

# 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

schomaker@ai.rug.nl

# 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

564

Vervolg van bladz 563.

1903	<sup>7<sup>de</sup></sup> D <sup>o</sup>	Oost-Indische Postovereen.
Sept 9	56	Rapport RD 4 Sept 18 28, tot het verlaanen van een vroeloopig pentewer, aan den C. J. Amstman, die 18 Nov 1895 met verlof bij de heer van der Sande, F. de Groot, Raadslid Secretaris van de Algem. Rekenkamer in Ned. Indië. — Besluit geat
Sept 9	57	Rapport RD 4 Sept 18 52, tot het verlaanen van pentewer aan eenige gewone militairen van het leger in Ned. Indië. — Besluit geat
Sept 18	10	Rapport RD 14 Sept 18 16, aflopen van de overloopt der Militairen bij de Ontslag Major en luitenant van het leger in Ned. Indië T. G. Oosterbeek. — Besluit geat
Oct 2	50	Rapport RD 18 Sept 18 5, om den Minister van Koloniën te maachten om aan den gequize den Kolonialen militair C. Verriest, die geveente drie maanden verlof hielde in de Strafgewarmerij te Utrecht, de helft van den wasschal pentewer te open uitheren. — Besluit geat
Oct 3	50	Rapport RD 29 Sept 18 47, tot het verlaanen van pentewer aan eenige gewone militairen van het leger in Ned. Indië. — Besluit geat
Oct 12	54	Rapport RD 6 Oct 18 18, maatching op den Minister van Koloniën om aan de luitenant van den gewonen Kolonialen militair J. G. King, bene den tijd dat hij, als militair, staet te verzuimen bij het 7 <sup>de</sup> Regt. Buffenarie van kechalinge oefeninge opgeroepen in Sept 1903 solde geveent, dat he kechalen de helft van het den tusschen tusschen at. — Besluit geat

Vervolg op bladz 565.

