

# CORPUS FACTORY

---

Adam Kilgarriff  
Lexical Computing Ltd., UK  
[adam@lexmasterclass.com](mailto:adam@lexmasterclass.com)

Siva Reddy  
IIIT Hyderabad, India  
[gvsreddy@students.iiit.ac.in](mailto:gvsreddy@students.iiit.ac.in)

Jan Pomíkálek  
Masaryk Uni., Brno, Cz  
[xpomikal@fi.muni.cz](mailto:xpomikal@fi.muni.cz)

Avinesh PVS  
IIIT Hyderabad, India  
[avinesh@students.iiit.ac.in](mailto:avinesh@students.iiit.ac.in)

# Problem

Lexicography requires large corpora  
**but**  
many languages *lack large corpora*

# Outline

- Introduce you to
  - Web Corpora Collection
  - Corpus Factory
- Method of Corpus Building
- Evaluation and Results
- What all can we do with Corpus Factory??
- Conclusions

# Web Corpora Collection

- Goal: Building large corpora (100 million words)
- Manual corpus-building
  - slow, labour intensive and expensive
- Solution: Using Web as corpus
- Is the Web a Corpus ?
  - Yes.
  - Question to be asked: Is it a good corpus for the task at hand ?

# Web Corpora Collection: cont..

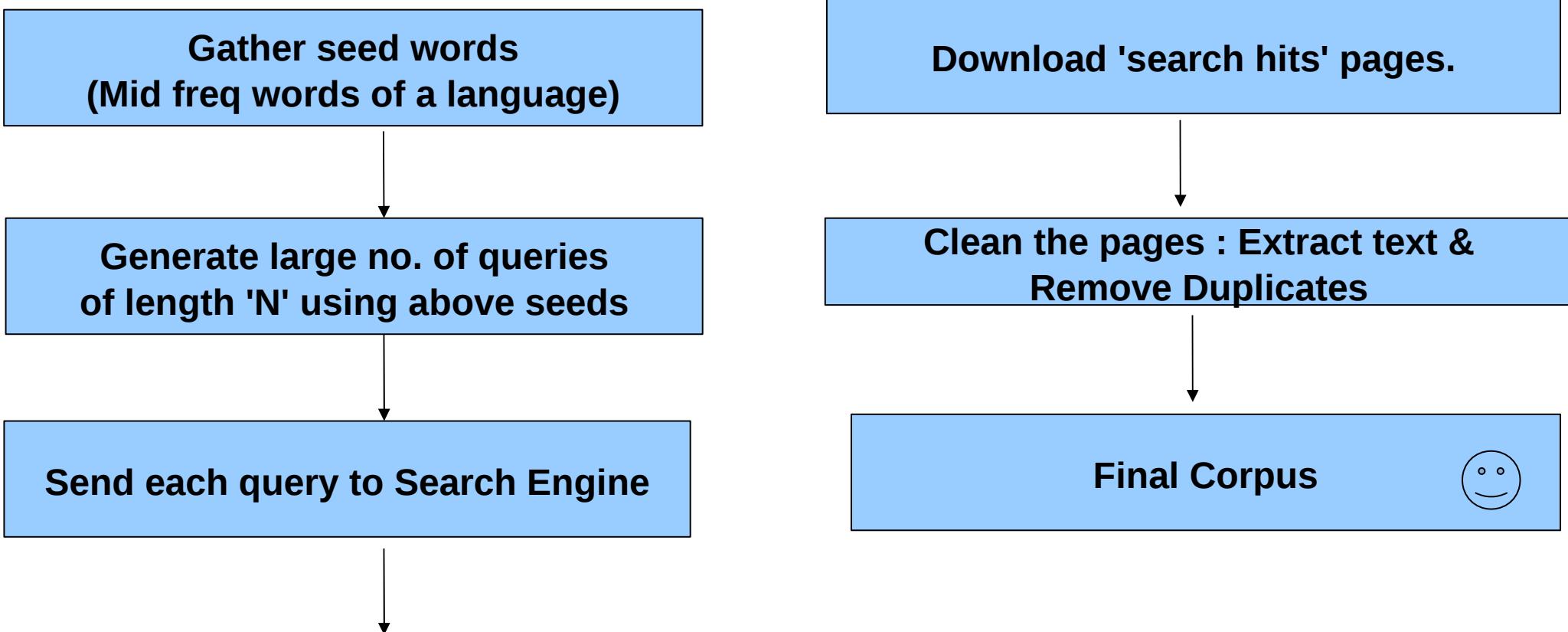
## ■ Lexicography and NLP

- ❑ Larger the corpus, the better it is.
- ❑ Many Domains

## ■ Why use Web ?

- ❑ Largest source of electronic text.
- ❑ Size of Web.
- ❑ Capture changes in language – New vocabulary.

