European Language Resource Coordination
Connecting Europe Facility

Kako zaista radi automatizirano prevođenje?

Marko Tadić
Sveučilište u Zagrebu, Filozofski fakultet

Zahvala za prilagođenu građu koja se pojavljivala na MT Maratonima, radionicama MT-a itd.

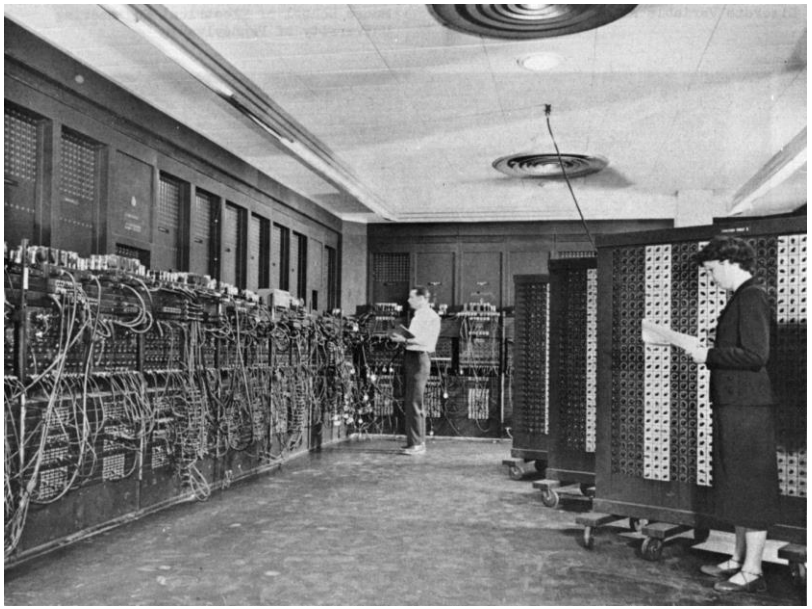
Strojno prevođenje (*Machine Translation*, MT)

Plan izlaganja

- Zašto strojno prevođenje: obujam, kakvoća, cijena?
- Zašto je MT teško?
- MT + ljudski prevoditelji = visoka kakvoća prijevoda
- Kako radi suvremeni statistički MT?
- Sve je u podatcima!
- I to u pravoj vrsti podataka!

Strojno prevođenje, kakvoća i cijena?

- Europa je višejezična!
- 24 službena jezika,
  24+2 jezika u CEF-u
- Toliko toga za prevesti!
- Cijena prevođenja?
- Može li MT pomoći?
- A što je s kakvoćom prijevoda?



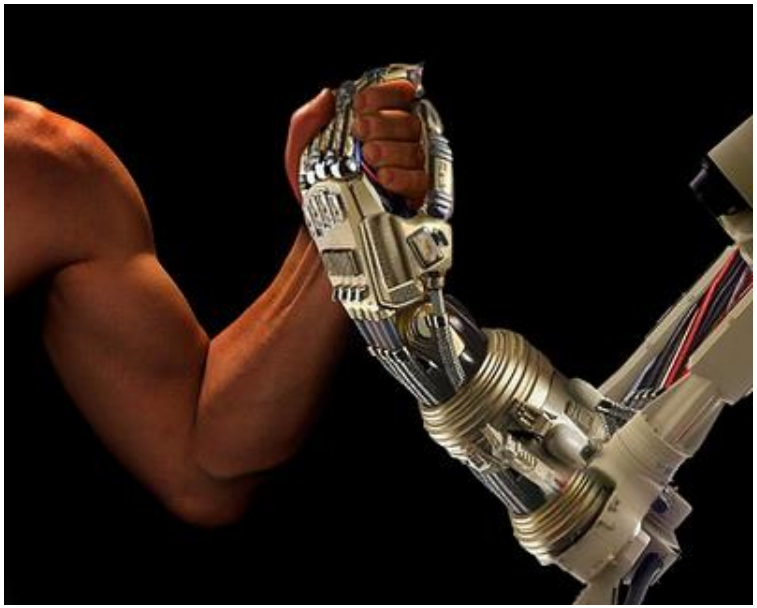
Slika: https://en.wikipedia.org/wiki/ENIAC#/media/File:Eniac.jpg
Licencija: javna domena

Zašto je strojno prevođenje teško?

- prirodni su jezici:
  - elegantni
  - učinkoviti
  - prilagodljivi
  - složeni

- jedna riječ/rečenica može imati više značenja
- isto se značenje može prenijeti na razne načine
- značenje riječi ovisi o okolini
- doslovna i prenesena značenja (metafore)
- jezik je blisko vezan s kulturom (različiti načini poimanja iste pojave)

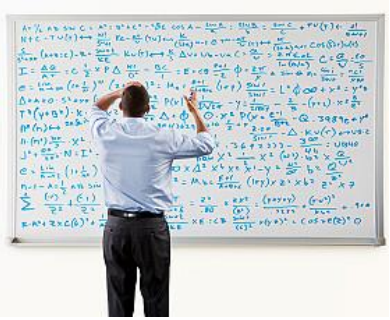
- strukturne osobine prirodnih jezika
  - red riječi u rečenici
  - morfologija
  - …



