Konekäännös suomalaisen kääntäjän näkökulmasta

Tommi Nieminen

31.1.2019

Johdanto

1. Miten konekäännös on vaikuttanut suomalaisen kääntäjän työhön?
2. Miten konekäännös tulee lähitulevaisuudessa vaikuttamaan suomalaisen kääntäjän työhön?
3. Onko suomi vaikea kieli konekäännöksen kannalta?

Käännösala Suomessa

Ennen konekäännöksen vaikutusten esittelyä on hyvä kerrata, **millaista tekstiä** ja **mille kielille** Suomessa käännetään ja **millaisia kääntäjäryhmiä** Suomessa on.

Käännösala Suomessa: taloudellisesti merkittävät kieliparit

Suuret julkiset organisaatiot ja yritykset

→

Korkeat laatuvaatimukset

Julkiset organisaatiot ja yritykset

→

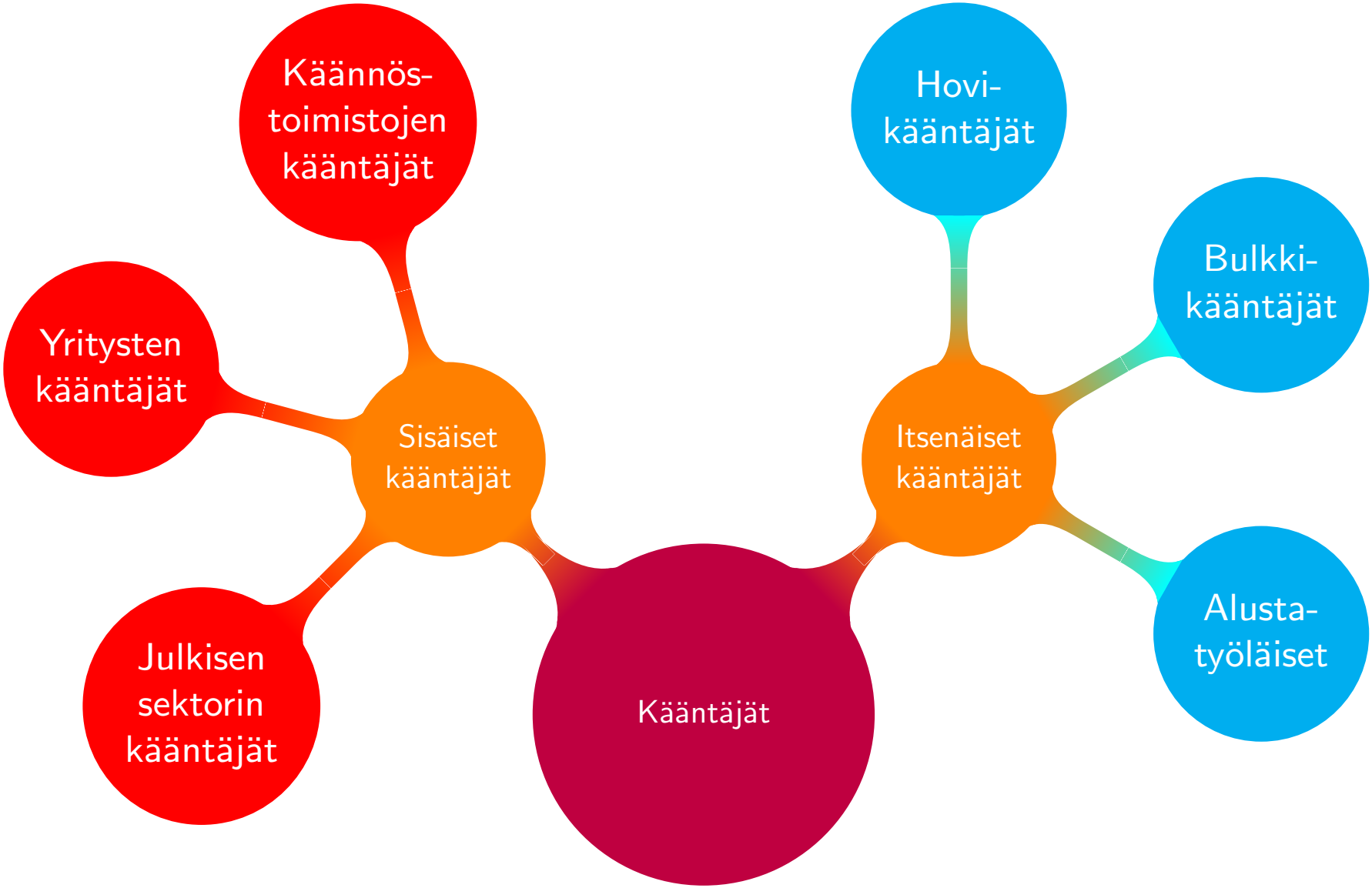
Korkeat laatuvaatimukset

Pääasiassa yritykset

→

Joustavat laatuvaatimukset

Käännösala Suomessa: kääntäjien ryhmät



Konekäännöksen tila

- Konekäännös perustuu nykyään lähes yksinomaan koneoppimismenetelmiin, jotka oppivat konekääntämään kaksikielisen opetusaineiston perusteella.
- Opetusaineiston riittävä laajuus on erittäin tärkeää laadun kannalta.
- Kaikille edellä mainituille kielipareille on runsaasti aineistoa, joista osa on vapaasti saatavilla ja valtaosa ainoastaan käännösten tilaajien ja käännöstoimistojen käytettävissä.

