column at the position of the second gap. This mimics the word list functionality of the “Extended search” and “Advanced search” interfaces. Please note that for this to work, you do need to enter @@ in the field where you want the substitution to take place. An empty field ([]) will match any term.

6.4.2 Saved Queries; Import and Export

Import previously saved queries here. To export a query: first run a search, then open the “History”, and select “Download as file” in the dropdown menu on the right.

7 Viewing Results

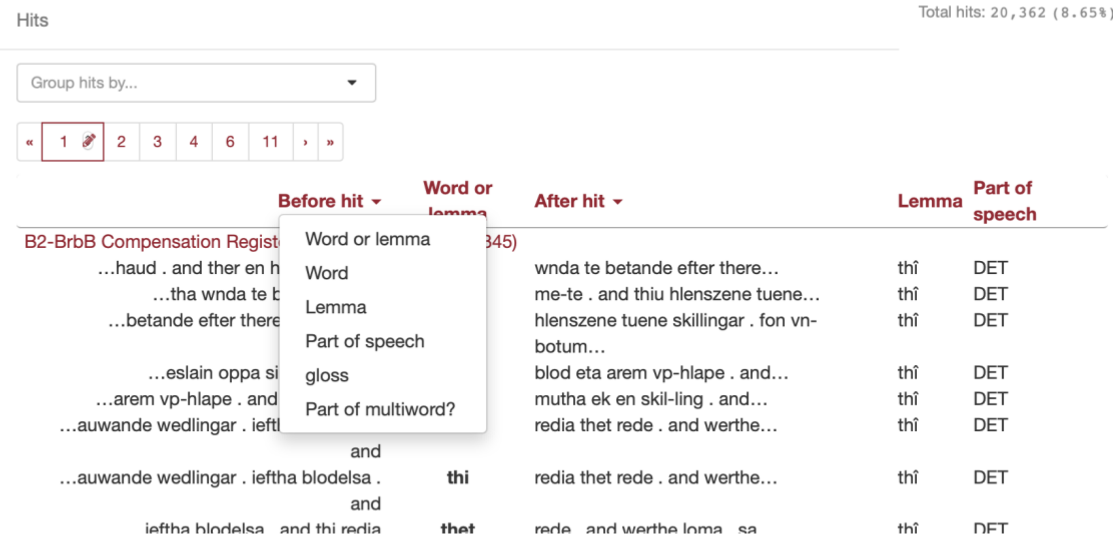
Results can be viewed in two ways: “Per hit” (hit is defined as one token or a group of tokens that matched the query), or “Per document” (each document listed contains at least one hit).



7.1 Per Hit View

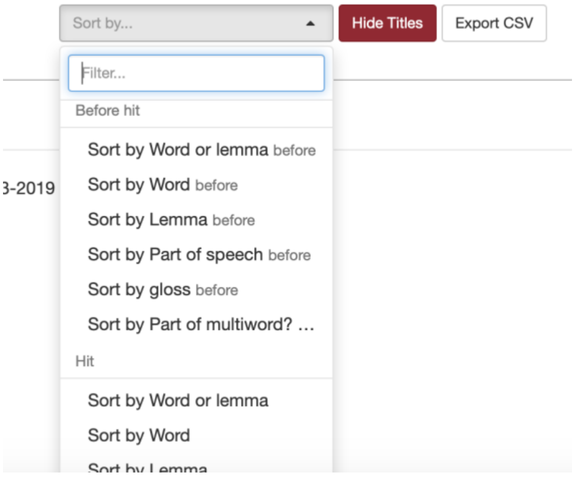
7.1.1 Sorting Options

Click on any of the column headings to sort the hits on values within the column, clicking again inverts the sort order. Extra sorting options are given when clicking on “Before hit” and “After hit”: you can sort by the attributes “Word or Lemma”, “Word”, “Lemma”, “Part of speech” (the other options in the list are not implemented).



