Text Mining for Historical Documents
Digitisation and Preservation of Digital Data

Caroline Sporleder

Computational Linguistics
Universität des Saarlandes

Wintersemester 2010/11

21.02.2011

Digitisation

Why digitise?

Advantages of Digitisation

- safeguard against data loss (e.g., City Archive of Cologne)
- easier access (for experts)
  - safe access to fragile documents
  - worldwide
  - instant and continuous
  - simultaneous
- virtual exhibitions (for laypersons)
  - draw in new customers
    $\Rightarrow$ increase visitor numbers (museum)
    $\Rightarrow$ stimulate tourism (country)
  - showcase artefacts typically not on display
    (approx. of a collection)
- (semi-)automatic data analysis

Why digitise?

Advantages of Digitisation

- safeguard against data loss (e.g., City Archive of Cologne)
- easier access (for experts)
  - safe access to fragile documents
  - worldwide
  - instant and continuous
  - simultaneous
- virtual exhibitions (for laypersons)
  - draw in new customers
    $\Rightarrow$ increase visitor numbers (museum)
    $\Rightarrow$ stimulate tourism (country)
  - showcase artefacts typically not on display (approx. 95% of a collection)
- (semi-)automatic data analysis

How to digitise?

For textual data

1. digital photographs / scanning
2. optical character recognition (OCR) or manual transcription

Optical Character Recognition

In the ideal case

word error rates of around 1% on typed text (Lopresti, 2005)

However

historical documents often pose problems

Optical Character Recognition

In the ideal case

word error rates of around 1% on typed text (Lopresti, 2005)

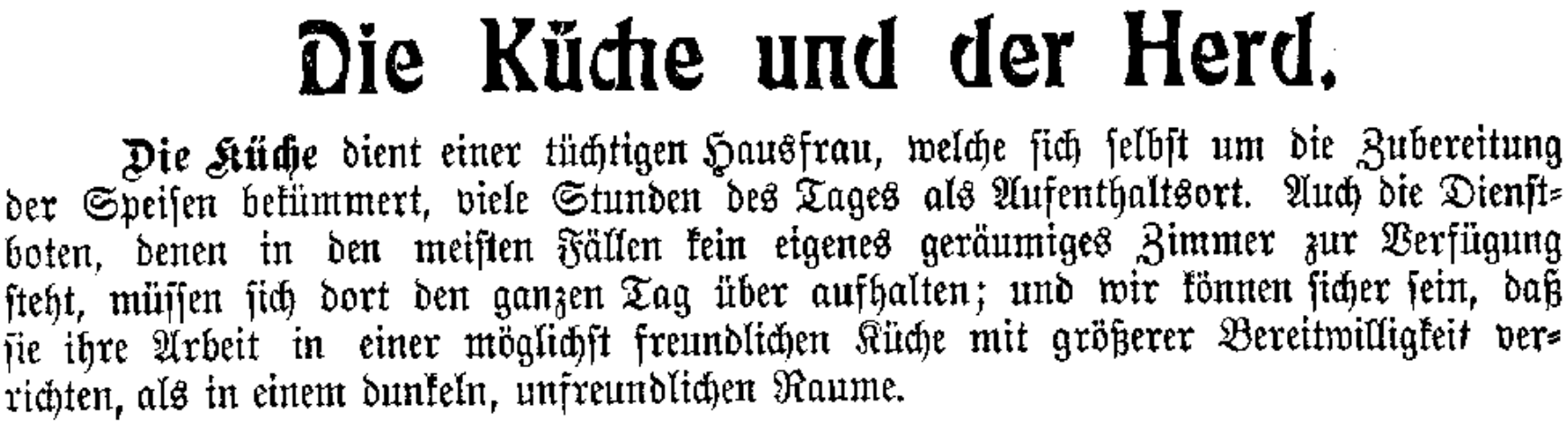
However

historical documents often pose problems

- bad print quality (stains, faded letters etc.)
- old-fashioned language (messes up language models)
- old-fashioned fonts

Example: Fraktur

Source: `http://www.suetterlinschrift.de/Lese/Schriftgeschichte/Fraktur2.htm`



Die Küche und der Herd.

Die Küche dient einer tüchtigen Hausfrau, welche sich selbst um die Zubereitung der Speisen bekümmert, viele Stunden des Tages als Aufenthaltsort. Auch die Dienstboten, denen in den meisten Fällen kein eigenes geräumiges Zimmer zur Verfügung steht, müssen sich dort den ganzen Tag über aufhalten; und wir können sicher sein, daß sie ihre Arbeit in einer möglichst freundlichen Küche mit größerer Bereitwilligkeit verrichten, als in einem dunkeln, unfreundlichen Raume.

