# Vorbespechung: Text Mining for Historical Documents / Historische Quellen mit (computer-)linguistischen Methoden analysieren

Caroline Sporleder and Martin Schreiber

Computational Linguistics &
Kultur- und Mediengeschichte
Universität des Saarlandes

Wintersemester 2008/09

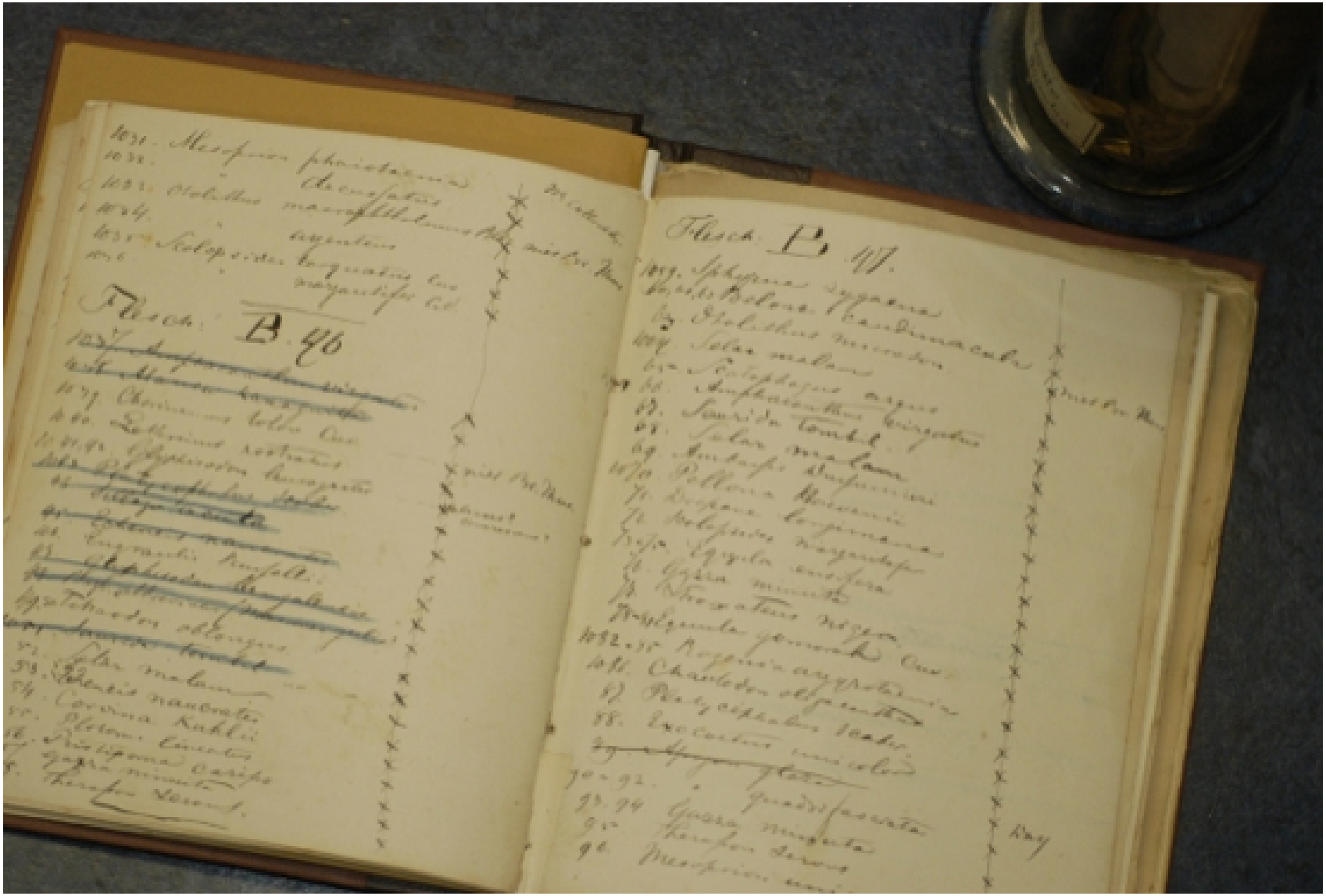
12.01.2009

# Motivation

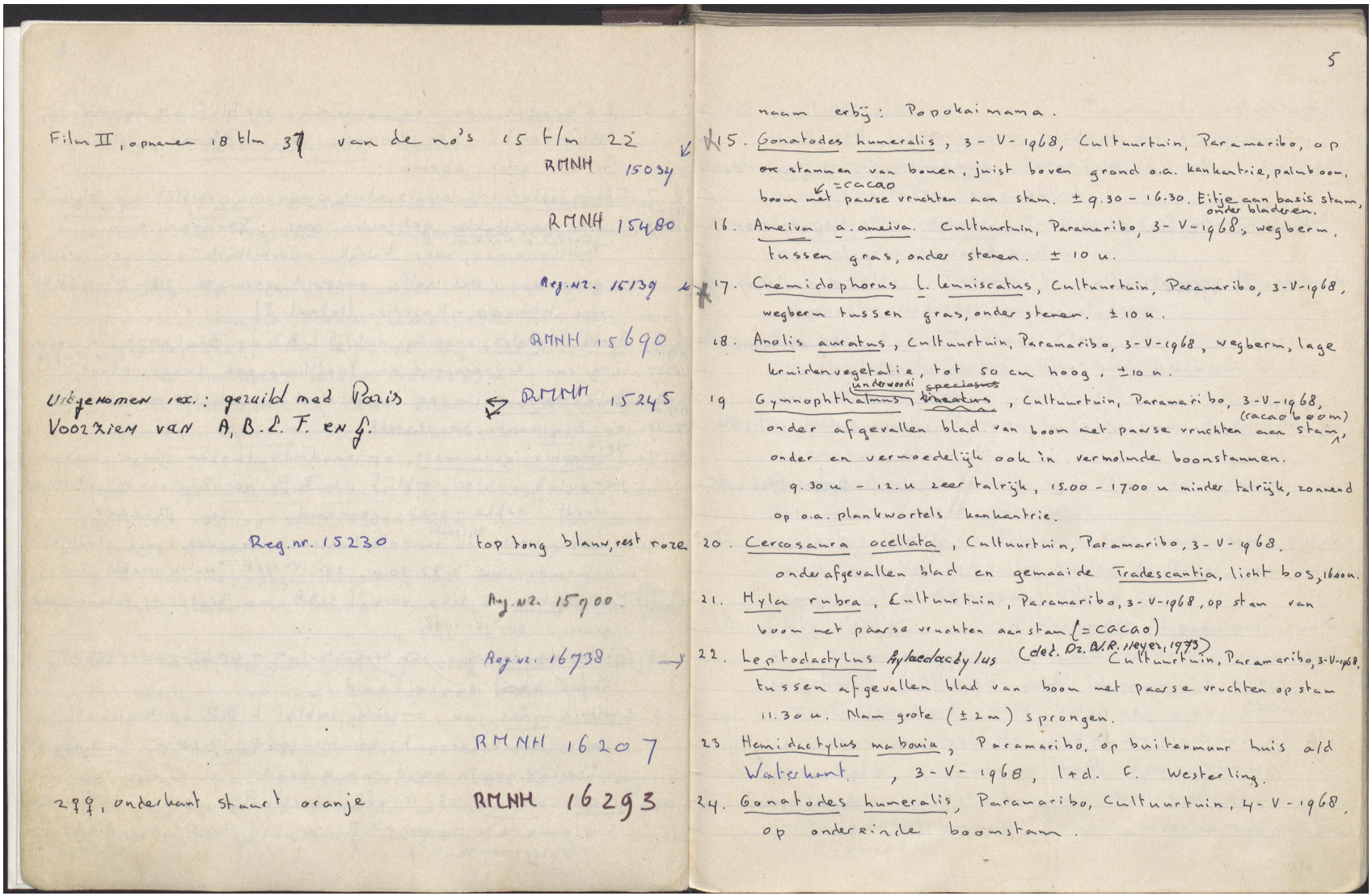
## Why is it relevant?

