



Complementing the BNC with a Corpus from the Web

American Association for Corpus Linguistics
BYU, Provo, UT, 12-15 March 2008
William H. Fletcher, U. S. Naval Academy

Goals of presentation

- Familiarize you with Web as Corpus community and my specific applications
- Outline approaches to retrieving, cleaning and standardizing Web data
- Describe my evolving online corpus databases at <http://webAsCorpus.org>
- Elicit feedback on specific user needs and wishes

WaC Community

- ACL Sigwac *Adam Kilgarriff & Marco Baroni*
- Annual Workshops
 - 2005 CL Birmingham
 - 2006 EACL Trento
 - 2007 UCL Louvain-la-Neuve CLEANEVAL
 - 2008 LREC Marrakesh
- WaC contact sites
 - WaCkywiki *Marco Baroni*
 - WaCwiki
 - Web genres wiki *Marina Santini*
 - Sourceforge depository *Stefan Evert*

WebAsCorpus.org - 1

Low-Threshold Entry to WaC

- **Web Concordancer** real-time KWIC concordances in 31 (*named*) languages using Live Search (*Yahoo! coming*)
- **Count Matching Webpages** reports how many webpages match a given set of search terms on LS and Y!, expresses as doc frequency, matches per M, and percent of total matches
 - Googleology enabler, but seizes the teachable moment.

WebAsCorpus.org - 2

Large English Web corpora *(100-500 MW)*

- 1-6-grams (fulltext*)
- Wildcard searchable (RegExp*)
- Results filterable against BNC and other large databases to isolate likely lexical innovations *(and spurious forms)*

**not on line yet*

WebAsCorpus.org - 3

English Web Corpora *(cont.)*

- 2006
 - 950,087 types
 - 97,198,272 tokens
- 2007
 - 3,123,996 types
 - 518,129,710 tokens
- 2008 *based on documents matching user queries*
 - 1,090,414 types
 - 148,505,932 tokens

WebAsCorpus.org - 4

Other Web Corpus resources

- WaCwiki for Web as Corpus community
- Word frequency lists as down-payment on Dutch Web Corpus (*initially 180 MW*)
 - Dutch (102,770 typ. / 1,605,346 tok.)
 - Afrikaans (62,785 typ. / 1,263,509 tok.)
- Free cool @webascorpus.org email addresses

WebAsCorpus.org - 5

Concept of Web Corpus 2007

- Minimum 1 gigawords of English
- representative: geographic, semantic, filetype (HTML + PDF)
- PoS-tagging comparable to BNC
 - CLAWS4 tagger, mapped onto BNC tagsets
 - post-tagging cleanup using UCREL / BNC templates
 - search by lemma or word form
- seamless integration with PIE
- query with wildcards & regular expressions
- filterable – show n -grams not in other datasets
- growing / self-renewing via actual user queries
- archiving of each release for replicability

Specific goals

– 1

- explore English beyond the BNC
 - recent and emerging usage
 - broader geographic representation
 - “long tail” victims of Zipf’s Law
- “dirty but texty”: reduce “garbage” – boilerplate, fragmentary content, non-English content – but err on side of inclusion



Specific goals

– 2

- prototype a Windows-based acquisition and processing system extensible to other languages
 - use open-source software where possible
 - Produce sharable apps / code
- deploy on shared LAMP host
 - unrestricted access yet inexpensive
 - learn to work around provider's policies

geographic concept

- “weighted” proportional distro of major English-speaking nations (*non-US 2x actual population proportion, with 10% oversampling for enough “keepers”*)
 - AU 10%
 - CA 13%
 - IE 2%
 - NZ 2%
 - UK 30%
 - US 43%
- reality: most country-specific hits – official sites or include geo. refs.
- consequently 1/2 documents fetched with no country specified for broader range of sources

semantic concepts

- for *breadth* selected terms from all semantic fields in UCREL's USAS, e.g.
 - A1.1.2 Damaging and destroying, general / abstract terms depicting / damage / destruction / demolition / pollution, etc
 - Prototypical examples: *armageddon, blemish, breakages, bulldoze, contaminating, crack...*
- for *interest* use ST from PIE and KF queries (*content words + phrases*)
- complement with pages matching user queries on WaC

filetype

- 90% HTML – general content
- 10% PDF
 - generally longer, higher-quality text
 - specific genres of interest
 - scholarly papers
 - print media
 - government documents



