SICHERUNG VON DATENQUALITÄT

Seminar vom 13. Mai 2015
Modul: Citizen Science in the Humanities: Methods and Trends

Dr Thomas Köntges
Citizen Science in the Humanities SS15

Digital Humanities
UNIVERSITÄT LEIPZIG

DATA QUALITY

IBM: HIGH QUALITY DATA IS

- Complete
- Accurate
- Available
- Timely

GS1: DATA QUALITY

- Complete
- Standards Based
- Consistent
- Accurate
- Time Stamped

COMPUTER SCIENCE

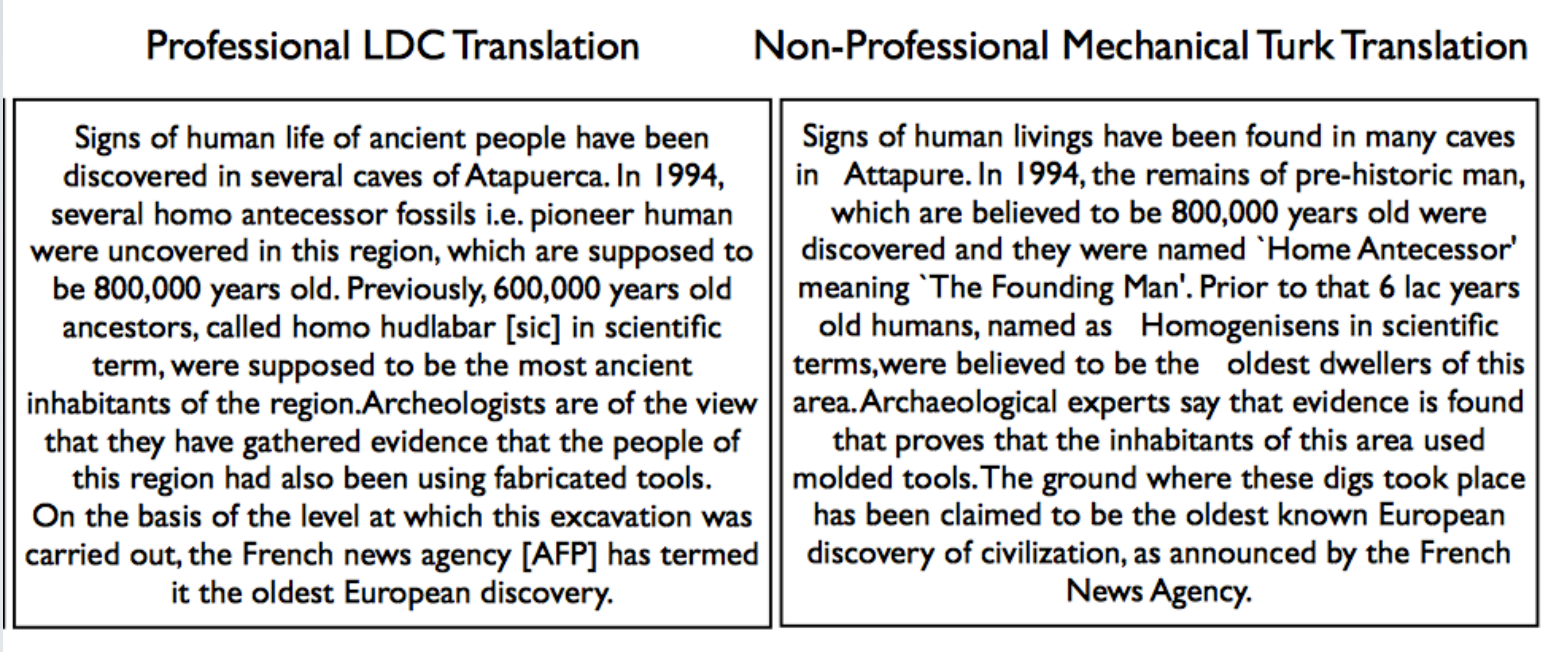
- Reliability
- Application Efficiency
- Validity = Readability for a machine
- Regulated by DQA processes

