Vorbespechung/Introductory Meeting: Text Mining for Historical Documents

Caroline Sporleder and Martin Schreiber

Computational Linguistics &
Kultur- und Mediengeschichte
Universität des Saarlandes

Wintersemester 2010/11

18.01.2011

Organisational Stuff

What is it?

Project Seminar

- a theoretical part (class presentations)
- a practical part (group work: design and implementation of a small or not so small project)
- interdisciplinary (history students and CS/Coli students)

Course Objectives

- overview of text mining techniques
- awareness of challenges of the cultural heritage domain
- hands-on experience with text mining techniques
- working in an interdisciplinary environment and communicating with non-experts

Requirements

No previous knowledge of text mining or NLP is necessary, however:

- interest in language (processing) and
- interest in history
- programming experience

. . . are essential!

There will also be **serious group work**!

Credits

Coli Students

- 5 CPs (150 hours)
- presentation (20-30 minutes)
- participation in group work (practical work / implementation)
- written report (one per group)

Computer Science Students

- 7 CPs (210 hours)
- presentation (20-30 minutes)
- written summary of the presentation or another topic from the theoretical part (around 5 pages)
- participation in group work (larger amount of practical work than Coli students)
- written report (one per group)

$\Rightarrow$ presentations should be prepared for the first day of the course

$\Rightarrow$ the written reports/summaries are due on 1.4.2011

Composition of Grades

- 70% project work and report (one grade per group, unless this would lead to serious unfairness)
- 30% for presentation (and summary)

Evaluation Criteria (1)

Presentations

Content

- main aspects of the paper(s) covered
- good explanations of content (examples, graphics)
- **main points of presentation understandable to non-experts**
- discussion of advantages/disadvantages of presented method
- use of additional (printed) references (optionally)

Form

- slides contain references
- slides are well-structured
- the topic was well-presented (interaction with audience etc.)
- the presentation is not too short
- the presentation is not (much) too long

Evaluation Criteria (2)

Group Work and Report

- project is well thought through (relevance for historians, use of existing methodology, design and implementation)
- report is well-written (motivation/relevance, references to published work, description of design)
- how well did the group work together?

When and Where

When: 21.2.-04.03.2011/11.3.2011
9:30(?)-18:00

Where: to be decided

Background

IT And Cultural Heritage (CH) (1)

A growing field:

- cultural heritage institutes possess large collections of data (primary data and metadata)
- more and more digitisation projects (national and international):
  - digitisation as a safeguard against data loss
  - governments have come to see CH as a valuable asset
  - digitised data can be accessed more easily

IT And Cultural Heritage (CH) (2)

Unlocking the value of CH collections

- efficient retrieval of information is crucial (e.g. automatic searching)
- but more is needed for sophisticated information access (e.g., automatic structuring of unstructured documents, linking of related documents, disambiguation of person names etc.)
- much primary and nearly all metadata are textual
  $\Rightarrow$ text mining and natural language processing (NLP) play a big role