Seuraavissa dioissa tarkastellaan konekäännöksen laatua englanti–suomi-kieliparin osalta.

Konekäännöksen lähihistoriaa (2017-)

WMT-konferenssia varten on kolmen vuoden ajan järjestetty konekääntimien vertailuja, joissa ihmiset arvioivat käännösten laatua lausekohtaisesti asteikolla 0–100. Englanti–suomi on yksi arvioiduista kielipareista.

English→Finnish

#	Ave %	Ave z	system
1	59.6	0.378	online-B
	57.8	0.305	HY-HNMT
3	51.6	0.090	online-G
	51.3	0.060	jhu-nmt-latt-resc
	49.3	−0.004	AaltoHnmtMultitask
6	46.4	−0.102	AaltoHnmtFlatcat
	46.7	−0.109	online-A
	45.8	−0.115	HY-SMT
	43.5	−0.192	HY-AH
	43.4	−0.204	jhu-pbmt
11	40.8	−0.298	TALP-UPC
12	8.0	−1.428	apertium

Lähde: Ondřej Bojar et al., "Findings of the 2017 Conference on Machine Translation (WMT17)"

English→Finnish

	Ave. %	Ave. z	System
1	64.7	0.521	NICT
	63.1	0.466	HY-NMT
3	59.2	0.324	UEDIN
	58.3	0.271	AALTO
	57.9	0.258	HY-NMT-2STEP
	57.4	0.238	TALP-UPC
	55.9	0.184	CUNI-KOCMI
	56.6	0.183	ONLINE-B
9	45.9	−0.212	ONLINE-A
	45.3	−0.233	ONLINE-G
11	42.7	−0.334	HY-SMT
	41.5	−0.369	HY-AH

Lähde: Ondřej Bojar et al., "Findings of the 2018 Conference on Machine Translation (WMT18)"

