Valentina Dani

Morfosyntaktiset, fraseologiset ja verbivalintaan liittyvät virheet tšekinkielisten ja venäjänkielisten suomenoppijoiden teksteissä ICLFI-aineiston perusteella

1 Johdanto

Tarkastelen tässä artikkelissa tšekinkielisten ja venäjänkielisten suomenoppijoiden morfosyntaktisia, fraseologisia sekä verbivalintaan liittyviä leksikaalisia virheitä Kansainvälisen oppijansuomen korpuksen (*International Corpus of Learner Finnish*, ICLFI) aineiston perusteella. Valitsin venäjän- ja tšekinkielisten suomenoppijoiden tekstit, koska molemmat kielet ovat slaavilaisia ja koska ICLFI:n venäjän- ja tšekinkielisten oppijoiden tekstien osakorpukset ovat tarpeeksi suuria lähdekielen vaikutuksen tutkimiseen (Jantunen –Brunni–Lehto–Airaksinen 2014, 67). Tutkimuksessani pyrin vastaamaan seuraaviin tutkimuskysymyksiin:

> 1) Millaisia morfosyntaktisia, fraseologisia sekä verbivalintaan liittyviä virheitä esiintyy tšekin- ja venäjänkielisten suomenoppijoiden B1-tason teksteissä?
>
> 2) Mitkä ovat yleisimmät virheiden makrokategoriat sekä virhekoodit tšekin- ja venäjänkielisten suomenoppijoiden B1-tason teksteissä?
>
> 3) Onko lähtökielen mukaan jaettujen suomenoppijoiden teksteissä eroja morfosyntaktisten, fraseologisten sekä verbivalintaan liittyvien virheiden laadussa ja määrässä, ja onko ero tilastollisesti merkitsevä?

Aineisto koostuu Kansainvälisen oppijansuomen korpuksen tšekinkielisten suomenoppijoiden sekä venäjänkielisten suomenoppijoiden teksteistä Eurooppalaisen kielitaidon viitekehyksen taitotasolla B1 (CEFR = *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*, suomeksi EVK) (Euroopan neuvosto 2001, suomeksi 2003). B1-taso on valittu siksi, että A-tasolla oppijat eivät vielä välttämättä pysty tuottamaan tekstejä, joiden morfosyntaktisia ja fraseologisia ominaisuuksia olisi aiheellista tarkastella (Thewissen 2013, 79). Tutkimus kuuluu korpusavusteiseen oppijankielen tutkimukseen ja suomi vieraana kielenä -alan tutkimukseen.

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

2 Teoreettinen tausta ja aiempi tutkimus

Lähestyn tutkimuskohdettani oppijankielen korpustutkimuksen, virheanalyysin ja lähdekielen vaikutuksen analyysin näkökulmasta. Tutkimukseni on pääosin kvantitatiivinen, vaikka aiempien korpustutkimusten tapaan pyrin analysoimaan tuloksia myös kvalitatiivisesta näkökulmasta.

Lähdekielen vaikutus eli niin kutsuttu siirtovaikutus kuvaa oppijan aiemmin opittujen kielten vaikutusta myöhemmin opittujen kielten omaksumiseen ja käyttöön (Jarvis–Pavlenko 2010, 1). Tutkimuksessani analysoin kahden slaavilaisen kielen puhujien tuottamia tekstejä tutkiakseni slaavilaisen L1:n vaikutusta oppijansuomeen.

Yvonne Breyerin mukaan (2011, 7), kielikorpus koostuu luonnollisesti esiintyvistä kieliaineistoista, joita tallennetaan sähköisesti. Oppijankielen korpukset sisältävät aitoja ei-äidinkielisten kirjoittajien tuottamia tekstejä, joita kerätään eksplisiittisten periaatteiden mukaan ja joista on saatavilla taustatietoja (Granger 2002, 4). Sylviane Grangerin, Gaëtanelle Gilquinin ja Fanny Meunierin (2015, 1) mukaan oppijankielen korpukset ovat edustavia oppijankielen otoksia ja sähköisinä mahdollistavat suurten tekstimäärien analysoinnin sähköisten työkalujen avulla. Oppijankielen korpustutkimus on kielitieteen ala, joka pyrkii yhdistämään korpustutkimusta (*Corpus Linguistics*, CL) toisen ja vieraan kielen tutkimukseen (*Second Language Acquisition Research*, SLA). Ala pyrkii oppijankielen tarkempaan kuvaukseen, josta on hyötyä muun muassa toisen ja vieraan kielen oppimisessa (Granger 2002, 3). Oppijankielen korpukset mahdollistavat ei-natiiviaineistojen vertailun natiiviaineistoon tai, kuten tässä tutkimuksessa, eri lähtökielisten oppijoiden aineistojen vertailun keskenään. Tällöin puhutaan kontrastiivisesta oppijankielen tutkimuksesta (*Contrastive Interlanguage Analysis*, CIA) (Meunier 2021, 27).

Laajojen oppijankielen sähköisten aineistojen avulla on mahdollista tutkia myös virheitä systemaattisesti (*Computer-aided Error Analysis*, CEA). *Virhe* voidaan määritellä miksi tahansa epäonnistuneeksi kielelliseksi tuotokseksi (Díez-Bedmar 2021, 93). Oppijankielen korpukseen voidaan lisätä annotointia (lemmatisointia, kieliopillista ja POS-annotointia, virheannotointia), jonka avulla voi helposti saada tilastollisia tietoja virheiden ja muiden oppijankielen ominaisuuksien määrästä ja laadusta. Virheluokitusta, jota virheannotointiin käytetään, voi muokata sen mukaan, minkälaisia virheitä aineistossa esiintyy.

Vaikka Geoffrey Leechin (1997, 2) mukaan annotointi tehostaa korpusten tutkimuskäyttöä, annotointimenetelmiä on myös kritisoitu, koska annotointiperiaatteet voivat vaikuttaa tutkimuksen tuloksiin ja tutkija saattaa joissakin tapauksissa tehdä johtopäätöksiä annotoinnin (eikä raaka-aineiston) perusteella (Tognini-Bonelli 2001, 73). Leechin (1997, 5) mukaan annotoinnin edut kuitenkin ylittävät haitat, koska annotointi mahdollistaa kielellisten ilmiöiden käsittelyn abstraktilla tasolla sekä antaa useammalle käyttäjälle pääsyn korpusaineistoon. Virheannotoitujen pitkittäisten ja pseudopitkittäisten oppijankielen korpusten avulla tutkijat voivat analysoida tarkkuuden kehitystä siirryttäessä taitotasolta toiselle. *Tarkka kielellinen tuotos* on virheetön kielellinen tuotos, joka noudattaa

kohdekielen sääntöjä (Thewissen 2021, 305). Tarkkuuden analyysi on osoittanut, että vaikka joidenkin virhetyyppien taajuus laskee oppijan siirtyessä ylemmälle taitotasolle, muilla virhetyypeillä saattaa olla U-muotoinen kehitys, jossa virhetyypin taajuus ensin nousee ja vasta sitten laskee.

Kansainväliseen oppijansuomen korpukseen pohjautuu useita oppijankielen tutkimuksia. Sisko Brunni, Jarmo Jantunen ja Valtteri Skantsi (2020) ovat tutkineet kaikkiin kielen tasoihin liittyvien (ortografisten, fonologisten, morfofonologisten, morfosyntaktisten, syntaktisten, leksikaalisten sekä fraseologisten) virheiden esiintymistä ja niiden kehitystendenssejä taitotasoilla A2–B2 ICLFI:n virheannotoidun osakorpuksen perusteella. Olli-Juhani Piri (2017) on tarkastellut korpuspohjaisessa pro gradu -työssään suomenoppijoiden kielitaidon tarkkuuden kehitystä morfosyntaktisten objektivirheiden analyysin avulla. Ilmari Ivaska (2015) on analysoinut edistyneiden suomenoppijoiden akateemista kieltä ja sen ominaisuuksia Edistyneiden suomenoppijoiden korpuksen (LAS2) perusteella. Marianne Spoelman (2013) puolestaan on tutkinut vironkielisten, saksankielisten ja hollanninkielisten suomenoppijoiden partitiivin käyttöä ICLFI-aineistossa.

Tšekinkielisten suomenoppijoiden tuotoksia ovat tutkineet muiden muassa Ivana Kováčová (2017), jonka tutkimus koskee tšekinkielisten suomenoppijoiden essiivin käyttöä ICLFI-aineiston perusteella, sekä Anna Geryšerová (2019), joka on tutkinut tšekinkielisten suomenoppijoiden passiivin käyttöä LAS2-korpuksen avulla.

