ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Specialised corpus construction for language professionals

Silvia Bernardini

School of Interpreters and Translators

University of Bologna

Italy

ELC - 2011 X Encontro de Linguística de Corpus

EBRALC - 2011 - V Escola Brasileira de Linguística Computacional

Corpora for language professionals

- Corpora can help to:
  - exercise and develop text analytical skills
    - reference and specialised (comparable)
  - produce more naturally-sounding translations
    - reference and specialised (comparable)
  - raise awareness of professional strategies
    - specialised (parallel)

➢ popular with linguists and translation trainers, but how about professionals?

Corpora in translation practice

- survey 2005-2006 (MeLLANGE)
- respondents:
  - professionals (74%)
  - students (26%)

1015 respondents

UK	France	Germany	Italy	Spain	undefined
567	125	25	19	4	275

Focusing on professionals

- **do you collect domain specific texts?**
  - 54.3% no
  - 45.7% yes
- **how do you use them?**
  - 50.8% read them
  - 49.2% search through with software
    - 65.7% search facility in word processor
    - 17.7% concordancer
    - 14.4% other

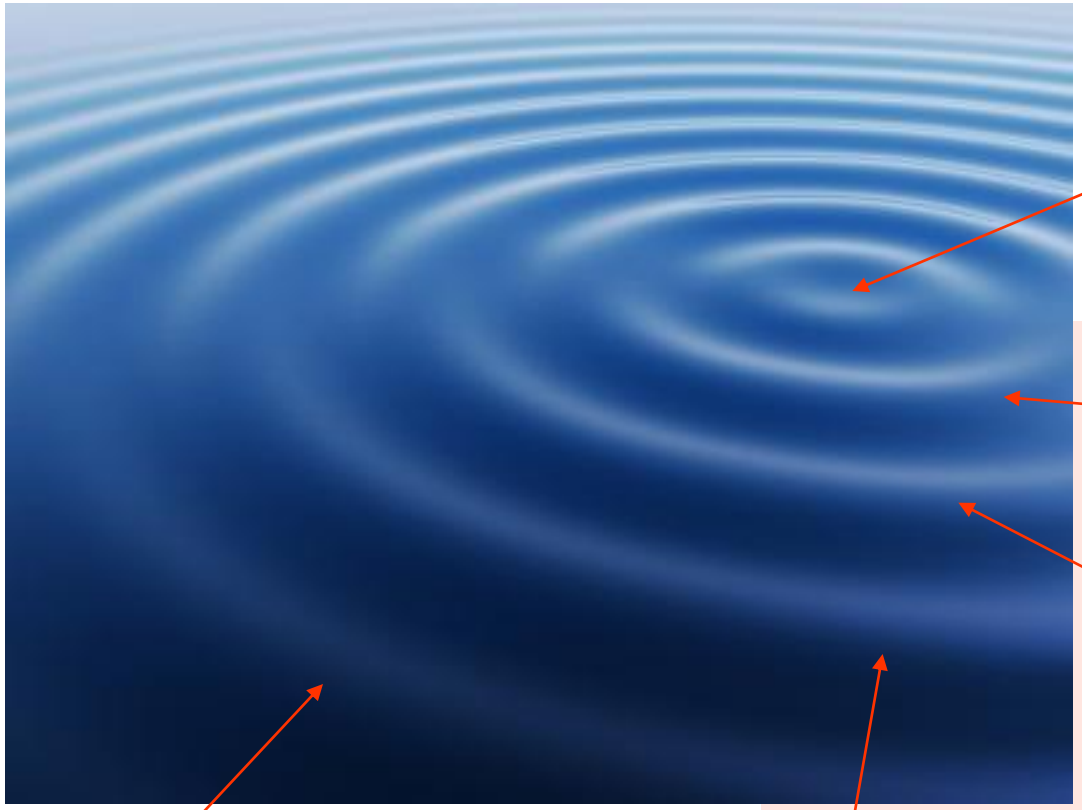
Interested?

- **if you do not use corpora, why?**
  - 42.0% never heard of them
  - 19.7% don't have time to build them
  - 17.1% don't know how to use a concordancer
  - 9.2% can't see any advantage over *Google*
  - 8.2% can't see any advantage over TMs
  - 3.7% other

Interested!

- **would you be interested in learning more about corpora?**
  - 83.7% yes
  - 16.3% no
- **would you be interested in a service providing specialised corpora?**
  - 79.7% yes
  - 20.3% no
- **would you be interested in a tool for extracting terms from specialised corpora?**
  - 80.0% yes
  - 20.0% no

The place of corpora in the documentation process



~~Translation Memories~~

Specialised manually built corpora

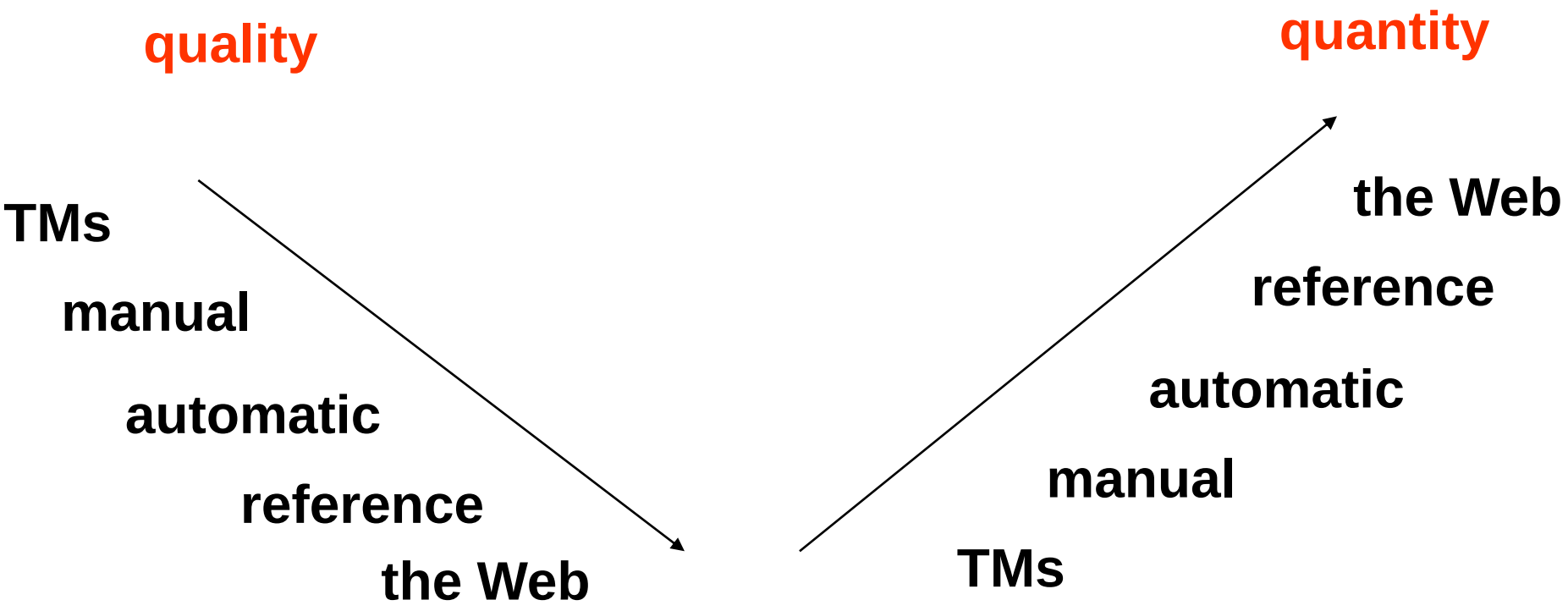
the Web

Specialised automatically built corpora

The web through Search Engines

Large general corpora

Quality vs. quantity



Time and effort required for:

