English-Swedish-Turkish Corpus

SND-ID: ext0078-1.

Creator/Principal investigator(s)

Beáta Megyesi - Uppsala University, Department of Linguistics and Philology

Éva Csató Johanson - Uppsala University, Department of Linguistics and Philology

Bengt Dahlqvist - Uppsala University, Department of Linguistics and Philology

Joakim Nivre - Uppsala University, Department of Linguistics and Philology

Eva Pettersson - Uppsala University, Department of Linguistics and Philology

Research principal

Uppsala University - Department of Linguistics and Philology

Description

We describe a syntactically annotated parallel corpus containing typologically partly different languages, namely English, Swedish andTurkish. The corpus consists of approximately 300 000 tokens in Swedish, 160 000 in Turkish and 150 000 in English, containing bothfiction and technical documents. We build the corpus by using the Uplug toolkit for automatic structural markup, such as tokenizationand sentence segmentation, as well as sentence and word alignment. In addition, we use basic language resource kits for the linguisticanalysis of the languages involved. The annotation is carried on various layers from morphological and part of speech analysis todependency structures. The tools used for linguistic annotation, e.g., HunPos tagger and MaltParser, are freely available data-drivenresources, trained on existing corpora and treebanks for each language. The parallel treebank is used in teaching and linguistic researchto study the relationship between the structurally different languages. In order to study the treebank, several tools have been developedfor the visualization of the annotation and alignment, allowing search for linguistic patterns.

Purpose:

The main goal of the project is to promote research and teaching in the Turkish language. More specifically, the aim is to build a language resource for Turkish, Swedish and English allowing contrastive studies between the involved languages.

Data contains personal data

No

Responsible department/unit

Department of Linguistics and Philology

Research area

Languages and literature (Standard för svensk indelning av forskningsämnen 2011)

Language and linguistics (CESSDA Topic Classification)

Keywords

Texts, Linguistics

Publications

Csató Johansson, Megyesi, Beáta, Dahlqvist, Bengt, Csató, Éva Á. & Nivre, Joakim, 'The English-Swedish-Turkish Parallel Treebank', Proceedings of Language Resources and Evaluation (LREC 2010)., 2010 http://uu.divaportal.org/smash/get/diva2:306475/FULLTEXT01.pdf
Download here | Swepub

If you have published anything based on these data, please notify us with a reference to your publication(s). If you are responsible for the catalogue entry, you can update the metadata/data description in DORIS.

Accessibility level

Access to data through an external actor
Access to data is restricted

Homepage

Link to description and demo of the corpus.

Contact for questions about the data

Beáta Megyesi

beata.megyesi@lingfil.uu.se

CLARIN Virtual Collection Registry

Add to collection
A virtual collection is connected to a specific research purpose and contains links to data resources from various digital archives. It is easy to create, access, and cite the collection.

Read more about virtual collections on the CLARIN website.

Download metadata

DataCite

DDI 2.5

DDI 3.3

DCAT-AP-SE 2.0

JSON-LD

PDF

Citation (CLS)