## Earlier approaches: BooTCat (2004) and Serge Sharroff (2006)



# Corpus Factory

- Minimize human intervention
  - Automate the process of collection
- Goals: Generate any language corpus
  - In a less time
  - With minimal labour, costs
  - Large and Clean Corpus over many domains.
- A factory for building large scale corpus of any language.

# Corpus Factory Goals

At each step of  
web corpora collection,  
we identify the **bottleneck** and solve it

# Step1: Gather Seed Words

- Sharoff used 500 common words drawn from word lists from pre-existing corpora
  - Eg: BNC for English, RNC for Russian
- Bottleneck: No pre-existing large general corpora for many languages.
  - Seed words from many domains required.

# Step1: Gather Seed word

- Wikipedia (Wiki) Corpora : Gather seeds from it
  - Articles from many domains
  - Cheap
  - 265 languages covered : More to come
- Extract text from Wiki.
  - Wikipedia 2 Text
- Tokenise the text.
  - Morphology of the language is important
  - Can use the existing word tokeniser tools.

# Step 1: Gather Seed word

## ■ Thai Word Segmentation

### □ Before tokenization

ป้า ณูหาของประเทศไทยม่าในภูมิภาคคីօនះទេ

(Gloss: Burma's problems in the region)

### □ After tokenization

ป้า ณูหา/ ของ/ ประเทศไทย/ พม่า/ ใน/ ภูมิภาค/ គីន/ នះ/ ទេ  
problem/ of/ Country/ Burma/ in/ Region/ is / ?

## ■ Lemma vs word form

ស្រែចង់តុល់ vs ស្រែចង់

# Step 1: Gather Seed word

- Build the word frequency lists
  - Sort based on document freq (Descending)
- Top words in the frequency list are the most frequent (Function) words of the lang.
  - Top 500 words (roughly)
  - Helps in identifying connected text.
- We select mid frequent words as seeds
  - 1000<sup>th</sup> to 6000<sup>th</sup> words (roughly)

# Step 2: Query Generation

- Bottleneck: Identifying length of a query
- Shorter length
  - Less number of queries are generated
  - Final corpora size is small.
- Longer Length
  - Data sparsity problems.
  - Less number of hits
- Reasonable length
  - Min hit count of most of the queries is 10.

# Step 2: Query Generation

	n=1	2	3	4	5	Best
Dutch	1300000	3580	74	5	-	3
Hindi	30600	86	1	-	-	2
Indonesian	29500	1150	78	9	-	3
Norwegian	49100	786	9	-	-	2
Swedish	55000	1230	33	7	-	3
Telugu	668	2	-	-	-	2
Thai	724000	1800	193	5	-	3
Vietnamese	1100000	15400	422	39	5	4

# Step 3: Collection

- Around 30,000 queries are generated in the previous step.
- Retrieve top 10 search hits of each query.
  - Yahoo Search API
- Download all pages of search hits.
- Downloaded pages contain unwanted (markup) text.

# Step 4: Cleaning

## ■ Body Text Extraction

- Bioler plate text like navigation bars, advertisements and other recurring material like legal disclaimers are removed
- Body Text Extraction algorithm (BTE, Finn et al. 2001)
  - Bioler plate generally is rich in markup
  - Body Text is light in markup.

## ■ Performed on all the downloaded pages to get plain text pages.

# Step 4: Encoding Issues

- Search engines generally normalize different encodings to UTF-8 before indexing.
- The pages downloaded may have different encodings.
- Example: In Thai, TIS-620 and UTF-8 are famous.
  - Identify encoding of the page.
  - Normalize it to UTF-8.
  - 'iconv' tool is used to normalize Thai.

