Electronic Concordancing for Study of Imagery in the Great Epics of India

15th World Sanskrit Conference
New Delhi, January 9, 2012

Les Morgan
les@growthhouse.org
mywhatever.com/sanskrit

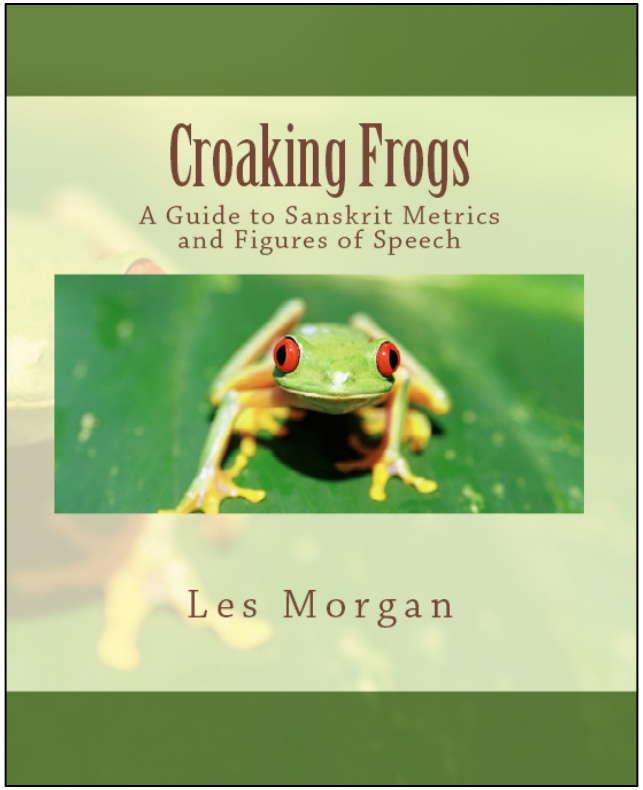
Ram Karan Sharma
ramkaransharma@yahoo.com

Presenters

- **Ram Karan Sharma**
  - Former President of the International Association of Sanskrit Studies
  - Author of *Elements of Poetry in the Mahābhārata*
- **Les Morgan**
  - Technologist with interest in multilingual applications
  - Designed bilingual Russian/English software for the International Space Station
  - Co-developer of the Vidyut keyboard for Devanāgarī on Windows computers
  - Author of *Croaking Frogs*, on Sanskrit metrics and figures of speech

Croaking Frogs

- Published January 2012
- Available at Amazon.com
- R. K. Sharma, Consulting Editor
- Traditional content on metrics and figures of speech presented in a modern format



Goals

- Create a complete enumeration of all objects of poetic images in the Indian Epics (*Mahābhārata* and *Rāmāyaṇa*)
- Make results easily available to others in electronic, reusable, form that can be used with other corpora
- Search tools can be used on any Unicode digital text, not just the Epics
- Project web site: mywhatever.com/sanskrit/epics

What is a “concordance”?

- A **concordance** brings together (“concords”) passages of a text that show the use of a word or concept
- Enables study of how a work uses language
- Shows how often a term is used
- Computer concordances let users interact directly with the texts they are studying
- We are making a concordance of poetic images

Our research methods

- Computer programs look for grammatical structures
- R. K. Sharma classifies results
- Disseminate findings using electronic publication methods that have the best potential for re-use of findings by other researchers
  - Work products will include XML files and other digital search aids

Challenges: Size of the Epics

- Immense size of the Epics defies analysis
- *Mahābhārata*
  - Longest epic poem in the world
  - Over 100,000 verses
  - 159,293 electronic edition lines
    - 8,659,001 characters (including spaces)
    - 1,062,237 strings (blank-delimited)
- *Rāmāyaṇa*
  - 24,000 verses (traditional count)
  - 38,083 electronic edition lines
    - 2,055,802 characters (including spaces)
    - 251,787 strings (blank-delimited)

Challenges: Technical

- Complexities of the Sanskrit language render some computer lexical tools useless
  - Word boundaries difficult to detect
- Multiple encoding methods for Devanāgarī (and its Romanization)
  - Not all encodings work on all software
  - CSX+ legacy encoding works on older software
  - Unicode is preferred for current use

Simile (*upamā*)

