# NEIRONU MAŠĪNTULKOŠANA MAZAJĀM VALODĀM – ATSKATS UZ 2018. GADU

Mārcis Pinnis

21.02.2019.

# NEIRONU MAŠĪNTULKOŠANA Tildē ~~MAZAJĀM VALODĀM~~ – ATSKATS UZ 2018. GADU

Mārcis Pinnis

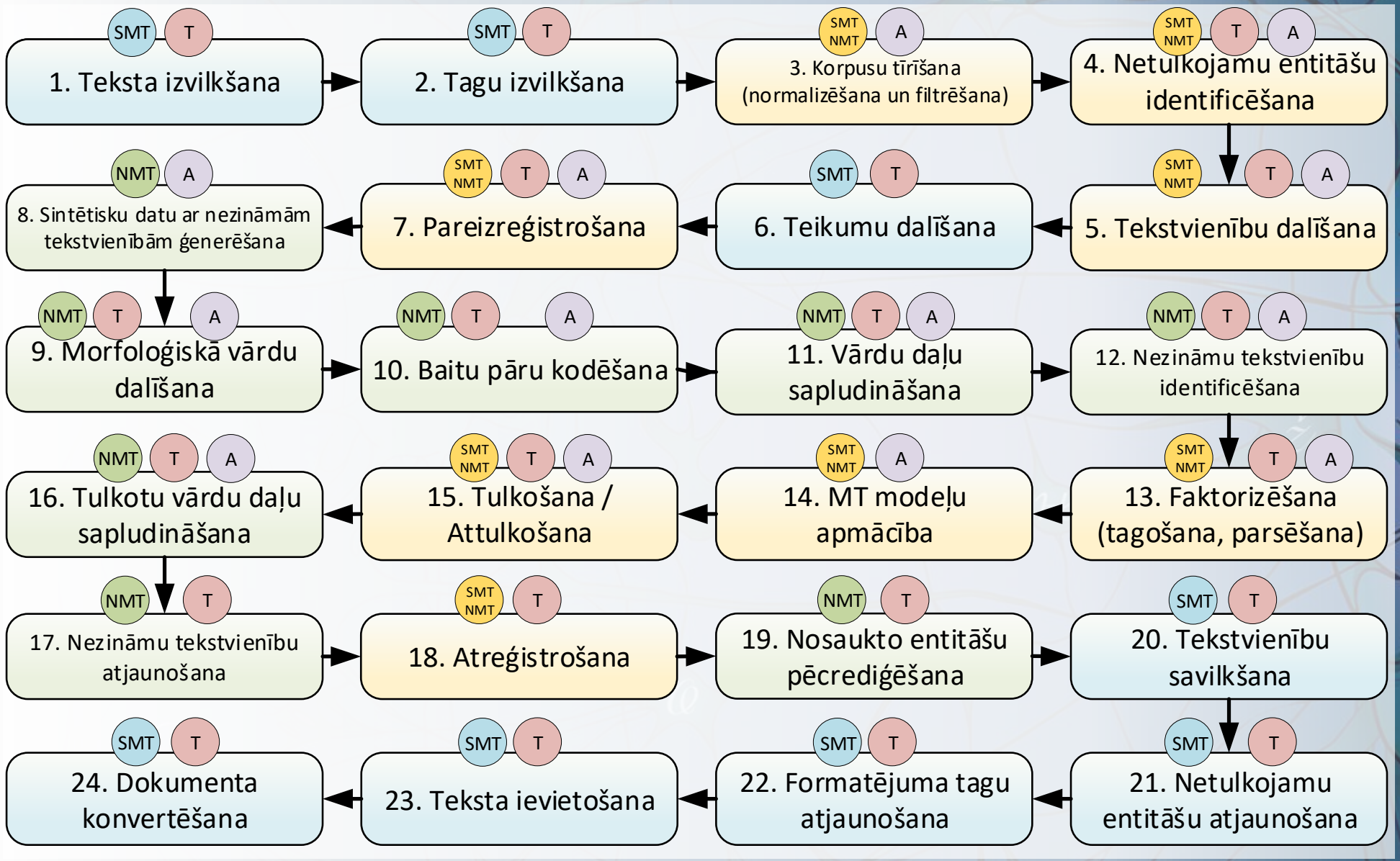
21.02.2019.