Venäjänkielisten suomenoppijoiden tuotoksia ovat tutkineet muiden muassa Annekatrin Kaivapalu (2005), jonka tutkimus koskee vironkielisten ja venäjänkielisten suomenoppijoiden suomen kielen monikkomuotojen muodostamista, sekä Veera Virtanen (2011), joka on tarkastellut venäjänkielisten suomi toisena ja suomi vieraana kielenä -oppijoiden perfektin ja imperfektin käyttöä.

3 Aineisto

Kuten edellä mainitsin, tutkimusaineistoni on osa Kansainvälistä oppijansuomen korpusta (ICLFI). Aineistoni koostuu tšekinkielisten ja venäjänkielisten suomenoppijoiden kirjoittamista teksteistä Eurooppalaisen kielitaidon viitekehyksen (EVK) taitotasolla B1 (kynnystaso). Kunkin tekstin tason on määritellyt vähintään kaksi koulutettua arvioijaa. Koska taitotaso on määritelty tekstikohtaisesti, saman opiskelijan eri tekstit saattavat saada eri EVK-arvoja (Jantunen 2011, 96). Niissä tapauksissa, joissa kahden arvioijan tasomäärittelyt poikkeavat toisistaan, tekstin on arvioinut myös kolmas arvioija.

Kaikki oppijat ovat nuoria aikuisia, jotka ovat opiskelleet suomea vieraana kielenä yliopistossa Tšekissä tai Venäjällä, siis kohdekielisen ympäristön ulkopuolella (suomen kielen opiskelusta toisena tai vieraana kielenä ks. esim. Ringbom 1980). Opiskelijat kirjoittivat tekstit pääasiassa kotona ja apuvälineiden (sanakirjan, kieliopin tai oppikirjan) avulla. Jotkut opiskelijat kirjoittivat tekstit luokassa. Pieni määrä kaikista teksteistä koostuu tenteistä. Tekstien aiheet koskevat yleensä opiskelijoiden arkielämää: opiskelijat kirjoittivat muun muassa kirja-arvosteluja, esseitä, kirjeitä fiktiiviselle ystävälle tai kertoivat

perheestään tai unelma-asunnostaan. Venäjänkielisten suomenoppijoiden osakorpuksessa esiintyy myös venäjänkielisten artikkeleiden referaatteja, kun taas tšekinkielisten suomenoppijoiden osakorpuksessa referaatteja ei esiinny. Aineistosta on poistettu kaikki tekstit, joiden kirjoittajat ovat korpuksen metatietojen mukaan oppineet suomen kieltä perheessä tai joiden vanhempien äidinkieli ei ole slaavilainen kieli. Näin pyritään vähentämään erilaisen oppimiskontekstin vaikutusta tuloksiin. Oppimiskontekstiin vaikuttavia tekijöitä on kuitenkin monenlaisia (Mukherjee–Götz 2015, 423), eikä vanhempien äidinkieli ole ainoa tekijä, joka on saattanut vaikuttaa siihen. Aineiston koko kuvataan taulukossa 1.

Taulukko 1. Korpusaineiston jakauma.

	Saneiden määrä	Tekstien määrä	Kirjoittajien määrä
Tšekinkielisten suomenoppijoiden kirjottamat tekstit (CZE)	28 608	178	40
Venäjänkielisten suomenoppijoiden kirjoittamat tekstit (RU)	21 028	102	56

4 Tutkimusmenetelmä

Aineistoon on manuaalisesti lisätty virheannotointi. Virhekoodit pohjautuvat osittain ICLFI:n virheluokitukseen (Brunni–Lehto–Jantunen–Airaksinen 2015, 144) sekä oppijantšekin CzeSL-korpuksen (*Czech as a Second Language with Spelling, Grammar and Tags*) virheluokitukseen (Štindlová–Rosen–Hana–Škodová 2012). Tarve oman virheluokituksen luomiseen johtuu siitä, että tutkimuksessani huomioin pelkästään morfosyntaktiset, fraseologiset ja verbivalintaan liittyvät virheet, kun taas ICLFI:n virhekoodisto kattaa kaikki kielen tasot (Jantunen ym. 2014, 71). Tšekinkielisten suomenoppijoiden tekstit on annotoitu ICLFI:ssä, mutta venäjänkielisten osakorpuksesta virheannotointi puuttui, ja se tehtiin tämän tutkimuksen tarpeisiin manuaalisesti. CzeSL-korpuksen virheluokitus täydentää ICLFI:n annotointia, koska tšekin kieli, kuten suomi, on morfologisesti rikas kieli, josta on olemassa oppijankielen virheannotoitu korpus.

Luomani virhekoodisto kattaa subjektin, objektin, verbin, predikatiivin ja adverbiaalin nollaesiintymät sekä niiden ylimääräiset esiintymät. Tämän lisäksi koodisto kattaa virheet subjektin, objektin, predikatiivin ja adverbiaalin sijavalinnassa. Virhekoodiston avulla huomioin myös kongruenssivirheet, virheet kieltomuodon muodostamisessa ja kieltolauseiden sijavalinnassa, verbiketjuvirheet, virheelliset verbivalinnat, fraseologiset virheet ja selittämättömät virheet. Virheet jakautuvat makrokategorioihin. Jokaiseen makrokategoriaan kuuluu yksi tai useampi virhekoodi. Virheluokitus on esitetty taulukossa 2.

Taulukko 2. Tutkimusaineiston virhekoodisto.

Makro-kategoria	**Makrokategoriaan liittyvät virhekoodit**	**Virhekoodin selitys**	**Makro-kategoria**	**Makro-kategoriaan liittyvät virhekoodit**	**Virhekoodin selitys**
Puuttuva pakollinen lauseenjäsen	<MISSING_SUB>	Puuttuva subjekti	*Ylimääräinen lauseenjäsen*	<EXTRA_SUB>	Ylimääräinen subjekti
	<MISSING_OBJ>	Puuttuva objekti		<EXTRA_OBJ>	Ylimääräinen objekti
	<MISSING_VERB>	Puuttuva verbi		<EXTRA_VERB>	Ylimääräinen verbi
	<MISSING_PRED>	Puuttuva predikatiivi		<EXTRA_PRED>	Ylimääräinen predikatiivi
	<MISSING_OTHER>	Puuttuva adverbiaali		<EXTRA_OTHER>	Ylimääräinen adverbiaali
Lauseen-jäsenen sijavalinta	<SUB_VÄÄRÄ SIJA_OIKEA SIJA>	Virheellinen subjektin sijavalinta	*NP-kongruenssi*	<AGR_HEAD+_DEP->	Kongruenssi-virhe – virheellinen sijavalinta lausekkeen määritteellä
	<OBJ_VÄÄRÄ SIJA_OIKEA SIJA>	Virheellinen objektin sijavalinta		<AGR_HEAD-_DEP+>	Kongruenssi-virhe – virheellinen sijavalinta lausekkeen pääsanalla
	<PRED_VÄÄRÄ SIJA_OIKEA SIJA>	Virheellinen predika-tiivin sijavalinta		<AGR_HEAD-_DEP->	Kongruenssi-virhe – virheellinen sijavalinta molemmilla lausekkeen jäsenillä
	<OTHER_VÄÄRÄ SIJA_OIKEA SIJA>	Virheellinen adverbiaalin sijavalinta	*Verbi-kongruenssi*	<AGR_VERBAL>	Kongruenssi-virhe – verbi ei kongruoi subjektin kanssa
Sijavalinta kielto-lauseessa	<NEG_CASEMARKING>	Virheellinen sijavalinta kielto-lauseessa	*Verbin kieltomuoto*	<NEG_VERBFORM>	Virheellinen kieltomuodon muodosta-minen

Verbiketjut	<VERBCHAIN_FORM>	Verbiketju-virhe – verbiketjun jälkimmäisellä verbillä virheellinen muoto	*Verbivalinta*	<V_LEX>	Leksikaalisesti virheellinen verbivalinta
	<VERBCHAIN_CASE>	Verbiketju-virhe – verbiketjun jälkimmäisellä verbillä virheellinen sijapääte	*Sanaluokka-valinta*	<POS>	Virheellinen sanaluokan valinta
	<VERBCHAIN_MISS>	Verbiketju-virhe – verbiketjun jälkimmäinen verbi puuttuu	*Yhdyssanat*	<COMPOUND>	Yhdyssana-virhe
Syntaktinen rakenne	<SYN_ST>	Virheellinen syntaktinen rakenne (lausetyyppi-virhe)	*Fraseologia*	<PHRASEO>	Fraseologinen virhe
Selittämätön	<UNC>	Selittämätön virhe	*Taivutus*	<INFL>	Taivutus

Annotointiprosessissa on avustanut äidinkieleltään suomenkielinen annotoija. Tarpeen vaatiessa olen konsultoinut Oulun yliopiston alkuperäistä annotointitiimiä, joka annotoi tšekinkielisten suomenoppijoiden osakorpuksen, sekä tarkistanut Kansainvälisen oppijansuomen korpuksen annotointimanuaalin. Alla annan muutaman esimerkin virhekoodiston käytöstä. Esimerkit 1 ja 2 kuvaavat aineiston verrattain yleisiä virhekoodeja (fraseologiset virheet ja virheellinen predikatiivin sijavalinta – oppija on käyttänyt partitiivin sijaan nominatiivia). Esimerkit 3 ja 4 kuvaavat virhekoodeja, jotka esiintyvät aineistossa harvemmin. Esimerkissä 3 oppija on käyttänyt inessiiviä illatiivin sijaan, ja esimerkissä 4 kyseessä on verbiketjuvirhe.