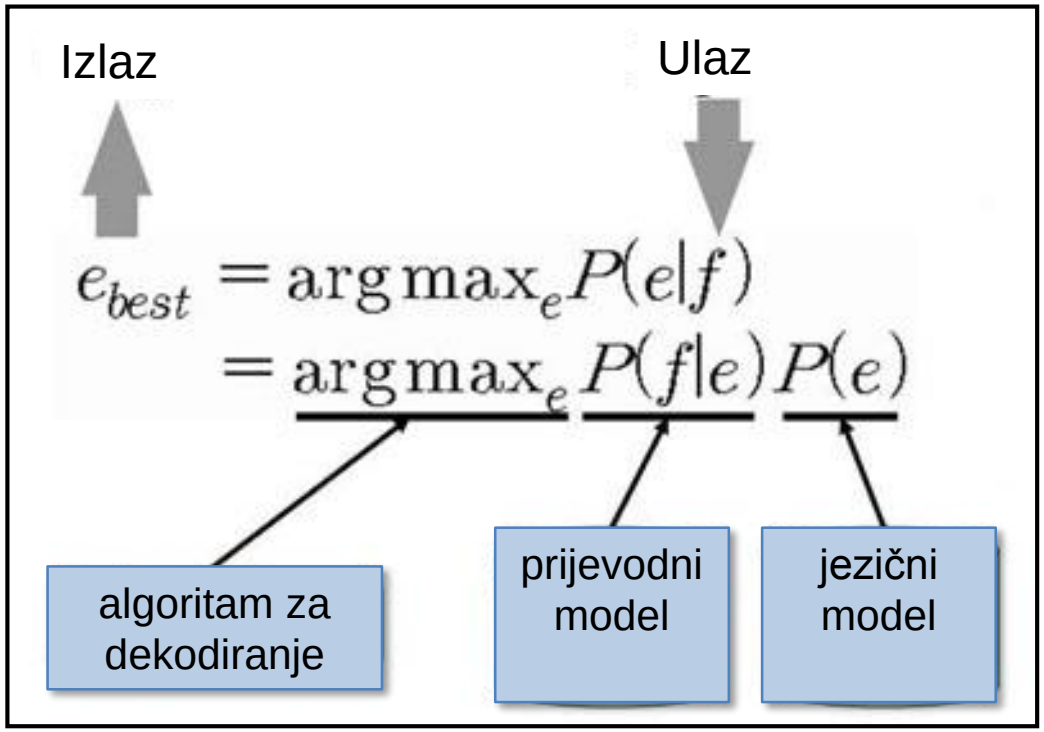
Slika: http://workingtropes.lmc.gatech.edu/wiki/index.php/File:Man-vs-machine.jpg

Prevođenje je složeno

- prevođenje s jezika na jezik je složeno
- ne možemo ga egzaktno izračunati
- pokušali smo: MT utemeljeno na pravilima + jezične tehnologije …
- Što sad pokušavamo?
- strojno učenje (*Machine Learning*, ML)
  - stroj uči iz podataka ⇒ podatci su važni
  - približna rješenja ⇒ nisu savršena, potrebna je pomoć
    - profesionalnih prevoditelja
    - naknadnoga uređivanja (*post-editing*)
    - automatiziranoga prevođenja (≠ automatsko prevođenje)



Kako radi suvremeno strojno prevođenje?



- Ne bojte se!
  Niti riječi danas o matematici.
- umjesto toga:
- priča o statističkome strojnom prevođenju u slikama …
- sve je u podatcima …

European Language Resource Coordination
Connecting Europe Facility

Kako radi suvremeno strojno prevođenje?

Statističko strojno prevođenje uči iz dvije vrste podataka:

- ljudski prijevodi
- tekstovi na ciljnome jeziku

- Što više podataka, to bolje!
- Ali: i prave vrste podataka!

GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
I. Von dem Unterschiede der reinen und empirischen Erkenntnis	*I. Of the difference between Pure and Empirical Knowledge*	*I. De la différence de la connaissance pure et de la connaissance empirique.*
Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.	Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.

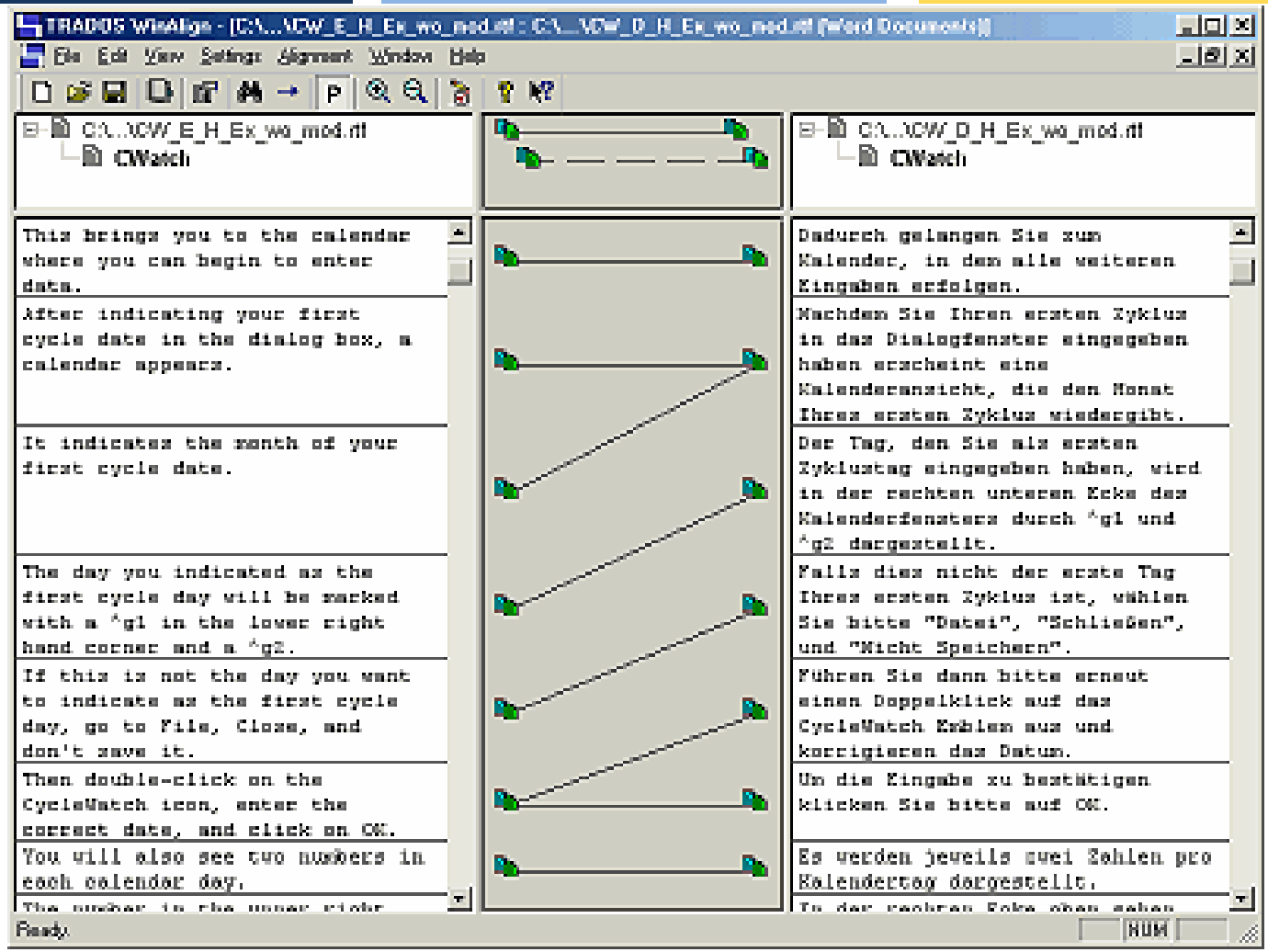
Što možemo/moramo naučiti iz podataka?