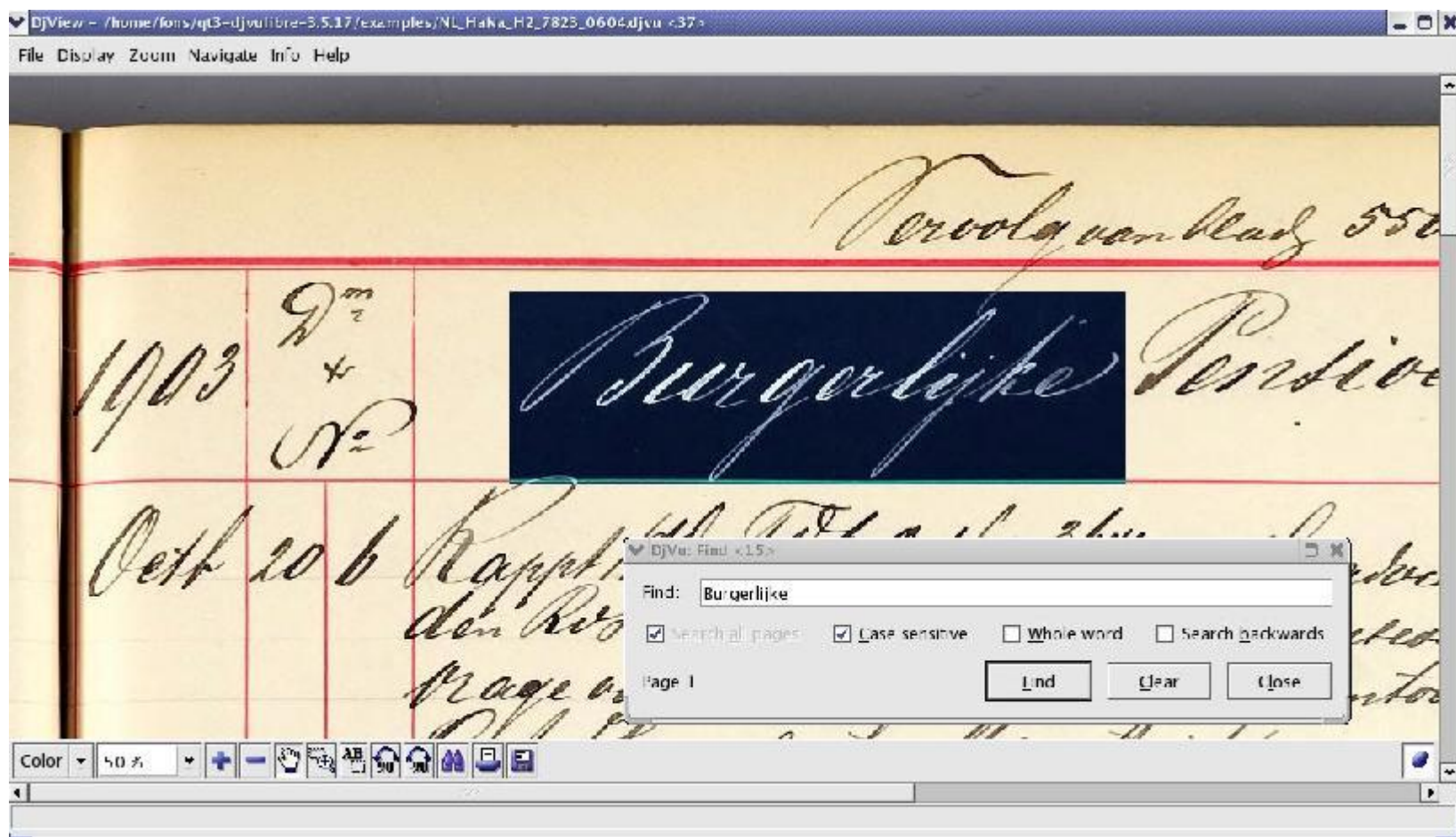
## Available data

- 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

*Noordbrabant, Noordholland en Utrecht*

Noord-brabant, Noord-holland en Utrecht

*Cursus bij het Kon. Instituut der Marine*

Cursus bij het Kon. Instituut der Marine

*burgemeester der gemeente Maasbree J.J.*

burgemeester der gemeente Maasbree J.J.

