NCN WORKSHOP ON FOUND DATA IN SPEECH AND LANGUAGE TECHNOLOGY

On 27–28 September 2017 the Nordic Clarin Network (NCN) is organizing the next in its series of thematic workshops at KTH Royal Institute of Technology.

Found data in speech and language technology

By "found data", we mean data that was not collected or recorded with speech or language technology in mind. This is a broad definition, which encompasses a wide range of linguistic materials, from broadcast materials to interviews; newspaper text to books, and YouTube videos to text-based chat conversations. Note that virtually all cultural heritage data is found data from our perspective. The concept is particularly pertinent for recordings of speech and spoken communication, as the difference, there, between lab recordings designed for speech analysis and recordings found in the wild is vast, to the extent that much speech technology that produces decent results on lab data (such as many automatic transcription methods) simply fails on found data. But similar issues arise with text, for example when using the web as a corpus resource. Note that although KTH, as organizers, are particularly interested in found *speech and multimodal data*, we welcome and strongly encourage contributions relating to text as well.

CONTRIBUTIONS

Suitable presentation/demo topics falling under this heading can be (but are not limited to):

- method development for the analysis of very large datasets whose properties are wholly or partially unknown;
- visualization of large and unwieldly data sets, in particular data sets for which inspection of the entire data set manually is simply not practicable, e.g. due to quantity;
- workflow organization/interface design assisting the analysis of such data sets, with a particular focus on efficient and affordable solutions, e.g. crowd sourcing/human computation and semi-automated methods;
- discussions on potential uses for found data, e.g. longitudinal historical studies (on longitudinal data) or correlations between language use and other factors; and
- improvement of language and speech technologies that can in turn be used to aid research on speech and text resources.

SCHEDULE

The meeting starts with lunch at **13.00** on **September 27th**, and ends with lunch at **noon** on **September 28th**.

VENUE

The workshop takes place at the KTH Royal Institute of Technology main campus in central Stockholm, in the Fantum lecture room on the premises of KTH Speech, Music and Hearing (Lindstedtsvägen 24). KTH Speech, Music and Hearing is one of the world's oldest, still existing speech technology departments, founded by renowned phonetician Gunnar Fant in 1951. It is the home of the CLARIN-SPEECH K-Centre.

ACCOMMODATION

We recommend Elite Hotel Arcadia which is walking distance from the venue. Bookings at Elite Hotel Arcadia can be managed by us.

REGISTRATION

Interested participants wishing to have their travel financed should contact their national coordinator for approval before they register:

Denmark: Bente Maegaard (bmaegaard æt hum.ku.dk)

Finland: Krister Lindén (krister.linden æt helsinki.fi)

Iceland: Eríkur Rögnvaldsson (eirikur æt hi.is)

Norway: Koenraad De Smedt (desmedt æt uib.no)

Sweden: Lars Borin (lars.borin æt svenska.gu.se)

Registration is at https://goo.gl/forms/dqmitLD95ICJWYXL2 before 2017-09-20.