TURUN YLIOPISTO

Venäjänkielisten internettekstien annotointi ja rekisterien vertailu suomen ja venäjän välillä

Nella Särkioja
Pro gradu -tutkielma
Kieliasiantuntijuuden tutkinto-ohjelma, digitaalinen kielentutkimus
Kieli- ja käännöstieteiden laitos
Humanistinen tiedekunta
Turun yliopisto
Toukokuu 2023

Pro gradu -tutkielma

Kieliasiantuntijuuden tutkinto-ohjelma, digitaalinen kielentutkimus
Nella Särkioja
Venäjänkielisten internettekstien annotointi ja rekisterien vertailu suomen ja venäjän välillä
Sivumäärät: 39 sivua, 3 sivua liitteitä

Tutkielman aiheena on internettekstien jakautuminen rekistereihin. Tutkielmassa käytetään valmiita rekisteriluokkia ja luodaan käsin venäjänkielinen aineisto, jossa venäjänkielisiä internettekstejä on luokiteltu rekistereihin. Tarkoituksena on kuvata venäjänkielistä aineistoa rekistereiden avulla sekä vertailla sitä suomenkielisten rekistereiden kanssa.

Tutkimus liittyy Turun yliopistossa käynnissä olevaan hankkeeseen Uutinen, mielipide vai jotain muuta? Erilaiset tekstit ja niiden automaattinen tunnistus monikielisestä internetistä, jossa kehitetään erilaisia automaattisia menetelmiä jaottelemaan internettekstejä rekistereihin. Tutkimuksen aineisto on koottu vapaasta internetistä ja annotoitu yhtenevin ohjeistuksin. Venäjänkielinen aineisto on annotoitu tutkielmaa varten ja vertailuun käytetään suomenkielistä, valmiiksi annotoitua aineistoa.

Tutkielman tavoitteena on tuottaa uutta tietoa rekistereistä tarkastelemalla ja vertailemalla rekisterien frekvenssejä sekä yksittäisiä esimerkkejä tarkastelemalla. Tärkeänä osana tutkielmaa on uusi, venäjänkielinen aineisto rekisteritutkimusta varten, joka on vapaasti saatavilla https://github.com/TurkuNLP/RuCORE. Suomenkielisiä ja venäjänkielisiä rekistereitä tarkastelemalla huomattiin paljon yhteneväisyyksiä mutta myös eroja kielten välillä.

Avainsanat: kielitiede, kieliteknologia, annotointi, rekisteri, rekisteriluokka

Sisällysluettelo

1 Johdanto

Tutkimus liittyy Turun yliopistossa käynnissä olevaan hankkeeseen *Uutinen, mielipide vai jotain muuta? Erilaiset tekstit ja niiden automaattinen tunnistus monikielisestä internetistä*, jossa kehitetään erilaisia automaattisia menetelmiä jaottelemaan internettekstejä rekistereihin. Rekisterit ovat tekstiluokkia, joihin tekstit voidaan jakaa. Rekisterit kuvaavat tekstien kielellisten piirteiden tilannekohtaista vaihtelua (Biber & Conrad 2019). Esimerkiksi onko kyseessä oleva teksti uutinen, käyttöohje vai mielipidekirjoitus. Tässäkin tutkimuksessa käytetyt rekisterit ovat syntyneet todellisten verkkotekstien pohjalta. Tutkimuksissa on todettu, että parempaan rekistereiden tunnistamiseen tarvitaan käsin annotoitua monikielistä dataa. Koska rekisterin avulla voidaan ennustaa tekstin kielellisiä piirteitä, voisi automaattinen rekisterintunnistus laajentaa internetissä olevan aineiston mahdollisuuksia kielitieteessä. (Laippala ym. 2019.)

Automaattisten monikielisten rekisterintunnistusohjelmien kehittämisen tueksi tarvitaan käsin annotoitua dataa mallien arvioimiseen sekä treenaamiseen. Annotoituja aineistoja voi myös hyödyntää kielentutkimuksen aineistona. Monikielisten rekisterintunnistusohjelmien kehittäminen hyödyttää erityisesti pieniä kieliä kuten suomea, sillä on osoitettu, että monikieliset kielimallit tuovat selvää hyötyä rekisterien tunnistamiseen etenkin pienillä kielillä, joilla on vähän treeniaineistoa.

Tutkielman tarkoituksena on kuvata venäjänkielisiä rekistereitä sekä vertailla niitä suomenkielisten rekistereiden kanssa. Vertaamalla aineistoja ja tarkastelemalla venäjänkielistä aineistoa on tarkoitus vastata seuraaviin tutkimuskysymyksiin. Miten venäjänkielisen ja suomenkielisen aineiston tulokset eroavat toisistaan esimerkiksi frekvenssien osalta? Onko rekistereillä yhteneväisyyksiä? Jos merkittäviä eroja löytyy, niin minkälaisia eroavaisuudet ovat? Millaisia ominaispiirteitä rekistereillä on ja onko kielten välillä eroa? Rekisterijaon lisäksi tarkastellaan esimerkkitekstejä havainnollistamaan rekisterijakoa. Tutkimuksen tarkoitus on lisätä tietoa eri kielten välisistä eroista ja yhtäläisyyksistä rekisterien suhteen.

Työn tavoitteena on annotoida venäjänkielisiä verkkotekstejä. Annotointi tarkoittaa tässä kontekstissa verkkotekstien systemaattista luokittelua konelukuisessa muodossa. Osana tutkielmaani olen annotoinut venäjänkielisiä verkkotekstejä rekistereittäin ja luonut siten venäjänkielisen aineiston hanketta varten eli tuottanut uutta tutkimustietoa. Annotoinnilla

tarkoitetaan tässä tutkimuksessa sitä, että olen määritellyt siis käsin rekisteriluokan suurelle määrälle venäjänkielisiä tekstejä.

Hankkeessa ja tässä tutkielmassa käytetyt rekisterit perustuvat Douglas Biberin ja hänen tutkimusryhmänsä kehittämään luokitteluun. Luokittelu perustuu laajaan aineistoon ja empiiriseen havaintoihin rekistereistä kyseisessä aineistossa (Biber & Egbert 2018, 10). Turku NLP on luonut tämän pohjalta oman ohjeistuksensa tekstien annotointiin eli luokitteluun (liite 1). Tutkielmassa annotoidut tekstit on luokiteltu tämän ohjeistuksen perusteella.

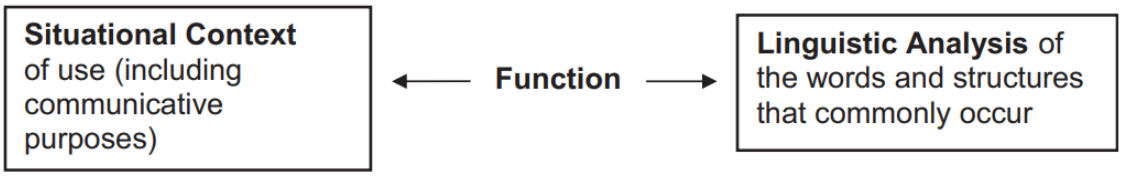
Tutkielman kannalta keskeisintä on 8 ylärekisteriä: kerronnallinen (NA), mielipide (OP), ohjeet (HI), vuorovaikutteinen (ID), informatiivinen kuvaus (IN), runollinen (LY), puhuttu (SP) ja informatiivinen suostuttelu (IP). Joitakin alaluokkia tarkastellaan tutkimuksen kannalta merkittävien rekistereiden osalta. Tämän lisäksi tutkielmassa havainnollistetaan esimerkkien avulla myös hylättyjä tekstejä sekä hybriditekstejä. Hylätyille teksteille ei ole ollut mahdollista antaa rekisteriluokkaa ja hybriditeksteille sen sijaan on annettu useampi rekisteriluokka.

Tutkielman luvussa kaksi esitellään tutkimuksen teoriataustaa tarkemmin. Luvussa kolme tarkastellaan tutkimuksessa käytettyjä aineistoja ja menetelmiä. Neljännessä luvussa esitellään tutkimuksen tulokset erityisesti määrällisten tulosten osalta. Tutkielman viidennessä eli analyysiluvussa tarkastellaan tarkemmin tutkimuksen tuloksia sekä havainnollistetaan rekisterijakoa esimerkkien avulla. Lopuksi viimeisessä luvussa on yhteenveto koko tutkimuksesta. Lisäksi tarkastellaan tutkimuksen keskeisiä ongelmakohtia sekä esitellään aiheita jatkotutkimukselle.

2 Teoriatausta

”Sana genre juontaa juurensa latinan sanaan *genus,* joka merkitsee ’ryhmää, jolla on yhteisiä ominaisuuksia’ ja elottomista puhuttaessa ’lajia, laatua’- - -” (Mäntynen ym. 2006, 13). Tekstillä voidaan tarkoittaa yhtä rajattua tekstikokonaisuutta tai sitten rajattomaa määrää tekstiä. Tekstilaji käsitteenä liittyy kiinteästi tekstiin, sillä tekstejä on vaikea käsitellä luokittelematta niitä ja siitä tarpeesta ovat syntyneet esimerkiksi arjen määritelmät eri tekstilajeista. Tekstilajilla tarkoitetaankin tekstin luokittelua joko tieteellisesti tai arkielämässä vaikkapa resepteihin, runoihin, mainoksiin, sanomalehtiartikkeleihin jne. Yhden tekstilajin sisällä on kuitenkin aina vaihtelua, eikä tekstilajienkaan rajat ole aina selviä. (Mäntynen ym. 2006, 9). Tässä tutkimuksessa tarkoitetaan tekstillä yhtä rajattua tekstikokonaisuutta. Kieliteknologiassa puhutaan rekistereistä tekstilajien sijaan. Näillä on kuitenkin paljon yhteistä. Rekisterille oleellista on tekstin konteksti sekä kielelliset piirteet, ja se miten nämä liittyvät toisiinsa (kaavio 1). Oleellista on siis tekstin tarkoitus ja sen kielellisten piirteiden analyysi (Biber & Conrad 2019, 6).

Kaavio 1: Rekisterianalyysin osat (Biber & Conrad 2019, 6)



Korpuslingvistiikka on kielitieteen osa, joka keskittyy tutkimaan korpuksia eli rajattuja aineistoja kirjoitetusta tai puhutusta kielestä. Kyseessä on siis korpuslingvistinen tutkimus. ”*Korpuksella tarkoitetaan kielentutkimuksessa laajaa sähköisessä muodossa olevaa tietokoneluettavaa tekstikokoelmaa, joka on strukturoitu ja edustava tutkimuksen tarpeita palvelevalla tavalla*” (Luodonpää-Manni ym. 2020, Luku 9 Korpusaineistot sivunumero?). Aiemmin korpuslingvistiikassa on keskitytty selkeästi rajattuihin aineistoihin. Kun aineistona on ”koko internet” eli internetistä satunnaisesti otetut tekstit tarkasti rajatun korpuksen sijaan, on paljon mahdollisuuksia, mutta myös ongelmia. Internetissä olevien tekstien kirjo on hyvin laaja ja se luo ongelmia tekstien luokitteluun. Missä menee raja rekisterien välillä? Perinteisten, painettujen tekstien kuten tietokirjojen tai sanomalehdessä olevan uutisen, rekistereiden välillä rajat ovat usein selvempiä.