# Segmentation on words

Real case



Arrondissement Rotterdam voor de Kan

Ideal case



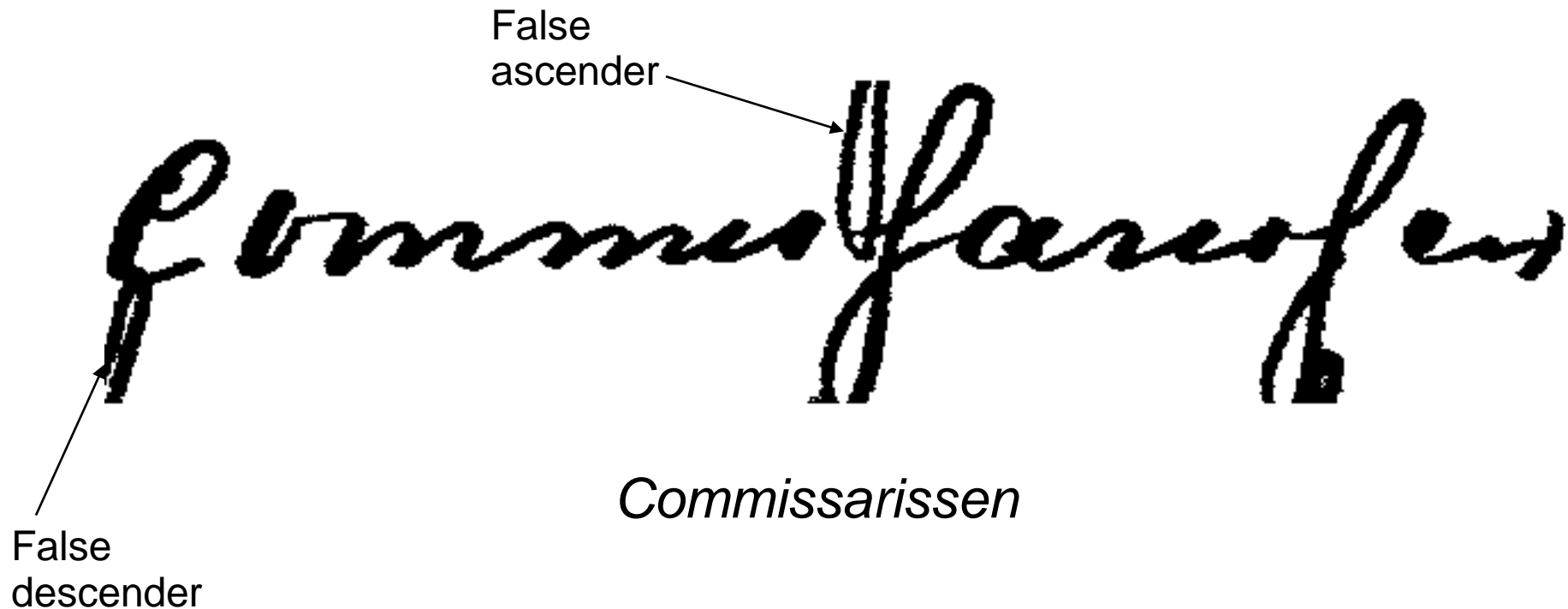
Arrondissement Rotterdam voor de Kan

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

Performance: 60% correctly segmented words from 100 lines

