Dutch LLMs

Challenges of training LLMs for mid-resource languages

Dr. ing. Pieter Delobelle - September 24, 2025

Pieter.ai

Pieter Delobelle

LLM Engineer 2024-2025
Aleph Alpha
R&D on translation, explainability, multi-node inference

Machine learning intern 2023
Apple
AurA project on LLM steering for toxicity reduction (ICML 2024)

PhD & Postdoc 2019-2024
KU Leuven (DTAI)
Fairer foundation models with Prof. De Raedt & Prof. Berendt

RobBERT, first author
Dutch BERT model pretrained in 2019, top 80 globally on Hugging Face

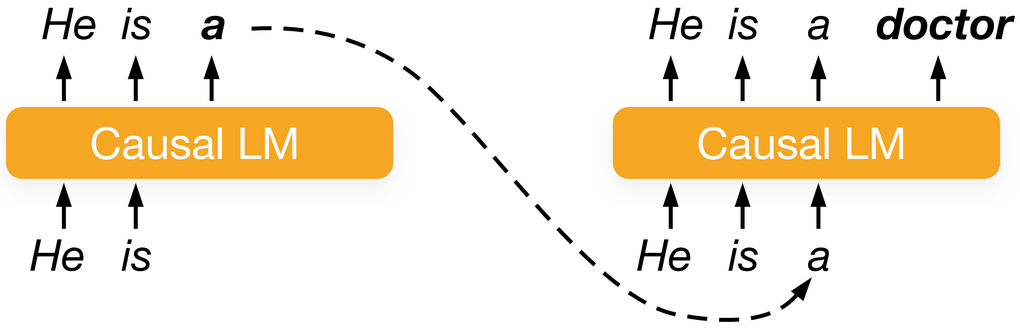
Trans-tokenization Research
Cross-lingual vocabulary transfer strategy (COLM 2024)

EU AI Office Expert
NLP expert for foundation models code of practice

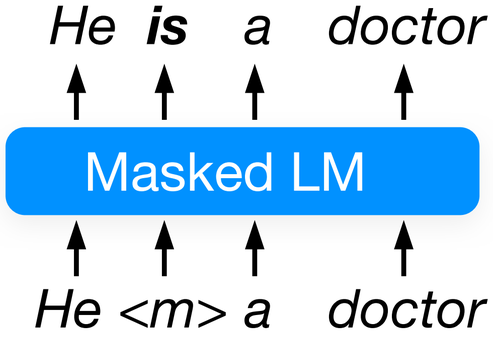
EU AI Office's Network of Evaluators Workshop, April 2025

