# Linguistic Disparities in Language Technology
## Lessons from West-Greenlandic

**Esther Ploeger**
Department of Computer Science
Aalborg University, Copenhagen

espl@cs.aau.dk

The start of a research project…

The start of a research project…

# West-Greenlandic (Kalaallisut)

# West-Greenlandic (Kalaallisut)


*Map: Alyson Hurt/NPR*

## Some facts:

- The official language of Greenland

# West-Greenlandic (Kalaallisut)


*Map: Alyson Hurt/NPR*

## Some facts:

- The official language of Greenland
- Family: Inuit-Yupik-Unangan

# West-Greenlandic (Kalaallisut)


*Map: Alyson Hurt/NPR*

## Some facts:

- The official language of Greenland
- Family: Inuit-Yupik-Unangan
- Speakers: approx. 50,000 – 60,000

# West-Greenlandic (Kalaallisut)


*Map: Alyson Hurt/NPR*

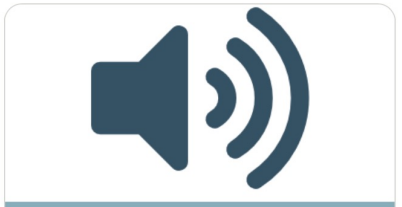
## Some facts:

- The official language of Greenland
- Family: Inuit-Yupik-Unangan
- Speakers: approx. 50,000 – 60,000
- Greenland was under Danish colonial rule from 1721 until 1953

# Greenlandic Language Secretariat

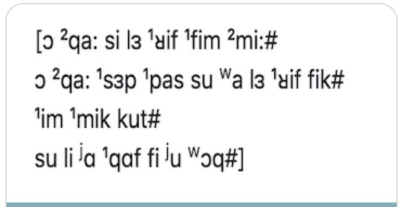
# Greenlandic Language Secretariat


Spellchecker

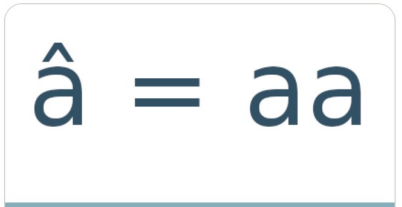


Speech Synthesis Martha

Corpus

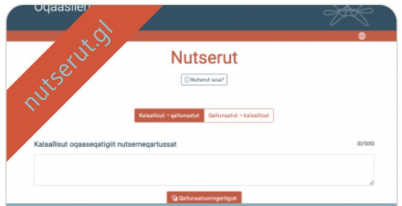



IPA Converter

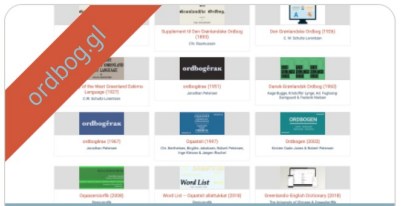



Kleinschmidt Converter

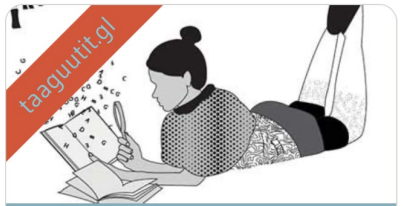

Katersat (worddatabase)




Nutserut (translator)




Dictionaries

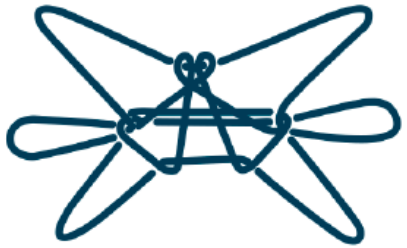



Taaguutaasivik (termbank)

https://oqaasileriffik.gl/en/resources/

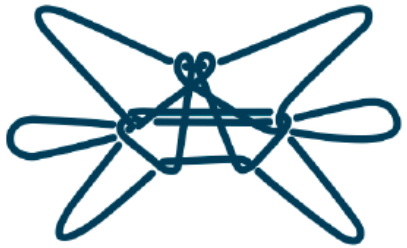
# Greenlandic Language Secretariat

What could we work on?



# Greenlandic Language Secretariat

"An area in need of improvement is danish-greenlandic law texts."



