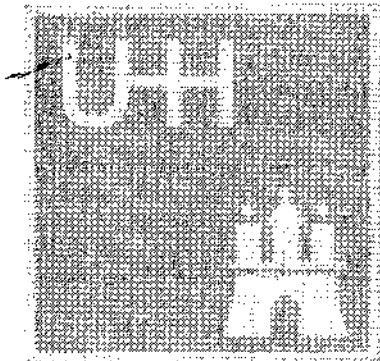


Challenges of Machine (Aided) Translation of Historical Texts

Walther v.Hahn
University of Hamburg • Department of Informatics
Natural Language Systems Group

WWW: http://nats-www.informatik.uni-hamburg.de/view/User/WaltherVHahn
E-Mail: vhahn@informatik.uni-hamburg.de

Why Translation of Historical Texts at all?

- Availability of resources for researchers in other languages.
  - Example: Historical documents of European history in Polish, Russian, French, German, etc.
- Availability of resources for researchers in various fields.
  - Example: Linguistics, social science, law, etc. within a cultural or historical area.
- Availability of resources for private foreign readers (e.g. out of genealogical or geographic interest).
- To allow for summarizing and indexing in other languages.
- Multilinguistic comparison of documents in academic education
- Not useful for historical linguists; however, their expertise is difficult to obtain and expensive for other scientists.

False Assumptions from Modern Texts

Standard Texts

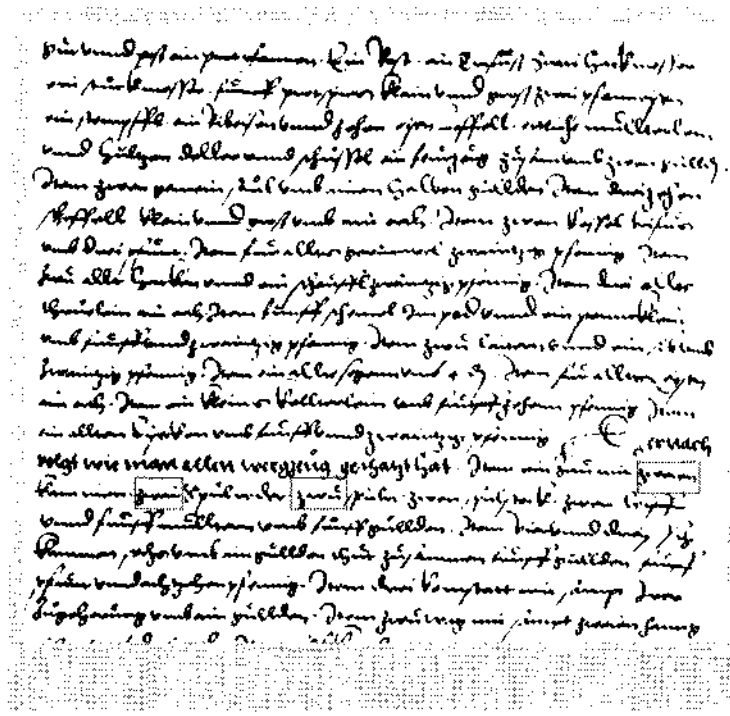
- have an author,
- are written in one language,
- have a well defined orthography,
- have definite grammatical rules,
- belong to a homogenous linguistic and historical layer
- are readable by every speaker of the language,
- are part of a vast homogenous corpus.

Orthography

- there is no orthography, not even writing rules within a document,
- several scripts and languages are mixed in one document,
- arbitrary abbreviations,
- illegible or ambiguous sections.

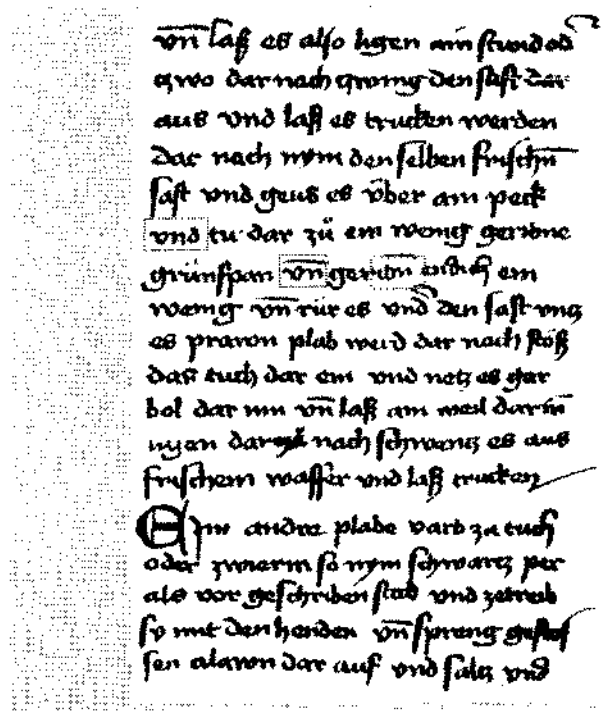
- All this makes OCR extremely unpromising. Informed transcriptions are mandatory for any sort of translation.
- Normalized orthography is standard for most historical linguistic approaches (there is no middle high German contemporary standard orthography, e.g.)