resource preparation

TMs

Manual

Automatic

Reference

The Web

searching and decision making

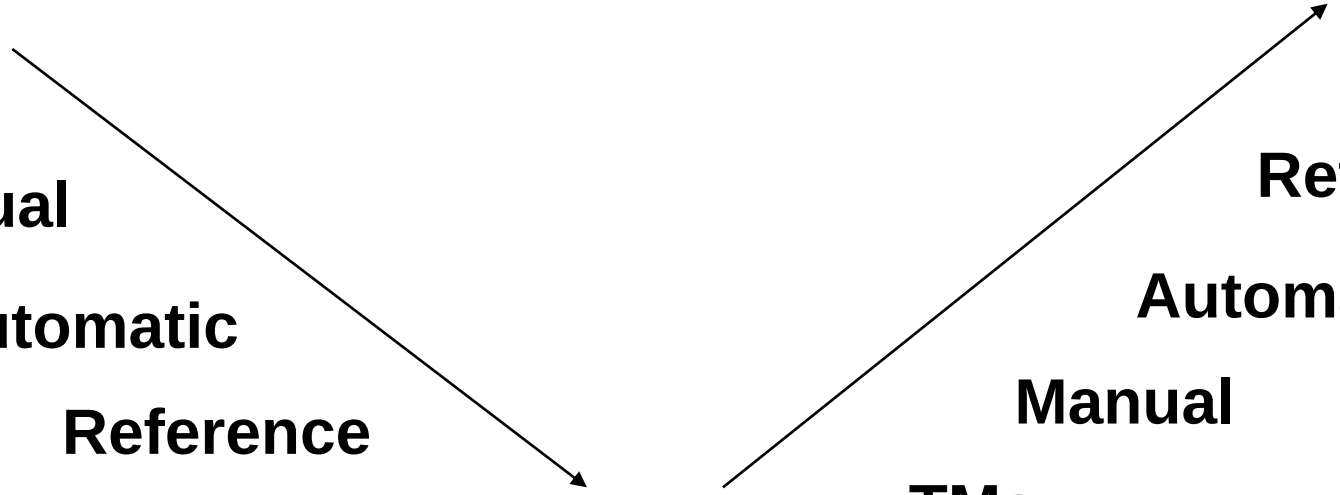
The Web

Reference

Automatic

Manual

TMs



The corpus

- A collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis. (Francis 1992(1982):17)

- A collection of naturally occurring language text, chosen to

First of all, is the Web a corpus?

1992:167)

- Finite-sized body of machine-readable text, sampled in order to be maximally representative of the language variety under consideration. (McEnery and Wilson 1996:23)

- A collection of (1) *machine-readable* (2) *authentic* texts […] which is (3) *sampled* to be (4) *representative* of a particular language or language variety. (McEnery et al. 2006:5)

The Web

- A mine of language data of unprecedented richness (Lüdeling et al 2007)
- A fabulous linguists' playground (Kilgarriff and Grefenstette 2003)
- [a] cheerful anarchy (Sinclair 2004)

Is the Web a corpus? Yes!

The definition of corpus should be broad. We define a corpus simply as “a collection of texts”. If that seems too broad, the one qualification we allow relates to the domains and contexts in which the word is used […]: *A corpus is a collection of texts when considered as an object of language or literary study.* The answer to the question “Is the web a corpus?” is yes.

Kilgarriff and Grefenstette (2003:334)

Is the Web a corpus? No!

The cheerful anarchy of the Web thus places a burden of care on a user, and slows down the process of corpus building. The organisation and discipline has to be put in by the corpus builder. […] users of a corpus assume that there is a consistency of selection, processing and management of the texts in the corpus.

Corpora should be designed and constructed exclusively on external criteria.

(Sinclair 2005)

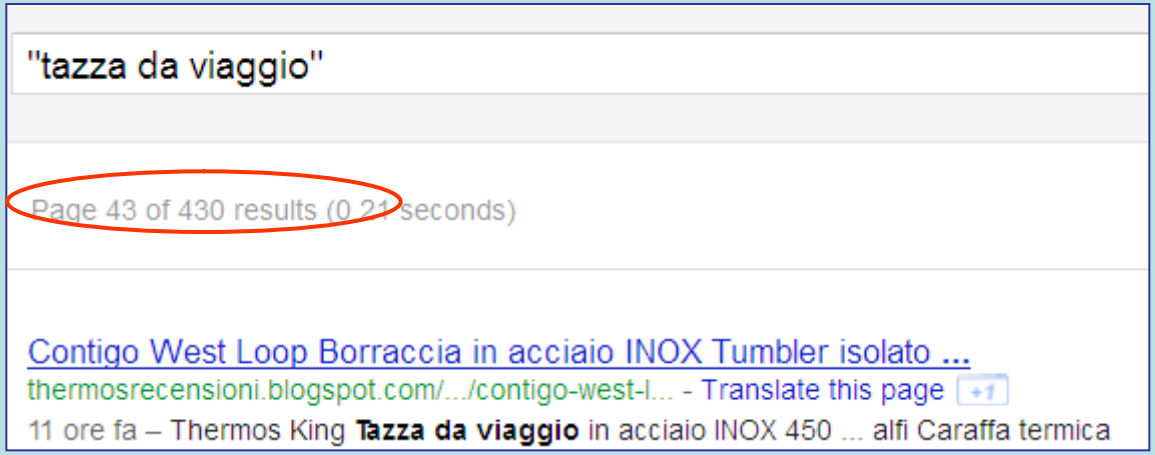
What we ***don't*** mean by *Web corpora / Web as Corpus*

- The Web as a corpus surrogate
- The mini-Web (or mega-corpus)
- The Web as a corpus supermarket
- *The corpus of the Web era*

– "Traditional" collections of texts taken from the Web
– Web as a corpus
– Saved locally

The i.e., trying to use the Web as we would use a corpus ate

Googleology is bad science!



SE post-processors?

