European Language Resource Coordination
Connecting Europe Facility

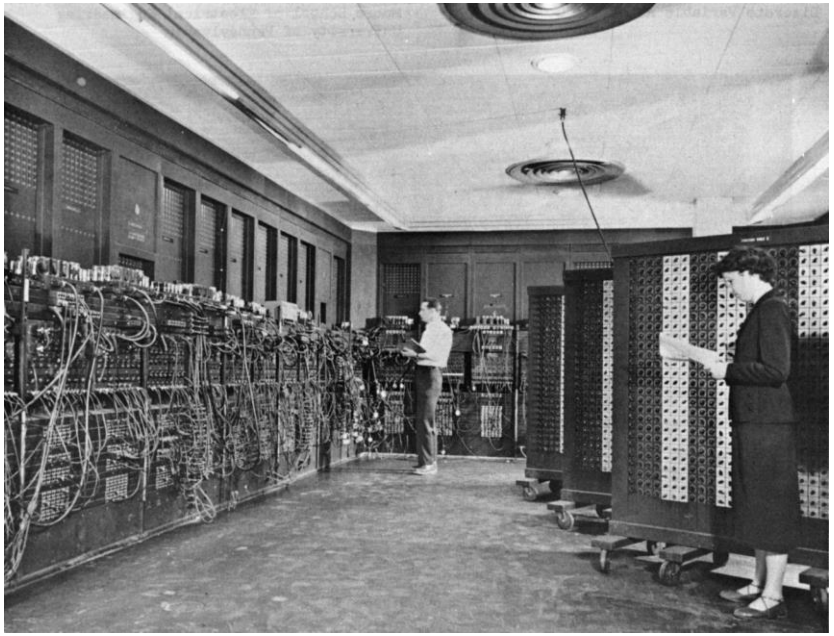
Automātiskā tulkošana: kā tā darbojas?

Inguna Skadiņa
Tilde/ELRC

Automatizētas tulkošanas risinājumi daudzvalodu Eiropai

- Daudzvalodu Eiropa:
  - 24 oficiālās valodas
  - 24+2 CEF valodas
- Daudz ....
- Cik tas maksā?

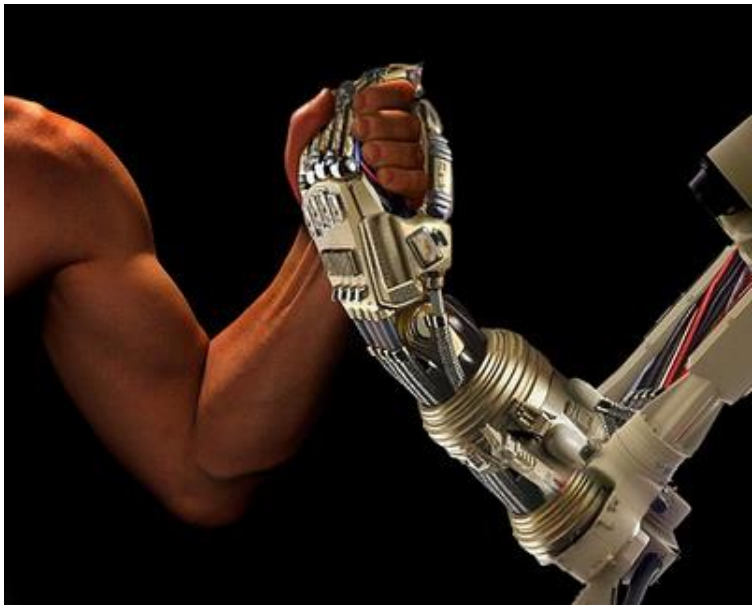
- Vai mašīntulkošana var palīdzēt?
- Un kā ar kvalitāti?



Attēls no https://en.wikipedia.org/wiki/ENIAC#/media/File:Eniac.jpg

Kāpēc tulkošana ir sarežģīts uzdevums?

- Dabiskās valodas ir vienkāršas un reizē sarežģītas:
  - gan vārdiem, gan teikumiem var būt vairākas nozīmes
  - ir daudz veidu, kā pateikt vienu un to pašu
  - nozīmi ietekmē konteksts
  - ....
- Vārdu secība teikumā
- Morfoloģija
- …



Attēls:: http://workingtropes.lmc.gatech.edu/wiki/index.php/File:Man-vs-machine.jpg
Licence: CC BY-NC-SA 3.0

European Language Resource Coordination
Connecting Europe Facility

Likumos balstīta MT

Valodneatkarīga reprezentācija (interlingva)

ANALĪZE (morfoloģskā, sintatiskā un semantiskā)

