Comparable Corpora BootCaT (CCBC)

Adam Kilgarriff, Avinesh PVS, Jan Pomikalek

Lexical Computing Ltd.

Just-in-time corpora

- Krista Varantola
- Translators, terminologists
- *In-domain terminology:*
  - Domain dictionaries
    - Don’t exist
    - Out of date
    - Not accessible
- Collect in-domain web pages
- Instant corpus

BootCaT (Bootstrapping Corpora and Terms)

- Baroni and Bernardini 2004
- User: input ‘seed terms’
- Send 3-at-a-time to a search engine
  - Returns search hits page
- Retrieve those pages
- A corpus!
  - Cleaning, deduplicating, linguistic processing
- Extract terms
  - Can use extracted terms as seeds, iterate

Works well

- Widely used
- More implementations
  - SkE has WebBootCaT, web front end
- Secret:
  - *piggybacks on search engines*
  - They do the donkey-work
    - on-domain, text-rich pages, no spam, …

Also in use for

- General language corpus
  - Long list of general seed words
    - Pioneer: Serge Sharoff
    - LCL: Corpus Factory
- ‘Varieties of Learner English’
  - General English, same queries *except*
    - Region=UK, US, Canada, Aus, China, Japan, Korea
  - Validation under way

The Sketch Engine

- Corpus query tool, since 2003
- Widely used by lexicographers
  - Commercial
    - OUP, CUP, Collins, Macmillan, Le Robert, Cornelsen, Shogukakan
  - National dictionary projects
    - Bulgaria, Czech Republic, Estonia, Netherlands, Slovakia, Slovenia
- Universities
  - Linguistics, language research, NLP, language teaching, teaching translation

55 languages and counting

Large corpora ready-to-use for

Arabic Bengali Bulgarian Chinese Czech Croatian Danish Dutch English Estonian Finnish French German Greek Gujarati Hebrew Hindi Indonesian Irish Italian Japanese Korean Latin Malay Malayalam Norwegian Persian Polish Portuguese Romanian Russian Serbian Setswana Slovak Slovene Spanish Swahili Swedish Tamil Telugu Thai Turkish Urdu Vietnamese

- Handles large corpora
  - Largest to date: 8 billion words
- Fast
- Web-based: no software to install
- Build ‘instant corpora’ from the web
- Load your own corpus
  - Quota of space on SkE server
- ***Word sketches***
  - One-page, automatic accounts of a word’s grammatical and collocational behaviour
- Free 30-day trial: sketchengine.co.uk

n Kilgarriff **corpus:** ukWaC Search

goal *(noun)* **ukWaC freq = 168345**

object_of	58924	3.0
score	8390	11.42
achieve	9422	10.05
concede	1421	9.43
accomplish	585	8.05
reach	1924	7.84
pursue	648	7.5
net	337	7.43
set	2413	7.42
attain	400	7.42
grab	406	7.39
pull	501	7.11
disallow	190	6.67
bag	186	6.64

and/or	16213	0.9
objective	858	7.01
aspiration	159	6.73
ambition	151	6.42
appearance	216	5.67
penalty	102	5.32
target	320	5.3
goal	315	5.3
dream	129	5.27
motivation	67	5.22
aim	227	5.15
try	34	5.13
vision	154	5.1
ideal	52	5.1

pp_after-i	336	3.6
minute	150	4.03
break	11	1.55
goal	17	1.18
ball	9	0.85

possessor	1934
poacher	14
opponent	39
striker	15
defender	11
visitor	67
opposition	16
government	169
charity	38
organization	21
client	61
administration	18
learner	10
organisation	71

WebBootCaT

- BootCaT integrated in SkE
- BootCaT a corpus
  - Clean, de-dupe, POS-tag, then
  - *Load into Sketch Engine*

WebBootCaT: Create corpus

Corpus ID volcano_en

Unique identifier of your corpus. May only contain letters, numbers, underscores and hyphens.

Language English

Creating BootCaT corpora is available only for those language which we can at least tokenise. All such languages are listed here.

Build word sketches ☑

This option only has effect if a pre-loaded sketch grammar is available for the selected language.

Input type

- (•) Seed words
- ( ) URLs

Select "URLs" to download data from specified URLs rather than use seed words for finding the URLs.

Seed words

```
volcanology volcanologist "volcanic
eruption" Eyjafjallajokull geodic tephra
magma stratigraphic tephrochronology
geochronological "volcanic ash" ablation
```

Use space as separator. Enclose multiword expressions into quotes (").

WebBootCaT: Create corpus

Please select URLs which you would like to process.