Challenges

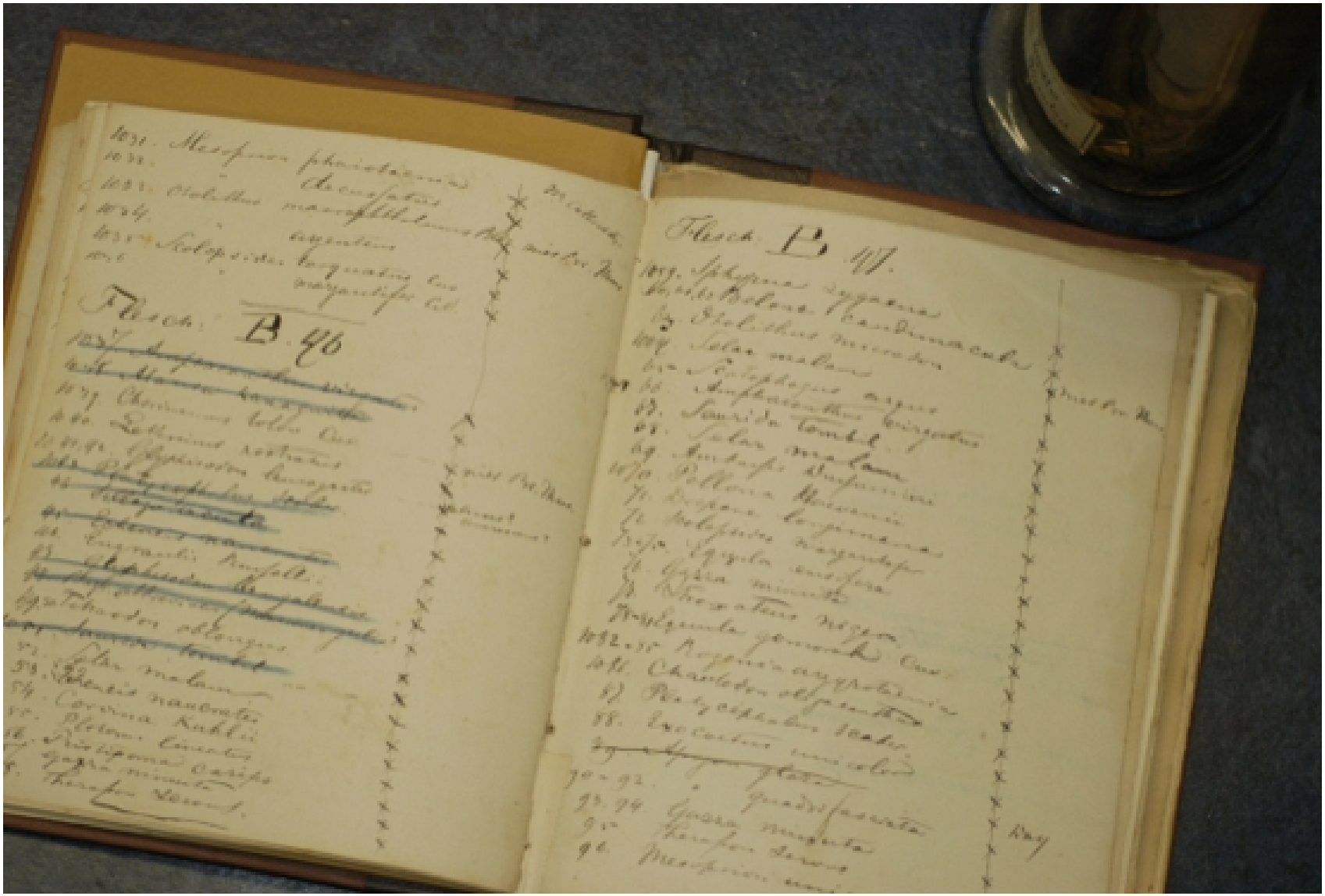
CH as a testbed for NLP technology

- noise introduced during digitisation
- archaic language, 'free' orthography
- few NLP resources for these domains
- multi-lingual and multi-modal content
- variety of data formats and knowledge representation standards (hamper interoperability between collections)

$\Rightarrow$ tools need to be very robust and make use of all available resources