- Subject of comparison (*upameya*)
- Object of comparison (*upamāna*)
- Shared property, “Tertium comparationis” (*upamānadharma*)
- Linking word or morpheme (*aupamyavācaka*) E.g., *iva*, *yathā*, -*vat*, etc.
- Sometimes effect is implicit with no linking word or mention of the shared property: E.g., “one having lotus-petal-like-eyes” (*kamalapatrākṣaḥ*)

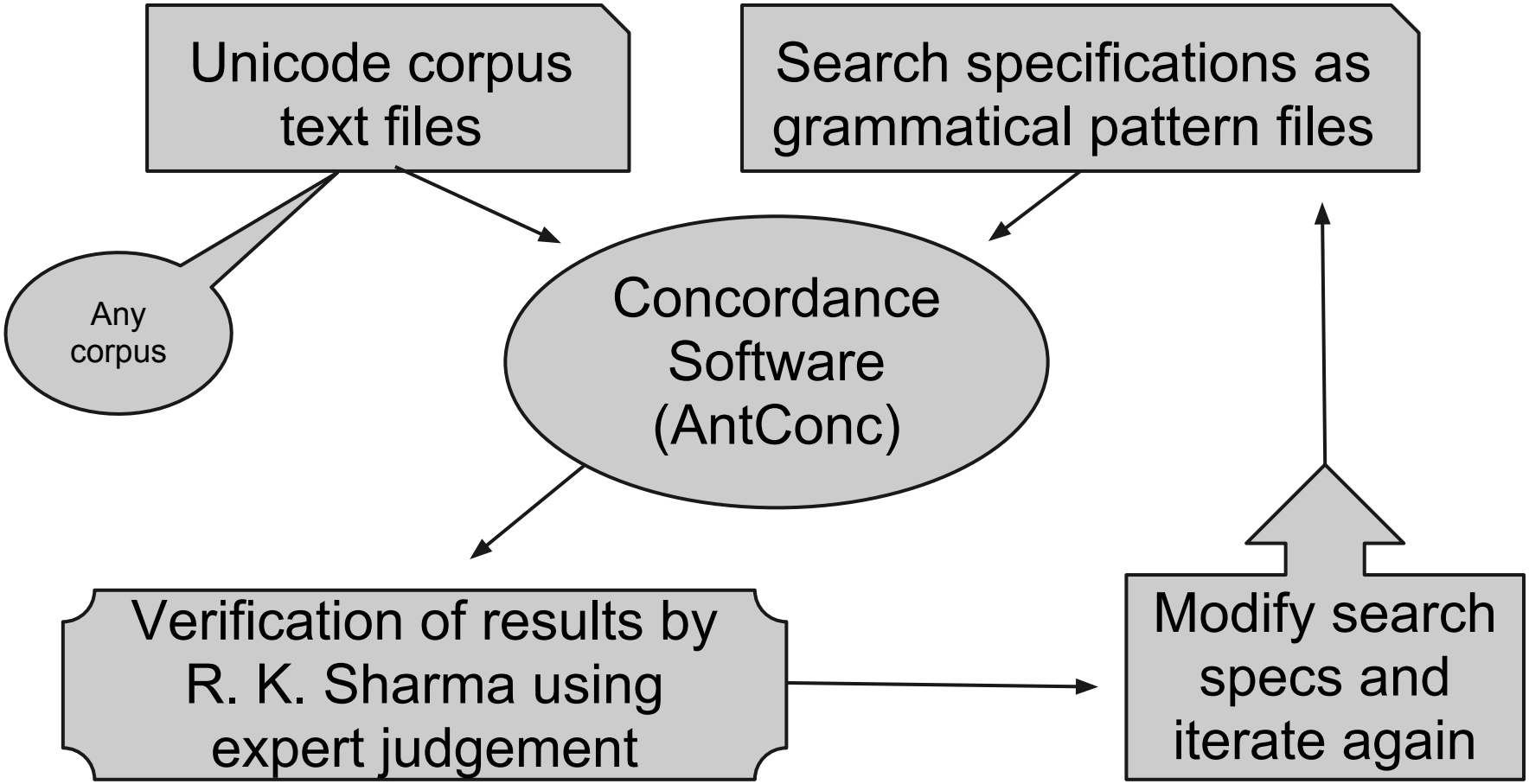
Poetic images

- Hanuman's **speed** is like that of the **mind**
  - Subject = Hanuman
  - Object = mind
  - Property = speed
- The warrior is as **strong** as an **elephant**
  - Subject = warrior
  - Object = elephant
  - Property = strength

Metaphor (*rūpaka*)