Language modeling

1. Autoregressive language modeling

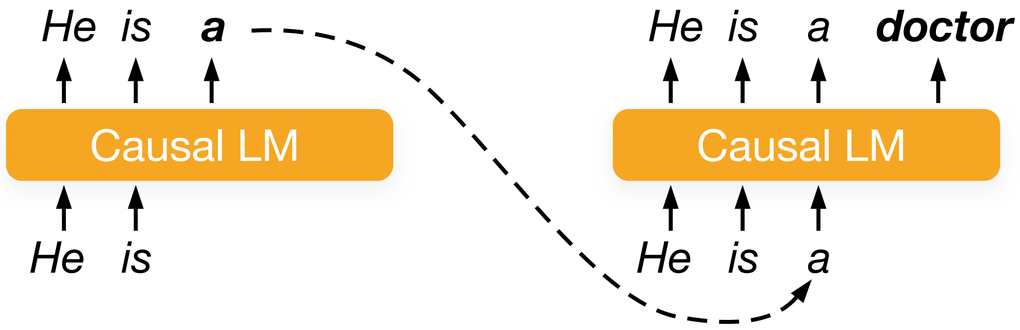


2. Masked language modeling

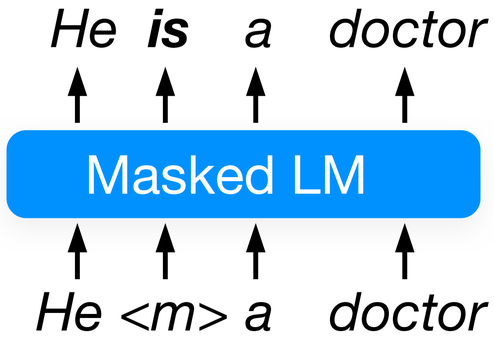


Language modeling

1. Autoregressive language modeling



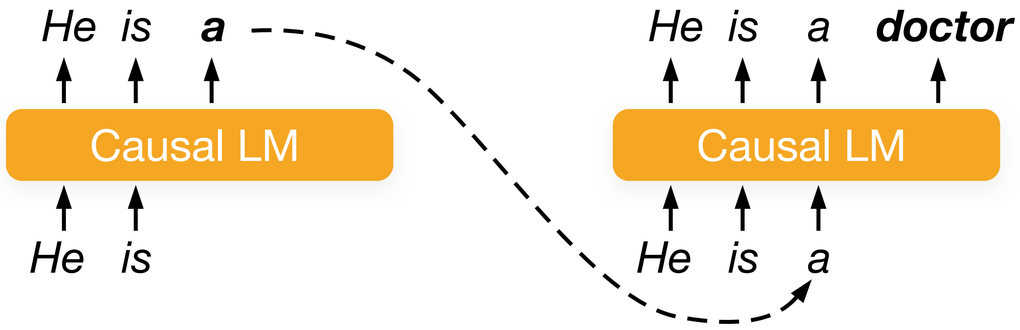
2. Masked language modeling



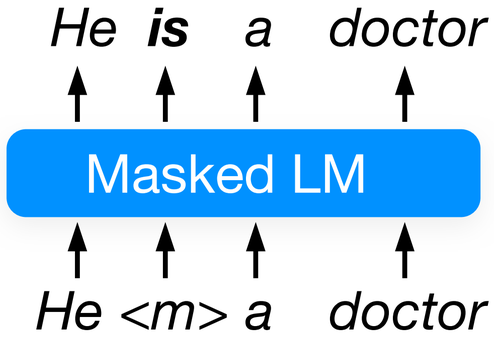
RobBERT

Language modeling

1. Autoregressive language modeling



2. Masked language modeling

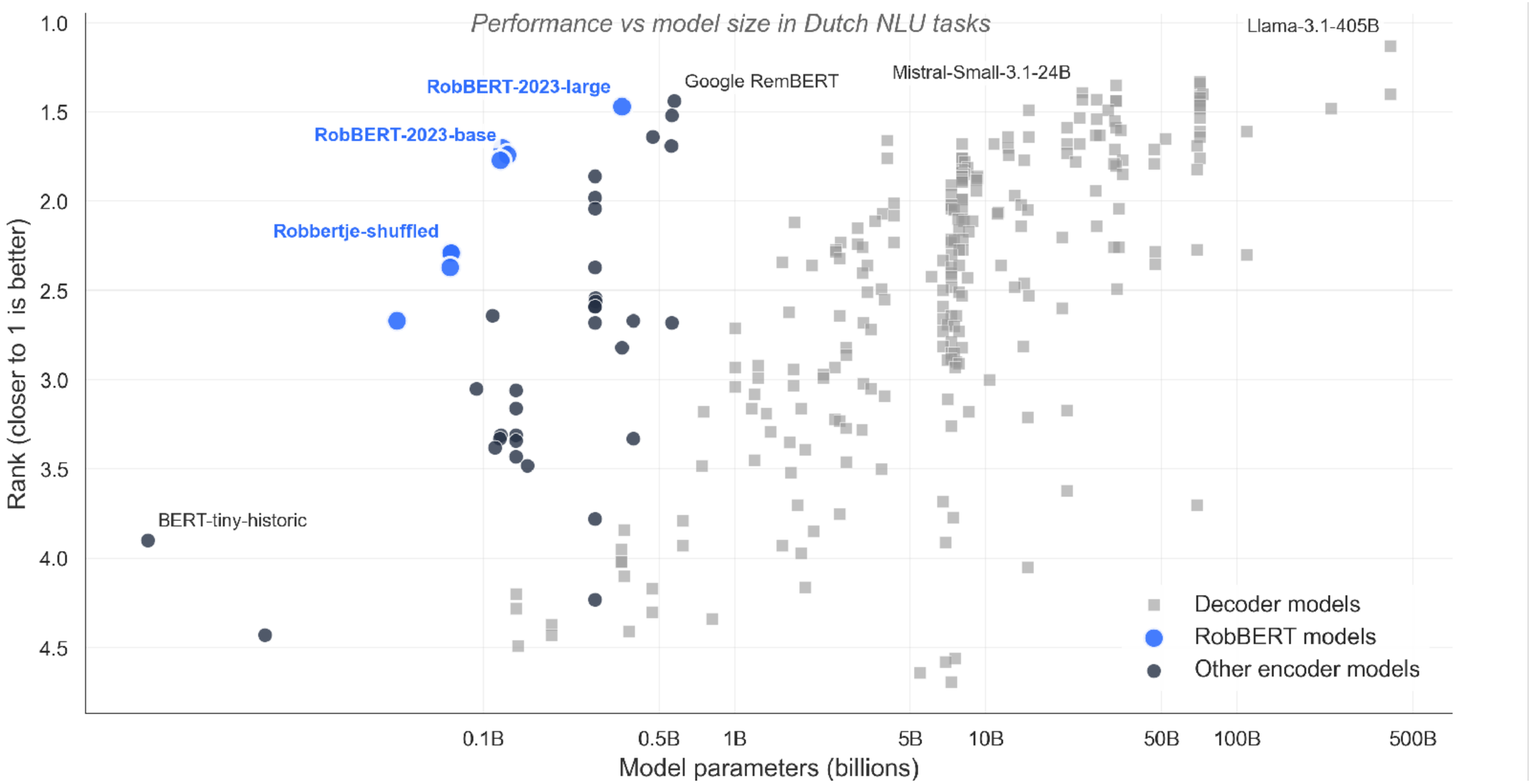


RobBERT

3M+ downloads

https://pieter.ai/robbert/

BERT-style models remain more efficient for Dutch NLU



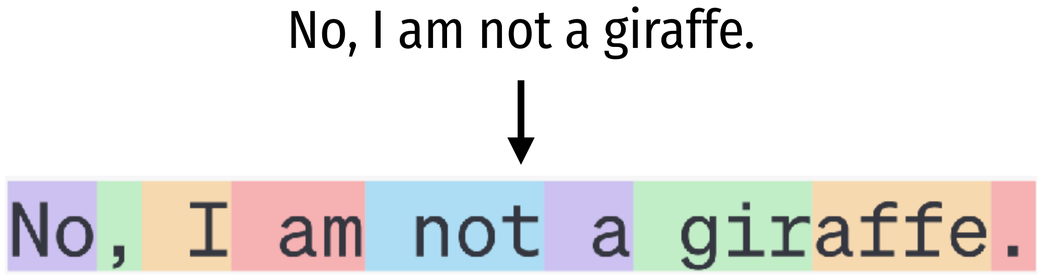
Tokenizing the training data

an example

No, I am not a giraffe.