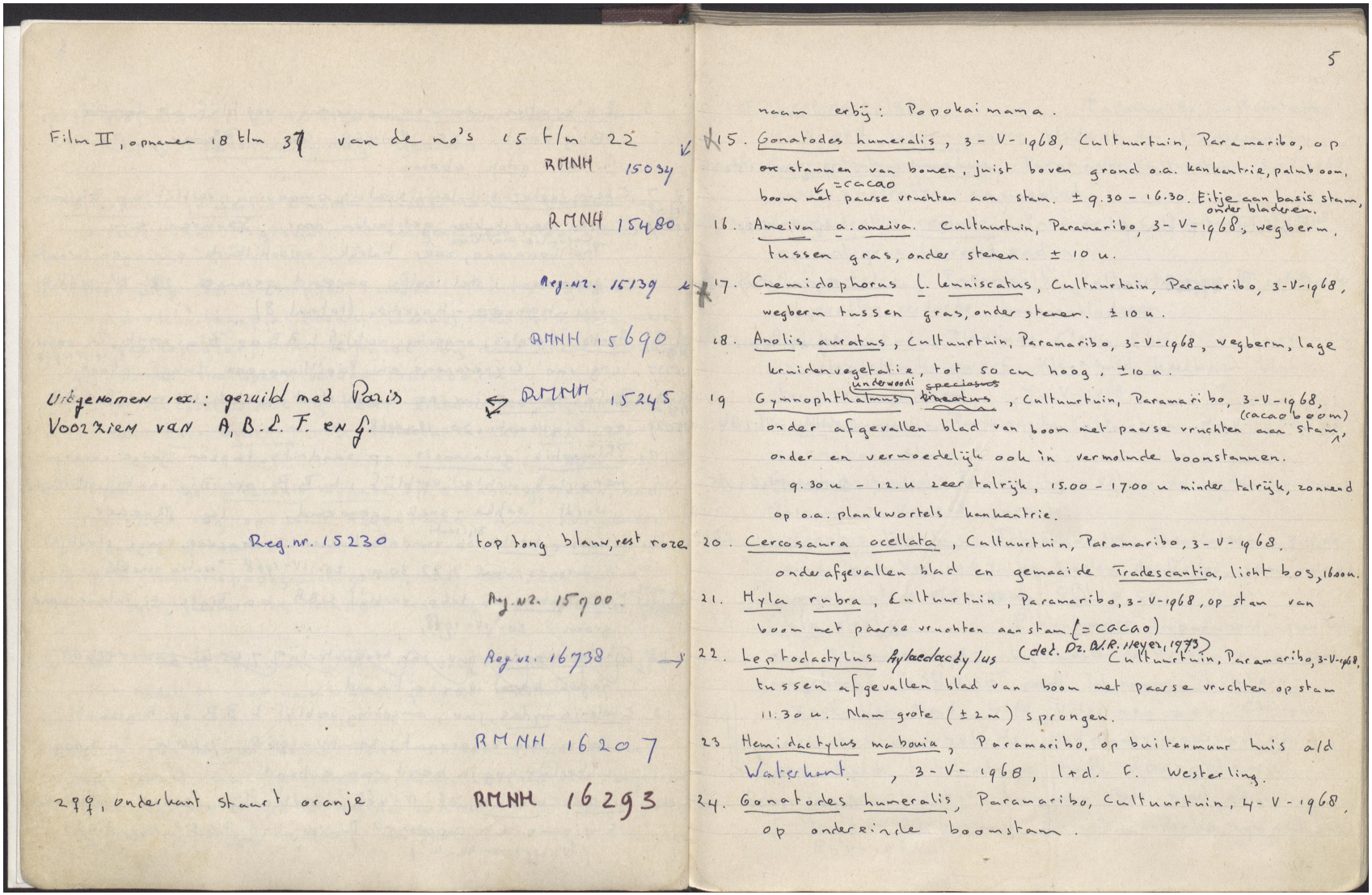
Example: From paper to information (1)

Original Manuscript



Example: From paper to information (2)

Digital Copy



5

naam erbij: Popokai mama.

Film II, opnamen 18 t/m 37 van de no's 15 t/m 22 RMNH 15034

15. Gonatodes humeralis, 3-V-1968, Cultuurtuin, Paramaribo, op ons stammen van bomen, juist boven grond o.a. kankantrie, palmboom, boom met paarse vruchten aan stam. ± 9.30 - 16.30. Eitje aan basis stam, onder bladeren. (=cacao)

RMNH 15480

16. Ameiva a. ameiva. Cultuurtuin, Paramaribo, 3-V-1968, wegberm, tussen gras, onder stenen. ± 10 u.

Reg.nr. 15139

17. Cnemidophorus l. lemniscatus, Cultuurtuin, Paramaribo, 3-V-1968, wegberm tussen gras, onder stenen. ± 10 u.

RMNH 15690

18. Anolis auratus, Cultuurtuin, Paramaribo, 3-V-1968, wegberm, lage kruidenvegetatie, tot 50 cm hoog, ± 10 u.

Uitgenomen ex.: gezuild met Paris Voorzien van A, B. L. F. en ♂. ⇔ RMNH 15245

19. Gymnophthalmus (underwoodi) lineatus (speciosus), Cultuurtuin, Paramaribo, 3-V-1968, onder afgevallen blad van boom met paarse vruchten aan stam, (cacaoboom) onder en vermoedelijk ook in vermolmde boomstammen. 9.30 u - 12.u zeer talrijk, 15.00 - 17.00 u minder talrijk, zonnend op o.a. plankwortels kankantrie.