(1) Minä pidän lukemisesta, koska kaikessa kirjassa on viisaus. <PHRASEO> (TS0042a)

(2) Kun työskennellään, on tärkeä oppia levätä. Se ei ole niin helppo, erityisesti nykyisin. <PRED_NOM_PAR> (TS0002i)

(3) Yliopiston jälkeen haluaisimme muuttua toiseen kaupunkiin täi Suomessa. <OTHER_INE_ILL> (VE0076)

(4) Kesällä monta ihmistä tulevat meren rannalle uida ja ottaa aurinkoa. <VERBCHAIN_FORM> (VE0052a)

Informanttikoodin kaksi ensimmäistä kirjainta viittaa oppijan äidinkieleen (TS = tšekki, VE = venäjä), numero viittaa tiettyyn anonymisoituun oppijaan ja lopun kirjain viittaa kyseisen oppijan tiettyyn tekstiin. Esimerkiksi koodi TS0002i viittaa siis anonymisoituun oppijaan 2, jonka äidinkieli on tšekki, ja kyseessä on hänen yhdeksäs aineistoon tallennettu tekstinsä (kirjain i). Jokaisesta tekstissä esiintyvästä virheestä lasketaan yksi piste siihen virhekoodiin, johon virhe kuuluu. Mikäli virheellä on useita tulkintavaihtoehtoja, kuten esimerkissä 5, jokaiseen virhekoodiin lasketaan 0,5 pistettä (mikäli tulkintavaihtoehtoja on kaksi) tai 0,33 pistettä (mikäli tulkintavaihtoehtoja on kolme).

(5) En haikaile mitään, joka kuvailee koko tilanteen ilmapiiri. <OBJ_NOM_PAR / OBJ_NOM_GEN> (TS0002j)

Virhekoodien pistemäärien laskemisen jälkeen kummankin osakorpuksen sanemäärä sekä kunkin virhekoodin pistemäärä syötetään Calc-laskentaohjelmaan (Cvrček 2019). Laskentaohjelmaa käytetään vertailemaan kielen ilmiöiden taajuutta eri aineistoissa. Sen avulla voidaan laskea tietyn ilmiön (tässä tapauksessa virhekoodin) odotuksenmukainen taajuus tietyssä aineistossa toisen aineiston perusteella. Mikäli aineistojen välillä on ero, sen tilastollinen merkitsevyys on laskettu khiin neliö (Chi2) -testin avulla. Chi2-testiä voi käyttää, vaikka tutkittava ilmiö jakautuisi aineistossa epätasaisesti (McEnery–Wilson 2001, 84). Tämä on erityisen tärkeää tämäntyyppisessä korpustutkimuksessa, koska sekä CZE- että RU-osakorpuksessa useammat tekstit ovat peräisin samoilta informanteilta. Tästä syystä ei ole mahdollista olettaa, että kielellinen ilmiö esiintyisi tasaisesti koko aineistossa.

En ole käyttänyt tutkimuksessani potentiaalisten esiintymien analyysia eli kielellisten ilmiöiden (tässä tapauksessa virheiden määrän) vertaamista suhteessa kaikkiin mahdollisiin esiintymiin (Thewissen 2013, 81) siitä syystä, että ICLFI:n automaattinen POS-annotointi ei vastaa virheluokitusta. Automaattisen POS-annotoinnin avulla olisi esimerkiksi mahdollista poimia aineistosta kaikki subjektit ja objektit ja suhteuttaa ne subjekti- ja objektivirheiden määrään, muttei olisi mahdollista käsitellä virheitä, jotka koskevat koko lausetta tai lauseen osaa, kuten virheelliset syntaktiset rakenteet ja fraseologiset virheet.

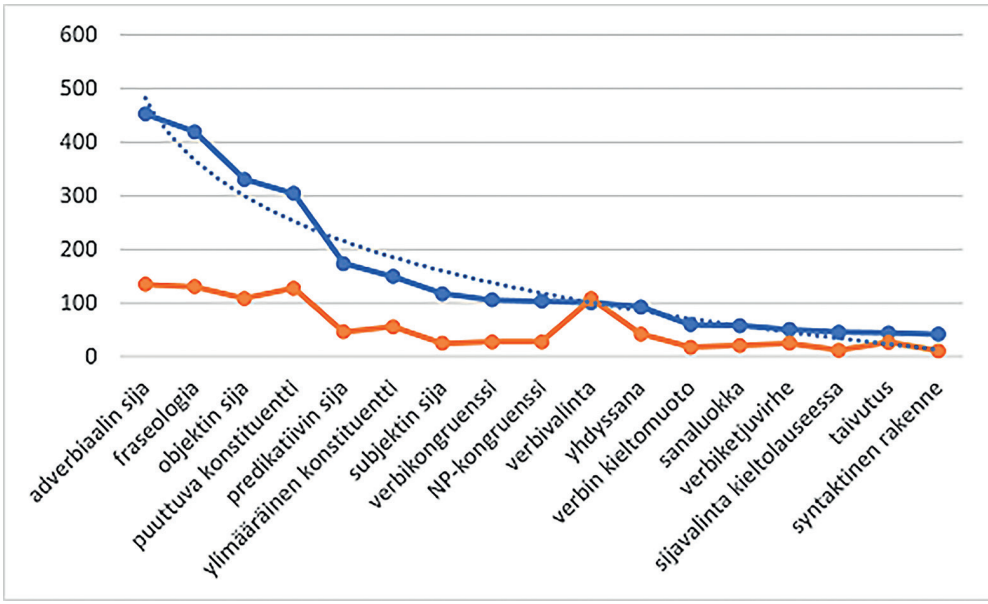
Käytetyn analyysin avulla on mahdollista saada selville, millaiset virheiden makrokategoriat, sekä virhekoodit yleensä, ovat aineistossa yleisimpiä, ja vertailla aineistoja keskenään ottaen huomioon niiden erilaisen sanemäärän.

5 Tulokset

Tässä luvussa esitetään kummankin osakorpuksen yleisimmät virhekoodit, niiden absoluuttiset ja suhteelliset määrät kummassakin osakorpuksessa sekä niiden esiintymistaajuuden tilastolliset erot Chi2-testin avulla. Tšekinkielisten suomenoppijoiden

osakorpuksessa viisi yleisintä virhekoodia absoluuttisen määrän mukaan ovat adverbiaalin sija, fraseologia, objektin sija, puuttuva pakollinen konstituentti (yleensä puuttuva omistusliite) ja predikatiivin sija. Venäjänkielisten suomenoppijoiden osakorpuksessa viisi yleisintä virhekoodia absoluuttisen määrän mukaan ovat adverbiaalin sija, fraseologia, puuttuva pakollinen konstituentti (yleensä puuttuva omistusliite), leksikaalisesti virheellinen verbivalinta ja objektin sija.

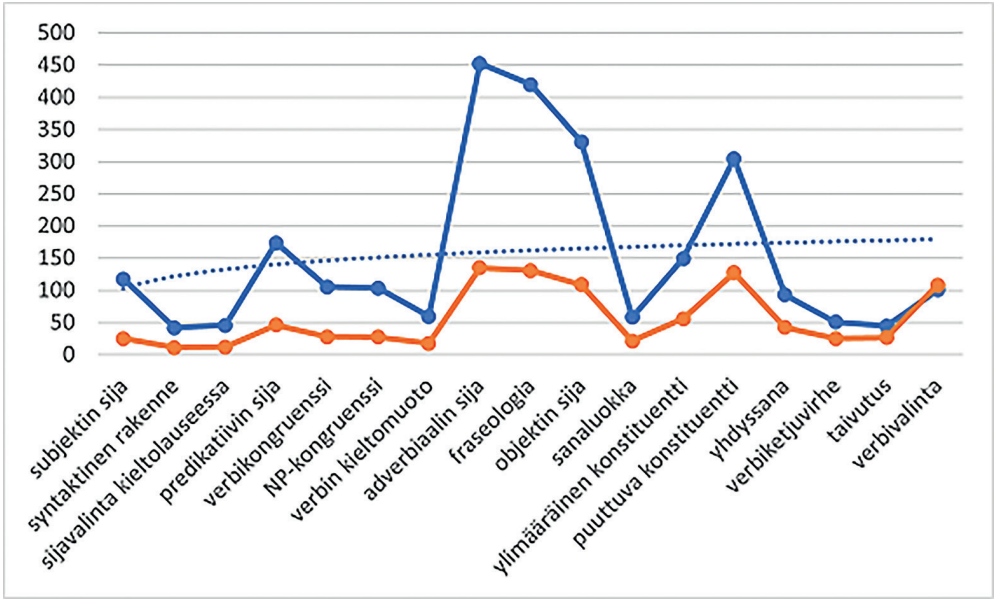
CZE-osakorpuksen (sininen käyrä) sekä RU-osakorpuksen (oranssi käyrä) virhekoodien absoluuttiset määrät esitetään kuviossa 1.