Internet tarjoaa suuren määrän aineistoa, joka luo paljon mahdollisuuksia luonnollisen kielen käsittelyn (Natural Language Processing) tutkimuksen saralla (Laippala, Kyllönen, Egbert, Biber & Pyysalo, 2019). Tietokoneet ja internet mahdollistavat paljon aiempaa suuremmat tutkimusaineisot ja automaattiset ohjelmat helpottavat tutkijoiden työtä (Luodonpää-Manni ym. 2020, 8). Internetin ongelmana on kuitenkin tietysti sen suuri koko, jolloin aineistoa on mahdotonta tarkastella käsin ja tähän tarvitaan kieliteknologian ratkaisuja. Digitaalistuvassa maailmassa internetissä vapaasti kaikkien saatavilla olevien tekstien merkitys korostuu ja onnistuneen luokittelun avulla tekstejä on mahdollista hyödyntää aiempaa paremmin. Rekisteri on yksi tärkeimmistä tekijöistä, mikä ennustaa tekstin kielellisiä piirteitä. Rekisteri myös vaikuttaa siihen, miten tulkitsemme tekstiä. (Biber, 2012.)

Vapaasti internetissä olevat tekstit ovat kaikkien käytössä, mutta on hyvä muistaa, ettei kaikki netissä olevat tekstit ole vapaasti saatavilla. Suuri osa kirjoista, lehdistä ja akateemisista julkaisuista ei ole ilmaiseksi kaikkien saatavilla. Kuitenkin ilmaiseksi on huikea määrä tietoa saatavilla ja suuri yleisö turvautuukin tiedonhaussaan näihin teksteihin. (Biber & Egbert 2018, 136.) Perinteisesti tutkimuksessa on hyödynnetty nimenomaan kaupallisesti julkaistuja tekstejä, kuten kirjoja ja lehtiä. Vapaasti internetissä olevien tekstien rekisterianalyysi keskittyy siis teksteihin, jotka ovat aiemmin jääneet tutkimusten ulkopuolelle. Internettekstien rekisterianalyysi asettaa siten kyseenalaiseksi myös perinteisen rekisteriluokittelun, jossa tekstit jaetaan rekistereihin julkaisumuodon mukaan (esimerkiksi lehdissä julkaistut tekstit ovat lehtiartikkeleita jne.) (Biber & Egbert 2018, 8–9.)

Maailmanlaajuinen internet on valtava aineisto ja kasvaa jatkuvasti. Rekisterien tunnistaminen on tärkeää, koska se antaa paljon tietoa tekstistä. Tämän vuoksi rekistereiden tunnistaminen on tärkeää tekstinprosessoinnin ja kieliteknologian kehityksen kannalta. Rekistereitä olisi mahdollista käyttää helpottamaan tiedonhakua, sillä nykypäivänä usein suurimpana ongelmana ei ole tiedon puute vaan tiedon löytäminen. Lisäksi rekistereitä voidaan käyttää kieltä automaattisesti käsittelevien sovellusten kehittämisessä. Tällaisia sovelluksia ovat esimerkiksi konekääntimet ja oikolukuohjelmat. Internetaineistoja käytetään myös yhä enemmän kielentutkimuksen aineistona. Systemaattinen rekisteriluokittelu auttaisi ymmärtämään paremmin internetissä käytettävää kieltä. (Egbert ym. 2015, 1817.)

Internettekstien rekisterianalyysissä rekistereihin jako ei siis ole ollenkaan yhtä selkeää kuin perinteisesti ulkoa määritelty jako. Internettekstit ovat myös vähemmän kontrolloituja ja arvaamattomampia verrattuna paperijulkaisuihin. Painettujen aineistojen rekisterit ovat

yleensä vakiintuneempia ja kontrolloidumpia toimittajien sekä kustantajien asettamien rajoitusten ja ohjeistusten takia. (Mehler ym. 2010, 9.) Tietysti myös joillekin internetteksteille on määritetty ulkoisesti luokka, esimerkiksi uutissivustojen tekstit ovat usein selkeästi uutisia, mutta tekstit kuitenkin luokitellaan rekistereittäin tekstien tyylin perustella sen mukaan, onko teksti objektiivinen ohje, mielipidekirjoitus, informatiivinen, narratiivinen vai vuorovaikutuksellinen teksti. Rekisteriluokkien jako ja rajat luokkien välillä eivät tällöin ole yhtä selkeitä kuin ”ylhäältä päin” tulevassa luokittelussa.

Rekisterijaottelussa lähdetään ensin siitä, onko kyse alun perin puhutusta vai kirjoitetusta tekstistä. Puhuttu teksti voi olla esimerkiksi haastattelu tai litteroitu puhe. Kirjoitetut tekstit taas jaetaan sen mukaan, onko kyse interaktiivisesta keskustelusta vai ei. Tekstit, jotka eivät olleet vuorovaikutteisia, jaettiin vielä kuuteen eri kategoriaan sen mukaan, mikä on niiden viestinnän tarkoitus: onko tekstin tarkoituksena kertoa tapahtumasta, informoida, ilmaista mielipidettä, kertoa faktoja tarkoituksena vaikuttaa mielipiteisiin, antaa ohjeita vai ilmaista itseään lyriikan keinoin. Narratiivisia tekstejä ovat esimerkiksi uutiset, blogit, historialliset artikkelit, novellit, romaanit ja lehtiartikkelit. Informatiivisia tekstejä ovat taas esimerkiksi tieteelliset artikkelit, objektiiviset kuvaukset, informatiiviset blogit ja tekniset raportit. Mielipideteksteissä ilmaistaan selkeästi oma mielipide. Sellaisia ovat esimerkiksi mielipideblogit, arvostelut, neuvot ja mainokset. Esimerkiksi myyntikuvaukset, suostuttelevat artikkelit tai esseet ovat taas tekstejä, joissa kerrotaan faktoja, mutta koitetaan vaikuttaa mielipiteisiin. Ohjeita ovat esimerkiksi reseptit ja tekniset ohjeistukset. Lyyrisiä tekstejä ovat puolestaan kappaleiden sanat, runot, rukoukset jne. Rekisteriluokkia luodessa huomattiin, että monet tekstit ovat niin kutsuttuja hybriditekstejä eli teksti saattaa kuulua kahteen tai useampaan eri rekisteriluokkaan samanaikaisesti. (Biber & Egbert 2018, 15-16.) Kirjoitetut tekstit siis jaetaan alaluokkiin sen perusteella, mikä on niiden viestinnällinen tarkoitus. Liite 2 kuvaa Biberin ja Egbertin tekemää jaottelua rekisteriluokkiin.

Rekisterien tutkimus eroaa aiemmasta kielentutkimuksesta siten, että näkökulma kieleen on tekstilähtöinen. Tutkimuksen tavoitteena on kuvata nimenomaan rekisterin kielellisiä piirteitä yleisten kielellisten piirteiden sijaan. Kvantitatiiviset havainnot jonkin kielellisen piirteen esiintymisestä kertovat niiden esiintymisestä yleisesti, mutta eivät kuvaa tekstilajien eroja. (Biber 2012, 33.) Rekistereiden avulla pystyy siis kuvaamaan paremmin ja tarkemmin jonkin tietyn tekstilajin kielellisiä piirteitä. Rekisteri siis antaa tietoa tekstin tuottamisesta,

tunnistamisesta sekä käyttötarkoituksesta (Mehler ym. 2010, 4). Tekstin rekisteri antaa siis monenlaista informaatiota.

3 Aineisto ja menetelmät

Turun yliopiston hankkeen aineisto on koottu vapaasta internetistä ja annotointi on suoritettu yhtenevin ohjeistuksin rekistereihin: informatiivinen suostuttelu, kerronnallinen, mielipide, informatiivinen kuvaus, ohjeet, vuorovaikutteinen, puhuttu, runollinen, hybridi tai konekäännetty. Lisäksi osalla rekistereistä on useampia alaluokkia. Puhuttu teksti voi olla haastattelu tai muu puhuttu teksti. Kerronnallisen rekisterin alaluokkia ovat: uutinen, urheiluuutinen, kerronnallinen blogi ja muu kerronnallinen teksti. Ohjeiden alaluokkia ovat resepti ja muu ohje. Informatiivisen rekisterin alle kuuluu tietosanakirjojen artikkelit, tutkimusartikkelit, henkilö- ja asiakuvaukset, usein kysyttyjen kysymyksien vastaustekstit, lakitekstit ja muut infomatiiviset kuvaukset. Mielipiderekisterin alaluokkia ovat: arvostelut, mielipideblogit, uskonnolliset blogit, neuvot ja muut mielipidetekstit. Informatiivisen suostuttelun alarekistereitä ovat myyntikuvaukset, uutisblogit sekä muut informatiivisen suostuttelun tekstit.

Aineisto koostuu suomenkielisestä aineistosta, joka on valmiiksi annotoitu, sekä tutkimuksessani annotoidusta venäjänkielisestä aineistosta. Kyseessä on siis määrällinen korpustutkimus, jossa tulosten pohjalta tehdään myös laadullista analyysia. Tässä tutkielmassa käytetyn venäjänkielisen aineiston koko rajattiin tutkielman määrittämän laajuuden mukaiseksi.

Turku NLP on luonut omat ohjeensa tekstien annotointiin, mutta annotointiohjeissa viitataan Biberin ja Egbertin luomaan rekisterijaotteluun, joka on luotu empiirisesti tarkastelemalla internettekstejä. Biber ja Egbertin rekistereiden tutkimus perustuu siihen, ettei rekisteriluokkia luoda ylhäältä käsin eli esimerkiksi kerätä blogipostauksia blogeista, vaan tekstit kerättiin internetistä ja rekisterit luotiin tekstien perusteella. (Biber & Egbert 2018, 12.) Näin saadaan todellisempi kuva siitä, millaisia tekstejä internetissä todella on sen sijaan, että kuvattaisiin vain ylhäältä päin rajattua osaa teksteistä. Rekisteriluokkien luomiseen käytetty aineisto pohjautuu CORE-nimiseen (Corpus of Online Registers of English) aineistoon, joka sisältää paljon verkkotekstejä. (Biber & Egbert 2018, 13.)