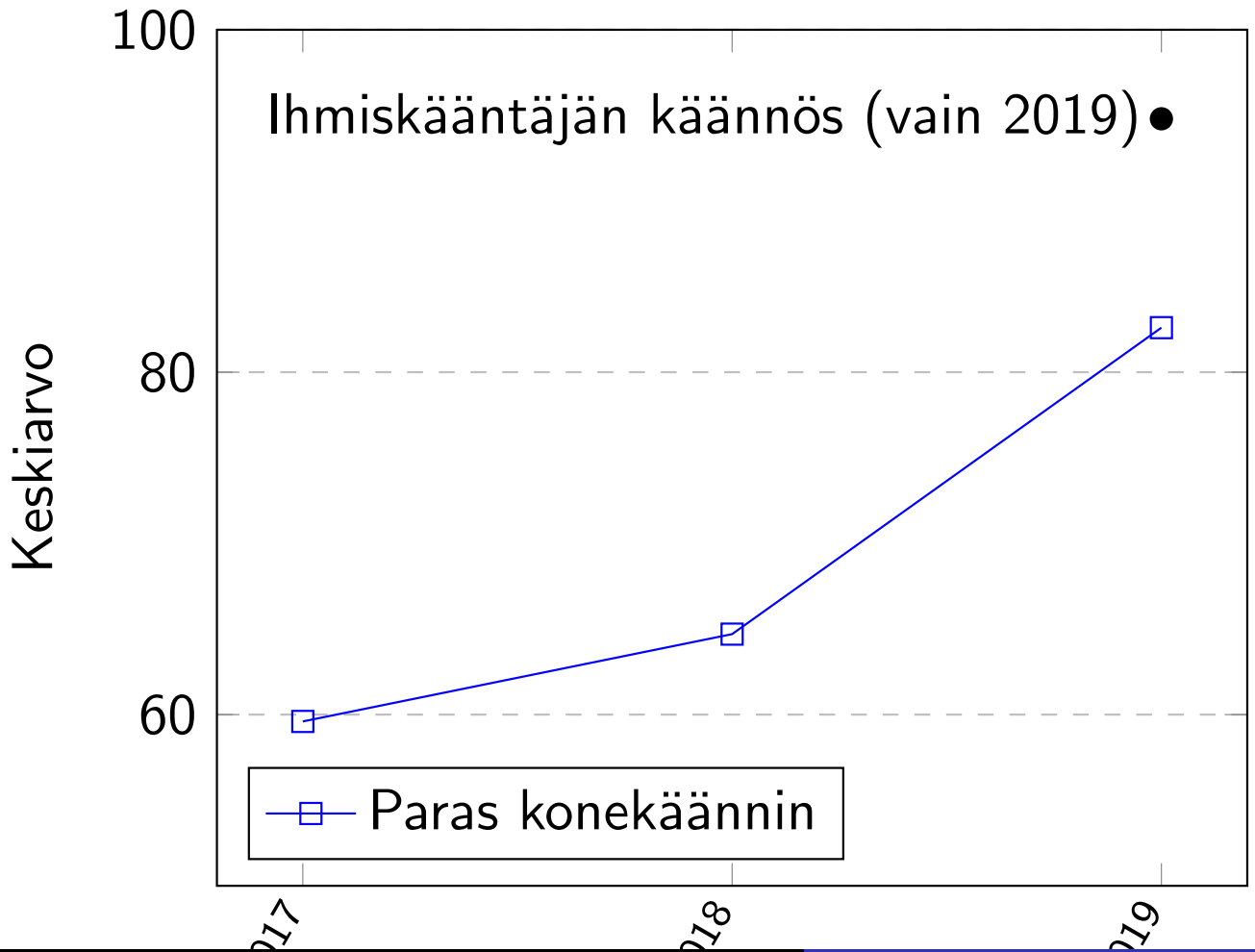
English→Finnish

Ave.	Ave. z	System
94.8	1.007	HUMAN
82.6	0.586	GTCOM-Primary
80.2	0.570	MSRA-NAO
70.9	0.275	online-Y
65.8	0.199	NICT
65.7	0.09	Helsinki-NLP
63.1	0.072	online-G
63.0	0.037	online-B
54.5	−0.125	TartuNLP-c
48.3	−0.384	online-A
47.1	−0.398	online-X
47.9	−0.522	Helsinki-NLP-rule-based
16.9	−1.260	apertium-uc

Lähde: Barrault et al., "Findings of the 2019 Conference on Machine Translation (WMT19)"

Konekäännöksen lähihistoriaa (2017-)

Eri vuosien WMT-tulokset eivät ole täysin vertailukelpoisia, mutta kehitys on ollut huimaa.



Konekäännöksen lähihistoriaa (2017-)

NMT löi läpi WMT-kilpailussa vuonna 2016, ja 2017 oli viimeinen vuosi, jolloin kilpailussa nähtiin vielä sääntöpohjaisia ja tilastollisia konekääntimiä.
Erot järjestelmien tuottamissa käännöksissä olivat selvät (suurimmat virheet on merkitty esimerkkeihin).

Esimerkki 1: lähdeteksti

"This isn't a good experience for people and we plan to address the issue," a Facebook spokesperson said in an e-mailed statement.

Konekäännöksen lähihistoriaa (2017)

Esimerkki 1: sääntöpohjainen konekäännös (RBMT)

”Tämä ei ole hyvä kokemus ihmisille ja me suunnittelemme ~~esittämään kysymyksen~~ asian selvittämistä, Facebookin ~~puhehenkilö~~ edustaja sanoi ~~e-mailed~~ sähköpostitse lähetetyssä julkilausumassa.

Tyypilliset ongelmat: taivutus, kontekstiin sopimattomat termit ja rakenteet, tuntemattomat sanat (tavuviiva sanassa ”e-mailed”).

Konekäännöksen lähihistoriaa (2017)

Esimerkki 1: tilastollinen konekäännös (SMT)

"Tämä ei ole ~~Ei~~ hyvä kokemus ihmisille ja aiomme puuttua asiaan", Facebookin tiedottaja sanoi ~~-~~ sähköpostitse lähetetyssä julkilausuma~~an~~ssa.

Tyypilliset ongelmat: taivutus, koherenssi, kadonneet sanat (tavuviiva sanassa "e-mailed"), merkityksen katoaminen.

Konekäännöksen lähihistoriaa (2017)

Esimerkki 1: neuroverkkokonekäännös (NMT)

”Tämä ei ole hyvä kokemus ihmisille ja me aiomme käsitellä asiaa”, Facebookin tiedottaja sanoi sähkö~~isessä~~postitse lähetetyssä lausunnossa.

Tyypilliset ongelmat: kontekstiin sopimattomat termivalinnat, merkityksien muuttuminen (tavuviiva sanassa ”e-mailed”).

Nykytila

Englanti–suomi-kieliparin NMT on kehittynyt huimasti vuoden 2017 jälkeen. Tämä näkyy WMT19:n testiaineiston pisimmistä käännöksistä.

