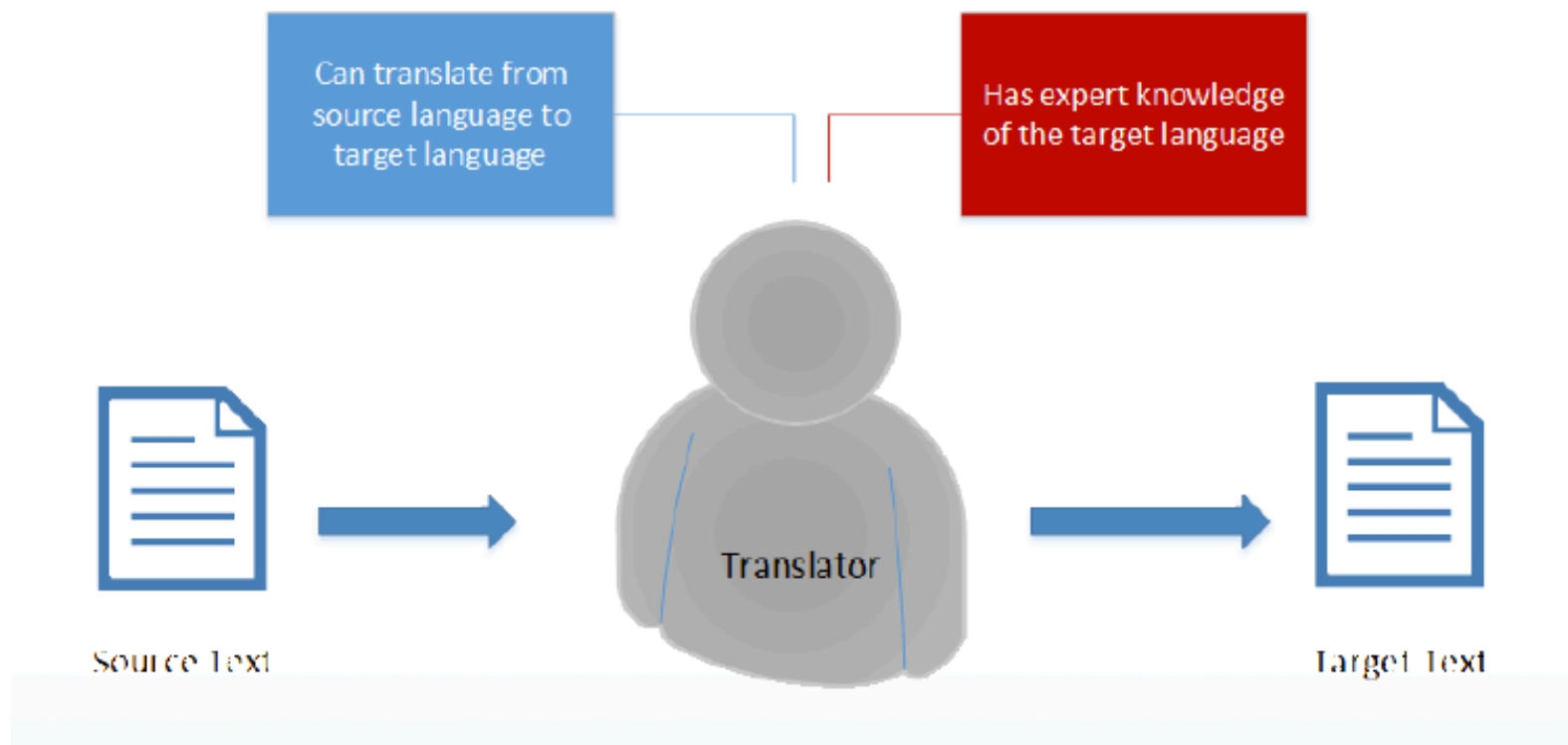
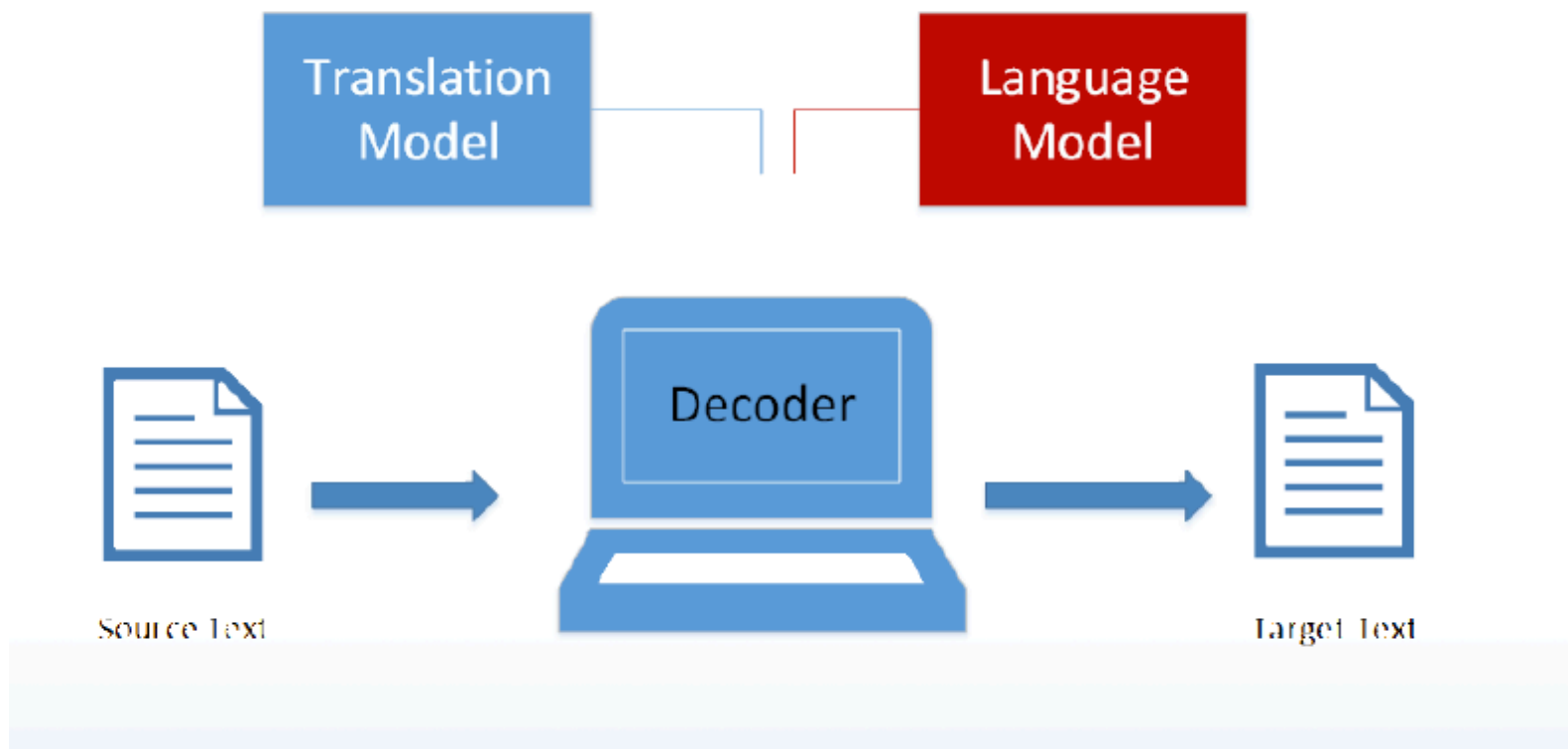
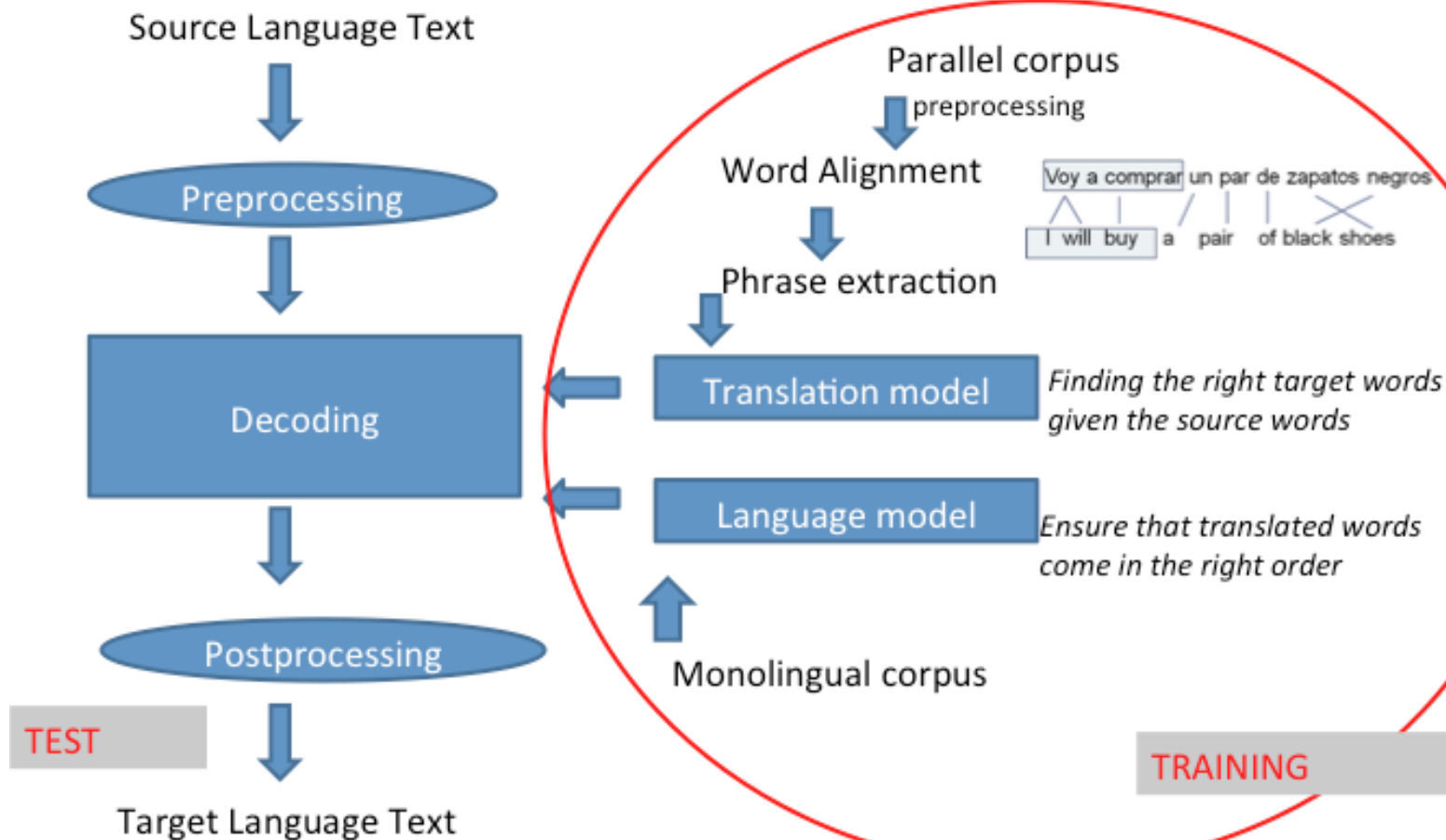