# Step 4: Encoding Issues

- Very bad hit on Indian Languages
  - Many encodings <--> font pairs exist
  - Cannot be retrieved by UTF-8 query
  - Solution
    - Generate Queries in the native encoding
    - Padma Plugin

# Step 5: Filtering

- Pages may contain unconnected text which is not desirable.
  - Eg: Menu of a hotel, list of names etc.
- Connected text in sentences reliably contains a high proportion of function words (Baroni, 2007)
- We determine the ratio of function words to non function words from wiki corpora.
- Discard the pages if this ratio is not met.

# Step 5: Near Duplicate Detection

- Pages may contain duplicates.
- Duplicates are removed using Broder et al (1997) similarity measure.
  - Two documents are similar if the similarity is greater than a threshold.
  - Similarity is based on the number of overlaps in their n-grams.
- Duplicate pages are removed.

Final Corpus of the  
desired language  
is obtained.

# Wiki and Web Corpora

	<b>Wiki Corpora</b>	<b>Web Corpora</b>
Dutch	30.0 m	108.6 m
Hindi	2.5 m	30.6 m
Indonesian	8.5 m	102.0 m
Norwegian	19.1 m	94.9 m
Swedish	9.3 m	114.0 m
Telugu	0.2 m	3.4 m
Thai	6.2 m	81.8 m
Vietnamese	9.5 m	149.0 m

Table 4: Sizes of Wiki and Web Corpora (in millions of words)

# Web Corpora Statistics

	Unique URLs Collected	After Filtering	After Duplicate Removal	Web Corpora Size	
				MB	m Words
Dutch	97,584	22,424	19,708	739	108.6
Hindi	71,613	20,051	13,321	424	30.6
Indonesian	79,402	28,987	27,051	708	102.0
Norwegian	258,009	66,299	62,691	628	94.9
Swedish	168,511	31,683	28,842	719	114.0
Telugu	37,864	6,178	5,131	107	3.4
Thai	120,314	23,320	20,998	1200	81.8
Vietnamese	106,076	27,728	19,646	1200	149.0

Table 3: Web Corpora Statistics

# Evaluation

- Corpus evaluation is a complex matter.
- Good Corpus??
  - ❑ If it supports us in doing what we want to do.
  - ❑ Only after using the corpus we get to know.
- The other strategy generally used is by comparison
  - ❑ comparing one corpus with another  
i.e. comparing frequency lists of the two corpora

## Evaluation: cont..

- For each of the languages, we have two corpora available:
  - the Web corpus and the Wiki corpus.
- Hypothesis: Wiki corpora are more ‘informational’
  - Informational --> typical written
  - Interactional --> typical spoken

# Comparing Wiki and Web Corpora

	Wiki Corpora	Web Corpora
Dutch	30.0 m	108.6 m
Hindi	2.5 m	30.6 m
Indonesian	8.5 m	102.0 m
Norwegian	19.1 m	94.9 m
Swedish	9.3 m	114.0 m
Telugu	0.2 m	3.4 m
Thai	6.2 m	81.8 m
Vietnamese	9.5 m	149.0 m

Table 4: Sizes of Wiki and Web Corpora (in millions of words)

# Comparison

- First and second person pronouns are strong indicators of interactional language.
  - For English: *I me my mine you your yours we us our*
- Ratio of common 1<sup>st</sup> and 2<sup>nd</sup> person pronouns of web and wiki corpora per million corpus are calculated.

# Results

Dutch			
Word	Web	Wiki	Ratio
ik	5786	2526	2.28
je	4802	975	4.92
jezelf	96	9	10.03
kij	188	37	5.06
jou	102	19	5.16
jouw	99	19	5.05
jullie	367	112	3.28
me	599	294	2.03
mezelf	41	5	6.89
mij	768	344	2.23
Total	14221	4771	2.98

Results prove that  
Web Corpora are more interactional.

# What all can we do with Corpus Factory

- > Platform to create any language resource.
- > Corpus Factory + Sketch Engine
- > Future Goal: Provide resources to as many languages as possible.

