

NEURALE AUTOMATISCHE VERTAALSISTEMEN

Lieve Macken, ELRC workshop, 8 juli 2021

OVERZICHT

- Kwaliteit van NMT-systemen
 - Artificiële neurale netwerken
 - Word embeddings
 - Architectuur
 - Belang van data
 - Mens versus machine?
- Josef van Genabith, "Data or no Data, that is the Question: Learning MT without Translation Data?". 2021. Presentation at JIAMCATT (19.04.2021)
 - Juan Antonio Pérez-Ortiz, Mikel L. Forcada and Felipe Sánchez-Martínez, "How neural machine translation works", 2021. Preprint version of book chapter (<https://multitrainmt.eu> project)
 - Koehn, P. (2020). Neural Machine Translation. Cambridge: Cambridge University Press.

KWALITEIT VAN NMT-SYSTEMEN

Officiële resultaten van de WMT 2020 News Translation task

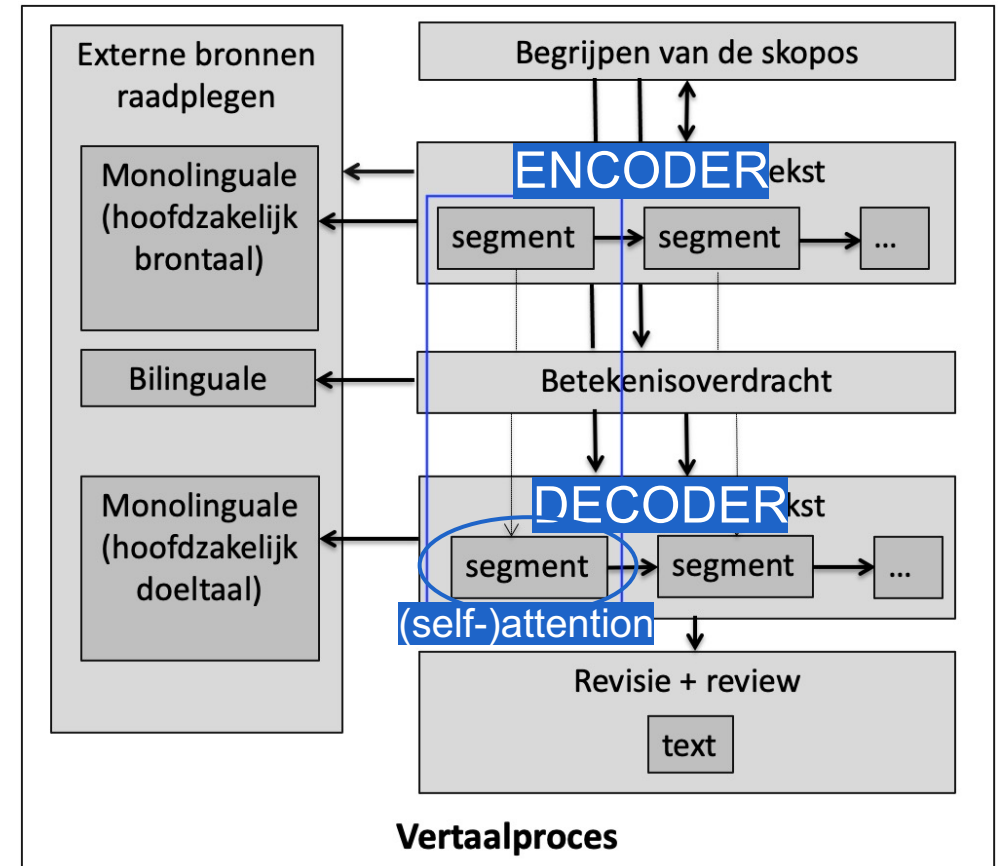
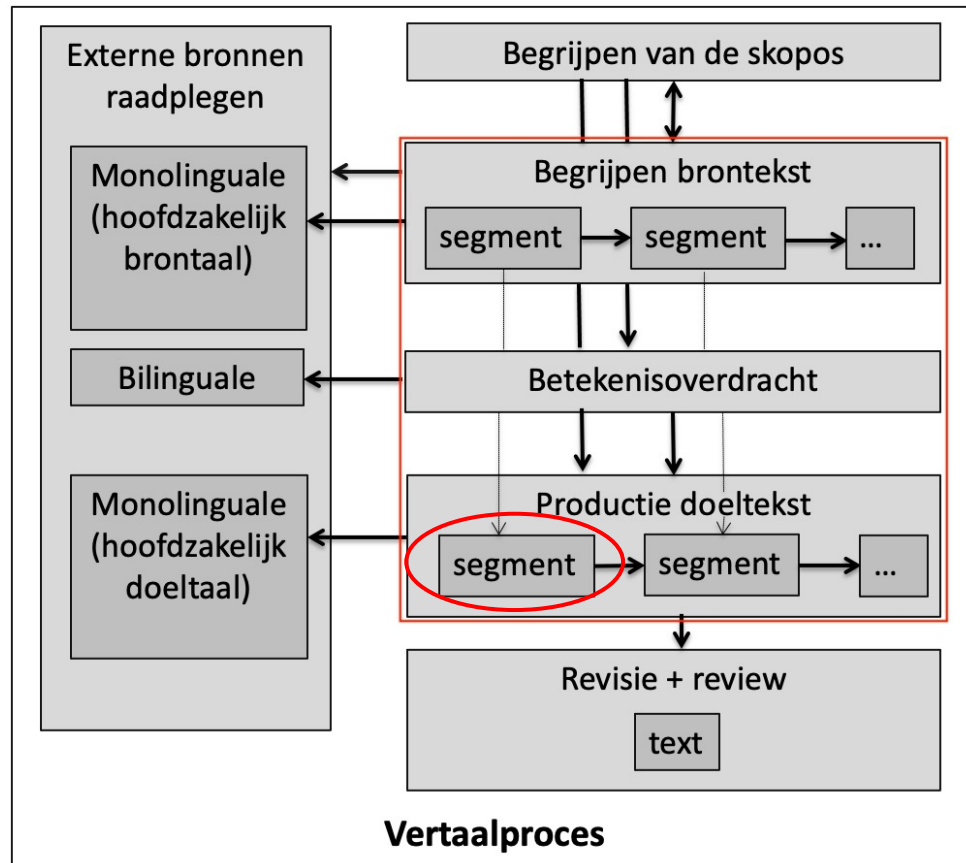
German→English

Ave.	Ave. z	System
82.6	0.228	VolcTrans
84.6	0.220	OPPO
82.2	0.186	HUMAN
81.5	0.179	Tohoku-AIP-NTT
81.3	0.179	Online-A
81.5	0.172	Online-G
79.8	0.171	PROMT-NMT
82.1	0.167	Online-B
78.5	0.131	UEDIN
78.8	0.085	Online-Z
74.2	-0.079	WMTBiomedBaseline
71.1	-0.106	zlabs-nlp
20.5	-1.618	yolo