- → excludes SMT



Inventory
Nürnberg
1548

(copy from Germ. Nat. Museum Nürnberg)



Dyer Recipe
from
Heidelberg
15th century

(Copy from Univ. Lib. Heidelberg)

Syntax

- Sentences were longer in medieval times,
- Medieval German tends to build long sequences of adjectives and nouns, esp. in legal documents,
- syntactic means changed: word order, composition;

(Example: *...die machen viel Unruhe den ehrbaren frommen alten Leuten*) ← missing preposition

- inflexion system changed (Example: shift from strong to weak verb paradigm, loss of imperfect in favour of perfect).
- Syntactic change is a continuous flow, because there is no sequence of static rules.
- Similarly, textual rules changed: More scheduled texts
- → excludes RMT and SMT

Sebastian Virdung, Musica getutscht 1511

(Faksimile Kassel o.J.)

Excerpt from Albrants Horse Medicine 1270 AD

So du dy wuerm wellest vertreiben aus dem magen,

so nim ezzeich und ayerschal, wol gestozzen, und rost ab dem eysen und gepranten pheffer, wol gestozzen, und la daz mit einander la werden und gevz iz dem rosse in den hals. Oder sneid gachheil chlain und sevd in mit walendem wein und gevz iz dem ros in den hals.

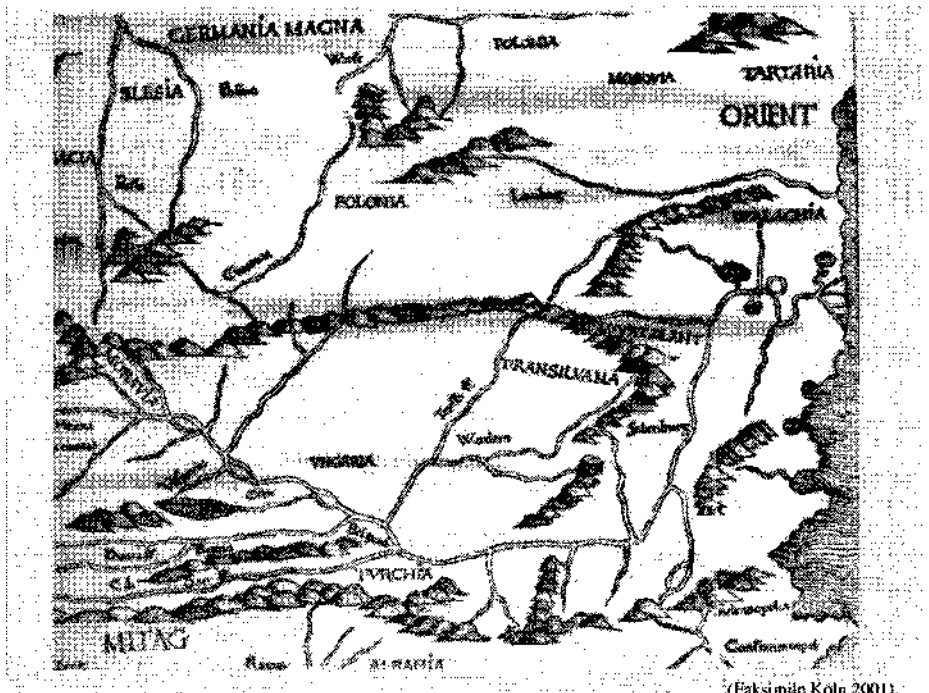
lau laß

Swelch ros wazzerreh ist,
dem trieffent dy naseloecher. Dem laz man an der hals oder.