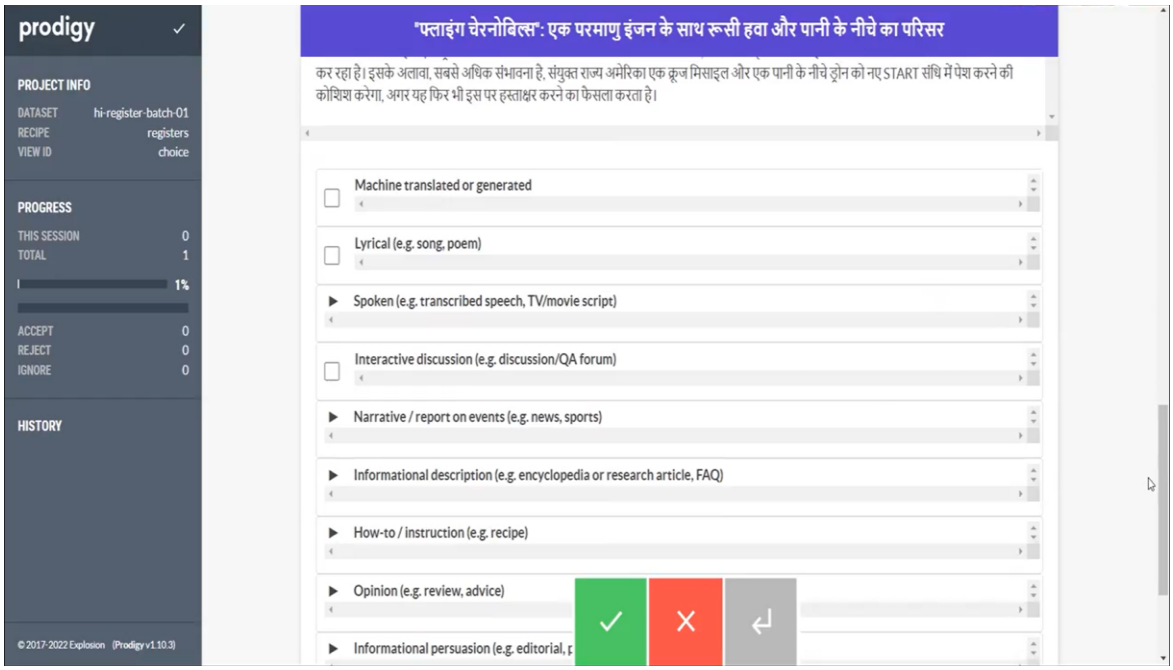
Omassa tutkimuksessa annotoin yhteensä 600 venäjänkielistä verkkotekstiä ja annoin rekisteriluokan 541 venäjänkieliselle tekstille. Tutkimuksen aineisto on luotu Prodigy-ohjelmistolla. Yhteensä 59 tekstille ei ollut mahdollista määrittää rekisteriluokkaa.

Hylätty teksti oli esimerkiksi vain lista linkkejä, teksti ei ollut venäjänkielinen tai teksti ei ollut koherentti kokonaisuus, jolloin rekisteriluokkaa ei voitu määrittää (kuva 1).

Kuva 1: Esimerkki tekstin hyväksymisestä.

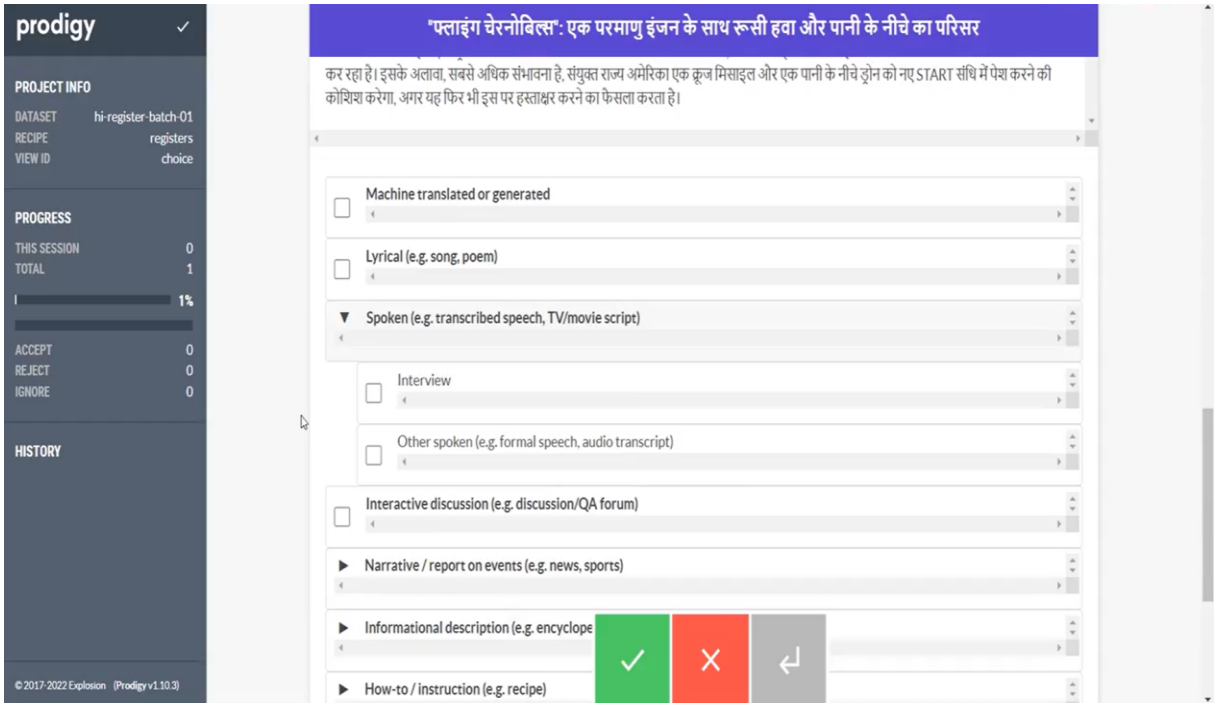
Kuvassa 1 nähdään ohjelma, jonka avulla teksti joko hylättiin tai hyväksyttiin annotointiohjeiden mukaisesti. Kuvan teksti ei ole venäjänkielisestä aineistosta. Vihreää painamalla teksti hyväksytään tai punaista painamalla tekstin voi hylätä. Hyväksytylle tekstille annetaan rekisteriluokka (kuva 2).

Kuva 2: Esimerkki tekstin luokittelusta



Toisille rekisteriluokille on määritelty myös alaluokkia, jolloin pitää valita mihin alaluokkaan teksti kuuluu. Esimerkki tästä kuvassa 3, jossa nuolesta painamalla avautuu puhutun rekisterin alaluokat, jotka ovat haastattelu tai muu puhuttu teksti.

Kuva 3: Rekisteriluokkien alaluokat



Tekstille on myös mahdollista antaa useampi luokka. Lopuksi luokitellut tekstit tallennettiin.

Suomenkielisenä aineistona tutkimuksessani käytän FinCOREa, joka on käsin annotoitu internetrekistereiden korpus. Suomenkielinen korpus sisältää tiedon kunkin tekstin rekisteristä samoin kuin tässä tutkielmassa luotu venäjänkielinen aineisto. Tämänkaltainen aineisto kuvaa hyvin internetin kielellistä vaihtelua. Suomenkielinen aineisto koostuu 2 237 suomenkielisestä dokumentista, ja se on koottu korpuksesta nimeltä Finnish Internet Parsebank, joka on 1,5 miljardin sanan suomenkielinen korpus. (Laippala ym., 2019.) Tutkimuksessani otan myös joitain rekistereitä tarkempaan, laadulliseen tarkasteluun. Tämän tutkimuksen tulokset on saatu aineistosta Coogle Colabin avulla.

4 Tulokset

Työssäni annotoin yhteensä 600 venäjänkielistä tekstiä. Hyväksyttyjä tekstejä eli tekstejä, joille oli mahdollista antaa rekisteriluokka, oli yhteensä 541. Kaikille teksteille ei ollut mahdollista antaa rekisteriluokkaa. Esimerkiksi listat linkeistä, toisella kielellä olevat tekstit tai dokumentit, jotka eivät sisältäneet yhtenäistä johdonmukaista tekstiä, hylättiin. Hylättyjä tekstejä oli yhteensä 59. Taulukosta 1 nähdään venäjänkielisten tekstien jakautuminen rekistereihin.

Taulukko 1

Venäjänkielisen aineiston jakautuminen rekistereihin	Määrä
Informatiivinen suostuttelu (Informational persuasion)	251
Kerronnallinen (Narrative)	113
Mielipide (Opinion)	69
Informatiivinen kuvaus (Informational description)	63
Ohjeet (How-to or instructions)	20
Vuorovaikutteinen (Interactive discussion)	19
Puhuttu (Spoken)	3
Runollinen (Lyrical)	2
Hybridi (Hybrid)	1
Konekäännetty (Machine translated or generated)?	0
YHTEENSÄ	541

Eniten tekstejä on luokiteltu informatiivisen suostuttelun rekisteriin. Suurin osa näistä oli myyntikuvauksia (description with intent to sell). Toiseksi yleisin rekisteri oli kerronnallinen, joista suurin osa oli uutisia.

Venäjänkielistä aineistoa vertailen suomenkieliseen, vastaavaan aineistoon. Taulukossa 2 esitellään suomenkielisen aineiston tuloksia niiden rekisterien osalta, mitkä ovat tässä tutkielmassa vertailun kannalta oleellisia.

Taulukko 2

Suomenkielisen aineiston jakautuminen rekistereihin	Määrä
Informatiivinen suostuttelu (Informational persuasion)	1470
Kerronnallinen (Narrative)	1913
Mielipide (Opinion)	1413
Informatiivinen kuvaus (Informational description)	1745
Ohjeet (How-to or instructions)	554
Vuorovaikutteinen (Interactive discussion)	1096
Puhuttu (Spoken)	92
Runollinen (Lyrical)	26
Konekäännetty (Machine translated or generated)	1415
Koko aineistossa tekstejä YHTEENSÄ	10754

Taulukossa 3 on molempien aineistojen rekisteriluokat prosentuaalisesti verraten tekstien kokonaismäärään. Ensimmäisessä sarakkeessa on venäjänkielisten tekstien tulokset ja toisessa sarakkeessa suomenkielisen aineiston tulokset.

Taulukko 3

Rekisteriluokka	***VEN.***	**SUOM.**
Informatiivinen suostuttelu (Informational persuasion)	46,4 %	13,7 %
Kerronnallinen (Narrative)	20,9 %	17,8 %
Mielipide (Opinion)	12,8 %	13,1 %
Informatiivinen kuvaus (Informational description)	11,6 %	16,2 %
Ohjeet (How-to or instructions)	3,7 %	5,2 %
Vuorovaikutteinen (Interactive discussion)	3,5 %	10,2 %
Puhuttu (Spoken)	0,6 %	0,9 %
Runollinen (Lyrical)?	0,4 %	0,2 %
Konekäännetty (Machine translated or generated)	0 %	13,2 %

Venäjänkielisestä aineistosta eniten tekstejä (46,4 %) kuuluu rekisterin ”informatiivinen suostuttelu” alle eli teksti on esimerkiksi myyntikuvaus, mainos tai vaikkapa lehden mielipidekirjoitus tai artikkeli. Suomenkielisessä aineistossa informatiivisen suostuttelun rekisterin alle kuului suhteessa koko aineistoon merkittävästi vähemmän tekstejä (13,7 %).