WMT 2019, MSRA-NAO

Huolimatta siitä, että Katalonian itsenäisyyttä kannattavat puolueet saavuttivat elintärkeän, joskin niukan voiton viime joulukuussa järjestetyissä aluevaaleissa, ne ovat ponnistelleet pitääkseen vauhtia yllä tänä vuonna monien tunnetuimpien johtajiensa ~~ollessa joko itse pakotettuja~~jouduttua lähtemään maanpakoon tai ollessa pidätettyinä odottamassa oikeudenkäyntiä roolistaan kansanäänestyksen järjestämisessä ja sitä seuranneessa itsenäisyysjulistuksessa.

Nykytila

WMT 2019, MSRA-NAO

Vaikka Ri toisti tuttuja Pohjois-Korean valituksia siitä, että Washington vastustaa ydinaseriisunnan ”vaiheittaista”lähestymistapaa, jossa Pohjois-Korea palkittaisiin sen toimiessa asteittain, hänen lausuntonsa vaikutti merkittävältä sikäli, että siinä ei hylätty yksipuolista ydinaseriisuntaa suoralta kädeltä, kuten Pjongjang on aiemmin tehnyt.

Tyypillisiä ongelmia: Raskas, paljon infiniittirakenteita ja sivulauseita sisältävä tyyli, joka mukailee liiaksi lähdetekstin rakennetta. Hyvin erilaisilla rakenteilla eri kielissä ilmaistujen asioiden kääntäminen (self-imposed exile).

Nykytila

Toisaalta:

WMT 2019, lähdeteksti

And yet, it goes beyond being an informed citizen when you find yourself on hour six of watching a panel of experts debate Bob Woodward's use of "deep background" sourcing for his book "Fear," Paul Manafort's $15,000 ostrich-leather bomber jacket ("a garment thick with hubris," The Washington Post said) and the implications of Stormy Daniels's lurid descriptions of Mr. Trump's, um, anatomy.

Nykytila

WMT 2019, GTCOM-Primary

Ja silti, se menee pidemmälle kuin on perillä kansalainen, kun löydät itsesi tunnin kuusi katsomassa paneeli asiantuntijoiden keskustelua Bob Woodward käyttö ”syvä tausta”hankinta kirjaansa ”Pelko”, Paul Manafort n 15000 dollaria strutsi-nahka pommikone takki (”vaate paksu hubris”, Washington Post sanoi) ja vaikutukset Stormy Daniels lurid kuvaukset Mr. Trump, um, anatomia.

Ongelmat: Epätavallisen lähdetekstin (rakenne, terminologia) vuoksi konekäännös hukkaa merkityksen täysin eikä voi edes ylläpitää kieliopillisuutta (mikä on yleensä NMT:n vahvuus).

Onko suomi vaikea kieli konekäännöksen kannalta?

- Suomea pidettiin pitkään vaikeasti konekäännettävänä kielenä monimutkaisen taivutuksen ja joustavan sanajärjestyksen vuoksi.
- WMT:n tulosten ja yleisten vaikutelmien perusteella taivutus ja sanajärjestys eivät tuota ongelmia NMT:lle.
- Suomessa merkityksiä ilmaistaan usein eri rakenteilla kuin sen tärkeimmissä lähde- ja kohdekielissä, mutta erot eivät ole niin suuria, että ne vaikuttaisivat olennaisesti laatuun.

.

Suomi ei siis ole erityisen vaikea kieli konekääntää. Sille voidaan luoda toimivia konekääntimiä helposti kieliriippumattomilla menetelmillä, kunhan aineistoa on kohtuullisen paljon saatavilla.

Konekäännöksen käyttäjät

Tällä hetkellä konekäännöstä käyttävät pääasiassa käännöstoimistojen kääntäjät, bulkkikääntäjät ja alustatyöläiset:

- Suurilla kansainvälisillä (ja joillain suomalaisilla) käännöstoimistoilla on omia konekääntimiä (SDL, Lingsoft).
- Suurilla kansainvälisillä käännösten tilaajilla on omia konekääntimiä (Microsoft).
- Useat alustat on rakennettu konekäännöksen hyödyntämistä varten (Unbabel, Lilt).
- Monet käännöstoimistot ja freelance-kääntäjät käyttävät jotain kolmannen osapuolen kehittämää konekäännintä (DeepL, Google jne.).

Monet yritykset ja julkiset organisaatiot ovat viime aikoina myös ryhtyneet kokeilemaan konekäännöksen käyttöä.

Konekäännöksen soveltaminen käännöstyössä

Yleisin tapa soveltaa konekäännöstä käännöstyöhön on edelleen konekäännöksen tarjoaminen sellaisenaan esikäännöksenä, käännösmuistin kautta tai konekäännöslaajennuksen kautta.

Konekäännöksen soveltamismenetelmät voidaan jakaa järjestelmälähtöisiin, jotka toimivat kääntäjästä riippumatta, ja kääntäjälähtöisiin, jotka edellyttävät kääntäjän vuorovaikusta.