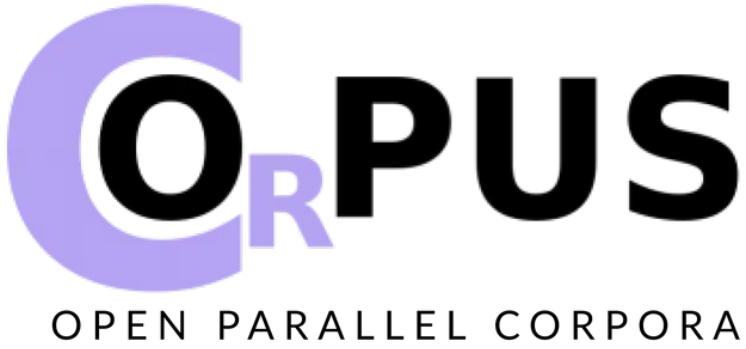
# A Tale of Three Obstacles

# A Tale of Three Obstacles

Data

# Obstacle 1: Data

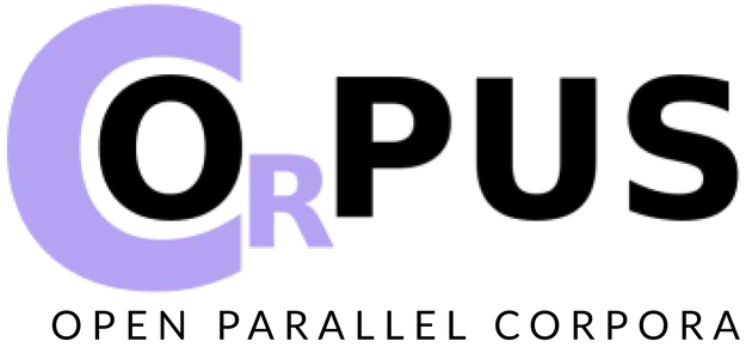
# Obstacle 1: Data




*Tiedemann (2009)*

**Size**: 291 lines
**Domain**: manuals

# Obstacle 1: Data




*Tiedemann (2009)*

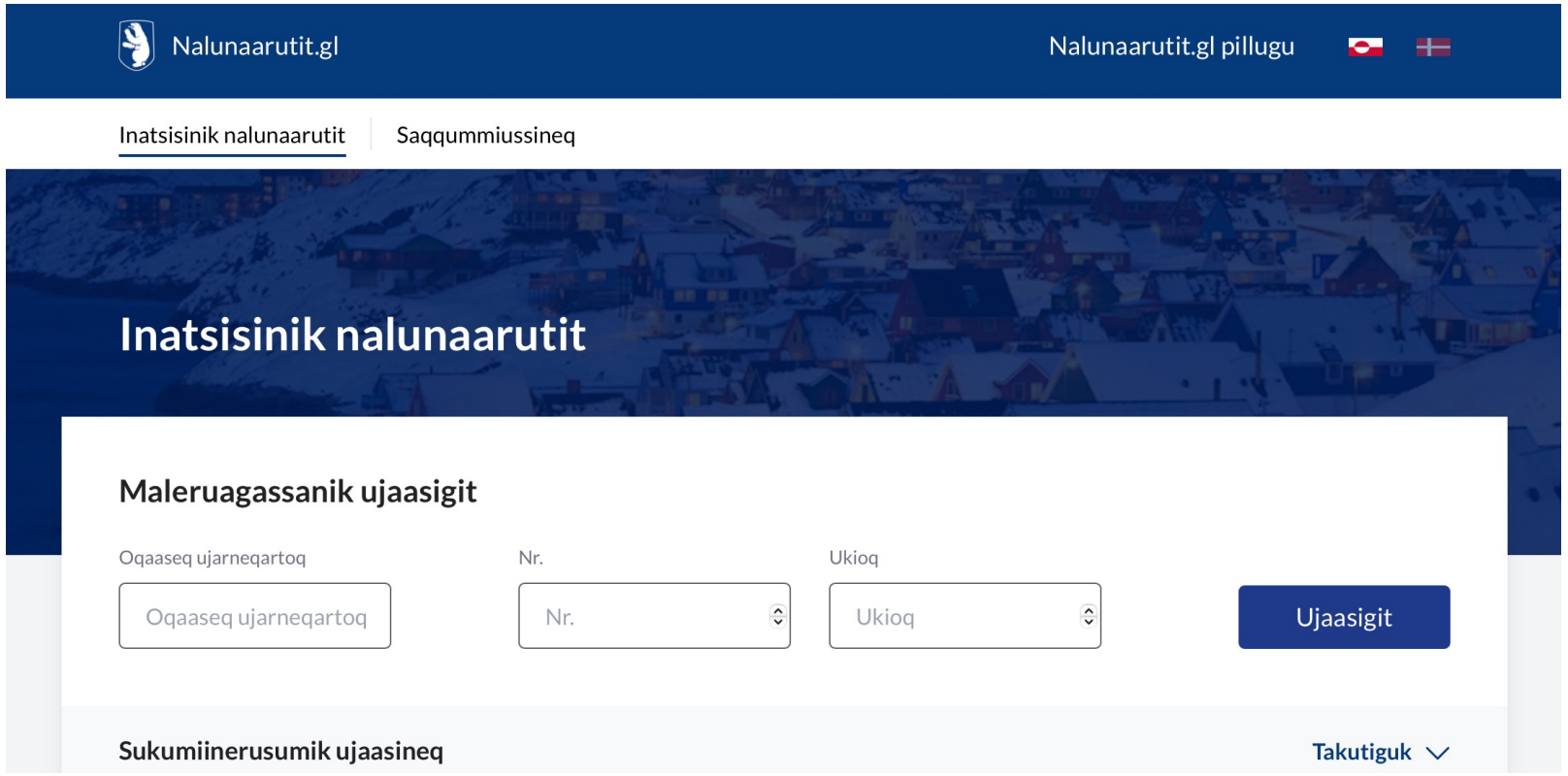
**Size**: 291 lines
**Domain**: manuals


*Kristensen-Mclachlan and Nedergard (2024)*

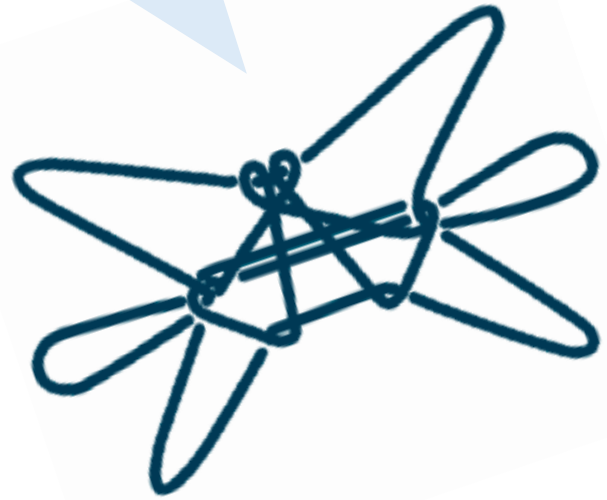
**Size**: 1.2m West-Greenlandic words; 2.1m Danish
**Domain**: news

# Obstacle 1: Data

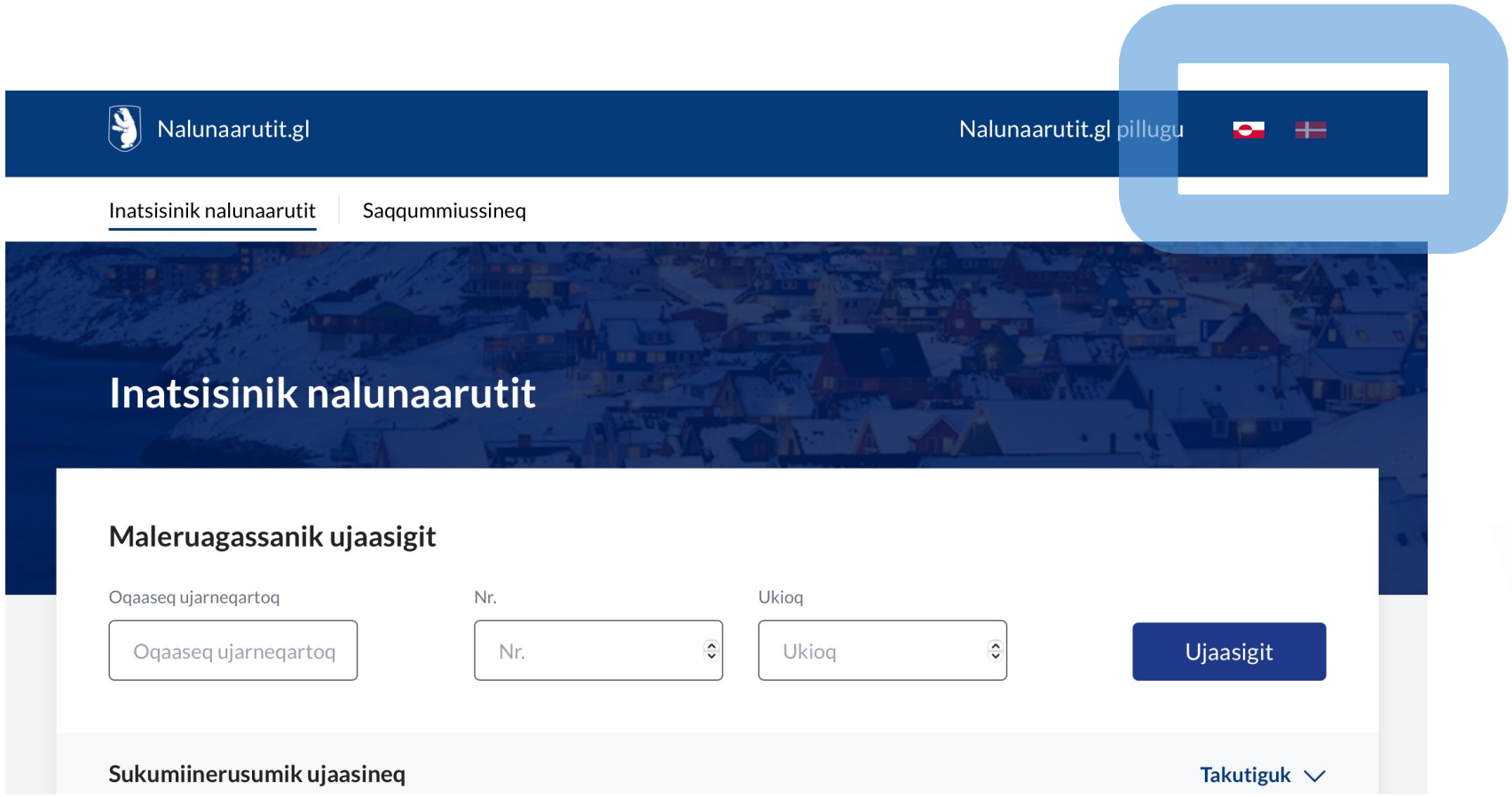
# Obstacle 1: Data



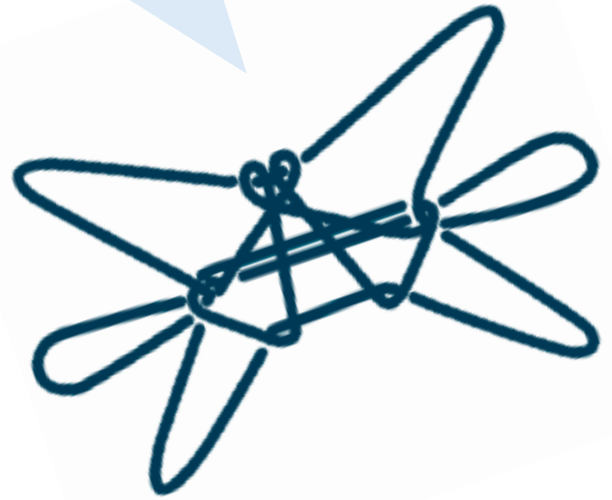

There is publicly available legal text!