Kerronnallista rekisteriä vertailtaessa ei huomata yhtä merkittävää eroa kielten välillä. Venäjänkielisessä aineistossa kerronnallisia tekstejä eli esimerkiksi uutisia, urheilu-uutisia, kerronnallisia blogeja, lehtiartikkeleita tai fiktiivisiä kertomuksia on 20,9 % kun suomenkielisiä tekstejä samassa rekisterissä on 17,8 % teksteistä. Suomenkielisessä aineistossa uutisia on 1359 eli yhteensä 71,0 % kerronnallisen rekisterin teksteistä on uutisia. Sama luku venäjänkielisen aineiston osalta on 71 tekstiä eli yhteensä 62,8 % kerronnallisen rekisterin teksteistä on uutisia.

Kolmanneksi eniten venäjänkielisessä aineistossa on mielipidekirjoituksia kuten arvosteluja, mielipideblogeja, uskonnollisia sivustoja, neuvoja tai muita mielipidekirjoituksia (12,8 %). Suomenkielisessä aineistossa mielipiderekisteriin kuuluvia tekstejä oli suhteellisesti lähes yhtä paljon (13,1 %). Melkein yhtä paljon venäjänkielisiä tekstejä kuuluu informatiiviseen rekisteriin eli ovat esimerkiksi wikipedia-artikkeleita, tieteellisiä julkaisuja, kuvauksia, kysymys-vastaus-palstoja, lainopillisia tekstejä tai muita informatiivisia tekstejä (11,6 %). Suomenkielisessä aineistossa erot rekisterien välillä ovat pienemmät. Informatiivisia kuvauksia teksteistä on (16,2 %), mikä on toiseksi eniten suomenkielisessä aineistossa ja hieman enemmän suhteessa venäjänkieliseen aineistoon verrattuna.

3,7 % teksteistä eli yhteensä 20 tekstiä on ohjeita kuten reseptejä tai muita objektiivisia kuvauksia jonkin asian tekemiseen. Suomenkielisessä aineistossa ohjeiden osuus on 5,2 %. Vuorovaikutteisia tekstejä eli esimerkiksi keskustelufoorumeita suomenkielisessä aineistossa on 10,2 % kun taas venäjänkielisessä aineistossa sama osuus on 3,5 %.

Vähiten tekstejä venäjänkielisessä aineistossa oli alun perin puhuttuun perustuvia tekstejä kuten haastatteluita tai elokuvien tekstityksiä (0,6 %) sekä runollisia tekstejä kuten lauluja tai sanoja (0,4 %). Suomenkielisessä aineistossa on myös selkeästi vähiten puhuttuja (0,9 %) sekä runollisia tekstejä (0,2 %).

Venäjänkielisessä aineistossa ei ollut yhtään konekäännettyä tekstiä tai sitten niitä ei tunnistettu, mutta suomenkielisessä aineistossa konekäännettyjä tekstejä on 13,2 % kaikista

teksteistä. Hybridit ekstejä venäjänkielisessä aineistossa oli 1 eli vain yhdelle tekstille on annettu useampi kuin yksi rekisteriluokka.

5 Analyysi

Tässä kappaleessa tarkastellaan sekä venäjänkielistä aineistoa että suomenkielistä aineistoa esimerkkien avulla, jotta saadaan käsitys siitä, minkälaisia tekstejä aineistot sisältävät ja voidaan vastata tämän tutkielman tutkimuskysymyksiin.

Tulokset kielten välillä poikkeavat toisistaan monilta osin. Yksi vaikuttava tekijä tulosten kannalta saattaa olla venäjänkielisen aineiston pieni koko verrattuna suomenkieliseen aineistoon, joten se on tärkeää huomioida analysoitaessa tuloksia. Venäjänkielinen aineisto on sen verran pieni verrattuna suomenkieliseen aineistoon, että varsinkaan hyvin pienet erot tulosten välillä eivät ole merkityksellisiä analyysin kannalta tässä tutkielmassa. Suomenkielisessä aineistossa tulokset rekisteriluokkien välillä ovat hieman tasaisempia verrattuna venäjänkieliseen aineistoon, mikä voi johtua suomenkielisen aineiston suuremmasta kokonaistekstimäärästä. Mielenkiintoista olisikin nähdä miten venäjänkielinen aineisto jakautuisi rekistereihin suuremmassa aineistossa.

5.1.1 Hylätyt tekstit

Ensimmäisenä tarkastellaan esimerkkejä teksteistä joka on hylätty eli kyseiselle tekstille ei ole voitu antaa rekisteriluokkaa.

Esimerkki 1:

```
reject 30 Ноября, Суббота 29 Ноября, Пятница 28 Ноября, Пятница 27 Ноября, Четверг 26 Ноября, Среда 25 Ноября, Вторник 24 Ноября, Понедельник 23 Ноября, Суббота 22 Ноября, Суббота 21 Ноября, Пятница 20 Ноября, Четверг 19 Ноября, Среда 18 Ноября, Понедельник 17 Ноября, Понедельник 15 Ноября, Суббота 14 Ноября, Пятница 13 Ноября, Четверг 12 Ноября, Вторник 11 Ноября, Вторник 10 Ноября, Воскресенье 09 Ноября, Воскресенье 07 Ноября, Пятница 06 Ноября, Четверг 05 Ноября, Среда 04 Ноября, Вторник 03 Ноября, Воскресенье 02 Ноября, Воскресенье 01 Ноября, Суббота
```

Esimerkin 1 suomennos:

hylätty torstai 6. marraskuuta lauantai 30. marraskuuta perjantai 29. marraskuuta perjantai 28. marraskuuta torstai 27. marraskuuta keskiviikko 26. marraskuuta tiistai 25. marraskuuta maanantai 24. marraskuuta lauantai 23. marraskuuta lauantai 22. marraskuuta perjantai 21. marraskuuta torstai 20. marraskuuta keskiviikko 19. marraskuuta maanantai 18. marraskuuta maanantai 17. marraskuuta lauantai 15. marraskuuta perjantai 14. marraskuuta torstai 13. marraskuuta tiistai 12. marraskuuta tiistai 11. marraskuuta sunnuntai 10. marraskuuta sunnuntai 9. marraskuuta perjantai 7. marraskuuta keskiviikko 5. marraskuuta tiistai 4. marraskuuta sunnuntai 3. marraskuuta sunnuntai 2. marraskuuta lauantai 1. marraskuuta

Kuten huomataan, niin ensimmäisen esimerkin teksti ei muodosta kokonaisia lauseita. Turun NLP:n annotointiohjeissa lukee seuraavalla tavalla: ” The purpose of rejecting documents is to focus the annotations on full texts. Reject when - - - the sentences don’t form a coherent text”. Kyseisessä esimerkissä ei ole koherentteja lauseita vaan teksti koostuu vain luettelosta päivämääriä. Tällaiselle tekstille ei ole mahdollista eikä järkevää antaa rekisteriluokkaa. Kun avataan nettisivu, josta kyseinen teksti on saatu, huomataan, että todellisuudessa kyseessä on lista linkkejä uutisteksteihin ja kunkin päivämäärän alla on linkkejä http://bestnews.lv/blog/2014-11.

Toinen esimerkki hylätystä tekstistä kuvaa kuinka teksti sisältää kieltä, joka ei ole kohdekieli (venäjä). Annotointiohjeiden mukaisesti se pitää myös hylätä, koska koherentin tekstin määrä on pieni verrattuna mitä kaikkea muuta tekstissä on. Tästä huomataan, että annotointiohjeissa on huomioitu monenlaisia erilaisia tilanteita, jolloin tekstiä ei ole järkevää annotoida eli teksti on järkevämpää hylätä. Usein tällaiselle tekstille olisi jopa mahdotonta määrittää rekisteriluokkaa.

Esimerkki 2:

reject 硫酸沙丁胺醇片 片剂 按C13H21NO3计算 2mg Китай - китайский - CFDA (药监局 - 中国食品和药物管理局) Купи это сейчас Некоторые документы для этого продукта в настоящее время недоступны, вы можете отправить запрос в нашу службу поддержки, и мы сообщим вам, как только мы сможем их получить. Послать запрос. Доступна с: 广西金嗓子药业股份有限公司 ИНН (Международная Имя): Salbutamol Sulfate Tablets дозировка: 按C13H21NO3计算 2mg Фармацевтическая форма: 片剂 Количество Авторизация: H45020187 Дата Авторизация: 2015-08-24 Поиск оповещений, связанных с этим продуктом Поделитесь этой информацией

Esimerkin 2 suomennos:

hylätty **硫酸沙**鄤剁胺醇/h4 > Kiina - kiina - CFDA (药监局 - **中国食品和**药物管理局) Osta nyt Joitakin tämän tuotteen asiakirjoja ei ole tällä hetkellä saatavilla. Voitte lähettää pyynnön tukitiimillemme ja ilmoitamme sinulle heti, kun asiakirjat ovat saatavilla. Lähetä pyyntö. Saatavilla: **广西金嗓子**药业股份有限公司 INN (kansainvälinen nimi): Salbutamolisulfaattitabletit annostus: **按**C13H21NO3计算 2mg Lääkemuoto: **片**剂 Valtuutus: H45020187 Päivämäärä: 24.8.2015 Hae tähän tuotteeseen liittyviä ilmoituksia Jaa nämä tiedot

Kyseessä on ilmeisesti sivu, jolta on poistettu tietoa. Joka tapauksessa annotoinnin kannalta teksti on luokiteltu hylätyksi, sillä se sisältää vain irrallisia sanoja ja lauseita, eikä muodosta järkevää ja koherenttia kokonaisuutta, joka olisi mielekästä tai aiheellista luokitella.

5.1.2 Konekäännetyt tekstit

Konekäännettyjä tekstejä ei joko ollut venäjänkielisessä aineistossa tai sitten niitä ei tunnistettu. On mahdollista, että konekäännettyjä tekstejä venäjänkielisessä aineistossa ei tunnistettu, sillä en ole äidinkieleltäni venäjänkielinen. Venäjänkielinen aineisto on pieni, joten on myös mahdollista, että kyse on sattumasta, ettei aineistoon sattunut yhtään konekäännettyä tekstiä. Todennäköisempää voisi kuitenkin olla, että konekäännettyjä tekstejä ei välttämättä ole venäjänkielisessä internetissä yhtä paljon kuin ehkä pienemmissä kielissä. Tutkimustulos voisi kertoa siitä, että venäjänkielisessä internetissä konekäännettyjä tekstejä ei ole yhtä suurissa määrin kuin suomenkielisessä aineistossa, mikä olisi loogista, kun vertaa suomen kielen ja venäjän kielen puhujamääriä.