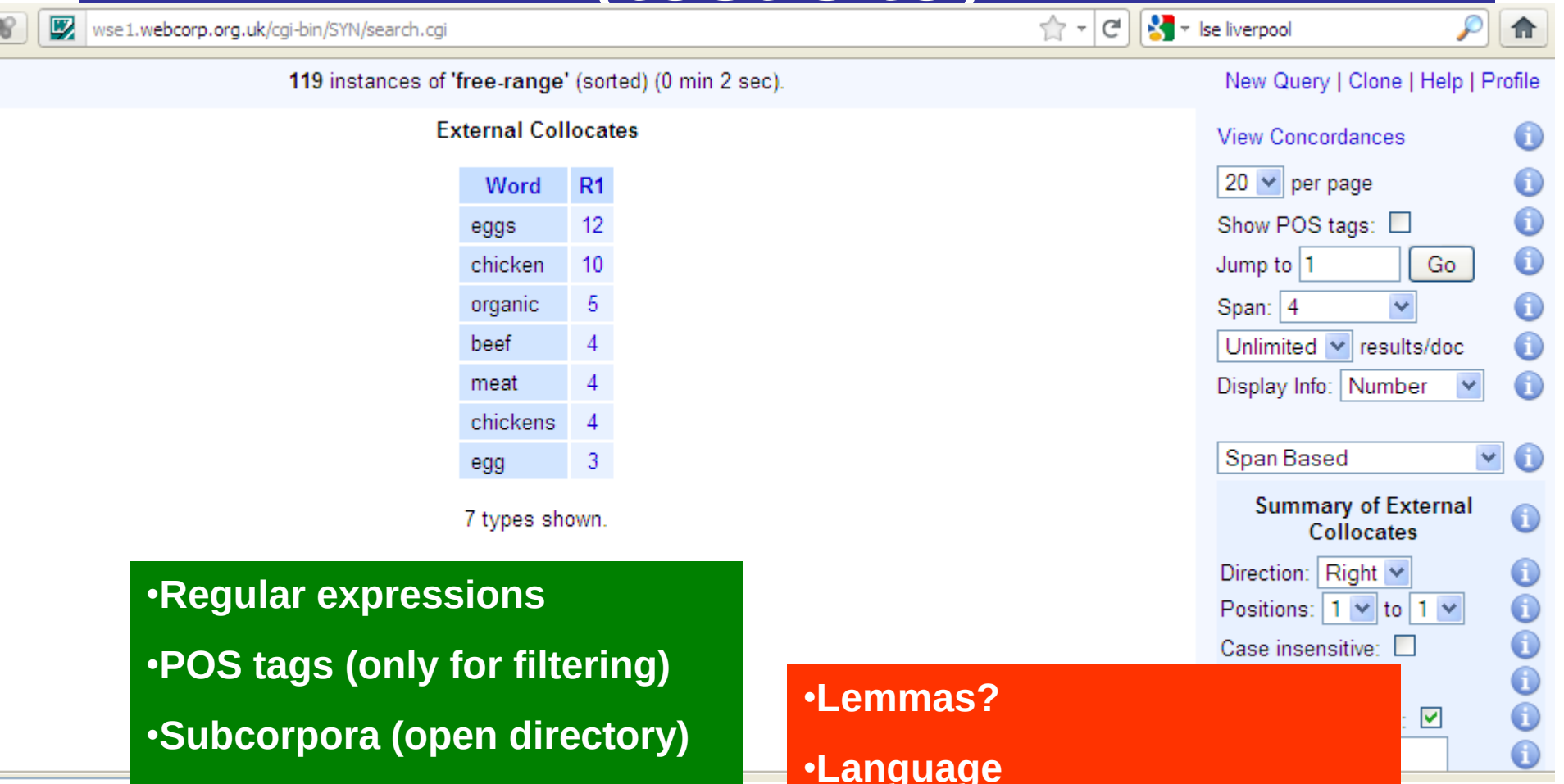
- e.g. *WebCorp*, *KWiCFinder*
  - Wildcards and tamecards
  - Concordance output
  - Collocation
- Not a solution, really
  - Slow
  - Same limits as SE
  - ➢ Not popular any longer

The mini-Web or mega-corpus

The mega-corpus/miniweb

- Baroni (2007): Effort spent by NLP community in developing Google-skills would be better spent building our own Google-sized corpora
  - (Early August 11 disaster: *Yahoo!* discontinued API service)
- Ultimate objective, build a linguist's search engine for the Web. In the meantime:
- Attempts:
  - LSE (Renouf et al. 2007)
  - The WaCky! effort

Linguist Search Engine (LSE) (test site)



- **Regular expressions**
- **POS tags (only for filtering)**
- **Subcorpora (open directory)**
- **Sorting**
- **Thinning**

- **Lemmas?**
- **Language detection/selection?**

Our own attempt
the *WaCky!* corpus pipeline

- Submit random word combinations to SE and obtain list of URLs (seeding)
- Crawling
- Code removal and boilerplate stripping
- Language filtering
- Near-duplicate detection
- Tokenization, POS-tagging and lemmatisation
- Indexing => querying
  - E.g.: ukWaC

Tools and size require Unix servers and expertise

A WaCky British English corpus (Baroni et al 2009)

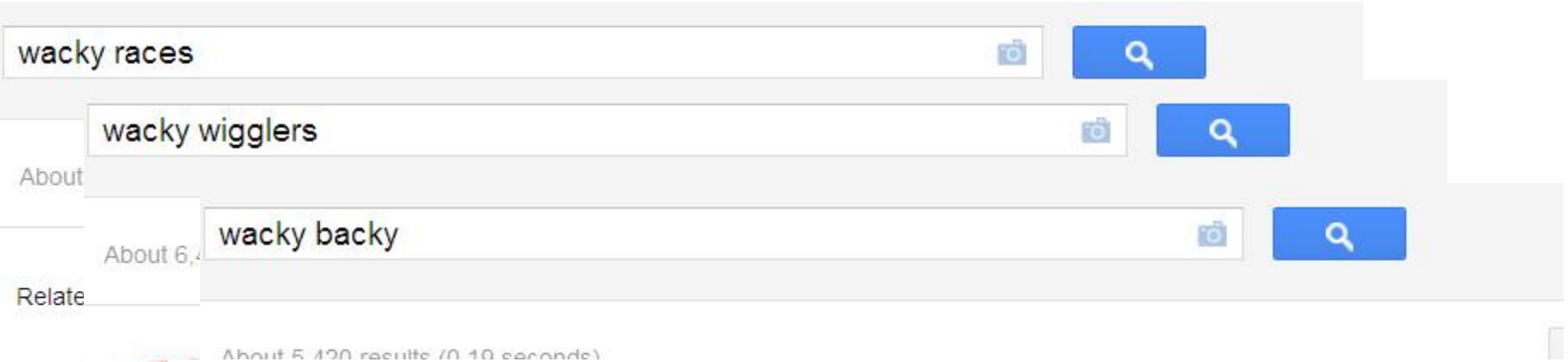
- Seeding
  - mid-frequency content words (BNC);
  - words from spoken text (BNC);
  - vocabulary list for foreign learners
- Crawl limited to UK domain and html
- Processing
  - Only files btwn 5 and 200kb kept
  - Perfect duplicates discarded
  - Code, *boilerplate*, files with unconnected text and pornographic pages removed
  - Near-duplicates removed

ukWaC info

- 2,000 seed word pairs
- 6,528 seed URLs
- 351 GB raw crawl size=>19 GB after document filtering=>12 GB after near-duplicate cleaning
- 30 GB with annotation (2.69 Million documents)
- 1,914,150,197 tokens (~2 billion words)
- 3,798,106 types