# Saturs

- Metodes
- Dalība Mašīntulkošanas konferences sacensībās
- Risināmās problēmas

# Metodes

# MT darbplūsma (Tilde MT)




TILDE

SMT — Kopš SMT ieviešanas

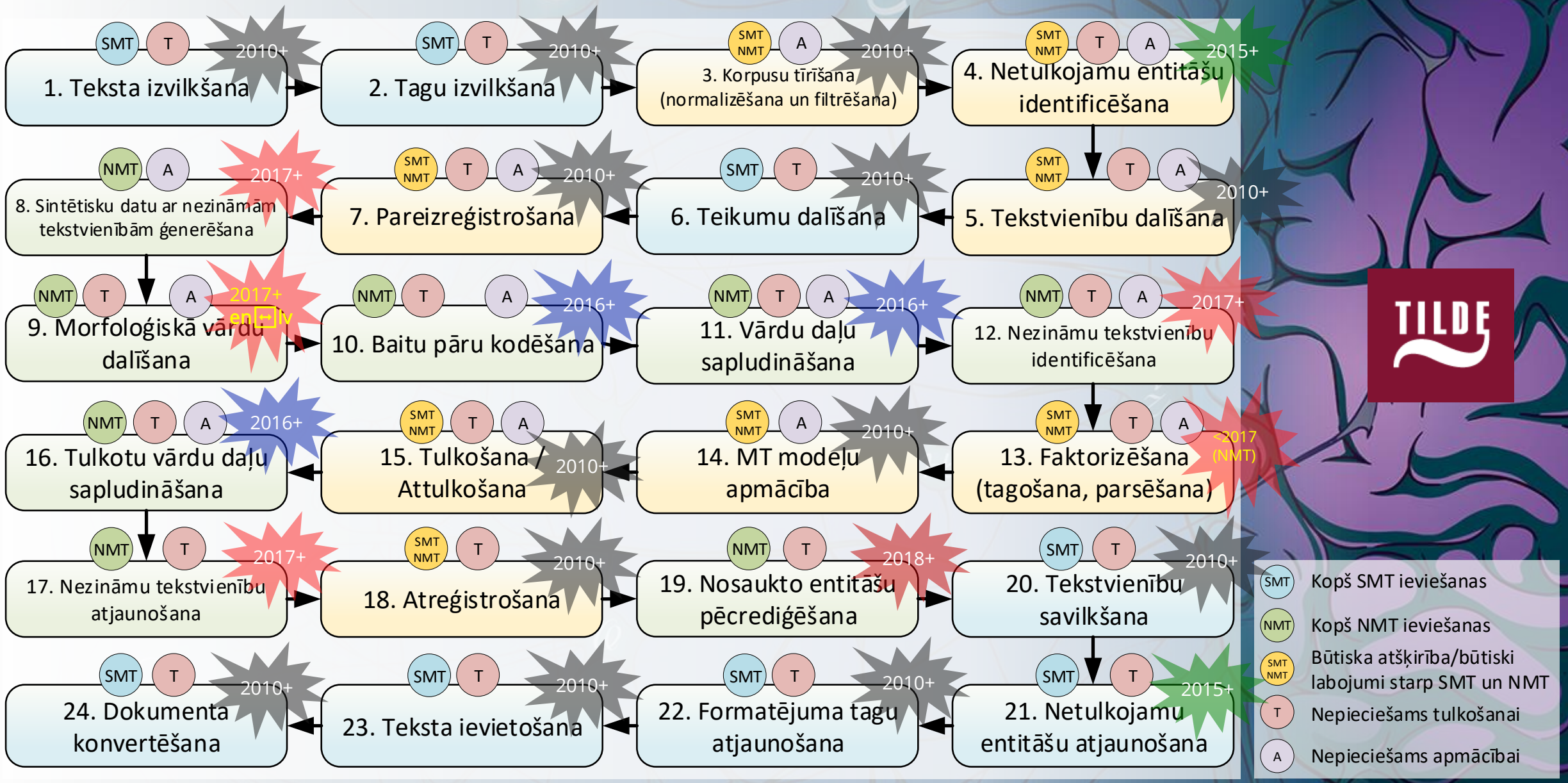
NMT — Kopš NMT ieviešanas

SMT NMT — Būtiska atšķirība/būtiski labojumi starp SMT un NMT

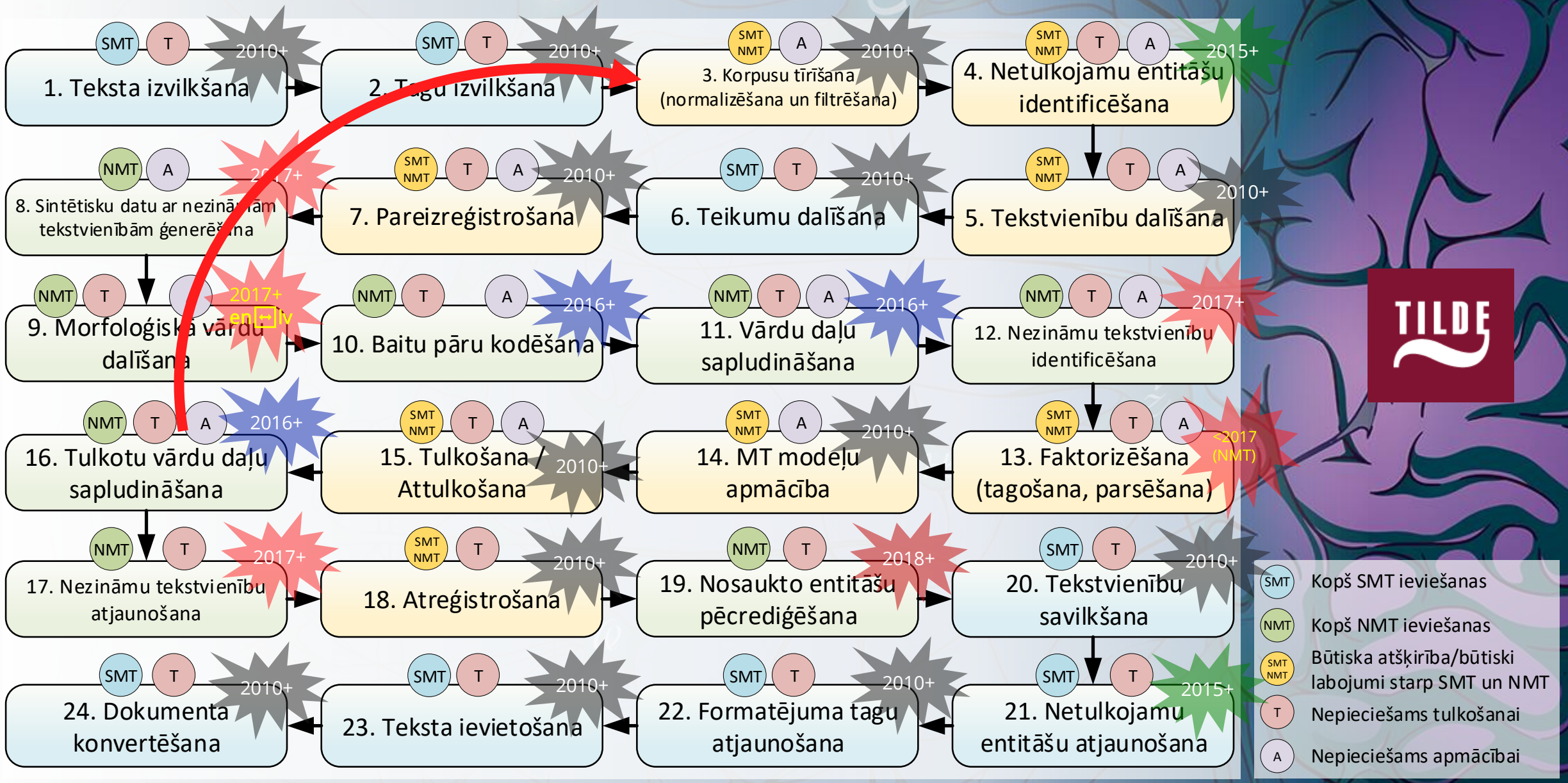
T — Nepieciešams tulkošanai

A — Nepieciešams apmācībai

# MT darbplūsma (Tilde MT)



# MT darbplūsma (Tilde MT) ar attulkošanu



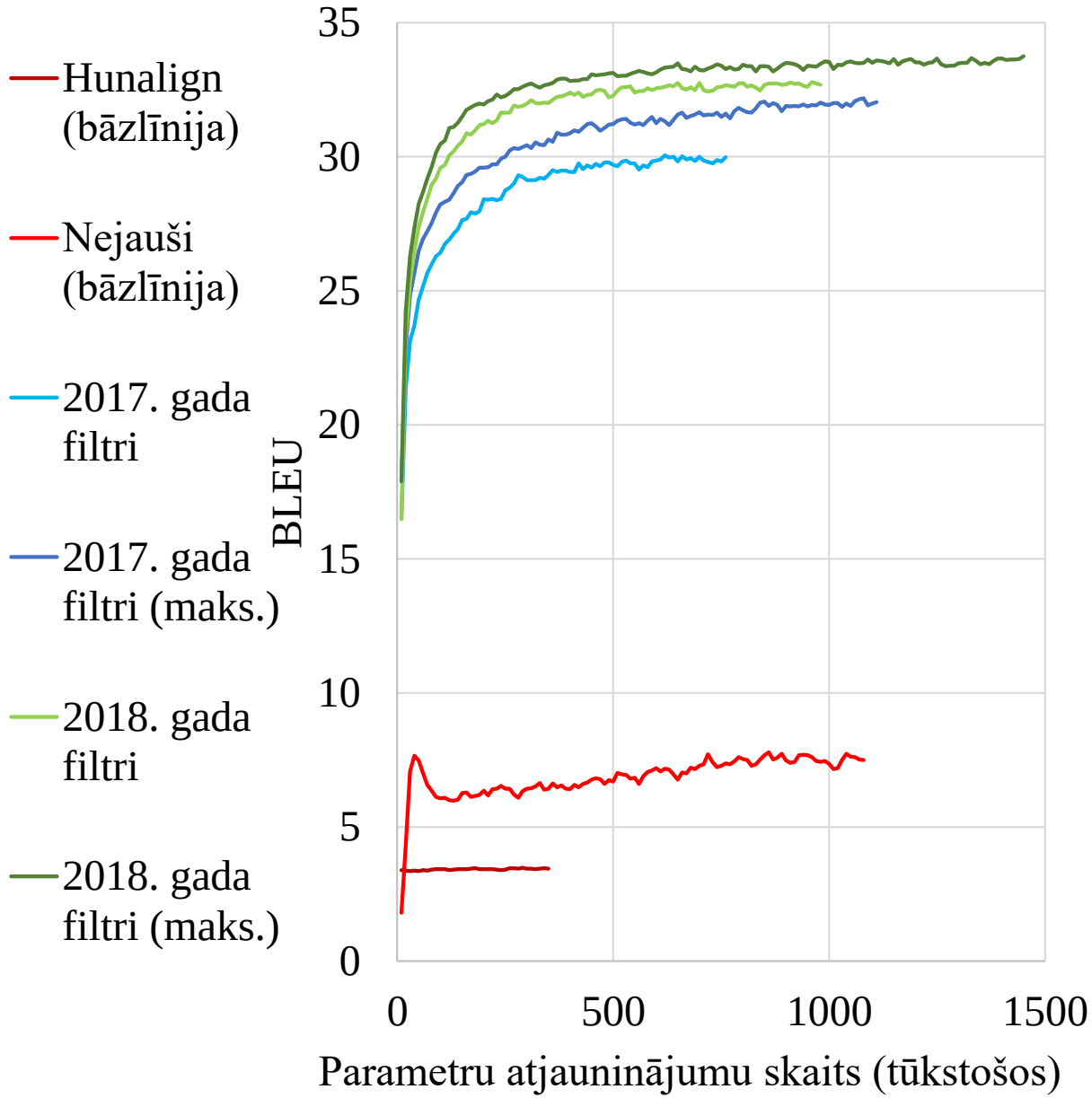