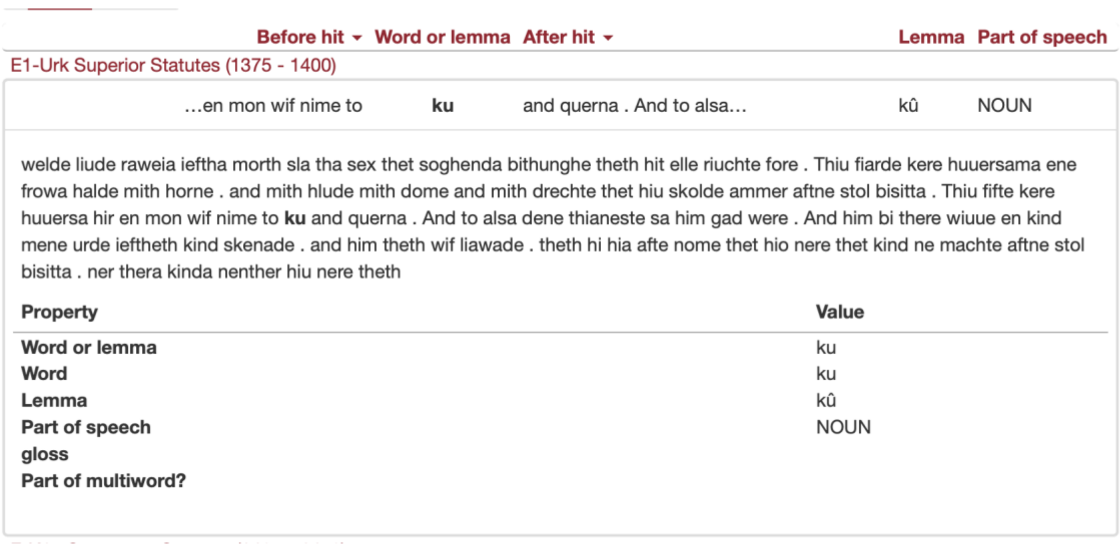
Before hit	Word or lemma	After hit	Lemma	Part of speech
B2-BrbB Compensation Regist	345)			
…haud . and ther en h		wnda te betande efter there…	thî	DET
…tha wnda te b		me-te . and thiu hlenszene tuene…	thî	DET
…betande efter there		hlenszene tuene skillingar . fon vn-botum…	thî	DET
…eslain oppa si		blod eta arem vp-hlape . and…	thî	DET
…arem vp-hlape . and		mutha ek en skil-ling . and…	thî	DET
…auwande wedlingar . ieftl and		redia thet rede . and werthe…	thî	DET
…auwande wedlingar . ieftha blodelsa . and	**thi**	redia thet rede . and werthe…	thî	DET
ieftha blodelsa . and thi redia	**thet**	rede . and werthe loma . sa	thî	DET

You can also sort the results by means of the dropdown menu at the bottom of the page, which offers additional sorting options. Please note again that “gloss” and “Part of multiw” are not implemented.



7.1.2 Hit Information

Click a row/hit to display more information about the particular hit. The window shows a larger portion of the textual context of the hit and the properties that have been recorded for the hit. Click the hit again to close the context window.



Before hit	Word or lemma	After hit	Lemma	Part of speech
E1-Urk Superior Statutes (1375 - 1400)				
…en mon wif nime to	**ku**	and querna . And to alsa…	kû	NOUN

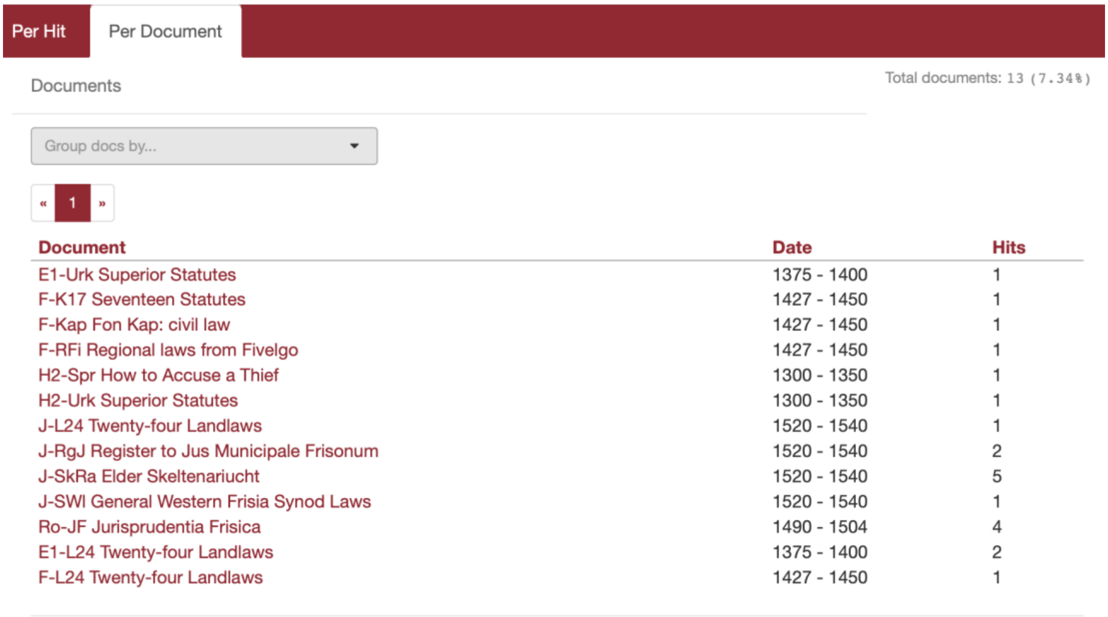
welde liude raweia ieftha morth sla tha sex thet soghenda bithunghe theth hit elle riuchte fore . Thiu fiarde kere huuersama ene frowa halde mith horne . and mith hlude mith dome and mith drechte thet hiu skolde ammer aftne stol bisitta . Thiu fifte kere huuersa hir en mon wif nime to **ku** and querna . And to alsa dene thianeste sa him gad were . And him bi there wiuue en kind mene urde iefthet kind skenade . and him theth wif liawade . theth hi hia afte nome thet hio nere thet kind ne machte aftne stol bisitta . ner thera kinda nenther hiu nere theth

Property	Value
Word or lemma	ku
Word	ku
Lemma	kû
Part of speech	NOUN
gloss	
Part of multiword?	

7.1.3 Show/Hide Titles

The document titles can can be toggled on or off by using the Show Titles / Hide Titles button below the Results view.