HUMANITIES

- Accurate = Solid argument
- Validity = Makes sense for a human
- Data is fuzzy

DATA QUALITY ASSURANCE

URDU TRANSLATION



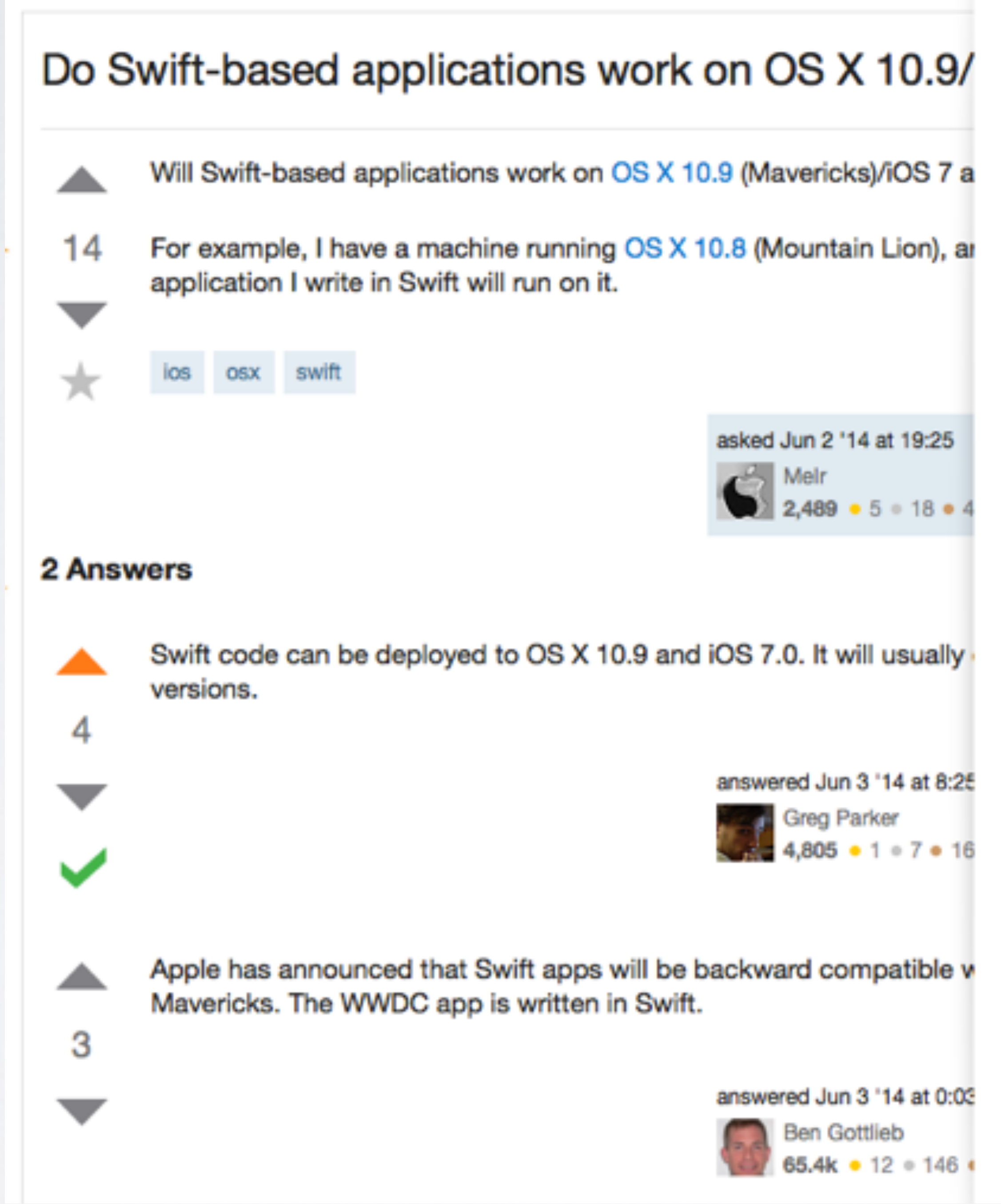
Professional LDC Translation

Signs of human life of ancient people have been discovered in several caves of Atapuerca. In 1994, several homo antecessor fossils i.e. pioneer human were uncovered in this region, which are supposed to be 800,000 years old. Previously, 600,000 years old ancestors, called homo hudlabar [sic] in scientific term, were supposed to be the most ancient inhabitants of the region.Archeologists are of the view that they have gathered evidence that the people of this region had also been using fabricated tools. On the basis of the level at which this excavation was carried out, the French news agency [AFP] has termed it the oldest European discovery.

Non-Professional Mechanical Turk Translation

Signs of human livings have been found in many caves in Attapure. In 1994, the remains of pre-historic man, which are believed to be 800,000 years old were discovered and they were named `Home Antecessor' meaning `The Founding Man'. Prior to that 6 lac years old humans, named as Homogenisens in scientific terms,were believed to be the oldest dwellers of this area. Archaeological experts say that evidence is found that proves that the inhabitants of this area used molded tools. The ground where these digs took place has been claimed to be the oldest known European discovery of civilization, as announced by the French News Agency.

P2P CONTROL?