# Datu tīrīšana

Datu kvalitāte (tīrība) ir svarīga situācijās, kad ir vēlme iegūt lietojamu (konkurētspējīgu) rezultātu.

Filtru uzdevums - atrast un izdzēst:

- neparalēlus datus;
- datus nepareizās valodās;
- sabojātus datus.

*Piemērs no vācu-angļu datu filtrēšanas uzdevuma WMT 2018*

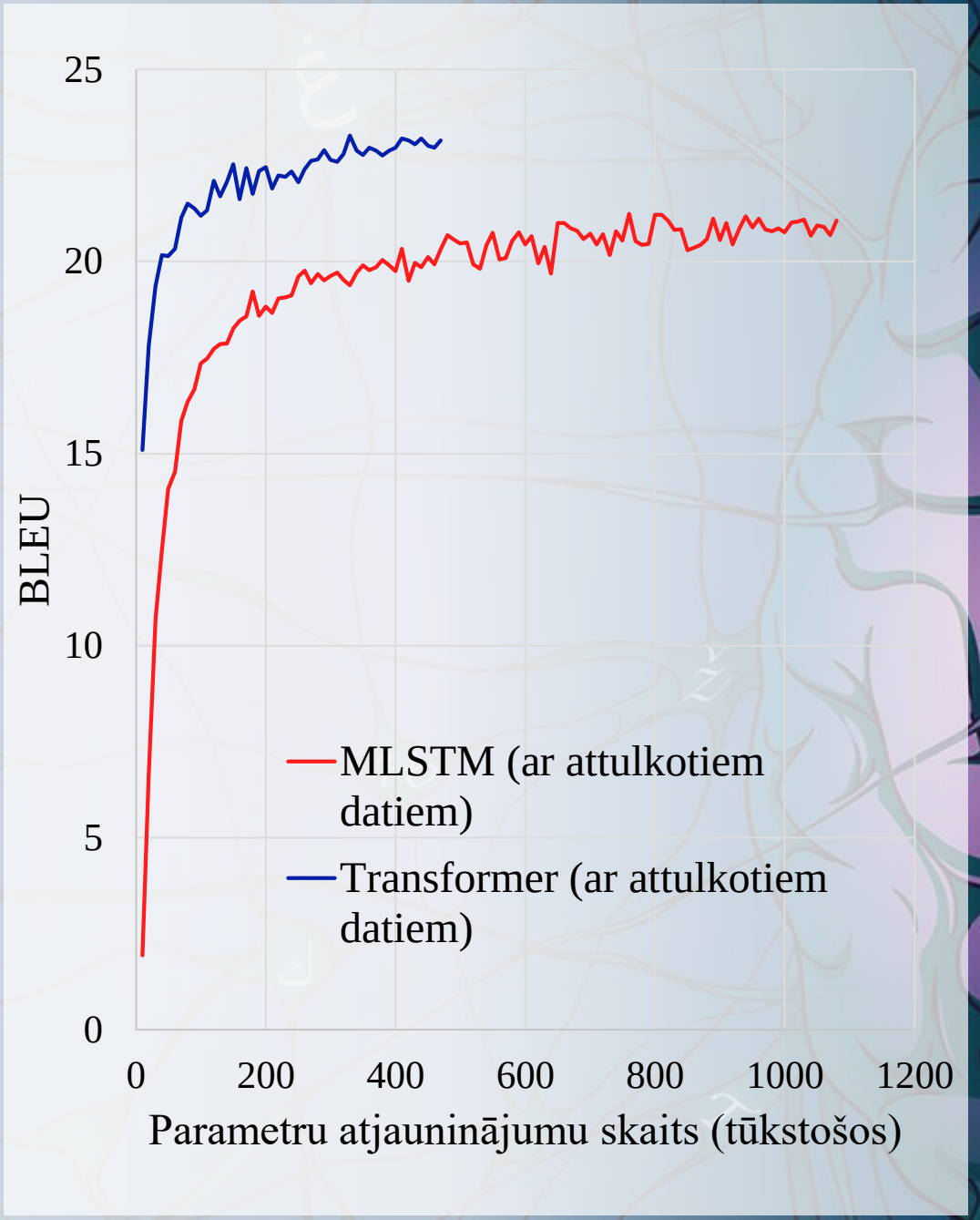


# Neironu tīklu arhitektūras

2017. gadā - MLSTM (*multiplikatīvās garās īstermiņa atmiņas rekurentie neironu tīkli*)

2018. gadā - Transformer (*jeb «pašuzmanības»*) neironu tīkli.