7.2 Per Document View



Total documents: 13 (7.34%)

Document	Date	Hits
E1-Urk Superior Statutes	1375 - 1400	1
F-K17 Seventeen Statutes	1427 - 1450	1
F-Kap Fon Kap: civil law	1427 - 1450	1
F-RFi Regional laws from Fivelgo	1427 - 1450	1
H2-Spr How to Accuse a Thief	1300 - 1350	1
H2-Urk Superior Statutes	1300 - 1350	1
J-L24 Twenty-four Landlaws	1520 - 1540	1
J-RgJ Register to Jus Municipale Frisonum	1520 - 1540	2
J-SkRa Elder Skeltenariucht	1520 - 1540	5
J-SWl General Western Frisia Synod Laws	1520 - 1540	1
Ro-JF Jurisprudentia Frisica	1490 - 1504	4
E1-L24 Twenty-four Landlaws	1375 - 1400	2
F-L24 Twenty-four Landlaws	1427 - 1450	1

Click on any of the column headings to sort the hits on values within the column, clicking again inverts the sorting order. You can also sort the results by means of the dropdown menu at the bottom of the page, which offers additional sorting options.

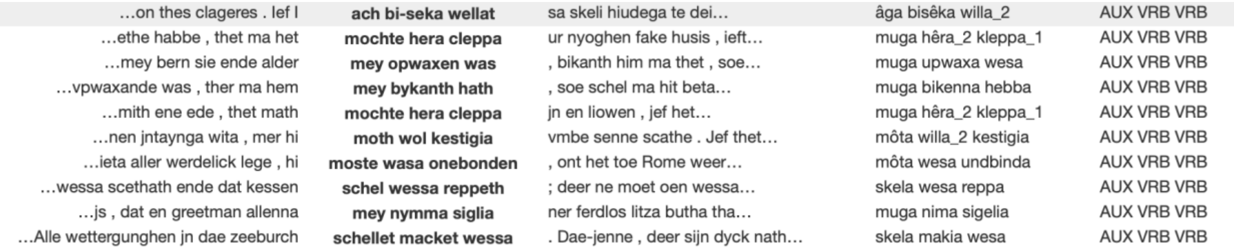
Click on a document title to open the document in a new window. Hits from the current query will be highlighted in the opened document.

7.2.1 Grouping Results

Results can be grouped by properties of the hits or of the documents in which those hits occur. Grouping is facilitated by the dropdown menus “Group hits by” and “Group docs” by. You can group the hits by their part-of-speech tags or you can group the hits and the documents by means of the collection (metadata filtering) to which they belong.

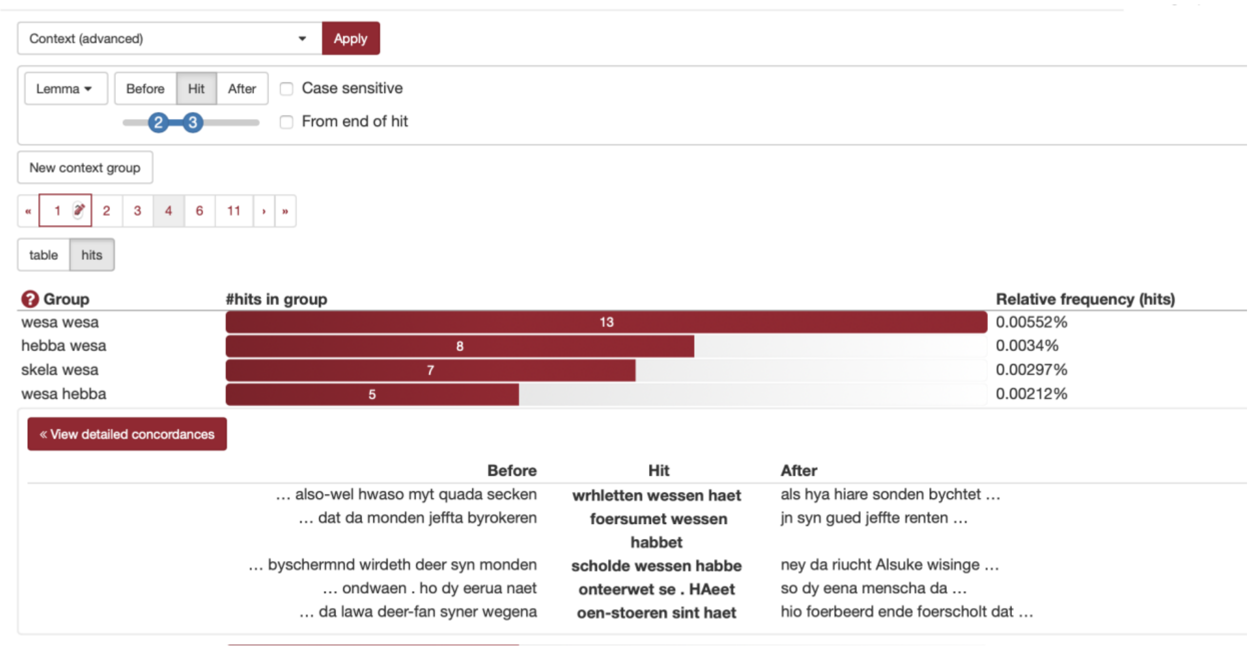
In the “Per hit” view, advanced grouping options are available by selecting the option “Context (advanced)”. This option allows you to group the results by up to 5 tokens before or after the hits. It also allows you to group the results based on (parts of) the hits.

