Text-image alignment for historical handwritten documents

S. Zinger

Video Coding and Architectures Research group

Eindhoven University of Technology

the Netherlands

s.zinger@tue.nl

J. Nerbonne

Center for Language and Cognition Groningen,

University of Groningen

the Netherlands

j.nerbonne@rug.nl

L. Schomaker

Artificial Intelligence Department,

University of Groningen

the Netherlands

schomaker@ai.rug.nl

Electronic Imaging 2009

San Jose, USA

18-22 January

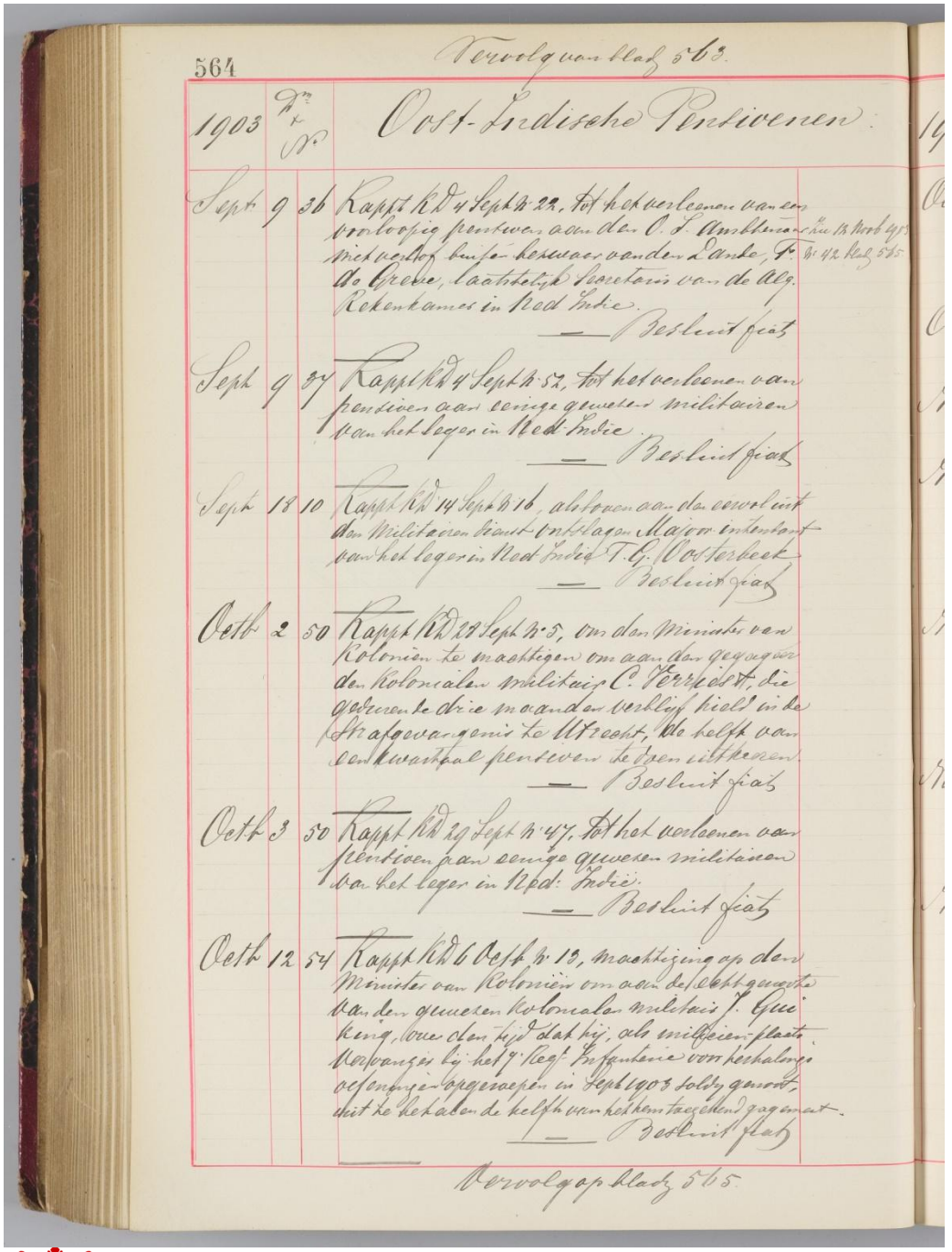
Expected result of SCRATCH – SCRipt Analysis Tools for the Cultural Heritage

enable search through the handwritten text like Google does for the texts in electronic form