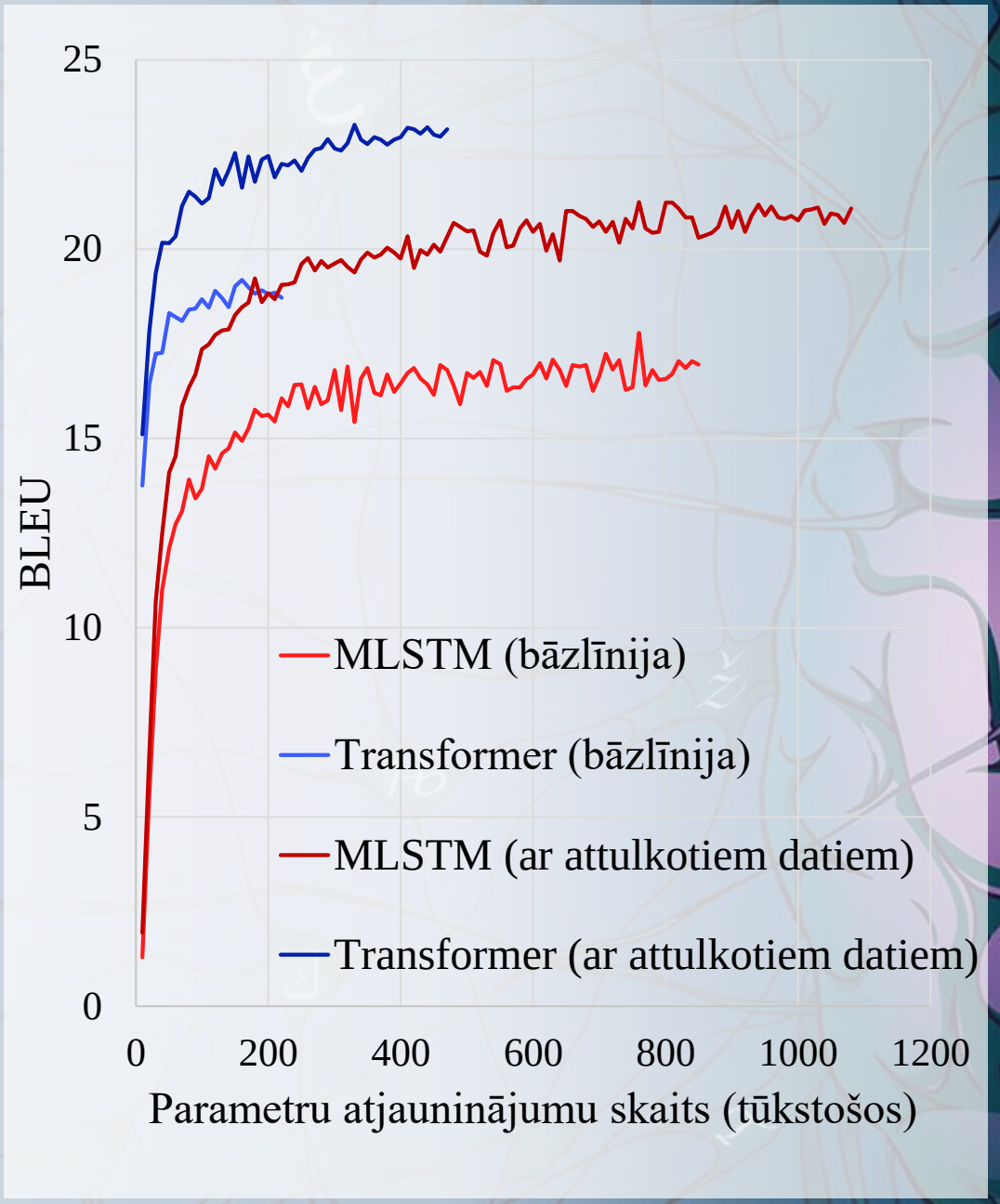
*Piemērs no Tildes angļu-igauņu WMT 2018 eksperimentiem*



# NMT pielāgošana jomai

**Attulkošana** - efektīvākā NMT sistēmu pielāgošanas metode, situācijās, kad jomas paralēlo datu nav.

*Piemērs no Tildes angļu-igauņu WMT 2018 eksperimentiem*



# Kur valodu specifika?

2017. gadā mēs pētījām:

