eTranslation työkaluna Euroopan komission käännöstoimessa – käytännön kokemuksia

Erkka Vuorinen
Laatuvastaava
Euroopan komissio,
käännöstoimen pääosasto,
suomen kielen osasto
24.10.2018

Esityksen sisältö

- Taustaa
- Komission konekääntämisen historiaa
- eTranslation
- eTranslation-käännösten laadusta
  - Konekäännösten analyysi
  - Käyttäjäkysely
- Jatkonäkymiä

Taustaa

EU:n (toimi)elinten käännösorganisaatiot

Komissio

Parlamentti

EU:n neuvosto

Talous- ja sosiaalikomitea

Euroopan keskuspankki

Käännöskeskus

Alueiden komitea

EU:n tuomioistuin

Tilintarkastus-tuomioistuin

Euroopan investointipankki

Komission käännöstoimen pääosaston (DGT) käännöstuotanto

- Koko DGT vuonna 2017:
  n. 2 028 000 sivua
  (joista ulkoistettuja 30,6 %)

- Vuoden 2018 arvio: 2,1 milj. sivua
  - kesäkuussa 2018 kaikkien aikojen ennätys: 314 000 sivua.

- 23 kieliosastoa (+ iirin yksikkö), n. 1 560 kääntäjää

Suomen osaston käännöstuotanto

- Viime vuosina käännetty n. 80 000 s. vuodessa (v. 2017 ulkoistettiin 23,6 %)
- 55 virkamieskääntäjää + 2 sopimussuhteista kääntäjää (lokakuu 2018)

Mitä komissiossa käännetään?

- Yli 30 tekstiluokkaa
  - lainsäädäntö ja sen perustelut, sopimukset, raportit, tiedonannot, ilmoitukset, kirjeenvaihto kansallisten viranomaisten ja kansalaisten kanssa, lehdistötiedotteet, esitteet ja muu pr-materiaali, nettisivut, puheet, artikkelit, sisäisen hallinnon lomakkeet ja asiakirjat...
- 50–60 prosenttia *suomennoksista* puhdasta lainsäädäntöä
- Myös ei-juridista aineistoa runsaasti
- Erittäin laaja aihepiirien kirjo

Komission konekääntämisen historiaa

Konekäännösjärjestelmät (1)

Systran (sääntöpohjainen): 1975–2010

- Ei kattanut kaikkia EU:n virallisia kieliä
- Joissakin kielipareissa tyydyttäviä tuloksia

MT@EC (tilastopohjainen): kesäk. 2013–

- Useissa kielipareissa varsin hyviä tuloksia (esim. EN-FR, EN-PT)
- Joissakin kielipareissa vaatimattomia tuloksia (EN-FI, EN-ET, EN-HU, EN-DE...)
- Edelleen käytössä/käytettävissä

Konekäännösjärjestelmät (2)

eTranslation (neuroverkkopohjainen): marrask. 2017–

- Asteittainen käyttöönotto
  - EN-FI ensimmäisten kieliparien joukossa, mutta ensisijaiseksi MT-resurssiksi työnkulkujärjestelmään (MT@EC:n tilalle) vasta heinäkuussa 2018
- Kesällä 2018 kaikki kieliparit *englanti – muu virallinen kieli* käytössä
- Käyttäjiä:
  - muut EU:n toimielimet
  - jäsenvaltioiden viranomaiset, EU:n online-palvelusivustot (IMI, SOLVIT, TED…) ja tätä kautta kansalaiset ja yritykset

European Commission

eTranslation

eTranslation-käännin

- Taustalla EU:n valtava käännös-tekstikorpus (Euramis-keskusmuisti)
  - yli miljardi käännössegmenttiä eri kielillä
  - kasvaa kuukausittain n. 2,5 miljoonalla segmentillä

- EN—FI-käännin ”koulutettu” n. 10 miljoonaa uniikkia käännössegmenttiä käsittävällä aineistolla
  - Käytetty avoimen lähdekoodin NMT-välineistöjä: *Nematus*, *OpenNMT*, *Marian* (tällä hetkellä tärkein)
  - Perustana olevaa datakorpusta on tarkoitus päivittää (täydentää, siivota) n. 4—6 kuukauden välein

eTranslation-käännin – tunnuslukuja

EN-FI

- **BLEU**: 0,5409
- **Translation error/edit rate (TER)**: 0,4855

(MT@EC: BLEU: 0,4269; TER: 0,5610)

eTranslation-käännösten laadusta (EN-FI)