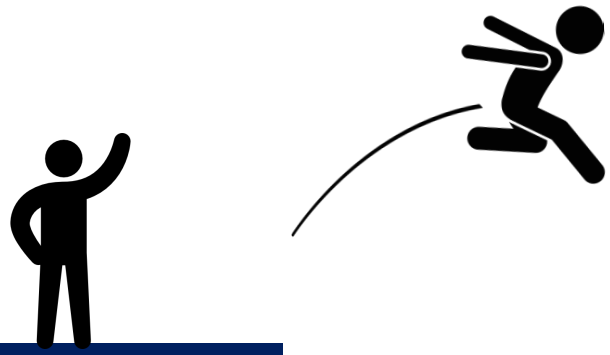
# Obstacle 1: Data




There is publicly available legal text!



# A Tale of Three Obstacles





Data

# A Tale of Three Obstacles

Data

Tools

# Obstacle 2: Tools

## Sentence Alignment

# Obstacle 2: Tools

## Sentence Alignment

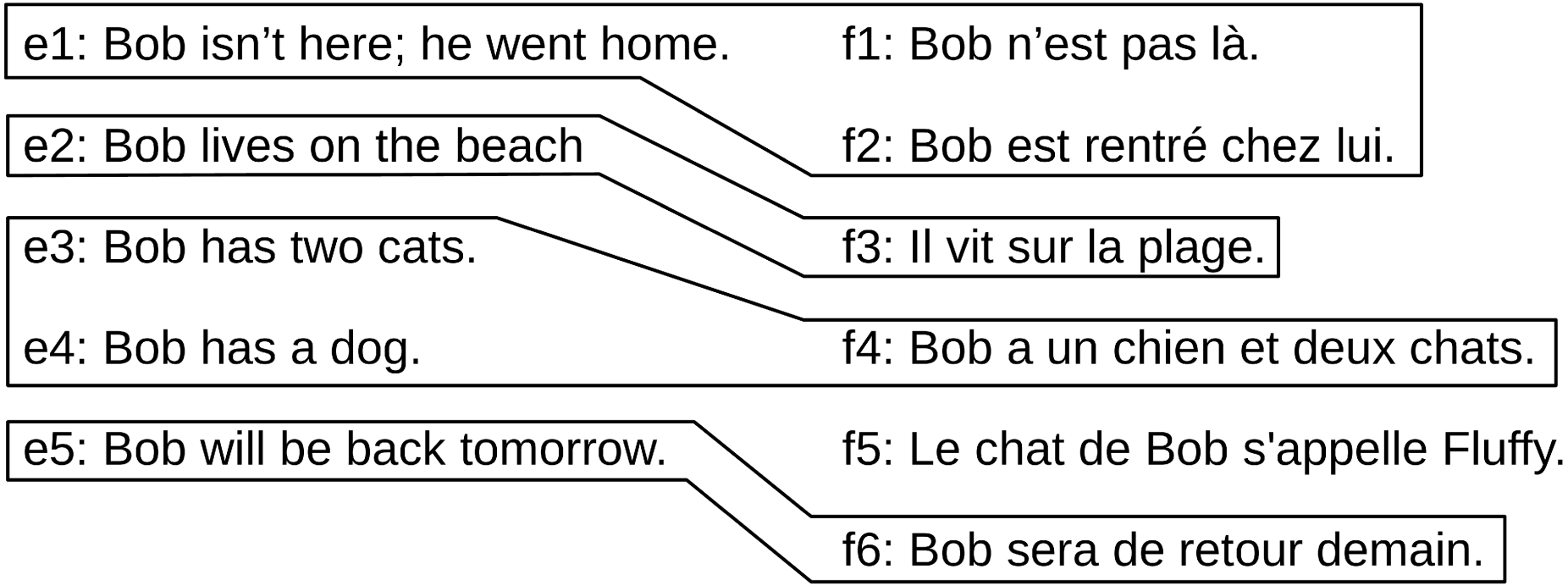
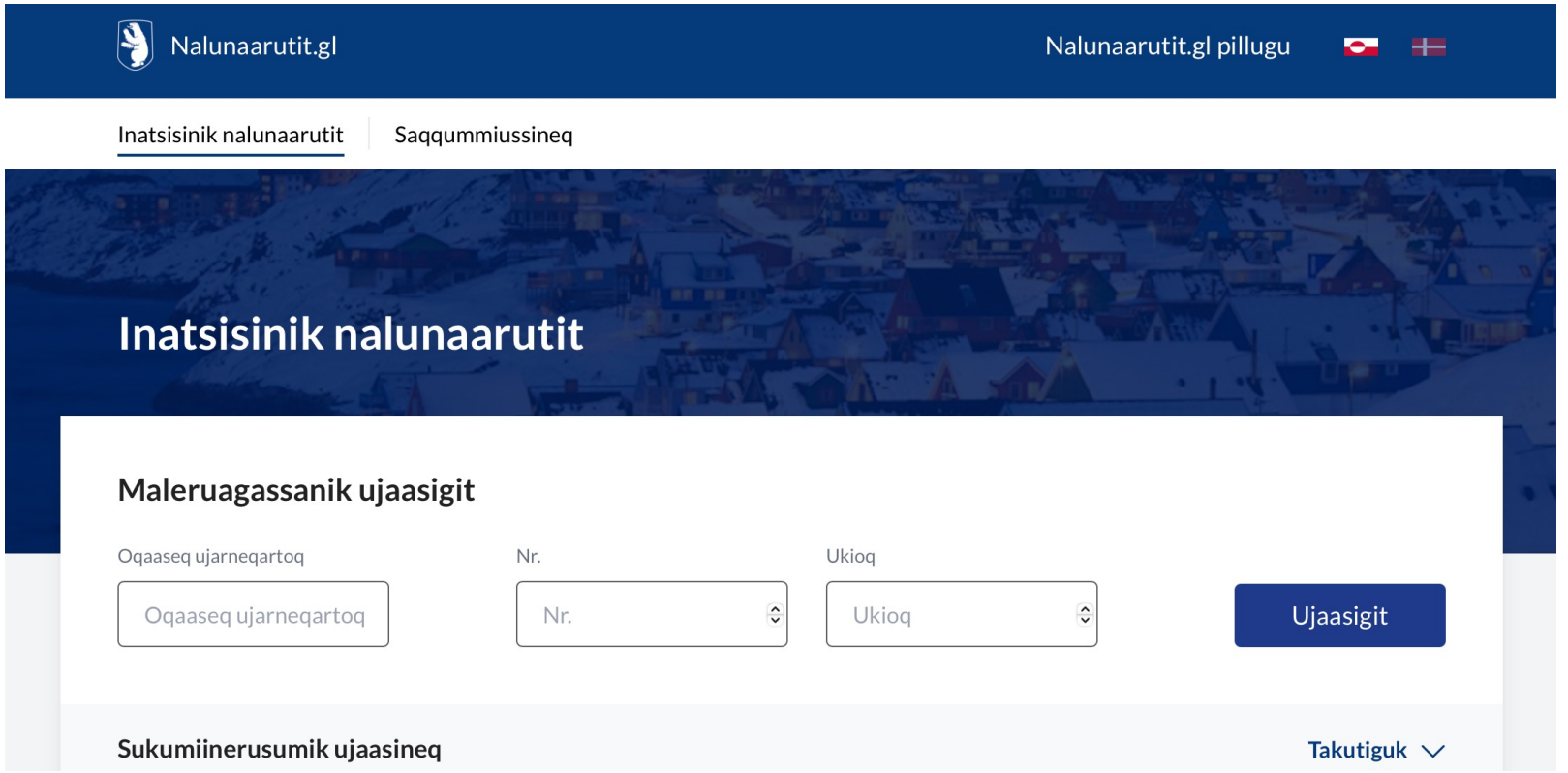



**Image:** *Vecalign: Improved Sentence Alignment in Linear Time and Space* (Thompson & Koehn, EMNLP-IJCNLP 2019)

# Obstacle 2: Tools



# Obstacle 2: Tools

§ 1. Inatsit manna suliffeqarfinnut inunnullu makkununnga atuuppoq:
1) Suliffeqarfiit inuillu nunat allat aningaasaannut nuutsitsisarnermik inuussutissarsiuteqartut.
2) Suliffeqarfiit inuillu attartortitsinernut aningaasalersuinermik imaluunniit taarsigassarsisitsisarnermik inuussutissarsiuteqartut.
3) Illunik nioqquteqarnermi akunnermiliuttartut.
4) Suliffeqarfiit inuillu illunik nioqquteqarnermi akunnermiliuttartutut kiffartuussinernik inuussutissarsiuteqartut.