Reg.nr. 15230 top long blauw, rest roze

20. Cercosaura ocellata, Cultuurtuin, Paramaribo, 3-V-1968, onder afgevallen blad en gemaaide Tradescantia, licht bos, 16 o.m.

Reg.nr. 15900

21. Hyla rubra, Cultuurtuin, Paramaribo, 3-V-1968, op stam van boom met paarse vruchten aan stam (=cacao)

Reg.nr. 16738 →

22. Leptodactylus fylaedactylus (det. Dr. W.R. Heyer, 1973) Cultuurtuin, Paramaribo, 3-V-1968, tussen afgevallen blad van boom met paarse vruchten op stam 11.30 u. Nam grote (± 2 m) sprongen.

RMNH 16207

23. Hemidactylus mabouia, Paramaribo, op buitenmuur huis a/d Waterkant, 3-V-1968, 1 ♂ d. F. Westerling.

2 ♀♀, onderkant staart oranje RMNH 16293

24. Gonatodes humeralis, Paramaribo, Cultuurtuin, 4-V-1968, op onderzijde boomstam.

Example: From paper to information (3)

Transcript

1 ex. Phyllobates femoralis At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij Phyllobates femoralis.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 Eleutherodactylus zeuctotylus 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, l [plus] d M. S. Hoogmoed.

Finding Information: Problems with keyword search

Aim: find all specimens of "Phyllobates femoralis"

1 ex. Phyllobates femoralis At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij Phyllobates femoralis.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 Eleutherodactylus zeuctotylus 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, l [plus] d M. S. Hoogmoed.

Finding Information: Problems with keyword search

Aim: find all specimens of "Phyllobates femoralis"

1 ex. Phyllobates femoralis At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij Phyllobates femoralis.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 Eleutherodactylus zeuctotylus 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, l [plus] d M. S. Hoogmoed.

Segmenting Fieldbook Entries

⇒ Number, Genus, Species, Subspecies, Biotope, Location, Date/Time, Remarks,

1 ex. Phyllobates femoralis At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij Phyllobates femoralis.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 Eleutherodactylus zeuctotylus 1 Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

Beyond keyword search

Aim: find all specimens of "Phyllobates femoralis"
Query: Genus=Phyllobates and Species=femoralis

1 ex. Phyllobates femoralis At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij Phyllobates femoralis.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 Eleutherodactylus zeuctotylus 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

Beyond keyword search

Aim: find all specimens of "Phyllobates femoralis"
Query: Genus=Phyllobates and Species=femoralis

1 ex. Phyllobates femoralis At base of tree on small island, primary forest, 20.45-22.00 u. RMNH 23865

Lithodytes lineatus, Brownsberg, aan voet, onder stuk rot hout, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal strepen heldergeel, tekening op dijen vuurrood, veel feller als bij Phyllobates femoralis.

Gonyocephalus auritus Meyer, 3 ex. (1 juv.), Misool. Hoedt 1867.

RMNH 17656 Eleutherodactylus zeuctotylus 1 [vrouw] Lelygebergte, 4 km N.O. van airstrip, distr. Marowijne, Suriname, 19-VIII-1975, onder stuk hout, 610m, 1 [plus] d M. S. Hoogmoed.

Topics

Main Areas for Presentation

- Pre-processing
- Preservation of Digital Data
- Non-Standard Language
- Semantic Web
- Metadata
- Information Extraction
- Text Mining
- Multi-Modal Data
- Personalisation

A detailed list of topics can be found on the web `http://www.coli.uni-saarland.de/courses/tm-hist11/readings.html`.