Konekäännöksen soveltaminen käännöstyössä

Järjestelmälähtöiset menetelmät

- konekäännös käännösmuistissa
- konekäännös erillisestä laajennuksesta
- räätälöitävä konekäännös
- mukautuva konekäännös.

Kääntäjälähtöiset menetelmät

- termien ja sääntöjen määritys konekäännökseen
- ennakoiva tekstinsyöttö konekäännösehdotuksesta
- konekäännöksen mukautuminen kääntäjän syötteeseen (interaktiivinen konekäännös)
- laadun automaattinen arviointi.

Konekäännöksen soveltaminen käännöstyössä

Kehittyneempiä menetelmiä kehitetään ja otetaan käyttöön hiljalleen.

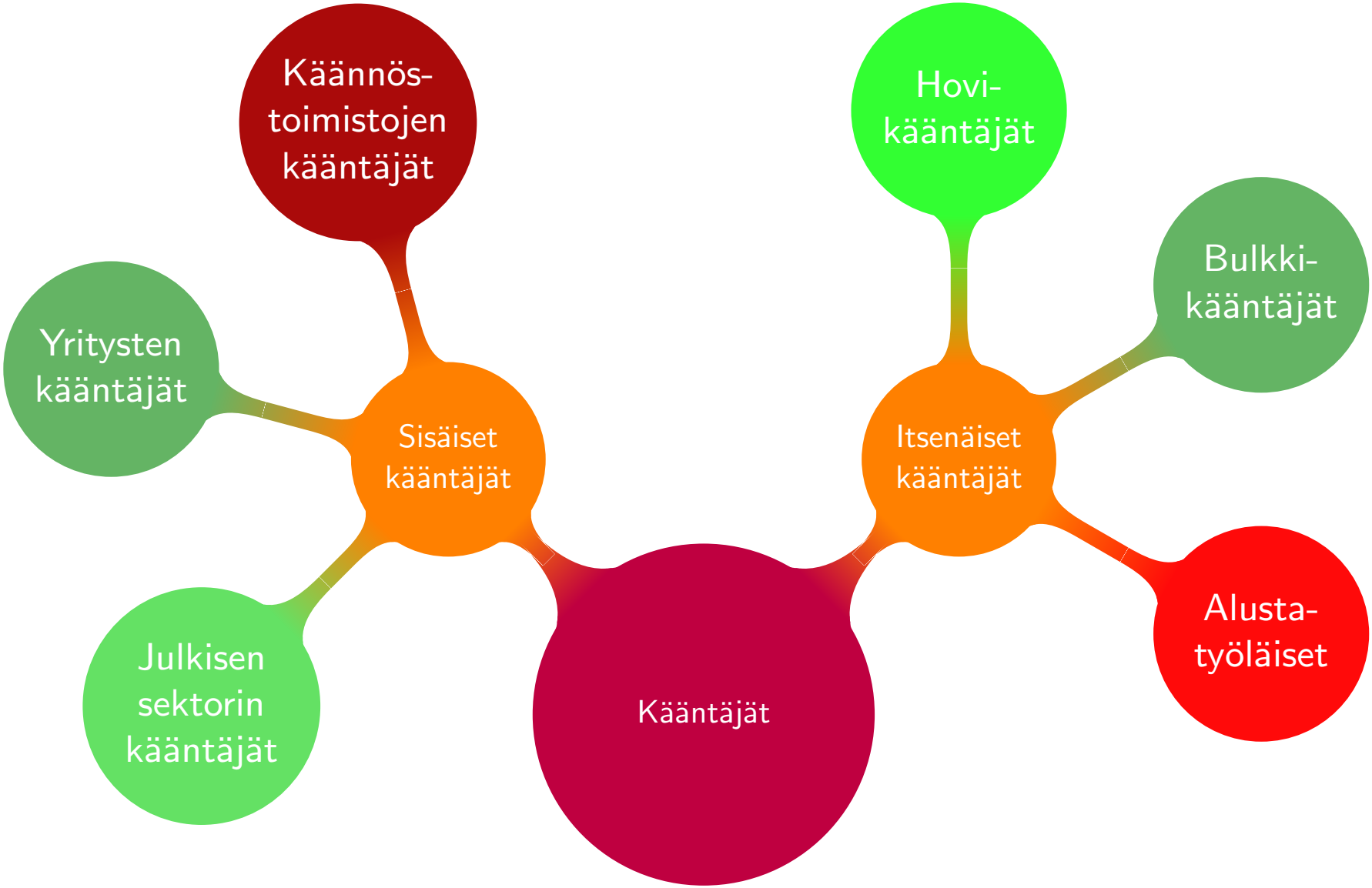
Kääntäjien tulevaisuuden kannalta olennaista on kääntäjälähtöisten menetelmien kehittyminen, sillä ne edellyttävät uusien taitojen opettelemista. Toisaalta kääntäjälähtöiset menetelmät vahvistavat kääntäjien asemaa asiantuntijoina.