Cancel | < Back | OK

Query: "volcanic eruption" Eyjafjallajokull geodic

no results found

Query: stratigraphic tephra volcanology

- [x] http://en.wikipedia.org/wiki/Volcanology
- [x] http://www.unige.ch/sciences/terre/mineral/volcano/Recherche/tephraChaiten.html
- [x] http://www.earth-prints.org/bitstream/2122/2883/3/bozza%20IC-%20giaccio.pdf
- [x] http://tripatlas.com/Volcanology
- [x] http://publications.esc.cam.ac.uk:8080/509/
- [x] http://www.answers.com/topic/volcanology
- [x] http://www.therafoundation.org/articles/volcanology/
- [x] http://www.earth-prints.org/handle/2122/4827
- [x] http://www.geo.mtu.edu/~raman/papers2/CentralAmer/Alvarado/soto%20arenal.pdf
- [x] http://maps.thefullwiki.org/Volcanology

Invert selection

Query: "volcanic eruption" stratigraphic tephra

- [x] http://www.archaeologywordsmith.com/lookup.php?category=&where=headword&terms=tephra
- [x] http://en.wikipedia.org/wiki/List_of_potentially_dangerous_volcanic_eruptions

volcano_en: WebBootCaT: Downloading data...

31%

Successfully processed files	8
Files remaining	46
Data downloaded	669 kB
Tokens retrieved	25,336
Tokens per file (avg)	3,167
Time elapsed	0:43
Estimated time remaining	1:34
Average file processing time	2.1 s

Errors	13
- unable to retrieve	8
- invalid content-type	0
- file size out of range	2
- number of words out of range	3
- keywords filter applied	0
- unable to convert to text	0
- duplicate	0

Cancel processing

```
Processing http://www.answers.com/topic/tephrochronology-1
- Content-type: text/html;charset=UTF-8
- File type: html
- Bytes read: 43468
- Number of words (approx.): 138
  - Too few words (min: 300)
Processing http://geology.geoscienceworld.org/cgi/content/abstract/15/9/809
- Content-type: text/html
- File type: html
- Bytes read: 29773
- Number of words (approx.): 172
  - Too few words (min: 300)
Processing http://tripatlas.com/Vesuvius
```

ukWaC	English	1,565,
FinnishWac	Finnish	144,
frWaC	French	1,628,

Sh

My corpora

Corpus ID	Corpus name	Language	Size	
brewing	brewing	English	351,078	
camera	camera	English	80,435	
ergative_en2	ergative_en2	English	1,037,229	
tufs_anatomy	tufs_anatomy	English	94,111	
ukwac_10m	ukwac_10m	English	1,002,520	
volcano2_en	volcano2_en	English	946,819	
volcano_en	volcano_en	English	77,031	
volcanology	volcanology	English	73,906	
ergative_fr3	ergative_fr3	French	663,014	
volcan_fr	volcan_fr	French	660,178	
heidelberg-anatomy	heidelberg-anatomy	German	18,094	
sanskrit	sanskrit	Hindi	5	
turkish1	turkish1	Turkish	17,554	

Create corp

Sketch grammar development corpora

How big a corpus do we get?

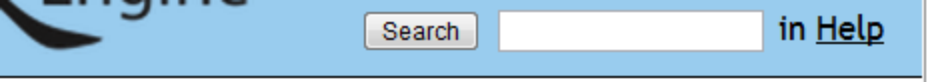
Lg		Q's sent	Volcanoes			Stradivarius		
			Url	**Doc**	**KW**	**Url**	**Doc**	**KW**
En	B	10	84	51	244	70	46	230
		50	318	180	679	230	150	1230
		250	941	515	1580	808	483	5326
	Y	10	67	39	152	60	47	148
		50	281	176	445	267	196	1071
		250	867	527	1232	937	649	3700
Fr	B	10	79	45	150	74	52	264
		50	246	152	461	225	145	1020
		250	755	506	1445	612	379	3815
	Y	10	79	36	118	82	60	720
		50	285	154	695	257	156	1155
		250	994	527	1737	843	510	2317

Observation

- Specialist domain, L1
- Specialist domain, L2
- *Matching terminology*

Going multilingual

- Translate seeds
  - English: `volcanology volcanologist "volcanic eruption" seismographs Eyjafjallajokull geodic "deformation monitoring" tephra magma stratigraphic tephrochronology geochronological "volcanic ash" ablation rhyolitic`
  - French: `vulcanologue volcanologie "éruption volcaniq ue" sismographes Eyjafjallajokull "surveillance de la déformation" géodiques tephra magma téphrochronologie stratigraphique géochronologiques "de cendres volcaniques" ablation rhyolitiques`
- BootCaT for French



volcanic *(adjective)*

volcano2_en freq = 2207

and/or	531	2.0
deadly	18	10.0
explosive	24	9.82
active	23	9.57
major	22	9.44
recent	17	9.41
Icelandic	12	9.07
felsic	9	8.99
subglacial	7	8.43
geothermal	7	8.42
hot	7	8.3
Quaternary	7	8.27
igneous	7	8.18
porous	5	8.17
future	5	8.0

modifies	2110	3.8
eruption	414	10.94
ash	241	10.9
activity	160	10.57
rock	115	9.6
hazard	45	9.18
field	47	9.0
cloud	42	8.95
glass	36	8.83
complex	29	8.69
center	30	8.68
vent	29	8.61
gas	33	8.49
event	33	8.47
plume	27	8.47

volcanique *(adjective)* volcan_fr

modifié	869	5.6
éruption	105	10.87
roche	98	10.72
activité	70	10.7
édifice	44	10.51
cendre	59	10.43
explosivité	15	9.09
risque	16	9.06
île	17	9.02
verre	15	9.0
manifestation	12	8.74
bombe	12	8.73
arc	13	8.71
faciès	12	8.66
gaz	14	8.64

pp_sous-x	2	5.6
pression	2	8.

pp_à-x	8	1.9
travers	2	10.
l	2	7.

pp_au-x	10	1.
cour|cours	2	8.