§ 1.  Denne lov finder anvendelse på følgende virksomheder og personer:
1) Virksomheder og personer, der erhvervsmæssigt udøver virksomhed med valutaveksling.
2) Virksomheder og personer, der erhvervsmæssigt udøver finansiel leasing eller udøver udlånsvirksomhed.
3) Ejendomsmæglere.
4) Virksomheder og personer, der i øvrigt erhvervsmæssigt leverer samme ydelser som ejendomsmæglere.

# Obstacle 2: Tools

§ 1. Inatsit manna suliffeqarfinnut inunnullu makkununnga atuuppoq:
1) Suliffeqarfiit inuillu nunat allat aningaasaannut nuutsitsisarnermik inuussutissarsiuteqartut.
2) Suliffeqarfiit inuillu attartortitsinernut aningaasalersuinermik imaluunniit taarsigassarsisitsisarnermik inuussutissarsiuteqartut.
3) Illunik nioqquteqarnermi akunnermiliuttartut.
4) Suliffeqarfiit inuillu illunik nioqquteqarnermi akunnermiliuttartutut kiffartuussinernik inuussutissarsiuteqartut.

§ 1. Denne lov finder anvendelse på følgende virksomheder og personer:
1) Virksomheder og personer, der erhvervsmæssigt udøver virksomhed med valutaveksling.
2) Virksomheder og personer, der erhvervsmæssigt udøver finansiel leasing eller udøver udlånsvirksomhed.
3) Ejendomsmæglere.
4) Virksomheder og personer, der i øvrigt erhvervsmæssigt leverer samme ydelser som ejendomsmæglere.

# Obstacle 2: Tools

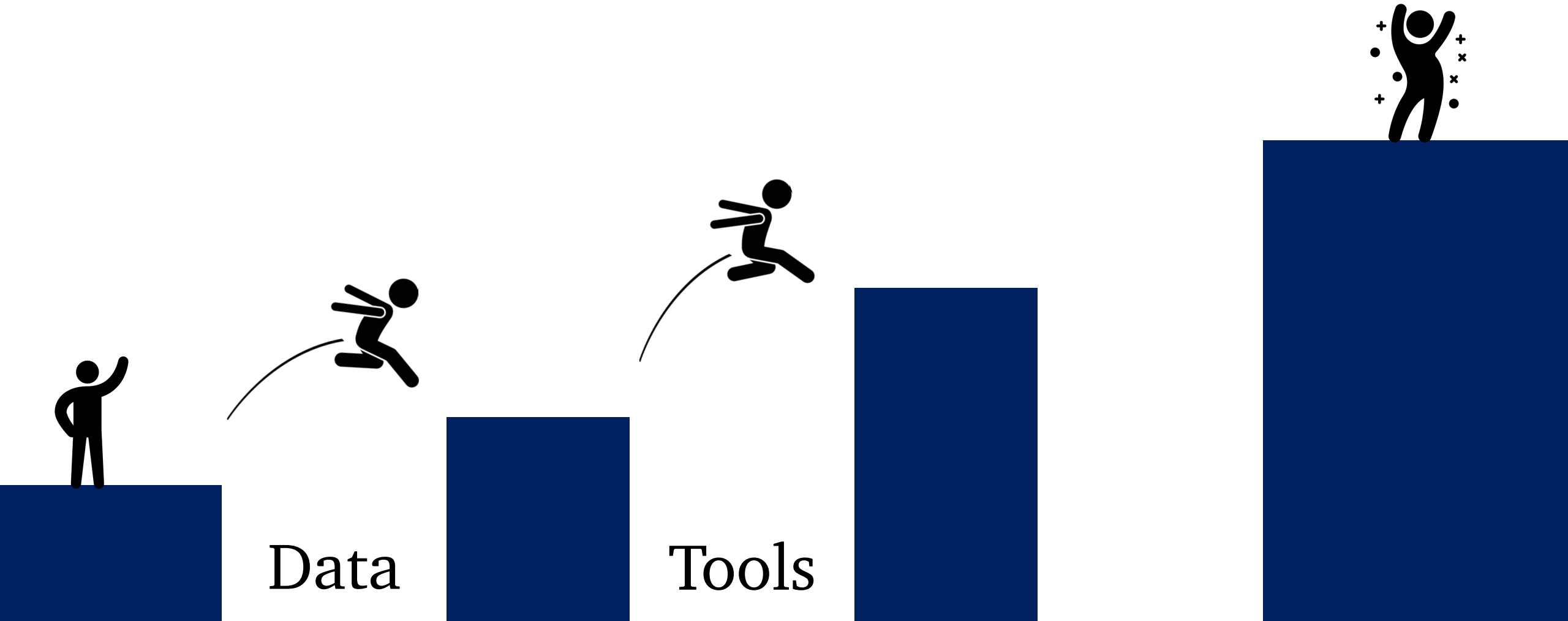
§ 1. Inatsit manna suliffeqarfinnut inunnullu makkununnga atuuppoq:
1) Suliffeqarfiit inuillu nunat allat aningaasaannut nuutsitsisarnermik inuussutissarsiuteqartut.
2) Suliffeqarfiit inuillu attartortitsinernut aningaasalersuinermik imaluunniit taarsigassarsisitsisarnermik inuussutissarsiuteqartut.
3) Illunik nioqquteqarnermi akunnermiliuttartut.
4) Suliffeqarfiit inuillu illunik nioqquteqarnermi akunnermiliuttartutut kiffartuussinernik inuussutissarsiuteqartut.

§ 1. Denne lov finder anvendelse på følgende virksomheder og personer:
1) Virksomheder og personer, der erhvervsmæssigt udøver virksomhed med valutaveksling.
2) Virksomheder og personer, der erhvervsmæssigt udøver finansiel leasing eller udøver udlånsvirksomhed.
3) Ejendomsmæglere.
4) Virksomheder og personer, der i øvrigt erhvervsmæssigt leverer samme ydelser som ejendomsmæglere.