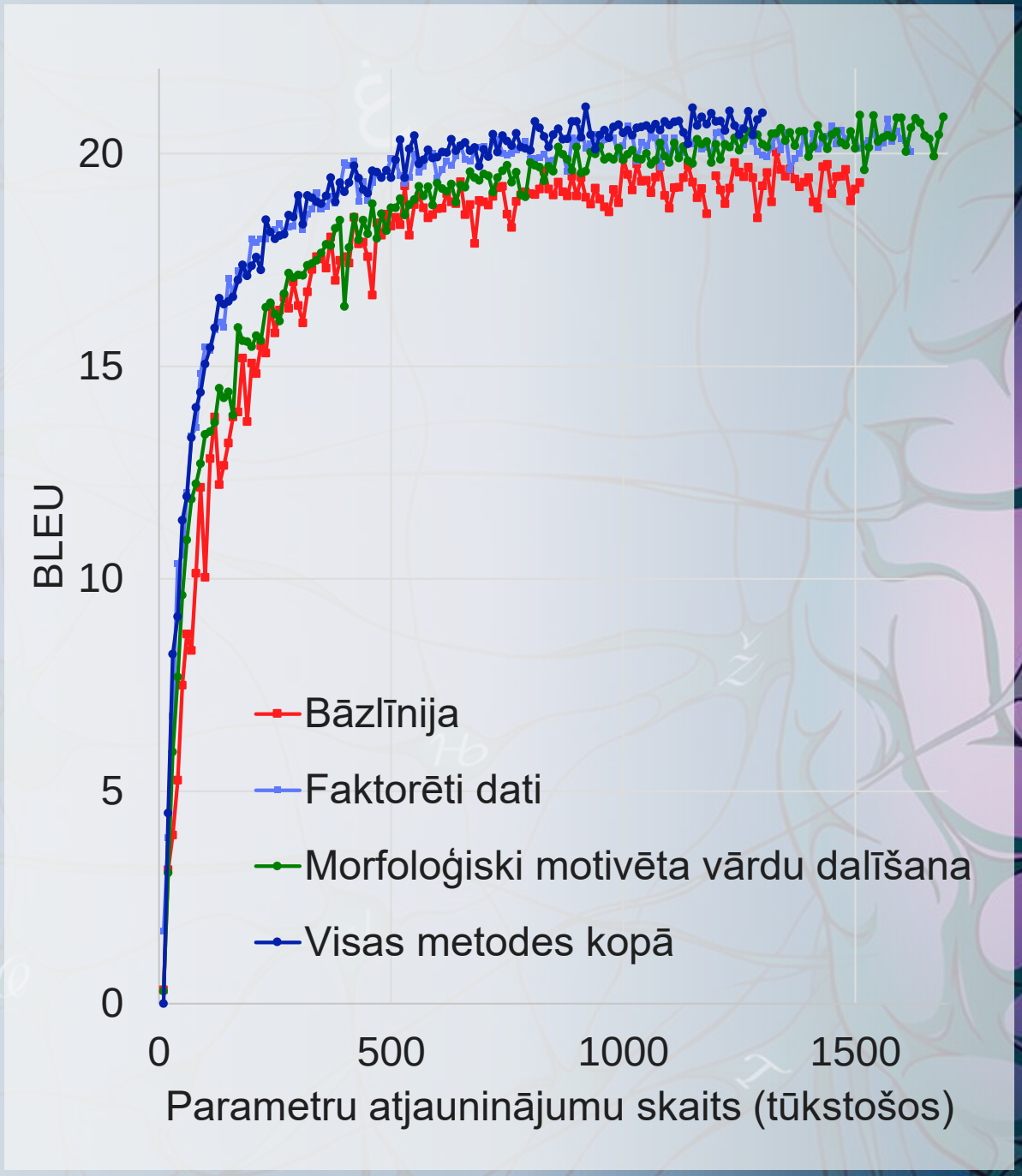
- morfoloģiski motivētu vārdu dalīšanu
- faktorētu datu izmantošanu

Ieguvumi - ~+1 BLEU

Tīrāku datu un attulkošanas (labākas pielāgošanas) ieguvums ir lielāks (≥5 BLEU)

Jaunāku modeļu ieguvums lielāks (≥2 BLEU)

Vispirms jāatrod labākās no valodām neatkarīgās metodes...



# Dalība Mašīntulkošanas konferences sacensībās

# Sacensības

**Uzdevums**: izmantojot organizatoru sagatavotus datus (*ierobežotu datu scenārijs*) vai citus datus (*neierobežotu datu scenārijs*), izstrādāt mašīntulkošanas sistēmu ziņu tulkošanai, ar kuru ir jāpārtulko novērtēšanas datu kopa; pārtulkotā datu kopa ir jāiesniedz vērtēšanai.

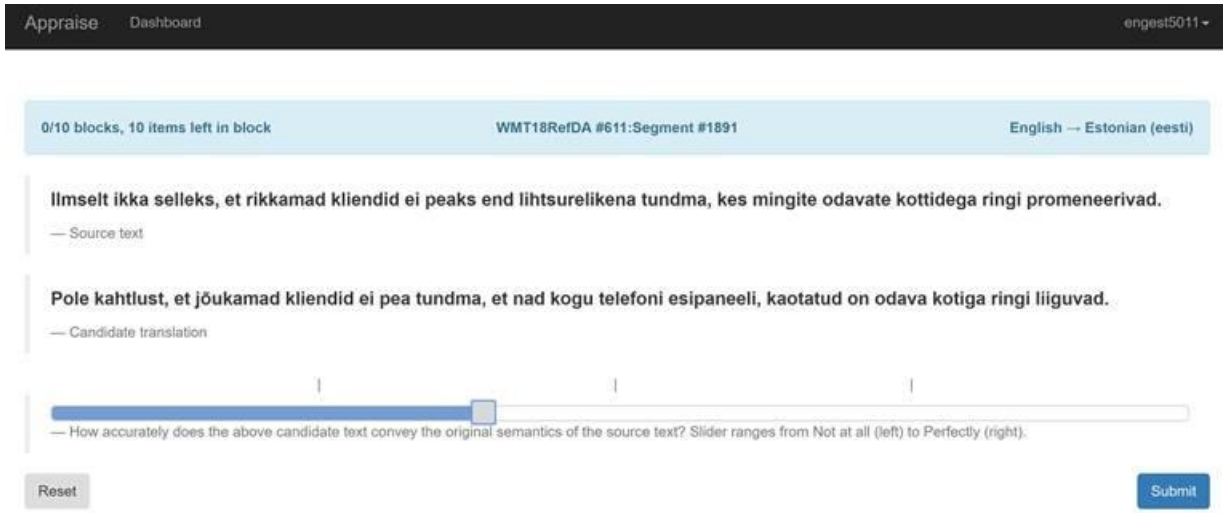
**Dalībnieki**: mašīntulkošanas jomas pētnieki un izstrādātāji no pētniecības centriem un uzņēmumiem no visas pasaules

# Sacensības

**Vērtēšana**: Tulkojumi tiek vērtēti, izmantojot:

1) *automātiskas vērtēšanas metodes*,
2) *manuālas vērtēšanas metodes, iesaistot cilvēkus* (tulkotājus, redaktorus, valodu ekspertus, kā arī «pūli»).