- Downloadable for research purposes
  - the corpus, seed words, tools, advice…
  - http://wacky.sslmit.unibo.it/doku.php?id=download
- Searchable
  - through a simple command line interface
  - through the SketchEngine ($)

A WaCky example
Results for wacky+NOUN (>2)




```
as they introduce the <wacky world> of kitchen science ,
 it’s dangerous . The <wacky world> of micronations
t be happening in the <wacky world> of dance over the next
pen my wallet in that <wacky world> of leather . Sle
 first drawn into the <wacky world> of Alice Cooper as a
y Park Experience the <wacky world> of Alice and all her
rough the wonderful - <wacky world> of pyramids ! I come
oods FRijj ( UK ) The <Wacky World> of FRijj Interactive
sightful and slightly <wacky world> of The Coral . The set
 : US : 1968-1970 The <Wacky World> Of Jonathan Winters
```

Mini Web / Mega corpus advantages

- Up-to-date
  - Grammaticalization, language change, new genres…
- Very very large
  - Infrequent structures / patterns / words (Zipf)
  - Web frequencies correlate better with judged plausibility [of bigrams] than corpus frequencies (Keller and Lapata 2003)
- Low-cost
  - The BNC cost over 1 million pounds!
- Can be the result of community effort
  - True Web 2.0 spirit: take from the community and then give back
  - Data, tools and heuristics in the public domain

Mini Web / Mega corpus challenges

- Cleaning techniques
  - Boilerplate stripping too greedy or too lax
- Web-tuned annotation tools
  - Dealing with non-standard uses
- Indexing and querying systems; interfaces
  - At the moment: CWB/CQP only
  - Many interfaces, but e.g. none can handle subcorpora
  - A whole new approach to data display / manipulation
- (Automatic) text classification
  - Understanding of corpus contents
  - Sub-corpus construction

The Web as a corpus supermarket

Select and download texts (semi-)automatically

- ⇒ Replicable results
- ⇒ Control over corpus contents (?)
- ⇒ Control over search methods
- ⇒ Linguistically sophisticated searches

Like traditional corpora

- ⇒ Size
- ⇒ Up-to-dateness
- ⇒ Understanding of contents/structure
- ⇒ Variety of contents
- ⇒ (Un)reliability
- ⇒ Noise