- koja je rečenica prevedena kojom rečenicom: sravnjivanje rečenica
- koja je riječ prevedena kojom riječju: sravnjivanje riječi + prijevodna vjerojatnost
- kako izgleda tekst na dobrome ciljnom jeziku: jezični model

GERMAN	ENGLISH	FRENCH
Einleitung	Introduction	Introduction
I. Von dem Unterschiede der reinen und empirischen Erkenntnis	*I. Of the difference between Pure and Empirical Knowledge*	*I. De la différence de la connaissance pure et de la connaissance empirique.*
Daß alle unsere Erkenntnis mit der Erfahrung anfange, daran ist gar kein Zweifel; denn wodurch sollte das Erkenntnisvermögen sonst zur Ausübung erweckt werden, geschähe es nicht durch Gegenstände, die unsere Sinne rühren und teils von selbst Vorstellungen bewirken, teils unsere Verstandestätigkeit in Bewegung bringen, diese zu vergleichen, sie zu verknüpfen oder zu trennen, und so den rohen Stoff sinnlicher Eindrücke zu einer Erkenntnis der Gegenstände zu verarbeiten, die Erfahrung heißt? Der Zeit nach geht also keine Erkenntnis in uns vor der Erfahrung vorher, und mit dieser fängt alle an.	That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare to connect, or to separate these, and so to convert the raw material of our sensuous impressions into a knowledge of objects, which is called experience? In respect of time, therefore, no knowledge of ours is antecedent to experience, but begins with it.	Que toute notre connaissance commence avec l'expérience, cela ne soulève aucun doute. En effet, par quoi notre pouvoir de connaître pourrait-il être éveillé et mis en action, si ce n'est par des objets qui frappent nos sens et qui, d'une part, produisent par eux-mêmes des représentations et, d'autre part, mettent en mouvement notre faculté intellectuelle, afin qu'elle compare, lie ou sépare ces représentations, et travaille ainsi la matière brute des impressions sensibles pour en tirer une connaissance des objets, celle qu'on nomme l'expérience? Ainsi, chronologiquement, aucune connaissance ne précède en nous l'expérience et c'est avec elle que toutes commencent.

European Language Resource Coordination
Connecting Europe Facility

Sravnjivanje rečenica



Sravnjivanje riječi

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

Sravnjivanje riječi

				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	60. Cantonese Wonton Soup	1.50	2.75
蕃	茄	蛋	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
蛋		花	湯	64. Egg Drop Soup	1.10	2.10
雲		蛋	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

Kako naučiti prijevode riječi?

- sustav za sravnjivanje riječi zna mnogo o nazivima kineskih juha
- ali ne zna ništa drugo

- zna samo ono što je već vidio u podatcima za učenje
- kao i ljudi

- Ukoliko imamo skup sravnjenih riječi dvaju jezika, jesmo li time dobili prijevodni rječnik?
- Jesmo. I to izgleda stvarno jednostavno …

