»Epigrāphia Carnāṭica« digital

DHd-AG *OCR*

Halle 08. 11. 2019

Dr. Michael Dahnke

Fahrplan

I) Projektvorstellung: *A.)*–*D.)*

II) Herausforderungen: *E.)*

III) Lösungsansätze: *F.)*

IV) Diskussion

A.) Rahmendaten

1. Projektname: *»Epigrāphia Carnāṭica« digital*

2. Kooperation des *Instituts für Digital Humanities*, des *Cologne Center for eHumanities* (*CCeH*), des *Data Center for the Humanities* (DCH) der Philosophischen Fakultät, alle drei Universität zu Köln, und des Lehrstuhls für moderne Indologie am *Institut für Indologie und Tibetologie* der Ludwig-Maximilians-Universität München

3. Antragsstellung: November 2019 DFG

4. Warum Vorstellung: Bislang primär mit Fachwissenschaftlern zu indologischen Themen gesprochen, aber kaum mit OCR-Fachleuten. Ergo: I n h a l t l i c h e u n d t e c h n i s c h e R ü c k f r a g e n a l l e r A r t e r b e t e n .

B.) Zwei Projektziele

1. Erstellung von OCR-Modellen für OCR der dravidischen Sprache Kannaḍa mit Inschriften des zweiten Bandes der *Epigrāphia Carnāṭica*: Śravaṇabeḷagoḷa; Kannaḍa eine der vier dravidischen Hauptsprachen Südindiens. Anstoß für digitale Erschließung des gedruckten kulturellen Erbes des alten Königreiches Mysuru.

2. Anschließende Prüfung der Modelle auf Tauglichkeit mit Textdigitalisierung von Inschriften des zweiten Bandes der *Epigrāphia Carnāṭica*: Śravaṇabeḷagoḷa, und Weiterverarbeitung der digitalisierten Texte zur digitalen Volltextedition dieser Inschriften: *»Epigrāphia Carnāṭica« digital*.

C.) Geplante Ergebnisse

1. Mindestens drei ›gemischte‹ *Calamari*-Modelle für *i)* Kannaḍa, *ii)* die Transliteration in lateinischen Buchstaben, und *iii)* Rice' Übersetzungen ebenfalls in Antiqua;

2. Text-digitalisierte Sammlung von Inschriften des zweiten Bandes der *Epigrāphia Carnāṭica: Śravaṇabeḷagoḷa* in drei digitalen Textformen: *i)* glyphengetreue Textdigitalisierung der in *Epigrāphia Carnāṭica* gedruckten Transliterationen in Kannaḍa, *ii)* lateinische Transliterationen mit entsprechenden Diakritika und *iii)* englische Übersetzungen in Antiqua.

3. Neben Modellen und ground truth werden auch alle anderen Ergebnisse des Projektes spätestens am Ende dessen Laufzeit unter CC BY-NC-SA 4.0 Lizenz zur Verfügung gestellt.

Mit kostenfreier Verfügbarkeit der ground truth wird zudem Tatsache mindestens drei allein im deutschsprachigen Raum gegenwärtig verwendeter OCR-Programme Rechnung getragen: *Tesseract 4*, *OCRopus* und *Calamari*.

D.) Arbeitsplan

1. Aktuell Transkription von Inschriften mit Indologin und Modelltraining mit *OCR4all*, um PAGEXML anschließend für Probetraining mit *nasḫī* zu verwenden.

2. Nach Abgabe des Antrags weiteres Trainning mit OCR-Software, die auf *Calamari* aufsetzt, um

   ಅ) bei positivem Förderungsbescheid bereits entscheidenden Schritt vorangekommen zu sein oder
   ಆ) bei Ablehnung wenigstens bis dahin erzielte Ergebnisse als Artikel **Wo?** zu veröffentlichen und damit erneute Antragsstellung zu beflügeln.

3. Bei Modellentwicklung soll mindestens ausprobiert werden: *i)* Training mit mehreren Modellen, *ii)* Voting und *iii)* Erstellung eines oder mehrere Modelle für Transkription der ಕನ್ನಡ-Schrift in lateinische Buchstaben mit Diakritika (ISO 15919/IAST).