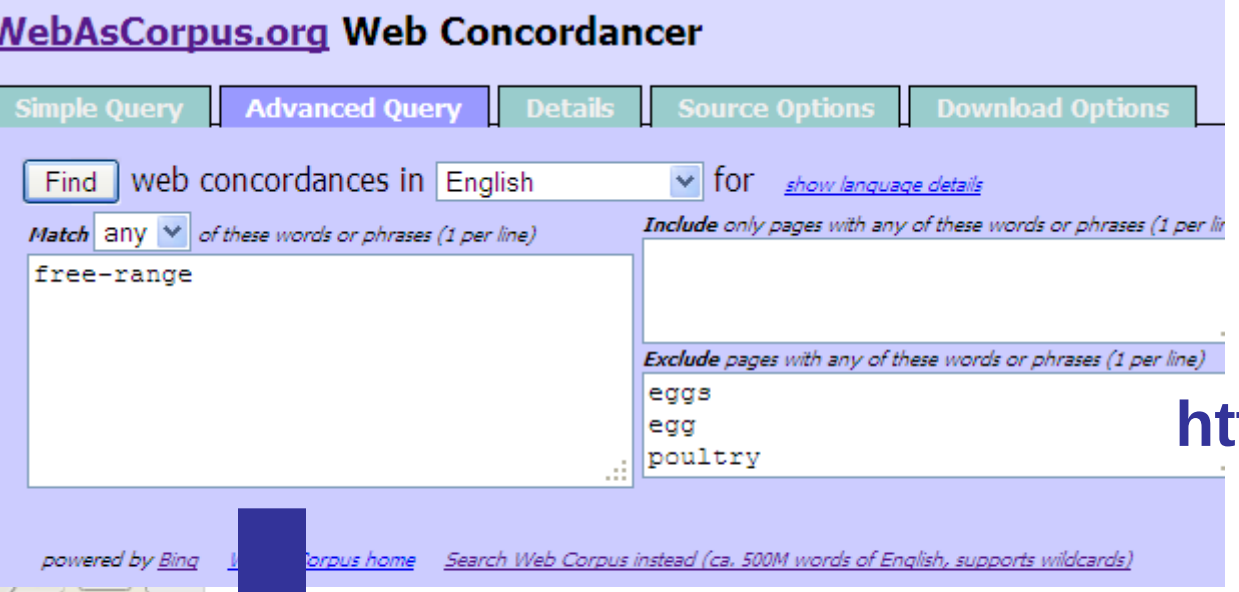
Like the Web

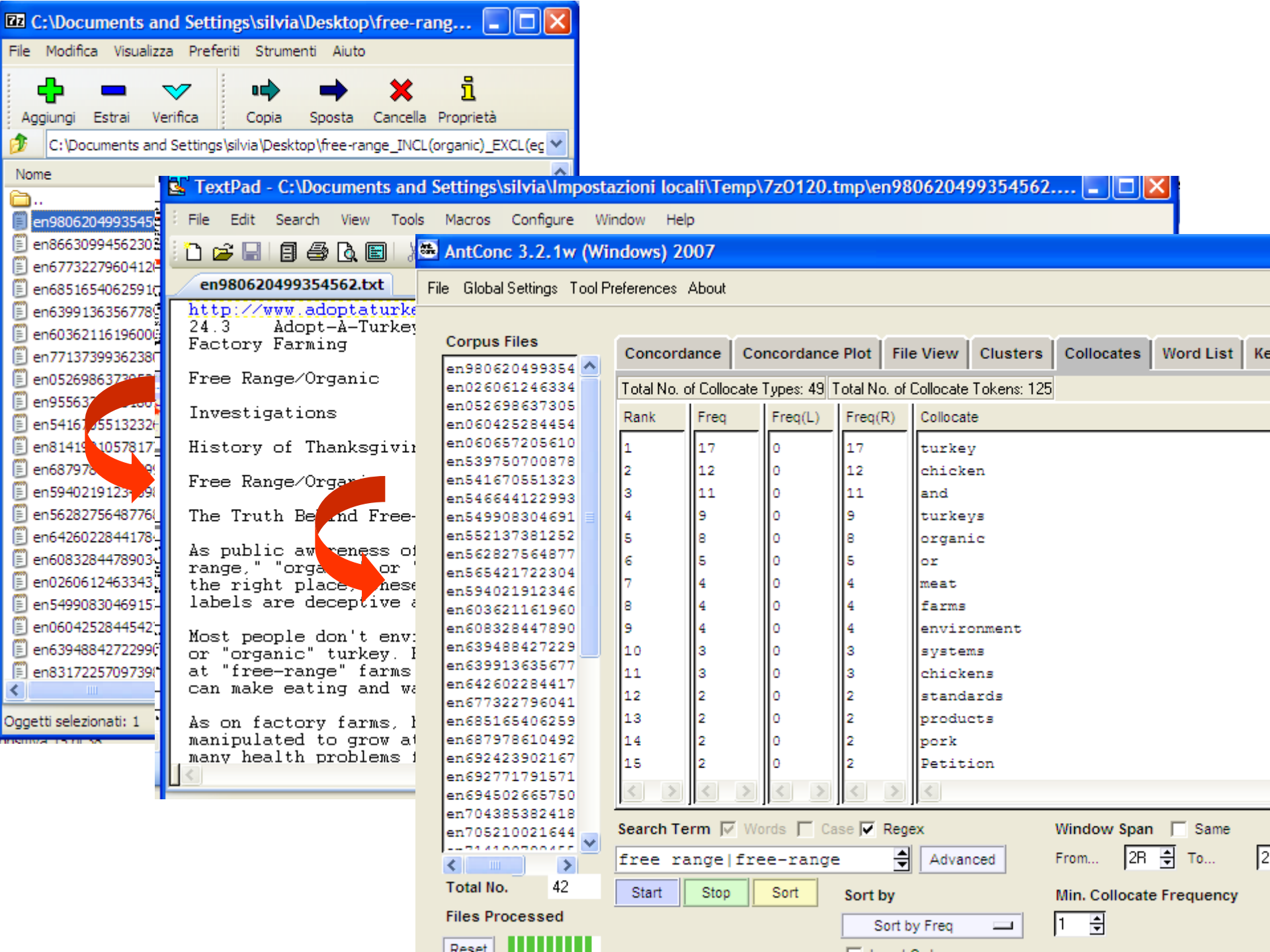
Two kinds of shopping

- to create/make available general corpora (~mega corpora)
  - Leeds Internet corpora (http://corpus.leeds.ac.uk/internet.html)
    - 12 languages (incl. Portuguese and English CC)
    - Lemmatised and pos-tagged
    - Indexed with the CWB and searchable online (CQP)
  - Fletcher's WaC (http://webascorpus.org/)
    - ~500M words of English (AU, CA, GB, IE, NZ, US)
    - Offers filtering option (OALD, BNC)
    - Phrase database linked to Web concordances
- to give users tools to build their own
  - Crucial for translators!