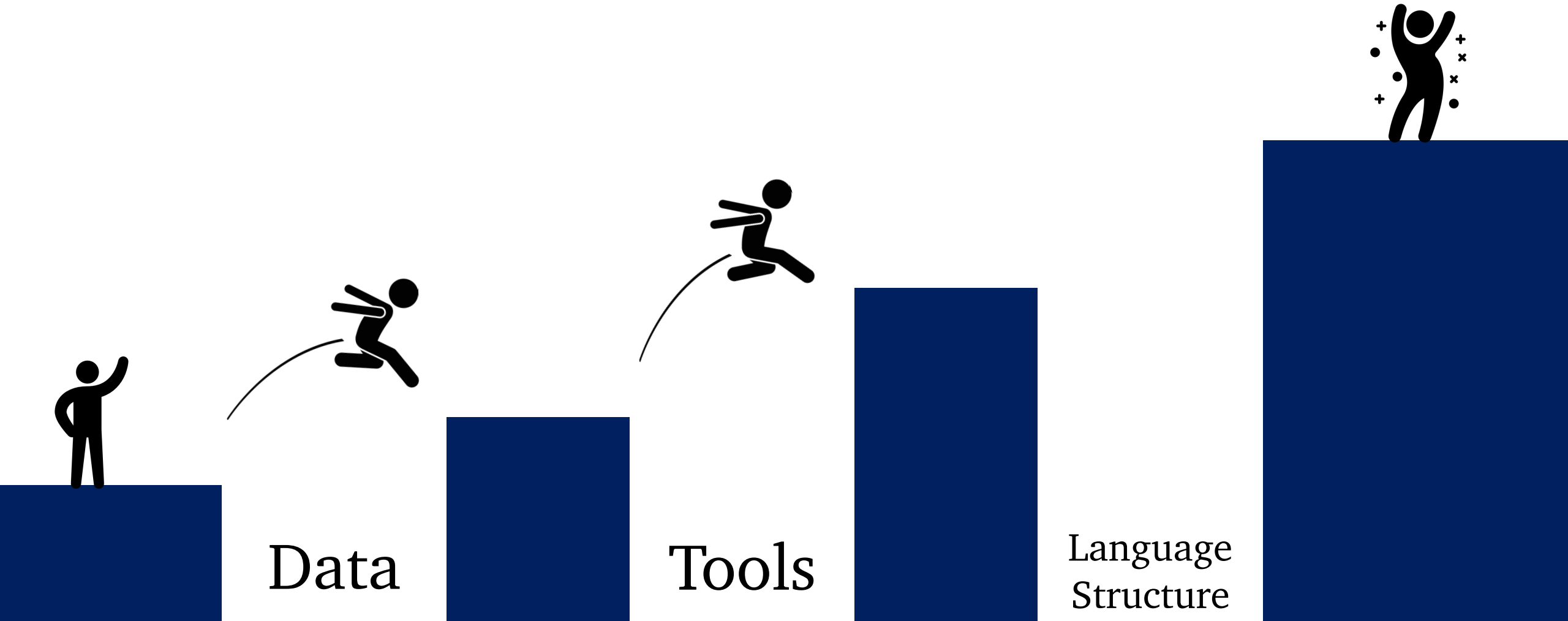
# Obstacle 2: Tools

| Split | # Lines | | # Words | |
|---|---|---|---|---|
| | *GL* | *DA* | *GL* | *DA* |
| Training | 39,936 | 39,936 | 663,734 | 929,904 |
| Validation | 1,000 | 1,000 | 16,594 | 23,021 |
| Testing | 1,000 | 1,000 | 16,665 | 23,846 |
| **Total** | **41,936** | **41,936** | **696,993** | **976,771** |

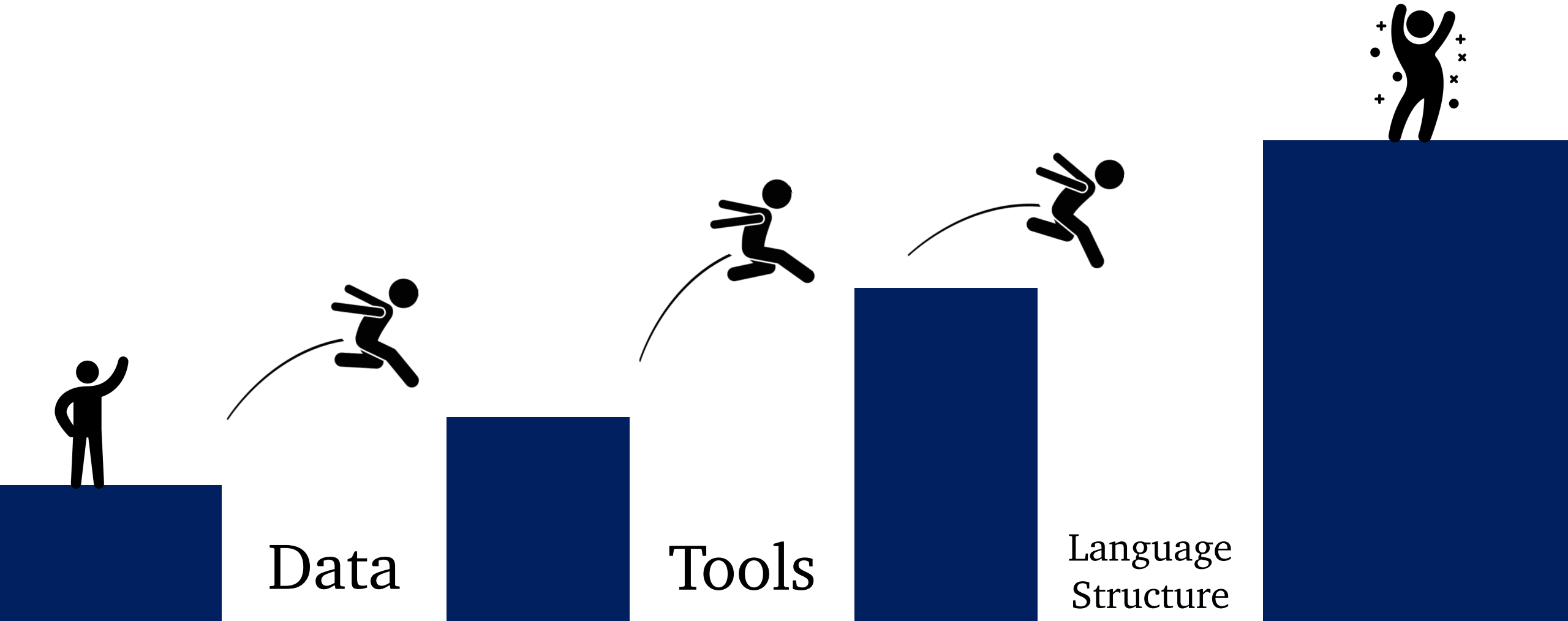
# A Tale of Three Obstacles



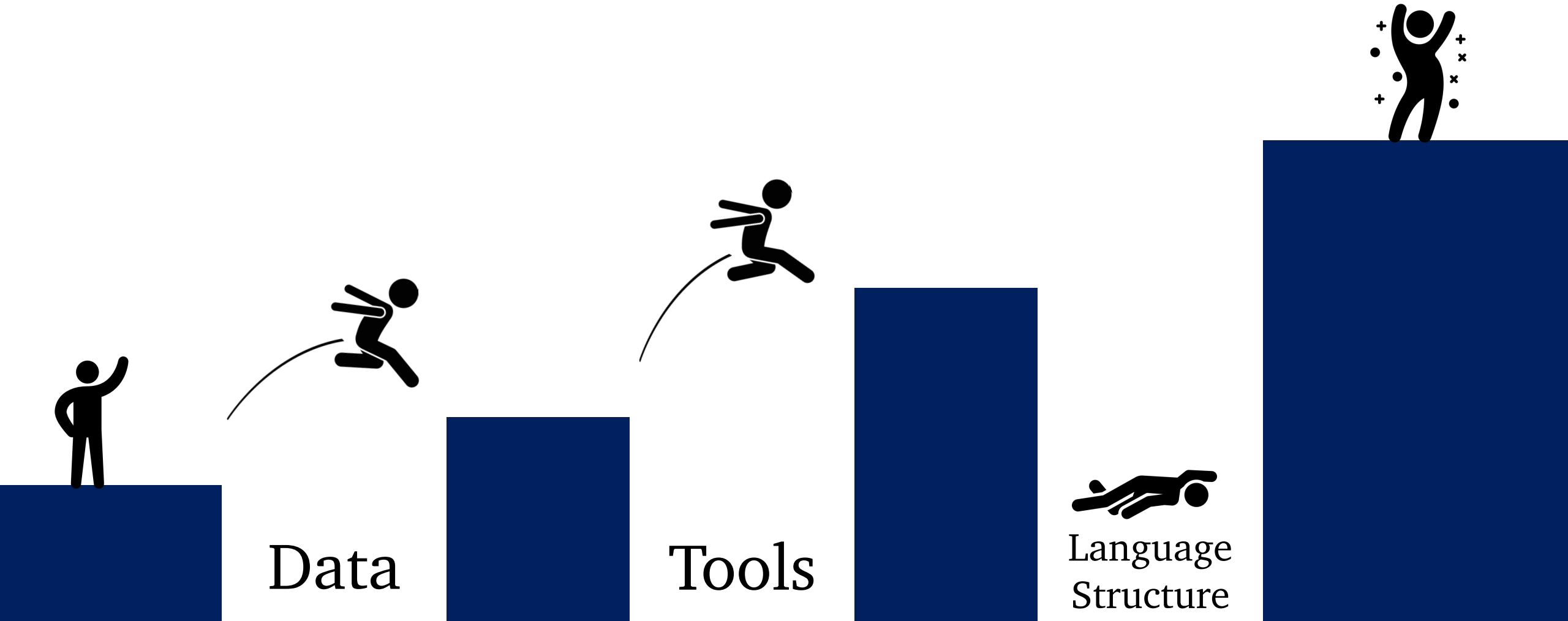
# A Tale of Three Obstacles



# A Tale of Three Obstacles



# A Tale of Three Obstacles



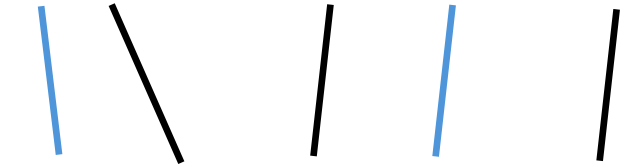
# Obstacle 3: Language Structure

# Obstacle 3: Language Structure

| | |
|---|---|
| English: | if he doesn’t have a dog |
| Danish: | hvis han ikke har en hund |