Queries to this search engine are words or lines of texts;

retrieved documents are images

Example of initial data



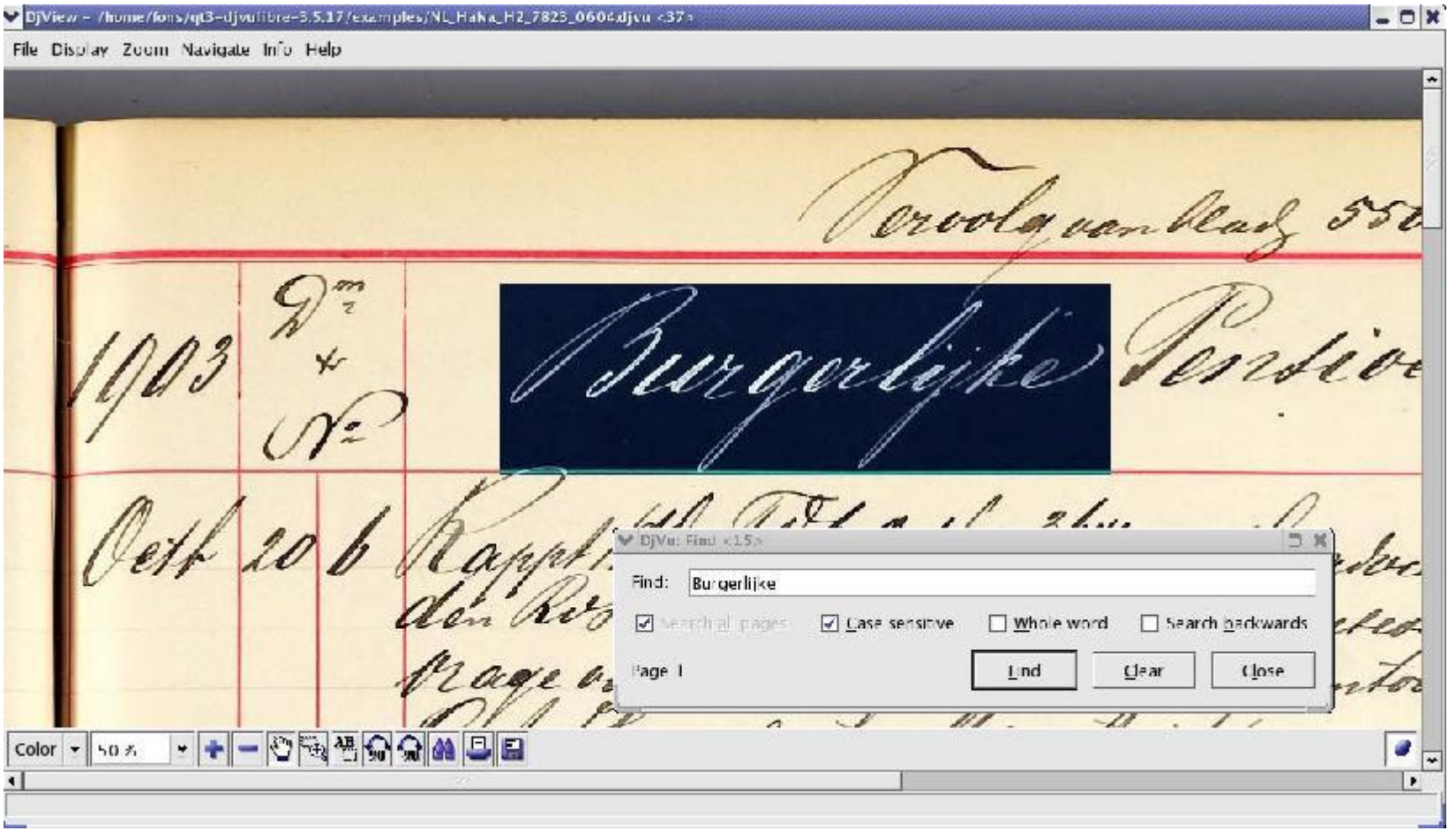
Available data

- 1094 scanned pages of the Royal decrees dated 1903 and provided by the Nationaal Archief in the Hague
- 24772 lines of handwriting manually annotated by ASCII text
- more than 10000 queries received by the National Archive for information retrieval from historical documents