European Language Resource Coordination
Connecting Europe Facility

Statističko strojno prevođenje

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

Aligned Data

Statističko strojno prevođenje

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

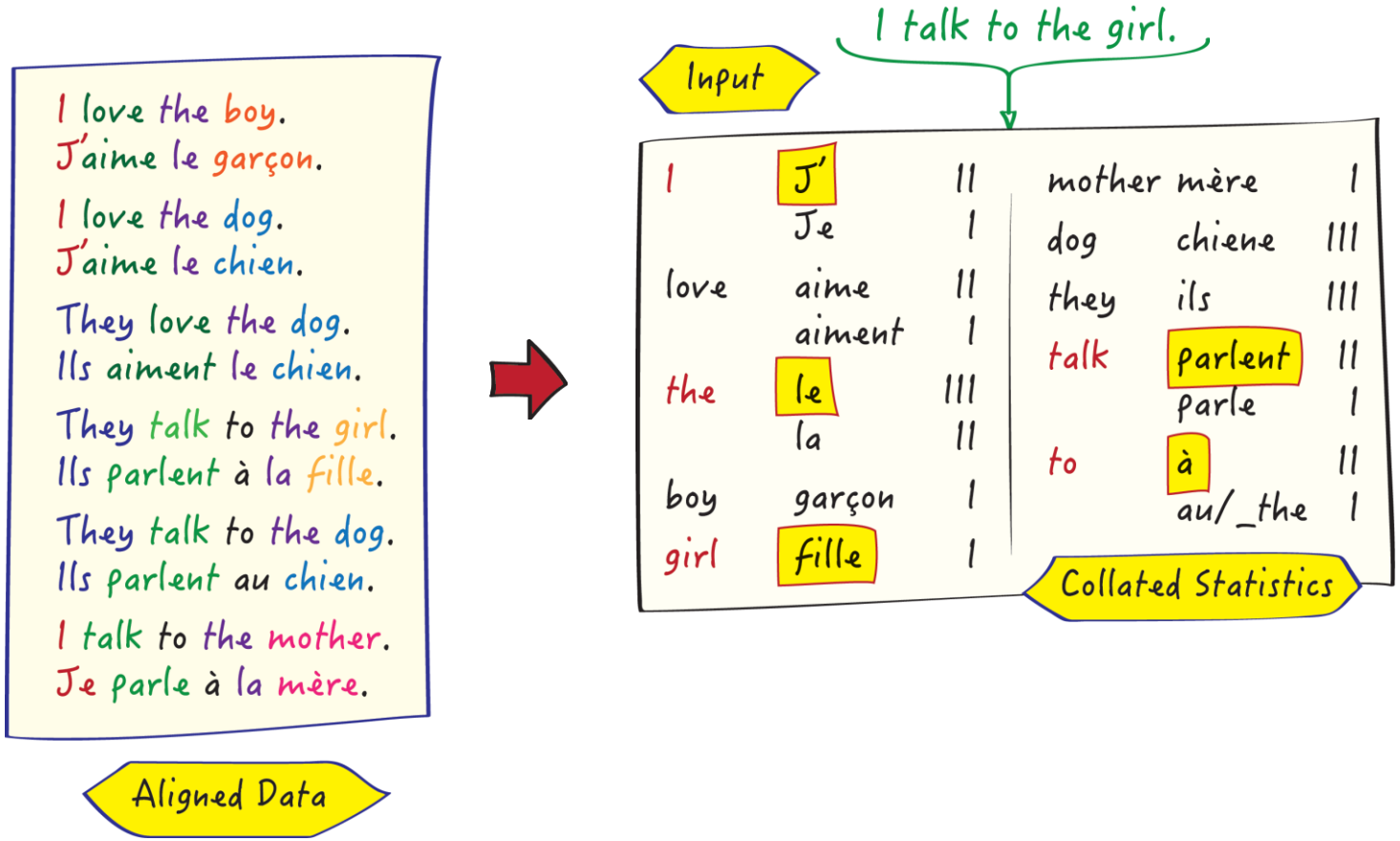
I talk to the mother.
Je parle à la mère.

Aligned Data

I	J'	II	mother	mère	I
	Je	I	dog	chiene	III
love	aime	II	they	ils	III
	aiment	I	talk	parlent	II
the	le	III		parle	I
	la	II	to	à	II
boy	garçon	I		au/_the	I
girl	fille	I			