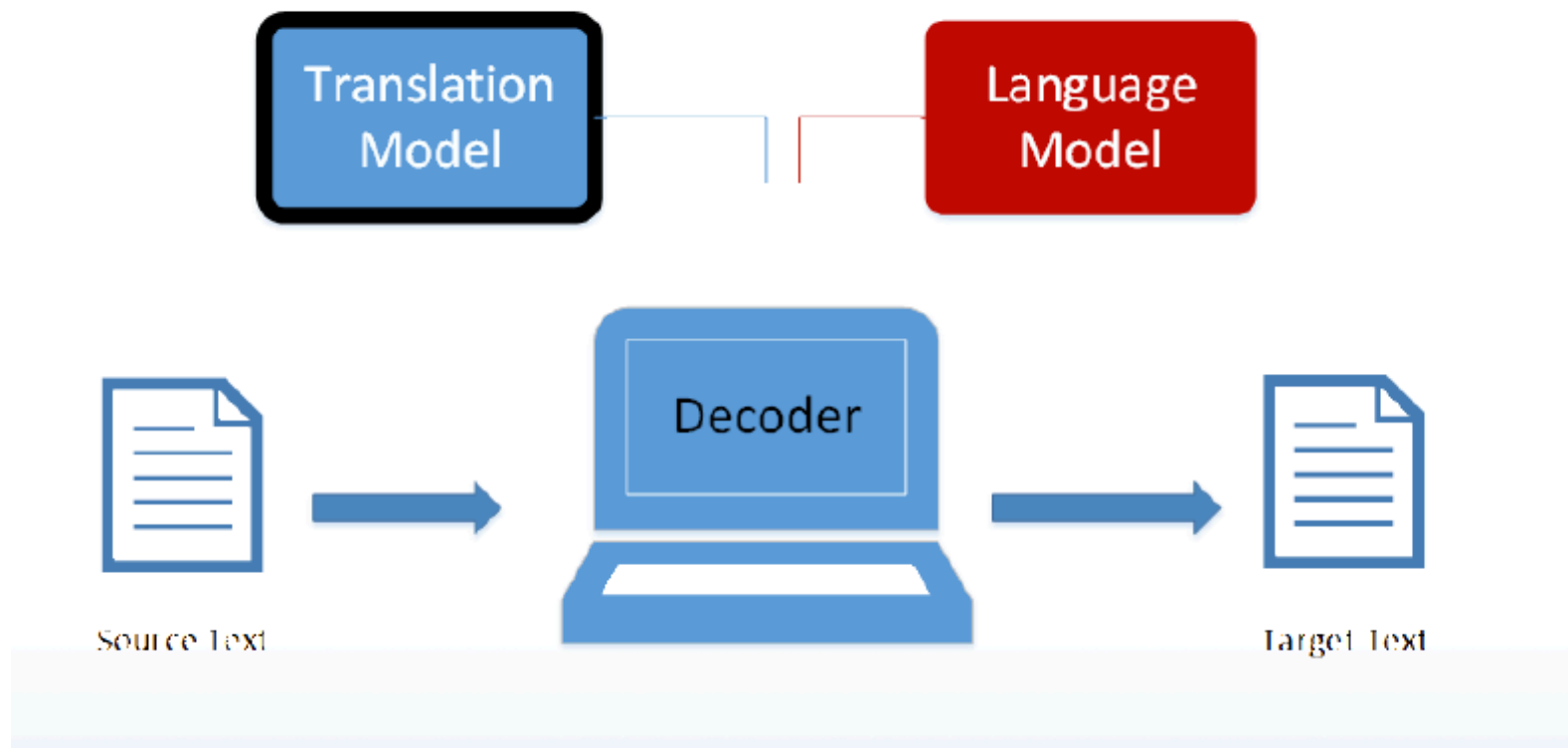
Automatisch Vertalen SMT

Lieve Macken









Hoe leert een computer vertalen?