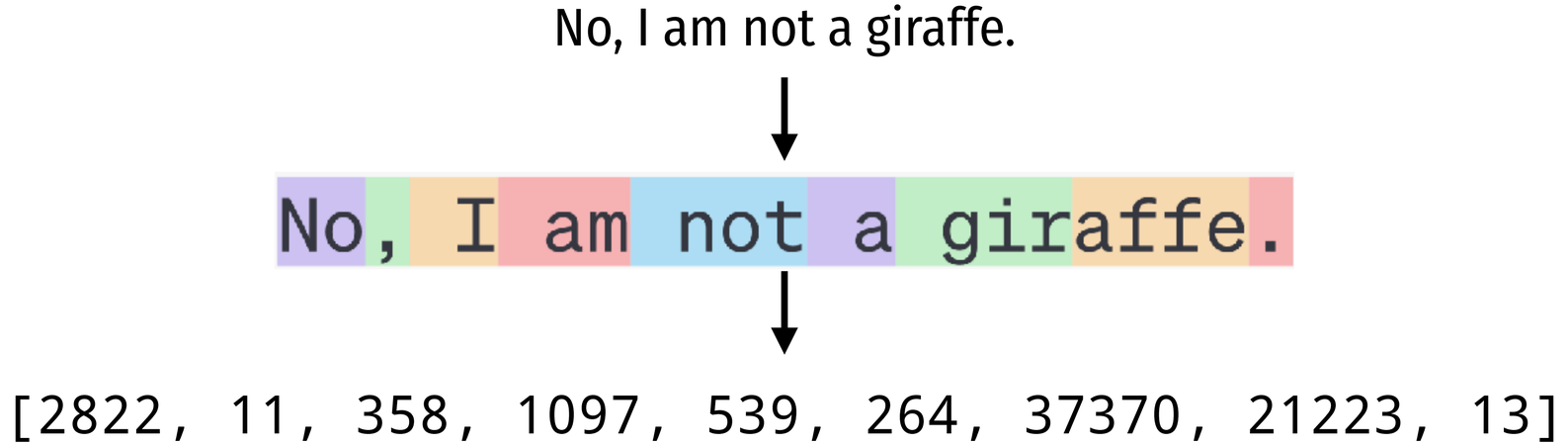
Tokenizing the training data

an example



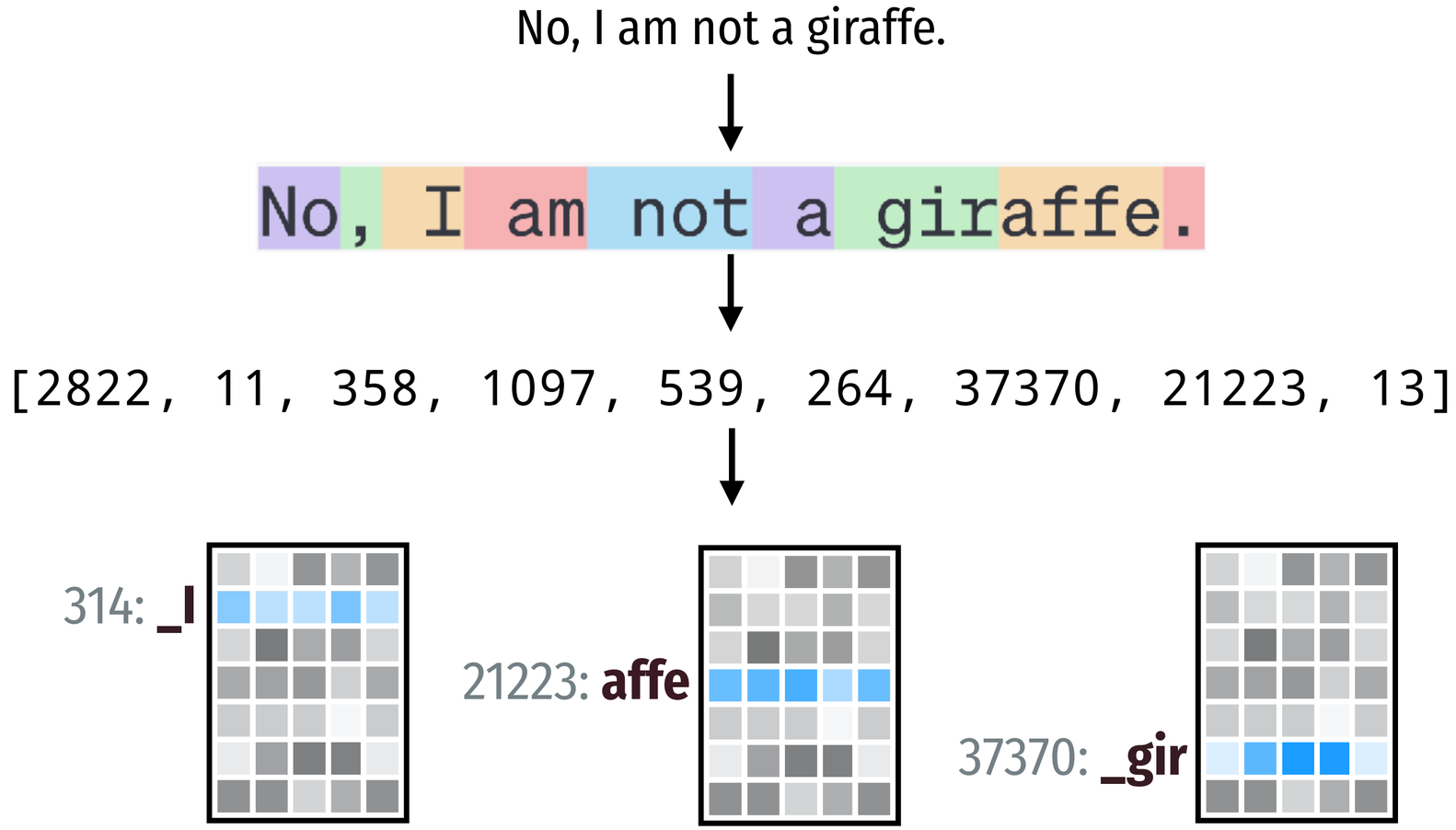
Tokenizing the training data

an example



Tokenizing the training data

an example

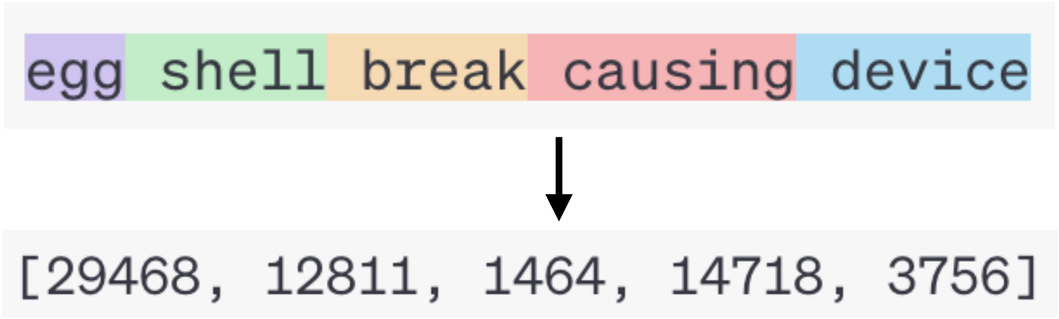


How does a tokenizer work?

Training LLMs beyond English

Tokenization

Translates a sequence of text to a sequence of numbers



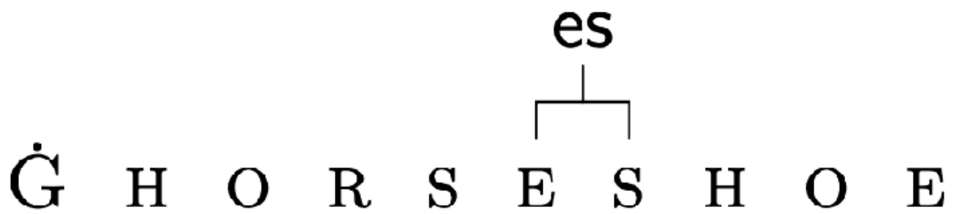
Typically not morphological, but preprocessing+occurrences:

- Splitting on spaces (including a separator _ or join ##)
- BPE or wordpiece for constructing a merge graph