Kotimaisten kielten keskuksen eli KOTUKSEN mukaan äidinkielenä Suomessa suomea puhuu noin 4,9 miljoonaa ja toisena kielenä yli puoli miljoonaa ihmistä, kun taas venäjän kieltä puhuu äidinkielenä noin 161,7 miljoonaa ihmistä ja toisena kielenä noin 110 miljoonaa ihmistä. On siis loogista, että venäjänkielinen internet on paljon suurempi, eikä siellä esiinny yhtä paljoa konekäännettyä aineistoa kuin puhujamääriltään pienemmissä kielissä. Yksi selitys voisi myös olla, että puhujamääriltään suuremmissa kielissä konekääntimet toimivat paremmin kuin pienissä kielissä, joten annotoidessa tekstejä sekä ihmisten että koneen on vaikeampi tunnistaa konekäännettyjä tekstejä tekstien joukosta. Seuraavat esimerkit ovat siis suomenkielisestä aineistosta.

Esimerkki 3:

```
Show other languages MBA sanoista Master in Business Administration ja
koostuu ohjelman opetussuunnitelma , joka tarjoaa perustan keskeiset
käsitteet liiketoiminnan , mukaan lukien hallinto , rahoitus , talous ,
myynti , markkinointi , henkilöstöjohtaminen ja toimitusketjun hallintaan .
MBA in Management on niille , jotka haluavat laajentaa heidän
urakehitystään . Opiskelijat tässä MBA tietenkin saada yleiskuva
liikkeenjohdon ja päästä tutkimaan kaikkia tärkeitä näkökohtia
liiketoiminnan kuten rahoitus , markkinointi , ryhmien ja organisaatioiden
dynamiikkaa , henkilöresurssit , liiketoiminnan etiikan ja sosiaalisen
vastuun , liikejuridiikan sekä strategia-ja kehitysjohtaja . Etäopetus tai
```

etäopetus on tapa tuottaa koulutusta ja opetusta , usein tapauskohtaisesti , opiskelijoille , jotka eivät ole fyysisesti läsnä kampuksella . Koulutus Yhdysvallat antaa pääasiassa julkisen sektorin kanssa ohjaus ja rahoitus tulevat kolme tasoa : valtion , paikallisten ja liittovaltion , tässä järjestyksessä .Yhteiset vaatimukset opiskella korkeakoulutasolla Yhdysvalloissa sisältää oman tunnustukset essee (tunnetaan myös lauselma tarkoituksesta tai henkilökohtainen lausuma) , opintosuoritusote , suositus / kirjaimilla ja kielikokeet West Lafayette on kaupunki Indiana rankattu väkirikkain kaupunki Indiana jossa asuu 29000 asukasta . Se isännöi Purduen yliopistosta , joka on arvostettu laitos ilmoittautumalla yli 40000 opiskelijaa . Etäopiskelu MBA Hallinto West Lafayette - Suorita MBA-tutkinto West Lafayette . Säästä aikaa ja ota yhteyttä kouluun suoraan tästä !

Esimerkki 4:

MT OS AWO valitsee uuden puheenjohtajan ja varapuheenjohtaja Jäsenet amerikkalaisen Vesiväylät Operaattorit (AWO) , kansallinen kaupan yhdistys Yhdysvaltain hinaaja , towboat ja proomu teollisuuden valittiin Timothy J. Casey puheenjohtajana ja George Foster varapuheenjohtajana perjantaina , huhtikuu 3 aikana AWO kevään yleissopimuksen Washington . Casey puheenjohtaja ja toimitusjohtaja K-Sea Kuljetus Oy , jonka pääkonttori sijaitsee East Brunswick , New Jersey , ja se varapuheenjohtaja , että viimeisen vuoden aikana . Foster puheenjohtaja JB Marine Service , Inc. , jonka pääkonttori sijaitsee St. Louis . Hänen selvitysosa että AWO Hallitus hänen valittiin puheenjohtaja , Mr Casey sanoi , että keskeiset painopistealueet seuraavan vuoden kuuluu varmistaa uusien hinaus aluksen tarkastus ohjelma kehittää asianmukaisesti . " Olemme työskennelleet kovasti kanssa rannikkovartiosto on kehittää järkevä , käytännöllinen ja ainutlaatuinen lähestymistapa aluksen tarkastus " , hän sanoi .Muita haasteita hän mainittu sisältää estää merenkulkijoiden ottaen tehdä toisen

matkan yhteen ilmoittautumisohjeita keskustaan saadakseen niiden Kuljetusalan Työntekijöiden tunnistaminen Credential ja kukistamiseksi ehdotuksen liittovaltion budjetti on sulkumaksu vero . Casey on toiminut puheenjohtaja , toimitusjohtaja ja johtaja K-Sea heinäkuusta 2003 lähtien . Aiemmin hän toimi puhemies , Chief Executive Officer ja johtaja EW Transportation LLC huhtikuussa 1999 . Casey liittyi EW Transportation LLC: n edeltäjä lokakuussa 1988 sen valvojan ja oli sen Chief Financial Officer vuodesta 1996 huhtikuuhun 1999 . Ennen , että Casey oli ohjain New York Hinaus Corp. , yhteisyrityksen kanssa Zapata Persianlahden Marine . Vuodesta 1982 vuoteen 1987 hän työskenteli Zapata Persianlahden Marine toimiin , liikenteen ja rahoitus . Mr.Casey on suorittaneet alemman of Arts in Business Administration / Finance alkaen Texas A & M yliopisto ja Master of Science in Transportation alkaen State University of New York , Maritime College . Hän on hallituksen jäsen merimiesten kirkko instituutti . K-Sea Transportation Corp . on suurin coastwise säiliöproomulla toimijalle , Yhdysvallat . Yhtiö tarjoaa öljytuotteiden meren kuljetus- , jakelu-ja logistiikkapalvelujen Yhdysvaltojen kotimaan meriliikenteen liiketoimintaa . Se on laaja asiakaskunta myös suurten öljy-yhtiöiden , öljy kauppiaiden ja jalostamot . K-Sea toimii paikkakunnalla New York , Philadelphia , Norfolk , Seattlen ja Honolulu .

Esimerkit tunnistaa helposti konekäännöksiksi selvien virheiden takia. Jotkin kohdat eivät myös ole kääntyneet ollenkaan. Rekisterin kieli ei myöskään ole sujuvaa eikä johdonmukaista. Se on myös suurilta osin vaikeasti ymmärrettävää tai täysin epäselvää.

5.1.3 Kerronnallinen rekisteri

Tässä kappaleessa käsitellään kerronnalliseen rekisteriin luokiteltuja tekstejä sekä suomenkielisestä että venäjänkielisestä aineistosta. Erityisesti tarkastellaan tärkeintä kerronnallisen rekisterin alaluokkaa eli uutista tekstinä. Suomenkielisestä aineistosta kerronnalliseen rekisteriin kuuluu 17,8 % teksteistä. Venäjänkielisessä aineistossa sama luku

on 20,9 % eli suunnilleen samaa luokkaa, kun rekisteriä verrataan koko tekstien määrään molemmissa aineistoissa.

Seuraavat esimerkit kuvaavat kyseisen rekisteriluokan kaikkein tyypillisintä tekstiä eli uutista. Suomenkielisessä aineistossa 1359 eli yhteensä 71,0 % kerronnallisen rekisterin teksteistä on uutisia. Sama luku venäjänkielisen aineiston osalta on 71 tekstiä eli yhteensä 62,8 % kerronnallisen rekisterin teksteistä on uutisia. Tämä tarkoittaa sitä, että myös venäjänkielisen aineiston osalta suurin osa kerronnallisen rekisterin teksteistä on uutisia. Tämän vuoksi kerronnallinen rekisteri ja juuri uutinen rekisterin alaluokkana on hyvä keino tarkastella sellaista rekisteriä, jossa ei ole kovin suuria jakaumaeroja suomen ja venäjän kielen välillä.

Annotointiohjeen mukaisesti uutinen on teksti, jonka on kirjoittanut toimittaja ja se on julkaistu uutistoimiston sivustolla – myös sääennustukset. Uutisia ovat myös yritysten ja yhdistysten omat uutiset sekä uutiskirjeet. Tekstin tarkoituksena on kertoa viimeaikaisista tapahtumista ja se on tyypillisesti ammattimaisesti kirjoitettu ja julkaistu mahdollisimman nopeasti tapahtumasta.

Esimerkki 5:

NE NA Suomi on EU:n kilpailukykyisin valtio – Miltä näyttää 10 valtion kärki? Maailman talousfoorumin raportin mukaan Suomi on EU:n kilpailukykyisin valtio. Suomi on Euroopan unionin kilpailukykyisin valtio, selviää Maailman talousfoorumin eilen julkaisemasta raportista. Suomen vahvuuksina ovat raportin mukaan korkea koulutustaso, yhteiskunnan tasa-arvoisuus sekä innovaatioihin panostaminen. Suomi otti Ruotsin paikan, joka putosi toiselle sijalle. Totuttuun tapaan kaikki Pohjoismaat pärjäsivät hyvin. Heikoimmin menee Bulgarialla, Romanialla ja Kreikalla. Raportin mukaan maat, joilla on vahvat sosiaaliset turvaverkot pärjäävät paremmin kuin eriarvoisemmat yhteiskunnat. Huolta herättää eriarvoisuuden lisäksi EU:n maiden välille aukeava innovaatiokuilu: patenttihakemusten määrä on Pohjoismaissa 16 kertaa korkeampi kuin muualla Euroopassa. Raportissa tarkastellaan Eurooppa 2020-strategiaksi nimetyn EU:n kasvustrategian tavoitteiden edistymistä. Strategiassa keskeisessä osassa ovat tietoon ja innovaatioihin perustuva talouskasvu. Kymmenen kilpailukykyisimmän valtion

kärki näyttää tältä: 1. Suomi 2. Ruotsi 3. Hollanti 4. Tanska 5. Saksa 6. Itävalta 7. Iso-Britannia 8. Luxemburg 9. Belgia 10. Ranska

Esimerkki 6:

NE NA Uutiset Kansalaisaloite tasa-arvoisen avioliittolain puolesta on kerännyt viimeisen kuuden kuukauden kuluessa jo yli 155 000 kannatusilmoitusta. Määrä on kolminkertainen eduskunnan asettamaan nimien vähimmäiskeruumäärään nähden, joten aloite tullaan käsittelemään suuressa salissa lähikuukausina. Tahdon2013-kampanja järjesti tiistai-iltana 3. syyskuuta verkossa avoimen kyselytunnin, jonka aikana kampanjan puheenjohtaja Senni Moilanen ja poliittisen vaikuttamisen koordinaattori Milla Halme vastasivat tukijoilta ja kampanjasta kiinnostuneilta tulleisiin kysymyksiin . Nämä miesmuusikot eivät juuri esittelyjä kaipaa! Samuli Putro on laulanut tiensä suomalaisten sydämiin Zen Café -yhtyeen kitaristina ja julkaissut myös kolme suuren suosion saavuttanutta sooloalbumia. Kannatusilmoitusten keruuaikaa on jäljellä alle kuukausi. Jo kampanjan ensimmäisen vuorokauden aikana saama valtava kannatusilmoitusten vyöry takasi sen, että kansalaisaloite tasa-arvoisesta avioliittolaista tullaan käsittelemään eduskunnassa . Silti päätimme, että kuuden kuukauden mittainen kannatusilmoitusten keräysaika käytetään kokonaisuudessaan. Seksuaaliselta suuntautumiseltaan ja sukupuoleltaan moninaisten ihmisoikeuksia poljetaan Venäjällä rajusti. Maailmalla seurataan MM-kisojen ohella myös tiiviisti maan ihmisoikeustilannetta. Vääräksi kokemansa sukupuolen korjaamaan onnistunut Sasha on kuitenkin elämäänsä onnellinen. Laura Birn on valmistunut Teatterikorkeakoulusta vuonna 2007. Hänet tunnetaan erityisesti elokuvarooleistaan. Tänä vuonna Laura nähdäänkin kotimaisissa elokuvissa Mieletön elokuu ja Leijonasydän. Hänet tullaan näkemään myös amerikkalaisessa A Walk Among the Tombstones -rikoselokuvassa.