Woordenschatverwerving in SMT



鱼汤
糖醋老鸭

yú tāng
táng cù lǎo yā

鸡汤	jī tāng	chicken soup
老鸭汤	lǎo yā tāng	duck soup
酸辣汤	suān là tāng	hot and sour soup
腰果鸡丁	yāo guǒ jī dīng	cashew chicken
辣子鸡丁	là zǐ jī dīng	spicy chicken
糖醋里肌	táng cù lǐ jī	sweet and sour pork
辣子猪肉丁	là zǐ zhū ròu dīng	spicy pork
糖醋鱼	táng cù yú	sweet and sour fish
红烧鱼	hóng shāo yú	fish in soy sauce
鱼汤	yú tāng	?
糖醋老鸭	táng cù lǎo yā	?

Google

鱼

Web

Images

Maps

Videos

News

More ▾



Google

鱼汤

Web

Images

Videos

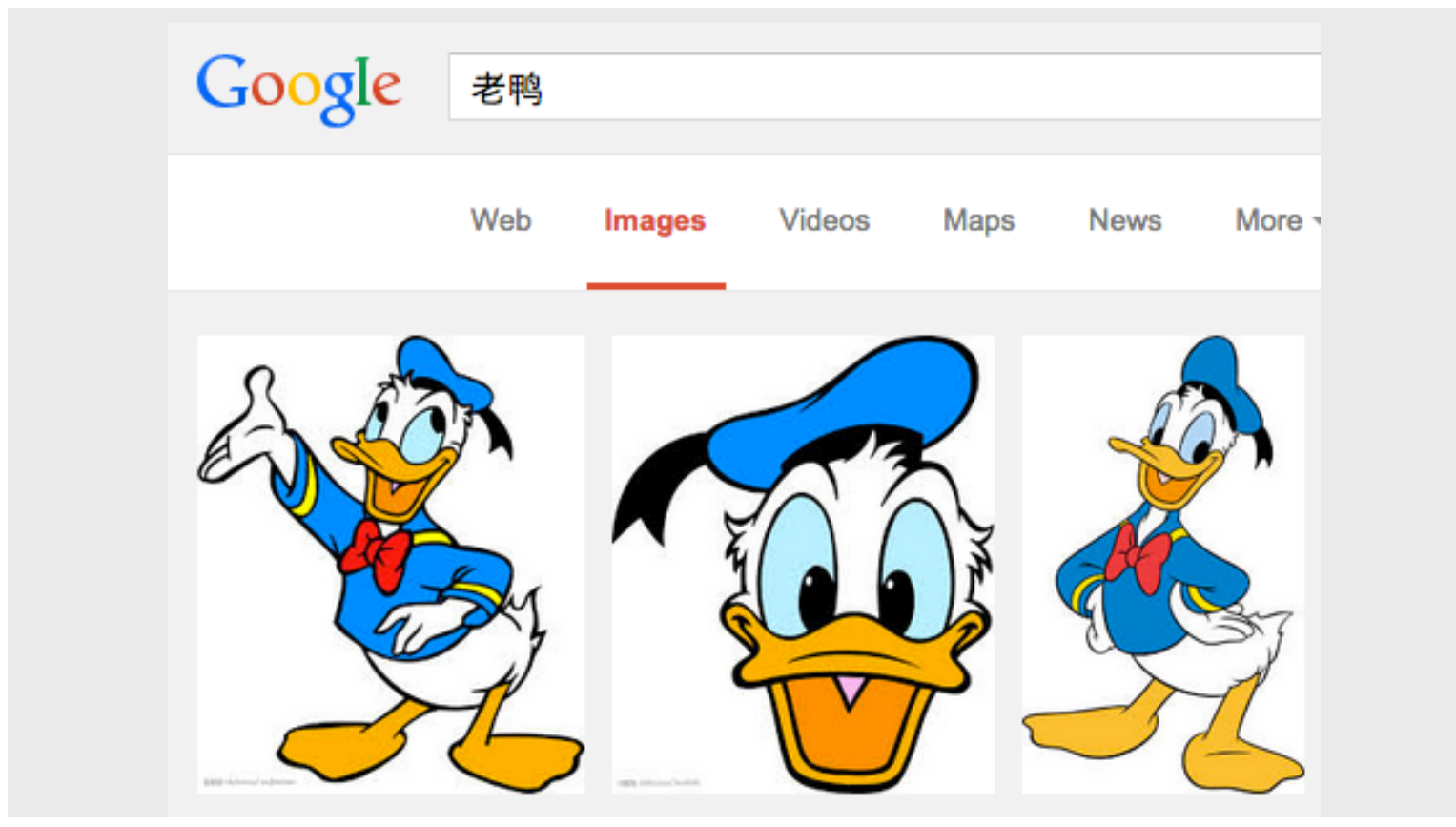
Maps

News

More ▾

Search tools





Google

糖醋老鸭



Web

Images

Videos

Maps

News

More ▾

Search tools



Co-occurrence frequency

鸡汤

jī tāng

chicken soup

老鸭汤

lǎo yā tāng

duck soup

酸辣汤

suān là tāng

hot and sour soup

...

糖醋里肌

táng cù lǐ jī

sweet and sour pork

糖醋鱼

táng cù yú

sweet and sour fish

红烧鱼

hóng shāo yú

fish in soy sauce

Co-occurrence frequency

鱼汤 = fish soup; 糖醋 = sweet and sour

鸡汤

jī tāng

chicken soup

老鸭汤

lǎo yā tāng

duck soup

酸辣汤

suān là tāng

hot and sour soup

...