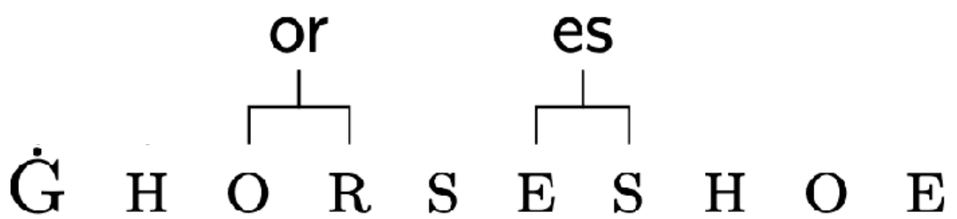
Example

Ġ H O R S E S H O E

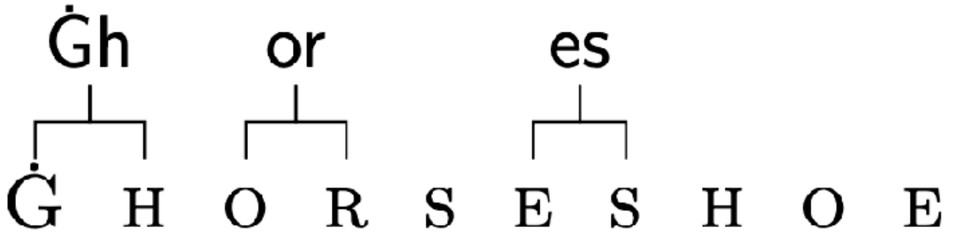
Example



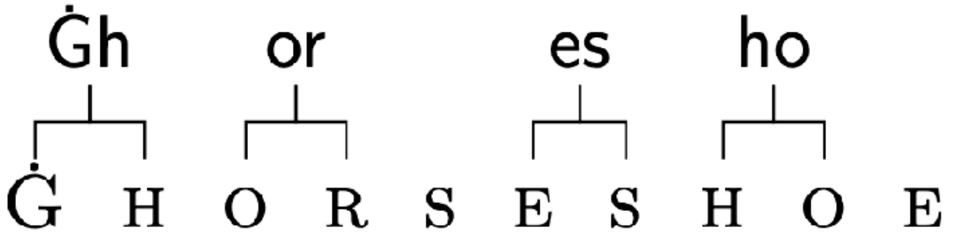
Example



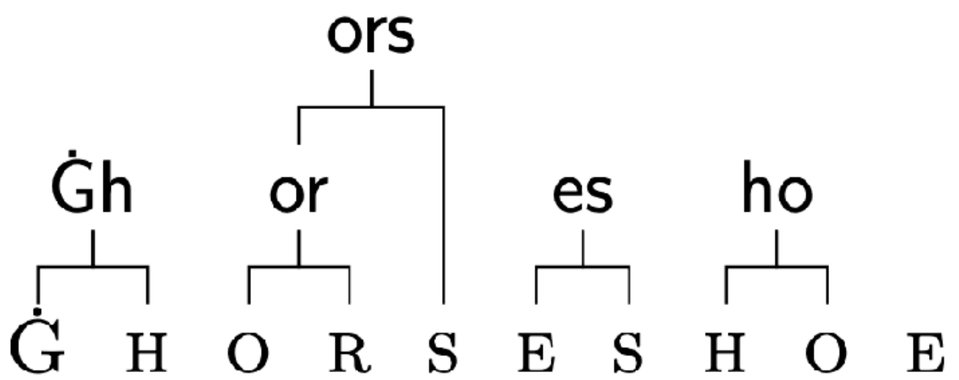
Example



Example

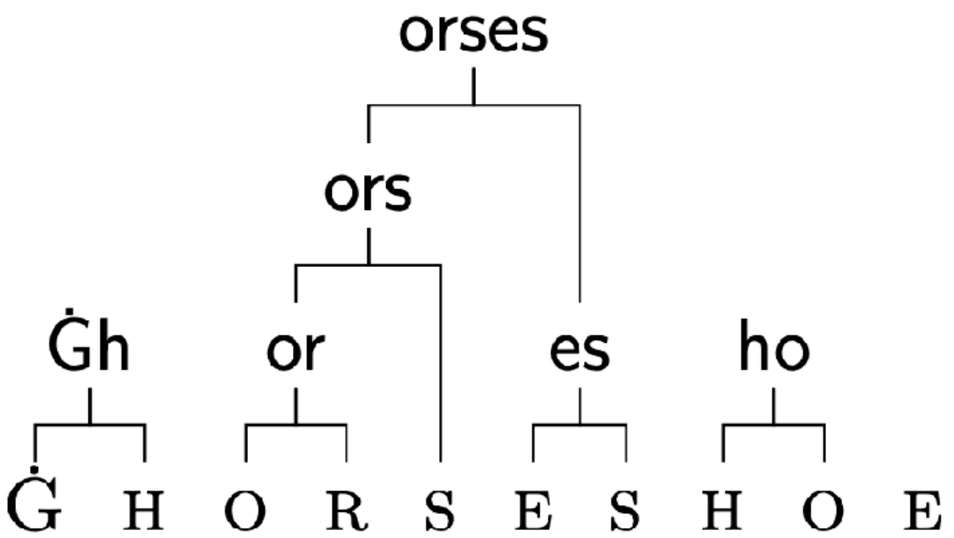


Example



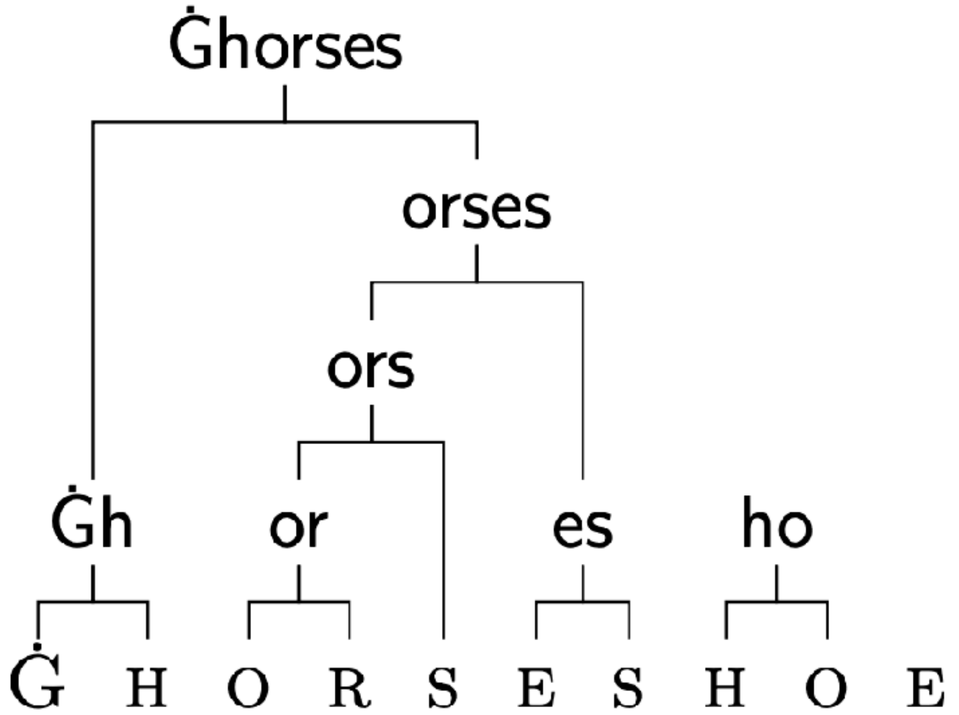
Slides credit of Thomas Bauwens: **BPE-Knockout**