A search for occurrences of three consecutive verbs or auxiliary verb forms produces hits like the following:



…on thes clageres . Ief I	**ach bi-seka wellat**	sa skeli hiudega te dei…	âga bisêka willa_2	AUX VRB VRB
…ethe habbe , thet ma het	**mochte hera cleppa**	ur nyoghen fake husis , ieft…	muga hêra_2 kleppa_1	AUX VRB VRB
…mey bern sie ende alder	**mey opwaxen was**	, bikanth him ma thet , soe…	muga upwaxa wesa	AUX VRB VRB
…vpwaxande was , ther ma hem	**mey bykanth hath**	, soe schel ma hit beta…	muga bikenna hebba	AUX VRB VRB
…mith ene ede , thet math	**mochte hera cleppa**	jn en liowen , jef het…	muga hêra_2 kleppa_1	AUX VRB VRB
…nen jntaynga wita , mer hi	**moth wol kestigia**	vmbe senne scathe . Jef thet…	môta willa_2 kestigia	AUX VRB VRB
…ieta aller werdelick lege , hi	**moste wasa onebonden**	, ont het toe Rome weer…	môta wesa undbinda	AUX VRB VRB
…wessa scethath ende dat kessen	**schel wessa reppeth**	; deer ne moet oen wessa…	skela wesa reppa	AUX VRB VRB
…js , dat en greetman allenna	**mey nymma siglia**	ner ferdlos litza butha tha…	muga nima sigelia	AUX VRB VRB
…Alle wettergunghen jn dae zeeburch	**schellet macket wessa**	. Dae-jenne , deer sijn dyck nath…	skela makia wesa	AUX VRB VRB

It is possible to group the hits by the second and third tokens of those hits.



□ Case sensitive

□ From end of hit

Group	#hits in group	Relative frequency (hits)
wesa wesa	13	0.00552%
hebba wesa	8	0.0034%
skela wesa	7	0.00297%
wesa hebba	5	0.00212%

Before	Hit	After
… also-wel hwaso myt quada secken	wrhletten wessen haet	als hya hiare sonden bychtet …
… dat da monden jeffta byrokeren	foersumet wessen habbet	jn syn gued jeffte renten …
… byschermnd wirdeth deer syn monden	scholde wessen habbe	ney da riucht Alsuke wisinge …
… ondwaen . ho dy eerua naet	onteerwet se . HAeet	so dy eena menscha da …
… da lawa deer-fan syner wegena	oen-stoeren sint haet	hio foerbeerd ende foerscholt dat …

Choose “Context Advanced” in the Group by dropdown list and the special grouping options will be displayed. After selecting the options click “Apply” to execute the grouping command.

Click a group’s bar to show the hits within that group. Click once more on the group to close it again.

Click on “View detailed concordances” to go to the normal hits view to see more detailed information for the hits in this particular group of results.

7.2.2 Exporting Results

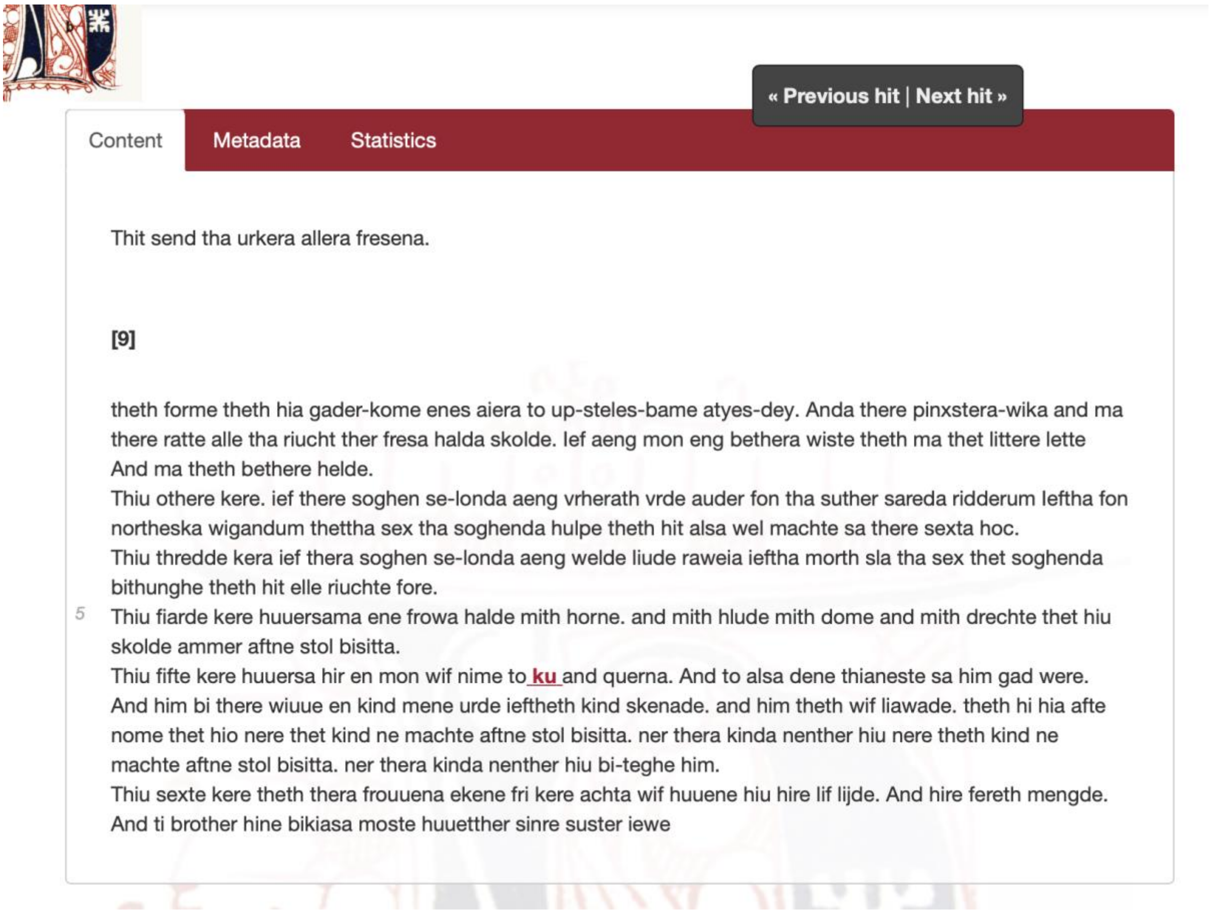
The search results can be exported by using the “Export CSV” button at the bottom right of the page.