糖醋里肌

táng cù lǐ jī

sweet and sour pork

糖醋鱼

táng cù yú

sweet and sour fish

红烧鱼

hóng shāo yú

fish in soy sauce

Beredeneerd gokje

糖醋老鸭 = sweet and sour duck

鸡汤

jī tāng

chicken soup

老鸭汤

lǎo yā tāng

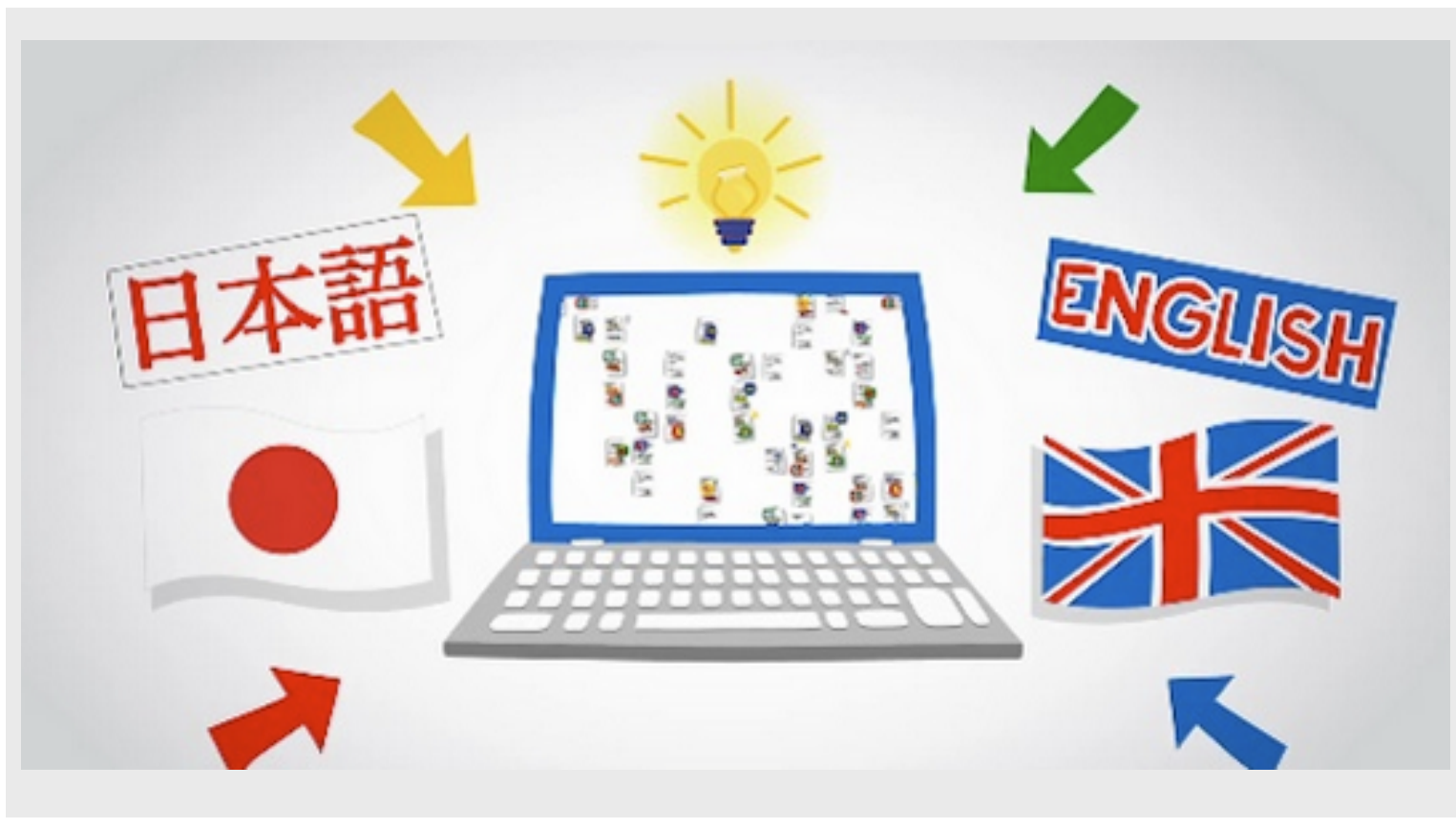
duck soup

酸辣汤

suān là tāng

hot and sour soup

...



Belangrijke begrippen

- Voorwaardelijke kans: $P(a|b)$
waarde tussen 0 en 1

Belangrijke begrippen

- Voorwaardelijke kans: $P(a|b)$
waarde tussen 0 en 1

$$P(\text{soup}|\text{tāng}) = \frac{\#(\text{soup}, \text{tāng})}{\#(\text{tāng})} = \frac{3}{3} = 1$$

Belangrijke concepten

- Voorwaardelijke kans: $P(a|b)$
waarde tussen 0 en 1

$$P(\text{soup}|\text{tāng}) = \frac{\#(\text{soup}, \text{tāng})}{\#(\text{tāng})} = \frac{3}{3} = 1$$

$$P(\text{chicken}|\text{jī}) = \frac{\#(\text{chicken}, \text{jī})}{\#(\text{jī})} = \frac{3}{4} = 0.75$$

Belangrijke concepten

- Voorwaardelijke kans: $P(a|b)$
waarde tussen 0 en 1

$$P(\text{soup}|\text{tāng}) = \frac{\#(\text{soup}, \text{tāng})}{\#(\text{tāng})} = \frac{3}{3} = 1$$

$$P(\text{chicken}|\text{jī}) = \frac{\#(\text{chicken}, \text{jī})}{\#(\text{jī})} = \frac{3}{4} = 0.75$$

$$P(\text{chicken}|\text{dīng}) = \frac{\#(\text{chicken}, \text{dīng})}{\#(\text{dīng})} = \frac{2}{3} = 0.67$$

Beperking: enkel “woorden”

$$P(\text{chicken} | \text{jī dīng}) = \frac{\#(\text{chicken}, \text{jī dīng})}{\#(\text{jī dīng})} = \frac{2}{2} = 1$$

Uniforme verdeling

jī tāng



chicken soup

lǎo yā tāng



duck soup

suān là tāng



hot and sour soup

yāo guǒ jī dīng



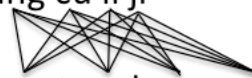
cashew chicken

là zì jī dīng



spicy chicken

táng cù lǐ jī



sweet and sour pork

là zì zhū ròu dīng



spicy pork

táng cù yú



sweet and sour fish

hóng shāo yú



fish in soy sauce

Na 1 iteratie

jī tāng

chicken soup

lǎo yā tāng

duck soup

suān là tāng

hot and sour soup

yāo guǒ jī dīng

cashew chicken

là zì jī dīng

spicy chicken

táng cù lǐ jī

sweet and sour pork

là zì zhū ròu dīng

spicy pork

táng cù yú

sweet and sour fish

hóng shāo yú

fish in soy sauce

Na 2 iteraties

jī tāng
chicken soup

lǎo yā tāng
duck soup

suān là tāng
hot and sour soup

yāo guǒ jī dīng
cashew chicken

là zì jī dīng
spicy chicken

táng cù lǐ jī
sweet and sour pork

là zì zhū ròu dīng
spicy pork

táng cù yú
sweet and sour fish

hóng shāo yú
fish in soy sauce

Tot convergentie

jī tāng
chicken soup

lǎo yā tāng
duck soup

suān là tāng
hot and sour soup

yāo guǒ jī dīng
cashew chicken

là zì jī dīng
spicy chicken

táng cù lǐ jī
sweet and sour pork

là zì zhū ròu dīng
spicy pork

táng cù yú
sweet and sour fish

hóng shāo yú
fish in soy sauce

Tijd voor dessert?

麻饼 má bǎng

Tijd voor dessert?

麻饼 má bǎng



Datagebaseerde methode

- Computer leidt alle kennis af uit data
- Meer data → meer evidentie → betere kwaliteit
- Kwaliteit ~ mate waarin te vertalen teksten lijken op trainingsmateriaal

Chinees-Nederlands

鸡汤

jī tāng

kippensoep

老鸭汤

lǎo yā tāng

eendensoep

酸辣汤

suān là tāng

zoetzure soep (heet)

...