English→German

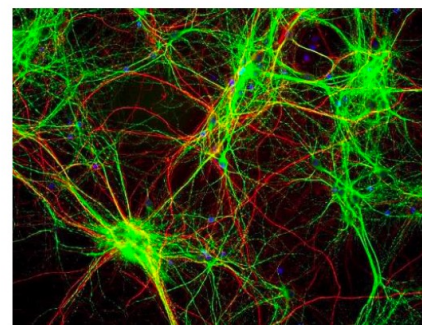
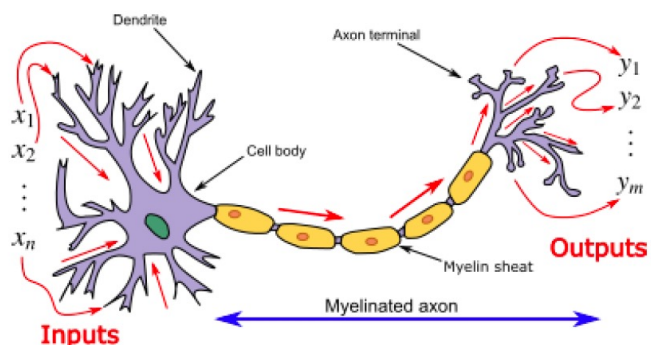
Ave.	Ave. z	System
90.5	0.569	HUMAN-B
87.4	0.495	OPPO
88.6	0.468	Tohoku-AIP-NTT
85.7	0.446	HUMAN-A
84.5	0.416	Online-B
84.3	0.385	Tencent-Translation
84.6	0.326	VolcTrans
85.3	0.322	Online-A
82.5	0.312	eTranslation
84.2	0.299	HUMAN-paraphrase
82.2	0.260	AFRL
81.0	0.251	UEDIN
79.3	0.247	PROMT-NMT
77.7	0.126	Online-Z
73.9	-0.120	Online-G
68.1	-0.278	zlabs-nlp
65.5	-0.338	WMTBiomedBaseline

VERTAALPROCES

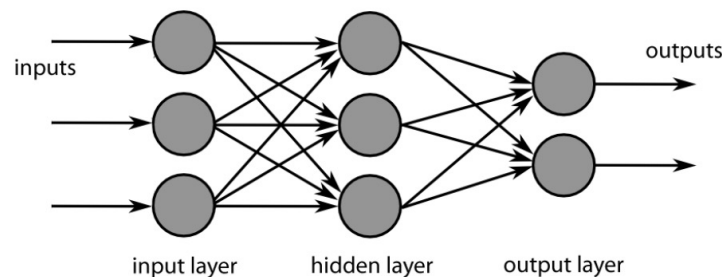
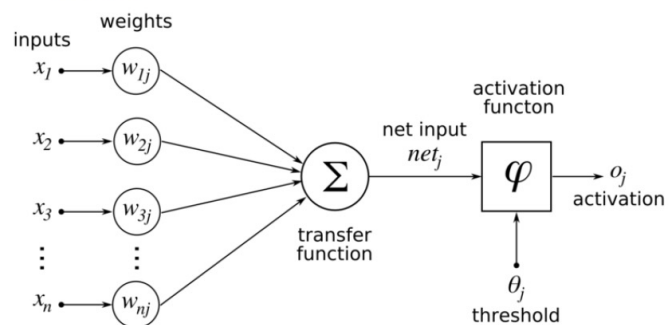


WAAROM IS DE KWALITEIT VAN NMT ZO GOED?