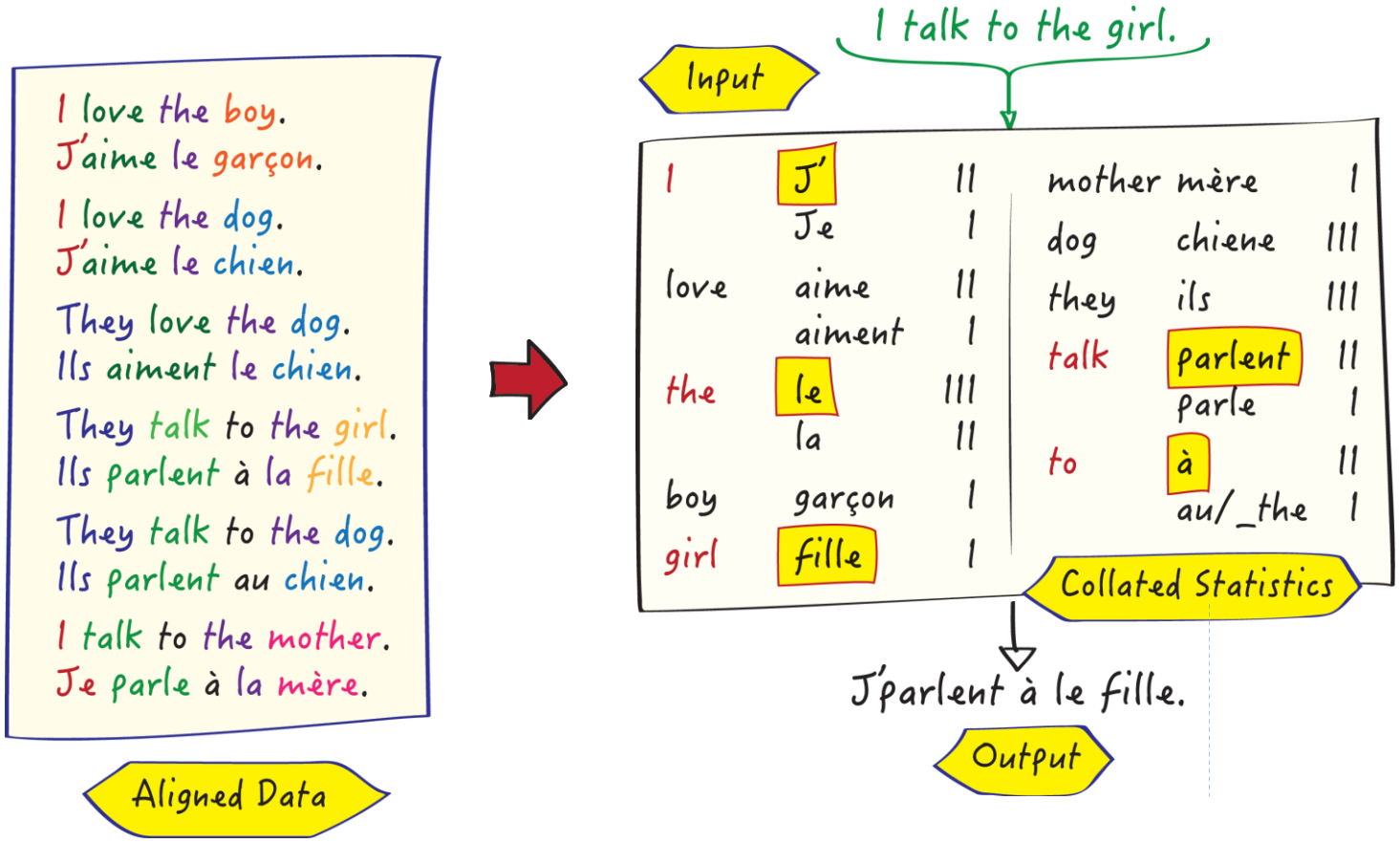
Collated Statistics

Statističko strojno prevođenje



European Language Resource Coordination
Connecting Europe Facility

Statističko strojno prevođenje



Statističko strojno prevođenje

I love the boy.
J'aime le garçon.
I love the dog.
J'aime le chien.
They love the dog.
Ils aiment le chien.
They talk to the girl.
Ils parlent à la fille.
They talk to the dog.
Ils parlent au chien.
I talk to the mother.
Je parle à la mère.

Aligned Data

I	talk	to	the	girl
J'	parlent		au	le	fille
2/3	2/3		2/3	3/5	1/1
Je	parle	à	la	fille
1/3	1/3	1/3	2/5	1/1

Kako odabrati prijevod?

Statističko strojno prevođenje

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

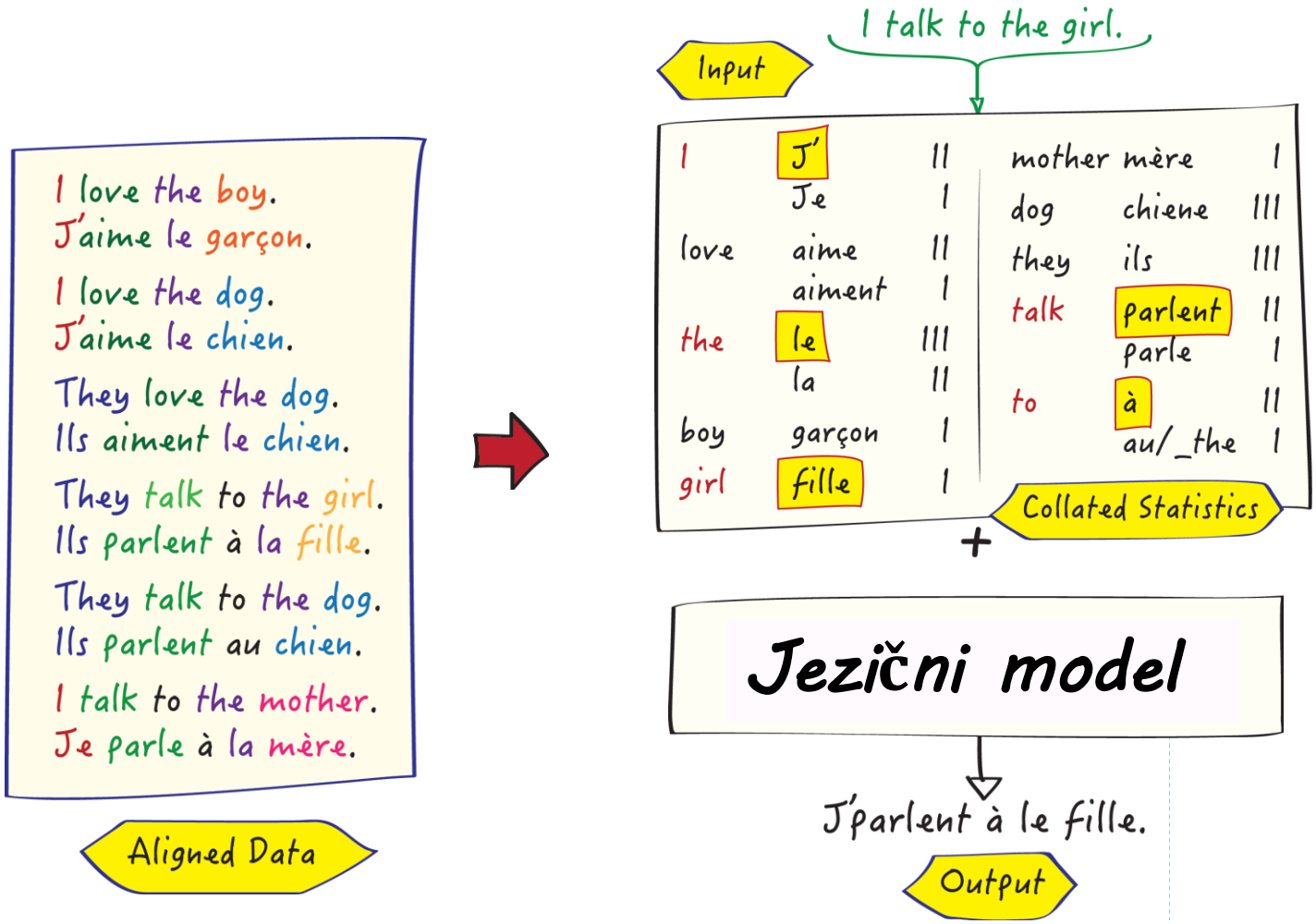
Aligned Data

Jezični model:

