Ajalooliste tekstide normaliseerimine

Gerth Jaanimäe, Tartu Ülikool

Normaliseerimine – Mis ja milleks?

- Tekstide teisendamine mittestandardselt kujult õigekirjanormidele vastavaks.
- Enamasti kasutatakse internetikeele või varasemalt kasutusel olnud kirjaviisis tekstide kontekstis.
- Normaliseeritud tekstidest saab otsida märksõnu ja neid uurida standardkeele jaoks väljatöötatud vahendite abil.

Uuritav andmestik

- 19. sajandil kirjapandud vallakohtuprotokollid.
- Rahvusarhiivi rahvahanke korras käsitsi piltidelt sisestatud tekstid.
- Umbes 3000000 tekstisõnet.
- Nendest käsitsi märgendatud umbes 70000 sõnet.

Näide

Harjumaa 23.08.1874

* Kautjallast Ann Keapa Pearna lesk kaebas Kautjallast pilleti peal soldat Hans Mättas peal, et Hans Mättas on tedda käggistanud ja ta naene ja naese emma on puu rondidega tedda wissand mis agga mitte temma pihta ei olle läind ja lubband tedda mahha lüa. Jääb polel teiseks kohtupääwaks kunni Hans Mättas ja ta naene ja naese emma saawad ette kutsutud.

- Kautjalast Ann Kääpa Päärna lesk kaebas Kautjalast pileti peal soldat Hans Mätase peale, et Hans Mätas on teda kägistanud ja ta naine ja naise ema on puu rontidega teda visanud mis aga mitte tema pihta ei ole läinud ja lubanud teda maha lüüa. Jääb pooleli teiseks kohtupäevaks kuni Hans Mätas ja ta naine ja naise ema saavad ette kutsutud.

Probleemid

- Palju varieeruvust
- Osa tekste on uues, osa vanas kirjaviisis ning osa n.ö. üleminekukirjaviisis.
  - Tulli ette Johann Alajaan, kes käes olewal aastal wäe teenistuse kohuse ...
- Põhja- ja lõunaeesti keel
- Erinevad murdejooned
- Eesti keelele omane morfoloogiline rikkalikkus

Normaliseerimismeetodid

- Mitmeid erinevaid
- Sõnastikud
- Reeglipõhised lähenemised
- Masintõlge
- jt

Masintõlge

- Vana ja uut kirjaviisi käsitletakse eraldi keeltena.
- Tähemärgipõhise masintõlke puhul käsitletakse sõnu kui lauseid ja tähemärke kui sõnu.
- Nii saab n.ö. tõlkida erinevaid täheühendeid ja nendega seonduvaid mustreid.
- Varasemalt katsetatud nt vanade sloveenikeelsete tekstide peal ja võrreldud tulemusi inglise, saksa, rootsi, islandi ja ungari keeleandmetel (Scherrer, Erjavec 2013, Pettersson & al 2013).
- Vaja treeningandmeid.

Probleemid

- Treeningandmeid võib olla liiga vähe.
- Vana kirjaviis on mitmeti mõistetav *kalla* tähendab nii *kala* kui ka *kalla.*
- Esimese probleemi lahendamiseks tekitati kunstlikult andmeid juurde ehk teisisõnu n.ö. hõbestandard.
- Tänapäeva kirjakeelele kõige lähedamal olevad tekstid teisendati mõne reegli abil vanasse kirjaviisi.

* Kasutada lihtsalt suuremat keelemudelit.

- Teise probleemi lahendamiseks kasutati sõnade asemel sõnapaare.
- Mis võib minna valesti järgmise lausega: *Ette tulli Jakob Kerra ja kaibas et Andri Uggur ollewat tedda ilma asjanda temma krami wälja pillanu nink mõllemba naisega tedda pesnud*.

Katsete tulemused

Igat meetodit rakendati testhulgal 10 korda. Järgnevad korrektsuste keskmised.

- Baastõlge (ilma andmete eeltöötluseta): 83.61%
- Hõbestandard: 83.93%
- Suur keelemudel: 85.5%
- Sõnapaarid: 79.58%

Tulevikuplaanid

- Hõbestandardi parendamine
  - Masintõlke kombineerimine teiste meetoditega

Viited

- Y. Scherrer, T. Erjavec, Modernizing historical Slovene words with character-based SMT, in: 4th Biennial Workshop on Balto-Slavic Natural Language Processing, 2013.
- E. Pettersson, J. Tiedemann, B. Megyesi, An SMT approach to automatic annotation of historical text, in: Proceedings of the workshop on computational historical linguistics, 2013.