Esimerkki 7:

Комитет по транспорту, организации дорожного движения и развития улично-дорожной сети наладил всесторонний контроль над работой общественного транспорта. Частные автоперевозчики - основные нарушители требований к качеству услуг, которые не всегда соблюдают условия заключенных договоров. Кроме того, МКУ «Муниципальная транспортная инспекция» реализует особый план работы, направленный на выявление и пресечение деятельности автоперевозчиков, оказывающих услуги незаконно. Инспекторы совместно с сотрудниками правоохранительных органов выписывают нарушителям административные штрафы. Как прокомментировал телеканалу «АТН» заместитель председателя Комитета по транспорту, организации дорожного движения и развитию улично-дорожной сети Сергей Яскевич: «Сегодня в Екатеринбурге несколько транспортных компаний незаконно занимаются перевозом пассажиров, за что не раз были привлечены к административной ответственности. К сожалению, после их ухода на маршрут выходят новые недобросовестные предприниматели, которым, также будут выписаны штрафы». Специалисты Комитета отмечают, что качество подвижного состава нелегальных автоперевозчиков невозможно проверить, что сказывается на уровне безопасности предоставляемой услуги. Муниципальная транспортная инспекция создана в целях сбора и анализа информации о качестве транспортного обслуживания в Екатеринбурге. Источник: Официальный портал Екатеринбурга

Esimerkin 7 suomennos:

Liikenteen ja tieverkon kehittämisen valiokunta on perustanut kattavan valvonnan julkista liikennettä varten. Yksityiset kuljettajat ovat palvelun laatuvaatimusten tärkeimpiä rikkojia, eivätkä he aina noudata tehtyjen sopimusten ehtoja. Lisäksi kunnan liikennetarkastusvirasto on laatinut erityisen suunnitelman, jonka tarkoituksena on tunnistaa ja tukahduttaa laittomasti palveluja tarjoavien kuljettajien toiminta. Tarkastajat antavat yhdessä lainvalvontaviranomaisten kanssa sakkoja sääntöjen rikkojille.

Liikenne ja tieverkkojen kehittämisen valiokunnan varapuheenjohtaja Sergej Jaskevits kommentoi ATN-televisiokanavalle: ”Nykyään useat Jekaterinburgin kuljetusyritykset kuljettavat matkustajia laittomasti, joista heidät on saatu vastuuseen useammin kuin kerran. Valitettavasti heidän lähdettyään tulee aina uusia, häikäilemättömiä yrittäjiä samoille reiteille ja he saavat myös sakot.” Komitean asiantuntijat huomauttavat, että laittomien tieliikenteen harjoittajien liikkuvan kaluston laatua ei voida tarkistaa, mikä vaikuttaa tarjotun palvelun turvallisuustasoon. Kunnan liikennetarkastusvirasto perustettiin keräämään ja analysoimaan tietoja kuljetuspalvelujen laadusta Jekaterinburgissa. Lähde: Jekaterinburgin virallinen sivusto

Esimerkki 8:

Петарда взорвалась в кармане школьника в лицее «Политэк»: его доставили в травмпункт ВолгодонскаФото: pixabay.comЧитайте также:Отдел муниципальной инспекции администрации Волгодонска возглавил Александр Бугай (25.11.2020 16:59)Три пациента скончались в ковидном госпитале Волгодонска за сутки (25.11.2020 12:56)За сутки у восьми волгодонцев подтвержден коронавирус (25.11.2020 11:12)25 ноября в Волгодонске Ростовской области произошел инцидент с участием школьника. 13-летний учащийся лицея «Политэк» пострадал от взрыва петарды.Как удалось выяснить «Блокноту», петарда разорвалась на перемене прямо в кармане ребенка. По предварительной информации, взрыв произошел из-за неудачной шутки одноклассника пострадавшего. Случившееся повергло в шок школьников, находившихся поблизости. Сначала детей испугал неожиданный звук взрыва, а после - кровотечение и ужас на лице ребенка.Как стало известно «Блокноту», в результате взрыва петарды школьник получил незначительную рваную рану в области бедра. Сразу после инцидента ребенок был доставлен на «скорой» в травматологический пункт БСМП, где ему зашили рану. В настоящий момент пострадавший находится в отделении детской хирургии. По словам источника, жизни и здоровью школьника ничего не

угрожает, к счастью, полученные травмы оказались не серьезными, и через пару дней он отправится домой.Напомним, в начале ноября еще один волгодонский школьник играл в опасные игры с петардами у сквера «Машиностроителей». Жительница Волгодонска стала свидетельницей, как школьник, стоя на остановке, поджигает петарды и закидывает их под колеса общественному транспорту. Причем, юный волгодонец, закинув петарду, мигом убегает с места происшествия. Водители попросту не успевают сделать даже замечание. Увидев опасные игры школьника, женщина сделала ему пару замечаний. Однако тот в ответ поджег петарду и кинул под ноги ее двухгодовалому ребенку.Ирина ЛитвиноваНовости на Блокнот-ВолгодонскНовости на Блокнот-Волгодонск

Esimerkin 8 suomennos:

Sähikäinen räjähti koulupojan taskussa Politekin lyseossa: oppilas vietiin Volgodonskin ensiapuun. Kuva: Pixabay.com Lue myös:Volgodonskin hallinnon kunnan tarkastusosastoa johti Alexander Bugaj (25.11.2020 16:59) Kolme potilasta kuoli vuorokauden aikana Volgodonskin sairaalassa koronaan (25.11.2020 12:56) Vuorokauden akana kahdeksalla volgodonskin asukkaalla vahvistettiin koronavirustartuna (25.11.2020 11:12)

Vahinko tapahtui 25. marraskuuta Volgodonskin koulussa, Rostovin alueella. Politek-lyseon 13-vuotias oppilas loukkaantui sähinkäisen räjähdyksessä. Sähinkäinen räjähti suoraan lapsen taskussa. Alustavien tietojen mukaan räjähdys tapahtui uhrin luokkatoverin epäonnistuneen pilan vuoksi. Tapaus järkytti lähistöllä olleita oppilaita. Ensin lapset pelästyivät räjähdyksen kovan äänen vuoksi ja sen jälkeen lapsen kasvoilla näkyvän kauhun ja verenvuodon takia. Räjähdyksen syy ei ole tiedossa, mutta sähinkäisen räjähdyksen seurauksena oppilas sai reiteen pienen repeämän. Välittömästi tapahtuman jälkeen lapsi vietiin ambulanssilla BSMP: n traumakeskukseen, jossa hänen haavansa ommeltiin. Uhri on tällä hetkellä lastenkirurgian osastolla. Lähteen mukaan oppilaan henki tai terveys ei ole välittömässä

vaarassa, sillä onneksi vammat eivät olleet vakavia. Oppilas pääsee kotiin luultavasti parin päivän kuluttua. Muistutamme lukijoita siitä, että marraskuun alussa toinen Volgodonskin koulupoika leikki vaarallisia leikkejä sähinkäisten kanssa lähellä Mashinostroitelej-aukiota. Volgodonskin asukas todisti bussipysäkillä seisovan koululaisen sytyttämässä sähinkäisiä tuleen ja heittävän niitä joukkoliikenteen alle. Lisäksi Volgogradin nuori asukas, joka heitti sähinkäisen, pakeni välittömästi paikalta. Kuljettajilla ei yksinkertaisesti ollut aikaa edes huomauttaa asiasta. Nähdessään oppilaan vaaralliset leikit, tempun todistanut asukas puuttui asiaan. Vastauksena oppilas kuitenkin sytytti sähinkäisen ja heitti sen asukkaan kaksivuotiaan lapsen jalkojen alle. Irina Litvinova Uutisia BlackNot-VolgodonskNewsissa Blacknot-Volgodonskissa

Kuten esimerkeistä huomataan, niin uutistekstien kieli on kielestä huolimatta hyvin asiallista, eikä sisällä esimerkiksi puhekielisyyttä kummallakaan kielellä. Esimerkeistä nähdään myös, että suomenkielisessä asiatekstissä lauseet pyritään pitämään melko lyhyinä ja pilkkomaan tekstiä helppolukuisemmaksi, kun taas venäjänkieliselle asiatekstille on tyypillistä todella pitkät lauserakenteet.

5.1.4 Informatiivinen suostuttelu

Informatiivisen suostuttelun rekisteriin kuuluvien tekstien määrä eroaa suomen ja venäjän välillä merkittävästi. Mielenkiintoista on venäjänkielisten tekstien suuri määrä informatiivisen suostuttelun rekisterissä, kun taas suomenkielisessä aineistossa rekisteri ei erotu joukosta yhtä merkittävästi. Suomenkielisessä aineistossa sen sijaan informatiivisia kuvauksia oli suhteessa hieman enemmän venäjänkieliseen aineistoon verrattuna.

Esimerkissä 9 on informatiivisen suostuttelun rekisterin myyntikuvauksen alaluokkaan luokiteltu teksti.

Esimerkki 9:

Пигмент для татуажа век Color King Сажа №818 Черный. Основа пигментов-глицерино-сорбитоловая. Палитра Color King состоит из 23 оттенков. Объем 10мл. Пигмент имеет низкую усадку цвета. Стойкость цвета от 2 лет. Цвета пигмента очень насыщенные и стойкие к выгоранию. Пигменты Color King изготовлены с добавлением в состав натуральных экстрактов. Благодаря этому пигменты обладают регенерирующим, восстанавливающим и противовирусным действием. Еще одно преимущество этих пигментов в том, что они устойчивы к ультрафиолетовому излучению. Не требует добавления корректора (исключение черный пигмент). Осуществляем доставку по Москве и России. Оплата при получении. Доставка в регионы курьерской службой. Ссылка на палитрутут Подробная инструкция по колористике пигментов по этойссылке

Esimerkin 9 suomennos:

Tatuoinnin väri on musta Color King Sasha nro 818. Väriaineen pohjana on käytetty glyseroli-sorbitoliseosta. Color King -paletti koostuu 23 sävystä. Tilavuus 10 ml. Väri kutistuu erittäin vähän. Värin kestävyys on 2 vuotta. Pigmenttivärit ovat erittäin vivahteikkaita, eivätkä ne haalistu Color King -värit valmistetaan lisäämällä luonnollisia uutteita koostumukseen. Tästä johtuen pigmenteillä on uudistuvia, korjaavia ja immunologisia vaikutuksia. Näiden pigmenttien toinen etu on, että ne ovat UV-suojattuja, eivätkä ne vaadi korjausta (mustan pigmentin lisäämistä). Toimitamme Moskovaan ja Venäjälle. Maksu toimituksen jälkeen. Toimitus alueille toimii kuriiripalvelulla. Linkki palettiin. Yksityiskohtaiset ohjeet pigmenttien värjäämiseen tällä linkillä.

