

Vorbesprechung: Text Mining for Historical Documents / Historische Quellen mit (computer-)linguistischen Methoden analysieren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

Why is it relevant?

- museums and archives possess 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

| Flock B 47. | |
|---------------------------------------|---|
| 1031. <i>Mesophorus phainopeplus</i> | ✓ |
| 1032. <i>Accipiter cooperii</i> | ✓ |
| 1033. <i>Otidites macrorhynchos</i> | ✓ |
| 1034. <i>Argentarius</i> | ✓ |
| 1035. <i>Scolopoides longirostris</i> | ✓ |
| 1036. <i>Motacilla alba</i> | ✓ |
| Flock B 46. | |
| 1037. <i>Accipiter cooperii</i> | ✓ |
| 1038. <i>Accipiter cooperii</i> | ✓ |
| 1039. <i>Chondestes grammacus</i> | ✓ |
| 1040. <i>Lathrotriccus euleri</i> | ✓ |
| 1041. <i>Cypselurus balasiensis</i> | ✓ |
| 1042. <i>Colaptes auratus</i> | ✓ |
| 1043. <i>Colaptes auratus</i> | ✓ |
| 1044. <i>Colaptes auratus</i> | ✓ |
| 1045. <i>Colaptes auratus</i> | ✓ |
| 1046. <i>Colaptes auratus</i> | ✓ |
| 1047. <i>Colaptes auratus</i> | ✓ |
| 1048. <i>Colaptes auratus</i> | ✓ |
| 1049. <i>Colaptes auratus</i> | ✓ |
| 1050. <i>Colaptes auratus</i> | ✓ |
| 1051. <i>Colaptes auratus</i> | ✓ |
| 1052. <i>Colaptes auratus</i> | ✓ |
| 1053. <i>Colaptes auratus</i> | ✓ |
| 1054. <i>Colaptes auratus</i> | ✓ |
| 1055. <i>Colaptes auratus</i> | ✓ |
| 1056. <i>Colaptes auratus</i> | ✓ |
| 1057. <i>Colaptes auratus</i> | ✓ |
| 1058. <i>Colaptes auratus</i> | ✓ |
| 1059. <i>Colaptes auratus</i> | ✓ |
| 1060. <i>Colaptes auratus</i> | ✓ |
| 1061. <i>Colaptes auratus</i> | ✓ |
| 1062. <i>Colaptes auratus</i> | ✓ |
| 1063. <i>Colaptes auratus</i> | ✓ |
| 1064. <i>Colaptes auratus</i> | ✓ |
| 1065. <i>Colaptes auratus</i> | ✓ |
| 1066. <i>Colaptes auratus</i> | ✓ |
| 1067. <i>Colaptes auratus</i> | ✓ |
| 1068. <i>Colaptes auratus</i> | ✓ |
| 1069. <i>Colaptes auratus</i> | ✓ |
| 1070. <i>Pellorneum capistratum</i> | ✓ |
| 1071. <i>Pellorneum capistratum</i> | ✓ |
| 1072. <i>Pellorneum capistratum</i> | ✓ |
| 1073. <i>Pellorneum capistratum</i> | ✓ |
| 1074. <i>Pellorneum capistratum</i> | ✓ |
| 1075. <i>Pellorneum capistratum</i> | ✓ |
| 1076. <i>Pellorneum capistratum</i> | ✓ |
| 1077. <i>Pellorneum capistratum</i> | ✓ |
| 1078. <i>Pellorneum capistratum</i> | ✓ |
| 1079. <i>Pellorneum capistratum</i> | ✓ |
| 1080. <i>Pellorneum capistratum</i> | ✓ |
| 1081. <i>Pellorneum capistratum</i> | ✓ |
| 1082. <i>Pellorneum capistratum</i> | ✓ |
| 1083. <i>Pellorneum capistratum</i> | ✓ |
| 1084. <i>Pellorneum capistratum</i> | ✓ |
| 1085. <i>Pellorneum capistratum</i> | ✓ |
| 1086. <i>Pellorneum capistratum</i> | ✓ |
| 1087. <i>Pellorneum capistratum</i> | ✓ |
| 1088. <i>Pellorneum capistratum</i> | ✓ |
| 1089. <i>Pellorneum capistratum</i> | ✓ |
| 1090. <i>Pellorneum capistratum</i> | ✓ |
| 1091. <i>Pellorneum capistratum</i> | ✓ |
| 1092. <i>Pellorneum capistratum</i> | ✓ |
| 1093. <i>Pellorneum capistratum</i> | ✓ |
| 1094. <i>Pellorneum capistratum</i> | ✓ |
| 1095. <i>Pellorneum capistratum</i> | ✓ |
| 1096. <i>Pellorneum capistratum</i> | ✓ |
| 1097. <i>Pellorneum capistratum</i> | ✓ |
| 1098. <i>Pellorneum capistratum</i> | ✓ |
| 1099. <i>Pellorneum capistratum</i> | ✓ |
| 1100. <i>Pellorneum capistratum</i> | ✓ |
| 1101. <i>Pellorneum capistratum</i> | ✓ |
| 1102. <i>Pellorneum capistratum</i> | ✓ |
| 1103. <i>Pellorneum capistratum</i> | ✓ |
| 1104. <i>Pellorneum capistratum</i> | ✓ |
| 1105. <i>Pellorneum capistratum</i> | ✓ |
| 1106. <i>Pellorneum capistratum</i> | ✓ |
| 1107. <i>Pellorneum capistratum</i> | ✓ |
| 1108. <i>Pellorneum capistratum</i> | ✓ |
| 1109. <i>Pellorneum capistratum</i> | ✓ |
| 1110. <i>Pellorneum capistratum</i> | ✓ |
| 1111. <i>Pellorneum capistratum</i> | ✓ |
| 1112. <i>Pellorneum capistratum</i> | ✓ |
| 1113. <i>Pellorneum capistratum</i> | ✓ |