糖醋里肌

táng cù lǐ jī

varkensvlees, zoetzuur

糖醋鱼

táng cù yú

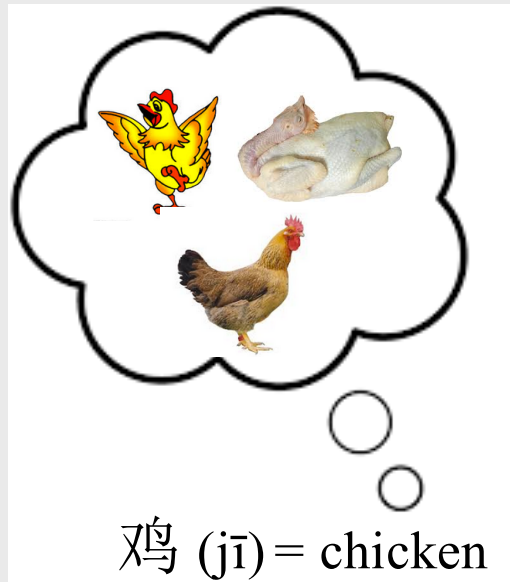
vis, zoetzuur

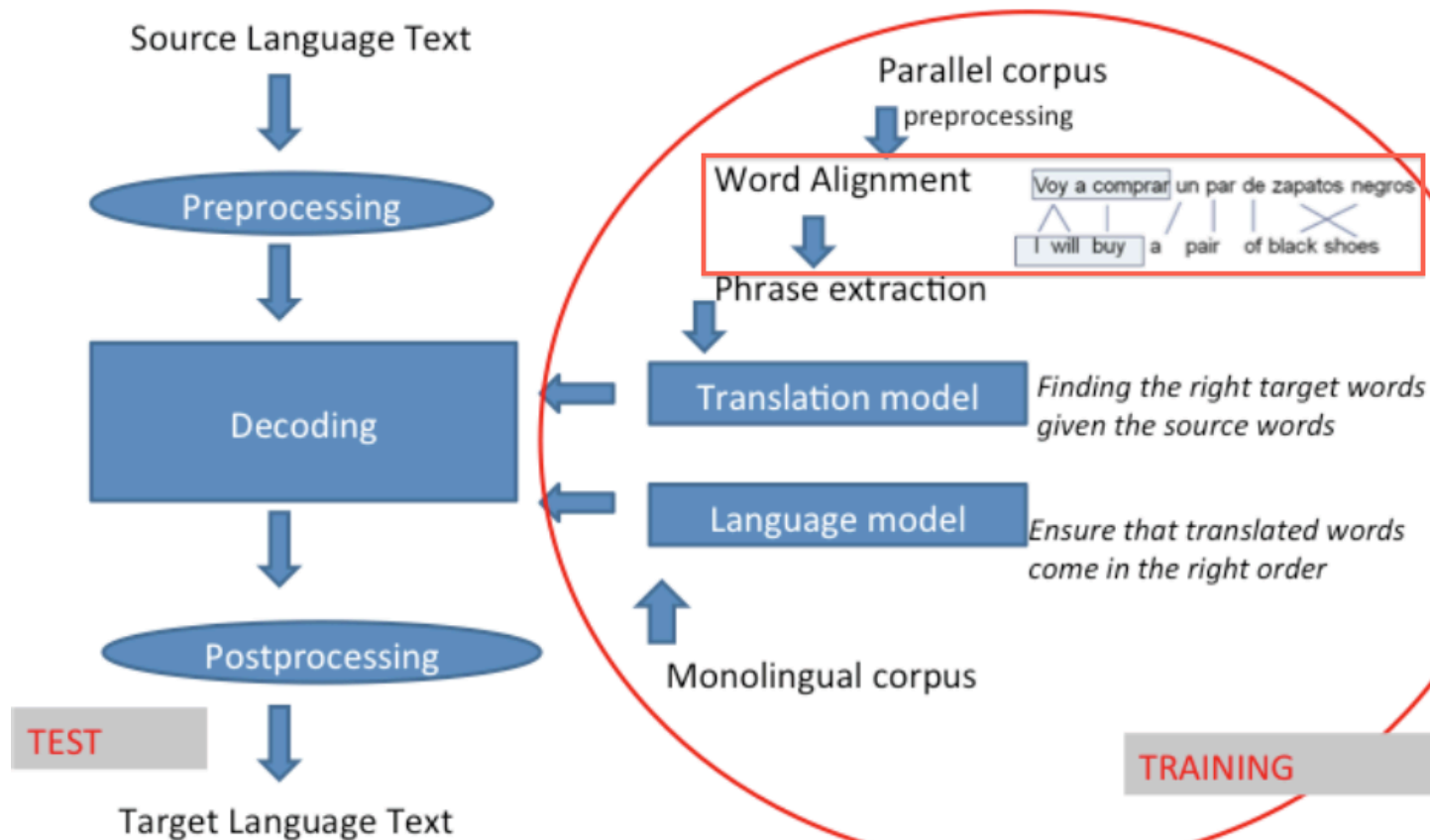
红烧鱼

hóng shāo yú

vis in sojasaus

Betekenis?







	dat	het	beginsel	van	vrije	mededinging	moet	worden	gerespecteerd
that									
the									
principle									
of									
open									
competition									
must									
be									
adhered									
to									



	dat	het	beginsel	van	vrije	mededinging	moet	worden	gerespecteerd
that									
the									
principle									
of									
open									
competition									
must									
be									
adhered									
to									



	dat	het	beginsel	van	vrije	mededinging	moet	worden	gerespecteerd
that									
the									
principle									
of									
open									
competition									
must									
be									
adhered									
to									



	dat	het	beginsel	van	vrije	mededinging	moet	worden	gerespecteerd
that									
the									
principle									
of									
open									
competition									
must									
be									
adhered									
to									

Phrase table

competition		concurrentie		0.810473	0.853619	0.20
competition		mededinging		0.840251	0.899487	0.27
competition		de mededinging		0.783431	0.899487	0
competition		de concurrentie		0.586288	0.853619	0
competition		vergelijkend onderzoek		0.654643	0.00	0.00
competition		wedstrijd		0.441456	0.420908	0.0097
competition		concurreren		0.100731	0.131467	0.00
competition		van de mededinging		0.106661	0.899487	0.00
competition		mededingingsvoorwaarden		0.220767	0.00	0.00
competition		de		1.88422e-05	0.0001257	0.0038995
competition		concurrentievoorwaarden		0.155522	0.00	0.00
competition		competition		0.978469	0.866799	0.00
competition		van concurrentie		0.230769	0.853619	0.00