Esimerkki 10:

Kysy tarjousta Tilaa uutiskirje Yritys 360 Event Service on yksi harvoista täyden mittakaavan palvelua tarjoavista tapahtumakalusteiden ja -tekniikan

vuokrausfirmoista Baltiassa ja Pohjoismaissa . Tarjoamme asiakkaillemme kokonaispalvelua , jolla varmistetaan yritys- ja yleisötilaisuuksien tekninen onnistuminen . Osallistumme tilasuunnitteluun , toimitamme esiintymislavat , teltat , kalusteet , somistustarvikkeet , korkealaatuisen AV- ja valaistustekniikan sekä paljon muuta – toiveiden ja tilauksen mukaan paikan päälle kuljetettuna ja käyttövalmiiksi asennettuna . Asiantuntevan henkilökuntamme kokemuspohja on yli 14 vuoden mittainen . Vuodesta 1998 alkaen olemme osallistuneet lähes 3000 tilaisuuden toteuttamiseen – niiden joukossa mm. Eurovision laulukilpailut Tallinnassa ja Flow Festival Helsingissä . Tapahtumareferenssilistamme puhukoon omaa kieltänsä . Vastaamme jokaisesta tekemästämme työstä selkeästi mustaa valkoisella - allekirjoituksin . Se takaa maksimaalisen laadun ja palvelukokemuksen .

Myyntikuvauksessa käytetään kuvailevaa kieltä ja termejä. Suomenkieliset lauseet ovat todella lyhyitä verrattuna venäjänkieliseen asiatekstiin.

5.1.5 Informatiivinen kuvaus

Informatiivisen kuvauksen rekisteriin kuuluvat tekstit, jotka kertovat tai kuvailevat jotain informatiivisesti ja niiden tavoitteena on tuottaa puolueetonta tietoa. Tällaisia tekstejä ovat esimerkiksi tietosanakirja-artikkelit, tieteelliset artikkelit, henkilö- ja asiakuvaukset, vastaukset usein kysyttyihin kysymyksiin, lainopilliset tekstit ja muut informatiiviset kuvaukset. On hyvä huomioida, että esimerkiksi tieteelliset artikkelit ovat usein erillisten tietokantojen takana vapaan internetin sijaan ja tämän tutkielman aineistojen tekstit on kerätty nimenomaan vapaasti internetissä olevista teksteistä.

Venäjänkielisessä aineistossa informatiivisen suostuttelun rekisteriin kuuluvia tekstejä on huomattavasti suurempi osuus kuin suomenkielisessä aineistossa. Venäjänkielisessä aineistossa informatiivisen kuvauksen rekisteriin kuuluu 11,6 % ja suomenkielisessä aineistossa vastaava luku on 16,2 %. Jos verrataan informatiivisen suostuttelun rekisteriin, niin venäjänkielisistä teksteistä siihen kuuluu 46,4 % teksteistä ja suomenkielisessä aineistossa 13,7 % teksteistä. Mielenkiintoista onkin, että informatiivisen kuvauksen rekisteriluokkaan kuuluvia tekstejä on suhteessa enemmän suomenkielisessä aineistossa kuin

venäjänkielisessä aineistossa. Tutkimustulos voisi viitata tapaan kertoa asioista eli että venäjänkielisessä internetissä olisi huomattavasti todennäköisempää kohdata informatiivisen suostuttelun rekisteriin kuuluva teksti kuin informatiivisen kuvauksen rekisteriin kuuluva teksti. Suomenkielisessä aineistossa sen sijaan on hieman todennäköisempää kohdata teksti, joka on informatiivinen kuvaus eikä suostuttelu.

Esimerkki 11:

Значение слова Шаба, посад Бессарабской губернии по словарю Брокгауза и Ефрона:Шаба, посад Бессарабской губернии (Шабалат) — посад Аккерманского уезда Бессарабской губернии, в 7 верстах от уездного города, на западном берегу Днестровского лимана. Посад основан швейцарскими колонистами (французами-кальвинистами), прибывшими сюда в 1824 г. по приглашению русского правительства. В 1841 г. колония наименована посадом, которому отведено 4074 десятины земли (по большей части весьма песчаной). Колонисты занялись разведением привезенных с собой виноградных лоз. Вначале этому сильно мешали сыпучие пески; впоследствии они укреплены посадкой акаций и удобрением. Со временем возникла рядом русская Ш. — колония, которой отведено правительством 3000 десятин земли для виноградарства. В настоящее время обе колонии слились. 3894 жителя; лютеранская церковь, 2 школы, несколько лавок. Благодаря мелиорации, ценность земли возросла до того, что в 1890 г. шабская дума продавала землю по 1100 руб. за десятину. В последние годы в Ш. развивается курорт для лечения виноградом и купаньем в лимане, выстроено много дач. О виноделии в Ш. — см. соотв. статью.

Esimerkin 11 suomennos:

Mitä Shaaba tarkoittaa? Bessarabian maakunta-alueen määritelmä Brokhausin ja Efronin sanakirjan mukaan: Shaba, Bessarabian maakunnassa (Shabalat) on Bessarabian maakunnan Akkermanin piirin asutusalue, joka sijaitsee 7 virstan päässä kaupungista, Dniesterin suiston länsirannalla. Alueen

perustivat sveitsiläiset siirtolaiset (ranskalaiset kalvinistit), jotka saapuivat tänne vuonna 1824 Venäjän hallituksen kutsusta. Vuonna 1841 siirtomaa nimettiin asutukseksi, jossa oli 4 074 kymmenystä maata (enimmäkseen hyvin hiekkaista). Siirtolaiset alkoivat kasvattaa viiniköynnöksiä, jotka he olivat tuoneet mukanaan. Viiniköynnösten kasvattamista hankaloitti hiekkainen maaperä ja myöhemmin sitä vahvistettiinkin istuttamalla akaasioita ja lannoittamalla maata. Hieman myöhemmin lähistölle ilmestyi venäläinen siirtomaa, jolle hallitus myönsi 3 000 hehtaarin maapalan viininviljelyä varten. Nykyään nämä alueet ovat yhdistyneet toisiinsa. 3 894 asukasta; luterilainen kirkko, 2 koulua, muutamia kauppoja. Maanparannusten ansiosta maan arvo nousi niin paljon, että vuonna 1890 Shaaban duuma myi maata 1100 ruplalla kymmenystä kohden. Viime vuosina alue on ollut suosittu lomakeskus, jossa rentoudutaan viiniköynnösten hoitamisella ja suistoalueella uimisella ja alueelle on rakennettu paljon kesämökkejä. Viininvalmistuksessa kaupungissa - katso seuraava artikkeli.

Informatiivisen kuvauksen rekisterin tekstit ovat hyvin asiallisia ja puolueettomia. Tekstissä kerrotaan faktoista. Teksti sisältää paljon asiaan liittyvää sanastoa.

5.1.6 Hybriditekstit

Analyysin lopuksi käsitellään vielä hybriditekstejä. Hybriditekstit ovat tekstejä, joille on annettu useampi kuin yksi rekisteriluokka. Tällainen teksti voi olla esimerkiksi teksti, jossa on myyntikuvaus ja arvostelu samassa tai teksti, jossa on kerronnallinen blogikirjoitus sekä resepti (esimerkki 12).

Esimerkki 12:

Category Archives : reseptivihosta " Vaimollani ja minulla on keittäjä " , hän sanoi , " mutta tärkeissä tilanteissa panen mielelläni kokinhatun

päähäni ja muut keittiömestarin varusteet ylleni valmistaakseni jonkin klassisen keittiömestarin tryffeliruokalajin . Omelette aux truffes tai munakokkelia jossa on pieniä tryffeliviipaleita , jotka muuttavat yksinkertaisen ruokalajin harvinaiseksi herkuksi ! Tai truffes fourrées , jolloin tryffelimuhennoksella täytetään ohut rapea taikinakuori , paistetaan uunissa ja tarjotaan madeirakastikkeen kanssa : tai tryffelit sous la cendre , paahdettuina hitaasti hiilloksella rautaisessa hiilipannussa . Entä oletteko koskaan kokeillut sellaista ruokaa kuin truffes au champagne ? Todella helppo valmistaa . Tryffelit pannaan kattilaan , niille kaadetaan kuivaa shampanjaa niin että ne peittyvät , ja keitetään aivan hiljaa kaksikymmentä minuuttia . Lopussa nestettä saisi olla jäljellä enää vain teelusikallinen tryffeliä kohti . Peitetään kevyellä paistostaikinalla ja paistetaan kuumassa uunissa kaksikymmentä minuuttia . ” M . Barbier ummisti silmänsä tätä ihanuutta ajatellessaan . Auto oli vähällä mennä ojaan . Minä onnistuin tarttumaan ohjauspyörään aivan viime hetkessä , mutta hän ei näyttänyt sitä huomaavan . Mittaa kattilaan riisiryynit , vesi ja suola ja kuumenna kiehuvaksi . Keittele miedolla lämmöllä tiiviin kannen alla 10 minuuttia . Sekoita joukkoon kerma ja maito ja hauduta hyvin miedolla lämmöllä kannen alla 30-40 minuuttia , kunnes puuro on valmista . Voitele uunivuoka (meidän on nelikulmainen , 20×25 cm) rasvaa säästämättä . Kalttaa , kuori ja rouhi mantelit ja lisää ne puuroon . Sekoita mukaan sahrami , sokeri tai hunaja , hienonnettu kardemumma ja jauhettu kaneli (mausteita pannaan vain ripaus , sillä maun on tarkoitus tuntua pelkkänä aavistuksena) . Puuron jäähdyttyä on aika sekoittaa mukaan munat . Jatka maidolla , jos taikina tuntuu liian paksulta . Kaada taikina (2-3 sentin kerrokseksi) voideltuun uunivuokaan . Paista 200-asteisessa uunissa 25-30 minuuttia . Tarjoa haaleana kermavaahdon ja hillon tai marjojen kanssa . Voit tarvittaessa lämmittää pannukakkua 175-asteisessa uunissa noin 15 minuuttia . Sahramipannukakun voi myös pakastaa . ”

Toinen esimerkki hybriditekstistä (esimerkki 13) on luokiteltu sekä konekäännetyksi tekstiksi että mielipiderekisterin alle luokitelluksi arvosteluksi.

