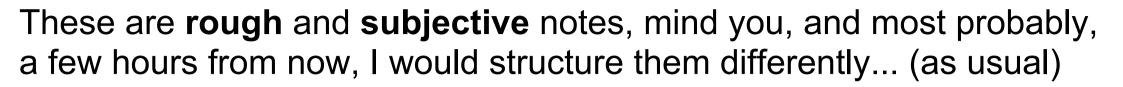
# Challenges in the Management of Large Corpora (CMLC-2) Closing remarks

Piotr Bański

Institut für Deutsche Sprache Mannheim banski@ids-mannheim.de



## **Points**

- Big data but is it big?
- Costs of data
- Costs of tools/storage/access
- User-friendliness

## ... but is it big?

big data -- a given, but of a sort (recall the ratio of storage to volume, from Marc's presentation)

maximizing the text is one thing:

 what follows causally (in a way...) is explosion of the size of the accompanying annotation layers

P. Bański: CMLC-2@LREC-2014

# **DATA** is costly

curation (AAC, IDS, other data centres mentioned today)

startling figures from DeReKo: the expenses for the acquisition and curation: fiction word = 25,000 \* 1 newspaper word

maximize the use and re-use of data:

- virtualization: Manatee, KorAP
- keep it in the raw form and use multiple annotations
- maximize the re-use for comparable and parallel uses (SketchEngine/Manatee, KorAP in a future project)
- keep the format **standardized** (or easy to transduce)
- curated data is costly, so... "just harvest"?

P. Bański: CMLC-2@LREC-2014

# Data is costly, so just harvest it?

Harvest the data from the Web as mentioned by Dirk Goldhahn, Steffen Remus, Uwe Quasthoff and Chris Biemann – there are pros and cons

- costs of cleaning the data are essential
- scrambled nature, sparse metadata -- "costs" for research (some research paths are closed)

## Cooostly, so...

#### maximize the re-use of the data

- virtualization / multiplying raw-text re-use: Manatee (Adam Kilgarriff, Pavel Rychlý and Miloš Jakubíček), KorAP (Marc Kupietz, Harald Lüngen, Piotr Bański and Cyril Belica)
- multiplying annotation layers for single documents

#### don't move the data:

- too big for download
- too dangerous, from the legal point of view

one solution --> put the computation near the data: KorAP

deal with the size: MapReduce (Dirk Goldhahn+colleagues and Jordi Porta)

## **Tools**

While the storage isn't costly, tools can still be:

- use open-source infrastructure technologies (e.g. Hadoop, XML databases like BaseX, etc.)
- produce open-source tools
  - Pressure from the community essential
  - Legislation: national-scale projects required to deliver opensource products

#### tool-maintenance costs:

- keep them modular for flexibility
- use techniques for horizontal scalability / clustering
- use the GPU for local applications (John Vidler, Andrew Scott, Paul Rayson, John Mariani and Laurence Anthony)

# Legal issues

legal issues concerning the raw data, annotations, and access:

- legal advice / analysis -- more and more important
- unified access procedures (Shibboleth single login, etc., the role of CLARIN-like initiatives in facilitating access -- sensible licensing)

P. Bański: CMLC-2@LREC-2014

## User-friendliness is essential

- using query languages raises the issue of user-friendliness (who do you produce this for?)
  - a new query language for an Ordinary Working Linguist? (nahh, too many goals, target data too varied)
     --> Corpus Query Lingua Franca?
  - an **overlay** on an existing powerful query language? (Vincent Vandeghinste and Liesbeth Augustinus)
- efficient indexing (Jordi Porta), scalable retrieval: a lot of innovative effort will be needed in this respect still, and...

while data-volume-wise, we may be now where IT was 10 years ago (cf. Adam and colleagues),

it looks like we're in for some exciting developments still! :-)

... and, finally...

Dinnah?