Translate

Instantvertaling uitschakelen



Engels Nederlands Frans Taal herkennen ▼



Nederlands Engels Frans ▼

Vertaal

that the principle of open competition must
be adhered to

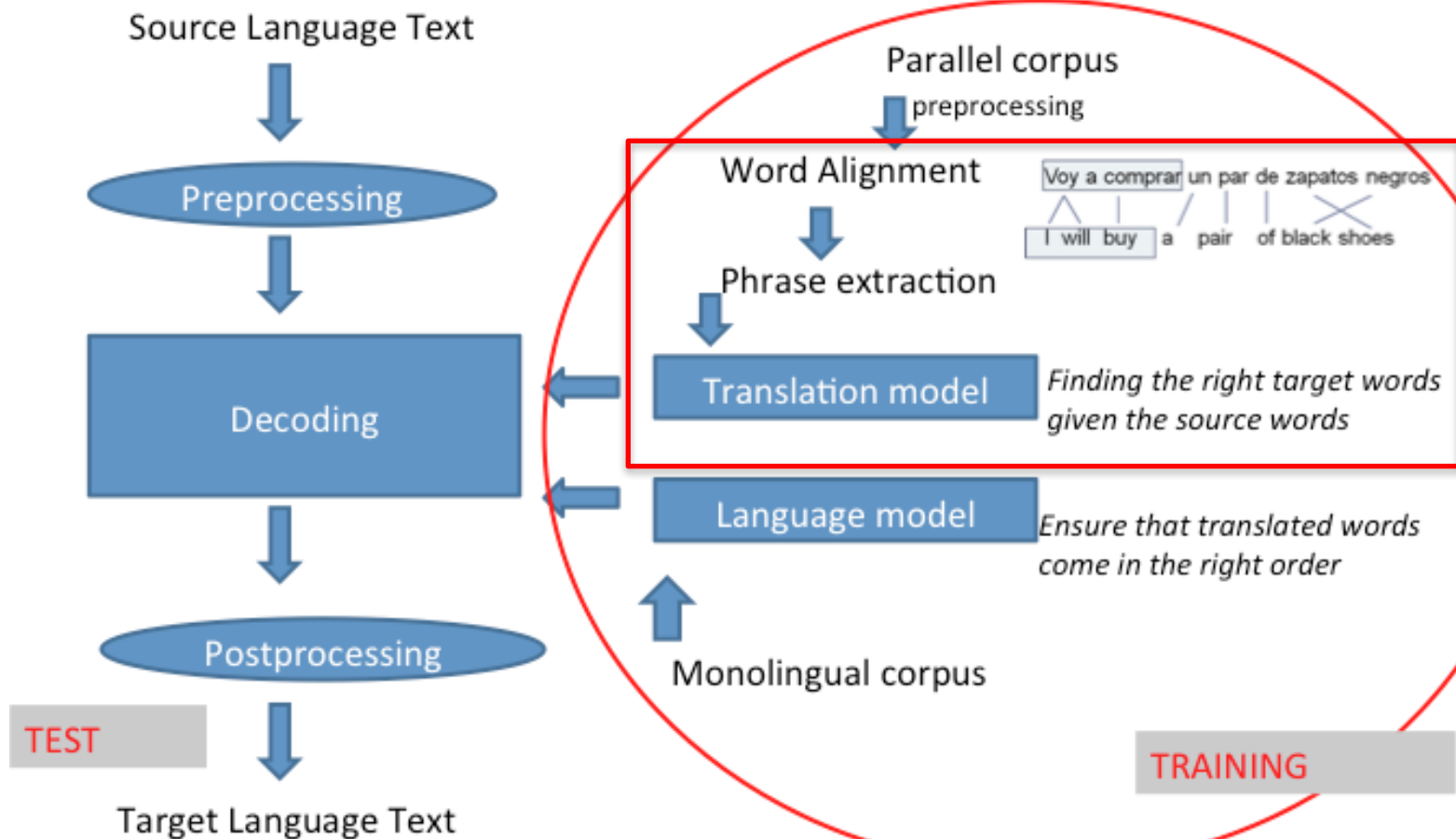


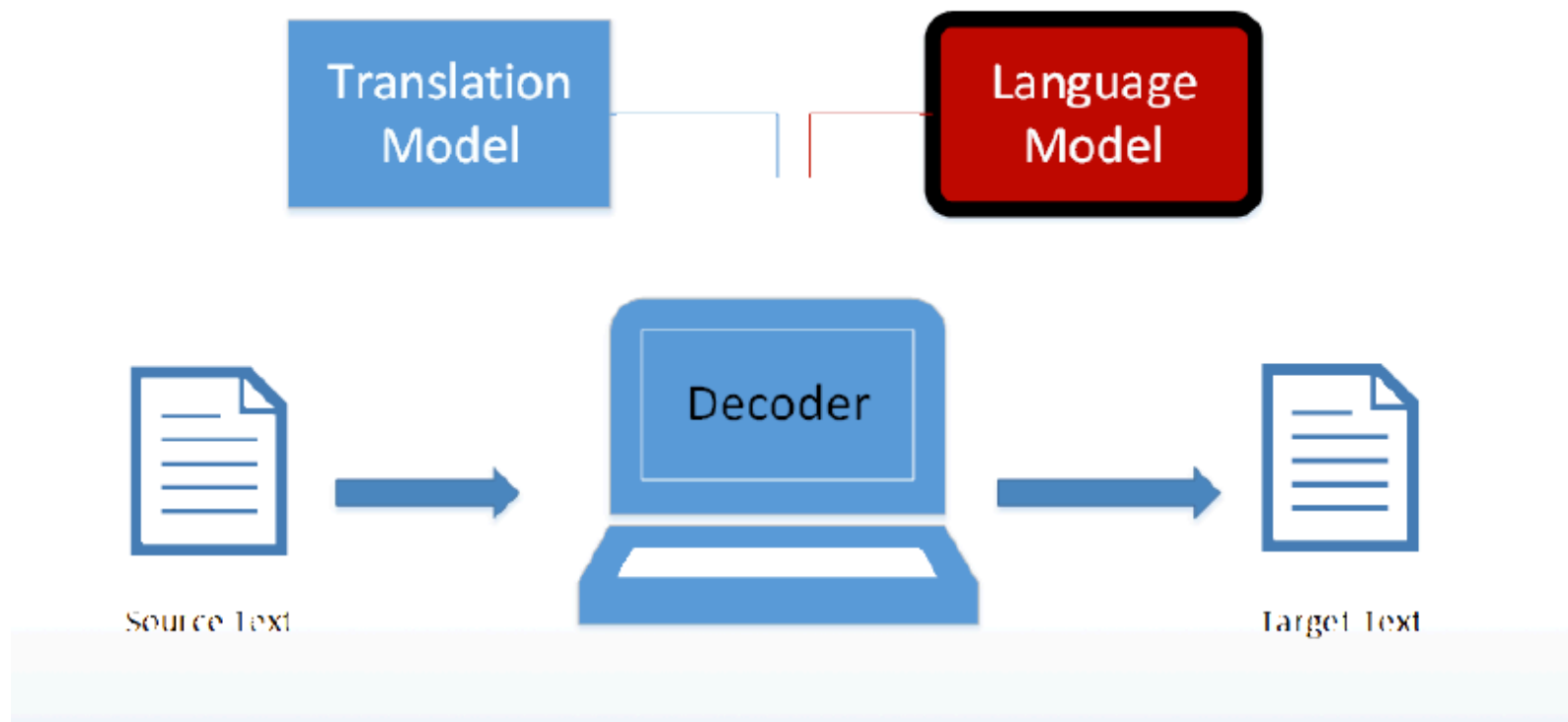
dat het beginsel van vrije mededinging moet
worden nageleefd

moet worden nageleefd
moeten worden gerespecteerd
dienen te worden nageleefd
acht moet worden genomen
moet worden gehouden

Deze vertaling verbeteren







Hoe leert een computer wat correct Engels/Nederlands is?

Taalmodel

I like Chinese ...

Taalmodel

I like Chinese food
New Year
tea
beer
.
be

Taalmodel

N-gram = sequentie van woorden

n-gram van lengte 1 = unigram (woord)

n-gram van lengte 2 = bigram

n-gram van lengte 3 = trigram

Taalmodel

Bigram: I like Chinese food

I like
like Chinese
Chinese food

Taalmodel

Trigram: I like Chinese food

I like Chinese
like Chinese food

Taalmodel

Hypothese: Indien een zin veel plausibele n-grammen bevat is het een plausibele (“correcte”) zin

“I want Chinese food”

“I want food Chinese”

“I want Chinese lunch”

Taalmodel

Hypothese: Indien een zin veel plausibele n-grammen bevat is het een plausibele (“goede”) zin

“I want Chinese food”	1
“I want food Chinese”	3
“I want Chinese lunch”	2

Taalmodel

N-gram probabiliteit (monolinguale corpora)

bigram

$$P(y \mid x) = \frac{\#("x \ y")}{\#("x")}$$

trigram

$$P(z \mid x \ y) = \frac{\#("x \ y \ z")}{\#("x \ y")}$$

Taalmodel

“I want Chinese food”

$$P(\text{want}|\text{I}) \times P(\text{Chinese}|\text{want}) \times P(\text{food}|\text{Chinese})$$

“I want food Chinese”

$$P(\text{want}|\text{I}) \times P(\text{food}|\text{want}) \times P(\text{Chinese}|\text{food})$$

“I want Chinese lunch”

$$P(\text{want}|\text{I}) \times P(\text{Chinese}|\text{want}) \times P(\text{lunch}|\text{Chinese})$$

Taalmodel: bi-gram probabiliteiten

I want Chinese food

$$P(\text{want}|\text{I}) \times P(\text{Chinese}|\text{want}) \times P(\text{food}|\text{Chinese}) = 0.32 \times 0.0049 \times 0.56 \\ = 0.0008781$$

	I	want	to	eat	Chinese	food	lunch
I	.0023	.32	0	.0038	0	0	0
want	.0025	0	.65	0	.0049	.0066	.0049
to	.00092	0	.0031	.26	.00092	0	.0037
eat	0	0	.0021	0	.020	.0021	.055
Chinese	.0094	0	0	0	0	.56	.0047
food	.013	0	.011	0	0	0	0
lunch	.0087	0	0	0	0	.0022	0

Figure 6.5 Bigram probabilities for 7 of the words (out of 1616 total word types) in the Berkeley Restaurant Project corpus of ~10,000 sentences.