Expected data

- 39 books of Kabinet der Koningin dated 1892-1913
- about 1121 pages in one book => 43700 pages
- 30 lines per page => more than a million lines

How the search engine should work - example



annotated lines

Noord-brabant, Noord-holland en Utrecht

Cursus bij het Kon. Instituut der Marine

burgemeester der gemeente Maasbree J.J.

Segmentation on words

Real case

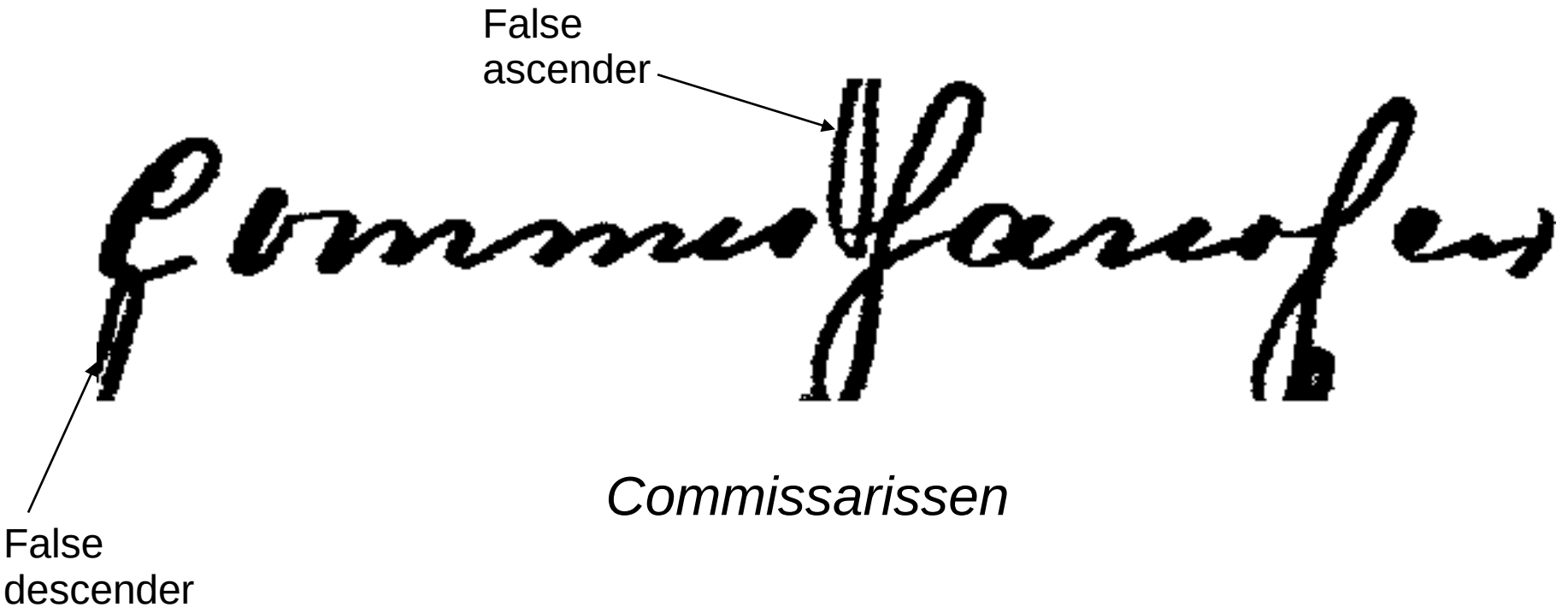
Ideal case

Arrondissement *Rotterdam* *voor* *de* *Kan*

Segmentation method: adaptive (medain value) threshold on the lengths of spaces in the line

