Processing of Very Large Text Collections

Miloš Jakubíček

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

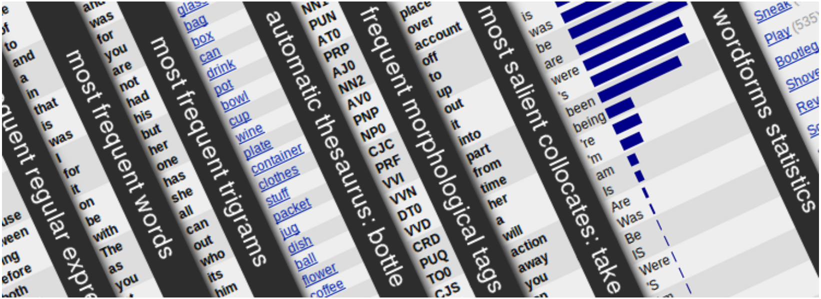
jak@fi.muni.cz

Workshop of Natural Language Processing Centre &
Seznam.cz, a.s.
May 16, 2014

Why to process natural language texts?

- **lots** of information, growing every day (web)
- need for **fast** and continuous knowledge mining
- **no time** for human intervention
- **large** data make statistical processing possible
- **real** data instead of false assumptions

Information in Text

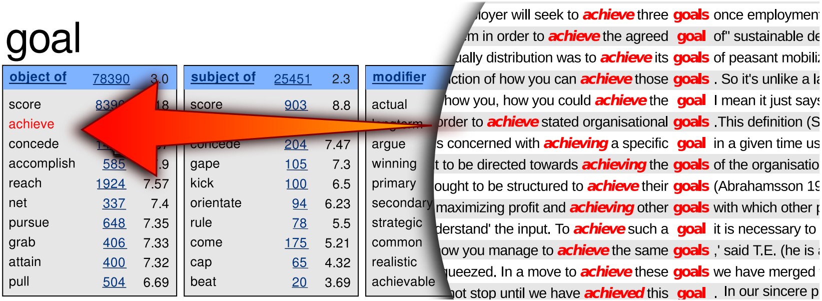


quent regular expre
most frequent words
most frequent trigrams
automatic thesaurus: bottle
frequent morphological tags
most salient collocates: take
wordforms statistics
e of to and a in that is was I for it on be with The as you
use ween ing efore both
and was for you are not had his but her one has she all can out who its him
glass bag box can drink pot bowl cup wine plate container clothes stuff packet jug dish ball flower coffee
NN1 PUN AT0 PRP AJ0 NN2 AV0 PNP NP0 CJC PRF VVI VVN DT0 VVD CRD PUQ TO0 CJS
place over account off to up out it into part from time her a will action away you
is was be are were 's been being 're 'm am Is Are Was Be IS Were 'S
Sneak Play (535) Bootleg Shove Rev So

Text collection = a text corpus

- text collection: usually referred to as **text corpus**
- **humanities** → corpus linguistics, language learning
- **computer science** → effective design of specialized database management systems
- **applications** → usage of *any text* as information source

Text Corpora as Information Source



goal

object of	78390	3.0
score	8390	[illegible]
achieve	[illegible]	[illegible]
concede	[illegible]	[illegible]
accomplish	585	[illegible]
reach	1924	7.57
net	337	7.4
pursue	648	7.35
grab	406	7.33
attain	400	7.32
pull	504	6.69

subject of	25451	2.3
score	903	8.8
[illegible]	[illegible]	[illegible]
concede	204	7.47
gape	105	7.3
kick	100	6.5
orientate	94	6.23
rule	78	5.5
come	175	5.21
cap	65	4.32
beat	20	3.69

modifier
actual
longterm
argue
winning
primary
secondary
strategic
common
realistic
achievable

loyer will seek to **achieve** three **goals** once employmen
m in order to **achieve** the agreed **goal** of" sustainable de
ually distribution was to **achieve** its **goals** of peasant mobili
ction of how you can **achieve** those **goals** . So it's unlike a la
how you, how you could **achieve** the **goal** I mean it just says
order to **achieve** stated organisational **goals** .This definition (S
s concerned with **achieving** a specific **goal** in a given time us
t to be directed towards **achieving** the **goals** of the organisatio
ought to be structured to **achieve** their **goals** (Abrahamsson 19
maximizing profit and **achieving** other **goals** with which other p
derstand' the input. To **achieve** such a **goal** it is necessary to
ow you manage to **achieve** the same **goals** ,' said T.E. (he is a
ueezed. In a move to **achieve** these **goals** we have merged
ot stop until we have **achieved** this **goal** . In our sincere p

So what is a corpus?

text corpus	=	lots of **text data**	+	lots of **metadata**	+	fast and complex **search/retrieval**
⇓		⇓		⇓		⇓
database	=	**gather** text data: web/proprietary	+	**apply** NLP tools to get metadata	+	effective **database management system** suitable for text