Example

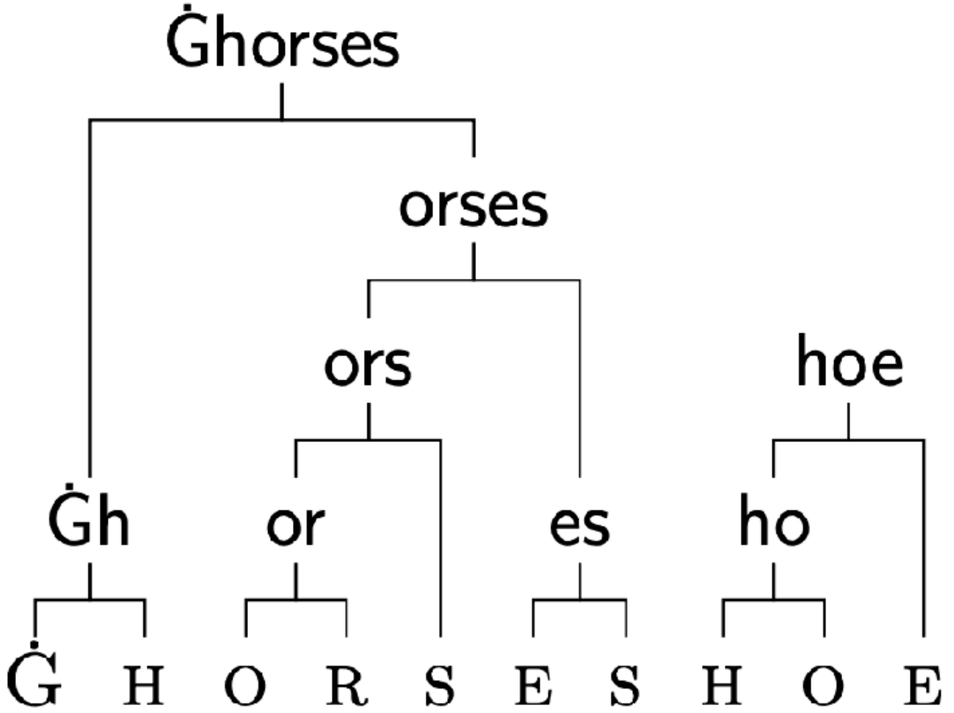


Slides credit of Thomas Bauwens: **BPE-Knockout**

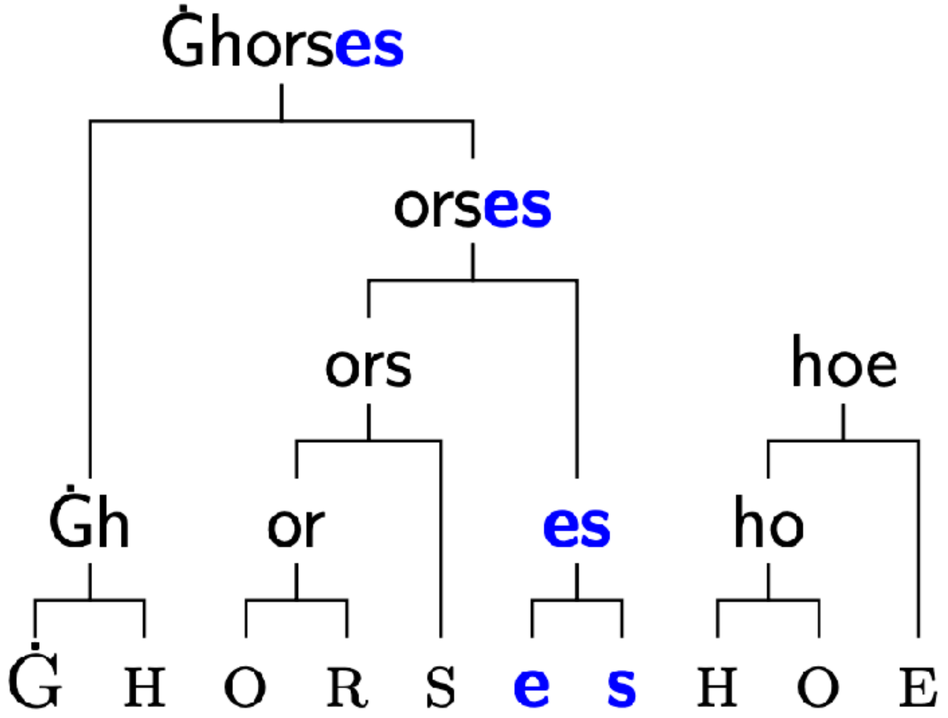
Example



Example

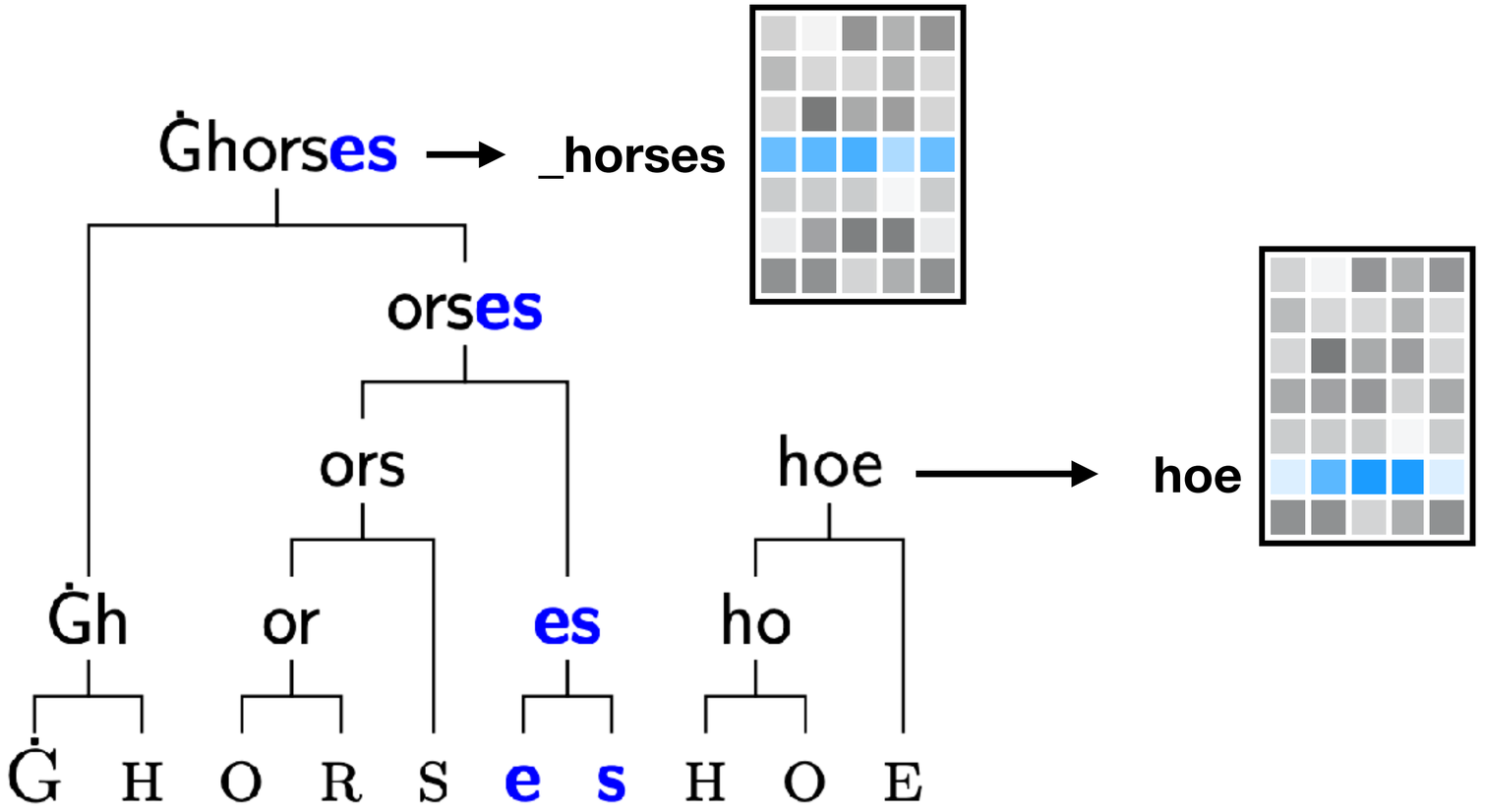


Problem: morphemes are merged together



https://pieter.ai/bpe-knockout/

Problem: morphemes are merged together



https://pieter.ai/bpe-knockout/

Morphology

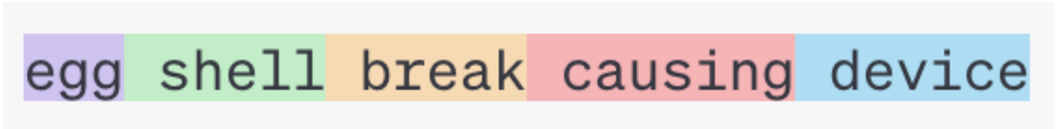
in English

```
egg shell break causing device
```

- ✔ splitting on words is logical
- ✔ most words end up being a token

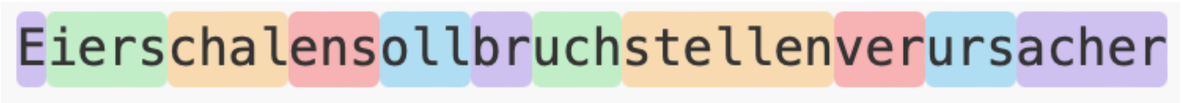
Morphology

in English



- ✔ splitting on words is logical
- ✔ most words end up being a token

in German



- ✖ doesn’t work for fusional and agglutinative languages
- ✖ splits are on unintuitive positions
- ✖ not following morphology
- ✖ shorter token types, so less meaning in embeddings

Fertility: the cost of poor tokenization

EN No, I am not a giraffe. That is an absurd thought.

DTAI

Fertility: the cost of poor tokenization

EN No, I am not a giraffe. That is an absurd thought.

DE Nein, ich bin keine Giraffe. Das ist ein absurder Gedanke.

NL Nee, ik ben geen giraf. Dat is een absurde gedachte.