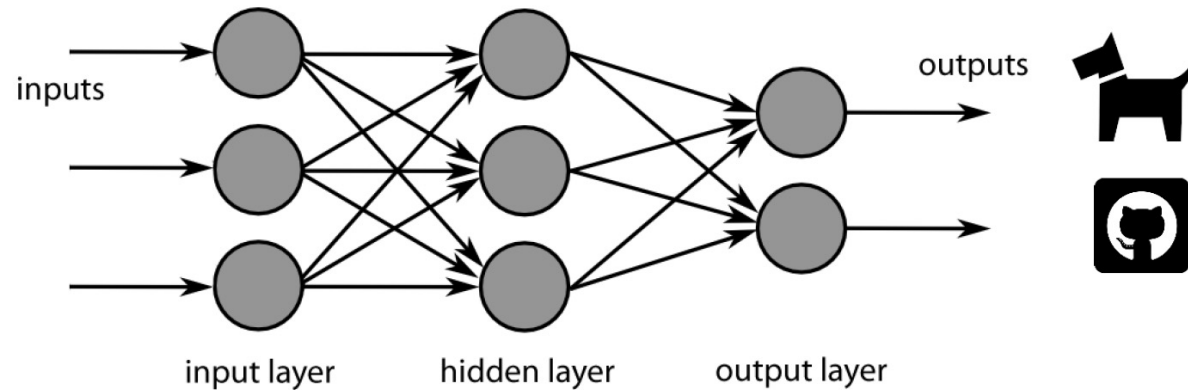
Artificiële neurale netwerken



Sources:
Wikimedia



ARTIFICIËLE NEURALE NETWERKEN


$$\begin{bmatrix} 7 \\ 22 \\ 4 \\ 112 \\ 34 \\ \vdots \\ 8 \end{bmatrix}$$


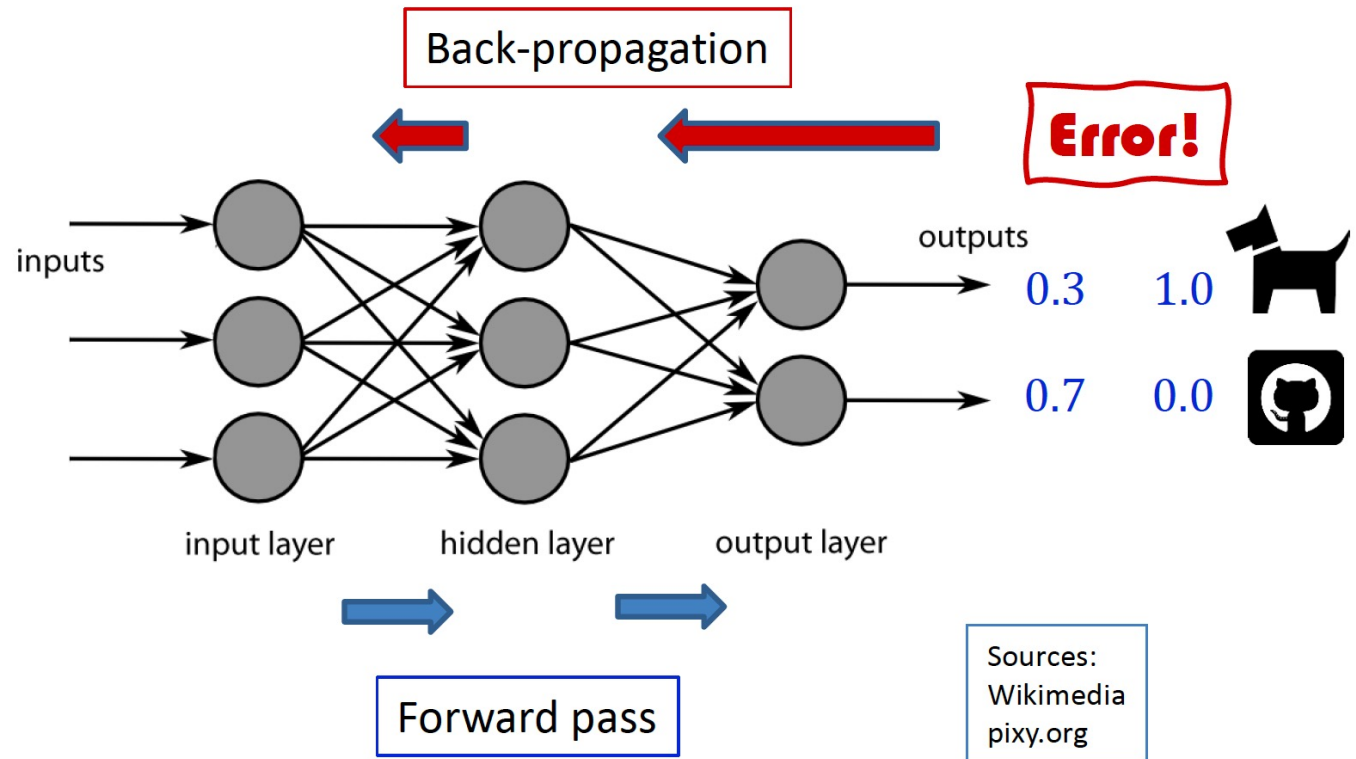
Sources:
Wikimedia
pixy.org

ARTIFICIËLE NEURALE NETWERKEN

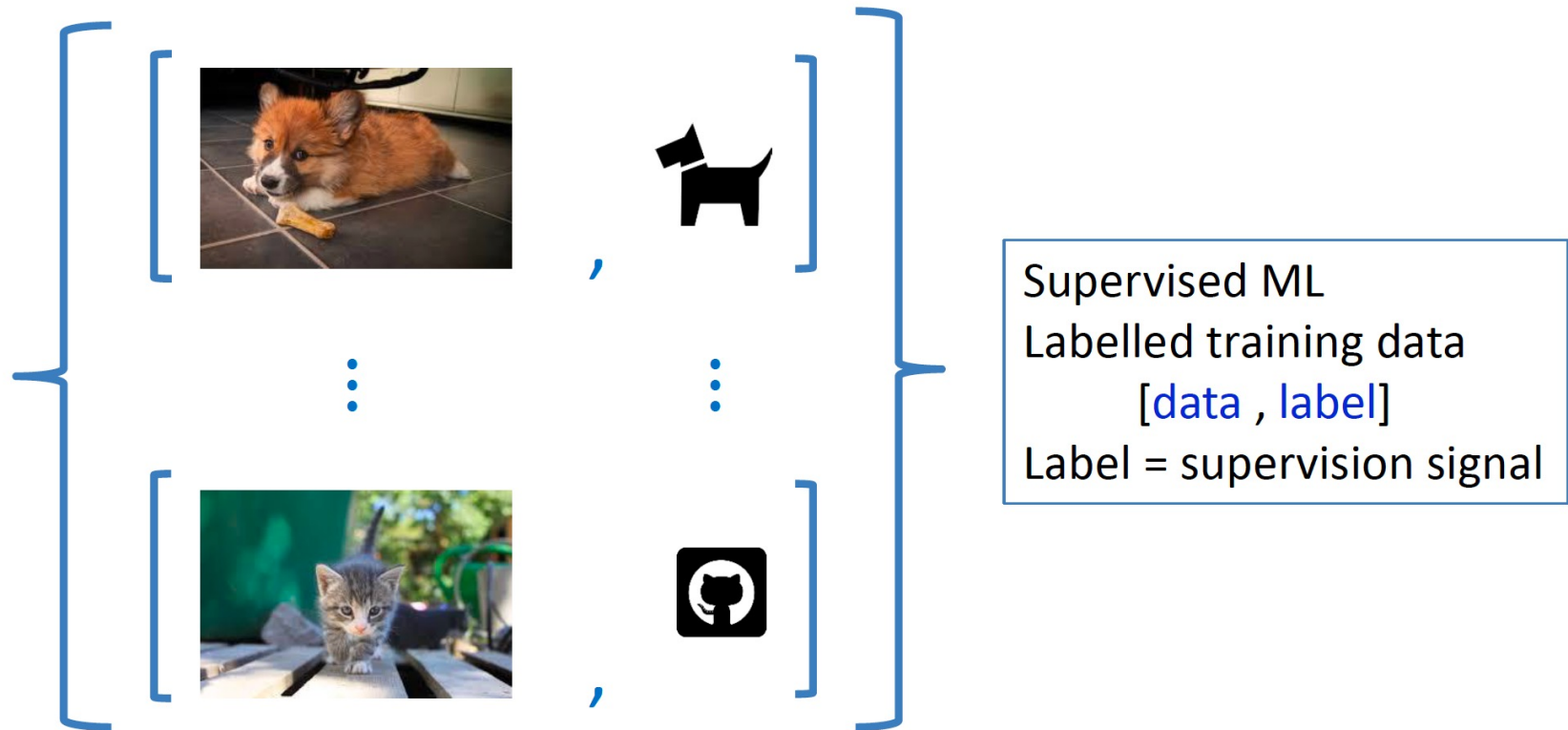


7
22
4
112
34
⋮
8

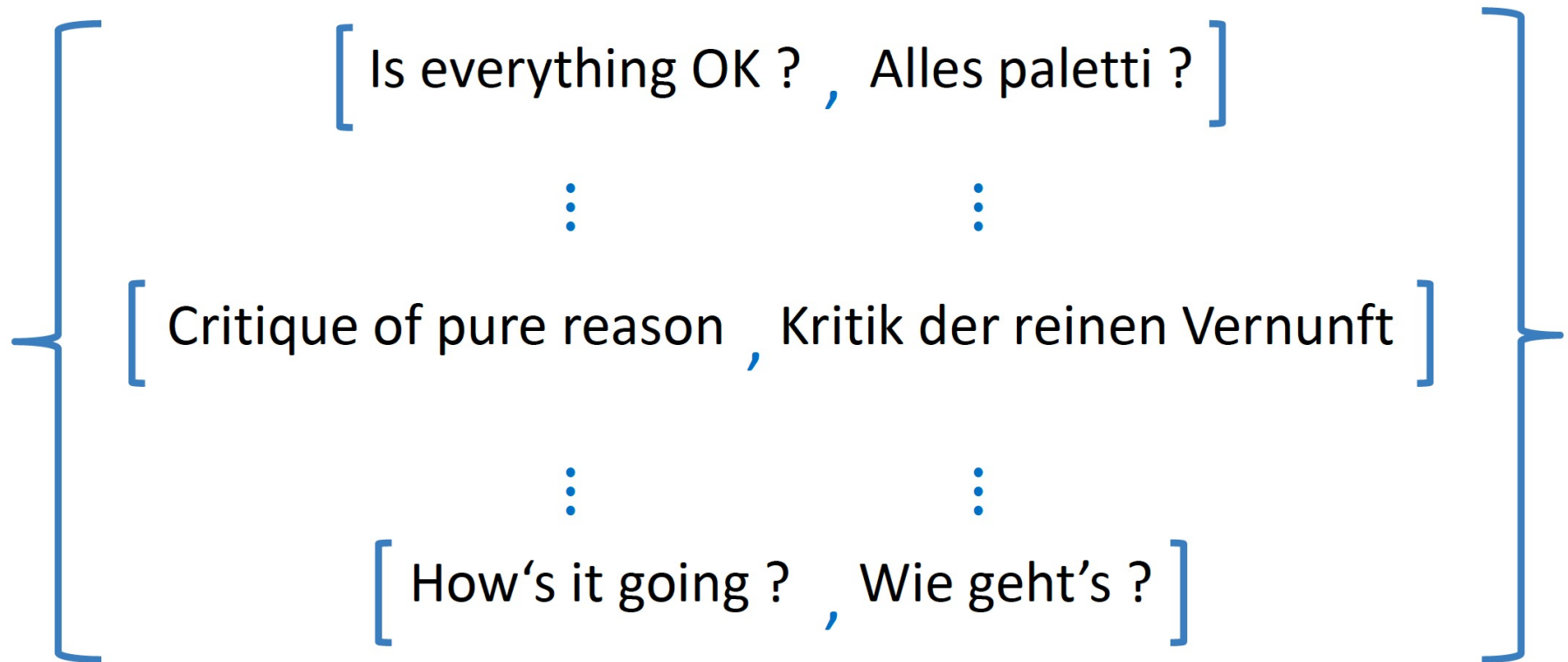
⋮
⋮



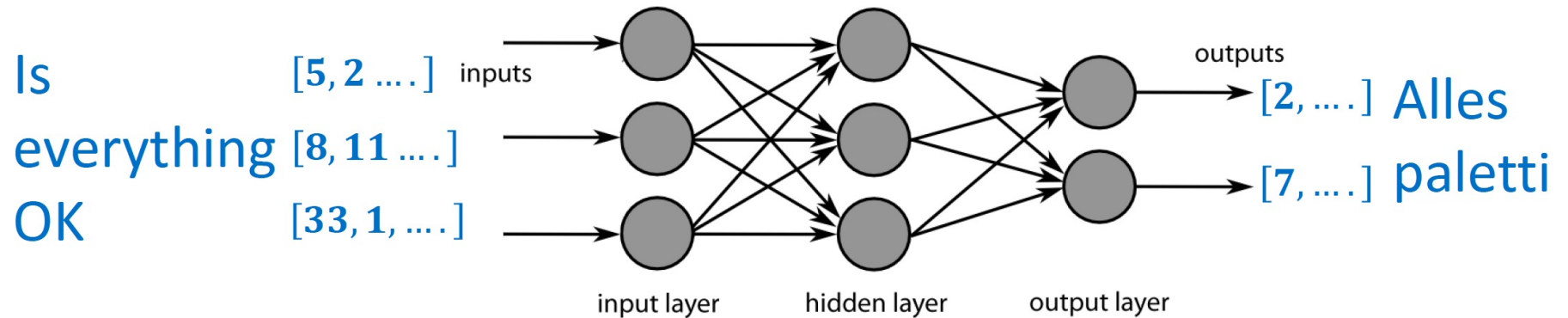
ARTIFICIËLE NEURALE NETWERKEN