| 1114. <i>Pellorneum capistratum</i> | ✓ |
| 1115. <i>Pellorneum capistratum</i> | ✓ |
| 1116. <i>Pellorneum capistratum</i> | ✓ |
| 1117. <i>Pellorneum capistratum</i> | ✓ |
| 1118. <i>Pellorneum capistratum</i> | ✓ |
| 1119. <i>Pellorneum capistratum</i> | ✓ |
| 1120. <i>Pellorneum capistratum</i> | ✓ |
| 1121. <i>Pellorneum capistratum</i> | ✓ |
| 1122. <i>Pellorneum capistratum</i> | ✓ |
| 1123. <i>Pellorneum capistratum</i> | ✓ |
| 1124. <i>Pellorneum capistratum</i> | ✓ |
| 1125. <i>Pellorneum capistratum</i> | ✓ |
| 1126. <i>Pellorneum capistratum</i> | ✓ |
| 1127. <i>Pellorneum capistratum</i> | ✓ |
| 1128. <i>Pellorneum capistratum</i> | ✓ |
| 1129. <i>Pellorneum capistratum</i> | ✓ |
| 1130. <i>Pellorneum capistratum</i> | ✓ |
| 1131. <i>Pellorneum capistratum</i> | ✓ |
| 1132. <i>Pellorneum capistratum</i> | ✓ |
| 1133. <i>Pellorneum capistratum</i> | ✓ |
| 1134. <i>Pellorneum capistratum</i> | ✓ |
| 1135. <i>Pellorneum capistratum</i> | ✓ |
| 1136. <i>Pellorneum capistratum</i> | ✓ |
| 1137. <i>Pellorneum capistratum</i> | ✓ |
| 1138. <i>Pellorneum capistratum</i> | ✓ |
| 1139. <i>Pellorneum capistratum</i> | ✓ |
| 1140. <i>Pellorneum capistratum</i> | ✓ |
| 1141. <i>Pellorneum capistratum</i> | ✓ |
| 1142. <i>Pellorneum capistratum</i> | ✓ |
| 1143. <i>Pellorneum capistratum</i> | ✓ |
| 1144. <i>Pellorneum capistratum</i> | ✓ |
| 1145. <i>Pellorneum capistratum</i> | ✓ |
| 1146. <i>Pellorneum capistratum</i> | ✓ |
| 1147. <i>Pellorneum capistratum</i> | ✓ |
| 1148. <i>Pellorneum capistratum</i> | ✓ |
| 1149. <i>Pellorneum capistratum</i> | ✓ |
| 1150. <i>Pellorneum capistratum</i> | ✓ |
| 1151. <i>Pellorneum capistratum</i> | ✓ |
| 1152. <i>Pellorneum capistratum</i> | ✓ |
| 1153. <i>Pellorneum capistratum</i> | ✓ |
| 1154. <i>Pellorneum capistratum</i> | ✓ |
| 1155. <i>Pellorneum capistratum</i> | ✓ |
| 1156. <i>Pellorneum capistratum</i> | ✓ |
| 1157. <i>Pellorneum capistratum</i> | ✓ |
| 1158. <i>Pellorneum capistratum</i> | ✓ |
| 1159. <i>Pellorneum capistratum</i> | ✓ |
| 1160. <i>Pellorneum capistratum</i> | ✓ |
| 1161. <i>Pellorneum capistratum</i> | ✓ |
| 1162. <i>Pellorneum capistratum</i> | ✓ |
| 1163. <i>Pellorneum capistratum</i> | ✓ |
| 1164. <i>Pellorneum capistratum</i> | ✓ |
| 1165. <i>Pellorneum capistratum</i> | ✓ |
| 1166. <i>Pellorneum capistratum</i> | ✓ |
| 1167. <i>Pellorneum capistratum</i> | ✓ |
| 1168. <i>Pellorneum capistratum</i> | ✓ |
| 1169. <i>Pellorneum capistratum</i> | ✓ |
| 1170. <i>Pellorneum capistratum</i> | ✓ |
| 1171. <i>Pellorneum capistratum</i> | ✓ |
| 1172. <i>Pellorneum capistratum</i> | ✓ |
| 1173. <i>Pellorneum capistratum</i> | ✓ |
| 1174. <i>Pellorneum capistratum</i> | ✓ |
| 1175. <i>Pellorneum capistratum</i> | ✓ |
| 1176. <i>Pellorneum capistratum</i> | ✓ |
| 1177. <i>Pellorneum capistratum</i> | ✓ |
| 1178. <i>Pellorneum capistratum</i> | ✓ |
| 1179. <i>Pellorneum capistratum</i> | ✓ |
| 1180. <i>Pellorneum capistratum</i> | ✓ |
| 1181. <i>Pellorneum capistratum</i> | ✓ |
| 1182. <i>Pellorneum capistratum</i> | ✓ |
| 1183. <i>Pellorneum capistratum</i> | ✓ |
| 1184. <i>Pellorneum capistratum</i> | ✓ |
| 1185. <i>Pellorneum capistratum</i> | ✓ |
| 1186. <i>Pellorneum capistratum</i> | ✓ |
| 1187. <i>Pellorneum capistratum</i> | ✓ |
| 1188. <i>Pellorneum capistratum</i> | ✓ |
| 1189. <i>Pellorneum capistratum</i> | ✓ |
| 1190. <i>Pellorneum capistratum</i> | ✓ |
| 1191. <i>Pellorneum capistratum</i> | ✓ |
| 1192. <i>Pellorneum capistratum</i> | ✓ |
| 1193. <i>Pellorneum capistratum</i> | ✓ |
| 1194. <i>Pellorneum capistratum</i> | ✓ |
| 1195. <i>Pellorneum capistratum</i> | ✓ |
| 1196. <i>Pellorneum capistratum</i> | ✓ |
| 1197. <i>Pellorneum capistratum</i> | ✓ |
| 1198. <i>Pellorneum capistratum</i> | ✓ |
| 1199. <i>Pellorneum capistratum</i> | ✓ |
| 1200. <i>Pellorneum capistratum</i> | ✓ |