Fertility: the cost of poor tokenization

EN `No, I am not a giraffe. That is an absurd thought.` → fertility = 1.09

DE `Nein, ich bin keine Giraffe. Das ist ein absurder Gedanke.` → fertility = 1.50

NL `Nee, ik ben geen giraf. Dat is een absurde gedachte.` → fertility = 1.50

DTAI

Fertility: the cost of poor tokenization

EN No, I am not a giraffe. That is an absurd thought. → fertility = 1.09

DE Nein, ich bin keine Giraffe. Das ist ein absurder Gedanke. → fertility = 1.50

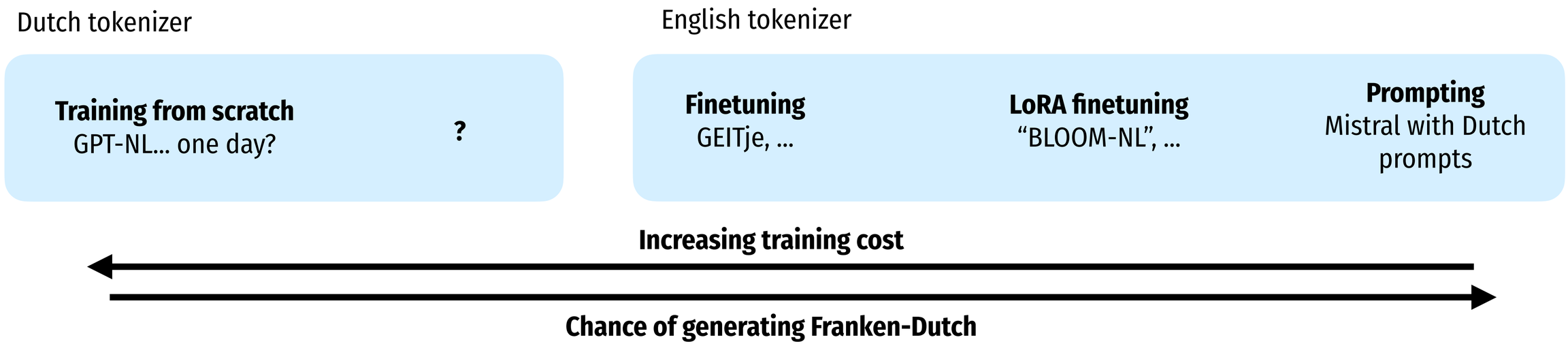
NL Nee, ik ben geen giraf. Dat is een absurde gedachte. → fertility = 1.50

Nee, ik ben geen giraf. Dat is een absurde gedachte. RobBERT's tokenizer → fertility = 1.20

DTAI

How does a tokenizer work?

Training LLMs beyond English


Dutch tokenizer
English tokenizer
Training from scratch
GPT-NL... one day?
?
Finetuning
GEITje, ...
LoRA finetuning
“BLOOM-NL”, ...
Prompting
Mistral with Dutch prompts
Increasing training cost
Chance of generating Franken-Dutch

Geitje-7b

First Dutch LLM

Geitje-7b

First Dutch LLM that got taken down by Brein

- Trained on ‘gigacorpus’
- A torrent with gigabytes of Dutch books
- Gigacorpus got taken down by Brein already

Ontwikkelaar haalt taalmodel GEITje offline na verzoek Stichting Brein - update

Het Nederlandse AI-taalmodel GEITje is offline gehaald op 'dringend verzoek' van Stichting Brein. GEITje zou volgens Brein deels getraind zijn op documenten uit de dienst Library Genesis, die afgelopen zomer is geblokkeerd.

Brein zegt dat het model is getraind met tienduizenden Nederlandstalige boeken die afkomstig zijn uit een illegale bron, namelijk Library Genesis, die afgelopen zomer op verzoek van Brein is geblokkeerd door Nederlandse accessproviders. De illegaal verkregen documenten en e-books waren waarschijnlijk terug te vinden in Gigacorpus, de dataset die afgelopen zomer door de maker zelf offline is gehaald. Gigacorpus bevatte naast boeken ook andere Nederlandstalige data, zoals wetsartikelen en uitspraken van Rechtspraak.nl.

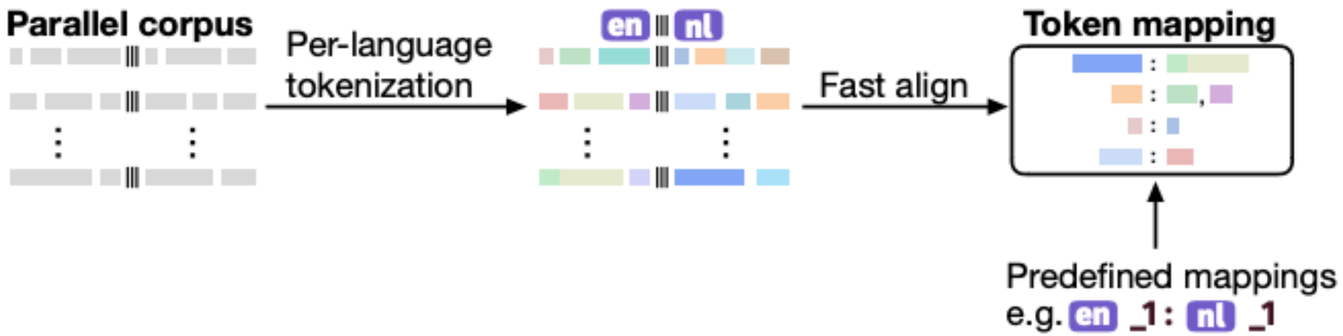
"Brein is niet tegen het trainen van AI, maar vindt wel dat de auteurs van al die muziek, boeken etc. daarvoor een eerlijke vergoeding moeten krijgen. Indien de oorspronkelijke makers niet willen dat hun materiaal voor het trainen van AI wordt gebruikt, dan moet dat ook gerespecteerd worden", schrijft de stichting.

De ontwikkelaar van GEITje verweerde dat tekstdatamining is toegestaan voor wetenschappelijke doeleinden en dat het model door wetenschappers wordt gebruikt, volgens Brein. De stichting wijst er echter op dat het model ook voor commercieel gebruik openbaar werd aangeboden op Huggingface.co. "De AI Act schrijft voor dat wetenschappers rechtmatig toegang moeten hebben tot materiaal om het te mogen gebruiken voor het trainen van AI. Dat is niet het geval als bij het trainen van een model gebruik is gemaakt van evident illegale bronnen", aldus Brein.

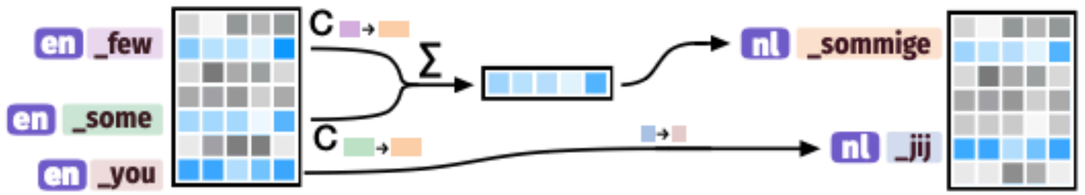
GEITje-maker Edwin Rijgersberg, op Tweakers bekend als E_Rijgersberg, bevestigt in een eigen post dat het taalmodel eind 2023 getraind is op gedeelten van het Nederlandse Gigacorpus. Brein heeft tegen Rijgersberg gezegd dat volgens de geldende wet- en regelgeving GEITje daarom offline gehaald moet worden.

https://tweakers.net/nieuws/231254/ontwikkelaar-haalt-taalmodel-geitje-offline-na-verzoek-stichting-brein.html

Trans-Tokenization: remapping the embeddings layer



(a) **Token alignment** is performed first based on a tokenized parallel corpus using a SMT-based alignment tool, to establish a probabilistic token mapping. We provide snippets of each stage of the full pipeline in Appendix D.