Fletcher's Web Concordancer

http://webascorpus.org/

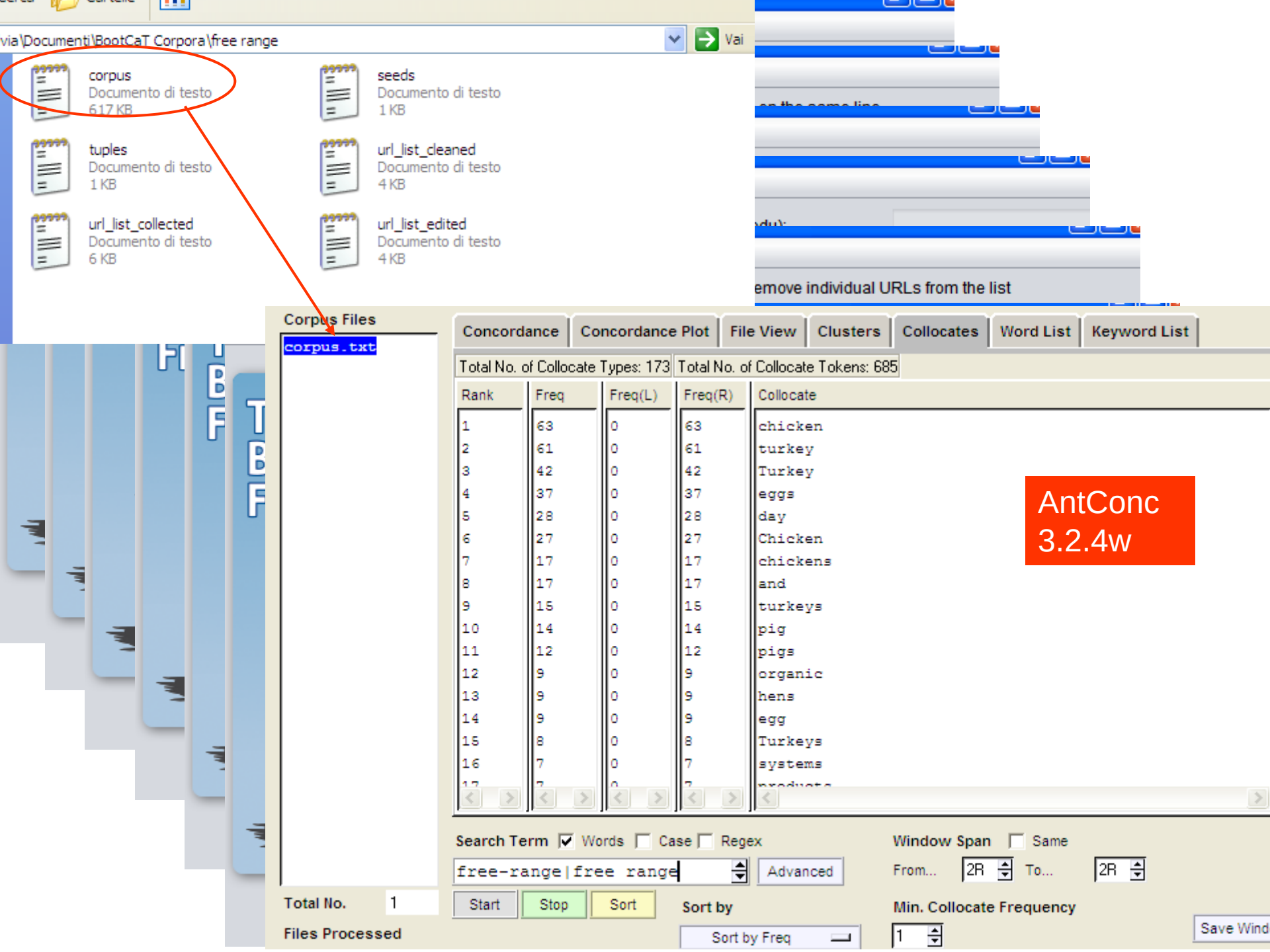



C:\Documents and Settings\silvia\Desktop\free-rang...
File Modifica Visualizza Preferiti Strumenti Aiuto
Aggiungi Estrai Verifica Copia Sposta Cancella Proprietà
C:\Documents and Settings\silvia\Desktop\free-range_INCL(organic)_EXCL(eg
Nome
Oggetti selezionati: 1
TextPad - C:\Documents and Settings\silvia\Impostazioni locali\Temp\7zO120.tmp\en980620499354562....
File Edit Search View Tools Macros Configure Window Help
en980620499354562.txt
http://www.adoptaturke
24.3 Adopt-A-Turkey
Factory Farming
Free Range/Organic
Investigations
History of Thanksgivi
Free Range/Organ
The Truth Be nd Free-
As public aw reness o
range," "orga or
the right place, hese
labels are deceptive
Most people don't envi
or "organic" turkey.
at "free-range" farms
can make eating and wa
As on factory farms,
manipulated to grow a
many health problems
AntConc 3.2.1w (Windows) 2007
File Global Settings Tool Preferences About
Corpus Files
en980620499354
en026061246334
en052698637305
en060425284454
en060657205610
en539750700878
en541670551323
en546644122993
en549908304691
en552137381252
en562827564877
en565421722304
en594021912346
en603621161960
en608328447890
en639488427229
en639913635677
en642602284417
en677322796041
en685165406259
en687978610492
en692423902167
en692771791571
en694502665750
en704385382418
en705210021644
Total No. 42
Files Processed
Reset
Concordance Concordance Plot File View Clusters Collocates Word List Ke
Total No. of Collocate Types: 49 Total No. of Collocate Tokens: 125
Rank Freq Freq(L) Freq(R) Collocate
1 17 0 17 turkey
2 12 0 12 chicken
3 11 0 11 and
4 9 0 9 turkeys
5 8 0 8 organic
6 5 0 5 or
7 4 0 4 meat
8 4 0 4 farms
9 4 0 4 environment
10 3 0 3 systems
11 3 0 3 chickens
12 2 0 2 standards
13 2 0 2 products
14 2 0 2 pork
15 2 0 2 Petition
Search Term Words Case Regex
free range|free-range
Advanced
Start Stop Sort
Sort by
Sort by Freq
Window Span Same
From... 2R To...
Min. Collocate Frequency
1

BootCaT
http://bootcat.sslmit.unibo.it/

- Basic pipeline
  - Select initial seeds (terms, keywords)
  - Query SE for random seed combinations
  - Retrieve pages and format as text (corpus)
  - Extract new seeds via corpus comparison
  - Iterate

  - Designed for translation students
  - Also used for reference corpus building
    - Leeds Internet Corpora


via\Documenti\BootCaT Corpora\free range
Vai
corpus
Documento di testo
617 KB
seeds
Documento di testo
1 KB
tuples
Documento di testo
1 KB
url_list_cleaned
Documento di testo
4 KB
url_list_collected
Documento di testo
6 KB
url_list_edited
Documento di testo
4 KB
emove individual URLs from the list
Corpus Files
corpus.txt
Concordance
Concordance Plot
File View
Clusters
Collocates
Word List
Keyword List
Total No. of Collocate Types: 173
Total No. of Collocate Tokens: 685
Rank Freq Freq(L) Freq(R) Collocate
1 63 0 63 chicken
2 61 0 61 turkey
3 42 0 42 Turkey
4 37 0 37 eggs
5 28 0 28 day
6 27 0 27 Chicken
7 17 0 17 chickens
8 17 0 17 and
9 15 0 15 turkeys
10 14 0 14 pig
11 12 0 12 pigs
12 9 0 9 organic
13 9 0 9 hens
14 9 0 9 egg
15 8 0 8 Turkeys
16 7 0 7 systems
AntConc
3.2.4w
Search Term
Words
Case
Regex
free-range|free range
Advanced
Window Span
Same
From...
2R
To...
2R
Total No.
1
Start
Stop
Sort
Sort by
Sort by Freq
Min. Collocate Frequency
1
Files Processed
Save Wind

Corpora of the Web era

Why a Wikipedia corpus?

- Opportunity
  - lots of text, multilingual coverage, convenient format (xml, Wikipedia
- Practical/di
  - translator
  - web form
- Theoretica
  - linked Wi
    - indepe
    - ST in l
    - ST and a heavily edited TT
  - how does our traditional notion of translation relate to collaborative web-based multilingual text production?

1. Our corpus typologies pre-date the Web

2. Our objects of study often ignore new developments

Turning Wikipedia into a comparallel corpus

Corpus structure (IT/EN but replicable)

1. Two large, independent **monolingual corpora**
   - all of Wikipedia IT + all of Wikipedia EN
2. A smaller **comparable corpus**
   - all entries available both in IT and EN
3. A (much) smaller set of **parallel segments**
   - Translation Memory style
   - 1:1 matches only
   - linked to whole texts in the comparable corpus providing browsable co-texts

What we aim for

- A corpus
  - consisting of all explicitly linked bi-articles (in Italian and English)
  - allowing browsing of article pairs and
  - on-the-fly building of thematic subcorpora
- Guidelines, pipelines, tools
  - for other language pairs
  - for future dumps (“monitor” Wikipedia corpus?)

Our starting point

```
<page>
  <title>Belo Horizonte</title>
  <id>82198</id>
  <revision>
```

Belo Horizonte

From Wikipedia, the free encyclopedia

Coordinates: 19°55′8.88″S 43°56′19.2″W

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(November 2010)*

Belo Horizonte (Portuguese pronunciation: [ˌbɛloriˈzõtʃi],[1] *Beautiful Horizon*) is the capital of and largest city in the state of Minas Gerais, located in the southeastern region of Brazil. It is the third largest metropolitan area in the country. Belo Horizonte (also known as "Belô", "Beagá", or "BH") has a population of over 2.4 million, or almost 5.4 million in the official Metropolitan Area.

The region was first settled in the early 18th century, but the city as it is known today was planned and constructed in the 1890s, in order to replace Ouro Preto as the capital of Minas Gerais. The city features a mixture of contemporary and classical buildings, and is home to several modern Brazilian architectural icons, most notably the Pampulha Complex. In planning the city, Aarão Reis and Francisco Bicalho sought inspiration in the urban planning of Washington, D.C.[2] The city has employed notable programs in urban revitalization and food security, for which it has been awarded international accolades.

The city is built on several hills and is completely surrounded by mountains.[3] There are several large parks in the immediate surroundings of Belo Horizonte. The "Parque das Mangabeiras", located six kilometres south-east from the city centre in the hills of the Serra do Curral, affords a view over the city. It has an area of 2.35 km^2 (580 acres), of which 0.9 km^2 (220 acres) is native forest. The "Mata do Jambeiro" nature reserve extends over 912 hectares (2,250 acres), with vegetation typical of the Atlantic forest. More than one hundred species of bird inhabit the reserve, as well as ten different species of mammals.

Contents [hide]

1 Geography
- 1.1 Surrounding cities and metropolitan area
- 1.2 Geology and geomorphology
- 1.3 Hydrology
- 1.4 Climate

2 History

3 Demographics
- 3.1 Religion

4 Economy

5 Education
- 5.1 Educational institutions

Belo Horizonte

— Municipality —

The Municipality of Belo Horizonte

From the top, left to right: view of the city with the Curral Mountains in the background, seen from the downtown at night, September 7 Square, Rui Barbosa Square, Church of Saint Francis of Assisi, and Administrative City of Minas Gerais

```
|leader_title = [[Mayor]]
|leader_name = [[Marcio Lacerda]] ([[Brazilian Socialist Party|PSB]])
|established_title = Founded
|established_date = 1701
|established_title2 = [[Municipal corporation|Incorporated]] (as city)
```

In practice…

1. Download Wikipedia dumps (18/03/10)
2. Extract XML files
3. Keep
   - references to entries in other languages
   - categories
4. Clean texts of markup and boilerplate (using *WikiExtractor*)

In practice (cont'd)