- Hard to detect automatically
- Identity is implicit, with no explicit linking word
- “Duryodhana is the great tree of furious temper…” (*duryodhano manyumayo mahādrumaḥ*)

Identification Process



Results to date

- Computer methods have found 15,099 lines containing general terms suggestive of poetic images
- This does not include more detailed searches for specific types of images

Mahābhārata	11,158
Rāmāyaṇa	3,941
Total lines	15,099

metaphor_keywords_v5.txt - WordPad

File Edit View Insert Format Help

ivācalam
upamā
upamān
upamāna
upamānaṁ
upamām
upamāsu
opama[\s]
opamā[\s]
opamaḥ
opamāḥ
opamān
opamau
opamai
opamam
opamām
opamadar
opame
opamayā
opamavarc
opamasattv
opamat

Create file with

169 search terms

suggestive of images

For Help, press F1 NUM

start Wi... Pa... GA... Mic... Nor... Se... GA... tao... Jas... Ant... me... ind... EN Norton 11:47 AM

AntConc 3.2.1w (Windows) 2007

File Global Settings Tool Preferences About

Corpus Files

MBh01.txt
MBh02.txt
MBh03.txt
MBh04.txt
MBh05.txt
MBh06.txt
MBh07.txt
MBh08.txt
MBh09.txt
MBh10.txt
MBh11.txt
MBh12.txt
MBh13.txt
MBh14.txt
MBh15.txt
MBh16.txt
MBh17.txt
MBh18.txt
ram_01_bala
ram_02_ayo
ram_03_ara
ram_04_kis
ram_05_sun
ram_06_yud
ram_07_utta

Concordance | Concordance Plot | File View | Clusters | Collocates | Word List | Keyword List

Hit	KWIC	File
1	001007c kālaḥ kamalapatrākṣa śaṁsaitat pṛcchato mama 01001008 sūta u	MBh01.txt
2	ñe mahābhāgāḥ sūryapāvakavarcasaḥ 01001013c kṛtābhiṣekāḥ śucayaḥ kṛta	MBh01.txt
3	aye 01001037a yathartāv ṛtuliṅgāni nānārūpāṇi paryaye 01001037c dṛśya	MBh01.txt
4	nte tāni tāny eva tathā bhāvā yugādi	
5	54c trīn agnīn iva kauravyāñ ja	
6	ryodhano manyumayo mahādrumaḥ; skand	
7	iṣṭhiro dharmamayo mahādrumaḥ; skand	
8	ṁ tad adbhutam ivābhavat 01001078a tatprītyā caiva sarveṣāṁ paurāṇāṁ	MBh01.txt
9	001083c āditya iva duṣprekṣyaḥ samareṣv api cābhavat 01001084a sa sar	MBh01.txt
10	001089a vimānapratimāṁ cāpi mayena sukṛtāṁ sabhām 01001089c pāṇḍavānā	MBh01.txt
11	īt praskandann iva saṁbhramāt 01001090c pratyakṣaṁ vāsudevasya bhīmer	MBh01.txt
12	iyam akṣatriyo yathā 01001100e gāndhārarājasahitaś chadmadyūtam amant	MBh01.txt
13	tatra yad yad yathā jñātaṁ mayā saṁjaya tac chṛṇu 01001101c śrutvā h	MBh01.txt
14	jayaṁ; śakrāt sākṣād divyam astraṁ yathāvat 01001110c adhīyānaṁ śaṁsi	MBh01.txt
15	āstravidusaḥ śakrapratimatejasaḥ 01001165a dharmeṇa pṛthivīṁ jitvā ya	MBh01.txt
16	devāhvayaḥ supratimaḥ supratīko bṛhadrathaḥ 01001175c mahotsāho vinī	MBh01.txt
17	7c pratibimbam ivādarśe paśyanty ātmany avasthitam 01001198a śraddad	MBh01.txt
18	201c navanītaṁ yathā dadh	dānā
19	padāṁ brāhmaṇo yathā 010	ṭhā
20	adām 01001202c yathaitān	3a
21	amāḥ 01002012c yathā deś	201

search terms

15,099 lines

Search Term ☑ Words ☐ Case ☑ Regex

devat.g Advanced

Start Stop Sort

Concordance Hits 15099

Search Window Size 50

Total No. 25

Kwic Sort

☑ Level 1 0 ☐ Level 2 0 ☐ Level 3 0

Save Window

Exit

Search accuracy goals

- Minimize false positives
  - Some lines are selected that should not be
  - We cannot claim that every line we find contains an image
- Minimize false negatives
  - Some lines are not selected that should be
  - We cannot claim that every image has been found