1) Konekäännösten analyysi suomen kieliosastossa (9/2018—)

Ensi vaiheessa:

- eTranslation- ja MT@EC-käännösten vertaileva virheanalyysi
- Tekstiotos: 6 erityyppistä tekstiä:
  - lehdistötiedote (CETA-sopimus 1 vuotta)
  - esite (luomuviljely)
  - tiedotusvälineille tarkoitettu esittelyteksti (EU-Aasia-yhteydet)
  - komission lausunto EP:n päätöslauselman johdosta (matkamittarien manipulointi)
  - 2 säädöstä
    - ehdotus neuvoston päätökseksi (EU:n kanta Minamatan yleissopimuksen osapuolten konferenssin kokouksessa)
    - komission täytäntöönpanoasetus (*Wolbachia*-bakteerin hyödyntäminen torjuntatarkoituksiin (hyttysissä))

Aineisto

- yht. runsaat 30 sivua, lähes 500 segmenttiä
- tekstejä, joiden käännösmuistivastaavuudet (matches) olivat pienet (enimmillään 18 %, vähimmillään 0,25 % 100-prosenttisia vastaavuuksia)
- analysoidut käännökset puhtaita konekäännöksiä
- havainto: suppeastakin aineistosta nousevat selvästi tyypilliset piirteet esille

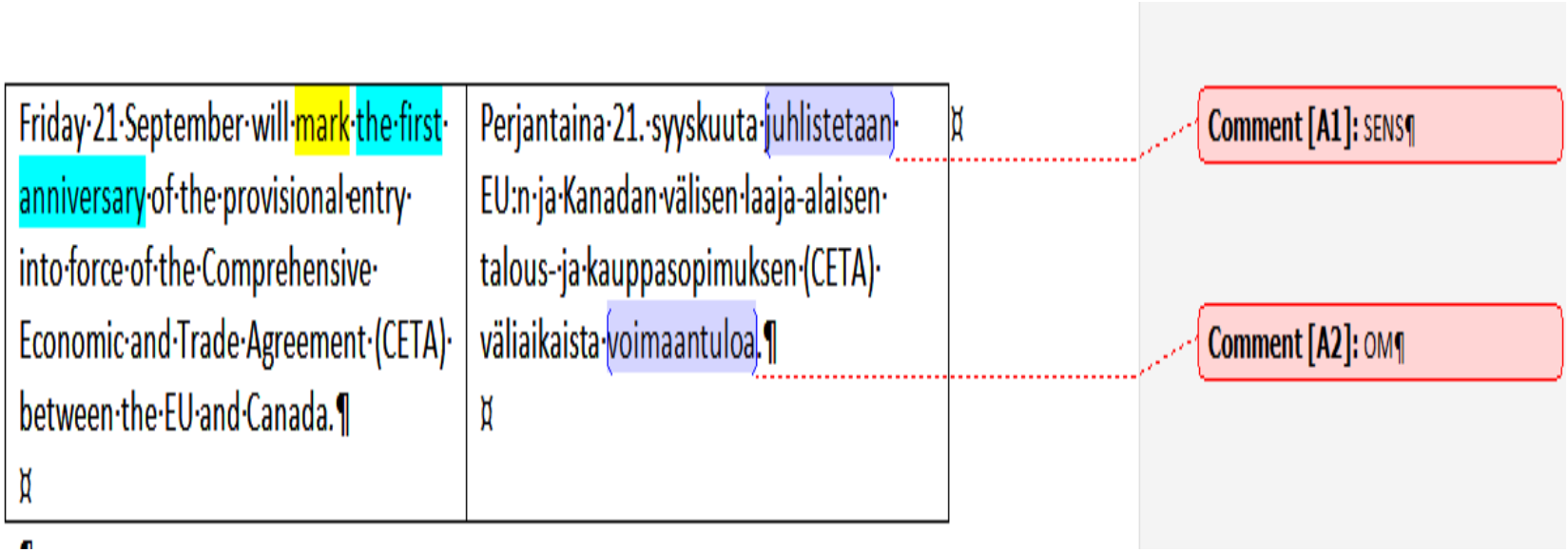
Ensimmäisen vaiheen lähestymistapa:

- Arvioitiin MT@EC-käännös (SMT) ja eTranslation-käännös (NMT) samalla tavalla kuin arvioitaisiin alihankkijan toimittama työ
- Määritettiin virheelliset kohdat sekä kulloisenkin virheen tyyppi ja vakavuus
- Virheluokat:

merkitys (SENS)	**kielioppi (GR)**
terminologia (TERM)	**selkeys ja idiomaattisuus(CL)**
poisjättö (OM)	**oikeinkirjoitus (SP)**
referenssien käyttö (RD)	**välimerkit (PT)**

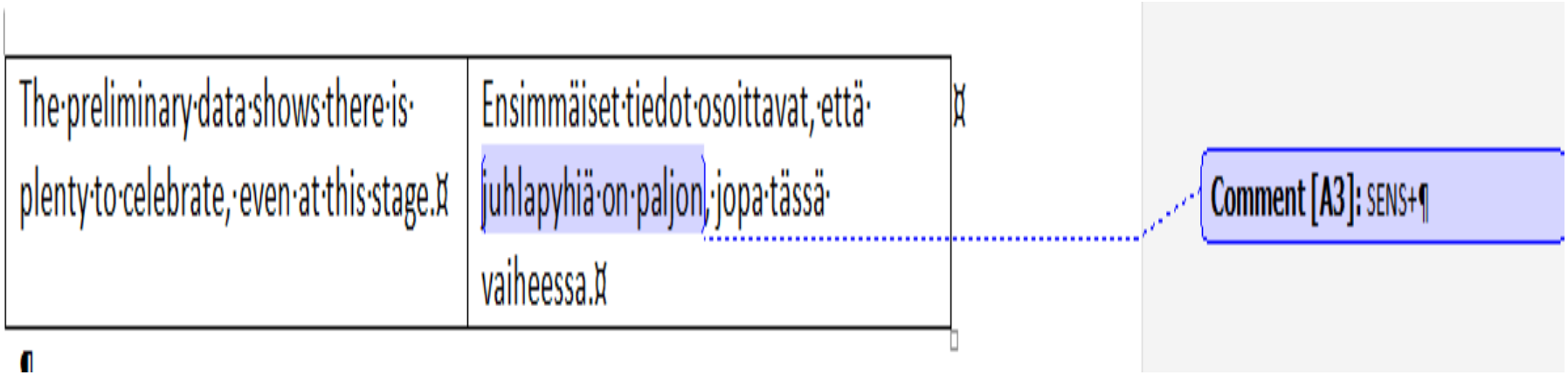
- Huom. virhetyypin ja virheen vakavuuden määrittäminen subjektiivista ja osin hankalaa → **tulokset ovat suuntaa antavia**, eivät eksakteja

Esim. (virheitä)



Vrt. Perjantaina 21. syyskuuta tulee kuluneeksi vuosi siitä, kun EU:n ja Kanadan laaja-alainen kauppasopimus (CETA) tuli väliaikaisesti voimaan.

Esim. (vakava virhe)



Vakava virhe: virhe tai puute, joka heikentää olennaisesti käännöksen käyttökelpoisuutta tai uskottavuutta

Havaintoja (1)

Yleistä:

eTranslationin tulokset eri virhemittareilla selvästi parempia kuin MT@EC:n

- vähemmän virheitä
- vähemmän vakavia virheitä
- vähemmän tolkuttomia segmenttejä

Eri virhetyyppi hallitsevana:

- eTranslation: merkitysvirheet
- MT@EC: kielioppivirheet

Virhetiheys vaihtelee huomattavasti teksteittäin.

Havaintoja (2)

eTranslation:

Vähemmän virheitä:

- keskim. n. –35 %

Vähemmän vakavia virheitä:

- keskim. n. –30 %

Vähemmän tolkuttomia segmenttejä (nonsense):

- keskim. n. –60 % (MT@EC: keskim. n. 20 % segmenteistä)

Nonsense-segmentit – esim.