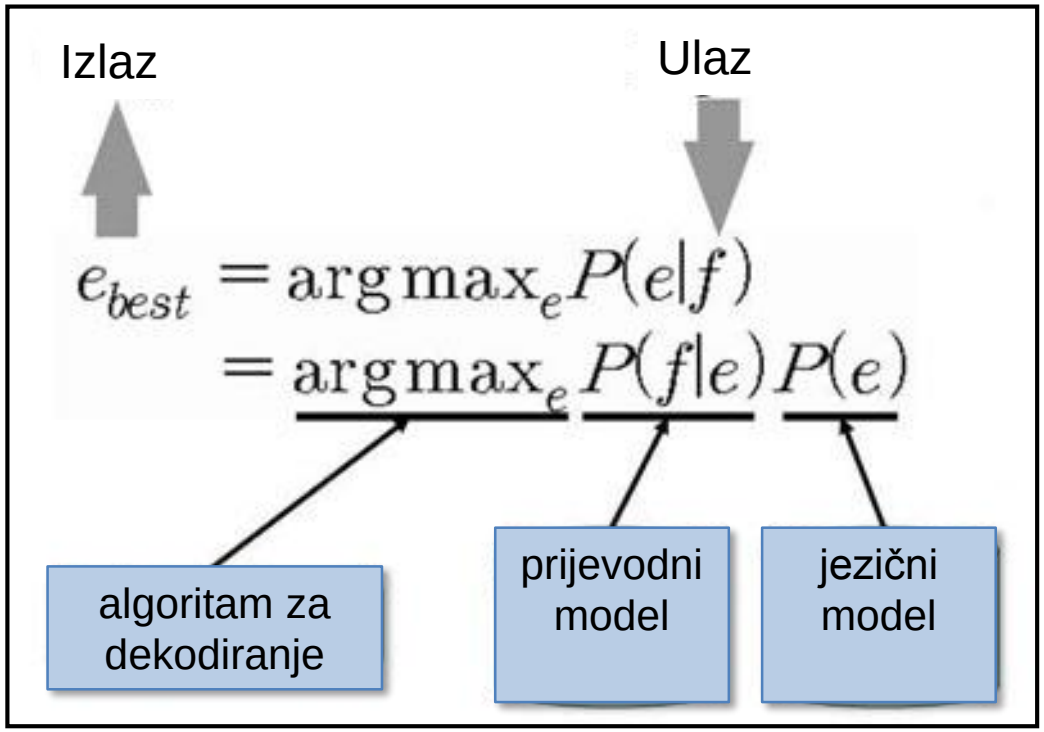
- Što je ispravno na ciljnome jeziku?
- Koje riječi mogu slijediti nakon kojih riječi, a koje ne mogu… gramatika?
- naučeno iz podataka…
  - *Je parle* ispravno …
  - *J' parlent* neispravno …
  - *la fille* ispravno …
  - *le fille* neispravno …
- *Je parle à la fille >> J' parlent à le fille*

European Language Resource Coordination
Connecting Europe Facility

Statističko strojno prevođenje



Kako radi suvremeno strojno prevođenje?



- Ne bojte se!
  Niti riječi danas o matematici.
- umjesto toga:
- priča o statističkome strojnom prevođenju u slikama …
- Sve je u podatcima …

SMT temeljen na frazama

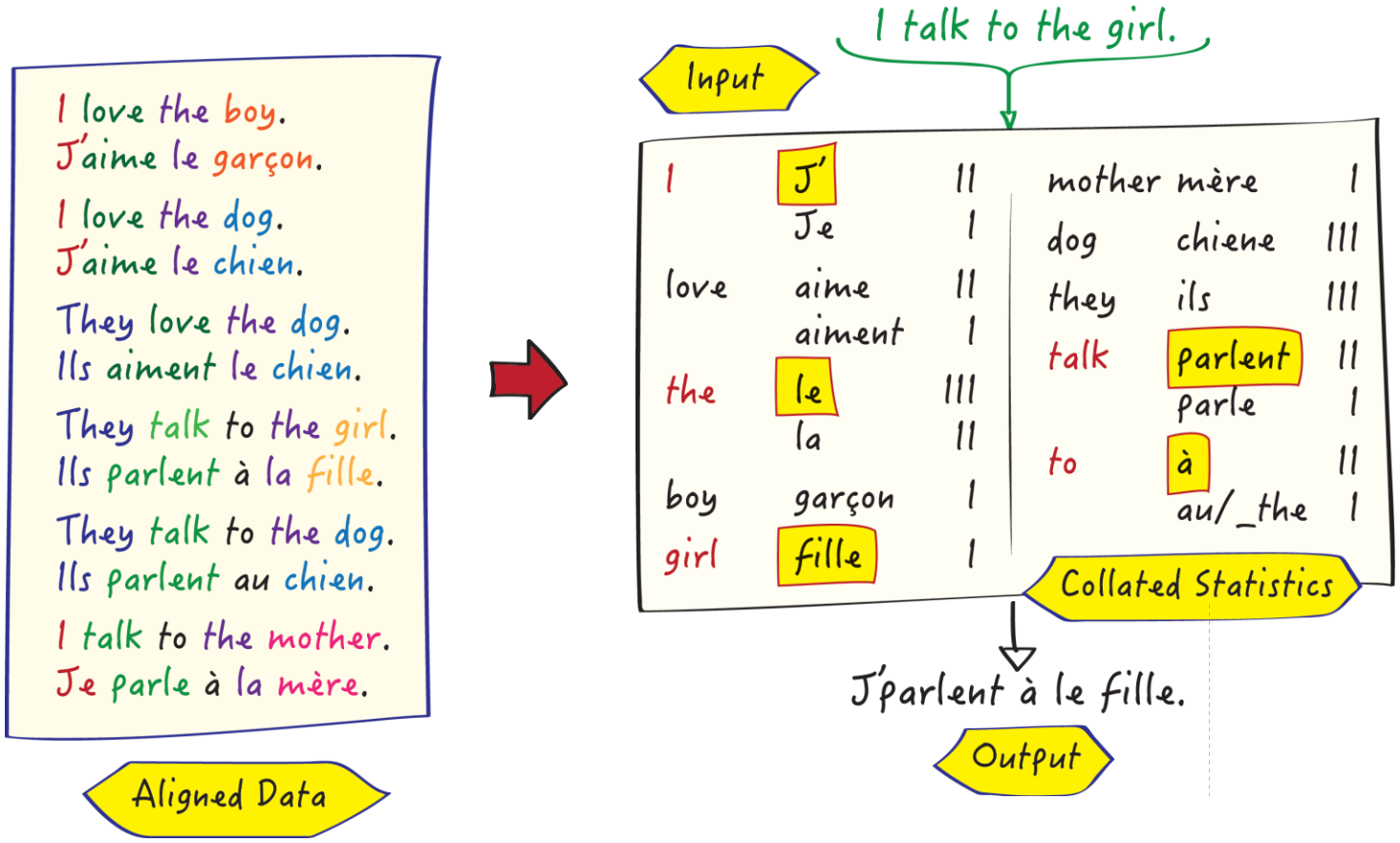
- do sad smo prevodili samo pojedinačne riječi
- gubi se okolina: npr. sročnost prema rodu (*le fille* …) itd.
- do neke to mjere može "popraviti" jezični model