(b) **Embedding mapping** is then performed, as the embedding table for the target language (e.g. Dutch, indicated by nl) is initialized from the embeddings of mapped tokens in the source language (e.g. English, indicated by en), while preserving hidden layers.

RobBERT-2023: a converted Dutch MLM

Configuration			Benchmark scores				
Lang.	**Model**	**Params**	**NLI**	**SA**	**NER**	**POS**	**PPL**
🇳🇱	BERTje (de Vries et al., 2019)	109 M	83.9	93.0	88.3	96.3	33.8
🇳🇱	RobBERT (Delobelle et al., 2020)	116 M	84.2	94.4	89.1	96.4	13.1
🇫🇷➪🇳🇱	**Converted camembert-base**						
	Tik-to-Tok + full finetuning	116 M	85.3	95.8	84.9	94.4	12.4
🇩🇪➪🇳🇱	**Converted gbert-base**						
	Tik-to-Tok + full finetuning	116 M	85.5	95.0	86.3	95.3	10.2
🇺🇸➪🇳🇱	**`robbert-2023-dutch-base`**						
	Tik-to-Tok only (no LM head)	116 M	85.0	95.5	78.6	93.8	∞
	Tik-to-Tok + embeddings ft.	116 M	85.4	95.6	86.0	95.1	9.9
	Tik-to-Tok + full finetuning	116 M	86.6	95.4	87.6	95.8	5.9
🇺🇸➪🇳🇱	**`robbert-2023-dutch-large`**						
	Tik-to-Tok + full finetuning	345 M	**89.2**	**97.0**	**89.5**	96.0	**4.9**
🌍	XLM-RoBERTa large (XLM-R)	560 M	87.9	96.5	**89.5**	**96.9**	5.5

pieter.ai/robbert-2023; dumbench.nl

Tweeties: a series of monolingual LLMs

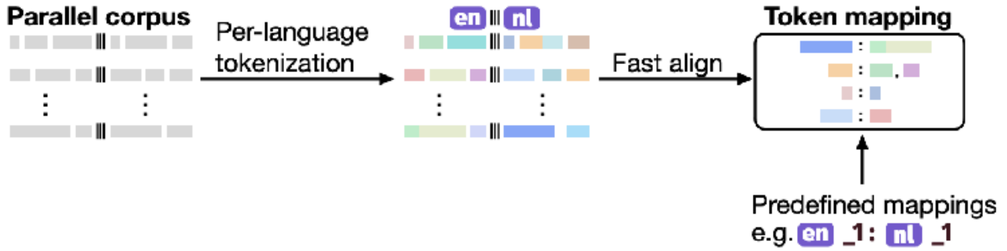
- Pretrained 7B models
- Monolingual LLMs for Dutch, Armenian, Italian and Tatar
- Mid-resource and low-resource languages

tweety-7b-dutch

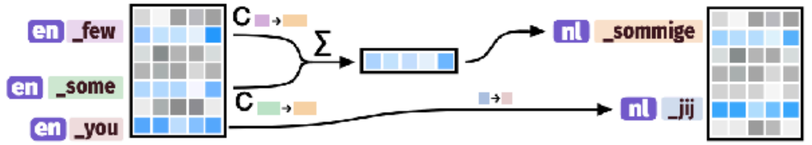
tweety-7b-tatar

tweety-7b-armenian

Tweeties: Trans-tokenized LLMs [preview only]



(a) **Token alignment** is performed first based on a tokenized parallel corpus using a SMT-based alignment tool, to establish a probabilistic token mapping. We provide snippets of each stage of the full pipeline in Appendix D.



(b) **Embedding mapping** is then performed, as the embedding table for the target language (e.g. Dutch, indicated by nl) is initialized from the embeddings of mapped tokens in the source language (e.g. English, indicated by en), while preserving hidden layers.

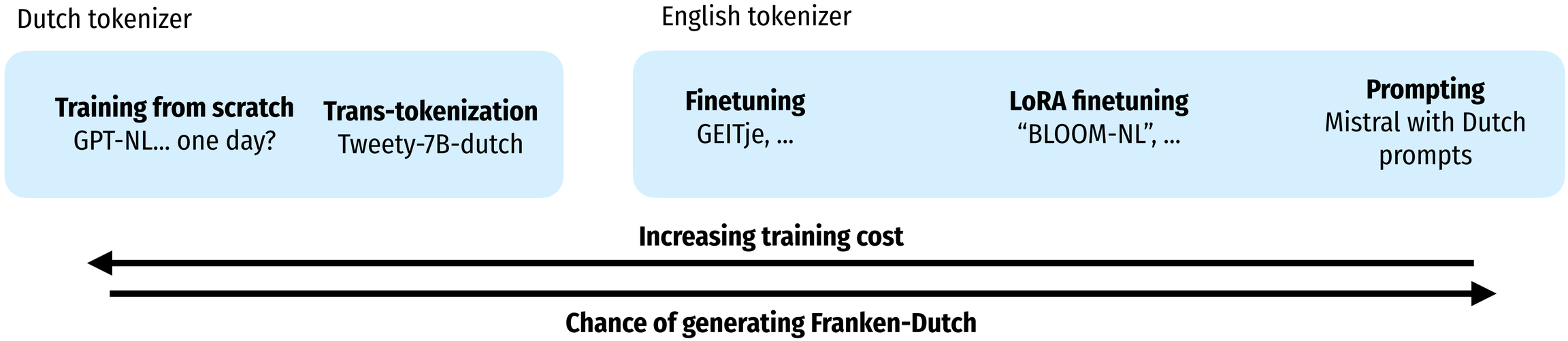
Figure 1: **Overview of our Trans-Tokenization method**

Because SMT-based alignment sometimes results in incorrect alignments, we discard any token alignment whose count is smaller than 10 (this can be increased for larger corpora). This ensures that the final mapping stays readable, avoiding a long tail of noisy mappings.

To deal with tokens whose mapping does not require real-world evidence (e.g. numbers, special characters, ...), we predefine a set of additional one-to-one mappings, which are

KU LEUVEN

pieter.ai/tweety-7b-dutch/


Dutch tokenizer
English tokenizer
Training from scratch
GPT-NL... one day?
Trans-tokenization
Tweety-7B-dutch
Finetuning
GEITje, ...
LoRA finetuning
“BLOOM-NL”, ...
Prompting
Mistral with Dutch prompts
Increasing training cost
Chance of generating Franken-Dutch

ChocoLlama

More effort to curate high-quality data

- OSCAR: 93 GB (28.6B tokens) - Common Crawl dump
- Open Subtitles: 5 GB (1.54B tokens)
- Wikipedia: 2.5 GB (769M tokens)
- Job Descriptions: 1.5 GB (462M tokens) - **TechWolf**
- Staatsblad: 1.4 GB (431M tokens) - **Bizzy**
- Project Gutenberg: 0.3 GB (92M tokens) - 970 books
- Legislation: 0.2 GB (62M tokens) - **ML6**

ChocoLlama

More effort to curate high-quality data

- OSCAR: 93 GB (28.6B tokens) - Common Crawl dump
- Open Subtitles: 5 GB (1.54B tokens)
- Wikipedia: 2.5 GB (769M tokens)
- Job Descriptions: 1.5 GB (462M tokens) - **TechWolf**
- Staatsblad: 1.4 GB (431M tokens) - **Bizzy**
- Project Gutenberg: 0.3 GB (92M tokens) - 970 books
- Legislation: 0.2 GB (62M tokens) - **ML6**