Corpora

- **text type**
  - *general language* (gather domain independent information: common sense knowledge, global statistics, information defaults)
  - *domain specific* (gather domain specific information: terminology, in-domain knowledge, contrast to common texts)
- **timeline**
  - *synchronic*: one time period / time span (→ what is up now?)
  - *diachronic*: different time periods / time spans (→ what are the trends?)
- **language, written/spoken, metadata annotation type, ...**

Corpora

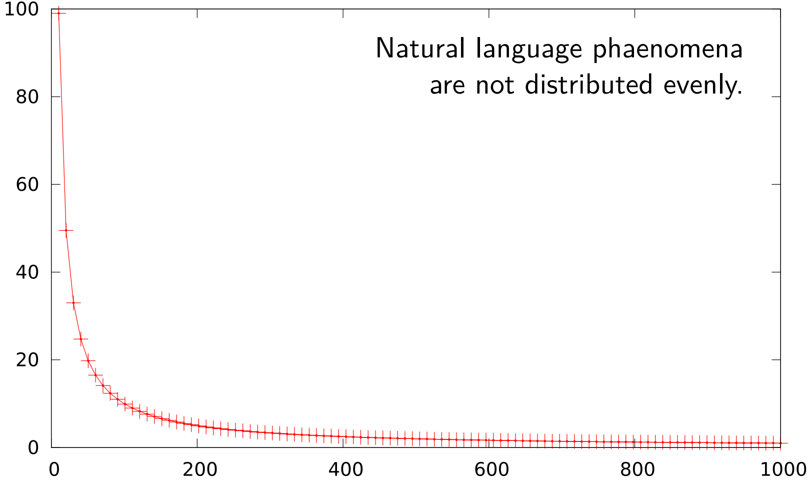
So is there any property one should aim at for all corpora?

Corpora

So is there any property one should aim at for all corpora?

Yes – the size. The bigger, the better.

Why does size matter so much?



Natural language phaenomena
are not distributed evenly.

Corpora now

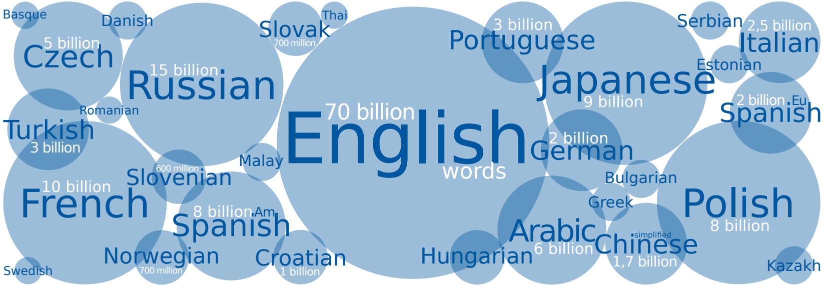
Corpora at NLP Centre:

- **LARGE**: billions ($\sim 10^{10}$) of words
- **COMPLEX**: muti-level multi-value annotation, wide range of languages

Corpora now

Corpora at NLP Centre:

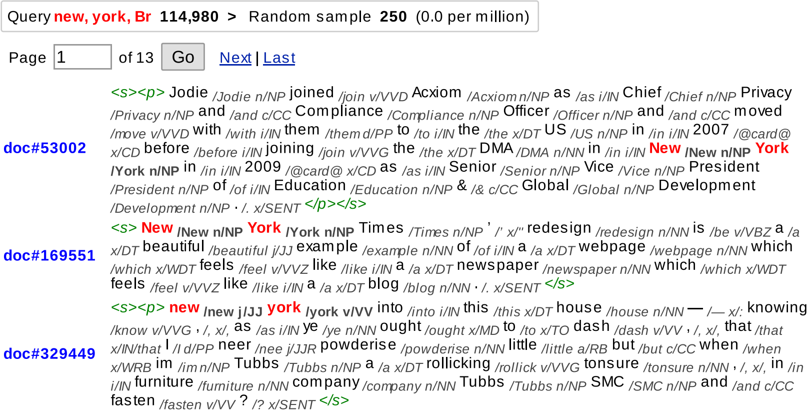
- **LARGE**: billions (~10^{10}) of words



Corpora now

Corpora at NLP Centre:

- **COMPLEX**: muti-level multi-value annotation, wide range of languages



Query **new, york, Br** **114,980** **>** Random sample **250** (0.0 per million)