5. Only keep articles with EN<=>IT link
6. Metadata:
   - text id (= article's title in lang. A)
   - text target (= matching article's title in lang. B)
   - categories
7. POS-tag and lemmatise (TreeTagger)
8. Index with the Corpus WorkBench
   - *Comparapedia* EN
   - *Comparapedia* IT

Aside: Categories from Wikipedia to Comparapedia

- Original Wikipedia categories
  - inserted by humans
  - richer in EN than in IT
  - some work done in NLP to give them structure
    - YAGO; DBPEDIA; WIKINET
- Our “quick and dirty” approach
  - lowercase
  - keep only lexical words => keywords
  - sort in alphabetical order
  - migrate EN keywords to matching IT article

From categories to keywords

EN

<text_keywords 1942 20th-century 21st-century academics academy adams albans albert alumni applied arts astronomers astronomical births british caius calculating cambridge college commanders companions copley cosmologists department disease einstein empire english fellows former freedom gold gonville hall hertfordshire honorary honour laureates living lucasian mathematics medal members motor national neuron order oxford people philosophers physicists physics pontifical presidential prize prodigies professors pupils recipients religious royal school science sciences skeptics society st theoretical trinity university wolf writers>

WIKIPEDIA
The Free Encyclopedia

Article | Discussion | Read | Edit | View history | Search

Stephen Hawking

From Wikipedia, the free encyclopedia

Main page
Contents
Featured content
Current events
Random article
Interaction
About Wikipedia

Categories: 1942 prodigies | Comm Society of Arts | F Applied Mathema Oxford | People fr Religious skeptics

WIKIPEDIA
L'enciclopedia libera

Stephen Hawking

Da Wikipedia, l'enciclopedia libera.

Pagina principale
Ultime modifiche
Una voce a caso
Vetrina
Aiuto
Comunità
Portale Comunità

Stephen William Hawking (Oxford, 8 gennaio 1942) è un matematico e astrofisico britannico, fra i più importanti e conosciuti del mondo. Pur essendo condannato all'immobilità dall'atrofia muscolare progressiva (e non come si pensava, dalla sclerosi laterale amiotrofica), ha occupato la cattedra lucasiana di matematica[1] all'Università di Cambridge (la stessa che fu di Isaac Newton) per trent'anni, dal 1979 al 30 settembre 2009[2]. È membro della Royal Society e del Mensa. Noto soprattutto per i suoi studi sui buchi neri, è oggi uno fra i cosmologi più autorevoli.

Nel 1974 ha dimostrato che, dal punto di vista termodinamico, i buchi neri sono corpi neri e obbediscono alle leggi della termodinamica: posseggono una temperatura e un'entropia definite dal loro campo gravitazionale e dalla loro superficie. Quindi i buchi neri dovrebbero irradiare particelle con una temperatura e un'entropia definite. Questa

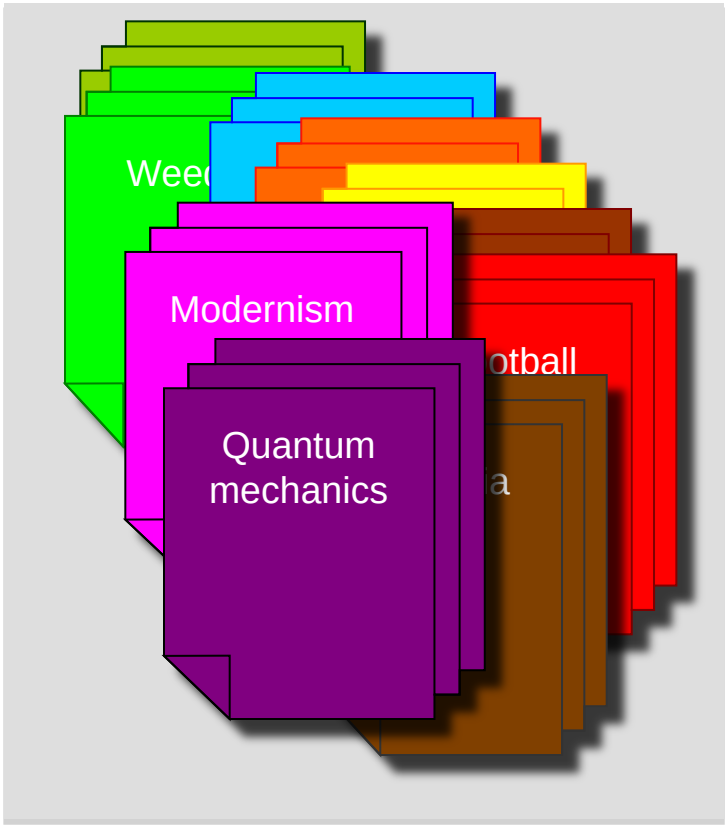
Categorie: Matematici britannici | Astrofisici britannici | Nati nel 1942 | Nati l'8 gennaio | Bambini prodigio | Fisici teorici | Saggisti britannici | Divulgatori scientifici britannici | Membri della Royal Society | [*altre*]

Quick corpus facts

	Comparapedia EN	Comparapedia IT
Articles	426,273	426,057
Tokens	274,344,165	139,975,783

- Corpus structure – pseudo xml
  - <text id="*title*" target="*target_title*" keywords="*kw1 kw2 kwn*">
  - the actual text in vertical format

    (positional attributes: word, pos, lemma)


Comparapedia EN
Modernism
Quantum mechanics
Comparapedia IT
Modernismo
Meccanica quantistica
Path integral formulation
Integrale sui cammini
Comparable subcorpora
Matching text pairs

In practice: using Comparapedia

Comparapedia EN

<text
id="Path integral formulation"
target="Integrale sui cammini"
keywords="concepts field fundamental mechanics physics quantum statistical theory">

The path integral formulation of quantum mechanics is a description of quantum theory which generalizes the action principle of classical mechanics. […]

Comparapedia IT

<text
id="Integrale sui cammini"
target="Path integral formulation"
keywords="concepts field fundamental mechanics physics quantum statistical theory">

L'integrale sui cammini (o "path integral") rappresenta una formulazione della meccanica quantistica che descrive la teoria quantistica generalizzando il principio di azione della meccanica classica. […]

?

How *parallelizable* is Comparapedia?