Taalmodel: bi-gram probabiliteiten

I want food Chinese

$$P(\text{want}|\text{I}) \times P(\text{food}|\text{want}) \times P(\text{Chinese}|\text{food}) = 0.32 \times 0.0066 \times 0 = 0$$

	I	want	to	eat	Chinese	food	lunch
I	.0023	.32	0	.0038	0	0	0
want	.0025	0	.65	0	.0049	.0066	.0049
to	.00092	0	.0031	.26	.00092	0	.0037
eat	0	0	.0021	0	.020	.0021	.055
Chinese	.0094	0	0	0	0	.56	.0047
food	.013	0	.011	0	0	0	0
lunch	.0087	0	0	0	0	.0022	0

Figure 6.5 Bigram probabilities for 7 of the words (out of 1616 total word types) in the Berkeley Restaurant Project corpus of ~10,000 sentences.

Taalmodel: bi-gram probabiliteiten

I want Chinese lunch

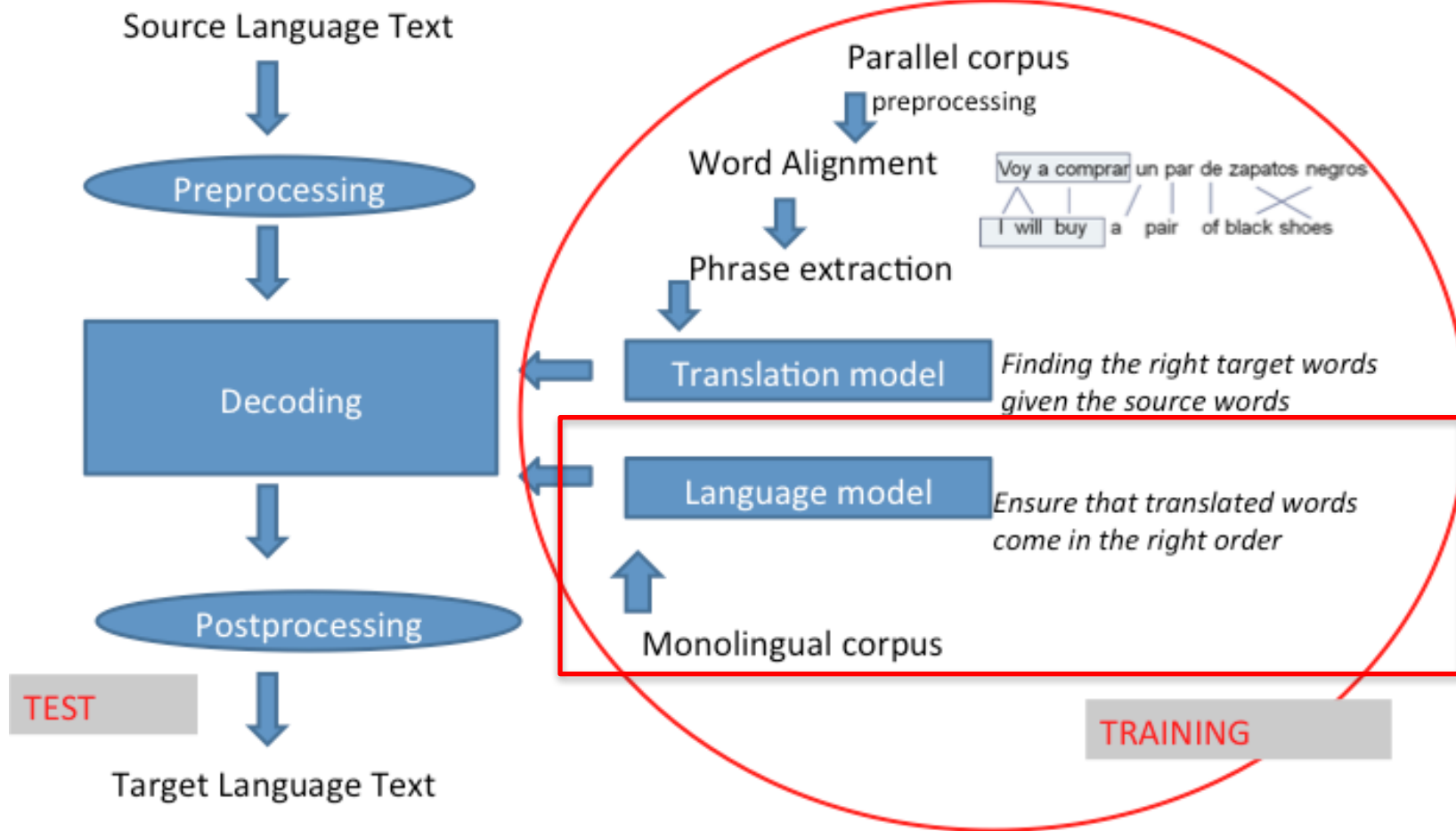
$$P(\text{want}|\text{I}) \times P(\text{Chinese}|\text{want}) \times P(\text{lunch}|\text{Chinese}) = 0.32 \times 0.0049 \times 0.0047 = 0.0000074$$

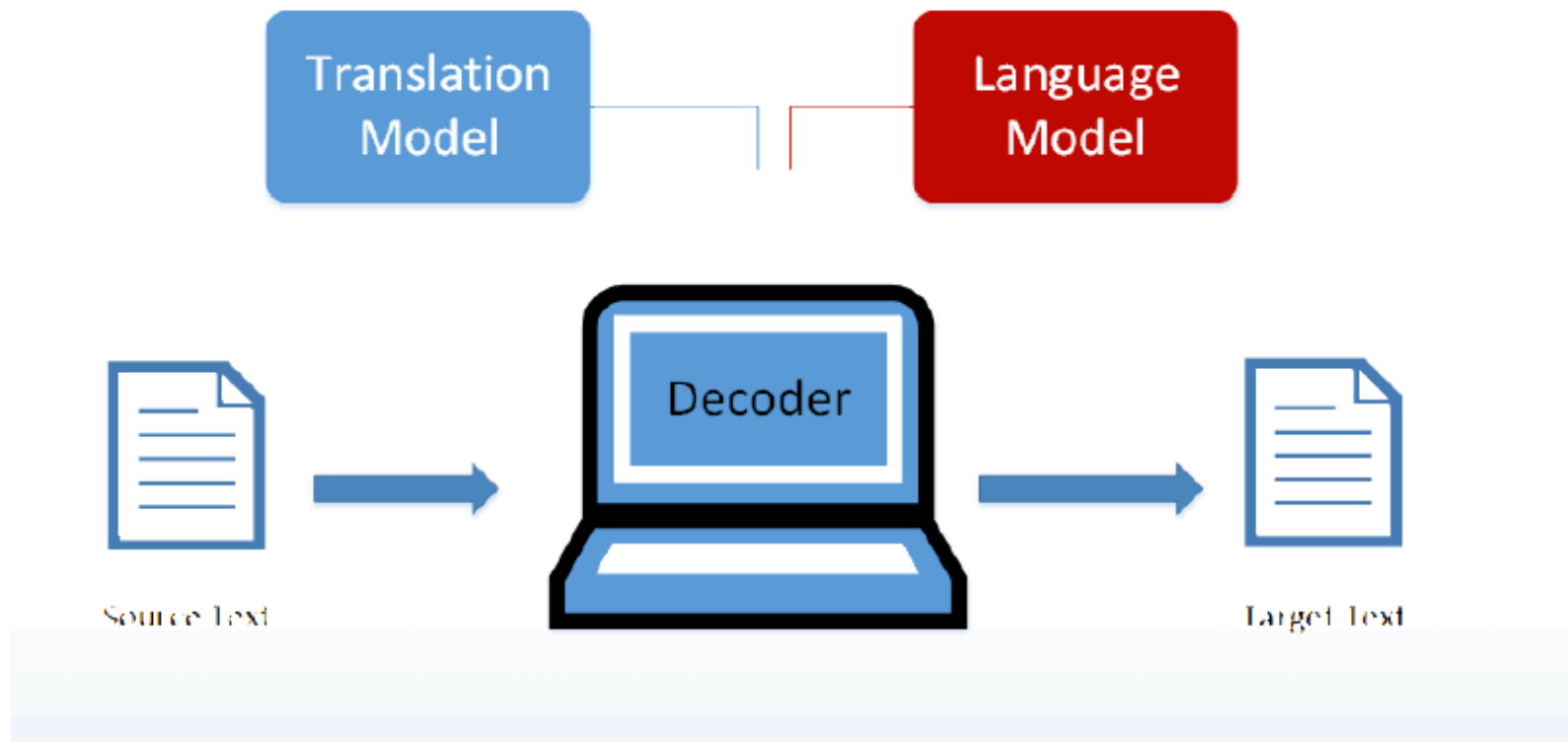
	I	want	to	eat	Chinese	food	lunch
I	.0023	.32	0	.0038	0	0	0
want	.0025	0	.65	0	.0049	.0066	.0049
to	.00092	0	.0031	.26	.00092	0	.0037
eat	0	0	.0021	0	.020	.0021	.055
Chinese	.0094	0	0	0	0	.56	.0047
food	.013	0	.011	0	0	0	0
lunch	.0087	0	0	0	0	.0022	0