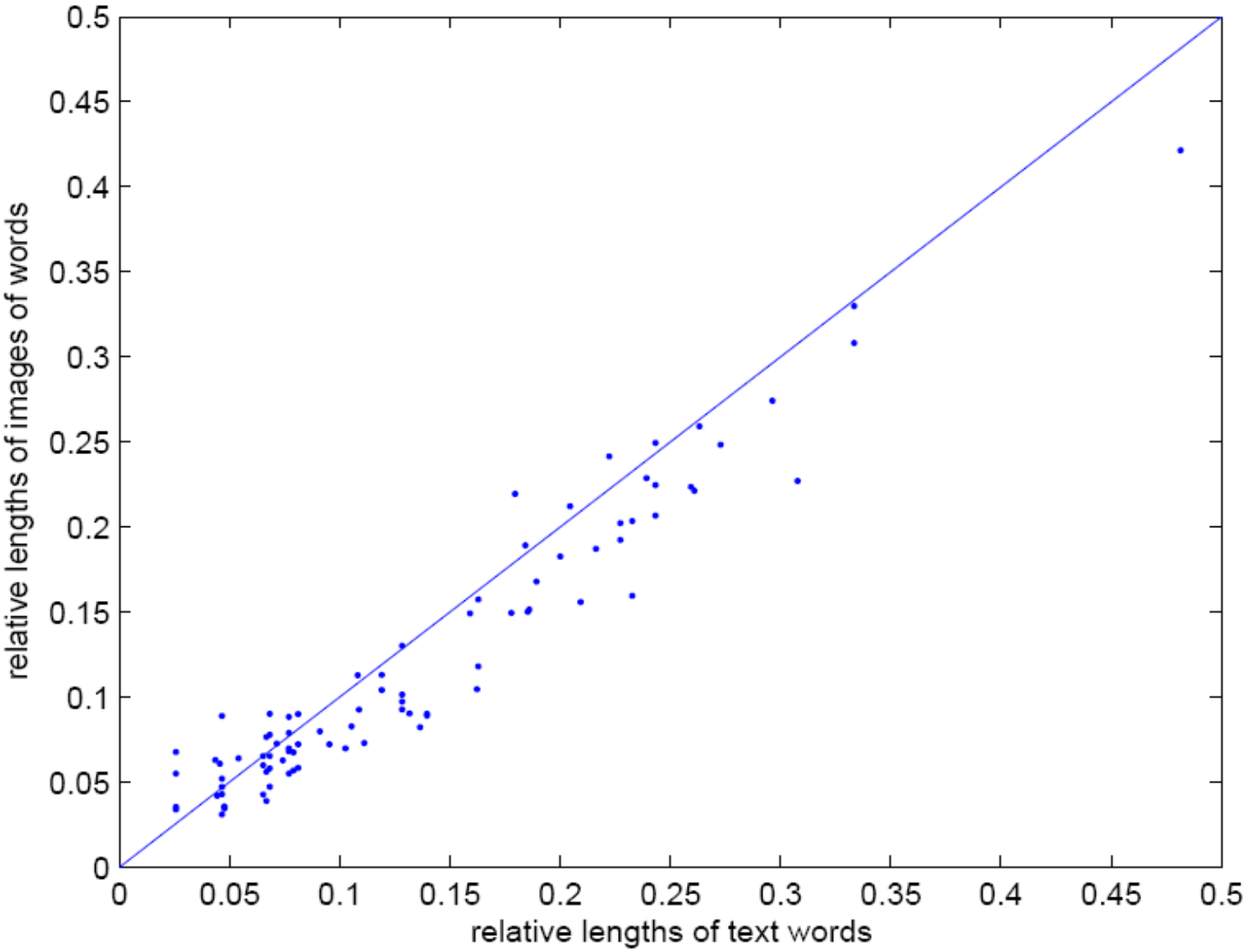
Performance: 60% correctly segmented words from 100 lines

Aligning words with their images



Using ascenders ans descenders for alignment will produce errors as a result of overlapping lines

Relative lengths of words



Correspondance of relative lengths of words in letters and in pixels for 87 words

