A digital text collection of Tundra Nenets

Nikolett Mus
Hungarian Research Centre for Linguistics, Budapest
mus.nikolett@gmail.com|nytud.hu

WG1, WG3, WG4

Introduction

- https://tundranenetsdata.nytud.hu/bonito
- a joint work with Réka Metzger (metzger.reka@gmail.com)

Tundra Nenets

- an endangered, indigenous, minority language
- c. 20,000 people
- in North-Eastern part of Europe and in North-Western part of Siberia
- Samoyedic branch of the Uralic language family
- an agglutinative-concatenating, nominative/accusative, head-final ((S)OV) language
- no (written or spoken) standard
- not any large, robust, balanced, and/or representative corpus
- a (relatively) huge amount of sources in print and/or on the web

Methodology & results

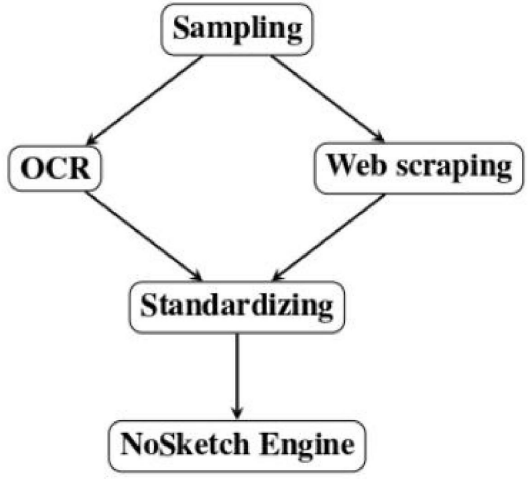

Figure 1: Workflow of text processing