Transfērs

ĢENERĒŠANA (semantiskā, sintaktiskā un morfoloģiskā)

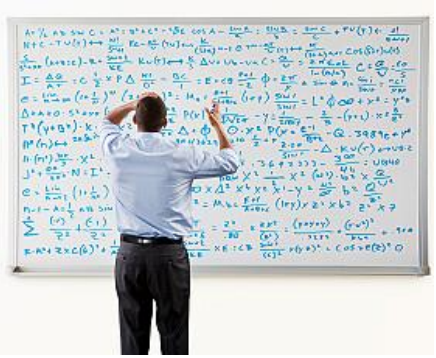
Tiešā tulkošana

Teksts avotvalodā

Teksts mērķvalodā

Vakvīza piramīda (Vauquois Pyramid)

Tulkošanas tehnoloģijas



- Tulkojumu nevar precīzi izskaitļot
- No likumos balstītas MT uz mašīnmācīšanos
  - Dators tulkojumus iemācās no datiem ⇒ dati ir ļoti svarīgi
  - Aptuvens risinājums ⇒ nav perfekts
    - profesionāli tulkotāji
    - pēcrediģēšana
    - automatizēta tulkošana ≠ automātiska

Statistiskās MT sistēmas

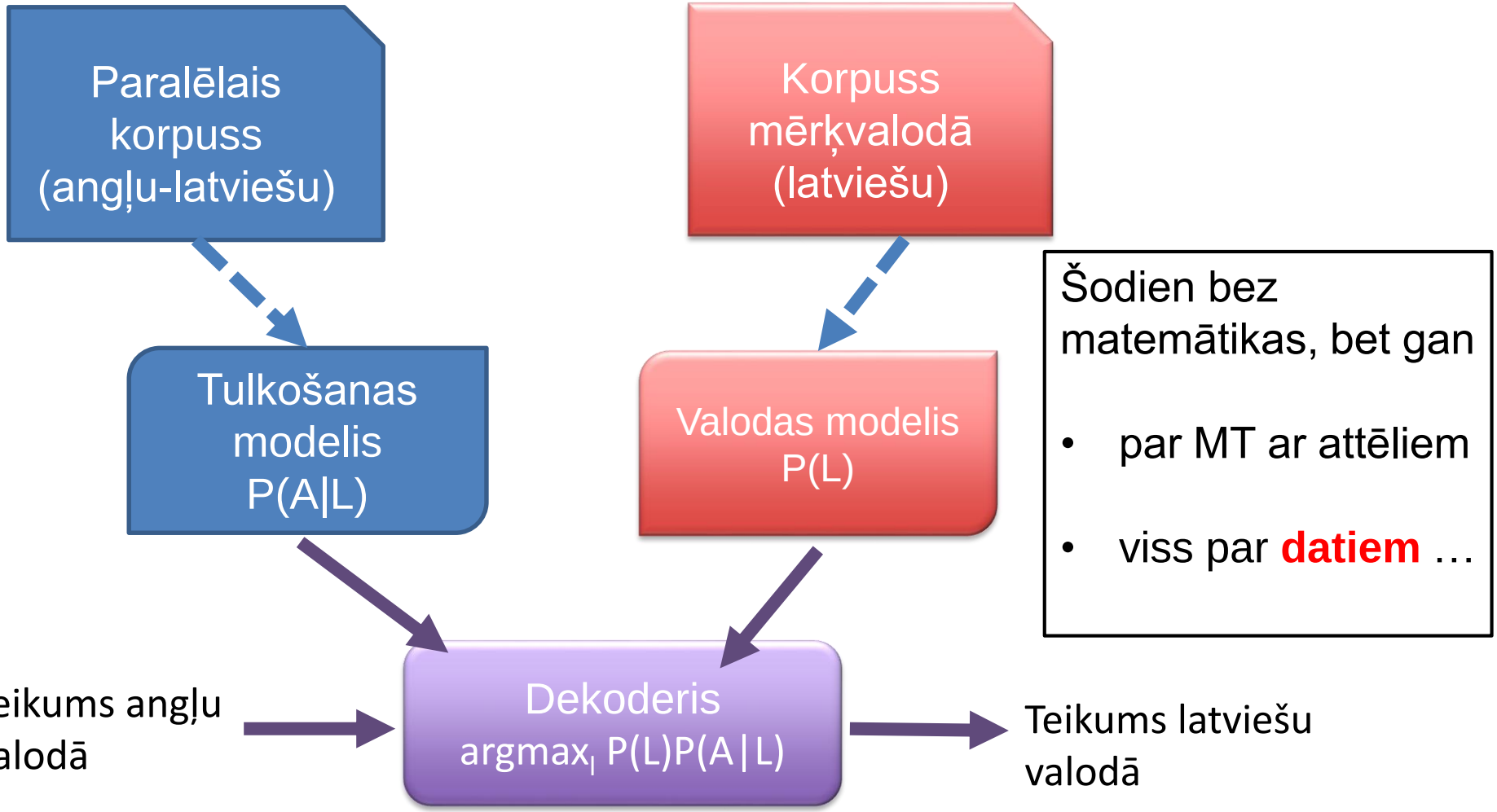
1947. gadā Vorens Vīvers (*Warren Weaver*) Norbertam Vīneram (*Norbert Wiener*) vēstulē izklāsta ideju par datora izmantošanu tulkošanā

… Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography... one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."...

Kad es redzu tekstu krieviski, es saku "Tas īstenībā ir rakstīts angliski, bet ir kodēts ar dīvainiem simboliem. Es to atkodēšu"

European Language Resource Coordination
Connecting Europe Facility

Kā darbojas mūsdienu MT?



Šodien bez matemātikas, bet gan

- par MT ar attēliem
- viss par **datiem** …

Kā darbojas mūsdienu MT?

Statistiskā MT iemācās tulkošanu no divu veidu datiem:

- Cilvēka veiktajiem tulkojumiem

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,	EIROPAS PARLAMENTS UN EIROPAS SAVIENĪBAS PADOME,
Having regard to the Treaty on the Functioning of the European Union, and in particular Article 114 thereof,	ņemot vērā Līgumu par Eiropas Savienības darbību un jo īpaši tā 114. pantu,
Having regard to the proposal from the European Commission,	ņemot vērā Eiropas Komisijas priekšlikumu,
After transmission of the draft legislative act to the national parliaments,	pēc leģislatīvā akta projekta nosūtīšanas valstu parlamentiem,
Having regard to the opinion of the European Central Bank (1),	ņemot vērā Eiropas Centrālās bankas atzinumu (1),
Having regard to the opinion of the European Economic and Social Committee (2),	ņemot vērā Eiropas Ekonomikas un sociālo lietu komitejas atzinumu
Acting in accordance with the ordinary legislative procedure (3),	saskaņā ar parasto likumdošanas procedūru (3),

- Tekstiem mērķvalodā
- Jo vairāk datu, jo labāk
- Svarīgi, lai tie būtu piemēroti dati!

Ko var iemācīties no datiem?

Kurš teikums ir kura tulkojums?		
Sastatījums teikuma līmenī	Take the cone on the square. Take the block. Take the green block.	Paņem konusu no kvadrāta. Paņem klucīti. Paņem zaļo klucīti.

Kā tulkojami vārdi un vārdu savienojumi?		
Vārdu sastatīšana Tulkojumu varbūtības	block green take	klucītis zaļš paņemt

Kāds ir pareizs teikums mērķvalodā?	
Valodas modelis	Paņem zaļo klucīti. ~~Paņem klucīti zaļo.~~

European Language Resource Coordination
Connecting Europe Facility

Sastatīšana teikuma līmenī

Take the cube.	Paņem kubu.
Take the green cube on the red circle.	Paņem zaļo kubu no sarkanā apļa.
Take the green cube, put the cube on the red circle.	Paņem zaļo kubu. Noliec kubu uz sarkanā apļa.
Put the cone on the red circle on the square.	Pārliec konusu no sarkanā apļa uz kvadrāta.

Sastatīšana vārdu līmenī

CLASSIC SOUPS

		Sm.	Lg.
57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
58.	Chicken Rice Soup	1.85	3.25
59.	Chicken Noodle Soup	1.85	3.25
60.	Cantonese Wonton Soup	1.50	2.75
61.	Tomato Clear Egg Drop Soup	1.65	2.95
62.	Regular Wonton Soup	1.10	2.10
63.	Hot & Sour Soup	1.10	2.10
64.	Egg Drop Soup	1.10	2.10
65.	Egg Drop Wonton Mix	1.10	2.10
66.	Tofu Vegetable Soup	NA	3.50
67.	Chicken Corn Cream Soup	NA	3.50
68.	Crab Meat Corn Cream Soup	NA	3.50
69.	Seafood Soup	NA	3.50

