Omsetjingsminne frå Doffin

Nasjonalbiblioteket, 2020-12-18

Dette korpuset inneheld data frå Doffin, den nasjonale kunngjeringsbasen for offentlege anskaffingar, forvalta av Direktoratet for Forvaltning og Økonomistyring (DFØ). Språkbanken fekk dataa i from av ein dump av ein XML-database.

Dumpen bestod av 41.143 dokumentpar (originalar og omsetjingar). 40.631 av desse var omsetjingar frå norsk til engelsk. Berre dei sistnemnde er inkluderte i korpuset. Av dei opphavleg norske dokumenta er 39.893 på bokmål og 736 på nynorsk.

Original og omsetjing vart først parallelliserte på dokumentnivå ved hjelp av ein intern dokumentidentifikator, deretter vart setningane identifiserte med NLTK Punkt Sentence Tokenizer og parallelliserte ved å nytte Hunalign. Dupliserte omsetjingar (eksakte duplikat) vart kasserte.

Totalt fann me 293.649 omsetjingseiningar (Translation Units – TU) for bokmål til engelsk, og 6.342 TUar for nynorsk til engelsk. Ein TU er eit omsetjingspar med ei originaltekst og ei parallellisert omsetjing, og svarar til ei meir eller mindre meiningsberande språkleg eining, typisk ei setning, overskrift eller liknande. Ein TU kan òg bestå ev eit enkeltord eller fleire setningar. Omsetjingseiningane for bokmål og nynorsk vert distribuerte som to separate filer, begge i TMX 1.4-format (ein variant av XML).

Translation Memory from Doffin

National Library of Norway, 2020-12-18

This corpus contains data from Doffin, the Norwegian web-based database for notices of public procurement and procurement in the utility sector, managed by The Norwegian Agency for Public and Financial Management. The Language Bank received the data in the form of an XML database dump.

The dump consisted of 41.143 document pairs (original and translation). 40.631 of these were translations from Norwegian to English. Only the latter are included in the corpus. Of the originally Norwegian documents, 39.893 were in Norwegian Bokmål and 736 in Norwegian Nynorsk.

Original and translation were first aligned on document level using an internal document identifier, then the sentences were extracted using the NLTK Punkt Sentence Tokenizer and aligned using Hunalign. Duplicate translations (exact duplicates) were discarded.

We recorded a total of 293.649 translation units (TUs) for Norwegian Bokmål to English, and 6.342 TUs for Norwegian Nynorsk to English. A TU is a translation pair with an original text and a parallelized translation, and corresponds to a more or less meaningful linguistic unit, typically a clause, a heading etc. A TU may also consist of a single word or several clauses. The translation units for the two languages are distributed as two separate files, both in TMX 1.4 format (a variant of XML).