Konekäännöksen vaikutukset eri ryhmissä

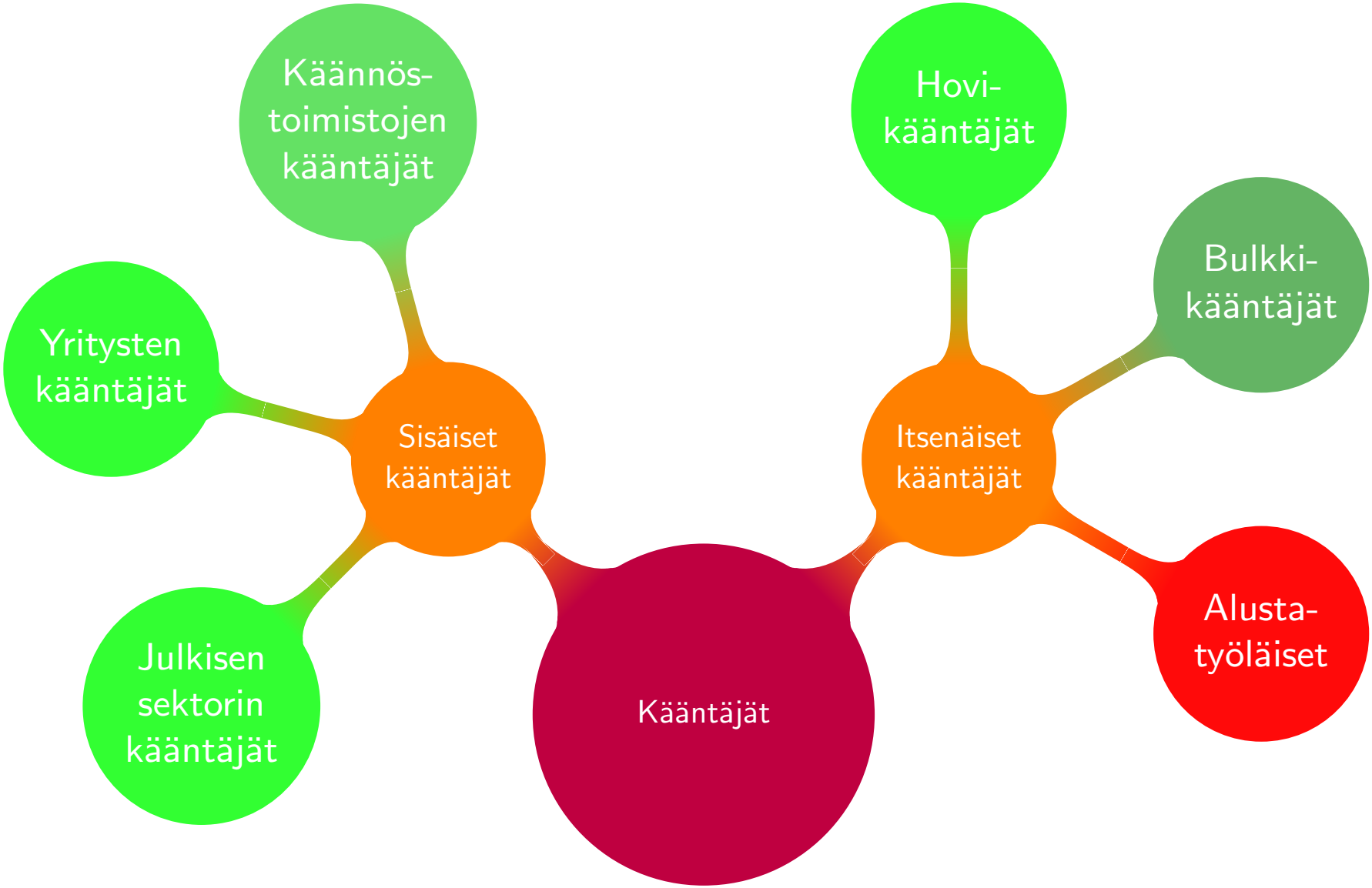
Kääntäjien ryhmien välillä on eroja, jotka vaikuttavat konekäännöksen hyödyntämiseen. Tärkeimmät erot ovat kääntäjien mahdollisuuksissa päättää omista työtavoistaan sekä laatu- ja tehokkuusvaatimuksissa.

Vapaus valita työkalut vaikuttaa siihen, miten helppoa konekäännös on omaksua osaksi työtä. Laatuvaatimukset taas vaikuttavat siihen, miten suurta osaa konekäännöksestä voi käyttää.