# Swedish in Sketch Engine

manuset in casa de Hällström-Stanisic , men **nycklarna** hade förväxlats så vi kom inte in .  
är om man inte letar efter den tappade **nyckeln** under gatlyktan där det är ljust i stället för  
judar -- men kontrollen över media är **nyckeln** till den judiska makten idag , inte kontrollen  
var presidentens ansvar , för han hade **nycklarna** . </s> <s> Hans uppgift var att hålla  
kyrka . </s> <s> Han håller prästadömets **nycklar** . </s> <s> Genom inspiration till dem som innehår  
<s> Genom inspiration till dem som innehår **nycklar** i kyrkan kallar han varje president för  
prästadömet återställdes med alla dess **nycklar** till Joseph Smith . </s> <s> Och jag bär  
mitt högtidliga vittnesbörd om att dessa **nycklar** har vidarebefordrats i våra dagar till  
Oavsett vad ditt företag , kommunikation är **nyckeln** för att få ditt budskap till teamet och  
om En kombination av väggar , lås och **nycklar** . </s> <s> En kombination av väggar , lås och  
<s> En kombination av väggar , lås och **nycklar** över till kombinationslås och slänga alla  
över till kombinationslås och slänga alla **nycklar** enda polisen har på honom är att han har  
enda polisen har på honom är att han har **nyckel** - då blir låssmeder lika överflödiga  
</s> <s> Läs inte detta som någon slags in till huset , men utan ytterligare bevis  
ringar vår hyresvärd och hon meddelar att till boken , alltså . </s> <s> Det är  
ordnade sig till det bästa **nyckeln** ligger i brevlådan . </s> <s> Hon trodde  
Verkligheten blir allt \_mer påtaglig . </s> <s> **Nyckeln** i låset och klev in i stugan . </s> <s>  
SD går till 20-30%. Tror jag . </s> <s> **Nyckeln** till förändring är inte att SD går  
för att stänga kyrkan , bröts axet av **nyckeln** är att även ett annat parti tar sitt  
en saga . </s> <s> Kistan är läst , och **nyckeln** , därför \_att ett hårt hopviket papper  
finns i gott förvar ; användas får den

# Telugu in Sketch Engine

- 15231.txt బాల్ బాయ్ను . అక్కడ **బంతి** అందించే వ్యక్తిగా  
9670.txt నాకివ్యని వాడి **బంతి** పగిలిపోయి ఎగరనప్పట్లూ  
25124.txt మ్యాచ్ లో చివరి **బంతి** కి పోరు కొట్టి  
9149.txt ఇప్పటికీ ముద్ద బంతి పూవు లాంటి తెలుగు  
9335.txt ఇనుము-రాయి తో చేసిన **బంతి** తన గరిమ బలాన్ని  
9335.txt క్షేత్రంలో జరిగితే అక్కడ **బంతి** పోటాను అన్న మాట  
12463.txt అభివర్ణించాడు . **బంతి** గతి మారుతుంది .  
19783.txt తీయాలనుకోండి . ఒక **బంతి** ఎగురుతున్నట్లు  
19783.txt దానికి ముందుగా **బంతి** వేరు వేరు స్థాయిలలో  
19783.txt గీసుకోవాలి . అంటే **బంతి** ఎగురుతున్నప్పుడు  
19783.txt అయిపోతుందనుకున్నారా ? . </p><p> ఆ **బంతి** ఒకవేళ నేలకు తగిలి  
19783.txt నేలకు తాకినప్పుడు ఆ **బంతి** ఏ మేరకు కుచించుకుపోతుంది  
35356.txt ఎక్కువగా పూస్తాయి . **బంతి** , చేమంతి , నంది వర్ధనం  
7076.txt పద్యాలు . </p><p> ముద్ద బంతి పూవులో ; ( మూగ  
15911.txt సాధారణం అయిపోయింది . ఆ **బంతి** చుట్టుపక్కల ఇళ్ళలో

# Future Work

Prepare corpora for all the official languages of the European Union, Korean, Tibetan and African languages.

- Encoding problems
- Estimate the Size of Web
- Bing Vs Google Vs Yahoo

# Conclusions

- Corpus Factory presents
  - A method for developing large general-language corpora which can be applied to many languages
  - Many to come

# Thank you

Lexical Computing Ltd,  
sketchengine.co.uk

[inquiries@sketchengine.co.uk](mailto:inquiries@sketchengine.co.uk)