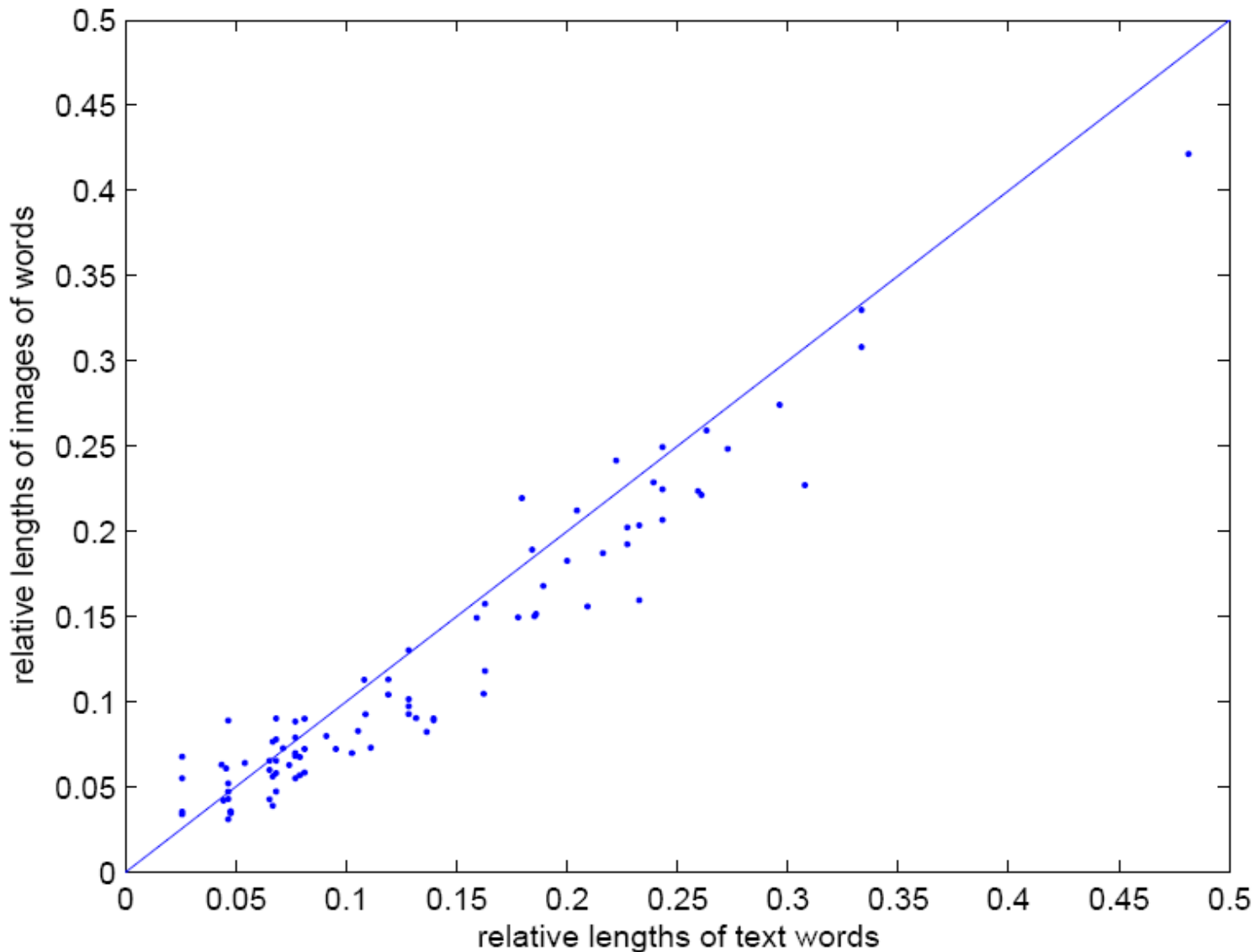


# Aligning words with their images



Using ascenders and 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

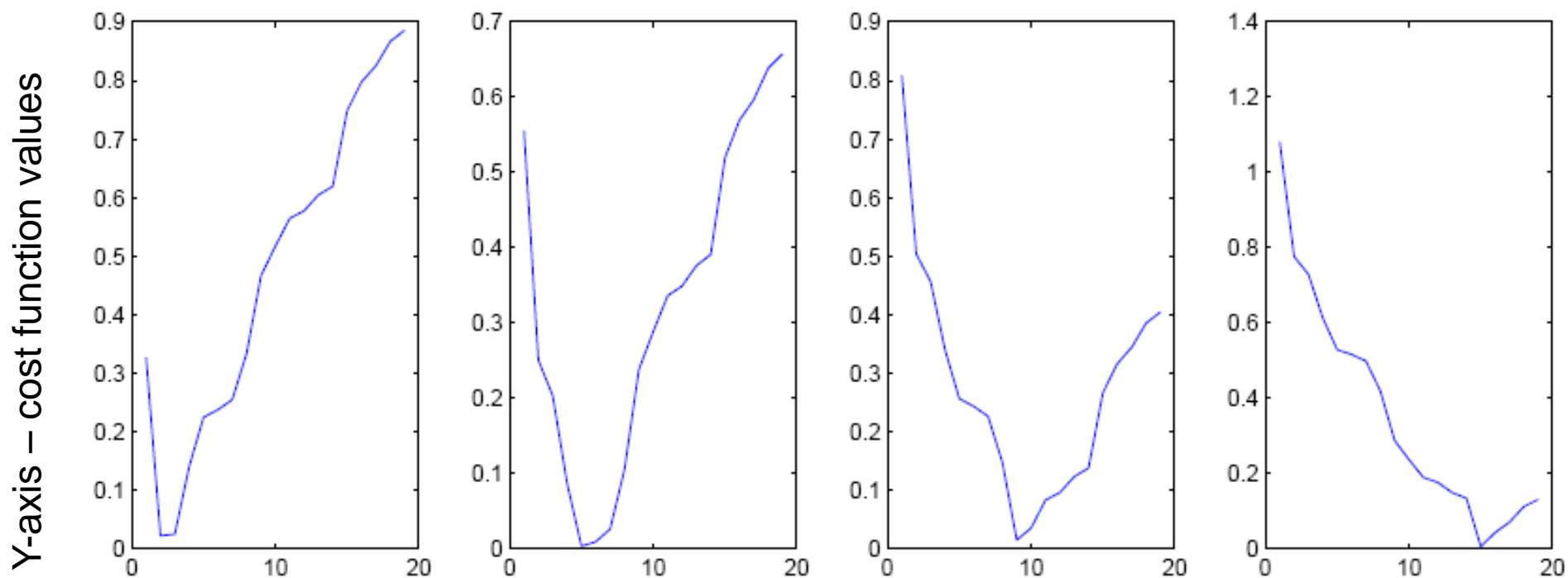
# Alignment as cost function minimization