- Naïve attempt at aligning 3 ita-eng Comparapedia article pairs

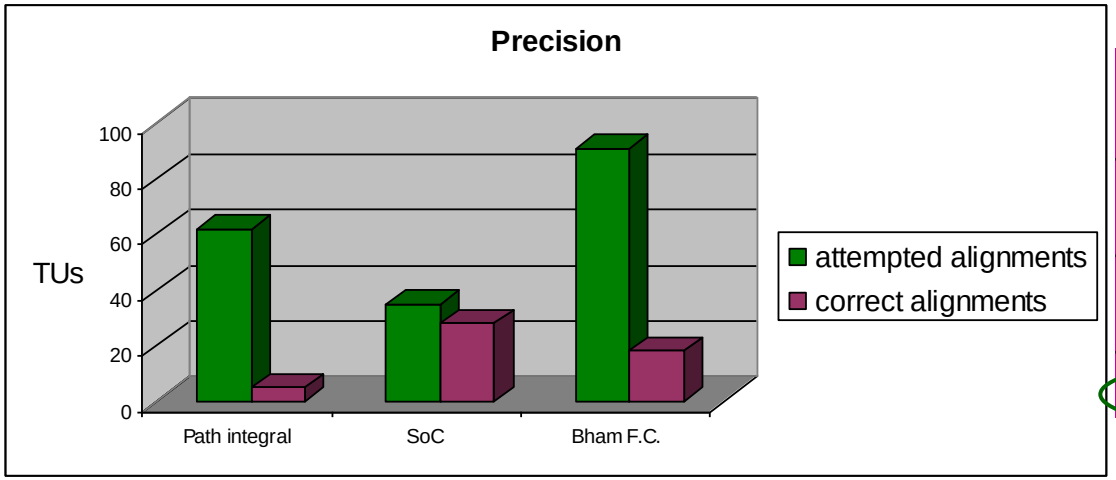
  1. scientific (quantum mechanics: *Path integral formulation*)
  2. technical (computing: *System-on-a-Chip, SoC*)
  3. cultural (Football: *Birmingham City F.C.*)

Translated?
Adapted?
Independent?

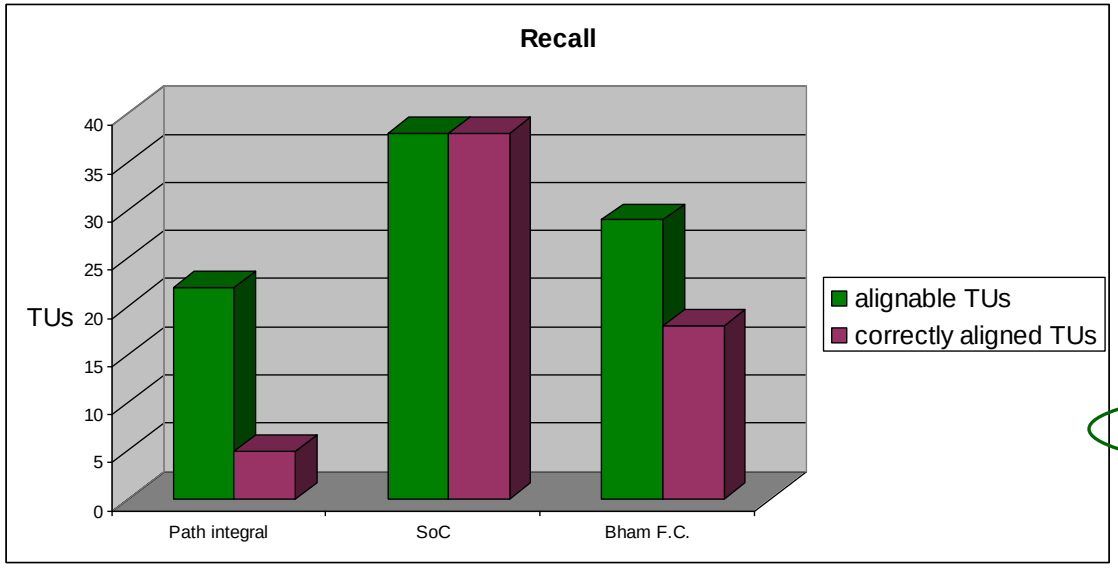
Method

1. Attempt alignment of text pairs
   - no structural information (text only)
   - Hunalign (Varga et al. 2005)
     - No dictionaries
2. Measure precision
   - translation units (TUs) correctly aligned/all TUs
3. Measure recall
   - TUs correctly aligned/manually alignable TUs

Results



	Path integral	SoC	Bham F.C.
attempted alignments	62	35	91
correct alignments	5	28	18
precision	8%	80%	19.7%



= = ≠

	Path integral	SoC	Bham F.C.
alignable TUs (1s)	22	38	29
correctly aligned TUs (1s)	5	38	18
recall	22.7%	100%	62%

Future work

English Segment	Italian Segment	Aligner confidence score
Fabrication	Realizzazione	9
Structure	Struttura di un SoC	9
Design flow	Progetto di un SoC	7
Peripherals including counter-timers, real-time timers and power-on reset generators.	Periferiche come contatori, orologi e altro	2.65625
Memory blocks including a selection of ROM, RAM, EEPROM and flash.	Un modulo di memoria contenente uno o più blocchi di tipo ROM, RAM, EEPROM o Memoria flash.	2
External interfaces including industry standards such as USB, FireWire, Ethernet, USART, SPI.	Connettori per interfacce standard come USB, FireWire, Ethernet, USART, SPI	1

Wikipedia as a multilingual *comparallel* corpus?

- hundreds of virtual, user-definable specialised comparable corpora
- thousands of matching bilingual text pairs
- potential for automatic alignment of matching text segments (translation units) from the corpus

- Implications
  - Practical: A more efficient resource for language professionals (optimising searches/matching)
  - Methodological: a new hybrid type of corpus reflecting the nature of collaborative writing/translation on the web
  - Descriptive/Theoretical: in what domains do different linguacultures distance themselves the most from each other?

Summing up

Web as Corpus prospects

- We are still very far from the dream of a search engine for linguists but
- Thanks to the Web, corpora are growing in size and variety and are more and more often available in the public domain

Web as Corpus prospects

- The different Web corpora we discussed offer several advantages
  - Up-to-dateness
  - Size (mega corpora) and task tuning (specialized corpora)
  - Convenience
    - Cost
    - Ease of collection
    - Under-resourced languages
  - Web-specific genres
  - Reference purposes
  - **Fewer copyright issues, more chances of sharing**

Web as Corpus prospects

- “Traditional” corpora (high-quality, time-consuming, costly) will continue to be built, but only if no suitable alternative (faster, cheaper) is available, i.e.
  - When selection based on external criteria is required
  - When register/genre has to be tightly controlled
  - For pre- or non-Web genres

The future of (web as) corpus linguistics *(a personal view)*

- Corpora of the future built through distributed community effort with open source tools and shared methods
  - possibility to improve corpus quality continually
  - [illegible]
  - [illegible]hose
- Fe[illegible] de[illegible]
  - [illegible]nes,
  - [illegible]
- …r[illegible]ge co[illegible]
  - will be able to develop and share their own corpora
  - ➢ we'll have more corpora in languages other than English

"Since it is theoretically suspect to aim at achieving a perfectly `balanced' corpus, it is practical to adopt a method of successive approximations. First, the corpus builder attempts to create a representative corpus. Then this corpus is used and analysed and its strengths and weaknesses identified and reported. In the light of this experience and feedback the corpus is enhanced by the addition or deletion of material and the cycle is repeated continually." (Atkins et al 1991)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

thank you

silvia.bernardini@unibo.it

ELC - 2011 X Encontro de Linguística de Corpus

EBRALC - 2011 - V Escola Brasileira de Linguística Computacional

References

- Too many…
- Ask me if you are interested in any in particular

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

thank you

silvia.bernardini@unibo.it

ELC - 2011 X Encontro de Linguística de Corpus

EBRALC - 2011 - V Escola Brasileira de Linguística Computacional