Files with the layout of Comma Separated Value scan be imported in other software applications, such as Excel or SPSS.

8 Information about Documents / Text Witnesses

8.1 Content

In order to see all the source text of a particular text witness or document as it is called in the seach engine, click (in the results window) on a text witness title (printed in red) to open the complete text in a new window. Hits from the current query (in this case “lemma=ku”) will be highlighted in the opened document.



Thit send tha urkera allera fresena.

[9]

theth forme theth hia gader-kome enes aiera to up-steles-bame atyes-dey. Anda there pinxstera-wika and ma there ratte alle tha riucht ther fresa halda skolde. Ief aeng mon eng bethera wiste theth ma thet littere lette

And ma theth bethere helde.

Thiu othere kere. ief there soghen se-londa aeng vrherath vrde auder fon tha suther sareda ridderum Ieftha fon northeska wigandum thettha sex tha soghenda hulpe theth hit alsa wel machte sa there sexta hoc.

Thiu thredde kera ief thera soghen se-londa aeng welde liude raweia ieftha morth sla tha sex thet soghenda bithunghe theth hit elle riuchte fore.

Thiu fiarde kere huuersama ene frowa halde mith horne. and mith hlude mith dome and mith drechte thet hiu skolde ammer aftne stol bisitta.

Thiu fifte kere huuersa hir en mon wif nime to **ku** and querna. And to alsa dene thianeste sa him gad were.

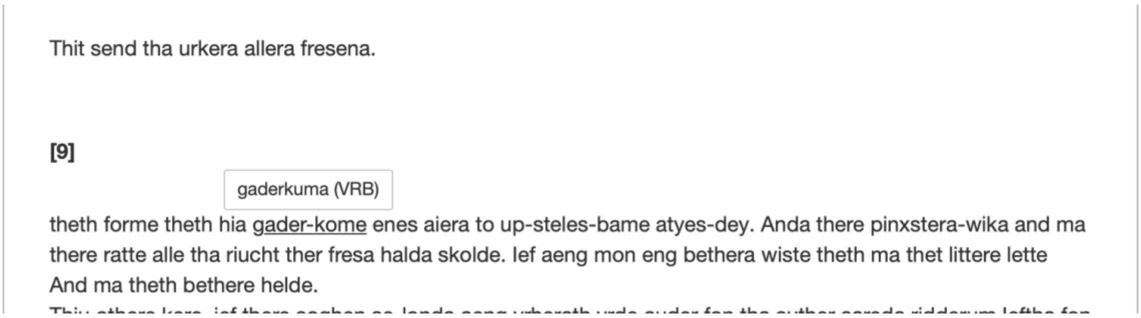
And him bi there wiuue en kind mene urde ieftheth kind skenade. and him theth wif liawade. theth hi hia afte nome thet hio nere thet kind ne machte aftne stol bisitta. ner thera kinda nenther hiu nere theth kind ne machte aftne stol bisitta. ner thera kinda nenther hiu bi-teghe him.

Thiu sexte kere theth thera frouuena ekene fri kere achta wif huuene hiu hire lif lijde. And hire fereth mengde.

And ti brother hine bikiasa moste huuetther sinre suster iewe

With the buttons “Previous hit | Next hit” it is possible to jump to the previous or next hit in the text.

When hovering with the mouse over the words in the text view, a small box with lemma and PoS-tag information will become visible (see also the screenshot below).



Thit send tha urkera allera fresena.