# Obstacle 3: Language Structure

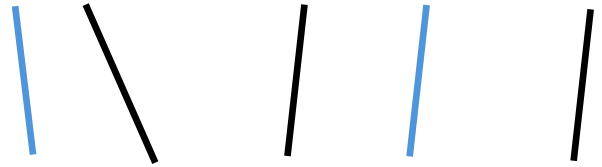
English: if he doesn't have a dog



Danish: hvis han ikke har en hund

# Obstacle 3: Language Structure

English: if he doesn't have a dog



Danish: hvis han ikke har en hund

West-Greenlandic: qimmeqanngikkuni

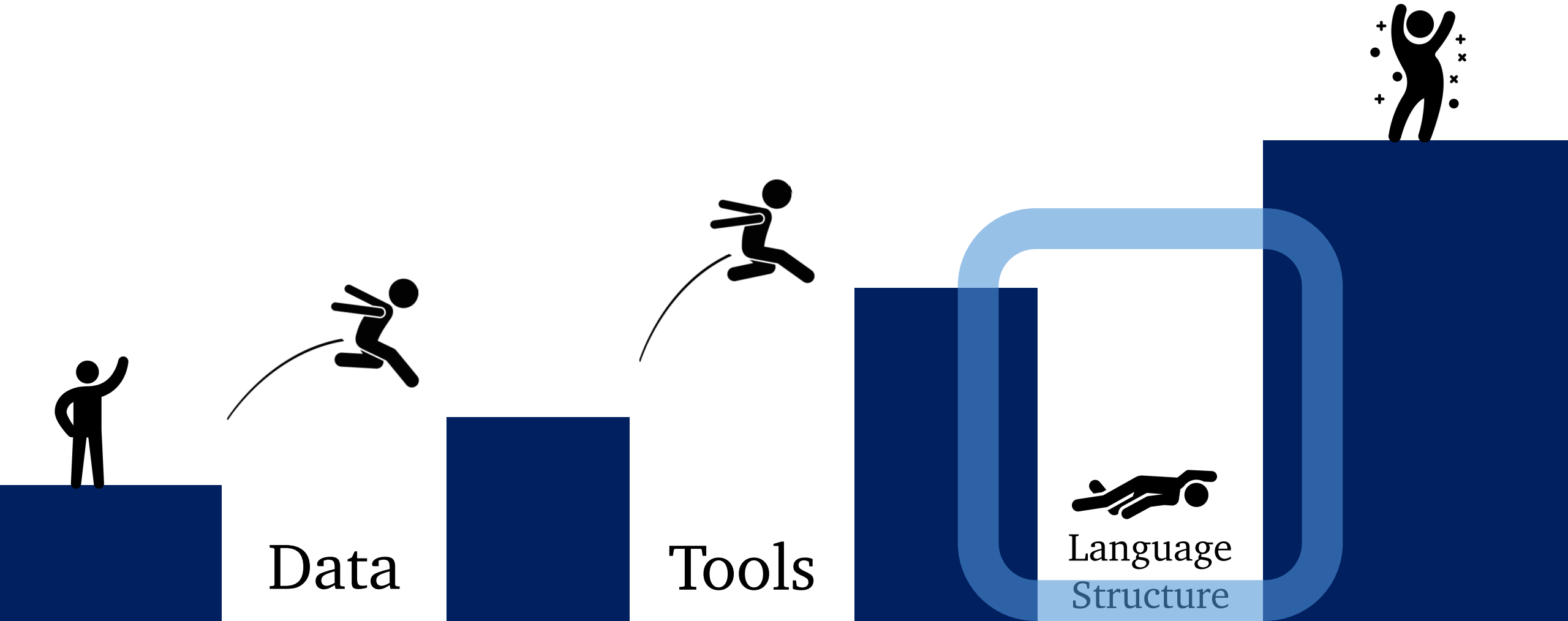
# Obstacle 3: Language Structure

English: if he doesn't have a dog

Danish: hvis han ikke har en hund

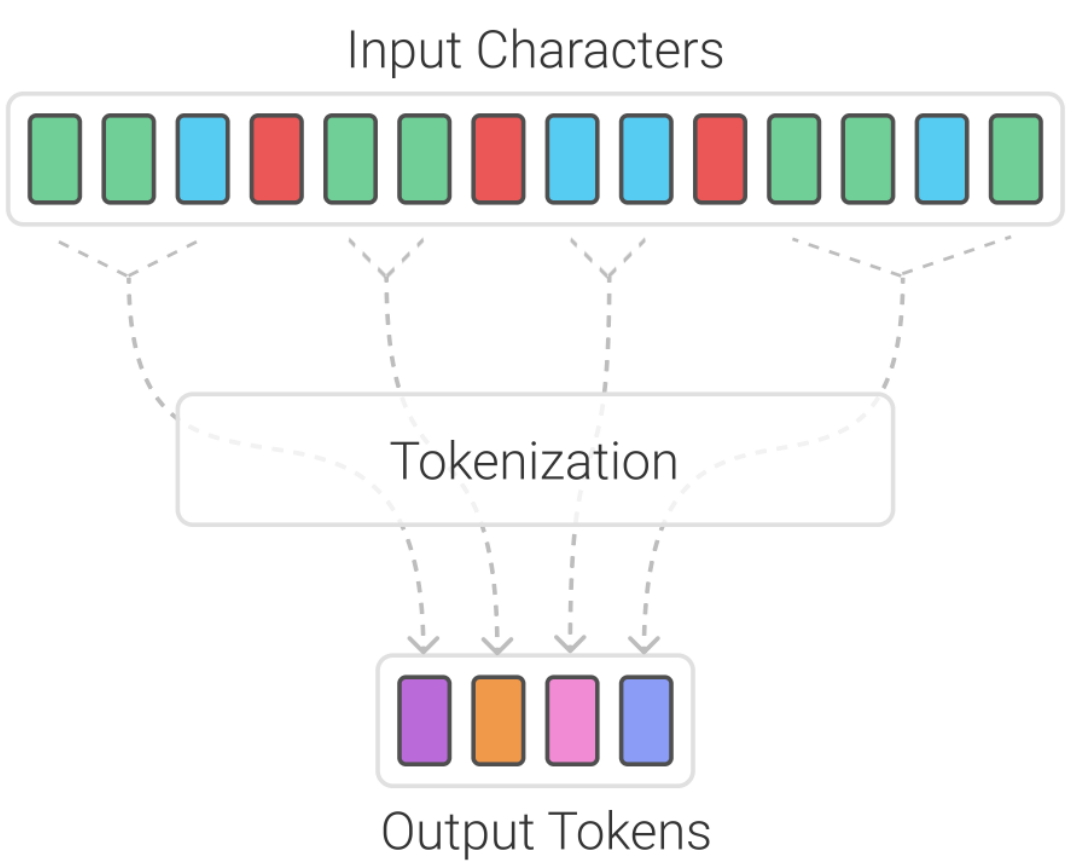
West-Greenlandic: qimmeqanngikkuni

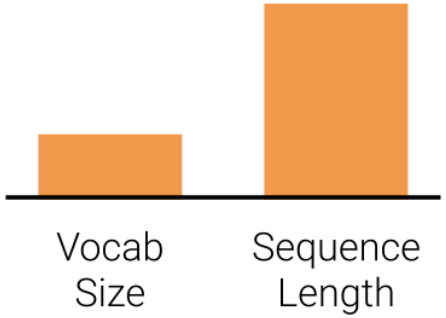
# A Tale of Three Obstacles



How should we segment text in a morphologically complex language like West-Greenlandic?

