

Automaatne parafraside leidmine ning sõnade ja lühifraaside tõlkimine paralleelkorpusete abil

Maarika Traat

Tehnilised andmed

- Taitjad:
 - Maarika Traat
 - Hendrik Nigul
 - Krista Liin
- Rahastamine
 - 2009 : 200 000 (160 000)

Eesmärk

- Luua veebiliidesega tööriist, mis väljastab sisestatud sõna või fraasi
 - tõlked
 - parafraasid
- Töö põhineb joondatud paralleelkorpusete kasutamisel
- Tööriist on abiks tõlkimisel, ühekeelse teksti kirjutamisel, leksikograafilises töös (nt tesauruse/wordneti täiendamisel)
- Eeskuju: Chris Callison-Burch'i ja Linear B sarnane tööriist <http://linearb.co.uk/>
- Koostöö masintõlke projekti inimestega

Tõlked

- Erinevus sõnaraamatust: tõlked vabamad, “ebatäpsemad”, laiema diapsooniga tõlgete valik
- Korpustest leitud reaalsed sõna/fraasi erinevates kontekstides kasutatud tõlkevasted
- Abiks tõlkimisel

Parafraasid

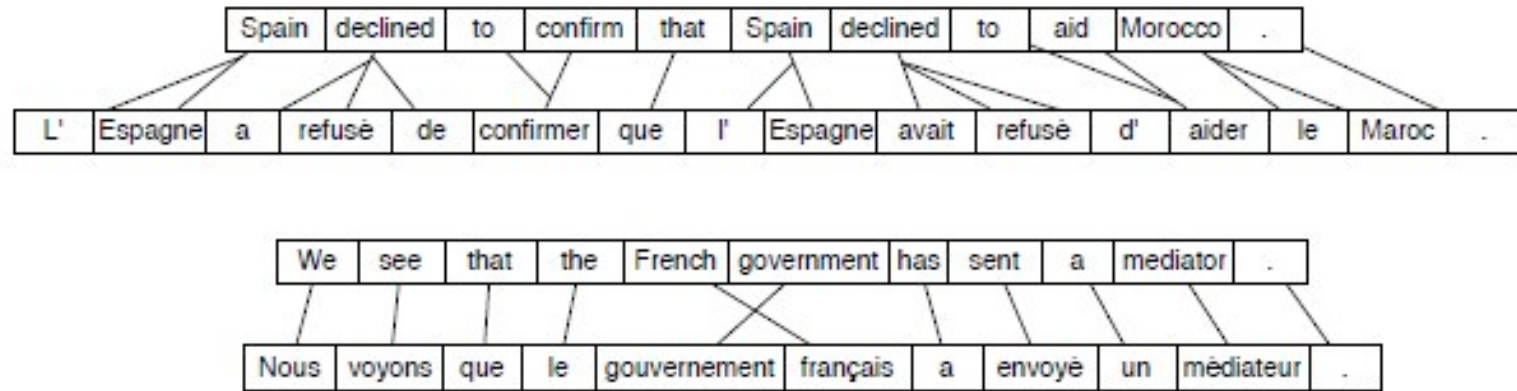
- Parafraasid ehk ümbersõnastused annavad sama informatsiooni edasi erineval moel.
 - Festivali "Kooliteater 2009" tänavuseks parimaks gümnaasiumitaseme teatritrupiks pärjati Hugo Treffneri Gümnaasiumi HTG66
 - Tänavuse kooliteatrite festivali võitis trupp HTG66
 - sai esimese koha, tuli esimeseks, tuli võitjaks
- Abiks ühekeelse teksti kirjutamisel
- Keeletehnoloogias: keele genereerimisel, mitmel dokumendil põhinevate sisukokkuvõtete tegemisel, masintõlkes, küsimustele vastamisel



Paralleelkorpused

- **Paralleelkorpus** – korpus, mis sisaldab mingit teksti originaalkeeles ja selle tõlget teise keelde või tõlkeid teistesse keeltesse.
- Paralleelkorpused joondatakse erinevatel tasanditel (lause, osalause, fraas, sõna)

Paralleelkorpuste kasutamine fraaside tõlkimiseks



declined – a refusé de, avait refusé d’