”WHY DO WE NEED BETTER CONNECTIVITY FOR EUROPE & ASIA?”

”MIKSI EUROPE YHTEYDESSÄ EESSI EUROOPAN YHTEYDESSÄ JÄRJESTELMÄN YHTEYDESSÄ?”

Havaintoja (3)

Virheitä kuitenkin paljon molemmissa:

- eTranslation: keskim. n. 1,5/segmentti (alin n. 1,1, ylin n. 2,3)
- MT@EC: keskim. n. 2,3/segmentti (alin n. 2, ylin n. 3,2)

Samoin vakavia virheitä paljon molemmissa:

- eTranslation: keskim. n. 30 % kaikista virheistä
- MT@EC: keskim. n. 30 % kaikista virheistä

Havaintoja (4)

Eniten esiintyvät virhetyypit:

- eTranslation: merkitysvirheet: n. 40 % kaikista virheistä
- MT@EC: kielioppivirheet: n. 35 % kaikista virheistä
  - erityisesti sijamuotojen, virkkeen osien järjestyksen ja yhdyssanamuodosteiden virheet

Mutta *vakavista* virheistä:

- eTranslation: merkitysvirheet: n. 55 %
- MT@EC: merkitysvirheet: n. 50 %

Kaiken kaikkiaan merkitysvirheiden kokonaismäärässä ei ollut kovin suurta eroa.

Havaintoja (5)

Virhetiheyden vaihtelu teksteittäin

eTranslation:

- komission lausunto: n. 1,1 virh./segmentti
- lehdistötiedote: n. 1,3 virh./segmentti
- komission täytäntöönpanoasetus: n. 2,3 virh./segmentti

eTranslation – laatuongelmia

- Merkityksen vääristyminen
- Lisätyt ja toisteiset ainekset
- Olennaisten tekstielementtien poisjätöt
- Terminologian sattumanvarainen vaihtelu (useita eri vastineita samassa käännöksessä)
- Elliptisten ja kontekstiltaan niukkojen tekstikohtien (esim. otsikot) käännökset sattumanvaraisia
- Kääntimen “keksimät” termit, sanat ja taivutukset
- Nimien ja lyhenteiden/lyhennesanojen vääristyminen
- Lähdetekstissä (esim. suorissa lainauksissa) käytetyn epämuodollisemman/puhekielisemmän ilmaisun aih. ongelmat

eTranslation – laatuongelmia

Erityisesti säädösten kääntämiseen liittyviä ongelmia:

- Käännöksen sotkeutuminen monimutkaisten rakenteiden takia
  - esim. monimutkaiset/-kerroksiset upotukset tai genetiivirakenteet
- Viittaukset toisten säädöskohtien sisältöön usein epätarkkoja
- Vanhentuneista vakioilmaisuista peräisin olevat elementit

Esim.

Merkityksen vääristyminen:

”Exports are up overall and many sectors have seen impressive increases.”

”Vienti on kaiken kaikkiaan yleistä, ja monilla aloilla on tapahtunut huomattavaa lisäystä.”

Lisätyt/toistetut ainekset:

”In the German region of Schleswig-Holstein, farmers and researchers (...) are comparing...”

”Saksan Schleswig-Holsteinin alueella Schleswig-Holsteinin osavaltion viljelijät ja tutkijat vertaavat...”

Esim.

Poisjätöt:

”However, three fungal diseases are causing problems in shallot cultivation.”

→ ”Salottisipulit aiheuttavat kuitenkin kolmisinakin ongelmia.”

Terminologiavaihtelu:

odometer readings: matkamittarilukemien, matkamittarin lukemien, matkamittarin lukemiseksi

Elliptinen/niukka konteksti:

Otsikko ”Organic inspiration for European agriculture”

→ ”EU:n maatalouden orgaaninen inspiraatio”

Esim.

Keksityt termit/sanat/taivutukset:

’organic cattle farm’ →’luomunauha’

’European blockchain network’ → ’eurooppalainen kauppalauttaverkko’

”vilja → *vilkkaan; hyttynen → *hyttynesissa

Vääristyvät nimet ja lyhenteet/lyhennesanat:

Wolbachia → **Wolbachialle;* Jean-Luc Millécamps → Jean-Luc *Milléleirin

EIP-AGRI → EIP-AGRI, mutta myös *Organic and AGRI

Puhekieli/suorat lainaukset:

Komissaari: ”This is something I intend to bring up with my Canadian counterparts at the Joint Committee next week.” →

”Aion tuoda yhteen kanadalaisten kollegoideni kanssa ensi viikolla.”