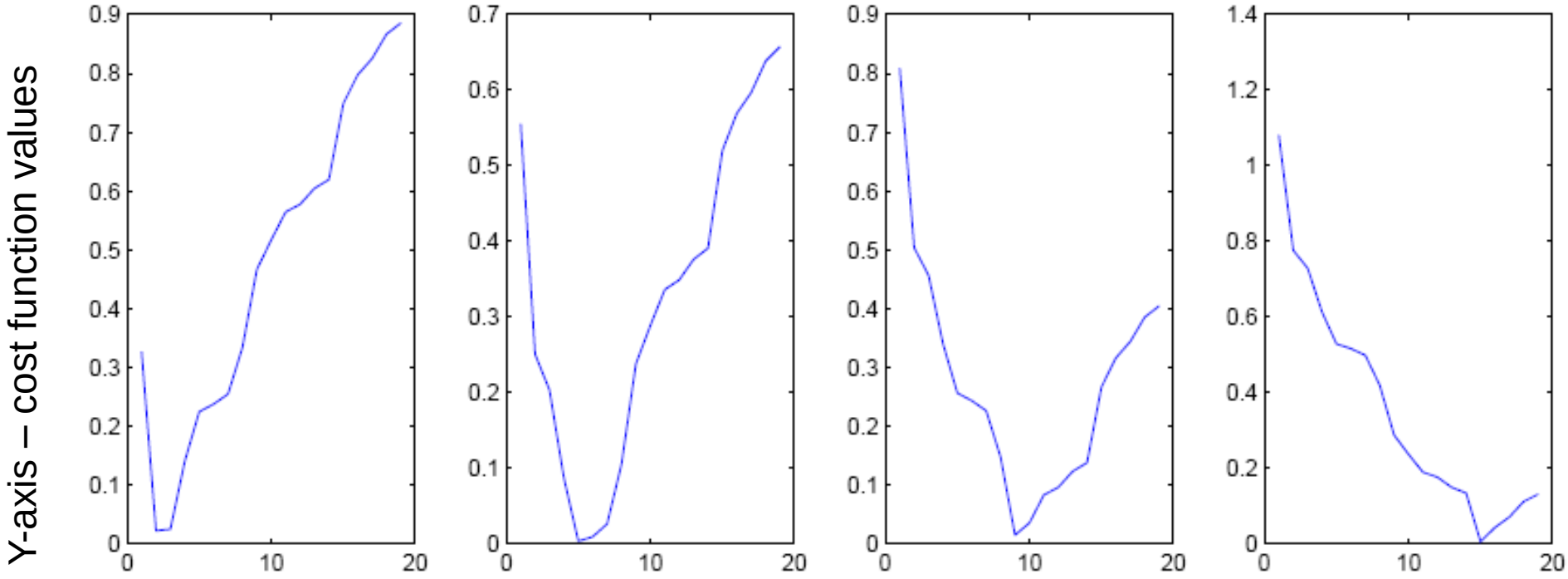
Relative length of word is its length divided by the length of the line that contains this word

Alignment as cost function minimization

- Right to left alignment
  - Fix the left boundary b of a word and find its optimal right boundary $i' = \arg\min f_n(i)$, where
    $f_n(i) = |h(i) - t_n|$ – cost function
    t_n – relative length of text word n,
    $h(i)$ – relative length of a handwritten word
  - Repeat the previous step for every word from the text annotation
- Exhaustive search for the best alignment – all possible combinations of spaces are considered to minimize the cost function