[9]

theth forme theth hia gader-kome enes aiera to up-steles-bame atyes-dey. Anda there pinxstera-wika and ma there ratte alle tha riucht ther fresa halda skolde. Ief aeng mon eng bethera wiste theth ma thet littere lette

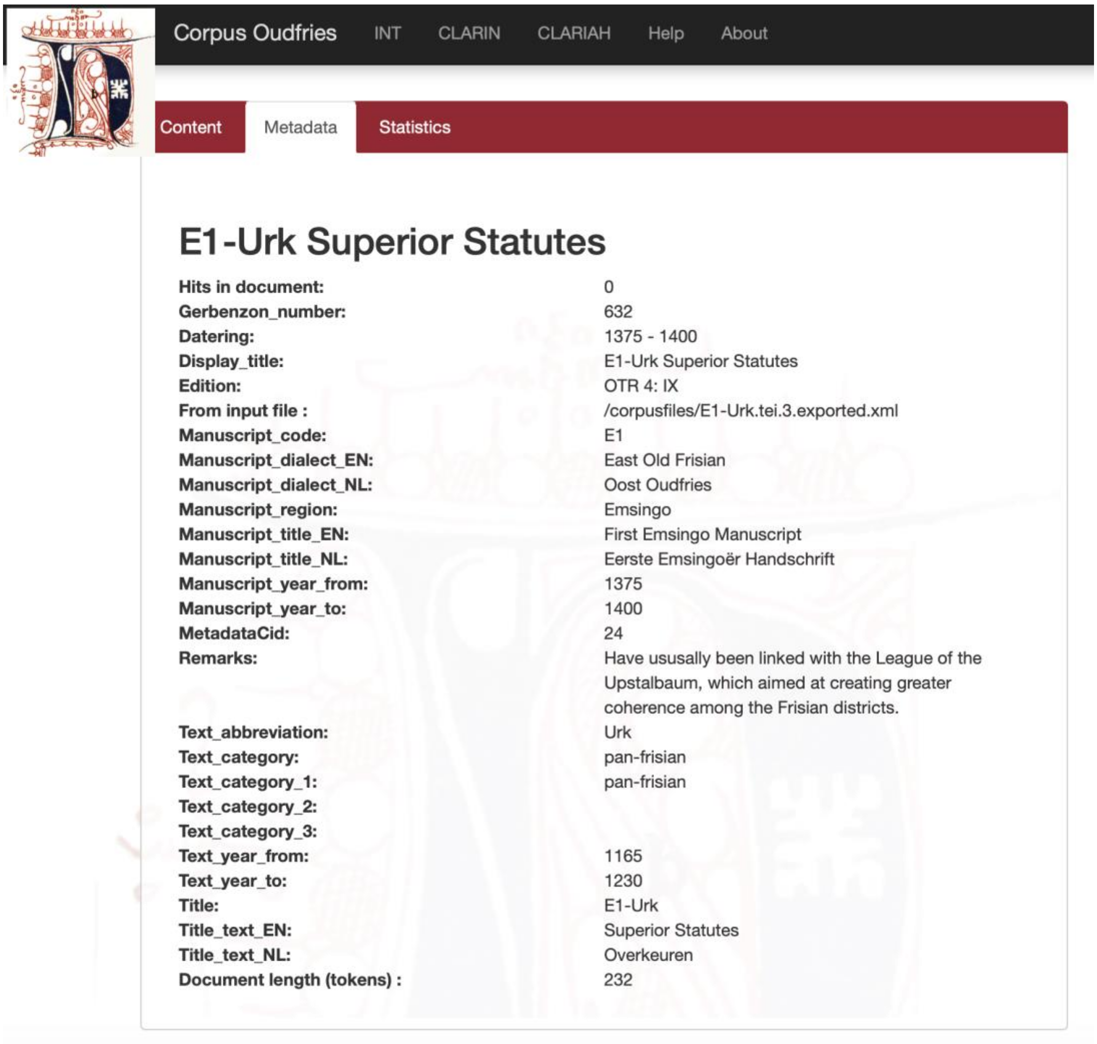
And ma theth bethere helde.

Thiu othere kere. ief there soghen se-londa aeng vrherath vrde auder fon tha suther sareda ridderum Ieftha fon

The document/text witness view will depend on the corpus and this functionality is only available when the corpus has not disabled the viewing of documents (unavailability is usually due to copyrights restrictions).

8.2 Metadata

In the “Metadata” tab, all metadata properties of the tekst witness are displayed. Some of these properties have been indexed which means they can be used as filtering or grouping options. These properties are: Title, Title_text_EN, Text_category, manuscript_title_EN, manuscript_region, manuscript_dialect_EN, manuscript_year, manuscript_year (“Date text witness”) and text_year (“Date text”).