Sastatīšana vārdu līmenī

			CLASSIC SOUPS	Sm.	Lg.
清	燉	雞 湯	57. House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞	飯	湯	58. Chicken Rice Soup	1.85	3.25
雞	麵	湯	59. Chicken Noodle Soup	1.85	3.25
廣	東	雲吞 湯	60. Cantonese Wonton Soup	1.50	2.75
蕃	茄	雪 湯	61. Tomato Clear Egg Drop Soup	1.65	2.95
雲	吞	湯	62. Regular Wonton Soup	1.10	2.10
酸	辣	湯	63. Hot & Sour Soup	1.10	2.10
雪	花	湯	64. Egg Drop Soup	1.10	2.10
雲	雪	湯	65. Egg Drop Wonton Mix	1.10	2.10
豆	腐	菜 湯	66. Tofu Vegetable Soup	NA	3.50
雞	玉	米 湯	67. Chicken Corn Cream Soup	NA	3.50
蟹	肉	玉米 湯	68. Crab Meat Corn Cream Soup	NA	3.50
海	鮮	湯	69. Seafood Soup	NA	3.50

Vārdu tulkojumu mācīšanās

- Vārdu sastatītājs daudz zina par ķīniešu zupām
- Bet nezin daudz ko citu …

- Zināšanas ir tikai par to, kas ir datos
- Līdzīgi cilvēkiem …
- Kopīga tēma …

- Vai no vārda līmenī sastatītiem tekstiem var iemācīties tulkojumus?
- Jā, pavisam vienkārši…

Statistiskā mašīntulkošana

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.

European Language Resource Coordination
Connecting Europe Facility

Statistiskā mašīntulkošana

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.

Biežumu statistika		
take	paņem	3
block	klucīti	2
	klucīša	1
green	zaļo	1
red	sarkano	1
	sarkanā	2
on	uz	2
put	noliec	2
	.....	

Statistiskā mašīntulkošana

Put the red block

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.

Biežumu statistika		
take	paņem	3
block	klucīti	2
	klucīša	1
green	zaļo	1
red	sarkano	1
	sarkanā	2
on	uz	2
put	noliec	2

European Language Resource Coordination
Connecting Europe Facility

Tulkošanas modelis

Put the red block

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.

Biežumu statistika		
take	paņem	3
block	klucīti	2
	klucīša	1
green	zaļo	1
red	sarkano	1
	sarkanā	2
on	uz	2
put	noliec	2

Noliec sarkanā klucīti

Statistiskā mašīntulkošana

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.

Put the red block

Noliec sarkano klucīti
2/2 1/3 2/3

Noliec sarkano klucīša
2/2 1/3 1/3

Noliec sarkanā klucīti
2/2 2/3 2/3

Noliec sarkanā klucīša
2/2 2/3 1/3

Statistiskā mašīntulkošana

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.

Put the red block

Noliec sarkano klucīti
2/2 1/3 2/3

Noliec sarkano klucīša
2/2 1/3 1/3

Noliec sarkanā klucīti
2/2 2/3 2/3

Noliec sarkanā klucīša
2/2 2/3 1/3

Kuru izvēlēties?

Valodas modelis

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.

Valodas modelis:

- Kas ir laba valoda?
- Kuri vārdi var kuriem sekot un kuri nevar?
- Kādi ir pareizie locījumi?
- To iemācās no datiem

- sarkano klucīti – labs teikums
- sarkano klucīša – nav labs teikums
- sarkanā klucīti– nav labs teikums
- sarkanā klucīša - labs teikums
- Put the red block -> noliec sarkano klucīti

Statistiskā mašīntulkošana

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.

Biežumu statistika		
take	paņem	3
block	klucīti	2
	klucīša	1
green	zaļo	1
red	sarkano	1
	sarkanā	2
	...	