Kuvio 1. Yleisimmät virhekoodit CZE- ja RU-osakorpuksessa.

Virhe-esiintymien absoluuttinen määrä on suurempi CZE-osakorpuksessa kuin RU-osakorpuksessa. Pitää kuitenkin ottaa huomioon, että CZE-osakorpus sisältää 36 % enemmän saneita. Jotta olisi mahdollista vertailla erikokoisia aineistoja, olen laskenut myös virhe-esiintymien suhteellisen taajuuden per 1000 sanetta. Siitä ilmenee, että vaikka adverbiaalin sija ja fraseologia ovat yleisimmät virhekoodit kummassakin osakorpuksessa, suurin suhteellinen ero CZE- ja RU-osakorpusten välillä koskee virheellistä subjektin sijavalintaa, joka esiintyy 4,11 kertaa per 1000 sanaa CZE-osakorpuksessa ja 1,19 kertaa per 1000 sanaa RU-osakorpuksessa (virhekoodi esiintyy 3,45 kertaa useammin CZE-osakorpuksessa). Toiseksi suurin suhteellinen ero on virheellinen syntaktinen rakenne -virhekoodilla, joka esiintyy 2,81 kertaa useammin CZE-osakorpuksessa kuin RU-osakorpuksessa. Sitä seuraa sijavalinta kieltolauseessa -virhekoodi, joka esiintyy 2,79 kertaa useammin CZE-osakorpuksessa kuin RU-osakorpuksessa, sekä predikatiivin sijavalinta-,

verbikongruenssi- ja NP-kongruenssi-virhekoodit, jotka esiintyvät 2,77 kertaa useammin CZE-osakorpuksessa kuin RU-osakorpuksessa. Leksikaalisesti virheellinen verbivalinta on ainoa virhekoodi, joka esiintyy sekä absoluuttisesti että suhteellisesti useammin RU-osakorpuksessa kuin CZE-osakorpuksessa (1,46 kertaa enemmän). Virhekoodien suhteelliset esiintymät CZE- (sininen käyrä) ja RU-osakorpuksissa (oranssi käyrä) esitetään kuviossa 2.



Kuvio 2. Virhekoodien suhteelliset esiintymät.

Tulosten analysoinnissa pitää myös ottaa huomioon, että niillä virhekoodeilla, jotka kuuluvat lauseenjäsenen sijavalinta-makrokategoriaan, on erilaisia arvoja riippuen siitä, mitä sijaa oppija on käyttänyt ja mikä olisi ollut oikea sijavalinta kontekstin perusteella. Esimerkiksi virhekoodi <OBJ_VÄÄRÄ SIJA_OIKEA SIJA> saattaa esiintyä tekstissä muodossa <OBJ_NOM_PAR>, jos oppija on käyttänyt nominatiivia partitiivin sijaan, tai <OBJ_NOM_GEN>, mikäli oppija on käyttänyt nominatiivia genetiivin sijaan (virhekoodissa esiintyy aina sija, jota oppija on käyttänyt virheellisesti, sekä oikea sija). Tästä syystä esitän suhteellisten esiintymien lisäksi CZE- ja RU-osakorpusten kymmenen yleisintä virhekoodia kaikkine variantteineen taulukossa 3 (määrät ovat absoluuttiset).

Molemmissa osakorpuksissa <PHRASEO> (fraseologiset virheet) on yleisin virhekoodi. Toiseksi yleisin virhekoodi on <MISSING_OTHER> (yleensä puuttuva pakollinen omistusliite) ja <PRED_NOM_PAR> (virheellinen predikatiivin sija – nominatiivi partitiivin sijaan). Kymmenen yleisimmän virhekoodin joukossa on CZE-osakorpuksessa <SUB_NOM_PAR> (virheellinen subjektin sijavalinta – nominatiivi partitiivin sijaan) sekä <NEG_VERBFORM> (virheellinen kieltomuodon muodostaminen), jotka puuttuvat

Taulukko 3. 10 yleisintä virhekoodia variantteineen CZE- ja RU-osakorpuksissa.

CZE-osakorpus	**Esiintymät**	**RU-osakorpus**	**Esiintymät**
<PHRASEO>	381	<PHRASEO>	112
<MISSING_ OTHER>	220	<MISSING_ OTHER>	95
<PRED_NOM_PAR>	122	<PRED_NOM_PAR>	37
<EXTRA_OTHER>	108	<COMPOUND>	35
<AGR_VERBAL>	99	<EXTRA_OTHER>	28
<OBJ_NOM_PAR>	84	<AGR_VERBAL>	21
<COMPOUND>	78	<POS>	19
<POS>	56	<VERBCHAIN_ FORM>	19
<NEG_ VERBFORM>	55	<OBJ_NOM_PAR>	18
<SUB_NOM_PAR>	43	<OBJ_NOM_GEN>	15

RU-osakorpuksen kymmenen yleisimmän virhekoodin listalta. RU-osakorpuksen kymmenen yleisimmän virhekoodin listalta löytyvät virhekoodit <VERBCHAIN_FORM> (verbiketjuvirhe) sekä <OBJ_NOM_GEN>, jotka puuttuvat CZE-osakorpuksen kymmenen yleisimmän virhekoodin listalta. Virhekoodeja havainnollistavat esimerkit 6–9.

(6) Siis ajanjärjestely on minua varten vaikea. <PHRASEO> (TS0002i)

(7) Gogol Bordellon tekstit ovat **vitsikkäät** mutta samalla **arvostelevat**. <PRED_NOM_PAR> (TS0007)

(8) Minusta, suomalaisten poliitikkojen pitää oppia käyttäää eleitä […] <VERBCHAIN_FORM> (VE0088d)

(9) Hän haluaa tehdä sovinto Tomin kanssa […] <OBJ_NOM_GEN> (VE0055)

Mitä tulee leksikaalisesti virheelliseen verbivalintaan, eniten virheitä tšekinkielisten suomenoppijoiden osakorpuksessa on verbeissä *mennä*, *alkaa*, *oppia*, *ottaa*, *auttaa*, *olla*, *nähdä*, *soittaa*, *tuntea* ja *kuulua*. Venäjänkielisten suomenoppijoiden osakorpuksessa eniten virheitä on verbeissä *olla*, *levätä*, *viettää*, *aloittaa*, *kehittää*, *mennä*, *täytyä*, *tapahtua*, *valmistua* ja *katsoa*. *Mennä*-verbin virheellisen käytön yhteydessä kyseessä on yleensä *mennä*-verbin valinta *tulla*-verbin sijaan, kuten esimerkissä 10.

(10) Punahilkka tuli aikaisin ja koputti ovelle. ”Mene sisään! Ovi on auki.” (TS0017d)

Alkaa-verbin virheellisen käytön yhteydessä kyseessä on yleensä ei-transitiivisen verbin käyttö transitiivisen verbin sijaan, kuten esimerkissä 11.

(11) Joulu aloittaa pitkän aikaa joulukuun edellä. (TS0007e)

Venäjänkielisten suomenoppijoiden *levätä*-verbin virheellinen käyttö saattaa johtua venäjän kielen interferenssistä: venäjän kielen verbillä *отдыхать* [otdychat] on 'levätä'-merkityksen lisäksi merkitys 'viettää lomaa'. Tällaista virheellistä käyttöä näkyy esimerkissä 12.

(12) Joka vuosi minä lepäsin siellä mummoni kanssa. (VE0088)

Viettää-verbin virheellisen käytön esiintyminen tilastoissa saattaa johtua siitä, että aineisto on verrattain pieni: kaikki *viettää*-verbin virheellisen käytön tapaukset ovat peräisin samalta opiskelijalta.

