New corpora in SkE

Vít Suchomel

vit.suchomel@sketchengine.co.uk

5th Sketch Engine Workshop
Bolzano, July 14, 2014

Recently added corpora – TenTen family

language	size [words]
Danish	2.06 G
Dutch	2.58 G
Finnish	1.40 G
Lithuanian	778 M
Latvian	534 M
Swedish	3.40 G
Russian	111 M
Ukrainian	2.19 G

- Crawling plan: all European languages, all major world languages

Recently added corpora – Aranea family

language	size [words]
French	108 M
German	111 M
Russian	111 M

- thanks to Vlado Benko
  - visit his talk on Thursday in the afternoon

Recently added corpora – WaC

language	size [words]
Bosnian	248 M
Croatian	1.21 G
Serbian	477 M

- thanks to Nikola Ljubešić

Other recently added corpora

- King Saud University corpus of classical Arabic
- TED talks (English, parallel multi-language corpora soon)
- ScienceBlog
- BLARC British Legal corpus

Data quality issues

- Better cleaning of mistakes in corpora
- Spam, computer generated text
- Compatibility of user and preloaded corpora
- Do let us know if you spot something wrong

Summary

- We gather and process big web corpora
- Data quality is important
- Have an interesting corpus? You can share it with others!