Esim. (säädökset)

Monimutkaiset rakenteet:

COMMISSION IMPLEMENTING DECISION (EU) .../... of XXX pursuant to Article 3(3) of Regulation (EU) No 528 of the European Parliament and of the Council on mosquitoes non-naturally infected with *Wolbachia* used for vector control purposes

→

COMMISSION IMPLEMENTING DECISION (EU) .../...
of XXX
pursuant to Article 3(3)
of Regulation (EU) No 528
of the European Parliament and of the Council
on mosquitoes
non-naturally infected
with *Wolbachia*
used for vector control purposes

Esim. (säädökset)

NMT:

KOMISSION TÄYTÄNTÖÖNPANOPÄÄTÖS (EU) .../...

Annettu XXX,

Euroopan parlamentin ja neuvoston asetuksen 528/2012 3 artiklan 3 kohdan mukaisesti sellaisten hyttysten esiintymisestä hyttynesissa, joissa Wolbachia on luonnostaan tartunnan vektorivalvonnassa

Vrt.

KOMISSION TÄYTÄNTÖÖNPANOPÄÄTÖS (EU) .../...,

annettu XXX,

vektorien torjuntatarkoituksiin käytettävistä *Wolbachia*-bakteeritartunnan muuten kuin luonnollisesti saaneista hyttysistä Euroopan parlamentin ja neuvoston asetuksen 528/2012 3 artiklan 3 kohdan nojalla

Esim. (säädökset)

Epätarkat viittaukset toisten säädöskohtien sisältöön:

NMT: Hyttyset ovat asetuksen asetuksen 528/2012 3 artiklan 1 kohdan g alakohdassa tarkoitettuja haitallisia organismeja, koska niillä voi olla ei-toivottu esiintyminen tai haitallinen vaikutus ihmisiin tai eläimiin

Vrt. ko. säädöskohta (määritelmä): ’haitallisilla eliöillä’ kaikkia eliöitä, myös taudinaiheuttajia, joiden esiintyminen ei ole toivottavaa tai joilla on haitallinen vaikutus ihmisiin, heidän toimintoihinsa tai heidän käyttämiinsä tai valmistamiinsa tuotteisiin taikka eläimiin tai ympäristöön;

Yleiset päätelmät analyysista

- eTranslationin EN-FI-käännösten laatu ei ole lähellä ihmiskääntäjän tuottaman käännöksen laatua (*pariteetti* edelleen kaukana).

- eTranslationin myötä on kuitenkin otettu suomenkin osalta selvä askel eteenpäin konekääntämisessä.

- Tilastopohjaisen kääntimen (MT@EC) tulokset jääneet kehitystyöstä huolimatta suomen osalta heikohkoiksi. Lisäksi jatkokehitysmahdollisuudet ovat hyvin rajallisia (perusongelmia vaikea ratkaista).

2) Käyttäjäkysely

- Tehtiin DGT:n suomen osastossa lokakuussa 2018.

- Kartoitettiin, missä määrin kääntäjät käyttävät neuroverkkopohjaista eTranslationia ja minkälaisia kokemuksia heillä on EN-FI-kääntimestä.

- Vastanneet: 45 kääntäjää (yli 80 % osaston kääntäjistä).

Käyttäjäkysely – tuloksia (1)