   Wenn ಅ):

   (a) Modellentwicklung Indologin/Indologe und M. D., bis Erkennungsrate $99+\%$ erreicht. Das erscheint angesichts Qualität der Digitalisate realistisch.
   (b) Mit fertigen Modellen Erkennung einer Sammlung weiterer, inhaltlich zusammenhängender Inschriften; Auswahl Prof. Zydenbos, stellvertretender Ordinarius für moderne Indologie, *Institut für Indologie und Tibetologie* LMU; Erkennung Indologin/Indologe und M. D.
   (c) Entwicklung eines Datenmodells auf Basis *EpiDoc* zur Auszeichnung der Texte: Indologin/Indologe und M. D.
   (d) Auszeichnung der glyphengetreu digital transliterierten Texte: Indologin/Indologe und M. D.

(e) Präsentation auf Webportal *»Epigrāphia Carnāṭica« digital* im Rahmen von *C-SALT | Cologne South Asian Languages and Texts*.

Wenn ೨):

(a) Modellentwicklung bis Erkennungsrate $99+\%$ erreicht.
(b) Damit Erkennung weiterer, inhaltlich zusammenhängender Inschriften; Auswahl Prof. Zydenbos.
(c) Präsentation der Ergebnisse in Fachzeitschrift/online **Wo?**.

E.) Herausforderungen

E.1) Abugida

1. Prinzip Abugida: Abugida ist Konsonantenschrift: Wörter bestehen in ಕನ್ನಡ (Kannaḍa) aus einem oder mehreren Konsonanten; von denen jeder wiederum mit mindestens einem Konsonant und maximal einem Vokal kombinierbar ist:

   ಕ ನ್ನ ಡ

   Ka nna ḍa

   Wenn Buchstaben wie in Beispiel in Grundform erscheinen, wird immer Vokal a mitgesprochen. Um aus ka (ಕ) ein ku (ಕು) zu machen, wird Diakritikon angefügt.

   Dasselbe gilt, wenn zwei Konsonanten direkt hintereinander: nna wie in Kannaḍa wird als ನ್ನ dargestellt. Zweiter, identischer Konsonant wird als Diakritikon dargestellt.

2. Nachfolgend wird von 37 verschiedenen Konsonanten als in den Inschriften der *Epigrāphia Carnāṭica* möglich ausgegangen:

   Zeichensatz der Kannaḍa-Schrift zur Darstellung der Kannaḍa-Sprache umfaßt laut Zydenbos 34 Konsonanten + »anusvāra (written ṃ in transliteration)« und »visarga (written ḥ in transliteration). It occurs exclusively in a few rare Sanskrit loan words.« [Zydenbos, Robert J.: *General information about the Kannada language*. http://lmu.zydenbos.net/Kannada-I/general-intro-131008.pdf. p. 7].

   Ebenfalls in Inschriften auftauchendes f (ಋ) in Alt(?) -Kannaḍa.

3. 16 Vokale [Zydenbos, Robert J.: *General information about the Kannada language*. http://lmu.zydenbos.net/Kannada-I/general-intro-131008.pdf. p. 7.