**Cilvēkvērtējums tiek uzskatīts par uzticamāku vērtējumu!**



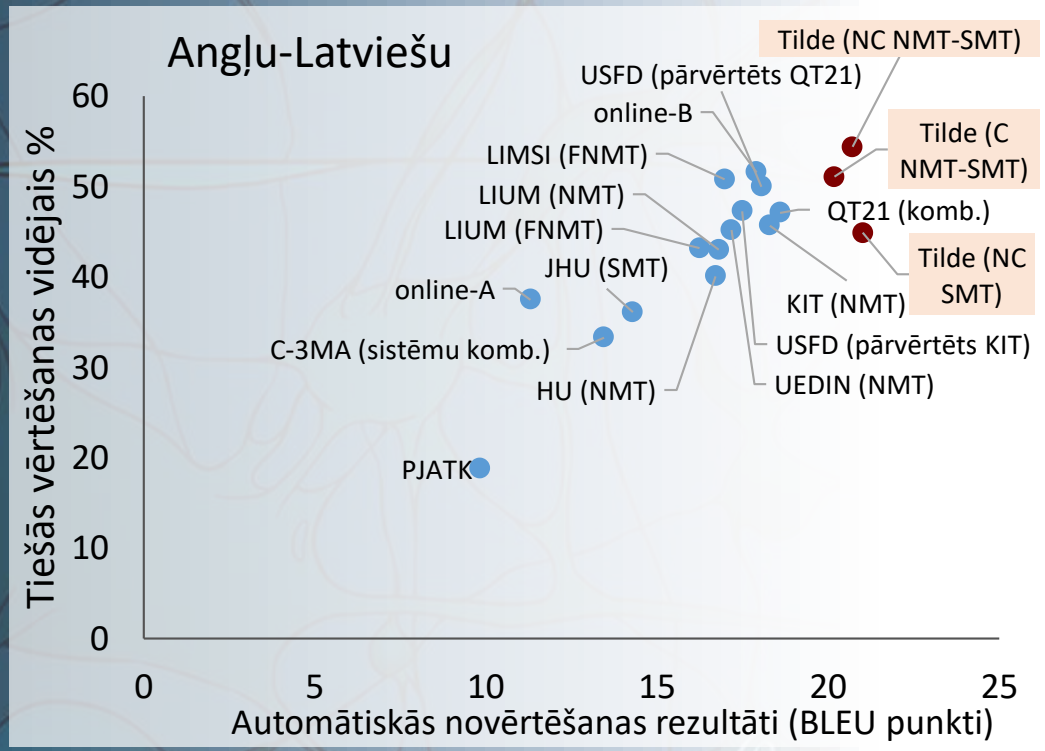

Attēls no WMT 2018 mašīntulkošanas sistēmu vērtēšanas sistēmas (http://wmt18.appraise.cf)

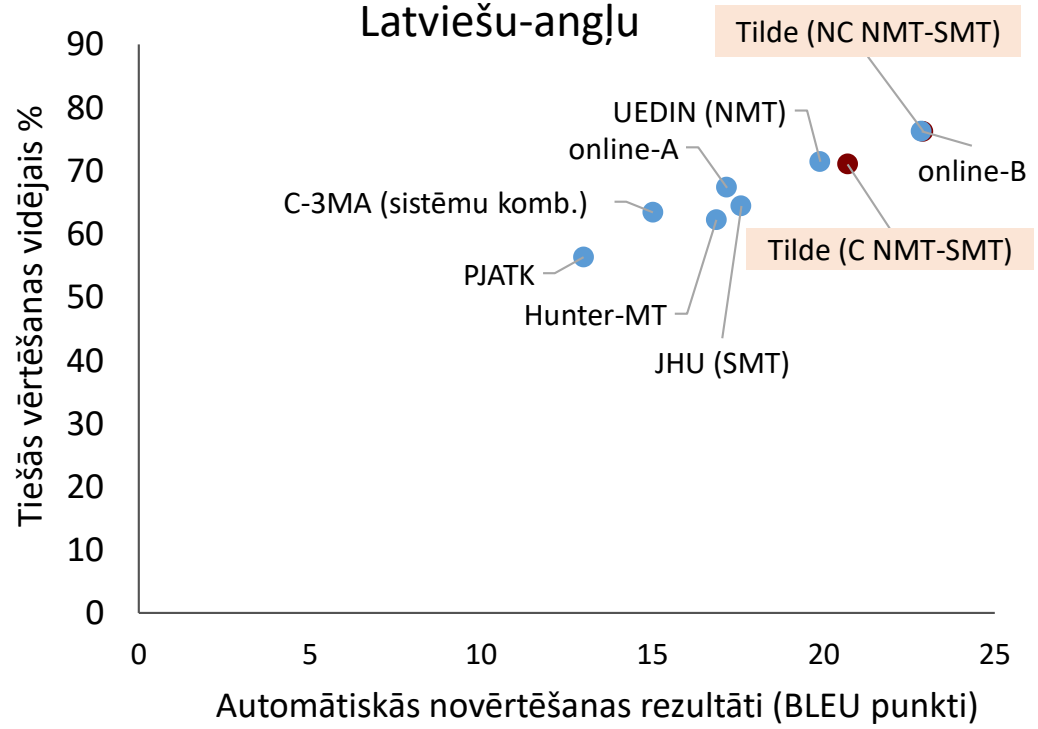
# Sacensības

Mēs sacensībās piedalījāmies:

- **2017. gadā**, izstrādājot **angļu-latviešu** mašīntulkošanas sistēmas ziņu jomai
- **2018. gadā**, izstrādājot **angļu-igauņu** mašīntulkošanas sistēmas ziņu jomai

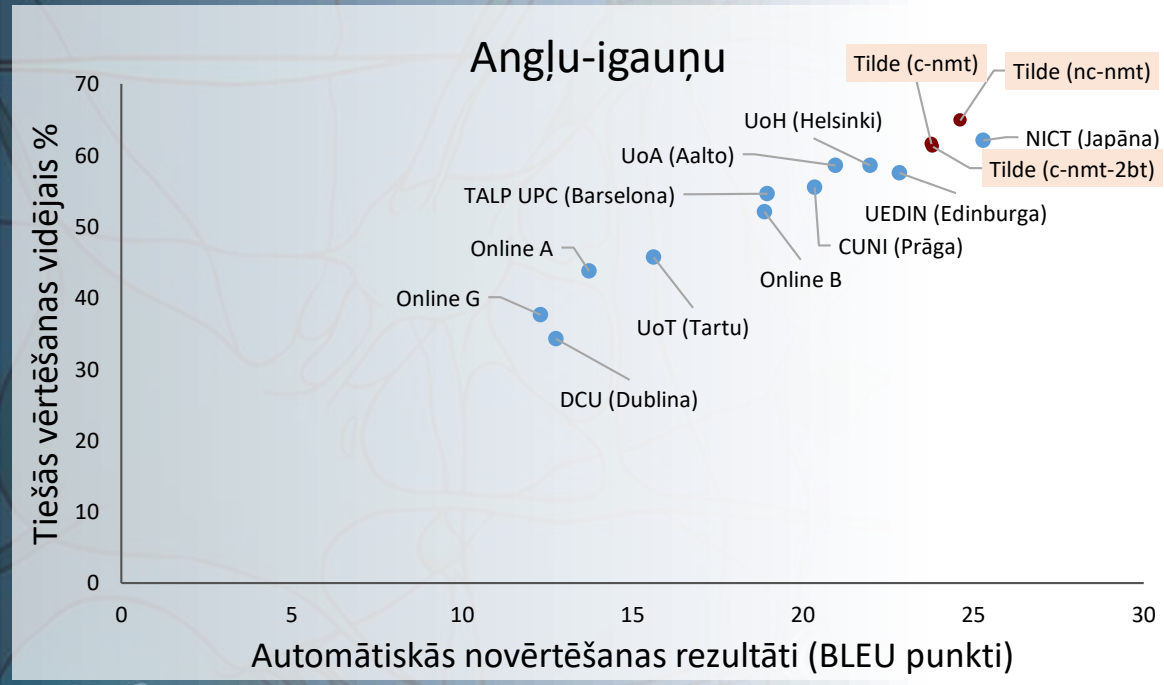
# Vislabākās angļu↔latviešu neironu mašīntulkošanas sistēmas WMT 2017 sacensībās