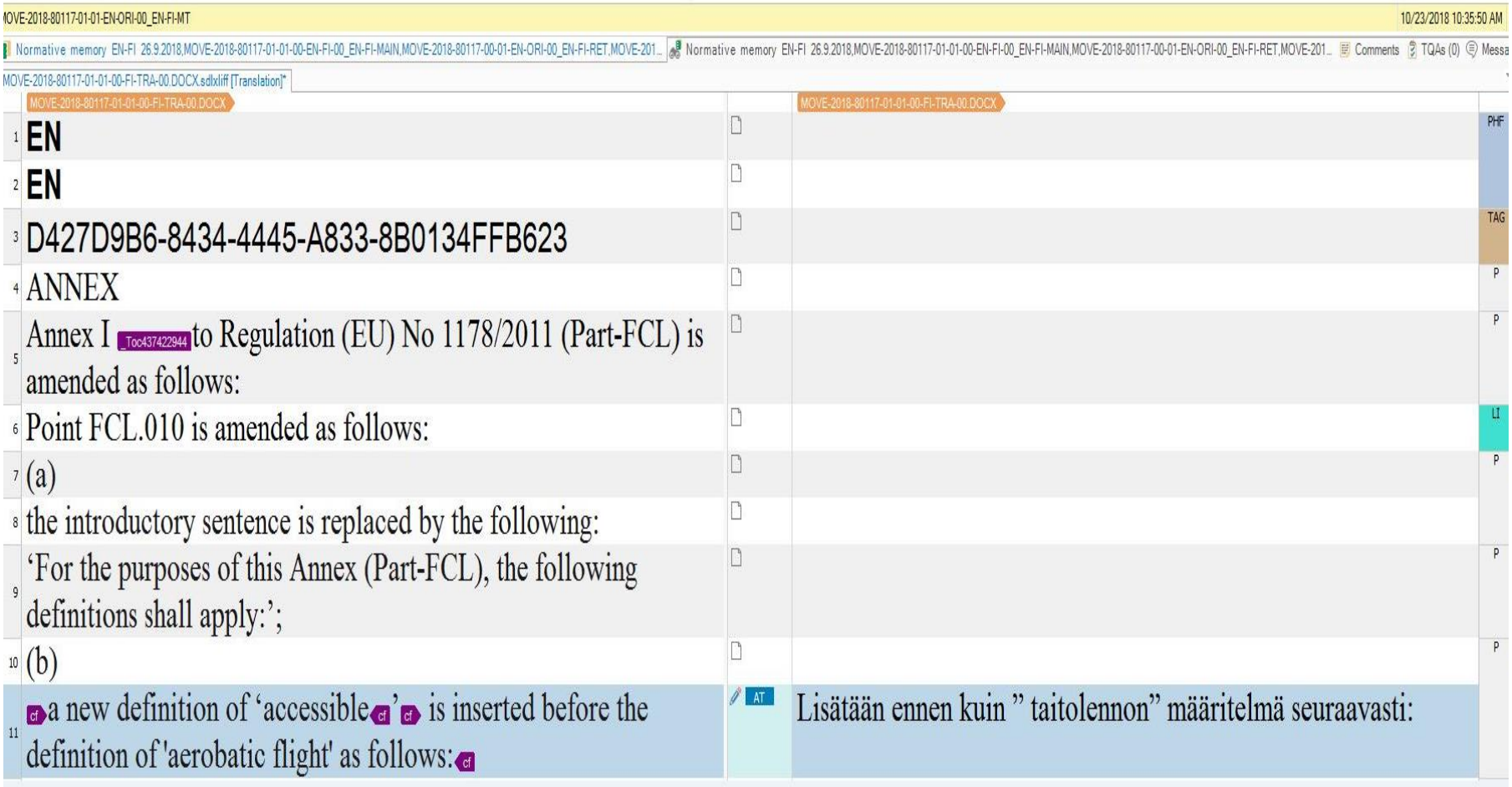
- eTranslationia käyttää kaikissa tai useimmissa käännöstoimeksiannoissa apuna 36 kääntäjää (n. 80 % vastaajista).
- 4 kääntäjää ei käytä lainkaan eTranslationia.
- Aiempaa tilastopohjaista MT@EC:tä käytti kaikissa tai useimmissa toimeksiannoissa 22 kääntäjää.
- MT@EC:n käyttö on käytännössä loppunut (MT@EC-käännöksen saa edelleen käyttöön erikseen tilaamalla).