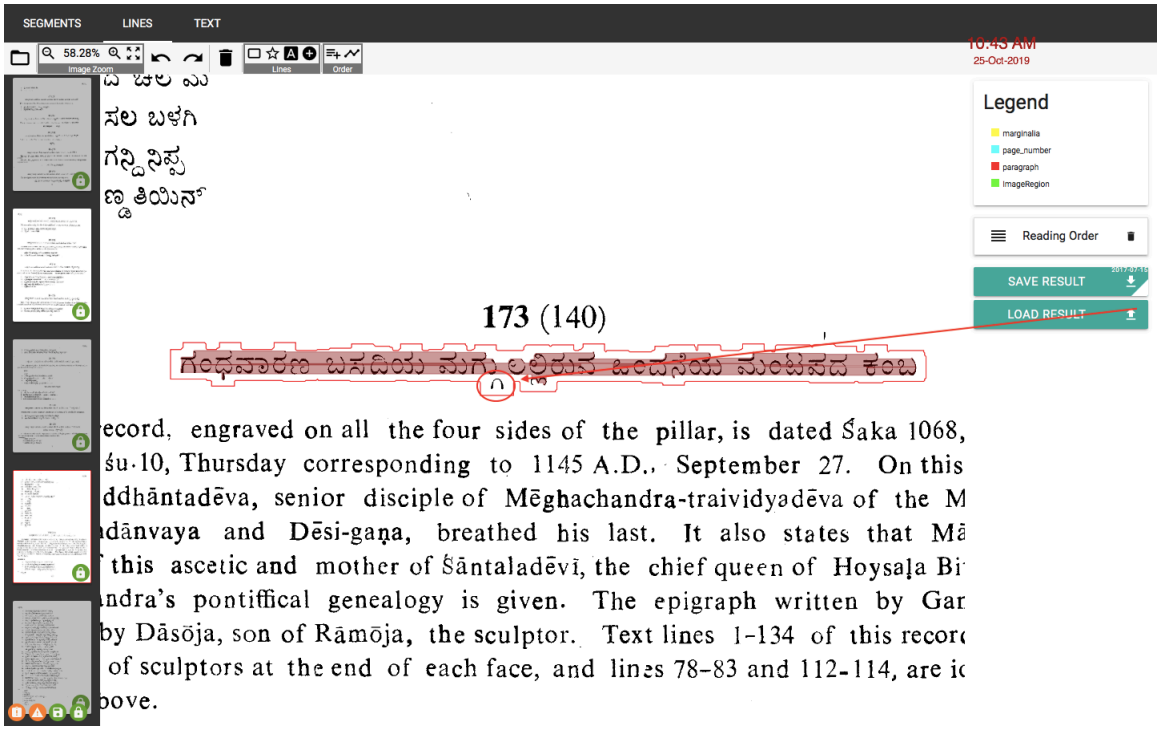
4. Anzahl der Kombinationen der Konsonanten theoretisch unbegrenzt und Codec der Abugida Kannaḍa damit theoretisch unendlich: Faktisch bislang folgendes Beispiel nachweisbar: ಸ್ತ್ರಾ, also »strā«, sprich ein Konsonant mit zwei Konsonanten als Diacriticon und einem Vokal.

5. Ausgehend von Kombination ನ್ಱ್, also e i n Konsonant mit z w e i Konsonanten als Diacritica und e i n e m Vokal: 37 x 37 x 37 x 16 = 810.448 mögliche Kombinationen.