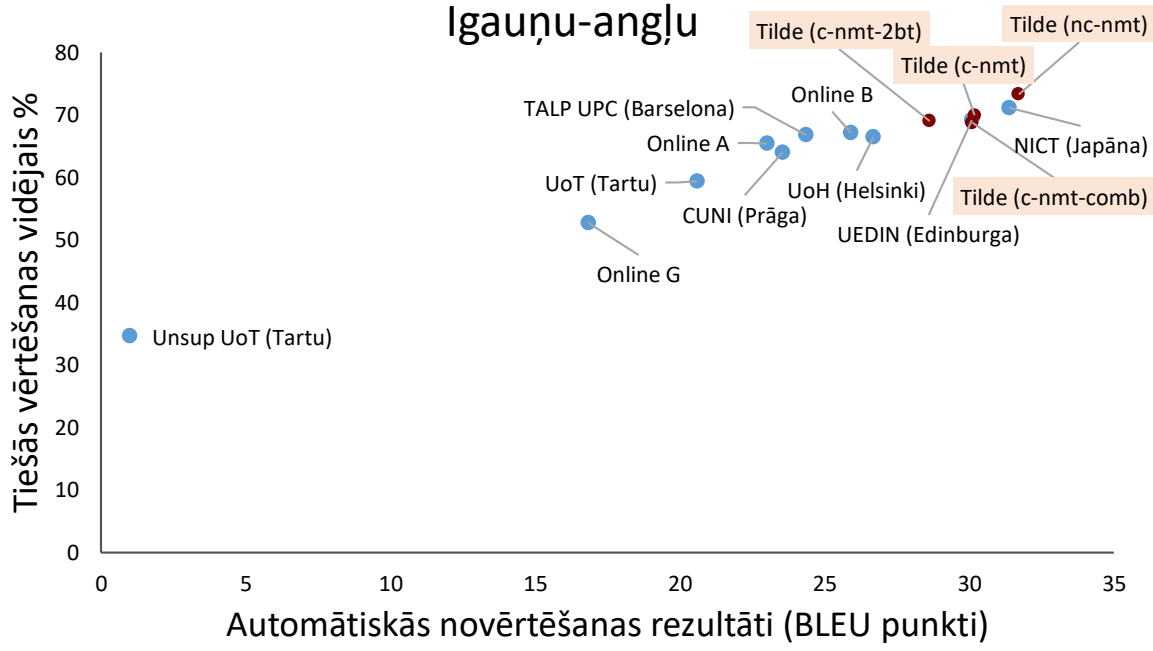



Mašīntulkošanas sistēmas apmācītas ar tā laika spējīgākajiem neironu tīklu modeļiem mašīntulkošanā - uzmanībās balsītiem rekurentiem neironu tīkliem ar multiplikatīvās garās īstermiņa atmiņas šūnām rekurentajos slāņos.

# Vislabākās angļu↔igauņu neironu mašīntulkošanas sistēmas WMT 2018 sacensībās

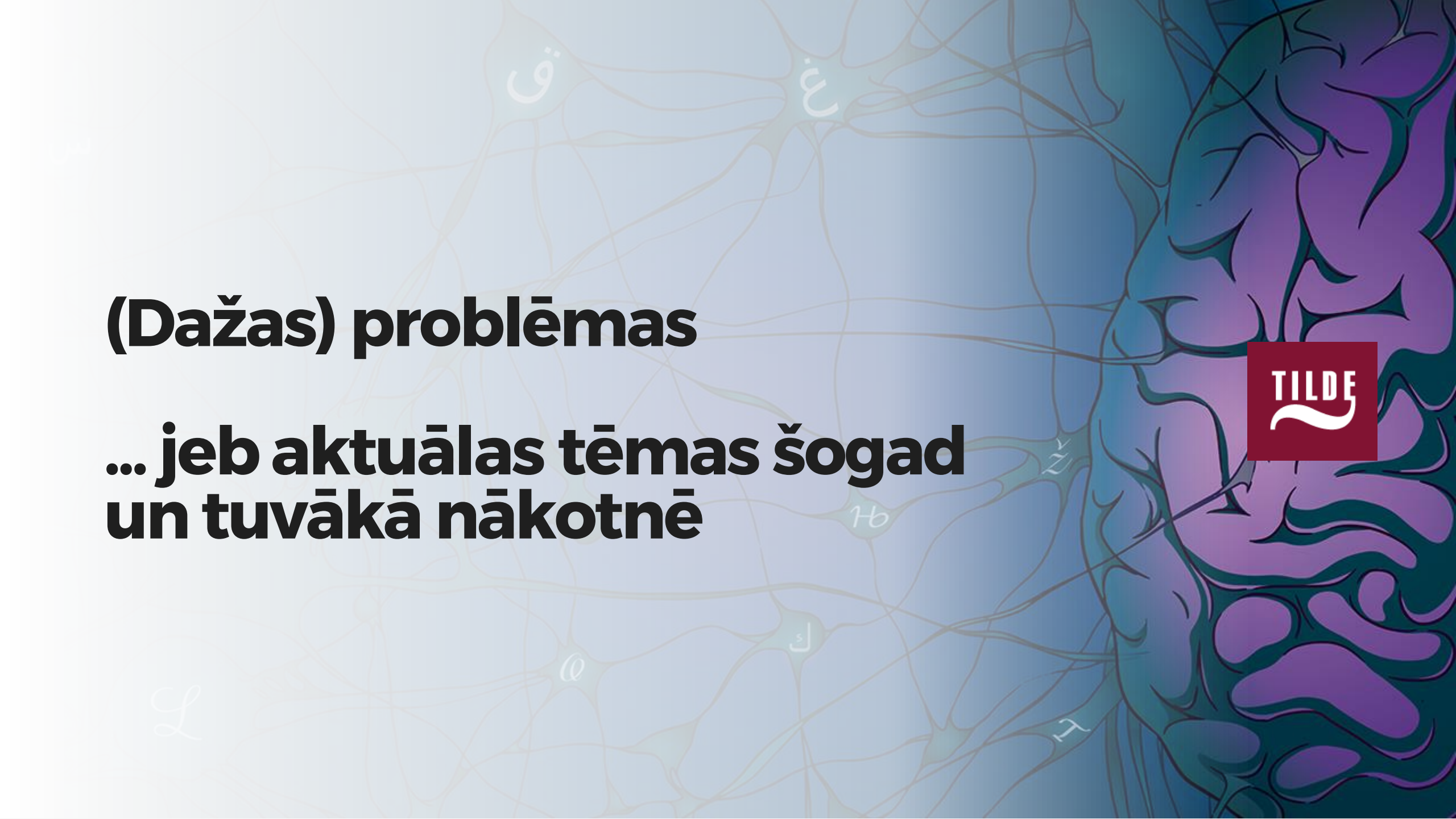



Mašīntulkošanas sistēmas apmācītas, izmantojot 2018. gada spējīgākos modeļus - Transformer modeļus.