Page 1 of 13 Go Next | Last

doc#53002 <s><p> Jodie /Jodie n/NP joined /join v/VVD Acxiom /Acxiom n/NP as /as i/IN Chief /Chief n/NP Privacy /Privacy n/NP and /and c/CC Compliance /Compliance n/NP Officer /Officer n/NP and /and c/CC moved /move v/VVD with /with i/IN them /them d/PP to /to i/IN the /the x/DT US /US n/NP in /in i/IN 2007 /@card@ x/CD before /before i/IN joining /join v/VVG the /the x/DT DMA /DMA n/NN in /in i/IN **New /New n/NP York /York n/NP** in /in i/IN 2009 /@card@ x/CD as /as i/IN Senior /Senior n/NP Vice /Vice n/NP President /President n/NP of /of i/IN Education /Education n/NP & /& c/CC Global /Global n/NP Development /Development n/NP . /. x/SENT </p></s>

doc#169551 <s> **New /New n/NP York /York n/NP** Times /Times n/NP ' /' x/'' redesign /redesign n/NN is /be v/VBZ a /a x/DT beautiful /beautiful j/JJ example /example n/NN of /of i/IN a /a x/DT webpage /webpage n/NN which /which x/WDT feels /feel v/VVZ like /like i/IN a /a x/DT newspaper /newspaper n/NN which /which x/WDT feels /feel v/VVZ like /like i/IN a /a x/DT blog /blog n/NN . /. x/SENT </s>

doc#329449 <s><p> **new /new j/JJ york /york v/VV** into /into i/IN this /this x/DT house /house n/NN — /— x/: knowing /know v/VVG , /, x/, as /as i/IN ye /ye n/NN ought /ought x/MD to /to x/TO dash /dash v/VV , /, x/, that /that x/IN/that I /I d/PP neer /nee j/JJR powderise /powderise n/NN little /little a/RB but /but c/CC when /when x/WRB im /im n/NP Tubbs /Tubbs n/NP a /a x/DT rollicking /rollick v/VVG tonsure /tonsure n/NN , /, x/, in /in i/IN furniture /furniture n/NN company /company n/NN Tubbs /Tubbs n/NP SMC /SMC n/NP and /and c/CC fasten /fasten v/VV ? /? x/SENT </s>

Corpora now

A big need for search/retrieval that is:

- **INTELLIGENT**: complex searching involving large amounts of metadata
- **VERY FAST**: parallel and distributed processing
- **ACCESSIBLE**: interfaces for automatic processing via third-party tools

Applications

- **information systems** (going beyond fulltext search)
- **information analytics** (opinion mining, marketing assessment)
- **intelligent text processing** (predictive and adaptive writing, correction tools, effective writing in mobile devices)
- **computer lexicography** (better dictionaries, larger dictionaries)
- **machine translation** (parallel corpora)
- **statistics** for enhancing NLP tools

What can we offer?

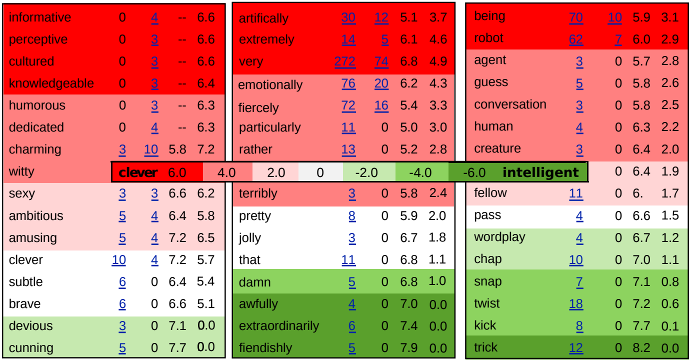
Ready-made tools for corpus building, management and effective search:

- **Building:** from own data/from the web, crawling, cleaning, deduplication
- **Management:** effective indexing in special DBMS
- **Search:** very fast evaluation of complex queries, keywords extraction, extraction of semantically related words, word sketches

Most of the tools are part of Sketch Engine, a product developed in collaboration with Lexical Computing Ltd.

Demo: Sketch Engine

compare and contrast words visually



informative	0	4	--	6.6
perceptive	0	3	--	6.6
cultured	0	3	--	6.6
knowledgeable	0	3	--	6.4
humorous	0	3	--	6.3
dedicated	0	4	--	6.3
charming	3	10	5.8	7.2
witty				
sexy	3	3	6.6	6.2
ambitious	5	4	6.4	5.8
amusing	5	4	7.2	6.5
clever	10	4	7.2	5.7
subtle	6	0	6.4	5.4
brave	6	0	6.6	5.1
devious	3	0	7.1	0.0
cunning	5	0	7.7	0.0