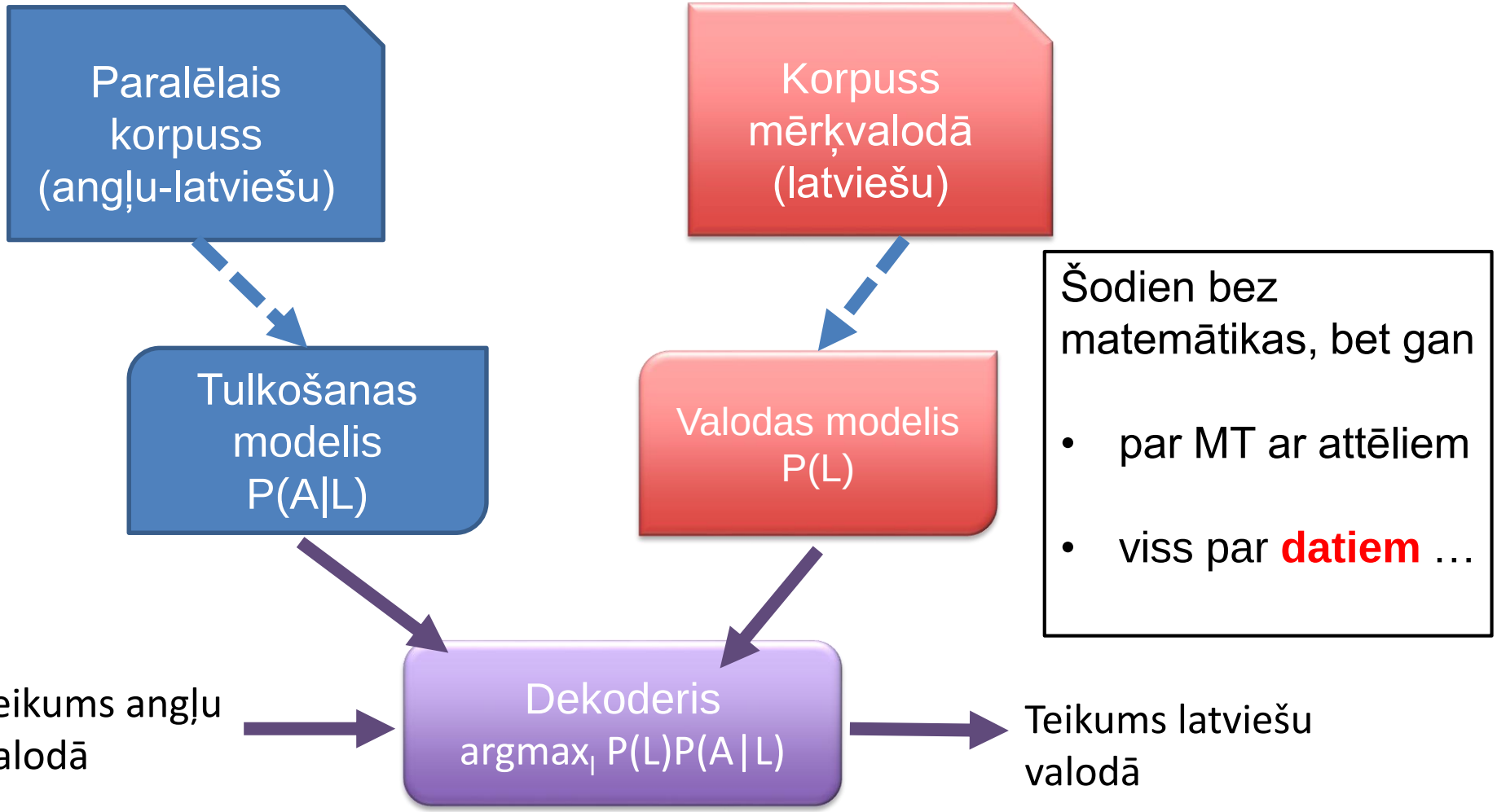
+

Valodas modelis

Tulkojums mērķvalodā

European Language Resource Coordination
Connecting Europe Facility

Kā darbojas mūsdienu MT?



Šodien bez matemātikas, bet gan

- par MT ar attēliem
- viss par **datiem** …

Frāzēs balstīta SMT

- Līdz šim: vārdu tulkošana
- Zaudēts konteksts: piemēram, saskaņojums (*paņem sarkanā klucīti* …) u.c.
- Daļēji "labo" valodas modelis

- Labāks modelis - ne tikai vārdu tulkojumi, bet arī frāžu tulkojumi

put the red square : noliec sarkano kvadrātu
on the square: uz kvadrāta

Statistiskā mašīntulkošana

Put the red square
on the green block

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the green square.	Paņem zaļo kvadrātu.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.
Put on the green block.	**Noliec** uz zaļā klucīša.

Biežumu statistika		
take	paņem	4
block	klucīti	2
	klucīša	1
green	zaļo	2
	zaļā	1
red	sarkano	1
	sarkanā	2
on	uz	3
put	noliec	3
square	kvadrāta	1
	kvadrātu	2

Noliec sarkanā kvadrātu uz zaļo klucīti

European Language Resource Coordination
Connecting Europe Facility

Statistiskā mašīntulkošana

Sastatījums	
Take the block.	Paņem klucīti.
Take the green block.	Paņem zaļo klucīti.
Take the green square.	Paņem zaļo kvadrātu.
Take the red square.	Paņem sarkano kvadrātu.
Put the square on the red block.	**Noliec** kvadrātu uz sarkanā klucīša.
Put on the red square.	**Noliec** uz sarkanā kvadrāta.
Put on the green block.	**Noliec** uz zaļā klucīša.

Biežumu statistika		
the red square	sarkano kvadrātu	1
on the green block	uz zaļā klucīša	1
on the red squire	uz sarkanā kvadrāta	1
put	noliec	3
	....	