Harvinaisimmat virhekoodien makrokategoriat tšekinkielisten suomenoppijoiden osakorpuksessa ovat virheellinen syntaktinen rakenne (lausetyyppi), sijavalinta kieltolauseessa, taivutus, verbiketjuvirheet ja virheellinen kieltomuodon muodostaminen. Taivutus-makrokategoriaa käytetään apukategoriana silloin kun tekstistä ei selviä, onko kyseessä sijavalintavirhe vai sijapäätteen muodostusvirhe, kuten esimerkissä 13.

(13) hän tykkää lukemista […] (TS0019i)

Harvinaisimmat virhekoodit venäjänkielisten suomenoppijoiden osakorpuksessa ovat syntaktinen rakenne, sijavalinta kieltolauseessa, virheellinen kieltomuodon muodostaminen, virheellinen sanaluokan valinta (esimerkiksi adjektiivi adverbin sijaan) ja subjektin sija.

Analyysin viimeisessä vaiheessa käytin Chi2-testiä selvittääkseni, ovatko virheesiintymien taajuuksien erot CZE- ja RU-osakorpuksessa tilastollisesti merkitseviä. Ensin analysoin yleisimmät virhekoodit: adverbiaalin sijavalintavirheet, fraseologiset virheet sekä objektin, predikatiivin ja subjektin sijavalintavirheet. Erot fraseologisten virheiden taajuudessa sekä objektin, predikatiivin ja subjektin sijavalintavirheiden taajuudessa ovat tilastollisesti merkitseviä Chi2-testin mukaan. Keskityin myös leksikaalisesti virheellinen verbivalinta -virhekoodiin, koska tämä on ainoa virhekoodi, jota esiintyy enemmän venäjänkielisten suomenoppijoiden teksteissä kuin tšekinkielisten suomenoppijoiden teksteissä. Tässäkin tapauksessa ero on tilastollisesti merkitsevä. Lopuksi analysoin virhekoodit, jotka esiintyvät harvemmin aineistossa: näitä ovat verbiketjuvirheet ja sijavalinta kieltolauseessa -virhekoodi. Ero verbiketjuvirheiden taajuudessa ei ole tilastollisesti merkitsevä, mutta tämä saattaa johtua verbiketjuvirheiden pienestä määrästä molemmissa aineistoissa. Sen sijaan ero kieltolauseiden sijavalintavirheiden taajuudessa on tilastollisesti merkitsevä, mutta kieltolausevirheiden määrä on pieni molemmissa aineistoissa, joten tässä tapauksessa ero saattaa johtua aineiston koosta. Taulukossa 4 esitän Chi2-testin arvot.

Taulukko 4. Chi2-testin arvot.

Virhekategoria	**Kategorian virhekoodien taajuus CZE-osakorpuksessa**	**Kategorian virhekoodien taajuus RU-osakorpuksessa**	**Chi2-testin p-arvo**	**Ero tilastollisesti merkitsevä?**
Adverbiaalin sijavalinta	452,5	135	$1{,}27 \times 10^{-21}$	Kyllä
Fraseologia	419,5	130,5	$5{,}13 \times 10^{-19}$	Kyllä
Objektin sijavalinta	331	102,5	$1{,}79 \times 10^{-15}$	Kyllä
Predikatiivin sijavalinta	173,5	46	$1{,}44 \times 10^{-10}$	Kyllä
Subjektin sijavalinta	117,5	25	$2{,}24 \times 10^{-9}$	Kyllä
Verbivalinta	100,5	108	0,005	Kyllä
Verbiketjut	50,5	25	0,113	Ei
Sijavalinta kieltolauseessa	45,5	12	0,001	Kyllä

6 *Tulkinnat ja yhteenveto*

Analyysin perusteella vastaan tutkimuskysymyksiin seuraavasti: Yleisin virhekoodi sekä venäjänkielisten että tšekinkielisten suomenoppijoiden osakorpuksessa on virheellinen adverbiaalin sija. Tämän virhekoodin yleisin variantti CZE-osakorpuksessa on <OTHER_PAR_GEN> (adverbiaalin kohdalla käytetty partitiivia genetiivin sijaan, kuten esimerkissä 14), toiseksi yleisin variantti on <OTHER_ADE_INE> (adverbiaalin kohdalla käytetty adessiivia inessiivin sijaan, kuten esimerkissä 15).

(14) […] jossa nukuimme teltassa koko viikonloppua. <OTHER_PAR_GEN> (TS0004c)

(15) Puutarhalla ovat puut, kukat ja kolme lammasta. <OTHER_ADE_INE> (TS0015f)

CZE-aineistosta nousee esiin, että ongelmia adverbiaalin sijavalinnassa aiheuttavat muun muassa ajanilmaukset (esimerkki 14) sekä paikallissijojen käyttö. <OTHER_VÄÄRÄ SIJA_OIKEA SIJA>-virhekoodiin kuuluvat myös sijavalintavirheet postposition yhteydessä (*suomea jälkeen* -tyyppiset virheet) sekä genetiivimääritteen sijavalintavirheet (*pidän antamisesta lahjoja* -tyyppiset virheet). <OTHER_VÄÄRÄ SIJA_OIKEA SIJA>-virhekoodi on aineiston yleisin. Ajanilmausvirheiden taajuus vastaa Klára Variksen (2010, 78) tuloksia, joiden mukaan nomineihin kohdistuu eniten ajanilmausvirheitä. Mitä tulee paikanilmauksiin, Tuija Määtän (2013, 135) mukaan oppijat sekoittavat hyvin usein illatiivin ja allatiivin keskenään.

Vaikka adverbiaalin sijavalintaan kohdistuu eniten sijavalintavirheitä myös venäjänkielisten suomenoppijoiden teksteissä, yleisimmät morfosyntaktiset virhekoodit koko RU-osakorpuksessa koskevat objektivirheitä. Yleisin morfosyntaktinen virhekoodi RU-osakorpuksessa on <OBJ_NOM_PAR> (nominatiivi partitiivin sijaan), toiseksi yleisin

on <OBJ_NOM_GEN> (nominatiivi genetiivin sijaan). Tämä vastaa aiempia tutkimustuloksia (Ranua–Ruotsalainen 2008, 160; Piri 2017, 55), joiden mukaan nominatiivi on tyypillisin virheellisesti tuotettu objektin sijamuoto.

<PHRASEO> (fraseologiset virheet) on toiseksi suurin makrokategoria sekä CZE-osakorpuksessa (381 tapausta) että RU-osakorpuksessa (112 tapausta). Tämä on kiinnostava havainto, koska oppijankielen tutkimuksen nykytendenssi on kiinnittää enenevässä määrin huomiota kielen idiomaattiseen ja fraseologiseen luonteeseen (Granger–Meunier 2008, 247). Fraseologisten virheiden kategoria on varsin laaja. Helena Kuuluvaisen (2015, 31) mukaan fraseologiset virheet muodostavat jatkumon sanastosta morfosyntaksiin. Kuuluvaisen mukaan fraseologisten virheiden pääryhmiin kuuluukin useampi virhetyyppikeskittymä: yksi sana – laajempi kokonaisuus, verbi ja sen täydennys, *kanssa*-sanan virheellinen käyttö, adpositio sijapäätteen tilalla, lausetyyppi, kysyvä sivulause ja muodollinen subjekti. Sekä CZE- että RU-osakorpuksesta löytyy fraseologisia virheitä, jotka vastaavat Kuuluvaisen kategoriointia: esimerkiksi kysyvä sivulause -tyyppiset fraseologiset virheet (esimerkki 16), epätyypilliset kollokaatit (esimerkit 17 ja 18), adpositio sijapääteen tilalla (esimerkki 19) ja *kanssa*-sanan virheellinen käyttö (esimerkki 20).

(16) ensin katsoin, jos jotain ei ollut tapahtunut heille. (TS0001e)

(17) ne ovat todella pieniä, täysin matalia […] (TS0001f)

(18) Siksi sanottiin, että Turandot oli kovin julma ja ankara. (VE0005)

(19) Jokaisella ihmisellä, joka haluaa seurata muotin mukaan, on iPhone. (VE0012a)

(20) hän lyhentää hänen odottamisensa suklaakeksin Delissan kanssa […] (TS0002j)

<PRED_NOM_PAR> (virheellinen predikatiivin sijavalinta, nominatiivi partitiivin sijaan) on yleisin <PRED_ VÄÄRÄ SIJA_OIKEA SIJA>-virhekoodin variantti sekä CZE- että RU-osakorpuksessa. Tämä osoittaa, että predikatiivin tapauksessa, samoin kuin objektin tapauksessa, nominatiivi on useimmiten virheellisesti valittu sijamuoto. Ongelmia predikatiivin sijavalinnassa aiheuttaa subjektin monikollisuus (esimerkki 21), jaollisuus (esimerkki 22) sekä tiettyjen lausetyyppien morfosyntaktiset ominaisuudet, esimerkiksi tuloslauseen predikatiivin sijavalinta (esimerkki 23).