Esimerkki 13:

Mielipide FLYMO EASICUT 5500 :sta Sen käyttäjät pitivät tuotetta FLYMO EASICUT 5500 hyvin käyttäjäystävällisenäHe pitivät sitä luotettavana . , Enimmäkeen samaa mieltä tässä kohtaa Jos haluat olla varma että FLYMO EASICUT 5500 on ratkaisu ongelmiisi , saat suurinta apua ja tukea toisilta Diplofix käyttäjiltä Keskiarvo pisteet mielipiteiden jakautumisesta on 7.58 ja tavallinen ero on 2.53 Korkea suorituskyky Käyttäjät ovat kysyneet seuraavia kysymyksiä : Onko EASICUT 5500 erittäin suorituskykyinen ? 12 käyttäjät vastaukset kysymyksiin ja tuotteen sijoitukset asteikolla 0-10 . Sijoitus on 10/10 jos FLYMO EASICUT 5500 on toimialallaan paras tekniseltä tasoltaan , tarjoaa parasta laatua tai tarjoaa suurinta sijoitusta ominaisuuksissaan.

6 Yhteenveto

Venäjänkielisen aineiston määrä on melko pieni, jolloin sattuman vaikutus voi olla suurempi. Aineiston tulisi olla suurempi, jotta kaikkien rekisteriluokkien tekstejä olisi tarpeeksi luotettavaan analyysin. Tässä tutkielmassa kuitenkin keskitytään niihin rekisteriluokkiin laadullisessa tarkastelussa, joita on suurempi määrä.

Kuten kielitieteessä harvemmin, luokittelu etenkään tekstien kohdalla ei ole helppoa. Ohjeet ovat selkeät, mutta haasteita asettaa internetistä löytyvien tekstien laaja skaala. Rajat rekistereiden välillä ovatkin todellisuudessa häilyviä. Toiset rekisterit ovat selkeämpiä kuin toiset. Esimerkiksi raja informatiivisen ja mielipidekirjoituksen välillä on joskus häilyvä. Missä menee objektiivisuuden ja subjektiivisuuden raja? Kiinnostavaa olisi myös, miten monen tekstin/sivun taustalla todella on taloudelliset vaikuttimet, sillä tekstillä ja rekisterillä on usein jokin tavoite, johon tekstillä yritetään päästä. Tällaisia tavoitteita voivat olla esimerkiksi tiedon lisääminen, vaikuttaminen tai tavaroiden myyminen. Erityisen kiinnostavaa on suuri ero informatiivisen suostuttelun rekisteriin kuuluvien tekstien osalta. Voisiko tämä kertoa erosta tavassa, jolla asioista kerrotaan Suomessa ja Venäjällä?

Suomessa on jo pitkään kouluissa ja muussa yleissivistävässä opetuksessa painotettu lähdekriittisyyttä sekä objektiivisuutta. Venäjällä samanlaista perinnettä opetuksessa ei ole – ehkä pikemminkin päinvastoin. Venäläiseen kulttuurin ja opetukseen ei kuulu yhtä olennaisena osana asioiden kriittinen tarkastelu, vaan pikemminkin ylemmältä taholta tulevaan tietoon kuuluu suhtautua totuutena. Mielenkiintoista olisi nähdä, näkyykö samanlainen ilmiö suuremmassa aineistossa ja minkälainen vaikutus olisi, jos äidinkieleltään venäläiset annotoisivat venäjänkieliset tekstit.

Suurena haasteena annotoinnissa on, etteivät tutkimuksen rekisterit ole selvärajaisia kuten ylhäältä päin luodussa rekisterijaottelussa. Esimerkiksi objektiivisuuden ja subjektiivisuuden raja on tuttu ongelma kieliteknologian ohella laajemminkin. Annotoinnissa pitää valita, onko teksti objektiivinen ohje vai subjektiivinen neuvo/mielipide ja raja näiden välillä on hyvin häilyvä. Sama pätee muutenkin mielipidekirjoitusten ja narratiivisten tai informatiivisten tekstien suhteen.

On myös mahdollista antaa teksteille useita eri luokkia. Itse kuitenkin kallistuin annotoinnissa pääsääntöisesti vain yhteen luokkaan, niin kuin useat muutkin annotoijat, eli huomattavasti suurin osa teksteistä saa vain yhden luokan siitä huolimatta, että tekstille olisi mahdollista

antaa useampi rekisteriluokka. Joillekin teksteille olisi nimenomaan tarkoitus antaa kaksi eri luokkaa. Esimerkki tällaisesta tekstistä olisi narratiivinen blogikirjoitus, joka sisältää reseptin. Venäjänkielisessä aineistossa ei ollut paljoa sellaisia tekstejä, joille olisi selkeästi kuulunut antaa kaksi eri rekisteriluokkaa.

Jää kuitenkin hieman epäselväksi, miksi venäjänkielinen aineisto jakautui rekisteriluokkiin paljon selvärajaisemmin kuin suomenkielinen aineisto. Voisi tutkia, vaikuttiko tuloksiin enemmän aineistojen kokoero tai oliko sillä merkitystä, että venäjänkielisen aineiston on annotoinut yksi ja sama (ei-äidinkielinen) henkilö. Asian tarkempi selvittäminen antaisi enemmän tietoa siitä, miten paljon eri henkilöiden tekemät annotoinnit eroavat toisistaan.

Avainsana-analyysi rekistereittäin olisi myös mielenkiintoinen aihe tutkia ja se kertoisi vielä laajemmin rekistereistä. Tulevissa tutkimuksissa mielenkiintoinen aihe olisi myös vertailla erikielisten aineistojen jakautumista rekistereittäin ja eri rekistereiden avainsanoja rekistereittäin ja eri kielten rekisterien välillä.

Lähteet

Biber, D. & Conrad, S. (2019). *Registers, Genres, and Styles: Fundamental Varieties of Language.* Cambridge, Cambridge University Press.

Biber, D. & Egbert, J. (2018). *Register Variation Online.* Cambridge, Cambridge University Press.

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus linguistics and linguistic theory.*

Egbert, J. & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora.*

Egbert, J., Biber, D. & Davies, M. (2015). Developing a Bottom-up, User-Based Method of Web Register Classification. *Journal of the association for information science and technology.*

Viittaus: (Egbert ym., 2015)

Laippala, V., Kyllönen, R., Egbert, J., Biber, D & Pyysalo, S. (2019). Toward Multilingual Identification of Online Registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics* https://aclanthology.org/W19-6130.pdf

(Laippala ym., 2019)

Luodonpää-Manni, M., Hamunen, M., Konstenius, R., Miestamo, M., Nikanne, U. & Sinnemäki, K. (2020). *Kielentutkimuksen menetelmiä I-IV.* Helsinki, Suomalaisen kirjallisuuden seura.

Viittaus: Luodonpää-Manni ym. 2020

Mehler, A., Sharoff, S. & Santini, M. (2010). *Genres on the Web: computational models and empirical studies.* Dordrecht, Spinger.

Viittaus: (Mehler ym., 2010)

Mäntynen, A., Shore S. & Solin, A. (2006). *Genre – tekstilaji.* Helsinki, Suomalaisen kirjallisuuden seura.

Liitteet

Liite 1. Annotointiohjeet

Web register annotation guidelines

The annotation task consists of two steps: deciding whether to accept or reject a document, and giving a register label / labels to the accepted documents.

When to accept or reject a document

When to give a document several labels

Please note that

- You can have a look at how the document website looks like by following the document url on the annotator
- The annotation decision should, however, base on the text on the annotator
- Some documents may be followed by a large number of comments. Please do not base your decision on those.

Quickstart

1. Is the web page **Machine translated or generated** from a template?
2. Is the web page **Lyrical**, such as songs or poems?
3. Is the web page originally spoken? (Texts composed of more than 50% spoken quotes classified as spoken)
    - If yes, is it an **Interview**?
    - If no, select **Other spoken** (e.g. formal speeches and TV/movie transcripts)
4. Is the web page **Interactive discussion** written by multiple participants in a discussion format (e.g. discussion or Q&A forum)? (Reader comments following e.g. an article or blog post are NOT included here)
5. Is the purpose of the document to narrate or report on EVENTS? If yes, select one of the following registers:
    - **News report**
    - **Sports report**
    - **Narrative blog** (e.g a travel blog or a personal blog)
    - **Other narrative** (e.g. fictional stories and magazine articles)
6. Is the purpose of the document to explain HOW-TO or INSTRUCTIONS?
    - If yes, is it a **Recipe**?

    - If no, select **Other how-to**. These are typically step-by-step, objective instructions on how to do something.

7. Is the purpose of the document to describe or explain INFORMATION? If yes, select one of the following registers:
    - **Encyclopedia article**
    - **Research article**
    - **Description of a thing or person**
    - **FAQ**
    - **Legal terms and conditions** (including long cookie texts)
    - **Other Informational description** (e.g. course materials and blogs for informing the reader)
8. Is the purpose of the document to express OPINIONS? If yes, select one of the following registers:
    - **Review**
    - **Opinion blog** (typically written by an amateur writer, such as a politician, to express their opinion)
    - **Denominational religious blog / sermon**
    - **Advice** (based on opinion; contrast with how-to text, which express objective instructions)
    - **Other opinion**
9. Is the purpose of the document to describe or explain FACTS WITH INTENT TO PERSUADE or MARKET? If yes, select one of the following registers:
    - Is the text a **Description with intent to sell** a product or service?
    - Is the text a **News & opinion blog or editorial**? These are typically written by a professional writer on a news-related topic, with well-structured argumentation.
    - If not, select **Other informational persuasion**. These are descriptive texts that also sell or promote a service, product or upcoming event, such as a hotel, a smartphone or a football game.

https://turkunlp.org/register-annotation-docs/

Table 2.1. *Visual representation of the key situational distinctions made in the final register framework*

Text	Text can be rated								Cannot rate*
Mode	Originally written							Originally spoken	
Participants	Single author or coauthors (non-interactive)						Multiple participants (interactive)		
Communicative Purpose	To narrate events	To describe information	To express opinion	To use facts to persuade	To explain instructions	To express lyrically			
General Register	Narrative	Info. description/ explanation	Opinion	Info. persuasion	How-to/ instruct.	Lyrical	Interactive discussion	Spoken	
Sub-registers	– News report – Sports report – Personal blog – Historical article – Travel blog – Short story – Novel – Biography – Magazine article – Obituary – Memoir	– Description of a thing – Informational blog – Description of a person – Research article – Abstract – FAQ (informational) – Legal terms – Course materials – Encyclopedia article – Technical report	– Opinion blog – Review – Religious blog/ sermon – Advice – Letter to editor – Self-help – Advertise	– Description with intent to sell – Persuasive article – Editorial	– How-to – Recipe – Instruction – FAQ (How-to) – Technical support	– Lyrics – Poem – Prayer	– Discussion forum – QA forum – Reader responses	– Interview – Transcript – Speech – Script	
Reader comments?									
Spoken quotes?									

"'Not enough text (mostly photos or graphics)" or "Site not found."