CCBC

- Input: L1, L1 seeds, L2
- Choose dictionary
  - Google as default
    - Google dictionary (25 lg pairs, limited API)
    - Google translate (1225 lg pairs, only 1 transl)
  - Option: edit translations
- Bootcat 2 corpora
- Bilingual word sketches

Bilingual word sketches (very first pass)

- For L1 nodeword *n*
  - For each of its translations n_1, n_2, …
    - For each collocate *c* in word sketch
      - For each of its translations c_1, c_2, …
        - Does c_i occur as collocate in word sketch for n_i?
        - If yes: output <*c*; n_i , c_i >
        - Add L1 and L2 examples sentences

corpus: volcano2_en Search

volcano *(noun)* volcano2_en freq = 2490

volcan *(noun)* volcan_fr freq = 1268

be	224	Close view of the eruption column of Mount St. Helens on May 18 , 1980 ; the volca photograph .
être	87	Le sommet du volcan est dégagé et l on peut apercevoir une épaisse colonne de vapeur blanche monter droit dans le
aller	6	Ce gonflement du volcan va générer des microséismes , une augmentation de l' inclinaison des pentes du volcan , une a caldeira sommitale .
produire	4	Le volcan a produit l' un des plus grands volumes de lave de partout dans le monde au cours du dernier millénaire , soit
venir	3	Les dynamismes : Les principales différences entre volcans viennent de la puissance de leur éruption , ou de leur explo
active	103	Is an ongoing eruption the only way to tell if a volcano is active ?
actif	64	Les scientifiques considèrent habituellement un volcan être actif s' il est actuellement éclatant ou montrant des signes de tremblement de terre ou les nouvelles émissions de gaz significatives .
eruption	89	Photo showing a cross section of the volcano before the 1980 eruption and the 330 Mt .
éruption	2	Lave La lave est une roche issue d \ ' un magma qui est émise sous une forme plus ou moins fluide par les volcans en éru
erupt	85	Volcanic ash : how far will it fall downwind from an erupting volcano ?
have	61	Pinatubo in June 1991 , this volcano has been an ideal location for remote sensing a volcano .
prendre	3	Puisque les coulées pyroclastiques successives se sont accumulées à la base de la montagne , le volcan a pris une forme
avoir	37	Des parties du premier cône du volcan ont été déplacées par des glaciers durant la Glaciation qui eut lieu il y a 14 0 0
Icelandic	37	Well folks , it was only a matter of time .... " Ice cap thaw may awaken Icelandic vo
islandais	2	Les effets géographiques des éruptions à l' échelle régionale Perturbations de la circulation aérienne 13 Pendant l' éru les autorités de régulation aéronautique ont convenablement joué leur rôle d' informateur par le biais des METARs (ME

Matching seeds – how?

- User translates
  - Yes but limited
- Bilingual dictionary
  - Yes but finding them??
  - Google dictionary
- Machine translations
- Wikipedia
  - Matching articles

Evaluation

- Extract terms for L1, L2
- Ask expert
  1. Are they terms
  2. Do the L1, L2 lists contain translations of each other?

- 3 lg-pairs
  - En-Fr, En-De, En-Cz
  - One expert for each pair
- 3 domains
  - Volcanoes
  - Stradivarius
  - Pancreatic cancer
    - Wikipedia: En and De only

Results

Who	Wds	Trans	Mwds
	Volcanoes, En		
E-Cz	29/30		10/85
E-De	29/30	10/29	16/85
E-Fr	29/30	19/29	24/85
	Stradivarius, En		
E-Cz	19/29		13/85
E-De	26/30	3/26	9/85
	Stradivarius, De		
E-De	16/30	2/16	6/84
	Cancer, En		
E-De	27/30	9/27	
	Cancer, De		
E-De	22/30	10/22	8/90
	Volcanoes, Fr		
E-De	27/30	19/27	5/83

In brief

- Words good
- Multiwords bad

Unithood and termhood

- To find terms
  - For multiwords only
    - Does it hang together?
    - *Unithood*
  - It it distinctive?
    - Keywords
    - *Termhood*
- We didn’t use termhood for multiwords but need to

Next steps

- Termhood for multiwords
- WebBootCaT from wikipedia
- From collocations to terms
  - More-than-2-word collocations
    - … deadline next Tuesday

Thank you

http://www.sketchengine.co.uk