- 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
$$f_n(i) = |h(i) - t_n| - \text{cost function}$$
$$t_n - \text{relative length of text word } n,$$
$$h(i) - \text{relative length of a handwritten word}$$
  - 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

*Landstede Algemeene Ordeelingen Verzekerung*

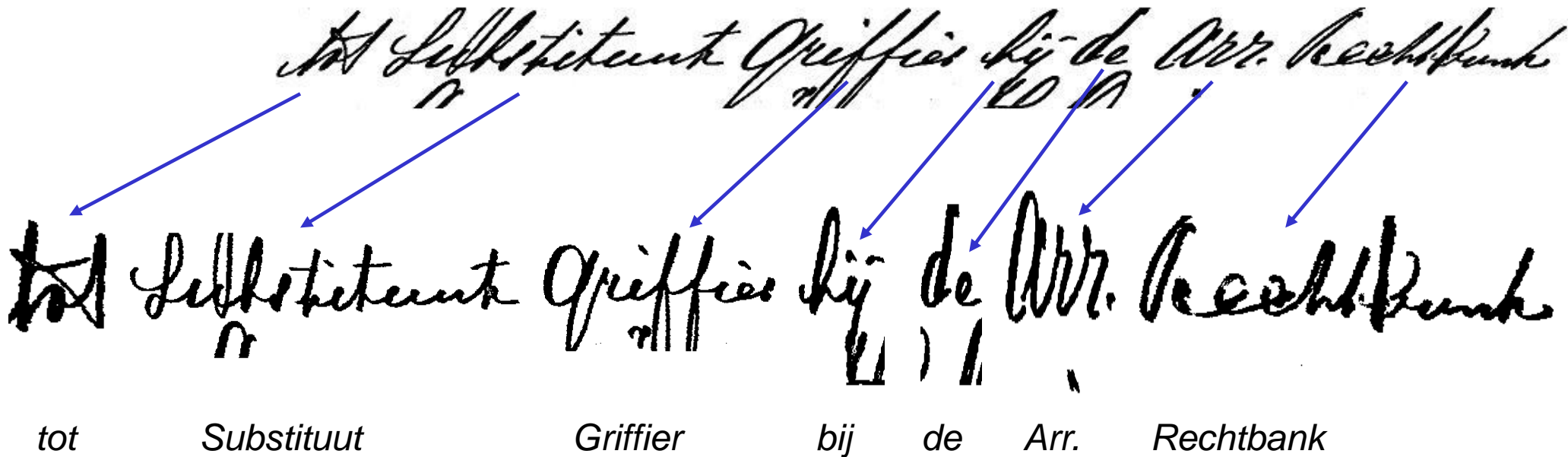
Handwritten line – 4 words



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



The image shows two lines of handwritten text in cursive script. The top line is the original scan, and the bottom line is the result of automatic alignment. Blue arrows point from the original text to the aligned text, showing how the words are correctly aligned despite the cursive style. Below the aligned text, the words are printed in a standard font for clarity.