Semantics

- obsolete words occur in texts,
- lexical semantics of known words changes over time,
- idioms change
- false friends over periods (Germ. *übel, wohl*),
- even technical terms change (Germ. *Verleger*),
- references might be wrong and might look like a translation error,

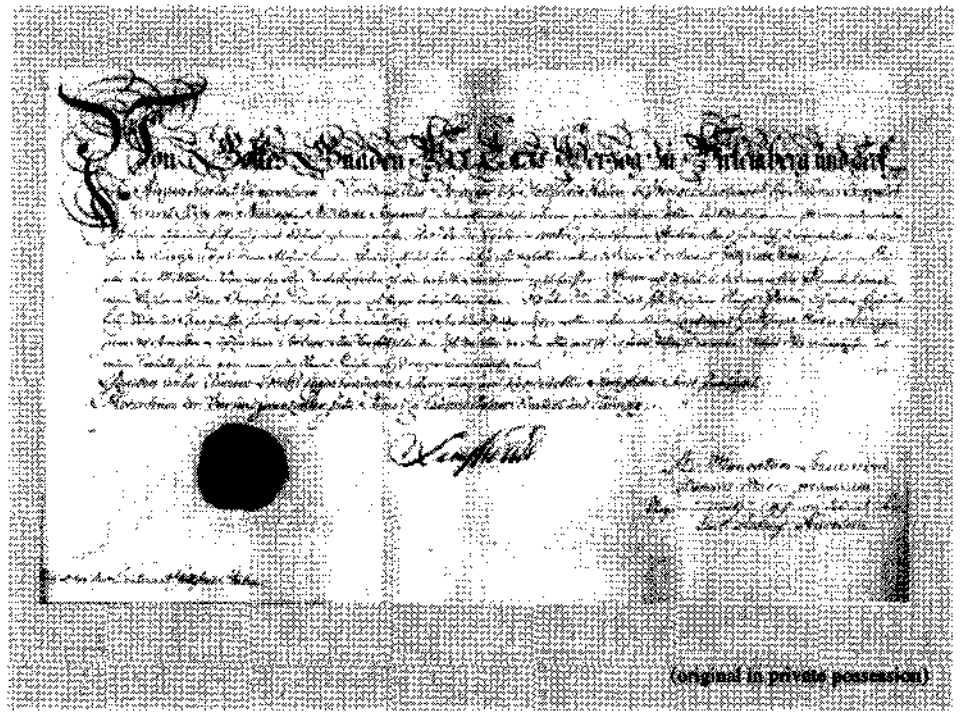
  See position of Constantinopel in Schedel's Nuremberg Chronicle
- *Is Istanbul* the correct translation of *Konstantinopel*?



(Faksimile Köln 2001)

Stylistic and Cultural Transfer?

- an „application" looks different in 14th century and 18th century,
- medieval texts often convey information about heterogenous fields (The *Physiologus* about Jesus Christ, stones and animals); is it an item in a corpus for biology, geology or theology?
- → excludes SMT



(original in private possession)

Translation into What?

- A replica of old documents in other old languages is historically impossible, even a „translation" into the same modern language is problematic.
- translation in a modern foreign language distort the document because of
  - changing language,
  - changing culture,
  - changing domains and facts
  - changing "Zeitgeist"
- modernizing syntax in the same language might be viable.

Modern „Translation“

Verfügung

Der Antragsteller Gottlieb Hahn, derzeit Leutnant beim Artillerie-regiment des Generalmajors von Nicolay möchte seine Laufbahn in einer anderen Position fortsetzen und daher kündigen. Ich will der Verbesserung seiner Karriere nicht im Wege stehen und daher der Auflösung des Vertrages zustimmen. Es sei vermerkt, dass Leutnant Hahn während der Tätigkeit bei uns bei allen Gelegenheiten seine Dienstverpflich-tungen erfüllt hat ...

Karl von Württemberg etc.

Is this the source for an English translation?

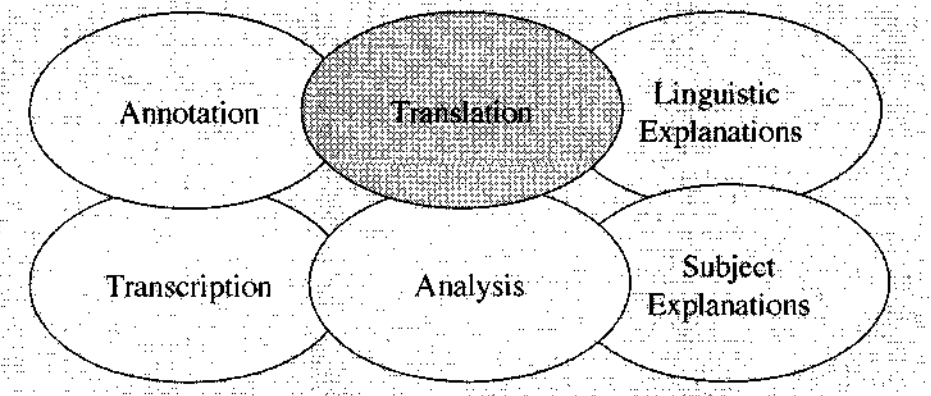
- the text is shorter
- sentences are shorter
- no unknown words, etc.
- completely different text form

Result

So, no translation?