Example of cost functions

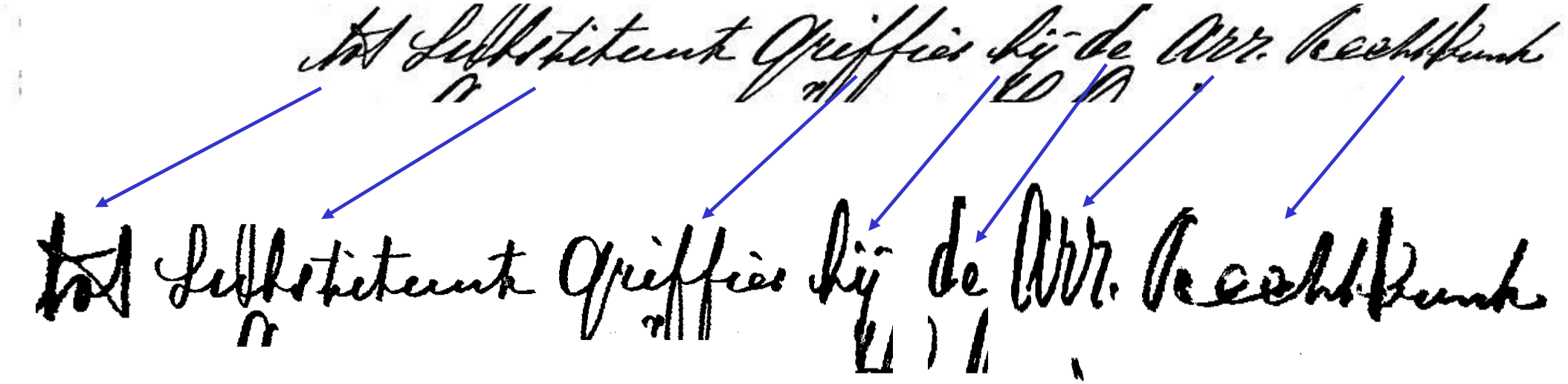
Handwritten line – 4 words



X-axis: ordinal number of boundary of the word on the image of the handwritten line