Do Swift-based applications work on OS X 10.9/

14

Will Swift-based applications work on OS X 10.9 (Mavericks)/iOS 7 a

For example, I have a machine running OS X 10.8 (Mountain Lion), ar application I write in Swift will run on it.

ios osx swift

asked Jun 2 '14 at 19:25
Melr
2,489 ● 5 ● 18 ● 4

2 Answers

4

Swift code can be deployed to OS X 10.9 and iOS 7.0. It will usually versions.

answered Jun 3 '14 at 8:25
Greg Parker
4,805 ● 1 ● 7 ● 16

3

Apple has announced that Swift apps will be backward compatible w Mavericks. The WWDC app is written in Swift.

answered Jun 3 '14 at 0:03
Ben Gottlieb
65.4k ● 12 ● 146

AUTOMATIC USE OF FEATURES

- Good English sentences from bad English sentences
- Sentence length?!?
- Similarity to other translations
- Worker-level features
- Rank-level features

FINE TUNING

- Parameter tuning
- Worker confidence

NECESSITY OF EVALUATING TEST SETS

- Compare a number of different professional translations

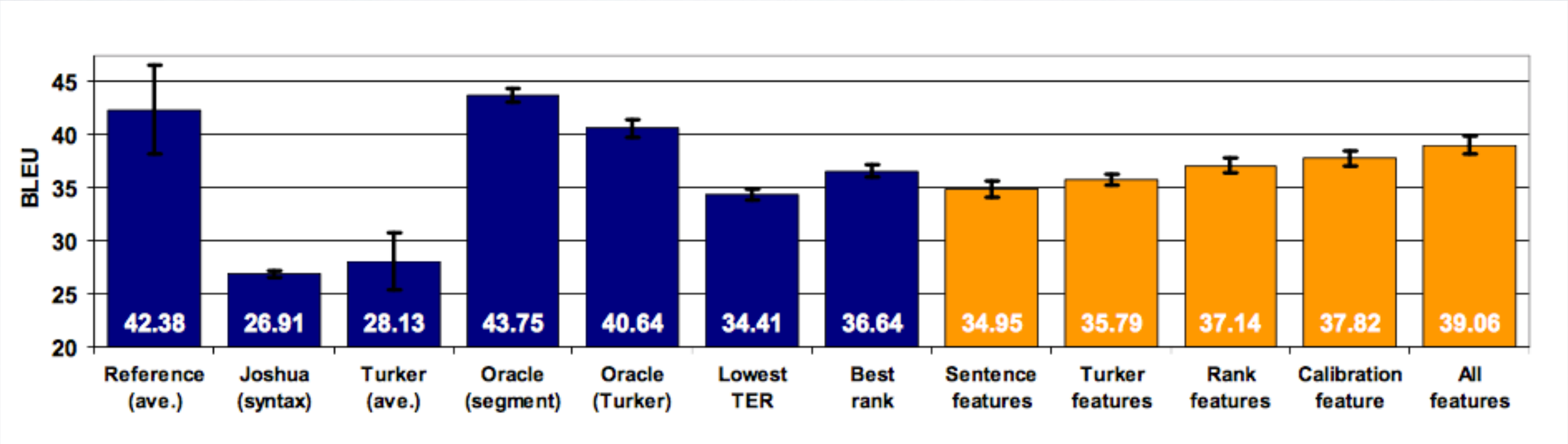
- Compare their BLEU scores (Bilingual Evaluation Understudy)

- Then compare non-professional to professional translations to evaluate

BLEU SCORES

- Professional translations had an average BLEU score of 42.38
- Non-professional translations had an average BLEU score of 26.91
- After Data-Quality-Assurances methods were applied. Non-professional translations could score as well as professional translations

BLEU SCORES


BLEU
45
40
35
30
25
20
42.38
26.91
28.13
43.75
40.64
34.41
36.64
34.95
35.79
37.14
37.82
39.06
Reference (ave.)
Joshua (syntax)
Turker (ave.)
Oracle (segment)
Oracle (Turker)
Lowest TER
Best rank
Sentence features
Turker features
Rank features
Calibration feature
All features

SCENARIO 1: HISTORIC DATA

- Contributors provide meta-data for historic pictures
- Which meta-data could be provided
- How do you assure data-quality?

SCENARIO II: TRANSCRIPTION

- Contributors transcribe text from digitised manuscript folia
- Which format can be used?
- What are the advantages / disadvantages?
- How do you assure data-quality

SCENARIO III: LINGUISTICS

- Contributors tree-bank sentences / determine syntactic structure of the sentences

- Which entry form is the most suitable?

- How do you assure data-quality?

SCENARIO IV: GEO-TAGGING

- Contributors find geo-references in texts and images and map them

- How do you assure data-quality?

QUESTIONS?

thomas.koentges@uni-leipzig.de