(21) Tavallisesti he ovat reilut, luotettavat, ahkerat ja ystävälliset. (VE0053a)

(22) kaikki on outo. (VE0088k)

(23) Tšaikovskin musiikista tuli hyvin suosittu. (VE0002a)

Subjektin sijavalintavirheet eivät ole kovin yleisiä CZE- eikä RU-osakorpuksessa. Subjektin sijavalintavirheet muodostavat 4,4 % kaikista virheistä CZE-osakorpuksessa ja vain 2,7 % kaikista virheistä RU-osakorpuksessa. Ero saattaa johtua siitä, että CZE-osakorpuksen tehtävänannot osittain eroavat RU-osakorpuksen tehtävänannoista:

CZE-osakorpuksen tehtävänantojen joukosta löytyy esimerkiksi ”Perheeni”- sekä ”Unelma-asuntoni”-teemat, joissa eksistentiaalilauseiden taajuus saattaa olla korkeampi, ja tämä on saattanut aiheuttaa isomman subjektin sijavirheiden määrän CZE-osakorpuksessa.

Leksikaalisesti virheellinen verbivalinta on ainoa makrokategoria, joka esiintyy useammin venäjänkielisten suomenoppijoiden teksteissä sekä absoluuttisesti että suhteellisesti. Tämäkin saattaa johtua tehtävänannoista, koska venäjänkielisten suomenoppijoiden kirjoitusten joukossa on venäjänkielisten artikkeleiden referaatteja suomeksi, kun taas tšekinkielisten suomenoppijoiden tehtävänantojen joukossa ei ollut referaatteja. Venäjänkielisten artikkeleiden kääntäminen suomeksi on saattanut lisätä verbivalintavirheitä. Tehtävänanto on saattanut vaikuttaa myös siihen, että venäjänkielisten suomenoppijoiden tuloksissa <V_LEX>-makrokategoriassa (leksikaalisesti virheellinen verbivalinta) esiintyy enemmän verbejä, joilla on Suomen sanomalehtikielen taajuussanastossa (CSC 2004) matalampi taajuus, kun taas tšekinkielisten suomenoppijoiden <V_LEX>-makrokategoriassa esiintyy hyvin frekventtejä verbejä. Tämäkin voi johtua siitä, että osa aineiston venäjänkielisistä suomenoppijoista on kääntänyt artikkeleita. Taulukossa 5 esitän kunkin osakorpuksen kymmenen yleisimmin virheellisesti valittua verbiä ja niiden sijoituksen Suomen sanomalehtikielen taajuussanastossa.

Taulukko 5. 10 yleisimmin virheellisesti valittua verbiä CZE- ja RU-osakorpuksissa ja niiden sijoitus Suomen sanomalehtikielen taajuussanastossa (CSC 2004).

CZE-osakorpuksen yleisimmin virheellisesti valitut verbit	**Verbin sijoitus sanomalehtikielen taajuussanastossa**	**RU-osakorpuksen yleisimmin virheellisesti valitut verbit**	**Verbin sijoitus sanomalehtikielen taajuussanastossa**
mennä	88	*olla*	1
alkaa	65	*levätä*	3695
oppia	740	*viettää*	652
ottaa	59	*aloittaa*	166
auttaa	550	*kehittää*	599
olla	1	*mennä*	88
nähdä	112	*täytyä*	641
soittaa	657	*tapahtua*	236
tuntea	219	*valmistua*	552
kuulua	96	*katsoa*	211

Verbiketjuvirheiden määrä on pieni sekä CZE-osakorpuksessa (1,9 % kaikista virheistä) että RU-osakorpuksessa (2,7 % kaikista virheistä). Verbiketjuvirheet koskevat sekä jälkimmäisen verbin tai deverbaalisubstantiivin muotoa että jälkimmäisen verbin tai deverbaalisubstantiivin sijavalintaa. Tšekinkielisten suomenoppijoiden teksteissä eniten

virheitä aiheuttaa kuitenkin jälkimmäisen verbin tai deverbaalisubstantiivin muoto eikä sen sijavalinta (39 tapausta), kuten esimerkissä 24.

(24) Kun työskennellään, on tärkeä oppia levätä. (TS0002i)

Tšekinkieliset suomenoppijat ovat käyttäneet A-infinitiiviä MA-infinitiivin sijaan kymmenessä tapauksessa ja A-infinitiiviä deverbaalisubstantiivin sijaan kymmenessä tapauksessa, mikä saattaa viitata yksinkertaistamiseen vastaavalla tavalla kuin nominatiivin suosiminen muiden muotojen sijasta. Yhdeksässä tapauksessa tšekinkieliset suomenoppijat ovat käyttäneet MA-infinitiiviä A-infinitiivin sijaan.

Myös venäjänkielisten suomenoppijoiden teksteissä eniten virheitä aiheutti verbiketjun jälkimmäisen verbin tai deverbaalisubstantiivin muoto (19 tapausta) (esimerkki 25).

(25) maanalaisten asukkaat jättävät kotinsa ja lähtevät etsiä uuden. (VE0002)

Venäjänkieliset suomenoppijat, kuten tšekinkielisetkin, ovat käyttäneet A-infinitiiviä MA-infinitiivin sijaan (11 tapausta), sekä A-infinitiiviä deverbaalisubstantiivin sijaan (viisi tapausta). Tanja Seppälä (2012) antaa pro gradu -työssään yleiskatsauksen verbiketjuvirheiden kategorioista ja frekvenssistä. Seppälän (2012, 78) mukaan kuitenkin oletus siitä, että MA-infinitiivin vaativat verbit yhdistyvät usein A- infinitiiveihin näyttää pätevän vain, kun kyseessä ei ole liikeverbi.

Sijavalinta kieltolauseessa on toiseksi harvinaisin makrokategoria sekä CZE-osakorpuksessa (1,7 % kaikista virheistä) että RU-osakorpuksessa (1,3 % kaikista virheistä). Taustalla saattaa olla kieltolauseiden välttäminen oppimisstrategiana (tästä ks. Šebesta 2012, 24). Kieltolausevirheiden yhteydessä frekventein virhekoodi CZE-osakorpuksessa on <OBJ_NOM_PAR> (14 tapausta). Tämä viittaa siihen, että oppijoilla saattaa olla tendenssi jättää objektia taivuttamatta myöntölauseiden lisäksi myös kieltolauseissa. Joissakin tapauksissa oppijat valitsivat genetiivin partitiivin sijaan objektin sijaksi (<OBJ_GEN_PAR>, yhdeksän tapausta). Sama tendenssi näkyy myös subjektin sijavalinnassa kieltolauseissa, joissa jaollisen subjektin yhteydessä on käytetty nominatiivia partitiivin sijaan, kuten esimerkissä 26.

(26) hänellä on omat ongelmat enemmän kuin tarpeeksi. (TS0004)

Ongelmia on osittain aiheuttanut myös predikatiivin sijavalinta kieltolauseessa – joissakin tapauksissa oppija on käyttänyt partitiivia nominatiivin sijaan, kuten esimerkissä 27.

(27) En ole varmaa, miten tämä sanoa suomeksi. (TS0001f)

Kyseessä saattaa olla kielteisen objektin sääntöjen yliyleistämisstrategia myös predikatiivin sijavalinnassa (yliyleistämisestä oppimisstrategiana ks. Rauto 2003, 78). Jotkut opiskelijat ovat saattaneet päätellä objektin sijavalinnan sääntöjen perusteella, että kielteisen lauseen predikatiivin tulee aina olla partitiivissa.

Toisin kuin tšekinkielisten suomenoppijoiden teksteissä, venäjänkielisten suomenoppijoiden kieltolausevirheistä löytyy vain kolme tapausta, joissa objektin sijaksi on

virheellisesti valittu nominatiivi, ja yksi tapaus, jossa predikatiivin sijaksi on virheellisesti valittu nominatiivi. Venäjänkielisten suomenoppijoiden teksteissä yleisin virheellisen kielto-objektin sijavalinta ei ole nominatiivi, kuten tšekinkielisten oppijoiden teksteissä, vaan genetiivi, kuten esimerkissä 28.

(28) Alex ei antanut Lisan kirjeen Mattille […] (VE0007)

Näissä tapauksissa oppijat ovat saattaneet yliyleistää myönteisen objektin sijavalinnan säännöt myös kielteisen objektin sijavalintaan.