E1-Urk Superior Statutes

Hits in document:	0
Gerbenzon_number:	632
Datering:	1375 - 1400
Display_title:	E1-Urk Superior Statutes
Edition:	OTR 4: IX
From input file :	/corpusfiles/E1-Urk.tei.3.exported.xml
Manuscript_code:	E1
Manuscript_dialect_EN:	East Old Frisian
Manuscript_dialect_NL:	Oost Oudfries
Manuscript_region:	Emsingo
Manuscript_title_EN:	First Emsingo Manuscript
Manuscript_title_NL:	Eerste Emsingoër Handschrift
Manuscript_year_from:	1375
Manuscript_year_to:	1400
MetadataCid:	24
Remarks:	Have ususally been linked with the League of the Upstalbaum, which aimed at creating greater coherence among the Frisian districts.
Text_abbreviation:	Urk
Text_category:	pan-frisian
Text_category_1:	pan-frisian
Text_category_2:	
Text_category_3:	
Text_year_from:	1165
Text_year_to:	1230
Title:	E1-Urk
Title_text_EN:	Superior Statutes
Title_text_NL:	Overkeuren
Document length (tokens) :	232

8.3 Statistics

The “Statistics” tab offers statistical information about a document, e.g. about the type-token ratio and vocabuary growth from the beginning to the end of the text.

9 Corpus Query Language

This section of the manual has been copied from the OpenSonar Manual developed by the IvdNt. It uses examples from the English language. The BlackLab software supports Corpus Query Language, a full-featured query language introduced by the IMS Corpus WorkBench (CWB) and also supported by the Lexicom Sketch Engine. It is a standard and powerful way of searching corpora.

This page will introduce the query language and show all features that BlackLab supports. If you want to learn even more about CQL, see CWB CQP Query Language Tutorial and Sketch Engine Corpus Query Language.

9.1 CQL Support

BlackLab currently supports (arguably) most of the important features of Corpus Query Language:

- Matching on token annotations (also called properties or attributes), using regular expressions and =, !=, !. Example: [word="bank"] (or just "bank")
- Case/accent sensitive matching. Note that, case-INsensitive matching is currently the default. To explicitly match case-/accent-insensitively, use "(?i)...". Example: "(?-i)Mr\." "(?-i)Banks"
- Combining criteria using &, | and !. Parentheses can also be used for grouping. Example: [lemma="man" & pos="VRB"]
- Match-all pattern [] matches any token. Example: "a" [] "day"
- Regular expression operators +, *, ?, {n}, {n,m} at the token level. Example: [pos="ADJ"]+
- Sequences of token constraints.
- Operators |, & and parentheses can be used to build complex sequence queries. Example: "happy" "dog" | "sad" cat"
- Querying with tag positions using e.g. <s> (start of sentence), </s> (end of sentence), <s/> (whole sentence) or <s> ... </s> (equivalent to <s/> containing ...). Example: <s> "The" . XML attribute values may be used as well, e.g. <ne type="PPRON"/> ("named entities that are persons").
- Using within and containing operators to find hits inside another set of hits. Example: "you" "are" within <s/>.
- Using an anchor to capture a token position. Example: "big" A:[]. Captured matches can be used in global constraints (see next item) or processed separately later (using the Java interface; capture information is not yet returned by BlackLab Server). Note that BlackLab can actually capture entire groups of tokens as well, similarly to regular expression engines.
- Global constraints on captured tokens, such as requiring them to contain the same word. Example: "big" A:[] "or" "small" B:[] :: A.word = B.word

9.2 Differences between CQL and CWB

BlackLab's CQL syntax and behaviour differs in a few small ways from CWBs. In future, the IvdNT aims towards greater compliance with CWB's de-facto standard (with some extra features and conveniences).

- Case-insensitive search is currently the default in BlackLab, although you can change this if you wish.
- If you want to switch case/diacritics sensitivity, use "(?-i).." (case sensitive) or "(?i).." (case insensitive, usually the default). CWBs %cd flags for setting case/diacritics-sensitivity are not (yet) supported, but will be added.
- If you want to match a string literally, not as a regular expression, use backslash escaping: "e\.g\.". %l for literal matching is not yet supported, but will be added.
- BlackLab supports result set manipulation such as: sorting (including on specific context words), grouping/frequency distribution, subsets, sampling, setting context size, etc. However, these are supported through the REST and Java APIs, not through a command interface like in CWB.
- Querying XML elements and attributes looks natural in BlackLab: <s/> means "sentences", <s> means "starts of sentences", <s type=A> means "sentence tags with a type attribute with value A". This natural syntax differs from CWBs in some places, however, particularly when matching XML attributes. While we believe our syntax is the superior one, we may add support for the CWB syntax as an alternative.
- We only support literal matching of XML attributes at the moment, but this will be expanded to full regex matching.
- In global constraints (expressions occurring after ::), only literal matching (no regex matching) is currently supported. Regex matching will be added soon. For now, instead of A:[] "dog" :: A.word = "happy|sad", use "happy|sad" "dog".
- To expand your query to return whole sentences, use <s/> containing (...). We don't yet support CWBs expand to, expand left to, etc., but may add this in the future.
- The implication operator -> is currently only supported in global constraints (expressions after the :: operator), not in a regular token constraints. We may add this if there's demand for it.
- We don't support the @ anchor and corresponding target label; use a named anchor instead. If someone makes a good case for it, we will consider adding this feature.
- backreferences to anchors only work in global constraints, so this doesn't work: A:[] [] [word =
- A.word]. Instead, use something like: A:[] [] B:[] :: A.word = B.word. We hope to add support for these in the near future, but our matching approach may not allow full support for this in all cases.