ARTIFICIËLE NEURALE NETWERKEN



ARTIFICIËLE NEURALE NETWERKEN

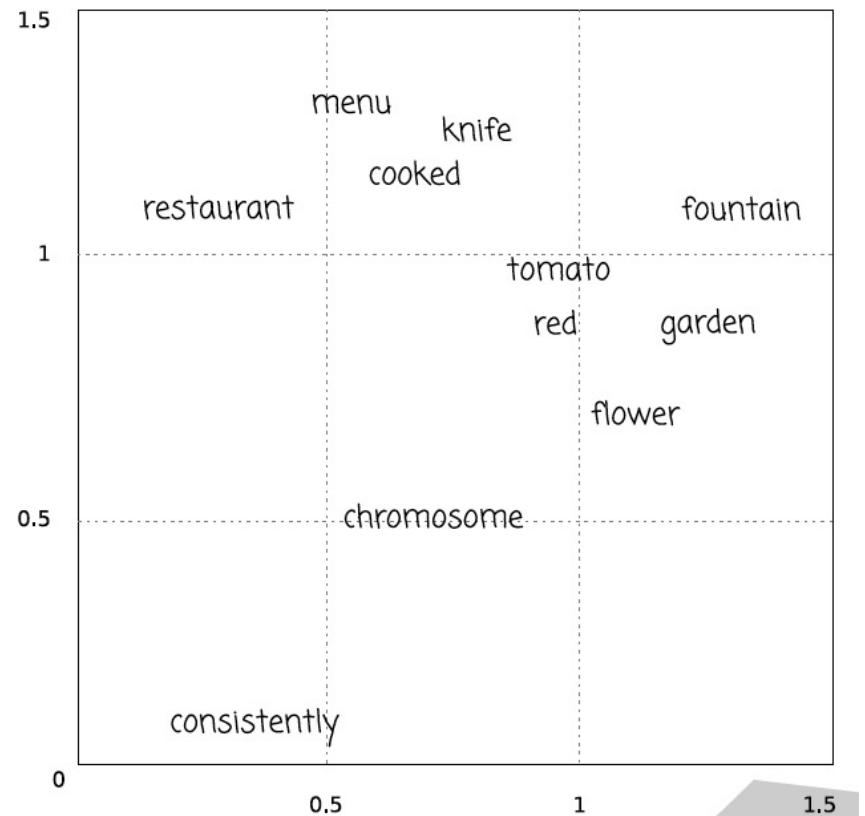


Sources:
Wikimedia

VAN WOORDEN NAAR VECTOREN

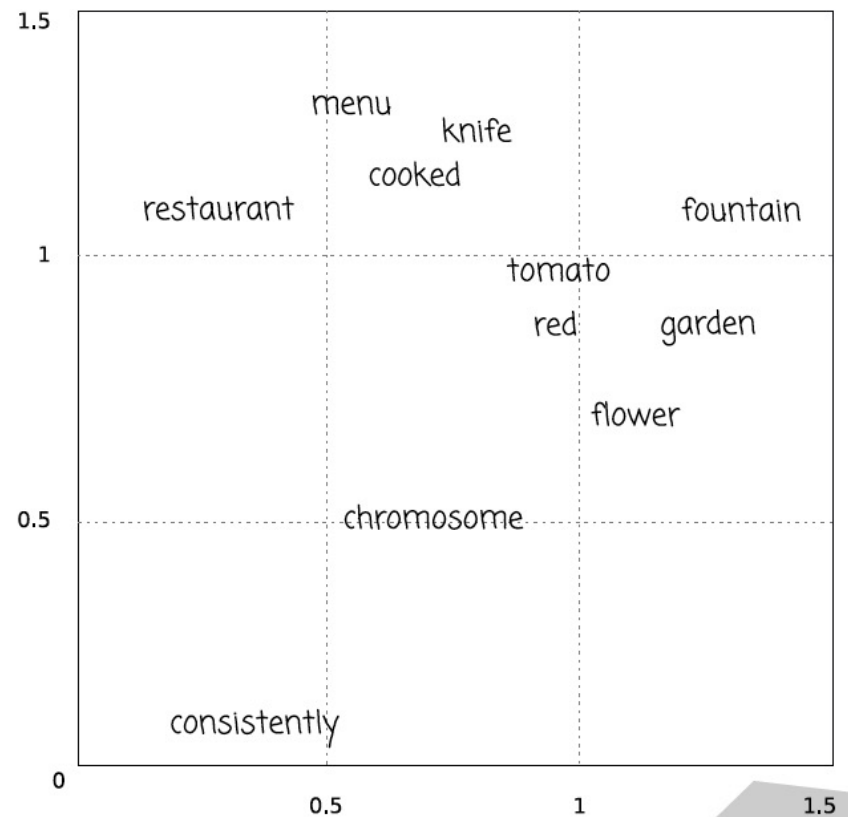
Word embeddings

restaurant	knife
red	menu
garden	cooked
fountain	chromosome
flower	consistently
tomato	



VAN WOORDEN NAAR VECTOREN

Word embeddings



restaurant = [0.25, 1.10]

menu = [0.60, 1.31]

...

VAN WOORDEN NAAR VECTOREN

Meer-dimensionele ruimte

poured = [0.25, 1.10, 0.32, ...]

rained = [0.27, 1.30, 0.31, ...]