Käyttäjäkysely – tuloksia (2)

- eTranslationin käyttäjistä 1/3 pitää sen tuottamia käännöksiä erittäin hyödyllisinä ja 2/3 jossain määrin hyödyllisinä.
- Lähes 3/4:n mielestä eTranslationin käännösten laatu on selvästi parempaa kuin MT@EC:n.
- Käyttäjistä
  - 60 % suosii *full segment* -moodia (= kääntäjä ehdottaa kokonaisia segmenttejä)
  - 40 % suosii *autosuggest*-moodia (= kääntäjä ehdottaa segmentin osia)

European Commission

Full segment -moodi



Autosuggest-moodi

Käyttäjäkysely – tuloksia (3)

Ruusuja

Voi **nopeuttaa työtä** vähentämällä kirjoitustarvetta

- tarjoaa usein kelvollista tavaraa, jota voi käyttää jopa sellaisenaan tai ainakin vähin muutoksin (etenkin lyhyissä segmenteissä)
- taivutusmuodot ja muu kielioppi nyt useimmiten kunnossa

Voi **antaa virikkeitä** käännösratkaisuihin ja muotoiluihin

- ”löytää” toisinaan asioita, joita käännösmuisti-retrieval ei ole löytänyt
- voi tarjota ideoita ja termejä, joiden avulla pääsee alkuun hankalan segmentin kanssa

Käyttäjäkysely – tuloksia (4)

Risuja

Kaunis pinta, **silkkoa sisältä?**

- Kieliopillinen oikeellisuus voi peittää merkityksen ja viittaussuhteiden vääristymisen
- ”Keksityt” sanat ja termit
- Tärkeitä elementtejä ja jopa kokonaisia segmenttejä voi jäädä kokonaan kääntämättä
- Virheet kuukausien nimissä, erisnimissä ja kirjaimin kirjoitetuissa numeroissa

Konekäännösehdotus voi **häiritä omaa ajattelua**

- Käännös voi lähteä väärään suuntaan.

Käyttäjäkysely – tuloksia (5)

Ruusuja & risuja

Termit

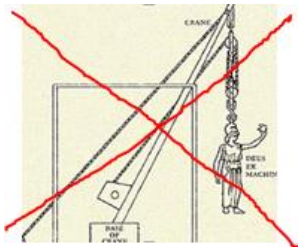
- Osan mielestä eTranslation tuottaa termejä luotettavammin ja yhdenmukaisemmin kuin MT@EC, osa on päinvastaista mieltä.

Käyttäjäkysely – tuloksia (6)

Havainnot pähkinänkuoressa

Myönteinen suhtautuminen (lisääntynyt käyttö)

Realismi (parantunut työkalu, ei *deus ex machina*)



Tarkkuus ja varovaisuus

Jatkonäkymiä

Mitä jatkossa?

- eTranslationin kehitystyö jatkuu
  - seuraavassa päivityksessä mukaan Euroopan parlamentin ja mahdollisesti unionin neuvoston dataa
  - datasiivoukset
  - konekäännöksen esi- ja jälkikäsittelysääntöjen mukauttaminen
- Erikoisalakohtaiset kääntimet (esim. *Public health* -käännin)
- Kielidatan kokoaminen jäsenmaista (ELRC) ja sisällyttäminen järjestelmään
- Käyttösovellusten laajentaminen Verkkojen Eurooppa -välineen (CEF) puitteissa
  - esim. avuksi EU:n puheenjohtajamaiden käännöstarpeisiin

Mitä jatkossa?

- Konekääntämisen tehokkuusvaikutusten selvittäminen DGT:ssä
  - suomen osasto (yhdessä ranskan osaston kanssa) paraikaa mukana tehokkuustutkimuksessa

- Kielikohtaiset laatuanalyysit, esim.
  - tarvittavien editointitoimenpiteiden yksityiskohtainen tarkastelu
  - mahdollisten lopputuotteen laadussa ilmenevien vaikutusten selvittäminen

KIITOS!

Hartelijk dank!

Muito obrigado!

Ďakujeme vám veľmi pekne!

Go raibh maith agaibh!

Tack så mycket!

Thank you!

Mulţumesc!

¡Muchas gracias!

Σας ευχαριστούμε πολύ!

Nuoširdžiai dėkojame!

Mockrát děkujeme!

Merci beaucoup !

Vielen Dank!

Suur tänu!

Liels paldies!

Nirringrazzjawk ħafna!

Grazie mille!

Mange tak!

Najlepša hvala!

Dziękujemy za uwagę!

Много Ви благодарим!

Köszönjük szépen!

Hvala lijepa!