Statistiskā mašīntulkošana

Put the red square on the green block

Biežumu statistika		
the red square	sarkano kvadrātu	1
on the green block	uz zaļā klucīša	1
on the red squire	uz sarkanā kvadrāta	1
put	noliec	3

Noliec sarkano kvadrātu uz zaļā klucīša

Frāzēs balstīta SMT

- Daudz labāka kā vārdos balstīta SMT!
- Standarta tehnoloģija izmanto *Google*, *Microsoft*, Tilde *Localisation & Translation Industry*

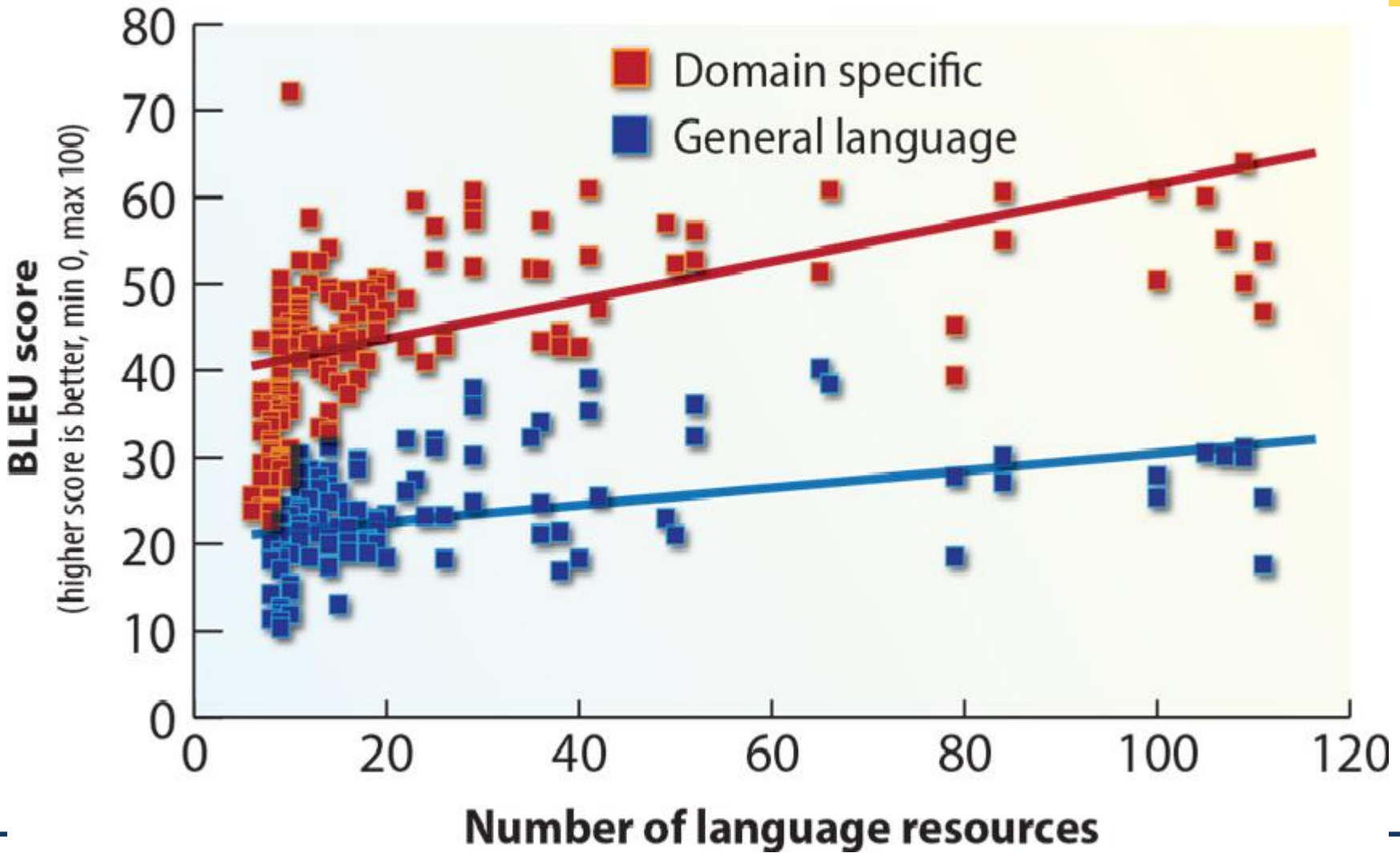
- *Moses* Open Source PB-SMT
- Visplašāk lietotā SMT sistēma
- EK atbalstīti pētījumi
- Izmanto EC DGT MT@EC un *hugo.lv*