Examples of search strategies

- Look for any line containing a simile
- Look for any specific object
- Look for a specific image

How to find an elephant

- Primarily a figure of might and vitality
- Vocabulary: *gaja*, *vāraṇa*, *kuñjara*, *mātaṅga*, *nāga*, *hastin*, etc.
- Named types and individuals: *Airāvata*, *abhipadma*, etc.
- Stock images, e.g., “furious like an elephant in rut” (*prabhinna iva vāraṇaḥ*)

Photo credit: Magnus Franklin. A keeper adorns his elephant after a bath

elephant.txt - WordPad

File Edit View Insert Format Help

gajān
gajānā
gajānāṁ
gajānām
gajaiḥ
gajaiś
gajakaṅkaṭa
gaje
gajebhy
gajebhyaś
gajena
gajendr
gajendrāḥ
gajendrān
gajo
gajaskandham
giric
giricaro
giricāriṇaḥ
hastin
hastinaḥ
hastino
hastī
hastīva
kuñjara
kuñjaraiḥ
kuñjarau
kuñjaro
kuñjarasye
madotkaṭa
madotkaṭaḥ
madotkaṭau
madotkaṭaṁ
madotkaṭam
madotkaṭaiḥ
mahāgaj
mahāgajam
mahāgajān
mahāhast
mātaṅga
mattavāraṇa
mattakuñjar
matto
mattam
nāgaḥ
nāgaś
nāgo
nāgo[\s]
mahānāg
prabhinna iva vārana

Define elephant terms

For Help, press F1 NUM

start Window... Pamela ... 2 Win... 2 Inte... Master ... AntCon... Microso... Jasc Pai... elephan... EN Norton 2:50 PM

AntConc 3.2.0w (Windows) 2006

File Global Settings Tool Preferences About

elephant terms

Corpus Files

MBh01.txt
MBh02.txt
MBh03.txt
MBh04.txt
MBh05.txt
MBh06.txt
MBh07.txt
MBh08.txt
MBh09.txt
MBh10.txt
MBh11.txt
MBh12.txt
MBh13.txt
MBh14.txt
MBh15.txt
MBh16.txt
MBh17.txt
MBh18.txt
ram_01_ba
ram_02_ay
ram_03_ar
ram_04_ki
ram_05_su
ram_06_yu
ram_07_ut

Concordance | Concordance Plot | File View | Clusters | Collocates | Word List | Keyword List

Hit	KWIC	File
1	ṁ rathāśvanaradantināṁ 01002014c yathāvac caiva no brūhi s	MBh01.txt
2	15a eko ratho gajaś caiko narāḥ pañca padātayaḥ 01002015c	MBh01.txt
3	tiḥ 01002020c gajānāṁ tu parīmāṇam etad evātra nirdiśet 01	MBh01.txt
4	ā rathāśvanaradantināṁ 01002149c nagarād dhāstinapurād bal	MBh01.txt
5	i 01003063c nānāgoṣṭhā vihitā ekadohanāḥ, tāv aśvinau duha	MBh01.txt
6	01003119C apramatto netum arhasīti 01003120A sa evam uktas	MBh01.txt
7	01003139a ya airāvatarājānaḥ sarpāḥ samitiśobhanāḥ 010031	MBh01.txt
8	apṛṣṭhe rejur airāvatodbhavāḥ 01003141a bahūni nāgavartmān	MBh01.txt
9	enāyāṁ cartum airāvataṁ vinā 01003142a śatāny aśītir aṣṭau	MBh01.txt
10	1003143c aham airāvatajyeṣṭhabhrātṛbhyo 'karavaṁ namaḥ 010	MBh01.txt
11	atā dṛṣṭaḥ sa airāvato nāgarājaḥ 01003174B yaś cainam adhi	MBh01.txt
12	devaṁ daityā nāgottamās tathā 01016029c cirārabdham idaṁ	MBh01.txt
13	asū 01025018c gajakacchapatāṁ prāptāv arthārthaṁ mūḍhaceta	MBh01.txt
14	mān samupaiti mahāgajaḥ 01025021a tasya bṛṁhitaśabdena kūr	MBh01.txt
15	aḥ pataty eṣa gajo jalam 01025022c dantahastāgralāṅgūlapād	MBh01.txt
16	rito yojanāni gajas tad dviguṇāyataḥ 01025024c kūrmas triy	MBh01.txt
17	5026c nakhena gajam ekena kūrmam ekena cākṣipat 01025027a	MBh01.txt
18	tvaṁ khādemau gajakacchapau 01025033a tato drumaṁ patagasa	MBh01.txt
19	deśān bahūn sagajakacchapaḥ 01026004c dayārthaṁ vālakhilyā	MBh01.txt
20	rkṣyaḥ saśākhāgajakacchapaḥ 01026019a na tāṁ vadhraḥ pariṇ	MBh01.txt
21	ḍas tāv ubhau gajakacchapau 01026027a tataḥ parvatakūṭāgrā	MBh01.txt