OCR Problems

Typical errors

- mixing up letter sequences that look similar (e.g., “ii”, “ü”, “n”, “il”, “li”)
- missing punctuation characters
- replacing letters by space or punctuation

This is problematic because

- it adversely affects retrieval results in keyword-based search
- it hampers further language processing (tokenisation, especially sentence splitting)
- OCR errors are more difficult to correct than typos because the erroneous words deviate often more strongly from the correct word because
  - errors can affect sequences of letters with no one-to-one mapping (e.g., “iii” $\rightarrow$ “m”)
  - errors can affect multiple positions in a word (e.g., “e” $\rightarrow$ “c”)

Correcting OCR Errors

Techniques used for spell checking don't work very well

- Levenshtein distance / edit distance (often high for OCR errors)
- dictionary look-up (problematic for historical data)

Possible solutions

- implement or learn heuristics (e.g., "c" $\longrightarrow$ "e")
- use taylor-made resources (e.g., historical dictionaries)
- combine predictions of different OCR systems (e.g., by voting) (see Volk et al., to appear)
- identify likely variants of a form and use statistics to detect errors (see Reynart, 2008)

Handwritten Documents

OCR infeasible

- word error rates of 50% and more (Rath et al., 2004)
- for historical documents probably even worse

Solution

- manual transcription
- partial manual transcription and search via visual similarity
- automatic alignment on paragraph-/sentence-level to simulate search in original document

Audio Data

Speech is similarly difficult to transcribe

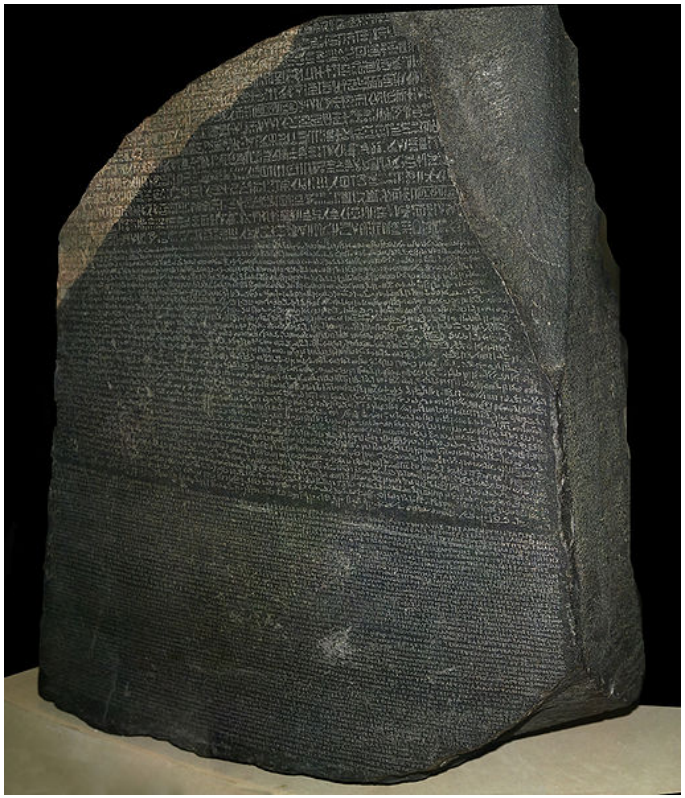
- automatic speech recognition challenging for open domains and multitude of speakers
- historical collections pose additional problems
  - quality of old recordings
  - speaker factors (emotion, non-native speech, old age)
  - possibly somewhat old-fashioned speech

Digital Data Preservation

Data Longevity

Rosetta Stone: approx. 2200 years old

Source: `http://en.wikipedia.org/wiki/File:Rosetta_Stone.JPG`



Data Longevity

Rhind Papyrus: approx. 3660 years

Source:

`http://en.wikipedia.org/wiki/File:Rhind_Mathematical_Papyrus.jpg`

Data Longevity

Floppy Disk: approx. 25 years

Source: `http://en.wikipedia.org/wiki/File:Floppy_disk_2009_G1.jpg`

Digital Data

Typically very short life cycle

Several prominent cases of digital data loss

- NASAs 1976 Viking Mars mission (stored on magnetic tape which deteriorated)
- BBC's Domesday Project, mid 1980s (outdated software, storage format)

Digital Data

Typically very short life cycle

- physical decay (bit rot) (avg. life span of a CD-Rom is 5 years (Horlings, 2003))
- outdated media (e.g., floppy disk)
- outdated formats (e.g., word processing, Word Perfect)
- loss of context (e.g., lost decryption keys)

Several prominent cases of digital data loss

- NASAs 1976 Viking Mars mission (stored on magnetic tape which deteriorated)
- BBC's Domesday Project, mid 1980s (outdated software, storage format)

Data Preservation (1)

Solution

very careful data management (not a solved problem yet)

- adherence to open standards such as XML (as opposed to proprietory formats)
- refreshing (i.e., periodically copying data to new storage media, which are ideally kept in different locations)
- migration (of data to newer storage formats)
- emulation (of outdated software on modern hardware)

Data Preservation (2)

But

challenging for most cultural heritage institutes

- very expensive and time consuming to implement
- amount of data
- heterogeneity of data, plethora of formats
  - databases
  - photos
  - scans
  - word processing documents
  - DOS, Windows, Mac, OS/2, Linux

$\Rightarrow$ most institutes ignore the problem for the time being, no long-term data management