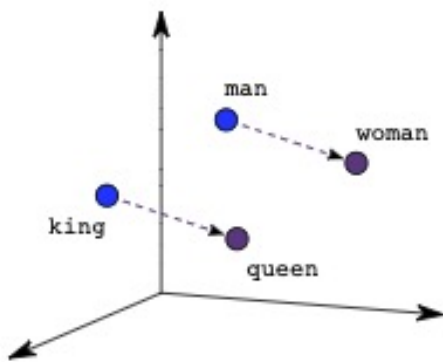
pouring = [0.25, 1.10, 0.12, ...]

raining = [0.27, 1.30, 0.11, ...]

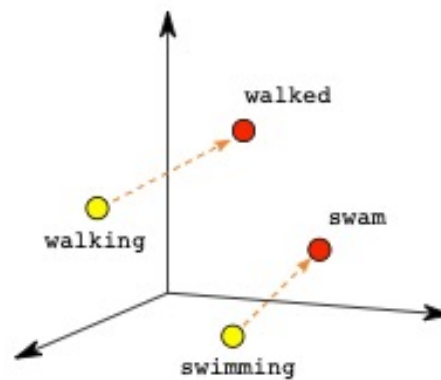
driving = [1.22, 0.89, 0.11, ...]

riding = [1.28, 0.83, 0.12, ...]

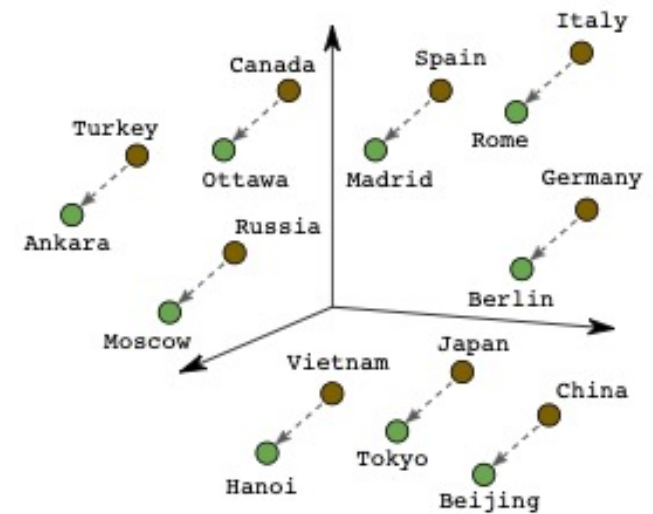
VAN WOORDEN NAAR VECTOREN



Male-Female



Verb Tense



Country-Capital

VAN WOORDEN NAAR VECTOREN

Word embeddings capteren 'betekenis'

Firth (1957) "You shall know a word by the company it keeps"

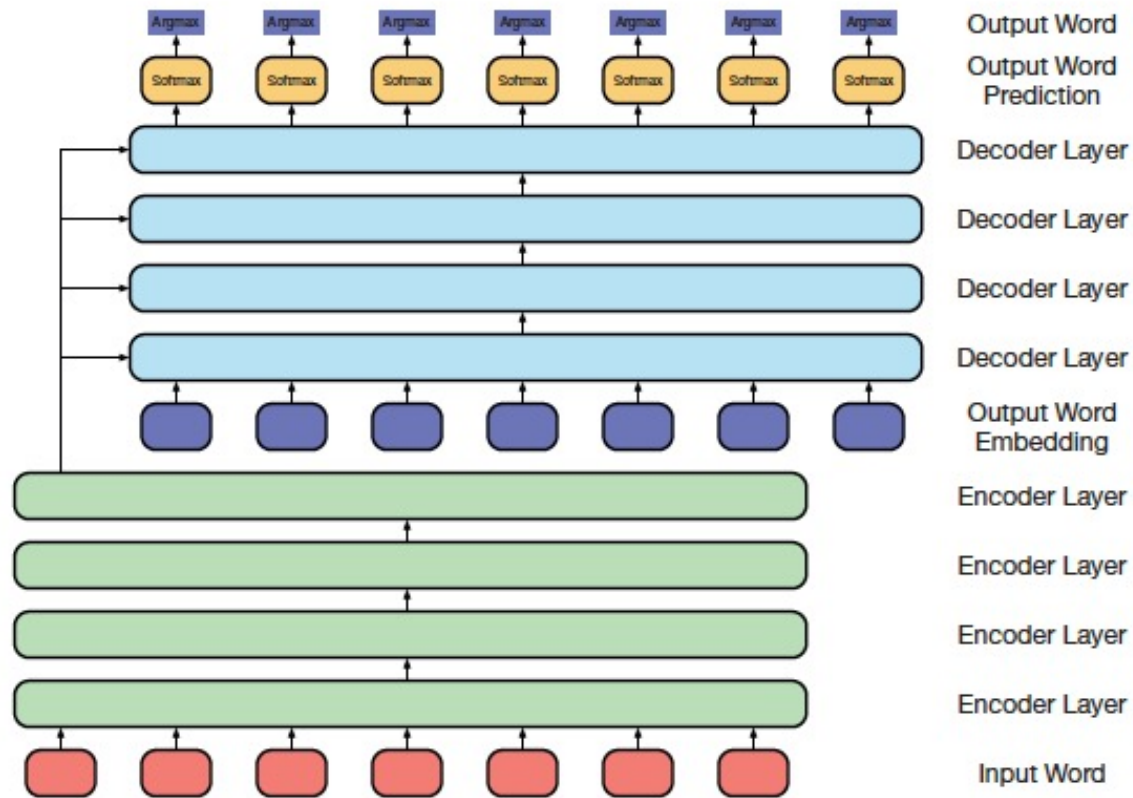
Word embeddings worden gebruikt als invoer in het NMT-systeem