9.3 Unsupported Features

The following features are not (yet) supported:

- intersection, union and difference operators. These three operators will be added in the future. For now, the first two can be achieved using & and | at the sequence level, e.g. "double" [] & [] "trouble" to match the intersection of these queries, i.e. "double trouble" and "happy" "dog" | "sad "cat" to match the union of "happy dog" and "sad cat".
- _ meaning "the current token" in token constraints. We will add this soon.
- lbound, rbound functions to get the edge of a region. We will probably add these.
- distance, distabs functions and match, matchend anchor points (sometimes used in global constraints). We will see about adding these.

- using an XML element name to mean token is contained within, like [(pos = "N") & !np] meaning "noun NOT inside in an tag". We will see about adding these.

9.4 Using Corpus Query Language

9.4.1 Matching Tokens

Corpus Query Language is a way to specify a "pattern" of tokens (i.e. words) you're looking for. A simple pattern is this one:

```
[word="mon"]
```

The query could be written even simpler without brackets, because "word" is the default property:

```
"man"
```

This simply searches for all occurrences of the word "man".

You can use lemma and PoS-tags seaches as well. For example, to find a form of word "search" used as a noun, use this query:

```
[lemma="search" & pos="NOUN"]
```

This query would match "search" and "searches" where used as a noun.

Use the "does not equal" operator (!=) to search for all words except nouns:

```
[pos != "NOUN"]
```

The strings between quotes can also contain wildcards, of sorts. To be precise, they are regular expressions,which provide a flexible way of matching strings of text. For example, to find "man" or "woman", use:

```
"(wo)?man"
```

And to find lemmata starting with "under", use:

```
[lemma="under.\*"]
```

Explaining regular expression syntax is beyond the scope of this document, but for a complete overview, see here. https://docs.microsoft.com/en-us/dotnet/standard/base-types/regular-expression-language-quick-reference

9.4.2 Sequences

Corpus Query Language allows you to search for sequences of words as well (i.e. phrase searches, but with many more possibilities). To search for the phrase "the tall man", use this query:

```
"the" "tall" "man"
```

It might seem a bit clunky to separately quote each word, but this allow us the flexibility to specify exactly what kinds of words we're looking for. For example, if you want to know all single adjectives used with man (not just "tall"), use this:

```
"an?|the" [pos="ADJ"] "man"
```

This would also match "a wise man", "an important man", "the foolish man", etc.

9.4.3 Regular Expression Operators

Corpus Query Language really starts to shine when you use the regular expression operators on whole tokens as well. If we want to see not just single adjectives applied to "man", but multiple as well:

```
"an?|the" [pos="ADJ"]+ "man"
```

This query matches "a little green man", for example. The plus sign after [pos="ADJ"] says that the preceding part should occur one or more times (similarly, * means "zero or more times", and ? means "zero or one time").

If you only want matches with two or three adjectives, you can specify that too:

```
"an?|the" [pos="ADJ"]{2,3} "man"
```

Or, for two or more adjectives:

```
"an?|the" [pos="ADJ"]{2,} "man"
```

You can group sequences of tokens with parentheses and apply operators to the whole group as well. To search for a sequence of nouns, each optionally preceded by an article:

```
("an?|the"? [pos="NOU"])+
```

This would, for example, match the well-known palindrome "a man, a plan, a canal: Panama!" (A note about punctuation: in BlackLab, punctuation tends to not be indexed as a separate token, but as a property of a word token - CWB and Sketch Engine on the other hand tend to index punctuation as a separate token instead. You certainly could choose to index punctuation as a separate token in BlackLab, by the way -- it's just not commonly done. Both approaches have their advantages and disadvantages, and of course the choice affects how you write your queries.)

9.4.4 Case- and Diacritics Sensitivity

CWB and Sketch Engine both default to (case- and diacritics) sensitive search. That is, they exactly match upper- and lowercase letters in your query, plus any accented letters in the query as well. BlackLab, on the contrary, defaults to *IN*sensitive search (although this default can be changed if you like). To match a pattern sensitively, prefix it with "(?-i)":

```
"(?-i)Panama"
```

If you've changed the default search to sensitive, but you wish to match a pattern in your query insensitively, prefix it with "(?i)":

```
[pos="(?i)nou"]
```

Although BlackLab is capable of setting case- and diacritics-sensitivity separately, it is not yet possible from Corpus Query Language. We may add this capability if requested.

9.4.5 Labeling Tokens and Capturing Groups

Just like in regular expressions, it is possible to "capture" part of the match for your query in a "group".

CWB and Sketch Engine offer similar functionality, but instead of capturing part of the query, they label a single token. BlackLab's functionality is very similar but can capture a number of tokens as well.

```
"an?|the" Adjectives:[pos="ADJ"]+ "man"
```

This will capture the adjectives found for each match in a captured group named "Adjectives".

BlackLab also supports numbered groups:

```
"an?|the" 1:[pos="ADJ"]+ "man"
```

9.4.6 Global Constraints

If you tag certain tokens with labels, you can also apply "global constraints" on these tokens. This is a way of relating different tokens to one another, for example requiring that they correspond to the same word:

```
A:[] "by" B:[] :: A.word = B.word
```

This would match "day by day", "step by step", etc.