- bolji prijevodni modeli:
- nisu izrađeni samo na prijevodima pojedinačnih riječi
- nego čitavih fraza:

– *the girl : la fille*
– *to the girl : a la fille*
– *I talk : Je parle*

European Language Resource Coordination
Connecting Europe Facility

„Obično” statističko strojno prevođenje



SMT temeljen na frazama

I love the boy.
J'aime le garçon.

I love the dog.
J'aime le chien.

They love the dog.
Ils aiment le chien.

They talk to the girl.
Ils parlent à la fille.

They talk to the dog.
Ils parlent au chien.

I talk to the mother.
Je parle à la mère.

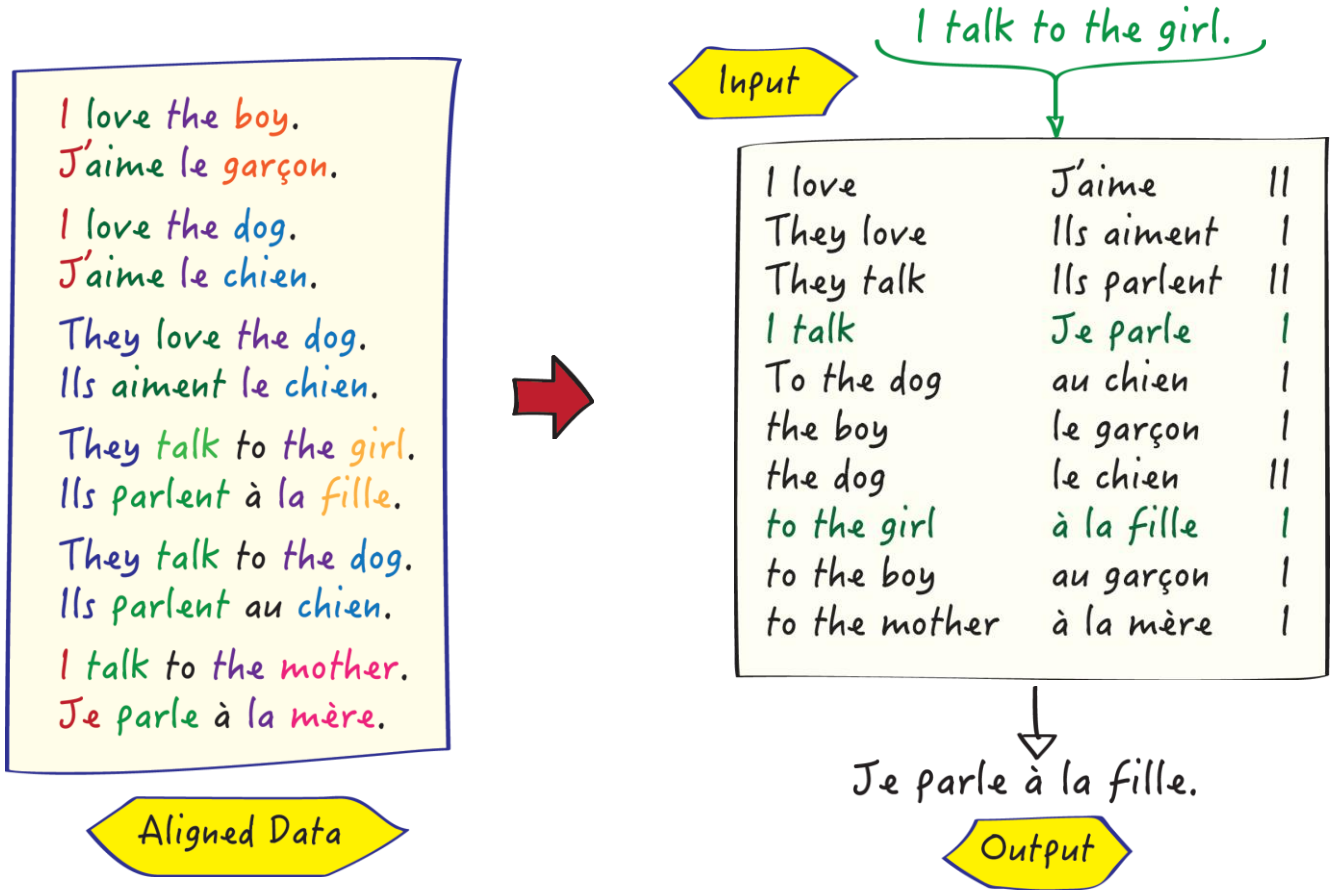
Aligned Data

Input

I talk to the girl.

I love	J'aime	II
They love	Ils aiment	I
They talk	Ils parlent	II
I talk	Je parle	I
To the dog	au chien	I
the boy	le garçon	I
the dog	le chien	II
to the girl	à la fille	I
to the boy	au garçon	I
to the mother	à la mère	I

SMT temeljen na frazama



SMT temeljen na frazama

- znatno bolje nego statističko prevođenje pojedinih riječi
- standardna tehnologija: Google, Microsoft, Baidu, globalna lokalizacijska i prijevodna industrija