Figure 6.5 Bigram probabilities for 7 of the words (out of 1616 total word types) in the Berkeley Restaurant Project corpus of ~10,000 sentences.

Taalmodel

“I want Chinese food”	0.0008781	1
“I want Chinese lunch”	0.0000074	2
“I want food Chinese”	0	3





that

dat

die

these

deze

deze regels

deze voorschriften

dat deze regels

rules

regels

must

moet

be

zijn

adhered

gehecht

to

naar

wordt voldaan

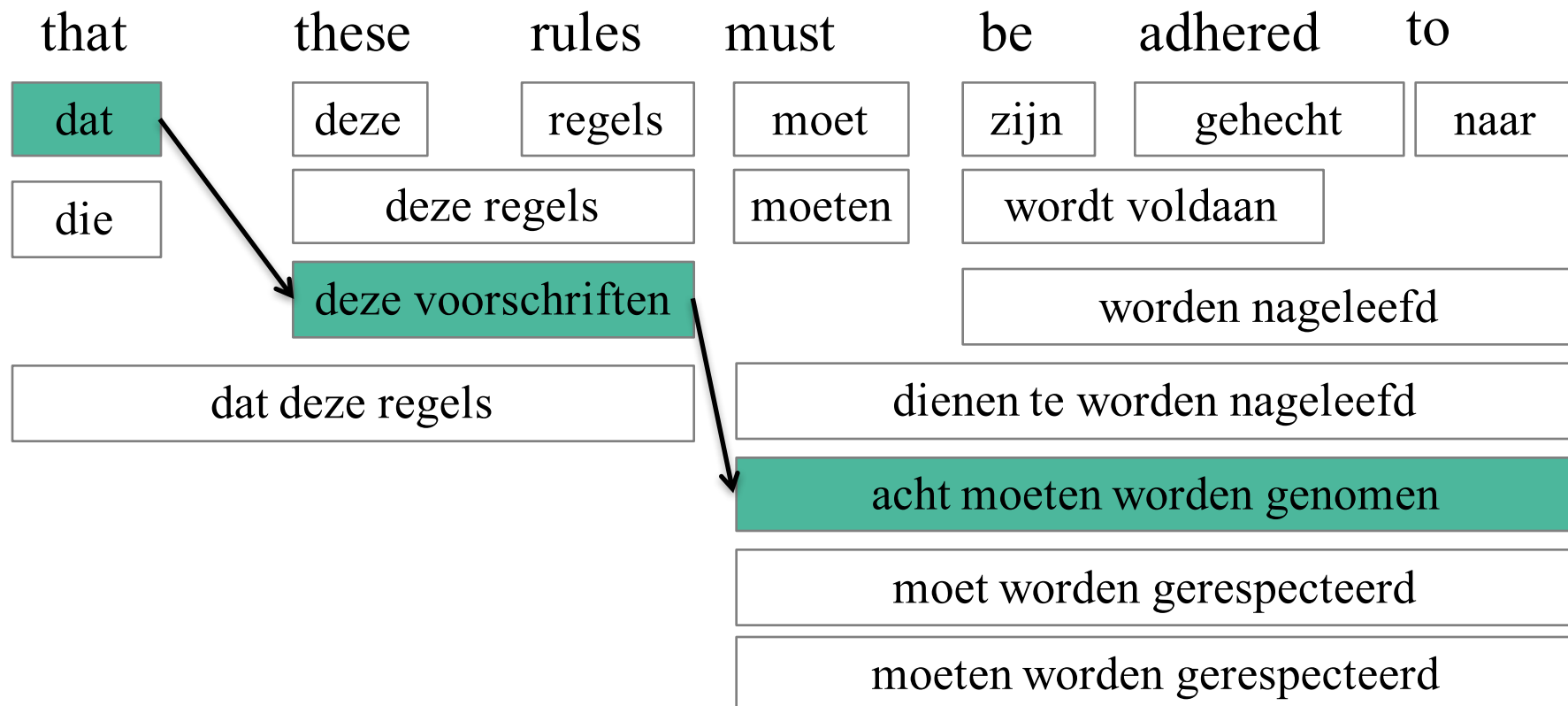
worden nageleefd

dienen te worden nageleefd

acht moeten worden genomen

moet worden gerespecteerd

moeten worden gerespecteerd



that

these

rules

must

be

adhered

to

dat

deze

regels

moet

zijn

gehecht

naar

die

deze regels

wordt voldaan

deze voorschriften

worden nageleefd

dat deze regels

dienen te worden nageleefd

acht moeten worden genomen

moet worden gerespecteerd

moeten worden gerespecteerd

that

dat

die

these

deze

deze regels

deze voorschriften

rules

regels

must

moet

be

zijn

adhered

gehecht

to

naar

wordt voldaan

worden nageleefd

dienen te worden nageleefd

acht moeten worden genomen

moet worden gerespecteerd

moeten worden gerespecteerd

dat deze regels

Vertaalmodel

- Hogere probabilliteit voor zinnen met dezelfde betekenis
- Probabiliteiten op basis van bilinguale corpora

Taalmodel

- Hogere probabilliteit voor grammaticaal correcte zinnen
- Probabiliteiten op basis van monolinguale corpora

Decoder

- Maakt gebruik van taal- en vertaalmodel
- Zoekt naar combinatie van frases met hoogste probabilliteit

Typische SMT fouten

Woorden ontbreken (scheidbare werkwoorden)

women return home with

vrouwen [~~keren~~] terug naar huis met ...

Verkeerde woordbetekenis

Episodes of personal violence could increase

Afleveringen van persoonlijk geweld zouden ...

Typische SMT fouten

Woordvolgorde (geen inversie)

... omdat het merk is minder bekend

Gebrek aan congruentie

Emissies van schepen zal worden gemonitord

Nederlandse samenstellingen

de windenergie sector in Europa

SMT-kwaliteit

Eeuwenlang hebben **bestonden** [included] Chinese huishoudens **uit** vele generaties ~~inbegrepen~~, en Chinese ouderen konden rekenen op hun kinderen ~~de zorg voor~~ **om voor hen te zorgen** [caring for them] als ze groeiden ~~broos~~ **zwakker werden** [as they grew frail].

Referenties

Andy Way and Mary Hearne (2011) On the Role of Translations in State-of-the-Art Statistical Machine Translation. *Language and Linguistics Compass* 5:227—248

Philipp Koehn (2010) Statistical Machine Translation. Cambridge University Press

Szymon Kloczek (2015) MT@EC. What's behind it?