Harvinaisimmassa makrokategoriassa <SYN_ST> (virheellinen syntaktinen rakenne, virheellisesti muodostettu lausetyyppi) eniten virheitä esiintyy eksistentiaalilauseissa (eksistentiaalilauseiden ominaisuuksista ks. VISK § 891), kuten esimerkeissä 29 ja 30.

(29) Musiikkijuhlassa olivat joka päivää paljon musiikkiryhmää […] (TS0002c)

(30) Naisella päällä on sininen mekko, miehellä ovat mustat housut ja vaalea paita. (VE0008b)

Chi2-testin tulokset osoittavat, että vaikka yleisimmät makrokategoriat sekä CZE-osakorpuksessa että RU-osakorpuksessa ovat adverbiaalin sijavalinta ja fraseologia, niiden esiintymistaajuudessa on kuitenkin merkitseviä tilastollisia eroja: adverbiaalin sijavalintavirheitä ja fraseologisia virheitä esiintyy enemmän CZE-osakorpuksessa. Leksikaalisia verbivalintavirheitä esiintyy enemmän RU-osakorpuksessa. Objektin, predikatiivin ja subjektin sijavalintavirheitä esiintyy enemmän CZE-osakorpuksessa. Toisaalta jos virhekoodin taajuus on pieni (sijavalinta kieltolauseessa, verbiketjut) tilastollisen merkitsevyyden arviointi on haastavampaa.

Tutkimukseni osoittaa, että tšekinkielisten ja venäjänkielisten B1-tason tekstien virhetendenssit vastaavat aiempia tutkimustuloksia: aineistossa näkyy nominatiivin yliedustuminen objektin sijana, kielteisen objektin sääntöjen yliyleistäminen myös predikatiivin sijavalinnassa, fraseologisten virheiden korkea taajuus ja moninaisuus sekä tiettyjen makrokategorioiden (verbiketjujen, kielteisten lauseiden) matala taajuus, mikä saattaa viitata näiden syntaktisten rakenteiden välttämiseen. Toisaalta se, että melkein kaikissa makrokategorioissa oppijaryhmien välillä esiintyy tilastollisesti merkitseviä eroja, saattaa viitata siihen, että annotointiperiaatteissa on saattanut olla eroja (tšekinkielisten oppijoiden tekstit annotoitiin Oulun yliopistossa, venäjänkielisten oppijoiden tekstit annotoitiin Prahan yliopistossa suomea ensikielenä puhuvan annotoijan avustuksella). Tästä syystä olisi suotavaa suorittaa IAA-testi (*Inter-annotator agreement test*, Artstein 2017), jonka avulla voi verrata annotointiperiaatteita systemaattisesti. Fraseologisten virheiden suuren määrän vuoksi olisi myös suotavaa luokitella kummankin osakorpuksen fraseologiset virheet Kuuluvaisen (2015, 31) käyttämien fraseologisten virheiden alakategorioiden mukaan ja vertailla tuloksia aiempaan tutkimukseen. Mahdollinen tuleva tutkimussuunta voisi olla myös tšekinkielisten ja venäjänkielisten suomenoppijoiden tulosten vertailu muiden suomenoppijoiden tuloksiin koko ICLFI-aineiston avulla. Näin olisi mahdollista tutkia, ovatko yllä mainitut virhetendenssit tyypillisiä kaikille suomi vieraana kielenä -oppijoille

vai pelkästään niille, jotka puhuvat äidinkielenään slaavilaista kieltä. Olisi myös mahdollista vertailla Venäjän venäjänkielisten ja Viron venäjänkielisten tuloksia keskenään. Näin saataisiin enemmän tietoja oppimiskontekstin vaikutuksesta suomen kielen omaksumiseen ja siitä, miten toisen suomalais-ugrilaisen kielen osaaminen vaikuttaa suomen kielen omaksumiseen.

Lähteet

Artstein, Ron 2017: Inter-annotator agreement. Nancy Ide ja James Pustejovsky (toim.): *Handbook of linguistic annotation*, 297–313. Springer, Dordrecht. http://dx.doi.org/10.1007/978-94-024-0881-2_11

Breyer, Yvonne Alexandra 2011: *Corpora in language teaching and learning. Potential, evaluation, challenges*. Peter Lang, Frankfurt am Main.

Brunni, Sisko – Jantunen, Jarmo – Skantsi, Valtteri 2020: Korpusavusteinen virheanalyysi tarkkuuden kehityksestä EVK:n taitotasoilla A2–B2. *Puhe ja kieli* 39 (3), 275–304. https://doi.org/10.23997/pk.76601

Brunni, Sisko – Lehto, Liisa-Maria – Jantunen, Jarmo – Airaksinen, Valtteri 2015: How to annotate morphologically rich learner language. Principles, problems and solutions. *Bergen Language and Linguistics Studies* 6, 133–152. https://doi.org/10.15845/bells.v6i0.812

CSC – Tieteen tietotekniikan keskus 2004: Suomen sanomalehtikielen taajuussanasto [tekstikorpus]. Kielipankki. Saatavissa http://urn.fi/urn:nbn:fi:lb-201405272

Cvrček, Václav 2019: Calc: Corpus calculator. Kaarlen yliopisto, Praha. Saatavissa https://korpus.cz/calc/ [viitattu 26.3.2021].

CzeSL = Šebesta, Karel – Bedřichová, Zuzanna – Šormová, Kateřina – Štindlová, Barbora – Hrdlička, Milan – Hrdličková, Tereza – Hana, Jiří – Petkevič, Vladimír – Jelínek, Tomáš – Škodová, Svatava – Poláčková, Monika – Janeš, Petr – Lundáková, Kateřina – Skoumalová, Hana – Sládek, Šimon – Pierscieniak, Piotr – Toufarová, Dagmar – Richter, Michal – Stráka, Milan – Rosen, Alexandr 2014: Czech as a Second Language with Spelling, Grammar and Tags [tekstikorpus]. Kaarlen yliopisto, Praha. Saatavissa http://www.korpus.cz [viitattu 26.8.2022].

Díez-Bedmar, María Belén 2021: Error Analysis. Nicole Tracy-Ventura ja Magali Paquot (toim.): *The Routledge handbook of second language acquisition and corpora*, 90–104. Routledge, New York and London. https://doi.org/10.4324/9781351137904-9

EVK 2003: *Eurooppalainen viitekehys. Kielten oppimisen, opettamisen ja arvioinnin yhteinen eurooppalainen viitekehys*. WSOY, Helsinki.

Geryšerová, Anna 2019: *Používání finského pasiva českými rodilými mluvčími na základě korpusu LAS2*. Pro gradu -tutkielma. Masarykin yliopisto, Brno. Saatavissa https://is.muni.cz/th/b8anh/Diplomova_prace.pdf [viitattu 15.3.2021].

Granger, Sylviane 2002: A bird's-eye view of learner corpus research. Sylviane Granger, Joseph Hung ja Stephanie Petch-Tyson (toim.): *Computer learner corpora, second language acquisition and foreign language teaching*, 3–33. Benjamins, Amsterdam. https://doi.org/10.1075/lllt.6

Granger, Sylviane – Gilquin, Gaëtanelle – Meunier, Fanny 2015: Introduction: learner corpus research – past, present and future. Sylviane Granger, Gaëtanelle Gilquin ja Fanny Meunier (toim.): *Cambridge handbook of leaner corpus research*, 423–442. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139649414

Granger, Sylviane – Meunier, Fanny 2008: Phraseology in language learning and teaching: where to from here? Fanny Meunier ja Sylviane Granger (toim.): *Phraseology in foreign language learning and teaching*, 247–252. Benjamins, Amsterdam & Philadelphia. https://doi.org/10.1075/z.138.19gra

ICLFI = Jantunen, Jarmo – Brunni, Sisko – Oulun yliopisto, suomen kielen oppiaine 2013: Kansainvälinen oppijansuomen korpus [tekstikorpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-20140730163

Ivaska, Ilmari 2015: *Edistyneen oppijansuomen konstruktiopiirteitä korpusvetoisesti: avainrakenneanalyysi*. Turun yliopisto, Turku. https://urn.fi/URN:ISBN:978-951-29-6204-4

Jantunen, Jarmo 2011: Kansainvälinen oppijansuomen korpus (ICLFI): typologia, taustamuuttujat ja annotointi. *Lähivõrdlusi. Lähivertailuja* 21, 86–105. http://dx.doi.org/10.5128/LV21.04

Jantunen, Jarmo – Brunni, Sisko – Lehto, Liisa-Maria – Airaksinen, Valtteri 2014: Oppijankieliaineistojen annotointi – esimerkkinä ICLFI:n annotoinnin prosessit, ongelmat ja ratkaisut. *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 7, 60–80. Saatavissa https://journal.fi/afinla/article/view/48160 [viitattu 28.3.2021].