MOSES CORE

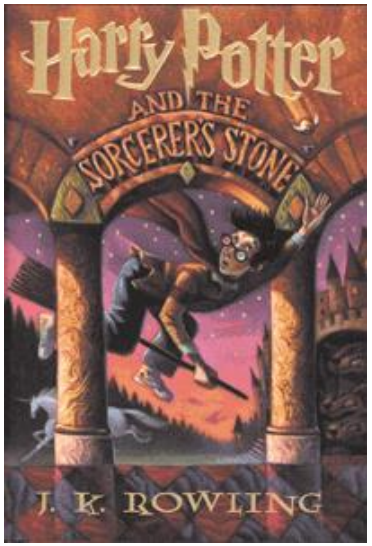
- Moses: sustav otvorenoga koda za SMT temeljen na frazama
- najraširenij SMT sustav u svijetu
- nastao na temelju istraživanja koje je poduprla EK
- koristi ga EK u DGT-jevoj platformi MT@EC

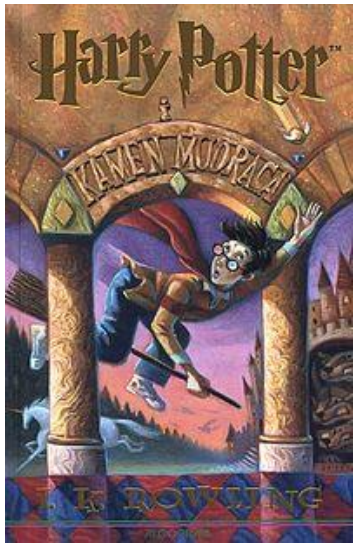
Strojno prevođenje i podatci

- u statističkome strojnom prevođenju sve ovisi o podatcima
- sustav za SMT iz podataka uči kako prevoditi
- podatci
  - prijevodi (dvojezični ili višejezični tekstovi, izvornici i prijevodi)
  - jednojezični podatci (tekstovi na ciljnome jeziku)
  - rječnici, zbirke termina, ontologije, popisi imena
- poput ljudi, SMT je dobar u onome što ga se naučilo

European Language Resource Coordination
Connecting Europe Facility

Strojno prevođenje i podatci



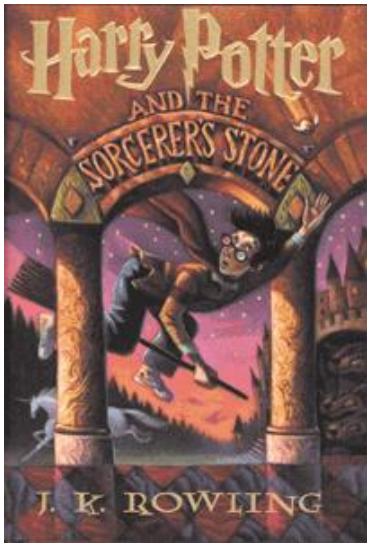


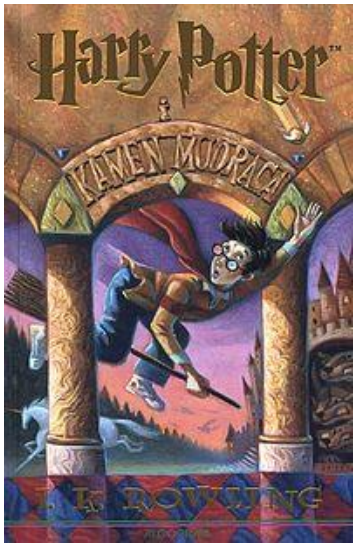

"Good," said Wizard Wink. "We'll start with Numbers."
"Oh no!" groaned Max and Mick.
The wizard boys didn't like school much. They wanted to play. Wizard Wink asked them the first question.
"If I had three crystal balls in one hand . . ." he said, "and four in the other . . . what would I have?"
"BIG 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."

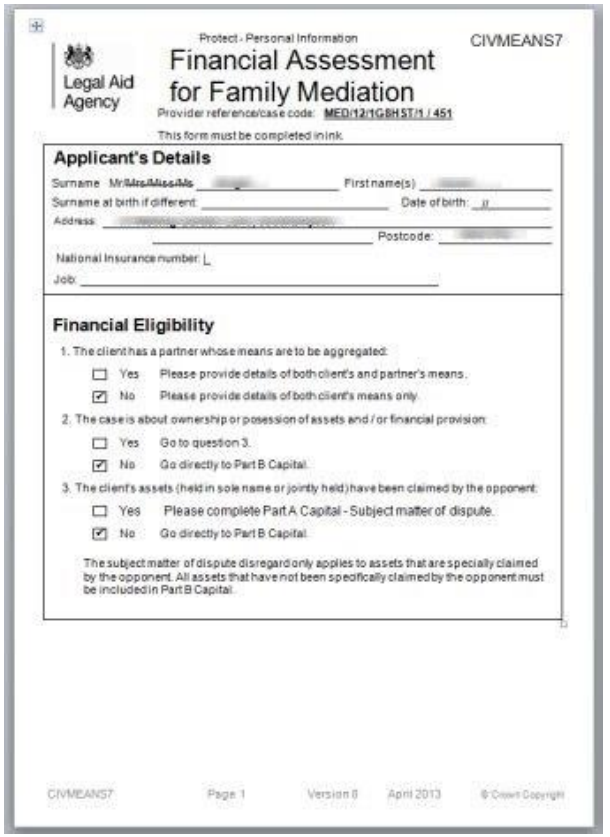
MOSES CORE

European Language Resource Coordination
Connecting Europe Facility

Strojno prevođenje i podatci






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

CEF.AT i podatci

- CEF.AT treba pravu vrstu podataka
- nacionalne vlade, javna administracija, javne službe, NVO-i
- CEF nudi usluge koje će omogućuju višejezično sudjelovanje pojedinačnih državljana kao i tijela javne vlasti država-članica EU-a

ELRC

- Pomozite nam pretvoriti CEF.AT u uspješnu
  - uslugu za državljane država-članica
  - uslugu za vas
  - potporu višejezičnosti
- Pomozite nam pronaći pravu vrstu podataka!
- Potpora hrvatskome jeziku potpora je ostalim europskim jezicima i obrnuto!

European Language Resource Coordination
Connecting Europe Facility

Zahvaljujem na pozornosti.