Digital Corpus of Sanskrit

- Digital corpus of Sanskrit (Since 1999) has more than 2,500,000 lexical items from about 200 Sanskrit texts.
- These Sanskrit texts are from five time slots, divided between the period from 500 BCE to 1900 CE.
- Texts included such as, *Upnishads, Mahabharata, Kamasutra, Arthashasthra, etc.*
- It's not a parallel corpora. The texts are available in Devanagri.
- Corpus is POS tagged.
- Created and maintained by Oliver Hellwig from the department of Classical Indology, South Asia Institute, University of Heidelberg, Germany.
- Licensed under, ' Creative Commons Attribution 3.0', imported license.
- http://kjc-fs-cluster.kjc.uni-heidelberg.de/dcs/index.php

META-DATA/ANNOTATIONS

- There are no recordings, only texts are used in this Corpus.

- Uses Frame net, a lexico-semantic resource to annotate texts with discourse semantic action scenario, so each frame has –

  target word > donor > recipient > theme

  (frame) (role) (role) (role)

- The whole corpus is wrapped up in XML.

- Query domain gives user three types of search options

  I. Basic search (offers word meaning in English and German)

  II. Searching in Scanned Sanskrit books (results always in Devanagri)

  III. Word clouds (generate frequency of searched word in any particular text)

Queries are only accepted in transliteration and not in Devanagri.

USAGE

- Created primarily to investigate the influence of time on the vocabulary of texts. To study development of Sanskrit vocabulary from a bird's eye perspective.
- For example, texts about Indian flora occurred in late parts of Sanskrit literature, there is a possibility of a high rate of non-IA words (especially from Dravidian languages).
- Useful in examining etymological trends in Sanskrit.
- Sanskrit (OIA) > Middle IA > Apabhramsa > Modern IA
- Uses ANOVA statistical model to capture variance.

REFERENCES

- Hellwig, O. (2009). Etymological trends in the Sanskrit vocabulary. *Literary and linguistic computing*, fqp034.

  (corpus construction)

- Frank, A., Hellwig, O., & Reiter, N. (2012). Semantic Annotation for the Digital Humanities-Using Markov Logic for Annotation Consistency Control. *Linguistic Issues in Language Technology*, *7*(1).

  (Main analysis of written descriptions of Nepalese rituals, using Sanskrit manuals about Nepalese rituals – both contain almost same South Asian religious material; for word sense annotation UKB system is used; for WSD, the UKB is adapted to the ritual domain and senses are acquired that may be characteristic of ritual domain from DCS)

- Narayana, A., & Thrigulla, S. R. TRANSITION OF AYURVEDIC EDUCATION: HERMITAGE TO ELECTRONIC AGE.

  (new sources of presenting Sanskrit and Ayurvedic literature)

ISLRN

Title	DSC
Full Title	Digital Sanskrit Corpus
Resource Type	Texts
Source/URL	http://kjc-fs-cluster.kjc.uni-heidelberg.de/dcs/index.php
Format/MIME type	text/xml
Size	> 2,500,000 lexical items
Access medium	online
Description	This corpus was designed as a general-purpose philological resource that covers Sanskrit texts from 500 BCE until 1900 CE. It is a searchable collection of lemmatized Sanskrit texts with more than 2,500,00 lexical items. They have been manually annotated with word senses from WordNet 2.0. It offers free internet access to a part of the database of the linguistic program Sanskrit Tagger, which has been under constant development since 1999.
Version	?
Media Type	Written Texts
Language	Sanskrit
Resourse Creator	Oliver Hellwig, University of Heidelberg
Distributor	Oliver Hellwig
Rights Holder	Oliver Hellwig