tot Substituut Griffier bij de Arr. Rechtbank

# Experiments on alignment: example of incorrect automatic alignment

Scan 810, line 19

1 <sup>van</sup> Hoofdbrievenbesteller per postkantore

oofdbrievenbesteller per

Hoofdbrievenbesteller ter

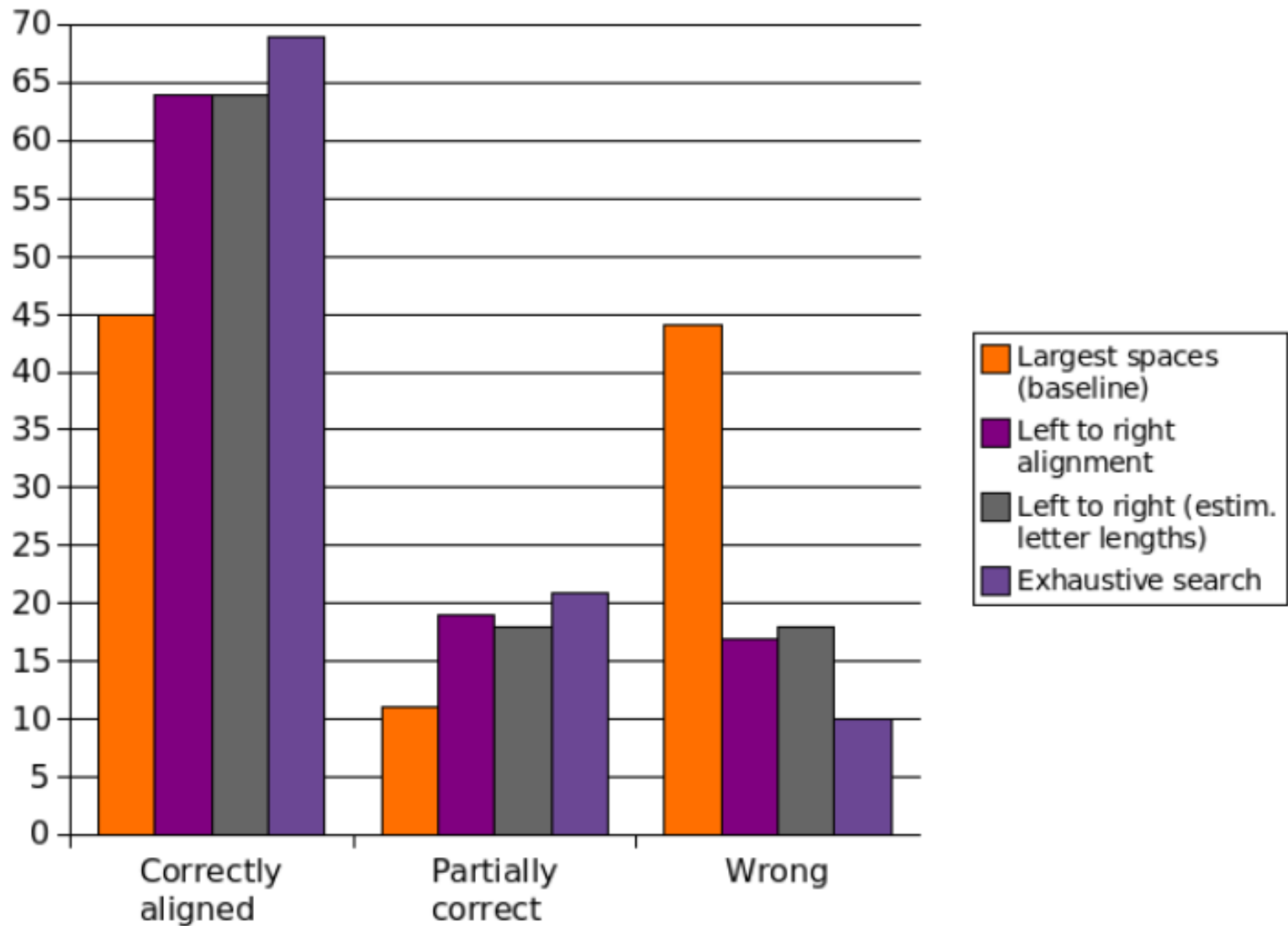
pos

pos

tkantore

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