- not in the modern sense,
- a translated transcription (word-to-word) does not meet the needs of most users,
- a replica of old texts in other old languages is historically impossible,
- A solution for scientists in other languages and other fields is
  - an integrating **tool for machine aided translation**, which allows for
  - a modern narrow paraphrase in L2, and
  - a broad documentation of the source text, target text, and translation.

Integration

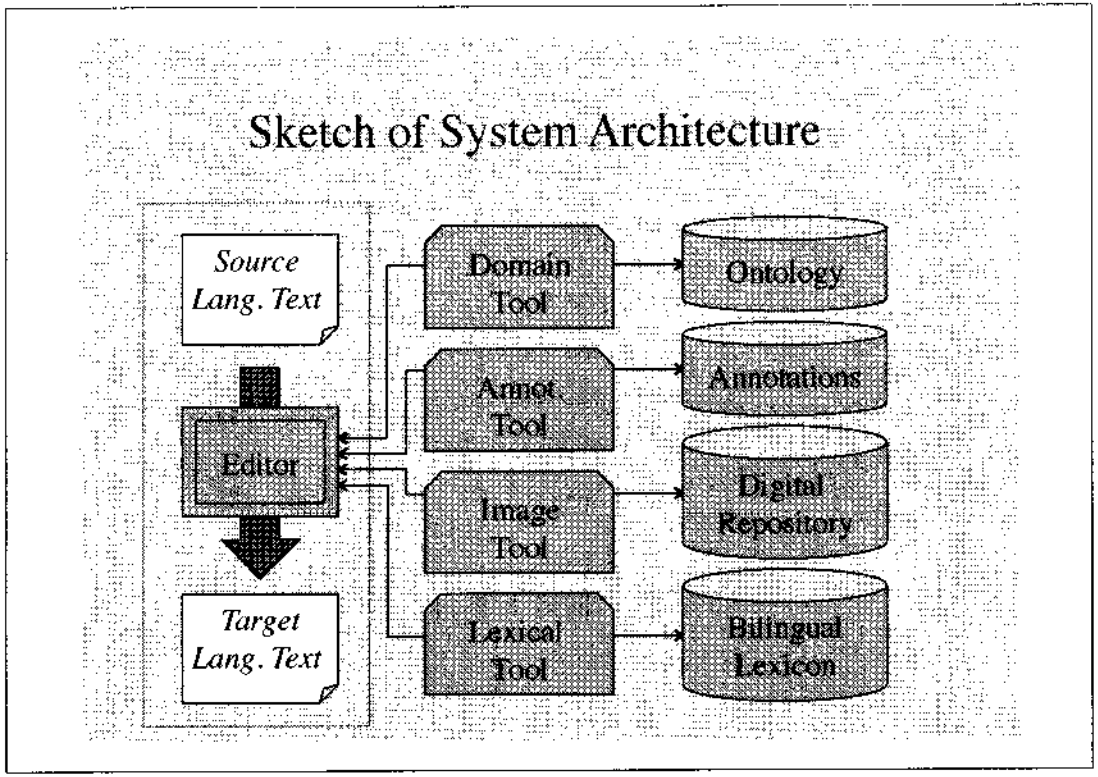


Functional Sketch of a Translation Tool for Historical Texts

- Displays to the translator
  - Digital source, ✔
  - Domain information,
  - Lexical correspondences via one single ontology, ✔
  - Target text marks for source annotations of languages, font styles, size, script, etc.
- Explanation generator for the translator ✔
- XML annotation support even for document images ✔

- Displays to the user
  - Source and target text ✔
  - Interactive target text comments
  - Meta data about the translation, ✔
  - Digital source, ✔

✔ = facility available in other contexts


Sketch of System Architecture
Source Lang. Text
Editor
Target Lang. Text
Domain Tool
Annot. Tool
Image Tool
Lexical Tool
Ontology
Annotations
Digital Repository
Bilingual Lexicon