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	**0.48**	**0.66**	**0.49**	**0.49**	**0.53**
llama-3-8B-rebatch	0.44	0.64	0.46	0.48	0.51
llama-3-8B-instruct	0.47	0.59	0.47	0.52	0.51
llama-3-8B	0.44	0.64	0.47	0.45	0.5
Reynaerde-7B-Chat	0.44	0.62	0.39	0.52	0.49
Llama-3-ChocoLlama-base	**0.45**	**0.64**	**0.44**	**0.44**	**0.49**
zephyr-7b-beta	0.43	0.58	0.43	0.53	0.49
geitje-7b-ultra	0.40	0.66	0.36	0.49	0.48
ChocoLlama-2-7B-tokentrans-instruct	**0.45**	**0.62**	**0.34**	**0.42**	**0.46**
mistral-7b-v0.1	0.43	0.58	0.37	0.45	0.46
ChocoLlama-2-7B-tokentrans-base	**0.42**	**0.61**	**0.32**	**0.43**	**0.45**
ChocoLlama-2-7B-instruct	**0.36**	**0.57**	**0.33**	**0.45**	**0.43
ChocoLlama-2-7B-base	**0.35**	**0.56**	**0.31**	**0.43**	**0.41**
llama-2-7b-chat-hf	0.36	0.49	0.33	0.44	0.41
llama-2-7b-hf	0.36	0.51	0.32	0.41	0.40

ChocoLlama

More effort to curate high-quality data

- OSCAR: 93 GB (28.6B tokens) - Common Crawl dump
- Open Subtitles: 5 GB (1.54B tokens)
- Wikipedia: 2.5 GB (769M tokens)
- Job Descriptions: 1.5 GB (462M tokens) - **TechWolf**
- Staatsblad: 1.4 GB (431M tokens) - **Bizzy**
- Project Gutenberg: 0.3 GB (92M tokens) - 970 books
- Legislation: 0.2 GB (62M tokens) - **ML6**

Model	ARC	HellaSwag	MMLU	TruthfulQA	Avg.
Llama-3-ChocoLlama-instruct	**0.48**	**0.66**	**0.49**	**0.49**	**0.53**
llama-3-8B-rebatch	0.44	0.64	0.46	0.48	0.51
llama-3-8B-instruct	0.47	0.59	0.47	0.52	0.51
llama-3-8B	0.44	0.64	0.47	0.45	0.5
Reynaerde-7B-Chat	0.44	0.62	0.39	0.52	0.49
Llama-3-ChocoLlama-base	**0.45**	**0.64**	**0.44**	**0.44**	**0.49**
zephyr-7b-beta	0.43	0.58	0.43	0.53	0.49
geitje-7b-ultra	0.40	0.66	0.36	0.49	0.48
ChocoLlama-2-7B-tokentrans-instruct	**0.45**	**0.62**	**0.34**	**0.42**	**0.46**
mistral-7b-v0.1	0.43	0.58	0.37	0.45	0.46
ChocoLlama-2-7B-tokentrans-base	**0.42**	**0.61**	**0.32**	**0.43**	**0.45**
ChocoLlama-2-7B-instruct	**0.36**	**0.57**	**0.33**	**0.45**	**0.43
ChocoLlama-2-7B-base	**0.35**	**0.56**	**0.31**	**0.43**	**0.41**
llama-2-7b-chat-hf	0.36	0.49	0.33	0.44	0.41
llama-2-7b-hf	0.36	0.51	0.32	0.41	0.40

Computerwetenschappers bouwen Vlaams AI-model ChocoLlama

06 februari 2025 16:48

All our models are publicly available

Model weights on Hugging Face

ChocoLlama/ChocoLlama-2-7B-base
Text Generation • Updated Dec 16, 2024 • 31 • 2

ChocoLlama/ChocoLlama-2-7B-instruct
Text Generation • Updated Dec 16, 2024 • 28 • 2

ChocoLlama/ChocoLlama-2-7B-tokentrans-instruct
Text Generation • Updated Dec 16, 2024 • 21 • 1

ChocoLlama/ChocoLlama-2-7B-tokentrans-base
Text Generation • Updated Dec 16, 2024 • 29

ChocoLlama/Llama-3-ChocoLlama-8B-base
Text Generation • Updated Dec 16, 2024 • 117 • 1

ChocoLlama/Llama-3-ChocoLlama-8B-instruct
Text Generation • Updated Dec 16, 2024 • 83 • 6

Tweeties/tweety-7b-dutch-v24a
Text Generation • Updated Aug 9, 2024 • 1.88k • 13

Tweeties/tweety-tatar-hydra-mt-7b-v24a
Text Generation • Updated Aug 9, 2024 • 13

Tweeties/tweety-tatar-hydra-base-7b-v24a
Text Generation • Updated Aug 9, 2024 • 14

Tweeties/tweety-7b-tatar-v24a
Text Generation • Updated Aug 9, 2024 • 40 • 11

Tweeties/tweety-7b-armenian-v24a
Text Generation • Updated May 27, 2024 • 4 • 1

Tweeties/tweety-7b-italian-v24b-llama3 private
Text Generation • Updated May 13, 2024

Piet

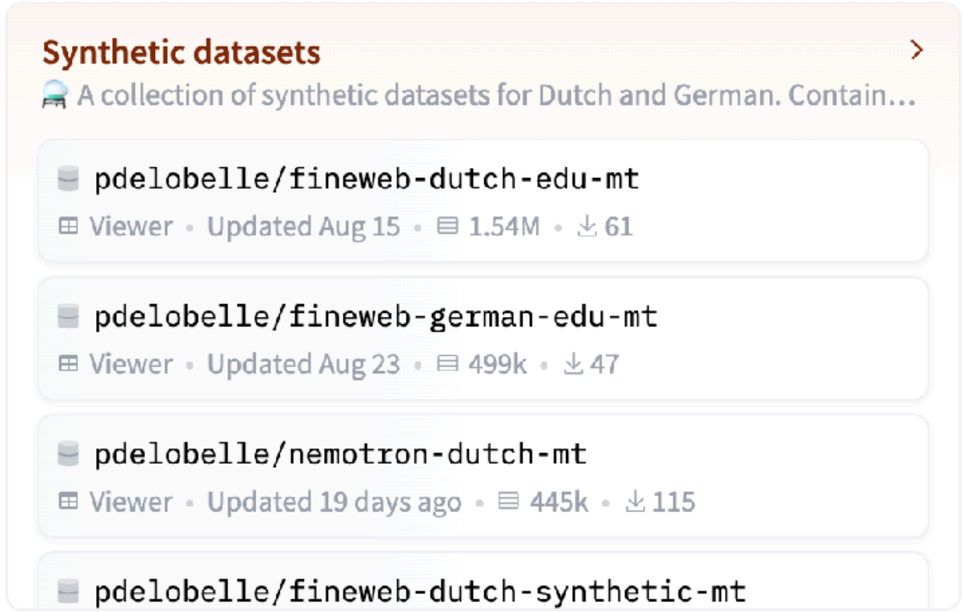
What's next?

What do we need for the next generation of Dutch LLMs?

What's next?

What do we need for the next generation of Dutch LLMs?

- **Data**: instruction-tuning, more domains, RL



What's next?

What do we need for the next generation of Dutch LLMs?

- **Data**: instruction-tuning, more domains, RL
- **Bigger models?** 🤷

What's next?

What do we need for the next generation of Dutch LLMs?

- **Data**: instruction-tuning, more domains, RL
- **Bigger models?** 🤷
- Better generative **benchmarks**

Slides available: pieter.ai/appearances.html