Experiments on alignment: example of correct automatic alignment

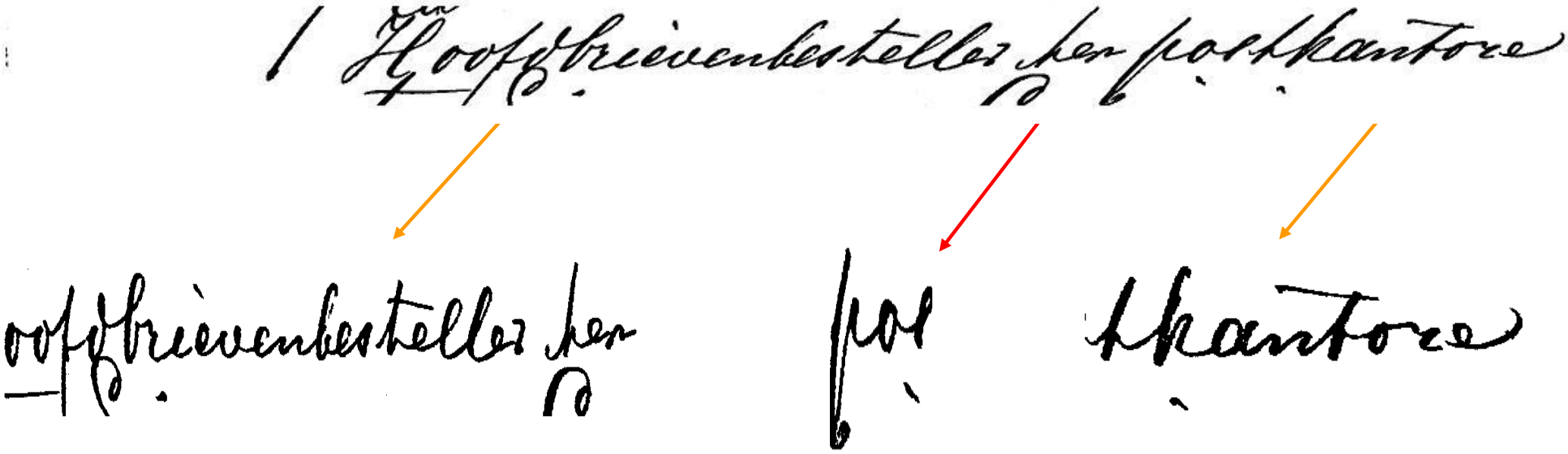
Scan 74, line 5



tot *Substituut* *Griffier* *bij* *de* *Arr.* *Rechtbank*

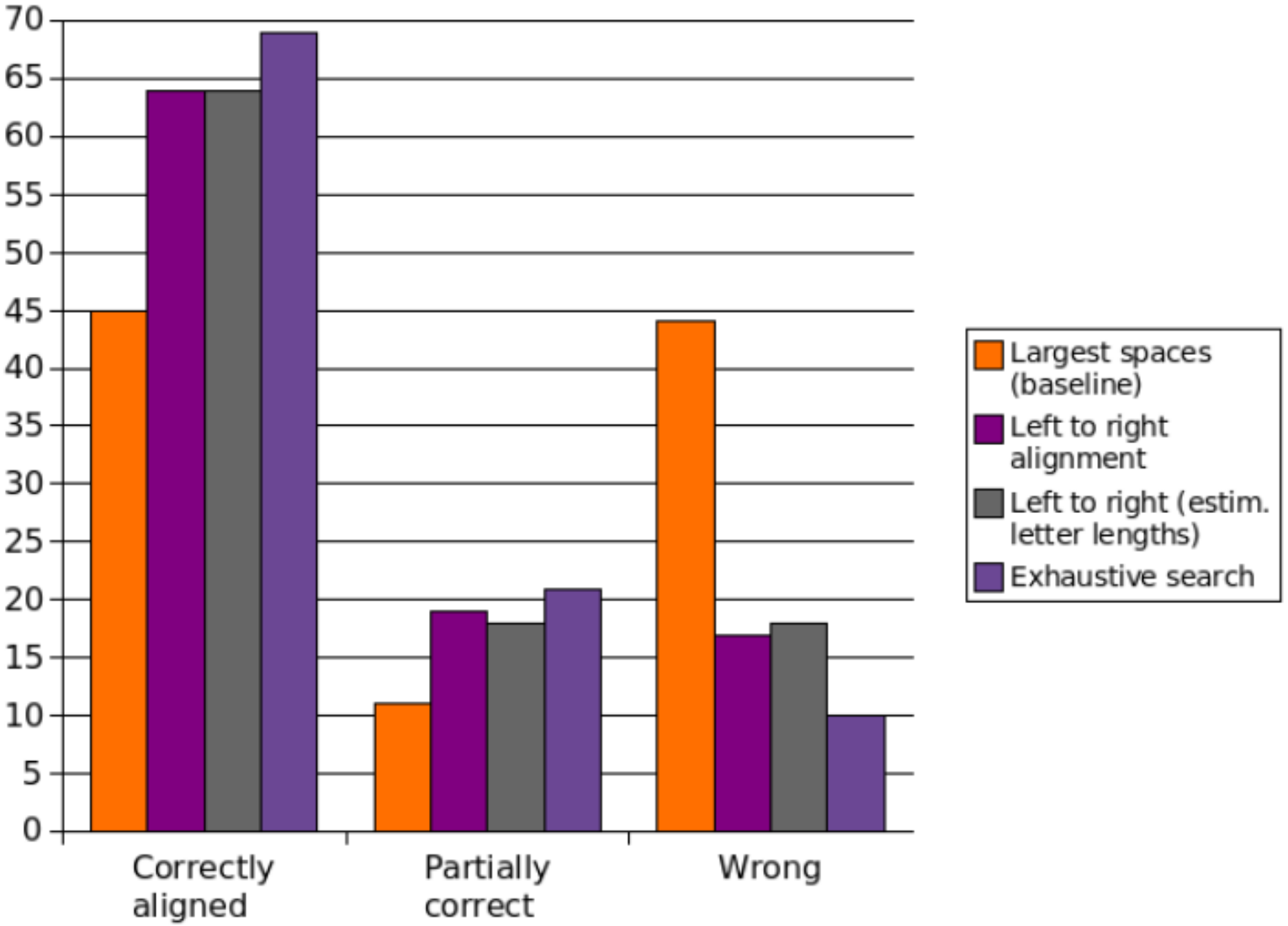
Experiments on alignment: example of incorrect automatic alignment

Scan 810, line 19



Hoofdbrievenbesteller ter *pos* *tkantore*

Alignment results on 100 lines



Conclusions and future work

- The best alignment is obtained by cost function minimization with exhaustive search
- Proposed method depends neither on actual words lengths nor on pattern recognition: variations in handwriting will not affect the alignment
- Text of transcriptions may be rendered and then aligned with image
- Upper-case letters or first characters of words can be detected and used to improve the alignment