Search Term ☑ Words ☐ Case ☑ Regex

Advanced

Start Stop Sort

Concordance Hits 2073

Search Window Size 50

Total No. 25

Kwic Sort

☑ Level 1 0 ☐ Level 2 0 ☐ Level 3 0

Files Processed

Reset

Save Window

Exit

2,073 elephant lines

start Wind... Pamel... 2 W... 2 In... Mast... WinT... Norto... AntC... Micro... Jasc ... EN Norton 10:20 AM

AntConc 3.2.0w (Windows) 2006

File Global Settings Tool Preferences About

Corpus Files

MBh01.txt
MBh02.txt
MBh03.txt
MBh04.txt
MBh05.txt
MBh06.txt
MBh07.txt
MBh08.txt
MBh09.txt
MBh10.txt
MBh11.txt
MBh12.txt
MBh13.txt
MBh14.txt
MBh15.txt
MBh16.txt
MBh17.txt
MBh18.txt
ram_01_ba
ram_02_ay
ram_03_ar
ram_04_ki
ram_05_su
ram_06_yu
ram_07_ut

Concordance | Concordance Plot | File View | Clusters | Collocates | Word List | Keyword List

Hit	KWIC	File
1	garāḍ iva saṁkruddhaḥ prabhinna iva kuñjaraḥ 03146016a dra	MBh03.txt
2	kesarīva yathotsiktaḥ prabhinna iva vāraṇaḥ 03157029c vyap	MBh03.txt
3	anīke 05023022c nāgaḥ prabhinna iva naḍvalāsu; caṅkramyate	MBh05.txt
4	arśayiṣyati 05050033a prabhinna iva mātaṅgaḥ prabhañjan pu	MBh05.txt
5	hedyaḥ sarvaśastrāṇāṁ prabhinna iva vāraṇaḥ 05149024c jajñ	MBh05.txt
6	viddhaḥ sravan raktaṁ prabhinna iva kuñjaraḥ 06088004c dad	MBh06.txt
7	o droṇaḥ satyasaṁdhaḥ prabhinna iva kuñjaraḥ 07020040c abh	MBh07.txt
8	āsanas tu saṁkruddhaḥ prabhinna iva kuñjaraḥ 07038028c ayo	MBh07.txt
9	rad arjunaḥ 07068052c prabhinna iva mātaṅgo mṛdnan naḍavan	MBh07.txt
10	udbhinnarudhiro rājan prabhinna iva kuñjaraḥ 09056060a tat	MBh09.txt