Jarvis, Scott – Pavlenko, Aneta 2010: *Crosslinguistic infl ence in language and cognition*. Routledge, New York.

Kaivapalu, Annekatrin 2005: *Lähdekieli kielenoppimisen apuna*. Jyväskylä studies in humanities 44. Jyväskylän yliopisto, Jyväskylä. Saatavissa https://jyx.jyu.fi/bitstream/handle/123456789/13439/9513923916.pdf?sequence=1 [viitattu 28.3.2022].

Kováčová, Ivana 2017: *Essiivin käyttö tsekkiläisten suomenoppijoiden kirjoitelmissa*. Pro gradu -tutkielma. Oulun yliopisto, Oulu. Saatavissa http://jultika.oulu.fi/files/nbnfioulu-201711093089.pdf [viitattu 11.3.2022].

Kuuluvainen, Helena 2015: *Fraseologiset virheet kansainvälisessä oppijansuomen korpuksessa*. Pro gradu -tutkielma. Oulun yliopisto, Oulu. Saatavissa http://jultika.oulu.fi/files/nbnfioulu-201509172008.pdf [viitattu 26.3.2022].

LAS2 = Turun yliopiston kieli- ja käännöstieteiden laitos 2012: Edistyneiden suomenoppijoiden korpus [tekstikorpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-201407167

Leech, Geoffrey 1997: Introducing corpus annotation. Roger Garside, Geoffrey Leech ja Tony McEnery (toim.): *Corpus annotation: linguistic information from computer text corpora*, 1–18. Longman, London.

McEnery, Tony – Wilson, Andrew 2001: *Corpus linguistics: an introduction*. 2. painos. Edinburgh University Press, Edinburgh.

Meunier, Fanny 2021: Introduction to learner corpus research. Nicole Tracy-Ventura ja Magali Paquot (toim.): *The Routledge handbook of second language acquisition and corpora*, 23–36. Routledge, New York and London. https://doi.org/10.4324/9781351137904-4

Mukherjee, Joybrato – Götz, Sandra 2015: Learner corpora and learning context. Sylviane Granger, Gaëtanelle Gilquin ja Fanny Meunier (toim.): *Cambridge handbook of leaner corpus research*, 423–442. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139649414

Määttä, Tuija 2013: Verbien *tulla* ja *mennä* rektioista ruotsinkielisten alkeistason suomenoppijoiden kirjallisissa tuotoksissa. *AFinLA-e: Soveltavan kielitieteen tutkimuksia* 5, 123–141. Saatavissa https://journal.fi/afinla/issue/view/1107 [viitattu 25.03.2022].

Piri, Olli-Juhani 2017: *Morfosyntaktiset objektivirheet oppijansuomessa: korpuspohjainen kuvaus suomenoppijoiden kielitaidon tarkkuuden kehityksestä*. Pro gradu -tutkielma. Oulun yliopisto, Oulu. http://urn.fi/URN:NBN:fi:oulu-201706012339.

Ranua, Minna-Mari – Ruotsalainen, Margit 2008: Syntaktisten virheiden vertailua suomi toisena ja suomi vieraana kielenä-oppijoiden teksteissä. Helena Sulkala, Maija-Liisa Halme ja Hannakaisa Holmi (toim.): *Tutkielmia oppijankielestä* III, 149–172. Studia humaniora ouluensia 5. Oulu University Press, Oulu.

Rauto, Eeva 2003: *Välikielen kehitys vieraskielisessä opetuksessa. Tutkimus muutoksista insinööriopiskelijoiden englannin kieliopin hallinnassa*. Jyväskylän yliopisto, Jyväskylä. http://urn.fi/URN:ISBN:951-39-1542-5

Ringbom, Håkan 1980: On the distinction between second-language acquisition and foreign-language learning. *Papers in Language Learning and Language Acquisition*, 37–44. AFinLa ry, Jyväskylä. Saatavissa https://journal.fi/afinlavk/issue/view/4074 [viitattu 25.3.2022].

Šebesta, Karel 2012: Parametry žákovských korpusů a CzeSL. Karel Šebesta ja Svatava Škodová (toim.*): Čeština – cílový jazyk a korpusy*, 13–33. Technical University of Liberec, Liberec.

Seppälä, Tanja 2012: *Oppijansuomen kolligaatit ketjuuntuvissa verbirakenteissa*. Pro gradu -tutkielma. Oulun yliopisto, Oulu. Saatavissa https://www.oulu.fi/sites/default/files/content/Seppala%20gradu.pdf [viitattu 26.3.2022].

Spoelman, Marianne 2013: *Prior linguistic knowledge matters. The use of the partitive case in Finnish learner language*. Acta Universitatis Ouluensis B Humaniora 111. Oulun yliopisto, Oulu. http://urn.fi/urn:isbn:9789526201146

Štindlová, Barbora – Rosen, Alexandr – Hana, Jirka – Škodová, Svatava 2012: CzeSL – an error tagged corpus of Czech as a second language. Piotr Pęzik (toim.): *PALC 2011 – Practical applications in language and computers (Łódź 13–15 April 2011)*, 13–15. Łódź studies in language 28. Peter Lang, Frankfurt am Main.

Thewissen, Jennifer 2021: Accuracy. Nicole Tracy-Ventura ja Magali Paquot (toim.): *The Routledge handbook of second language acquisition and corpora*, 305–317. Routledge, New York & London. https://doi.org/10.4324/9781351137904-27

——— 2013: Capturing L2 accuracy developmental patterns: insights from an error-tagged EFL learner corpus. *The Modern Language Journal* 97 (1), 77–101. https://doi.org/10.1111/j.1540-4781.2012.01422.x

Tognini-Bonelli, Elena 2001: *Corpus linguistics at work*. Studies in corpus linguistics 6. John Benjamins Publishing Company, Amsterdam / Philadelphia. https://doi.org/10.1075/scl.6

Varis, Klára 2010: *Ajanilmaukset Cefling-hankkeen koululaisaineistossa*. Pro gradu -tutkielma. Jyväskylän yliopisto, Jyväskylä. Saatavissa https://jyx.jyu.fi/bitstream/handle/123456789/25306/varis-gradu.pdf?sequence=4&isAllowed=y [viitattu 25.3.2022].

Virtanen, Veera 2011: Minä lienen tullut joskus Suomessa vielä? *Venäjänkielisten suomi toisena ja suomi vieraana kielenä -oppijoiden perfektin ja imperfektin omaksumisen ongelmista*. Pro gradu -tutkielma. Jyväskylän yliopisto, Jyväskylä. Saatavissa https://jyx.jyu.fi/bitstream/handle/123456789/27066/URN%3aNBN%3afi%3ajyu-2011052410910.pdf?sequence=1&isAllowed=y [viitattu 15.3.2022].

VISK = Hakulinen, Auli – Vilkuna, Maria – Korhonen, Riitta – Koivisto, Vesa – Heinonen, Tarja Riitta – Alho, Irja 2004: *Iso suomen kielioppi*. Verkkoversio. Suomalaisen Kirjallisuuden Seura, Helsinki. Saatavissa http://scripta.kotus.fi/visk [viitattu 3.6.2022].

Valentina Dani: Morphosyntactic-, phraseological- and verb choice-related errors in the Czech L1 and Russian L1 sections of the International Corpus of Learner Finnish

The present study aims at analysing the Czech L1 and Russian L1 subcorpora in the International Corpus of Learner Finnish. The study focuses on morphosyntactic, phraseological and verb choice errors found in texts rated at level B1 according to the Common European Framework of Reference for Languages, with a view to uncovering converging and diverging trends in error frequency and type. The findings indicate that the most common error types in both subcorpora concern adverbial case-marking and phraseology. The findings also reveal differences in error frequency in subject case-marking and syntactic structure on the one hand (these errors were found to be significantly more frequent in the Czech L1 subcorpus) and verb choice on the other hand (these errors were found to be more frequent in the Russian L1 subcorpus). The present study confirms previous findings of research on Finnish as both a second and foreign language, such as the prominence of time expression errors, the high frequency of phraseological

errors and the overrepresentation of the nominative case in object and predicative case-marking and of the A-infinitive in verb + verb constructions. The overrepresentation of both the nominative case and the A-infinitive in learners' texts show a tendency of simplification typical of a learner language, while the higher frequency of verb choice errors in the Russian L1 subcorpus indicates that assignment type influences error frequency and type, since the Russian L1 subcorpus, unlike the Czech L1 subcorpus, includes translation assignments.

Valentina Dani
dani.valentina.uni@gmail.com
Ústav Českého národního korpusu / Institute of the Czech National Corpus
Kaarlen yliopisto (Univerzita Karlova / Charles University)
https://orcid.org/0000-0001-6914-9434