Grootste voordeel → generaliseren

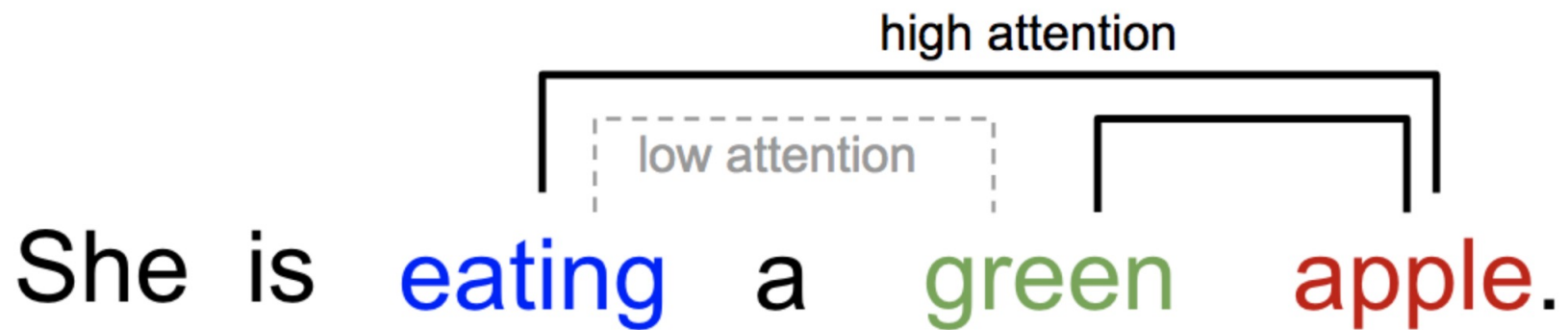
cat ≈ cats ≈ kitten

pouring ≈ raining ≈ driving ≈ riding

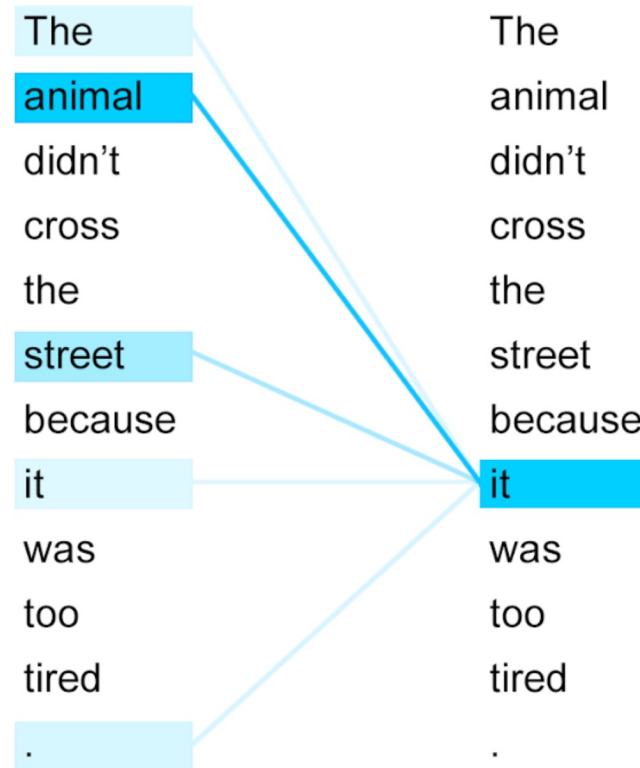
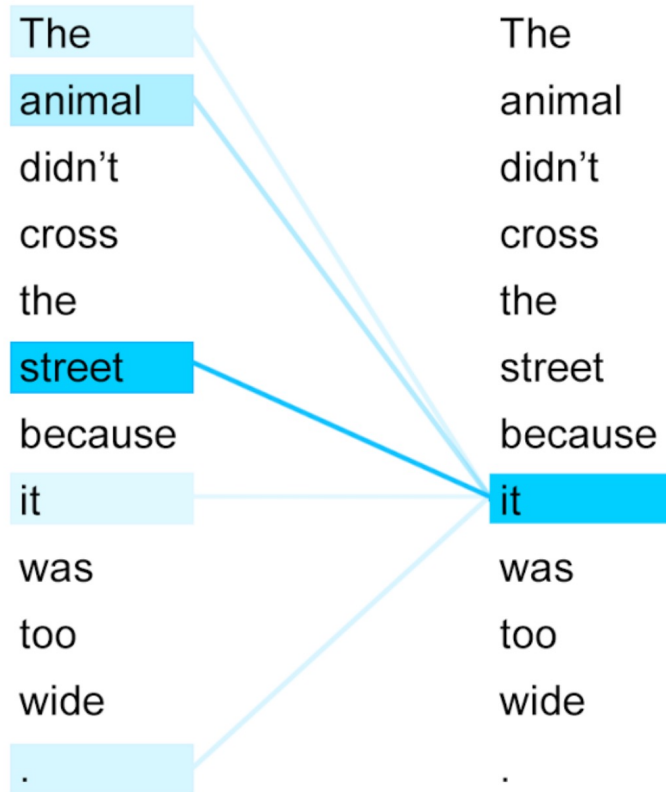
ENCODER-DECODER ARCHITECTUUR



ATTENTION



ATTENTION



Bron: <https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html>

ATTENTION

*The animal didn't cross the street because **it** was too tired.
L'animal n'a pas traversé la rue parce qu'**il** était trop fatigué.*

*The animal didn't cross the street because **it** was too wide.
L'animal n'a pas traversé la rue parce qu'**elle** était trop large.*

BELANG VAN DATA

A whopping 99% of our human DNA is the same as that of our closest primate cousins – chimps and bonobos. But while chimps tend to be male-led, bonobos take their lead from females.

eTranslation: trained on EC material; tested on news (2019)

99 % van ons DNA is dezelfde als die van onze nauwste primatencousins — Garnalen en „bonobos”. Hoewel pooiers meestal een man zijn, nemen Bonovbos hun lood af van vrouwelijke dieren.

eTranslation: trained on EC material; tested on news (2021)

Een bijna 99 % van ons menselijk DNA is hetzelfde als dat van onze naaste primate cousins — borstels en bonobos. Terwijl de chimps over het algemeen mrouwelijk zijn, nemen bonobos hun voorsprong bij vrouwtjes.

BELANG VAN DATA

A whopping 99% of our human DNA is the same as that of our closest primate cousins – chimps and bonobos. But while chimps tend to be male-led, bonobos take their lead from females.

eTranslation: trained on EC material; tested on news (2021)

Een bijna 99 % van ons menselijk DNA is hetzelfde als dat van onze naaste primate cousins — chimpansees en bonobos. Terwijl de chimps over het algemeen mannelijk zijn, nemen bonobos hun voorsprong bij vrouwtjes.

DeepL (2021)

Maar liefst 99% van ons menselijk DNA is hetzelfde als dat van onze naaste primate cousins - chimpansees en bonobo's. Maar terwijl chimpansees door mannen worden geleid, nemen bonobo's de leiding van de vrouwtjes.

BELANG VAN DATA



Tweet



Leunis Jacques
@JacquesLeunis



N'en déplaie à @GillesFavard les Belges ne vont pas tout de suite « rentrer chez eux manger des frites ». Et même si d'après @_johanmicoud « ils ont souvent le boulard » les 🇧🇪 eux ont dit non au racisme même en Russie et même si les 🇷🇺 ne voulaient pas #FINBEL @lequipedusoir



BELANG VAN DATA

N'en déplaie à @GillesFavard les Belges ne vont pas tout de suite « rentrer chez eux manger des frites ».

Heeft niet geklaagd bij @ GillesFavard dat de Belgen niet naar huis gingen om frites te eten.

106 / 2500



From French  To Dutch

[Advanced options](#)

