



Overview

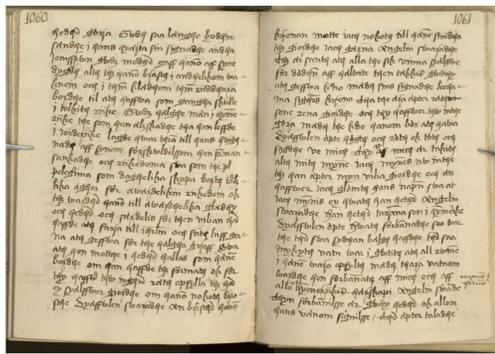
- How can computational linguistics support text-based research in the humanities and social sciences?
- Digital Philology is a subfield within Digital Humanities dealing with the digitization and automatic processing of text-based sources thereby allowing scholars in the humanities and social sciences to study large amounts of data more systematically.

SWE-CLARIN

- SWE-CLARIN is the Swedish branch of the "Common Language Resources and Technology Infrastructure" ERIC and focuses on the development of resources and tools that can be used for research in digital humanities and social sciences.
- Participants: The Swedish Language Bank, The Language Council of Sweden, Swedish National Data Service, Digisam, KTH, Linköping, Lund, Stockholm and Uppsala University.

From Quill To Bytes

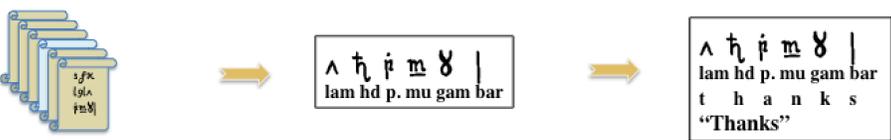
- Computational analysis of historical handwritten material, mainly medieval charters and manuscripts.
- Cooperation involving several subjects: image analysis, philology, computational linguistics, and history, in a series of projects.
- Word spotting and transcription.
- Scribe attribution and dating.



DECODE

- Automatic decoding of historical secret writings
 - Create a database of commonly occurring cipher types, codes and keys from Early modern times (1400-1850)
 - Automatically categorize different types of ciphers
 - Develop algorithms/tools to decrypt various types

Collection [Transliteration] Decryption

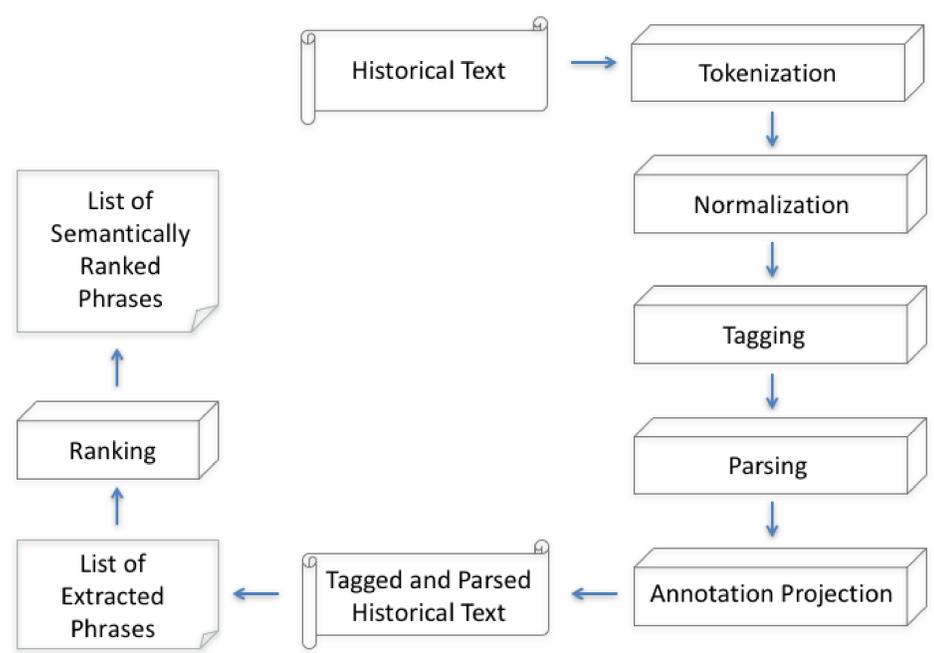


1. Cleartext or ciphertext?
2. Language id?
3. Transcription error?
4. Spelling error?
5. Historical spelling?

Language	No. of items	Number of items				
French	4	2210	1	1	1	1
Italian	1	230	1	1	1	1
Spanish	1	230	1	1	1	1
Portuguese	1	230	1	1	1	1
German	1	230	1	1	1	1
English	1	230	1	1	1	1
Dutch	1	230	1	1	1	1
Swedish	1	230	1	1	1	1
Other	1	230	1	1	1	1

HistSearch

- A tool for automatic information extraction from historical text
- Applied to the Gender and Work project:
 - Historians are interested in what men and women did for a living in the Early Modern Swedish society (appr. 1550-1800)
 - HistSearch automatically searches text for phrases describing working activities
 - Spelling normalization is a crucial step, translating the original spelling to a standardized spelling, prior to tagging and parsing



SWEGRAM

- SWEGRAM is a web-based tool for automatic annotation and quantitative linguistic analysis of Swedish texts.
- SWEGRAM allows users to create their own corpus or compare texts on various linguistic levels. We used SWEGRAM for the creation of the Uppsala Corpus of Student Writings consisting of 1.5 million tokens.



TEXT ID	ID	FORM	NORM	LEMMA	U-POS	C-POS	C-FEATS	U-FEATS	HEAD	DEPREL	Translation
2.2	1	Den	den	DET	DT	UTR SIN DEF	Definite=Def Gender=Com Numbers=Sing		3	det	The
2.2	2	kalla	kall	ADJ	JJ	POS UTR NEU SIN DEF NOM	Case=Nom Definite=Def Degree=Pos Numbers=Sing		3	amod	cold
2.2	3	vinden	vind	NOUN	NN	UTR SIN DEF NOM	Case=Nom Definite=Def Gender=Com Numbers=Sing		4	nsubj	wind
2.2	4	slåg	slog	VERB	VB	PRT AKT	Mood=Ind Tense=Past VerbForm=Fin Voice=Act		0	root	hit
2.2	5	mot	mot	ADP	PP				7	case	[against]
2.2	6	mina	min	DET	PS	UTR NEU PLU DEF	Definite=Def Numbers=Plur Poss=Yes		7	nmod:poss	my
2.2	7	kinder	kind	NOUN	NN	UTR PLU IND NOM	Case=Nom Definite=Ind Gender=Com Numbers=Plur		4	nmod	cheeks
2.2	8	.	.	PUNCT	MAD				4	punct	.

Word statistics

Amount	Mean	Median
7855	490.24	480.3
7147	446.69	445.3
410	254.3	27.5
39	244	2
0	0	0

Part-of-speech statistics

Part of speech	Amount	Share
VB	1636	20.83%
NN	1197	15.24%
PN	846	10.77%
AD	758	9.65%
PP	629	8.01%