seach for a specific image

10 lines found

Search Term ☑ Words ☐ Case ☑ Regex

prabhinna iva

Advanced

Start Stop Sort

Concordance Hits 10

Search Window Size 50

Total No. 25

Kwic Sort

☑ Level 1 0 ☐ Level 2 0 ☐ Level 3 0

Save Window

Exit

Files Processed

Reset

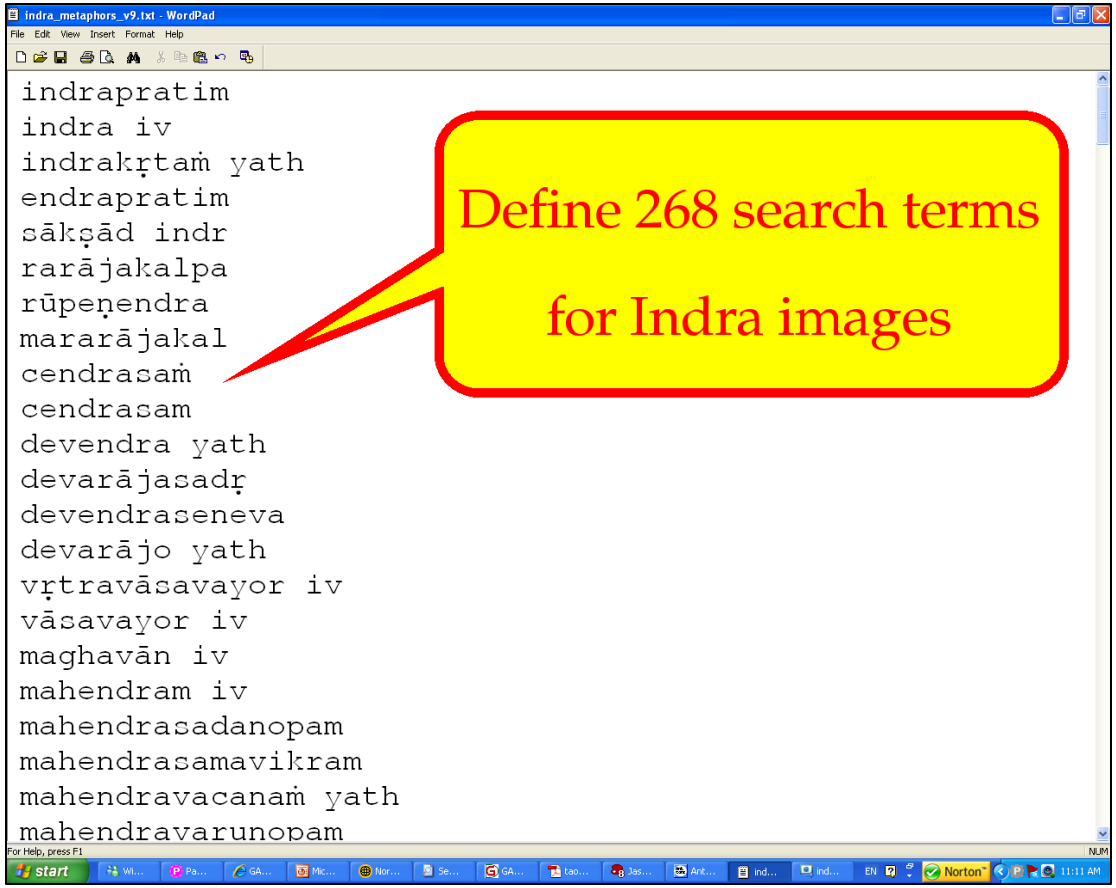
start Window... Pamela ... 2 Win... 2 Inte... Master ... AntCon... Microso... Jasc Pai... elephan... EN Norton 3:03 PM

Indra images

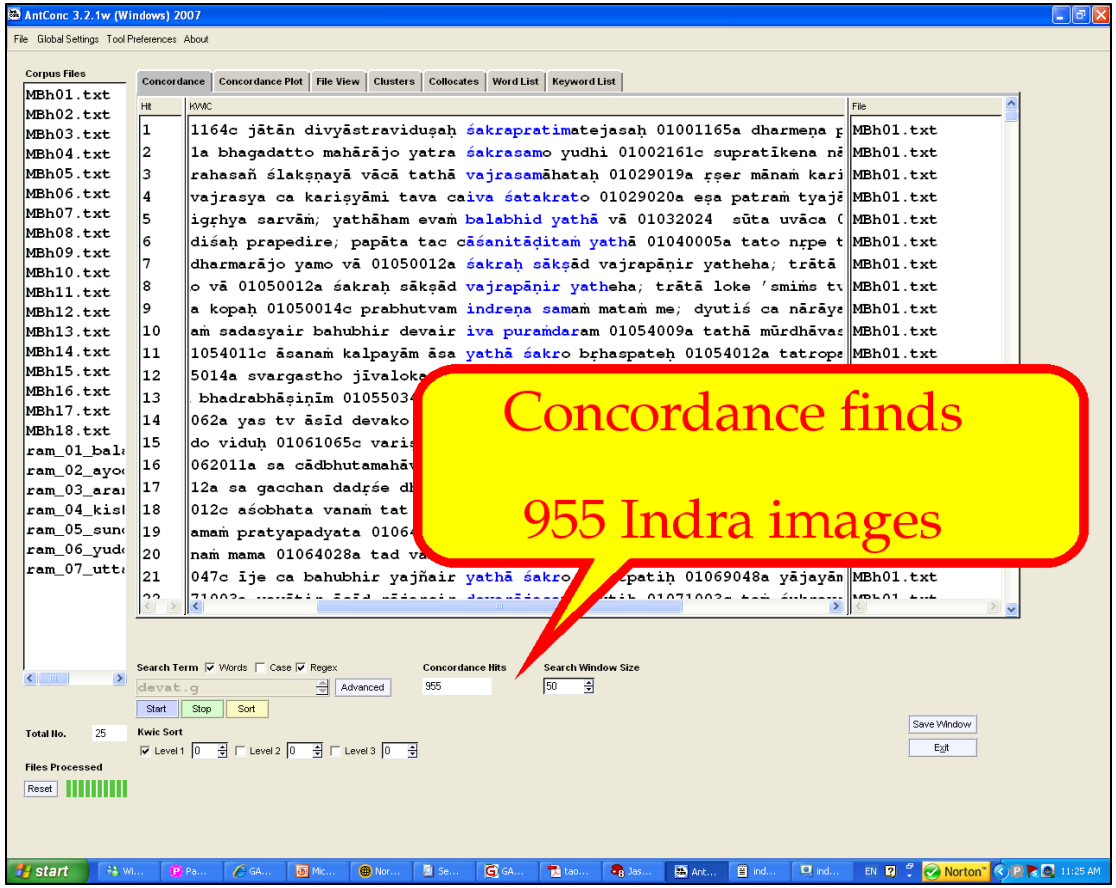
- Effectiveness of this method is shown by accurate identification of Indra images in the *Rāmāyaṇa* via lexical structures previously found in the *Mahābhārata*
- Files of lexical search terms can be distributed as independent work products for re-use on any other corpus