# Why Segment Text?



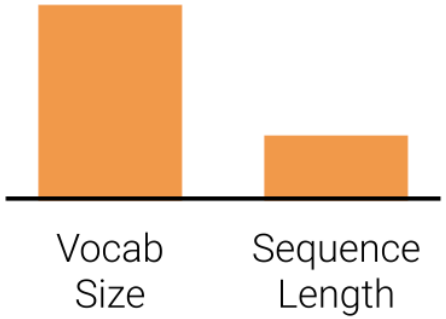




**Image:** https://www.thoughtvector.io/blog/subword-tokenization/

# BPE

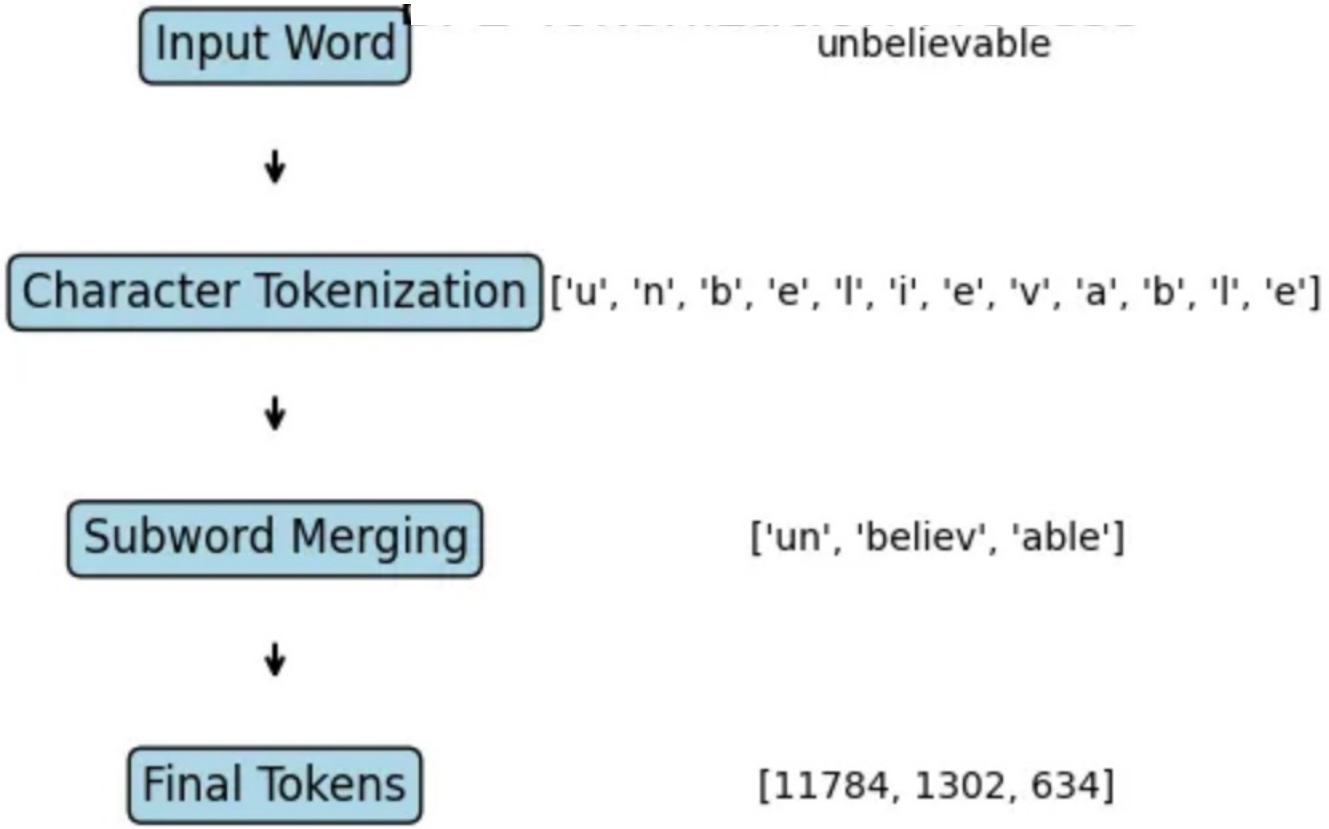
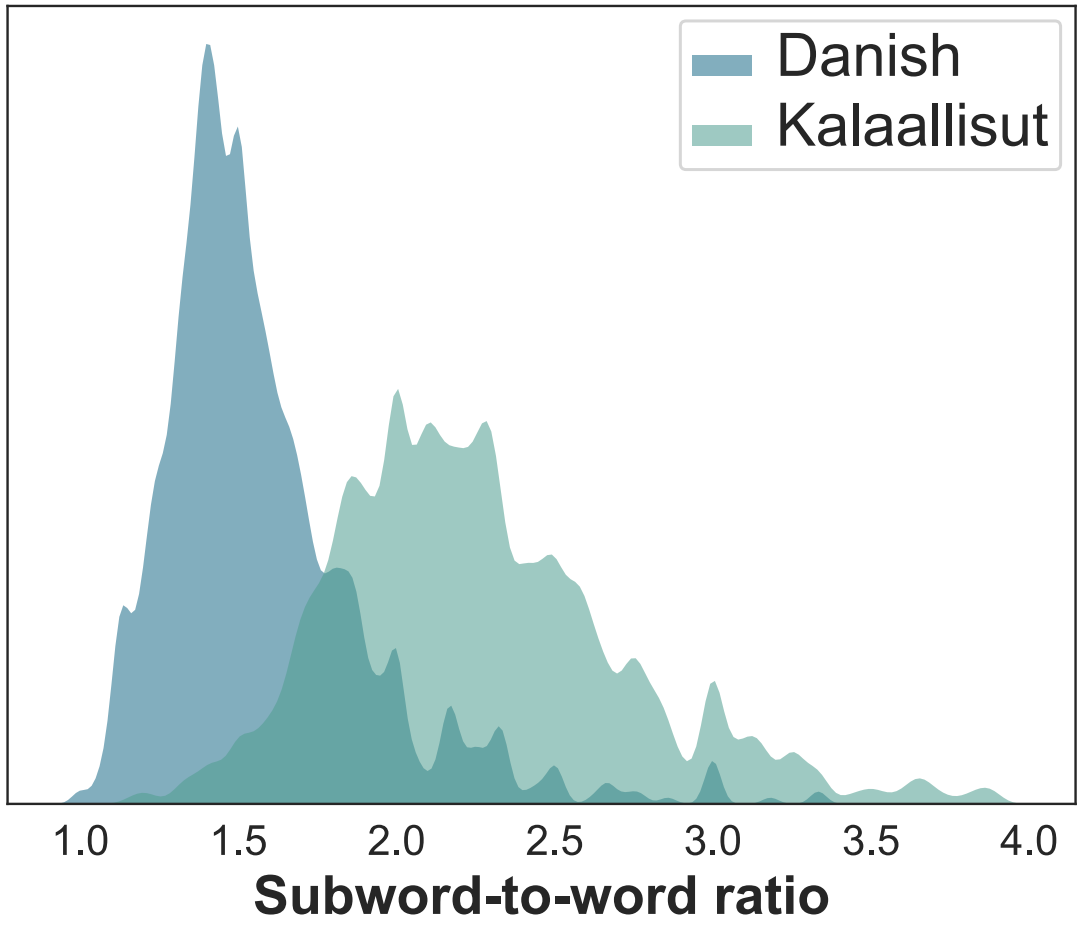



Image: https://python.plainenglish.io/efficient-tokenization-with-byte-pair-encoding-bpe-for-neural-networks-7cf4a54b5fd0

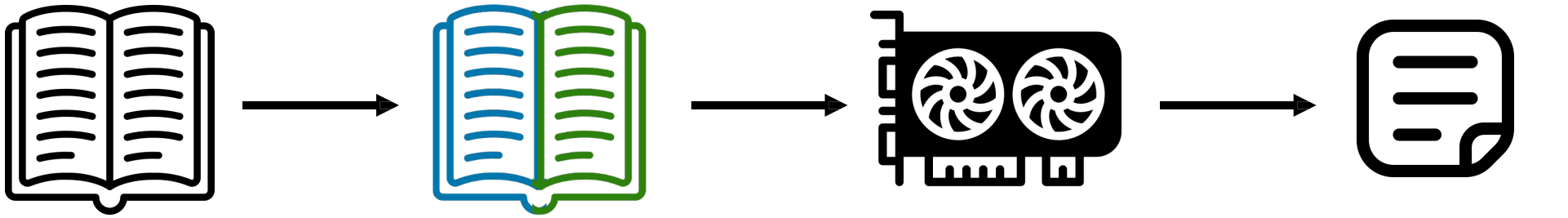
# BPE

Biased by language structure!