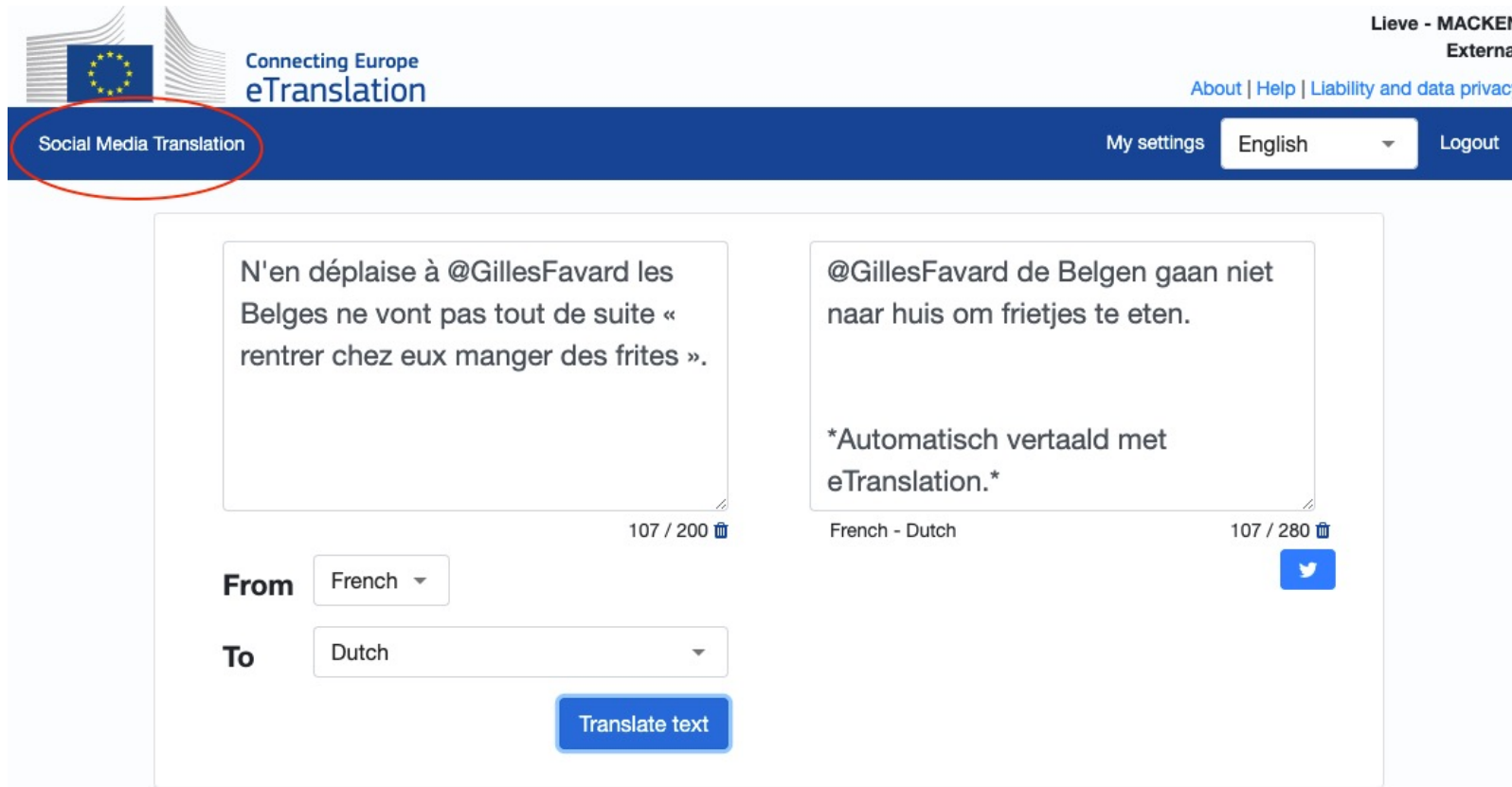
Domain EU Formal Language



E-mail me my translation

Translate text

BELANG VAN DATA



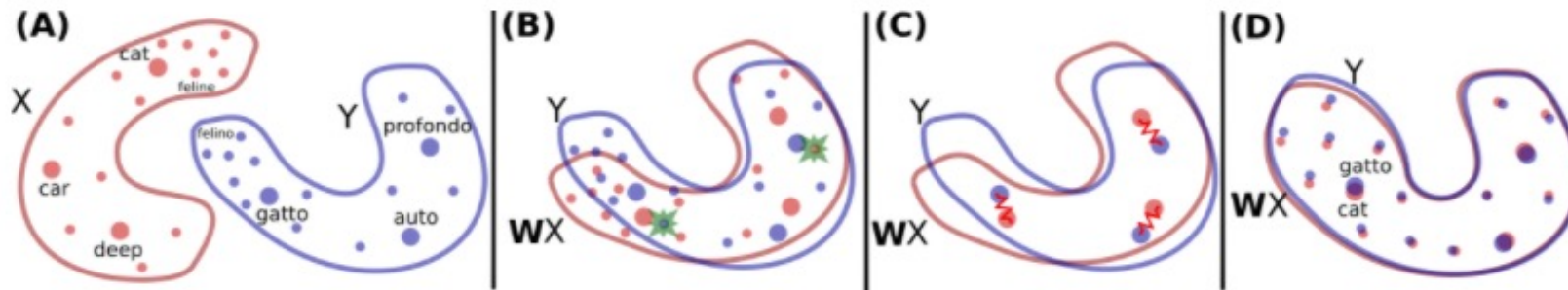
The screenshot displays the eTranslation web interface. At the top left, the logo for 'Connecting Europe eTranslation' is visible, featuring the European Union flag. To the right of the logo, the text 'Connecting Europe eTranslation' is displayed. Further right, the user's name 'Lieve - MACKEN' and status 'External' are shown. Below this, there are links for 'About | Help | Liability and data privacy'. A dark blue navigation bar contains the text 'Social Media Translation' (circled in red), 'My settings', a language dropdown menu set to 'English', and a 'Logout' button. The main content area shows a translation example. On the left, a text box contains the French text: 'N'en déplaie à @GillesFavard les Belges ne vont pas tout de suite « rentrer chez eux manger des frites ».' Below this text box, the character count '107 / 200' is shown. Underneath the text box, there are two dropdown menus: 'From' set to 'French' and 'To' set to 'Dutch'. A blue button labeled 'Translate text' is positioned below the dropdowns. On the right, a text box contains the Dutch translation: '@GillesFavard de Belgen gaan niet naar huis om frietjes te eten.' Below this text box, the character count '107 / 280' is shown. Underneath the text box, there is a blue Twitter icon and the text '*Automatisch vertaald met eTranslation.*'.

HET BELANG VAN DATA

- ~ 7000 talen
- Aantal MT-systemen
 $n \times (n-1) = 48,993,000$ MT-systemen
- Pivot-taal
 $(n-1) + (n-1) = 13,998$ MT-systemen
- Voldoende parallel trainingsmateriaal voor < 100 taalparen
Wat met de rest?

ONGESUPERVISEERDE MT

Veel meer monolinguale data in bron- en doeltaal beschikbaar dan parallel (vertaald) materiaal



Systeem voor woord-voor-woord-vertalingen

Werkt tot op zekere hoogte, maar kwaliteit daalt indien

- L1 & L2 verschillend scripts

- L1 en L2 data uit verschillende domeinen



MENS VS. MACHINE?

Masterproef Luca Desmet (2021) “An exploratory study of professional post-edits by English-Dutch DGT translators”