- museums and archives posess large collections of manuscripts and other data, which are an invaluable resource for researchers
- efficient retrieval of information is crucial (e.g. automatic searching)
- currently several large-scale digitisation projects are underway (e.g., producing digital images of hand-written documents, scanning / optical character recognition, transcription)
- but more is needed for sophisticated information access (e.g., automatic structuring of unstructured documents, linking of related documents, disambiguation of person names etc.)

# The Digitisation Process: Original Manuscript



# The Digitisation Process: Digital Copy



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

RMNH 15480

Reg.nr. 15139

RMNH 15690

Uitgenomen 1 ex.: geruild met Paris
Voorzien van A, B.L. F. en fg. RMNH 15245

Reg.nr. 15230 top long blauw, rest roze

Reg.nr. 15900

Reg.nr. 16738

RMNH 16207

2 ♀♀, onderkant staart oranje RMNH 16293

5

naam erbij: Popokai mama.

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

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

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

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

19 Gymnophthalmus (onderwoud) speciosus lineatus, 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.

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

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

22. Leptodactylus Pyllodactylus (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.

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

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

# The Digitisation Process: Transcription

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

Phyllobates femoralis, post Conini, Coeroenirivier, forrest, 25-VI-1968, 8.30-13.30 u. RMNH 26127-26129

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

Lithodytes lineatus, Brownsberg, at the base, under piece of wood, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal stripes light yellow, marks on the side bright red, much stronger than for Phyllobates femoralis.

# Finding Information: Problems with keyword search

**Aim:** find all specimens of "Phyllobates femoralis"

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

Phyllobates femoralis, post Conini, Coeroenirivier, forrest, 25-VI-1968, 8.30-13.30 u. RMNH 26127-26129

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

Lithodytes lineatus, Brownsberg, at the base, under piece of wood, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal stripes light yellow, marks on the side bright red, much stronger than for Phyllobates femoralis.

# Finding Information: Problems with keyword search

**Aim:** find all specimens of "Phyllobates femoralis"

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

Phyllobates femoralis, post Conini, Coeroenirivier, forrest, 25-VI-1968, 8.30-13.30 u. RMNH 26127-26129

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

Lithodytes lineatus, Brownsberg, at the base, under piece of wood, 13.07.1968, 8.45 u., RMNH 26076 Dorsolateraal stripes light yellow, marks on the side bright red, much stronger than for Phyllobates femoralis.

# This course

## Structure

- open for students from computational linguistics and history
- developing practical solutions for information access to historical documents
- theoretical part (presentations)
- practical part (interdisciplinary group work)
- presentation of the group projects at the beginning of next term

# Course Objectives

## CoLi Students ...

- hands-on experience with text mining techniques
- awareness of challenges of the cultural heritage domain

## History Students ...

- awareness of the possibilities and limitations of computer assisted analysis of historical sources

## Both ...

- working in an interdisciplinary environment and communicating with non-experts

# When and Where

When: 16.-28.02.2009
9:30(?)-18:00

Where: Geb. C7.2, Konferenzraum 2.11

# Requirements and Credits: History Students

## Bachelor/Master History: Projektmodul

- presentation/Referat (20-30 minutes)
- participation in small group work
- longer term paper / report (umfangreicherer Abschlussbericht bzw. umfangreichere Seminararbeit)

## Magister History

- presentation/Referat (20-30 minutes)
- participation in small group work

## Bachelor and Diplom "Historisch orientierte Kulturwissenschaften"

- presentation/Referat (20-30 minutes)
- participation in small group work
- shorter term paper / report (kürzerer Abschlussbericht bzw. kürzere Seminararbeit)

# Requirements and Credits: Coli Students

## Bachelor and Master: Projektseminar

- 5 CPs (30+120 hours)
- presentation (20-30 minutes)
- participation in small group work (practical work / implementation)
- written report (one per group)

# Topics for Presentations: Coli

- Digitisation Issues
  - **Detection and correction of OCR and transcription errors**
  - **Dealing with non-standard orthography**
  - Detection of inclusions in other languages
- Information Extraction
  - Named Entities Background (NER for historical documents)
  - **Named Entity Disambiguation and Linking**
  - Information Extraction Background
- Semantic Web
  - **Semantic Web Background**
  - **Inferring Meta-Data**
  - Ontologies
- Multi-Modal Data
  - Speech
  - Images and video