Kääntäjien ryhmät: vapaus päättää työtavoista



Kääntäjien ryhmät: laatuvaatimukset



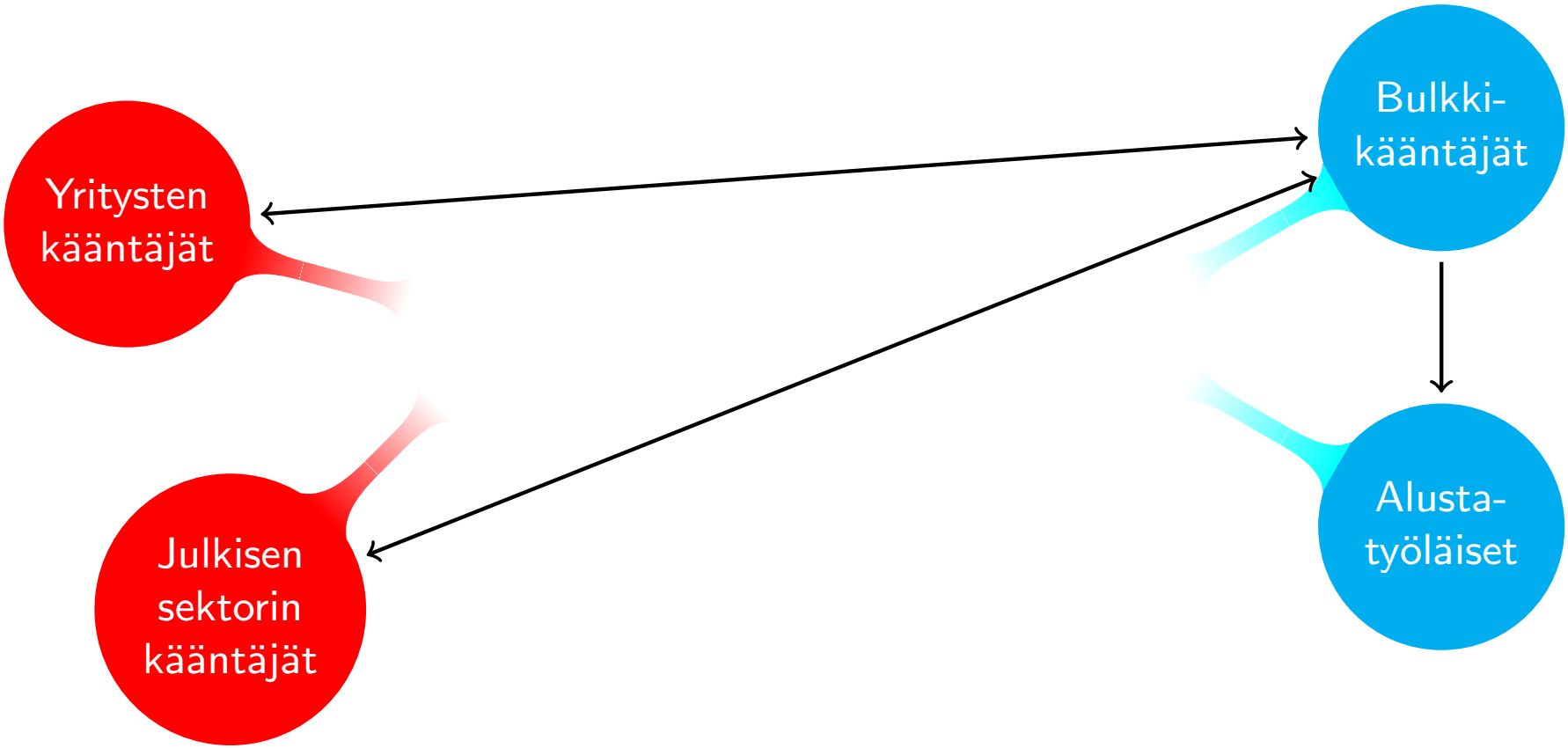
Konekäännöksen vaikutukset ryhmien välillä

Konekäännöksen yleistymiselle voi olla useita vaikutuksia eri kääntäjäryhmien suhteisiin:

1. Konekäännös laskee ulkoistettujen käännösten hintoja, joten käännösten ulkoistaminen on houkuttelevampaa.
2. Konekäännös lisää sisäisten kääntäjien tehokkuutta, jolloin ulkoistamisen tarve vähenee.
3. Osa bulkkikääntäjien töistä siirtyy alustakääntäjille.

Todennäköisesti kaikkia ilmiöitä esiintyy. Ratkaisevaa on, miten hyvin sisäisiä kääntäjiä työllistävät tilaajaorganisaatiot omaksuvat konekäännöksen ja kuinka paljon laadusta realistisesti voidaan joustaa eri aloilla.

Kääntäjien ryhmät: vaikutukset ryhmien välillä



Konekäännöksen merkitys sisäisille kääntäjille

Sisäisten kääntäjien kannattaa siirtyä hyödyntämään konekäännöstä ja erityisesti kääntäjälähtöisiä menetelmiä mahdollisimman nopeasti, sillä se

- vahvistaa heidän asiantuntijuuttaan
- estää ylhäältä päin pakotetut muutokset
- vähentää ulkoistuksen houkutteluvuutta.

