

COST Action CA16111
Action Title: International Ethnic and Immigrant Minorities Survey Data Network
(ETHMIGSURVEYDATA)

AGENDA COST WG1, WG2 &WG3 MEETING

Date: 30 April 2019

Location: Universitat Pompeu Fabra

Campus Ciutadella, Carrer de Ramon Trias Fargas, 25-27

Building: Roger de Llúria. Room: 40.249,

08005 Barcelona

Meeting type: WG1, WG2 &WG3 Meeting

Meeting title: QUESTION DATA BANK

10:00 -11:30 – Introduction: Data for Natural Language Processing (NLP)
What is a corpus? What is not a corpus? What are parallel corpora?
Example: The Norwegian-Spanish Parallel Corpus (NSPC)

11:30 – 12:00 – Coffee Break

12:00 – 13:00 – Creating a corpus (basic theory).
Framework for corpus building: all the steps to get your data ready for NLP.

13:00 – 14:00 – Lunch Break

14:00 – 16:00 – Creating a corpus (applied session)
Management of PDF files
R in NLP. Raw data preparation and references in R
Alignment (LF Aligner)
Annotation (BRAT software)
Corpus processing (SketchEngine)

****NOTE: Participants will need to bring their own laptop****