MS Live Search

- Powerful API supports
 - search by country
 - weighting result set by popularity, freshness and / or exactness of match, effectively many times 1000 hits
- License permits 10,000 queries per IP address (not total per license)
- Cache
 - HEAD returns doc size for pre-weeding
 - fetch typically much quicker than original
 - formats PDF usefully
 - sniffs charset and converts to UTF-8
- Yahoo! now meets these criteria too

Acquisition & processing – Plan A

concept – highly distributed parallel processing on 10-20 PCs

- ± 8000 seed ST assigned to worker PCs, which independently
 - query & fetch hits from LS
 - strip HTML, determine “keepers”
 - non-dupes, size
 - verify text-ness and English-ness
 - PoS-tag text
 - when all done, create n -grams of (un)tagged text
- worker PCs communicate with servers only to...
 - avoid previously processed or rejected documents
 - upload texts and n -grams for final databases

Acquisition & processing – Plan B

- Range of seed STs assigned manually to 3 worker PCs, which independently
 - queried & fetched hits from LS
 - cache id used to avoid multiple downloads
 - local database tracked doc stats
 - stripped HTML, but...
- all files and local databases copied manually to single PC
 - restrip, rehash to find true empties and dupes
 - rewinnow (HTML horrors)
 - tagging put on hold
- single PC processing bottleneck due to huge number of files per directory (>100k)

Bumps along the road

- PDF (*iFilter?*)
- PHP `strip_tags()`
- search by country codes
 - consistent, but not 100% reliable
 - excludes pages without specific geographic references
- encodings
 - hybrid encodings
 - mapping to CLAWS4 specs

Cleaning the data

– 1

- avoid garbage pages
 - ignore outsized pages ($< 5k / > 300k$)
 - eliminate / excerpt* pages with
 - too few / many words ($< 500 / * > 50,000$ words)
 - too short / long paragraphs ($< 13 / > 500$ w/¶)
- sample text starting / ending with first / last paragraph with at least 13 words

**not on line yet*

Cleaning the data

– 2

- (re)sniff language (*SE unreliable*) and assess “textiness” (*and uniqueness*) at paragraph / chunk level
- score by ratio of English markers
 - initial *th- wh-*
 - **distinctive** high-freq function words
 - *it if* but not *is* (*Dutch, Afrikaans...*)
 - *you* but not *I* (*Roman numeral*)
 - *and* but not *a* (*section, preposition, article*)
- eliminate documents / chunks that fall below threshold values

Cleaning the data

– 3

- setting thresholds
 - BNC texts range 20-35% markers/tokens exceptions:
 - real-estate ads (fragmentary)
 - rap song (*d-* for *th-*, other non-standard)
 - WaCuser tossed out all
 - docs below 15%: bilingual, lists, SE spam
 - chunks below 10%
 - 10-15% in “marginal” table; includes narrative and descriptive passages rich in content words
- finer threshold criteria essential to avoid self-fulfilling prophecy

Unique documents 2007

<i>filetype</i>	<i>count</i>	<i>words</i>
HTML	689,958	828,203,995
PDF	93,133	299,891,530
total	783,091	1,128,095,525

Distribution “long tail”

	1-grams	2-grams	3-grams	4-grams	5-grams	6-grams
total unique	3,123,996	57,140,986	210,320,192	359,073,268	440,426,238	471,511,994
1x only	57.0%	67.0%	79.5%	87.7%	92.5%	94.8%
2x only	14.0%	13.1%	10.2%	7.3%	5.1%	3.9%
3 or more	29.1%	19.9%	10.3%	5.0%	2.3%	1.3%

Full-text search

- 1

- Produce concordances instantaneously from WC data
- Currently only possible by proxy via SE
 - no wildcards / RegExp
 - Exalead API?
- Resource-intensive sophisticated search perhaps not feasible in affordable shared hosting environment
- MySQL FT not (yet) scalable to GW
- SQLite 3 FT enhanced by Google?

Complementing the BNC with a Corpus from the Web

Feedback

- critiques
- suggestions
- wishes

<http://KWICFinder.com>

<http://PhrasesInEnglish.org>

<http://pie.usna.edu>

<http://WebAsCorpus.org>

fletcher@usna.edu