Mašīntulkošana un dati

- SMT ir par datiem
- SMT no datiem iemācās tulkot
- Dati
  - tulkojumi (divvalodu dati)
  - vienvalodas dati (teksti mērķvalodā)
  - Vārdnīcas, terminoloģija, ontoloģijas, nosauktās entitātes
- Līdzīgi cilvēkiem, SMT ir laba tajā, ko iemācījusies

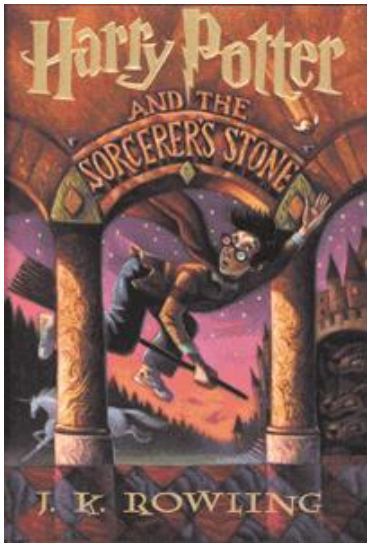
European Language Resource Coordination
Connecting Europe Facility

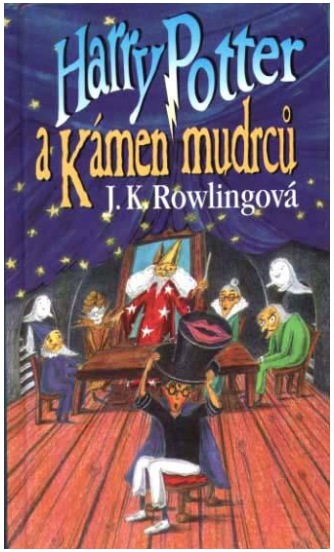
Valodas resursu ietekme uz tulkojuma kvalitāti