- 52% noodzakelijke aanpassingen

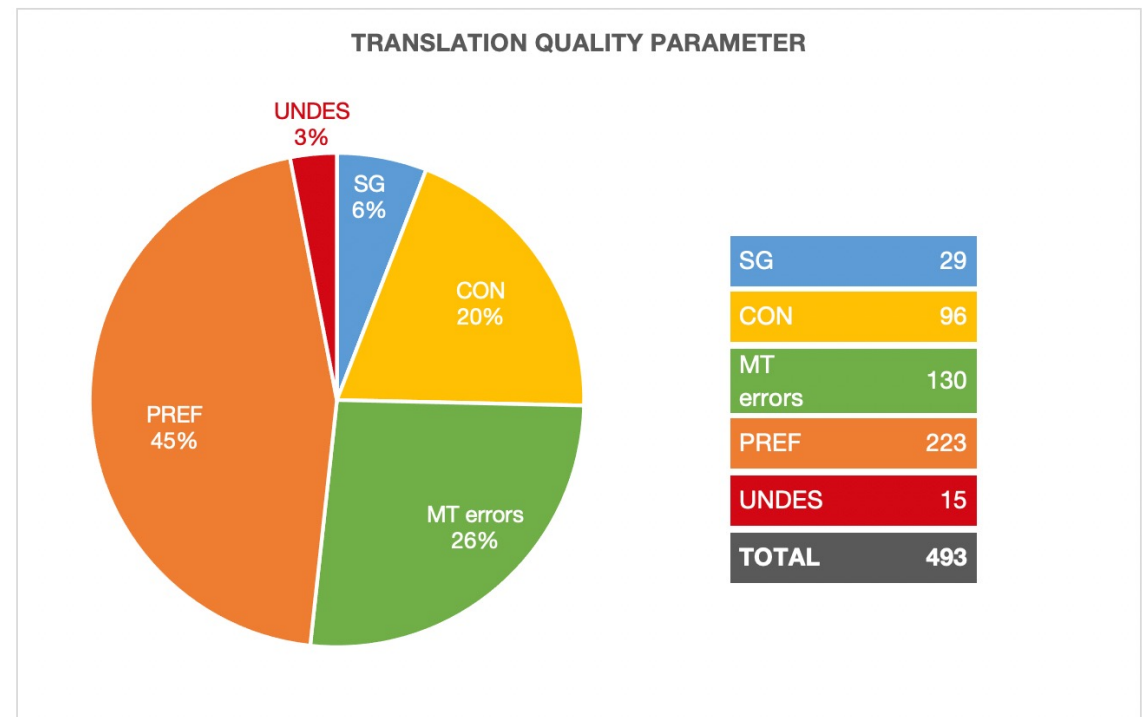
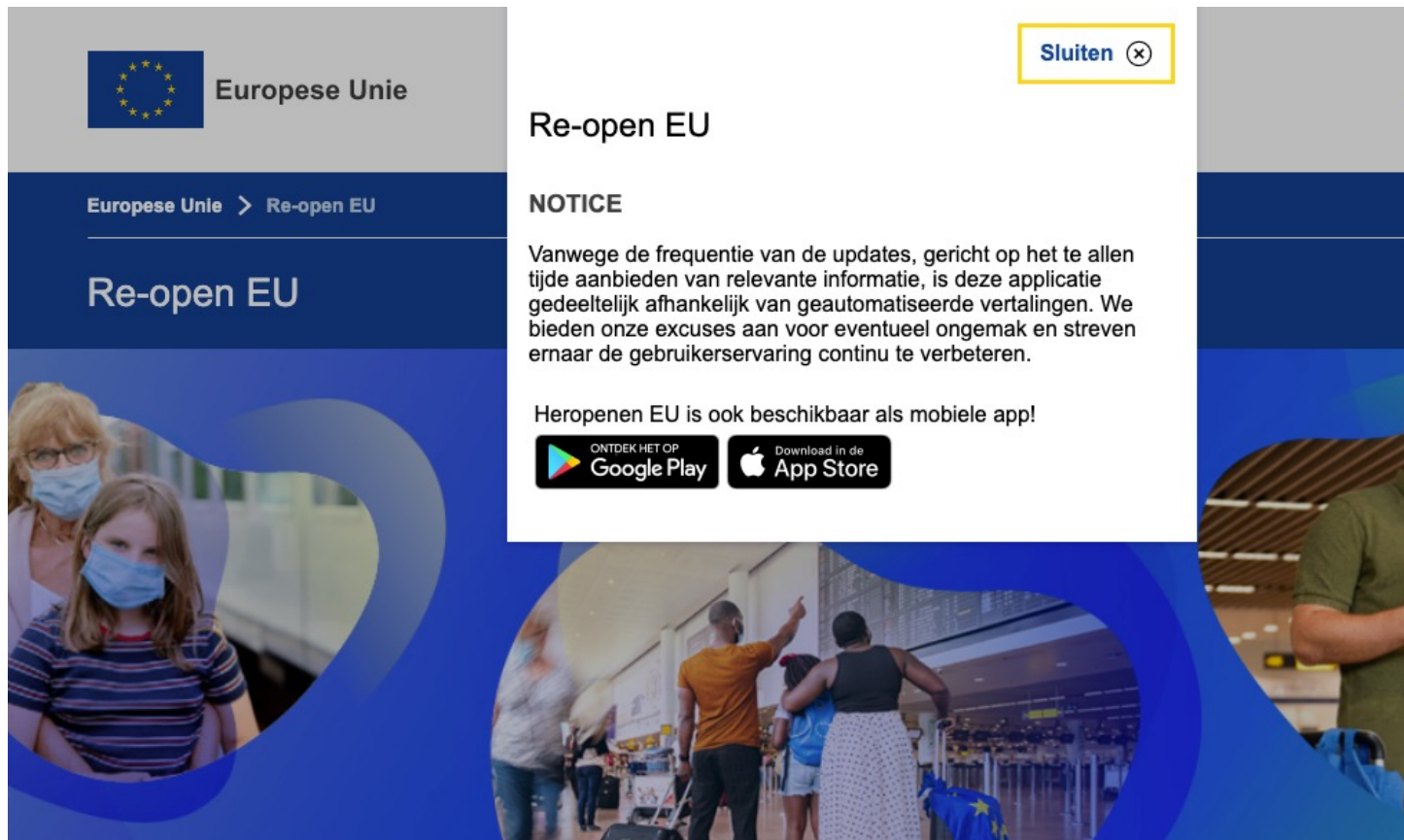


Figure 3. Distribution of the Translation Quality Parameter categories

MENS VS. MACHINE?





The image shows a screenshot of the 'Re-open EU' website. At the top left is the European Union flag and the text 'Europese Unie'. Below it, a breadcrumb trail reads 'Europese Unie > Re-open EU'. The main heading is 'Re-open EU'. The background features a blue circular graphic with images of people wearing masks and a person at an airport. A white notice box is overlaid on the right side, containing the following text:

Re-open EU

NOTICE

Vanwege de frequentie van de updates, gericht op het te allen tijde aanbieden van relevante informatie, is deze applicatie gedeeltelijk afhankelijk van geautomatiseerde vertalingen. We bieden onze excuses aan voor eventueel ongemak en streven ernaar de gebruikerservaring continu te verbeteren.

Heropenen EU is ook beschikbaar als mobiele app!

ONTDEK HET OP  **Google Play**  Download in de **App Store**

A yellow 'Sluiten' button with a close icon is located in the top right corner of the notice box.

MENS VS. MACHINE?

Are human translations always better than Machine Translations?

While generally raw machine translation is not as good as human translation, there are a number of cases where machine translation can actually be better than human translators.

In the LexisNexis Univentio case study, 13 million patents were translated from Japanese to English. Machine translated patents resulted in a 20% increase in meaningful search results when compared to human translations of patents. Machines translation delivered a significant and meaningful amount of added value.

Had the LexisNexis Univentio project been translated by human translators it would have taken 152,257 person years of effort and cost in excess of US\$ 40 billion. This project would have never been attempted due to time and cost requirements beyond thousands of times beyond the business value that such a task would deliver. With Language Studio™ this task was successfully

REFERENTIES

- Juan Antonio Pérez-Ortiz, Mikel L. Forcada and Felipe Sánchez-Martínez, “How neural machine translation works”, 2021.
Preprint version of book chapter (<https://multitrainmt.eu> project)
- Koehn, P. (2020). Neural Machine Translation. Cambridge: Cambridge University Press.
- Josef van Genabith, "Data or no Data, that is the Question: Learning MT without Translation Data?". 2021. Presentation at JIAMCATT (19.04.2021)

Lieve Macken

Docent vertaaltechnologie

VAKGROEP VERTALEN, TOLKEN EN
COMMUNICATIE

lieve.macken@ugent.be