# (Dažas) problēmas

# ... jeb aktuālas tēmas šogad un tuvākā nākotnē



# Retu vārdu/frāžu pareizāka tulkošana: nosauktās entitātes

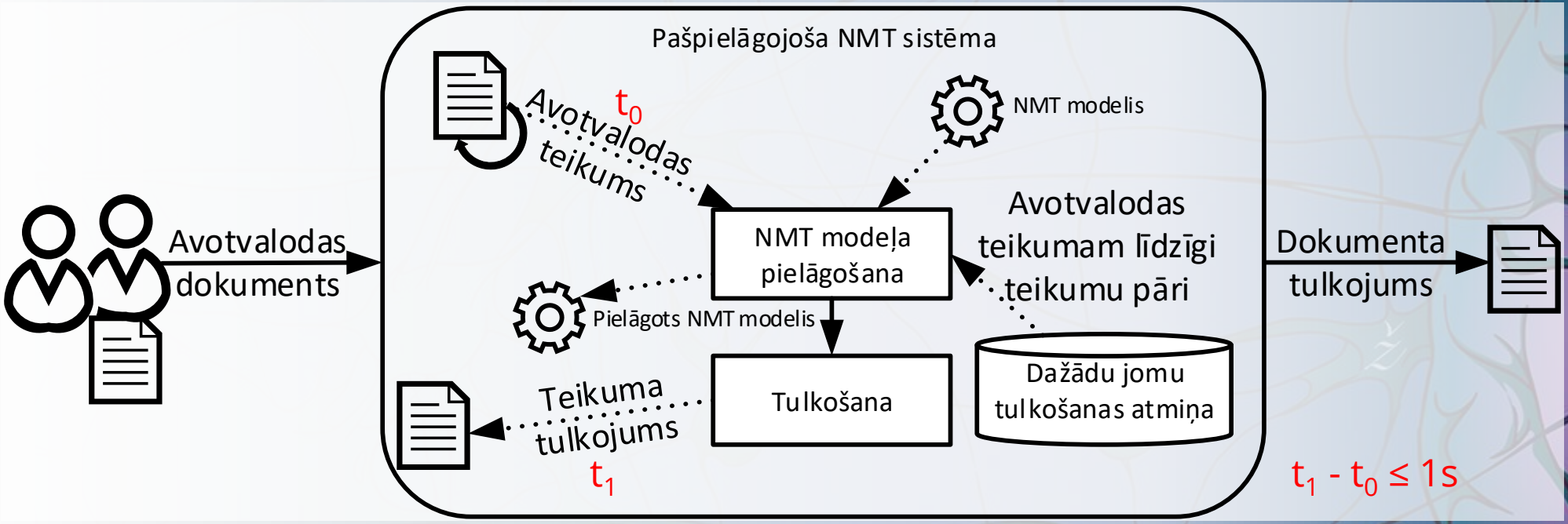
| Latviešu | Angļu |
|---|---|
| ***Google NMT (2019. gada 20. februāris)*** | |
| Šuplinska solīja, ka algas [...] | Šuplinska promised to pay [...] |
| [...] priekšsēdētāja Inga Vanaga. | [...] President Inga Vanaga. |
| Rīgas domes priekšsēdētājs Nils Ušakovs (S) [...] | Riga City Council Chairman Nils Ushakov (S) [...] |
| Ansis Ansbergs (LA) interesējās, | Ansis Ansberg (LA) was interested in |
| [...] izpildītājs Emīls Jakrins atbildēja [...] | [...] Emperor Yakrins answered [...] |
| ***Tilde NMT (WMT 2018)*** | |
| Šuplinska solīja, ka algas [...] | Shuplinska promised that wages [...] |
| [...] priekšsēdētāja Inga Vanaga. | [...] Chairman Inga Vanaga. |
| Rīgas domes priekšsēdētājs Nils Ušakovs (S) [...] | Riga City Council Chairman Nils Ushakov (S) [...] |
| Ansis Ansbergs (LA) interesējās, | Ansis Ansbergs (LA) was interested, |
| [...] izpildītājs Emīls Jakrins atbildēja [...] | [...] performer Emile Yakrin replied [...] |

# Retu vārdu/frāžu pareizāka tulkošana: terminoloģija

| Latviešu | Angļu |
|---|---|
| ***Google NMT (2019. gada 20. februāris)*** | |
| The corncrake, corn crake or landrail is a bird in the rail family. | Kukurūzas, kukurūzas krējums vai krastmala ir putns dzelzceļa ģimenē. |
| ***Tilde NMT (WMT 2018)*** | |
| The corncrake, corn crake or landrail is a bird in the rail family. | Kukurūzas krakšķis, kukurūzas krakšķis vai zemūdene ir putnu dzimtas putns. |

| | |
|---|---|
| corncrake, corn crake, landrail | grieze |
| bird | putns |
| rail family | dumbrvistiņu dzimta |

# Pielāgošana jomai - pašpielāgojošas NMT sistēmas




Sākotnējie rezultāti liecina, ka plašas jomas sistēmām, ieviešot paškpielāgojošas NMT sistēmas iespējas, tulkojumu kvalitāti var palielināt no +2 (ja līdzīgu datu tulkošanas atmiņā nav daudz) līdz pat +50 (ja līdzīgi jomas dati ir tulkošanas atmiņā sastopami) BLEU

# Kā gūt labumu no (daudz) datiem?

| Tulkošanas virziens | Datu kopapjoms (teikumu pāri milj.) | BLEU | Tiešā vērtējuma vidējais svērtais rezultāts |
|---|---|---|---|
| Angļu-igauņu | 3,65 | 23,54 | 61,6 |
| (WMT 2018) | 63,07 | 24,35 | 64,9 |
| Igauņu-angļu | 3,73 | 29,46 | 69,9 |
| (WMT 2018) | 67,91 | 30,94 | 73,3 |
| Angļu-latviešu | 6,19 | 20,18 | 51,1 |
| (WMT 2017) | 46,04 | 20,71 | 54,4 |
| Latviešu-angļu | 6,19 | 20,81 | 71,0 |
| (WMT 2017) | 45,71 | 23,02 | 76,2 |

Pašpielāgojošas NMT sistēmas?


Paldies!
TILDE