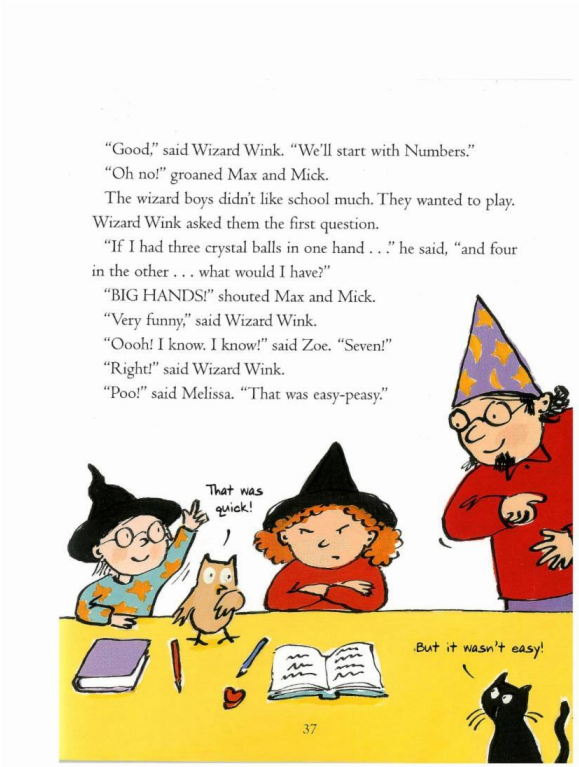


European Language Resource Coordination
Connecting Europe Facility

Mašīntulkošana un dati



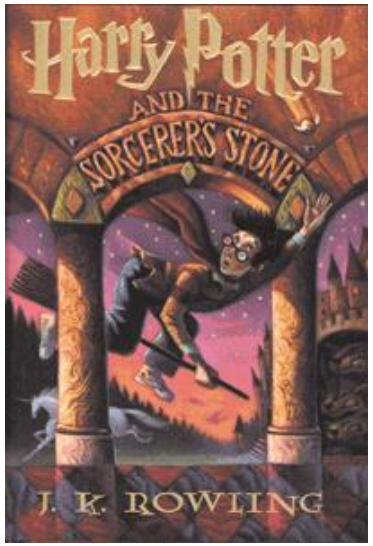


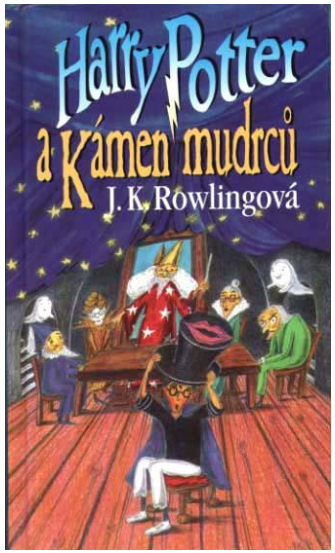

"Good," said Wizard Wink. "We'll start with Numbers."
"Oh no!" groaned Max and Mick.
The wizard boys didn't like school much. They wanted to play. Wizard Wink asked them the first question.
"If I had three crystal balls in one hand . . ." he said, "and four in the other . . . what would I have?"
"BIG HANDS!" shouted Max and Mick.
"Very funny," said Wizard Wink.
"Oooh! I know. I know!" said Zoe. "Seven!"
"Right!" said Wizard Wink.
"Poo!" said Melissa. "That was easy-peasy."

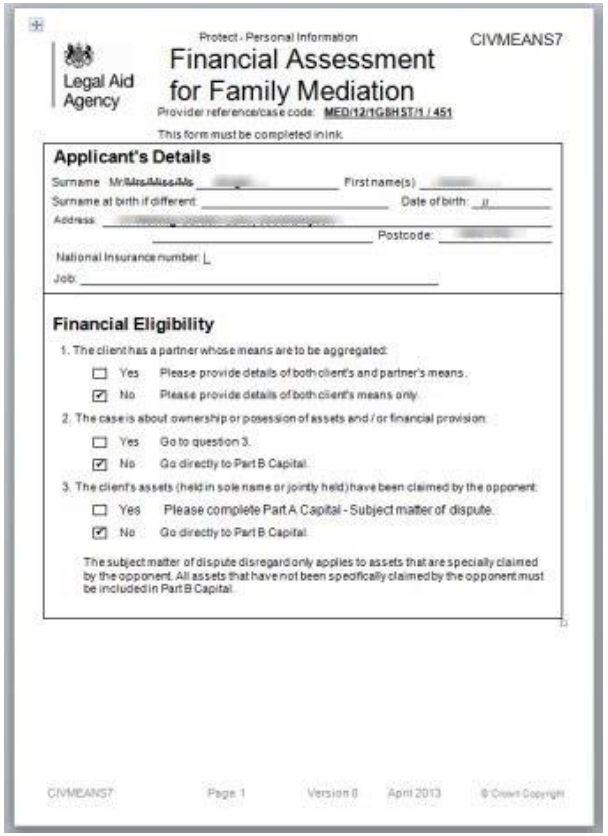
MOSES CORE

European Language Resource Coordination
Connecting Europe Facility

Mašīntulkošana un dati






Protect - Personal Information CIVMEANS7

Legal Aid Agency

Financial Assessment for Family Mediation

Provider reference/case code: MED/12/1GBHST/1 / 451

This form must be completed in ink

Applicant's Details

Surname: Mr/~~Mrs/Miss/Ms~~ First name(s):

Surname at birth if different: Date of birth:

Address: Postcode:

National Insurance number:

Job:

Financial Eligibility

1. The client has a partner whose means are to be aggregated:
   - ☐ Yes Please provide details of both client's and partner's means.
   - ☑ No Please provide details of both client's means only.
2. The case is about ownership or possession of assets and / or financial provision:
   - ☐ Yes Go to question 3.
   - ☑ No Go directly to Part B Capital.
3. The client's assets (held in sole name or jointly held) have been claimed by the opponent:
   - ☐ Yes Please complete Part A Capital - Subject matter of dispute.
   - ☑ No Go directly to Part B Capital.

The subject matter of dispute disregard only applies to assets that are specially claimed by the opponent. All assets that have not been specifically claimed by the opponent must be included in Part B Capital.

CIVMEANS7 Page 1 Version 8 April 2013 © Crown Copyright

MOSES CORE

Valodas resursi automatizētai tulkošanai

- Statistiskā mašīntulkošana «iemācās» tulkojumus no valodas resursiem (datiem)
- Eiropas automatizētās tulkošanas infrastruktūrai nepieciešami «pareizie» dati
- Šādi dati ir gan valsts sektora, gan publiskā sektora, gan nevalstisko organizāciju, gan privāto organizāciju rīcībā

Eiropas valodu resursu koordinācijas aktivitāte (ELRC)

- Nepieciešama palīdzība Eiropas automatizētās tulkošanas infrastruktūras sekmīgai izveidei:
  - pakalpojumi Eiropas iedzīvotājiem
  - pakalpojumi mums visiem
  - atbalsts daudzvalodībai
- Nepieciešama palīdzība pareizo datu atrašanā
- Atbalstot savu valodu, mēs atbalstām daudzvalodu Eiropu un otrādi