Tämä on luonnollisesti mahdollista vain organisaatioissa, joissa kääntäjillä on mahdollisuuksia vaikuttaa työmenetelmiinsä.

Konekäännöksen merkitys itsenäisille kääntäjille

Itsenäisten kääntäjien kohdalla tilaajat tulevat vääjäämättä edellyttämään konekäännöksen käyttöä lähes kaikille osa-alueille.

Siirtymäajasta tulee kuitenkin todennäköisesti pitkähkö, ja sen aikana kääntäjät voivat myös omaehtoisesti käyttää konekäännöstä työssään (jos sopimus asiakkaan kanssa sallii sen).

Konekäännöksen merkitys itsenäisille kääntäjille

Käännösten hinnoittelu on aina ollut melko sattumanvaraista. Korvaus perustuu merkki-, sana- tai sivumäärään, mutta työmäärä voi vaihdella valtavasti.

Käännösmuistien ja fuzzy-luokkien myötä hinnoittelu muuttui yhä monimutkaisemmaksi, ja konekäännöksellä on samankaltainen vaikutus.

Perimmäinen ongelma on, että käännöstyötä voi muun asiantuntijatyön tapaan hinnoitella reilusti vain käytettyjen tuntien mukaan.

Tuntikorvauksen ongelmana taas on vaikeus arvioida kustannuksia sekä vaikutus kääntäjän työmotivaatioon.

Konekäännöksen merkitys itsenäisille kääntäjille

Konekäännös näkyy hinnoittelussa yleensä könttäalennuksena uusien sanojen hintaluokassa.

Konekäännöksen hyödyllisyyydessä on kuitenkin suuria eroja töiden välillä. Esimerkiksi seuraavat asiat vaikuttavat siihen:

1. lähdetekstin rakenne (kokonaisia lauseita vai fragmentteja)
2. lähdetekstin laatu (kirjoittajan äidinkieli, kirjoitusvirheet)
3. lähdetekstin segmenttien pituus (ei mielellään liian lyhyitä tai pitkiä)
4. konekääntimen soveltuvuus aihepiiriin tai asiakkaalle
5. yhdenmukaisuusvaatimukset jo käännettyjen segmenttien kanssa.

Konekäännöksen merkitys itsenäisille kääntäjille

Konekäännöstöiden hinnoittelu on siis vieläkin epäjohdonmukaisempaa kuin käännösmuistitöiden.

Taustaoletuksena on, että korvausten suuruusvaihtelut tasaantuvat pidemmällä aikavälillä ja että keskiarvoinen korvaus on asiantuntijatyön edellyttämällä tasolla.

Käytännössä hintavaihtelut ovat jossain määrin ennustettavissa, joten sekä tilaajat että kääntäjät voivat harrastaa kermankuorintaa tai arbitraasia.

Selkein esimerkki tästä on se, että kääntäjät voivat yleensä tutustua lähdetekstiin ennen työn hyväksyntää, ja erittäin työläät työt on helppo tunnistaa ja jättää väliin. Tiettyjä aihepiirejä on myös syytä välttää (laki- ja lääketieteellinen teksti).

Konekäännöksen merkitys itsenäisille kääntäjille

Konekäännös lisää arbitraasin mahdollisuuksia, sillä tietyt aihepiirit ja asiakkaat ovat nopeampia kääntää kuin muut.

Siirtymäaikana on siis mahdollista saada enemmän taloudellista etua konekäännöksestä, mutta ennen pitkää hinnoittelumallin on muututtava.

Yksi vaihtoehtoinen hinnoittelumalli on laskea konekäännösalennus sen mukaan, kuinka paljon valmis käännös eroaa konekäännöksestä.

Toinen vaihtoehto on siirtyä aikaperusteiseen hinnoitteluun tai aika- ja sanaperusteiseen hybridihinnoitteluun. Nämä tavat soveltuvat erityisen hyvin käännösalustoille, joissa voidaan seurata ajankäyttöä tarkkaan.

Konekäännöksen merkitys itsenäisille kääntäjille

Tällä hetkellä ratioonalisin toimintatapa itsenäisille kääntäjille on hyödyntää konekäännöstä itsenäisesti mahdollisimman paljon (tottumisen ja tehokkuuden vuoksi).

Toinen prioriteetti on pyrkiä maksimoimaan oma korvaus harjoittamalla edellä kuvattua arbitraasia eli kieltäytymällä töistä, joihin konekäännös ei sovellu.

Töiden valikointi edesauttaa järkevämpien hinnoittelumallien syntymistä, ja siirtymäaikana se lisää omia tuloja ja kannustaa tilaajia käyttämään konekäännöstä vain sopivissa töissä.