Mahābhārata	720
Rāmāyaṇa	235
Total Indra lines	955

Example: 268 Indra keywords



Example: 955 Indra images



Current status of project

- Work is being done gradually in stages
- Grammatical pattern files are being developed for more objects of comparison
- Interim results are available for use by other researchers via web
- mywhatever.com/sanskrit/epics

Good News: Benefits

- Multiple researchers can use the grammatical pattern files to study other Unicode texts
- Concordance software is free and works well with Sanskrit
- Distribution costs via the web are minimal

Related technologies

- Semantic and Pragmatic annotation
  - TEI - Text Encoding Initiative
  - XML “tagging” of various other kinds
- Web Ontology Language (OWL)
  - Touted as the foundation for the next generation of intelligent web applications (“semantic web”)

Knowledge Management terms

- **Tagging** is the placement of computer codes (metadata) within a stream of text to flag specific concepts within a specific corpus
- **Ontologies** are formal specifications of how concepts relate to one another in meaningful ways (organized knowledge schemas)
- If both are available for a text, computer programs can make logical inferences about what the content “means” to humans

TEI: Interpretation tagging

- Once interpretation elements are defined, they can be linked to the text by the analysis attribute (**ana**) on any element:

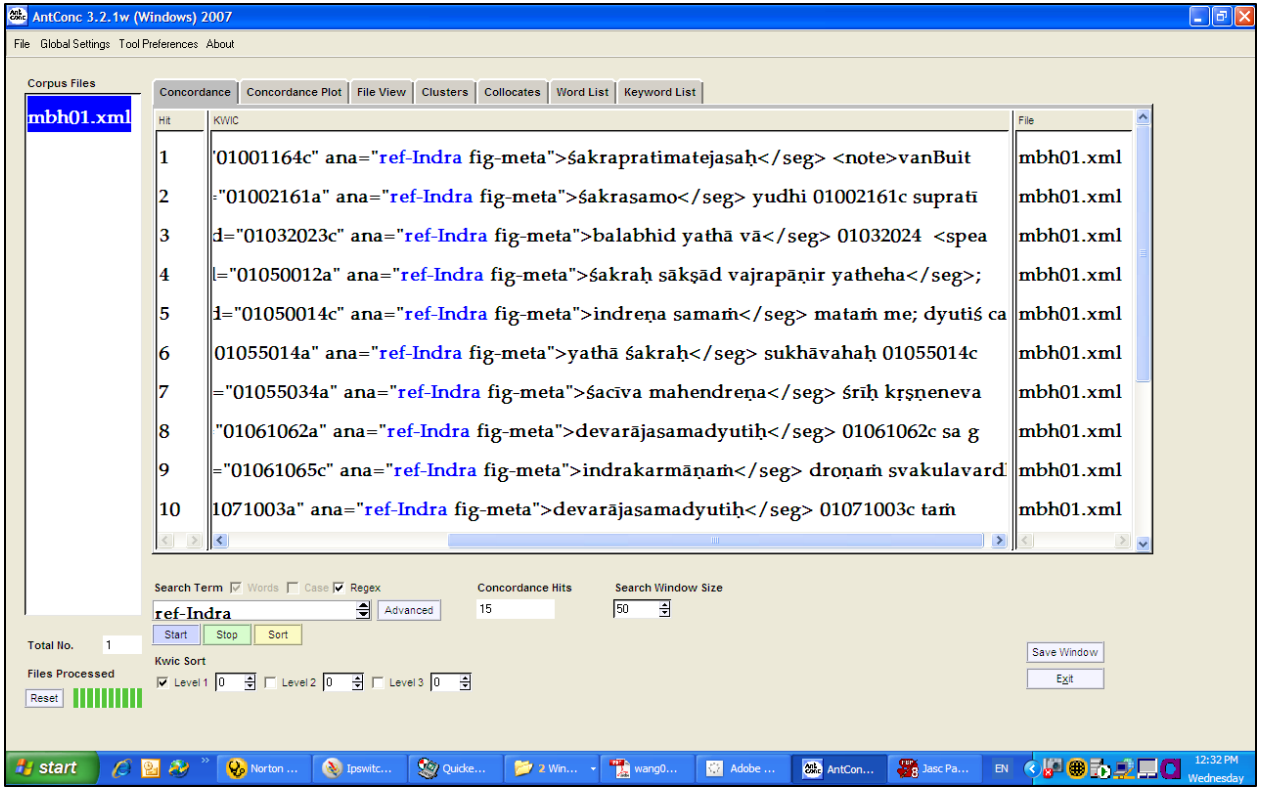
<seg id="MBH6.43.34" ana**=**"fig-sim ref-Indra ref-Vrt set-battle resp=rksharma">
vṛtravāsavayor iva</seg>