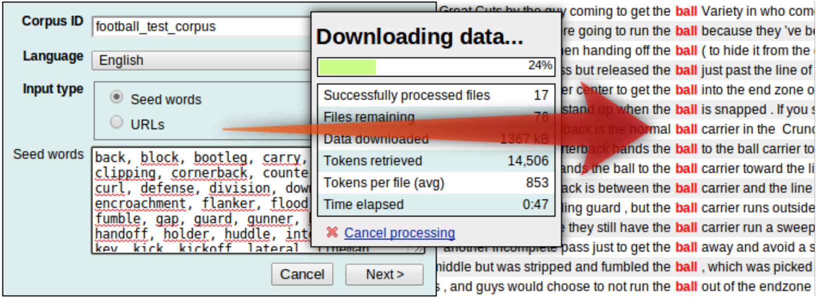
artifically	30	12	5.1	3.7
extremely	14	5	6.1	4.6
very	272	74	6.8	4.9
emotionally	76	20	6.2	4.3
fiercely	72	16	5.4	3.3
particularly	11	0	5.0	3.0
rather	13	0	5.2	2.8
terribly	3	0	5.8	2.4
pretty	8	0	5.9	2.0
jolly	3	0	6.7	1.8
that	11	0	6.8	1.1
damn	5	0	6.8	1.0
awfully	4	0	7.0	0.0
extraordinarily	6	0	7.4	0.0
fiendishly	5	0	7.9	0.0

being	70	10	5.9	3.1
robot	62	7	6.0	2.9
agent	3	0	5.7	2.8
guess	5	0	5.8	2.6
conversation	3	0	5.8	2.5
human	4	0	6.3	2.2
creature	3	0	6.4	2.0
		0	6.4	1.9
fellow	11	0	6.	1.7
pass	4	0	6.6	1.5
wordplay	4	0	6.7	1.2
chap	10	0	7.0	1.1
snap	7	0	7.1	0.8
twist	18	0	7.2	0.6
kick	8	0	7.7	0.1
trick	12	0	8.2	0.0

clever 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 **intelligent**

Demo: Sketch Engine

build specialised corpora instantly from the Web



Corpus ID: football_test_corpus

Language: English

Input type: ● Seed words ○ URLs

Seed words:

```
back, block, bootleg, carry,
clipping, cornerback, counte
curl, defense, division, dow
encroachment, flanker, flood
fumble, gap, guard, gunner, 
handoff, holder, huddle, int
key, kick, kickoff, lateral
```

Downloading data...

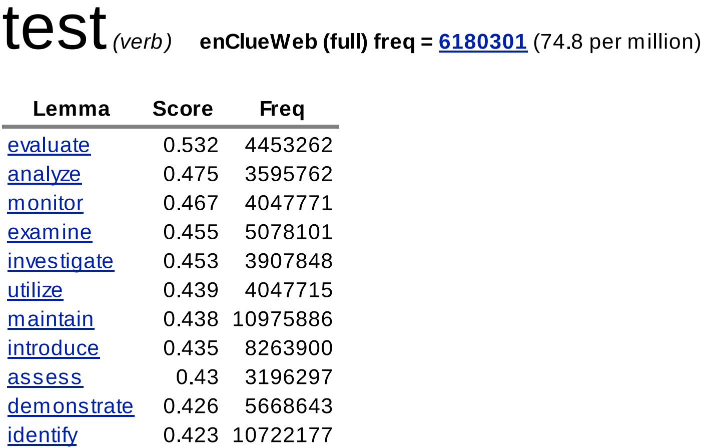
24%

Successfully processed files	17
Files remaining	76
Data downloaded	1367 kB
Tokens retrieved	14,506
Tokens per file (avg)	853
Time elapsed	0:47

```
Great Cuts by the guy coming to get the ball Variety in who com
                  re going to run the ball because they 've b
                 en handing off the ball ( to hide it from the
                ss but released the ball just past the line of
               er center to get the ball into the end zone o
              stand up when the ball is snapped . If you s
         back is the normal ball carrier in the  Crun
          terback hands the ball to the ball carrier to
           ands the ball to the ball carrier toward the li
        ack is between the ball carrier and the line
          ling guard , but the ball carrier runs outside
         they still have the ball carrier run a sweep
another incomplete pass just to get the ball away and avoid a s
iddle but was stripped and fumbled the ball , which was picked
s , and guys would choose to not run the ball out of the endzone
```

Demo: Sketch Engine

thesaurus



test *(verb)* **enClueWeb (full) freq = 6180301** (74.8 per million)

Lemma	Score	Freq
evaluate	0.532	4453262
analyze	0.475	3595762
monitor	0.467	4047771
examine	0.455	5078101
investigate	0.453	3907848
utilize	0.439	4047715
maintain	0.438	10975886
introduce	0.435	8263900
assess	0.43	3196297
demonstrate	0.426	5668643
identify	0.423	10722177

Conclusions

- Text corpora represent a **valuable information source** useful for many practical applications
- Corpora as text databases require **special solutions** that are fast and powerful
- There are number of **tools developed in the NLP Centre** for corpus building, management and efficient search