6. Faktisch soviel wahrscheinlich nicht.

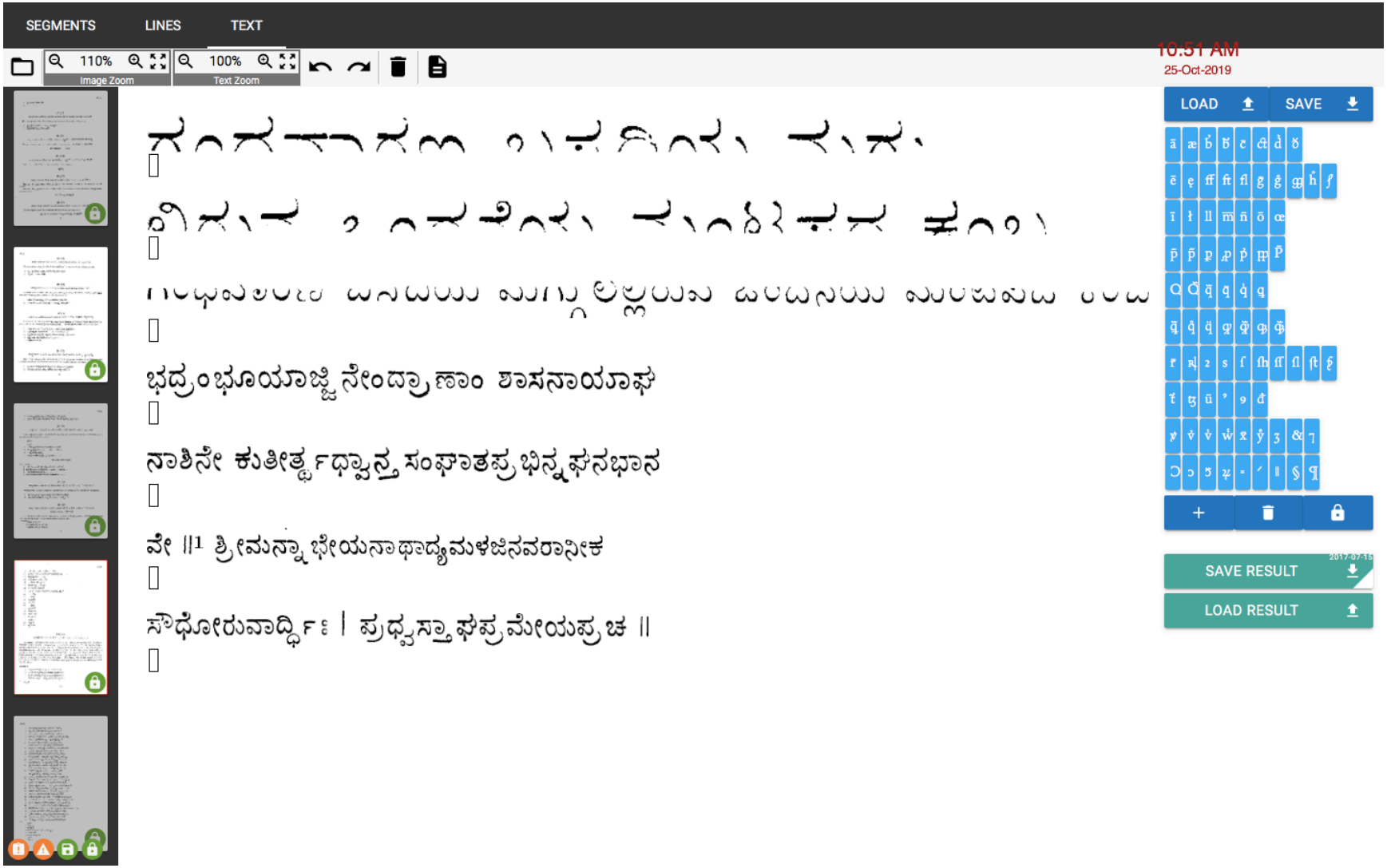
7. Aber selbst wenn nur von Kombination eines Konsonanten mit einem Vokal ausgegangen wird: 37 x 16 = 592 mögliche Kombinationen. Großer Codec, der jedes Glyph für Trainingszwecke mindestens einmal enthalten sollte.

E.2) Unsauber ausgeschnittene Diacritica



Benutzte Software: *OCR4all*, ›darunter‹ für Zeilensegmentierung *OCRopus*

Zudem ärgerliche Petitesse vieler tiefliegender und darob nicht ausgeschnittener Diacritica. Hat jemand ähnliches Problem?


SEGMENTS
LINES
TEXT
110%
Image Zoom
100%
Text Zoom
10:51 AM
25-Oct-2019
LOAD
SAVE
SAVE RESULT
LOAD RESULT

E.3) Publikationsmöglichkeiten

Hat jemand Vorschläge, wo man zu diesem Projekt fachspezifisch publizieren kann?

E.4) Trainingsmöglichkeiten

Hat jemand Möglichkeit, Training anzustoßen und was braucht er/sie dazu?

F.) Lösungsansätze

F.1) Abugida

Statt Konsonanten Silben trainieren? Also ಸ್ತ್ರಾ (»strā«) als Silbe interpretieren?

1. Gegenbeispiel:

   ಶ್ರೀ ಉದ್ಯಾನೈರ್ಜ್ಜಿ ತನನ್ದನಂ ಧ್ವನದಳಿವ್ಯಾಸಕ್ತರಕ್ತೋತ್ಪಲ[1]

   śri udyānair jjita-Nandanaṃ dhvanad-aḷi-vyāsakta-raktōtpala-[2]

   Im ersten markierten Fall ist a der zweiten Silbe »aḷi« inhärentes a des davorstehenden ದ: da-ḷi

[1] EC II.2 (1973; p5). http://idb.ub.uni-tuebingen.de/opendigi/EC_02_1973#p=113. Letzter Zugriff 08. 11. 2019.
[2] Ebd.