TEI: Feature structures

01001164a mahatsu rājavaṁśeṣu guṇaiḥ samuditeṣu ca
01001164c jātān divyāstravidusaḥ
<seg id="01001164c" ana="ref-Indra fig-sim">**śakrapratimatejasaḥ**</seg>

```
<fs>
      <f name="image-subject">
       <symbol value="kings"/>
      </f>
      <f name="image-object">
       <symbol value="Indra"/>
      </f>
      <f name="image-sharedProperty">
       <symbol value="splendor"/>
      </f>
     </fs>
<note>vanBuitenen trans.: "their [i.e., the kings] splendor was a match for Indra's"
     </note>
```

AntConc searches TEI tags



Web Ontology Language (OWL)

- Designed for use by applications that need to process the content of information instead of just presenting information to humans.
- OWL facilitates machine interpretability of content by providing additional vocabulary and formal semantics.
- http://www.w3.org/TR/owl-features/

OWL ontology terms

- Classes = groups of individuals that belong together because they share some properties.
  - Class(Devas) = (Indra, Śiva, Viṣṇu)
- Individuals = instances of classes
  - Indra is an instance of the class Devas
- Properties
  - Indra(hasProperty) = (valor, splendor, might)
  - Indra(hasOpponent) = (Vṛtra, Maya, Prahlāda)

Questions for discussion

- Who is working on semantic tagging of Sanskrit corpora?
- Has someone got a good method for multi-site, multi-user collaboration on electronic tagging and ontology development of this type?

Credits

- Electronic text of the critical edition of the *Mahābhārata* is John Smith's revision of Prof. Muneo Tokunaga's version, and is made available by the Bhandarkar Oriental Research Institute (BORI) in Pune.
  - http://bombay.indology.info
- Electronic text of the *Rāmāyaṇa* is John Smith's revision of Prof. Muneo Tokunaga's version.
  - http://bombay.indology.info
- AntConc concordance software was developed by Laurence Anthony, Waseda University, Japan
  - http://www.antlab.sci.waseda.ac.jp/
- Protégé Ontology Editor is distributed by Stanford University
  - http://protege.stanford.edu

TEI - Sanskrit Task Force report

- In 2004 John Smith proposed methods for Sanskrit word boundary issues
  - http://www.tei-c.org/Activities/Workgroups/CE/cew12.pdf

```
<choice>
<seg type="compound">
sarvavidvajjanāpriyam
</seg>
<seg type="analysis">
<seg type="level1">sarva</seg>
<choice>
<seg type="level1">vidvaj</seg>
<seg type="level3">vidvat</seg>
</choice>
<seg type="level2">jana</seg>
<seg>apriyam</seg>
</seg>
</choice>
```