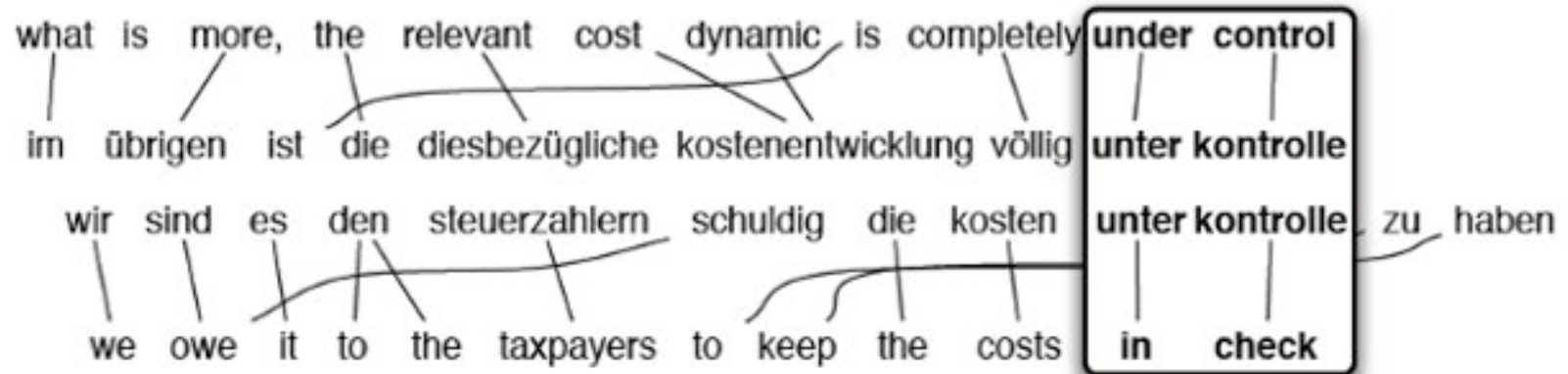
declined to confirm – a refusé de confirmer

declined to aid – avait refusé d’aider

French government – gouvernement français

<http://linearb.co.uk/>

Paralleelkorpuste kasutamine parafraside leidmiseks



- under control → unter kontrolle = unter kontrolle → in check
- <http://linearb.co.uk/>

Eesti keelt sisaldavad paralleel- korpused

- JRC Acquis – koostatud Acquis Communautaire'i baasil (Euroopa ühenduse seadustekstid; suurim paralleelkorpus)
 - **22 languages:** Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene and Swedish
- Opus – veebist kogutud paralleeltekstid, mitmed alamkorpused (subtiitrid, meditsiinalased dokumendid, arvutimanuaalid)

Tõlkemälud

- Tõlkemälu on lühikeste tekstilõikude (laused või lauseosad) kogu. Tõlkide töövahend, mis säästab neid uuesti tõlkimast tekstilõike, mille varasemad tõlked on olemas.
- DGT Multilingual Translation Memory (DGT-TM) – Acquis Communautaire'i TM (käsitsi joondatud)
- Kohalike riigiasutuste, tõlkebüroode tõlkemälud

Tegevusplaan esimeseks aastaks

- tutvuda Chris Callison-Burchi ja tema firma, Linear B loodud tööriistas "Searchable Translation Memories" (<http://linearb.co.uk>)
- määratleda, mida on vaja teha selleks, et seda programmi eesti keelele kohandada
- tutvuda olemasolevate mitmekeelsete paralleelkorpustega, mis sisaldavad eesti keelt (JRC-Acquis ja OPUS)
- koguda andmeid juurde, sealhulgas pidada tõlkebüroode ja tõlkemälusid kasutavate riigiasutustega läbirääkimisi nende tõlkemälude meie andmebaasi kaasamise võimaluste üle
- aasta lõpuks luua esialgne (baseline) programm sõnade ja lühifraaside tõlkimiseks eesti keelest enamräägitud võõrkeeltesse ja vastupidi, ning eestikeelsete parafraaside leidmiseks

Tänan tähelepanu eest!

Thank you for your attention!

Thank you for listening to me!

Many thanks for your attention!

I would like to thank you for your attention!

Vielen dank für ihre aufmerksamkeit!

Je vous remercie de votre attention!

Gracias por su atención!

Vi ringrazio per l' attenzione!

Tack för er uppmärksamhet!

Ik dank u voor uw aandacht!

اشكرکم علی حسن استماعکم

多谢你神