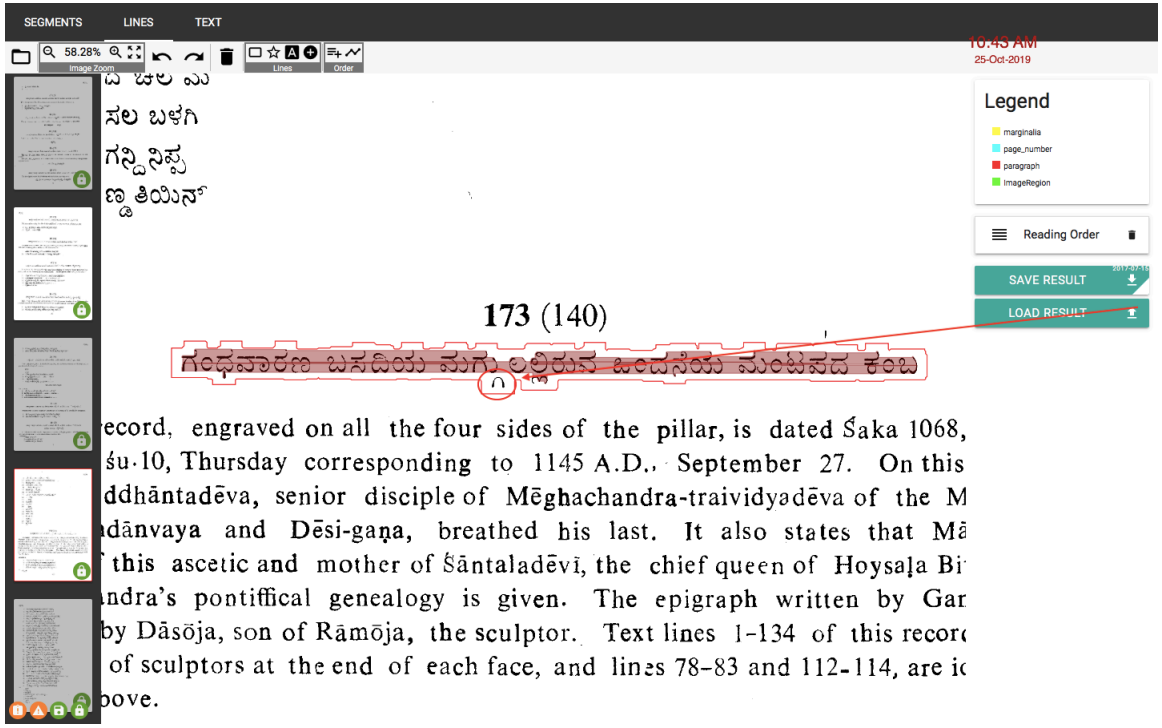
2. OCR-Algorithmus schneidet immer Scheiben (48 px hoch und 1 px breit) auf einem Bildzeilendigitalisat – das einer Textzeile Höhe und Länge entspricht – aus und berechnet Schnitte als Teile von Buchstaben. Kann man ob dessen Programm auch für Silben trainieren?

Wer hat Erfahrungen mit OCR-Training auf einer Konsonantenschrift?

Arabisch? Arabisch? Arabisch?

F.2) Unsauber ausgeschnittene Diacritica

Manuell nachsegmentieren. Unbefriedigend, zumindestens solange kein Transkriptions-ವಲ (non-Indian SHK/WHK etc.) zur Verfügung steht. Zeilen jetzt erst einmal herausgelöscht; bei größerem Corpus unbefriedigend, weil natürlich mindestens alle Textzeilen von automatischer Zeilenerkennung erfaßt werden sollen.



F.3) Publikationsmöglichkeiten

http://www.zfdg.de/

https://currentepigraphy.org/