The Digitisation Process: Digital Copy

5

Film II, orenstaan 18 klm 37 van de no's 15 t/m 22
RMNH 15039

RMNH 15480

Reg.nr. 15739

RMNH 15690

Uitgenomen no's: gezield med Paris
Voorzien van A.B.L.F. en f.

RMNH 15245

Reg.nr. 15230

top lang blauw, rest rose

Reg.nr. 15700

Reg.nr. 16738

RMNH 16207

RMNH 16293

289, onderkant stam + oranje

naam erbij: Paoukaimama.

15. Gonatodes humeralis, 3-V-1968, Cultuurtuin, Paramaribo, op orenstaan van bomen, juist boven grond o.a. kantoor, palenboom, boom met paarse vruchten aan stam. ± 9.30 - 16.30. Blote aan basis stam, ^{o.a. bladeren}.
16. Ameiva ameiva, Cultuurtuin, Paramaribo, 3-V-1968, wegberm, tussen gras, onder steen. ± 10 u.
17. Cnemidophorus l. lemniscatus, Cultuurtuin, Paramaribo, 3-V-1968, wegberm tussen gras, onder steen. ± 10 u.
18. Anolis auratus, Cultuurtuin, Paramaribo, 3-V-1968, wegberm, lage kruiden vegetatie, tot 50 cm hoog. ± 10 u.
19. Gymnophthalmus speciosus, ^(Cocaoboom) Cultuurtuin, Paramaribo, 3-V-1968, onder afgevallen blad van boom met paarse vruchten aan stam, onder en vermoedelijk ook in vermolde boomstammen. 9.30 u - 12.30 u zeer heldrig, 1500 - 1700 u minder heldrig, zonnebad op o.a. plantwortels kantoor.
20. Cercosaura occellata, Cultuurtuin, Paramaribo, 3-V-1968, onderafgevallen blad en gemaaid Tradescantia, licht bos, boom.
21. Hyla rubra, Cultuurtuin, Paramaribo, 3-V-1968, op stam van boom met paarse vruchten aan stam (= COCAO)
22. Leptodactylus hyloscirtus (det. Dr. B.R. HAYES, 1972) Cultuurtuin, Paramaribo, 3-V-1968, tussen afgevallen blad van boom met paarse vruchten op stam 11.30 u. Naar grote (± 2 m) sprongen.
23. Hemidactylus mabouia, Paramaribo, op buitenmuur huis o.a. Waterkant, 3-V-1968, l.d. F. Westerling.
24. Gonatodes humeralis, Paramaribo, Cultuurtuin, 4-V-1968, op onderkant 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

Gonocephalus 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*.

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: 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)

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