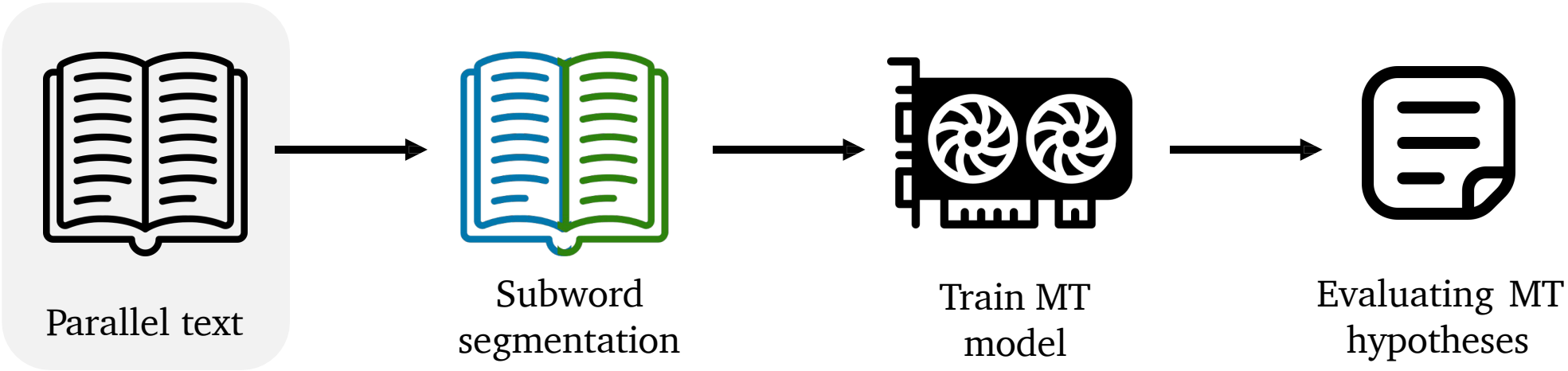


Previous work on neural polysynthetic language modelling suggests that BPE is suboptimal!

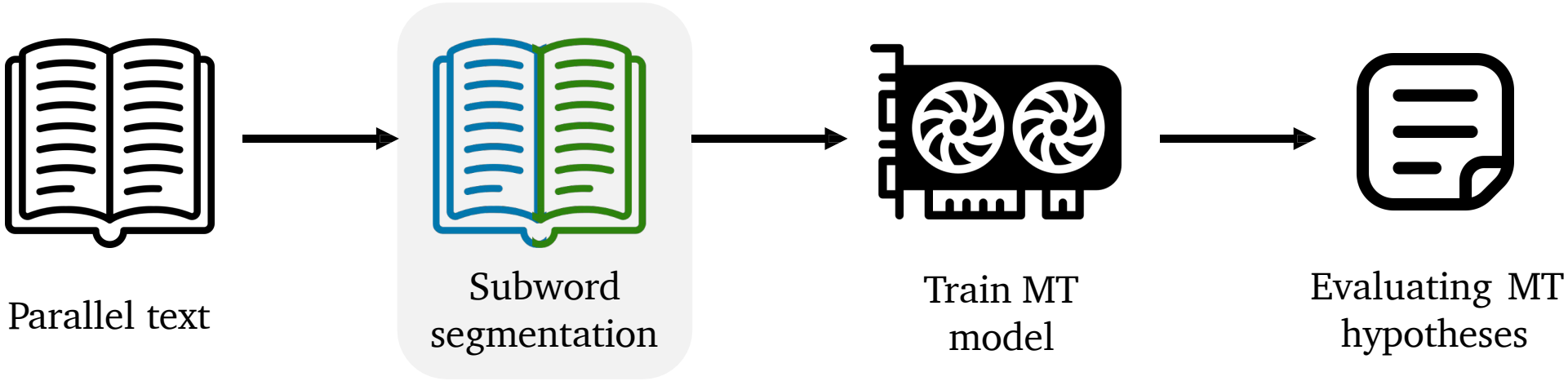
# Our Experiment



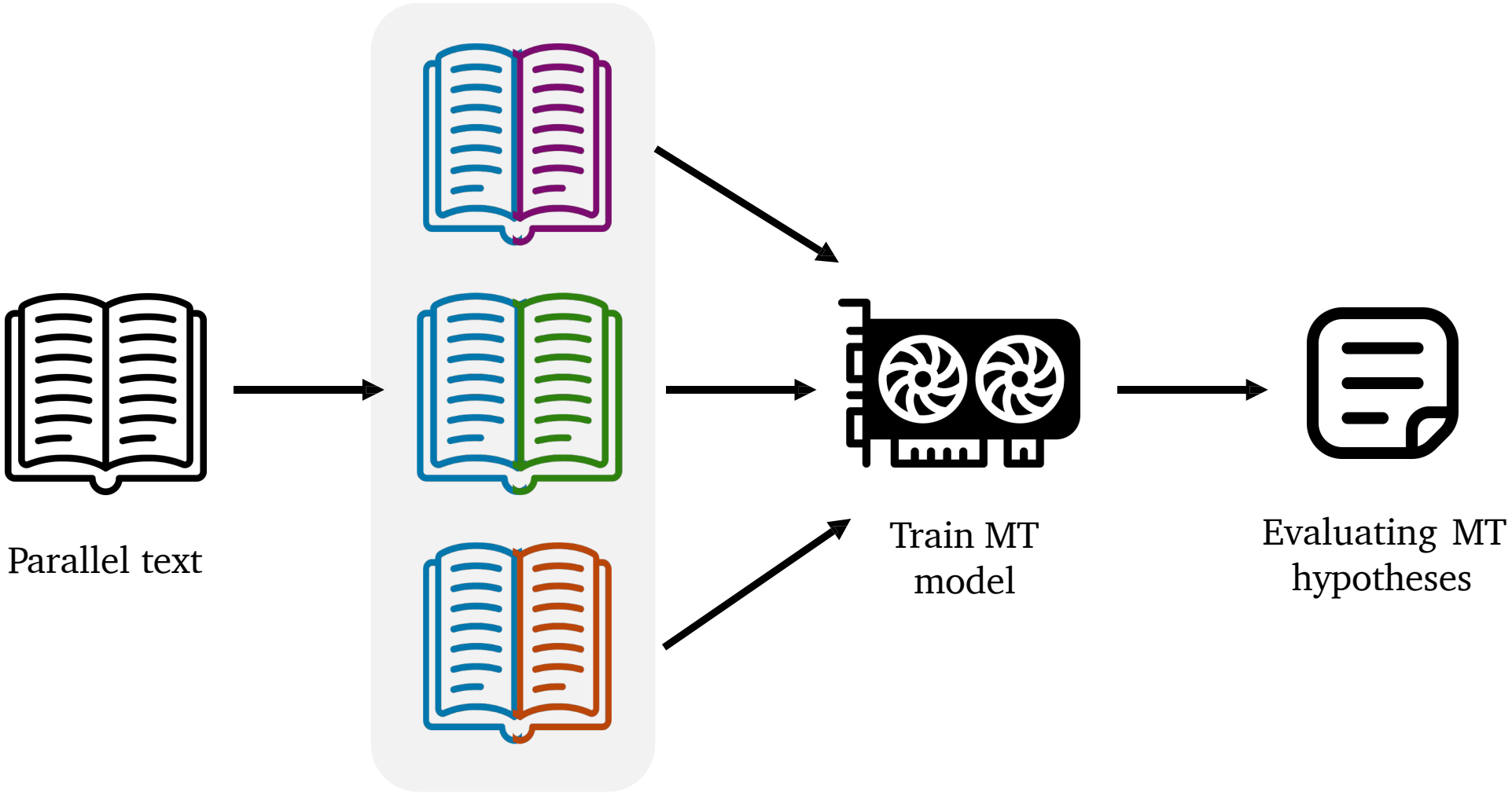
# Our Experiment



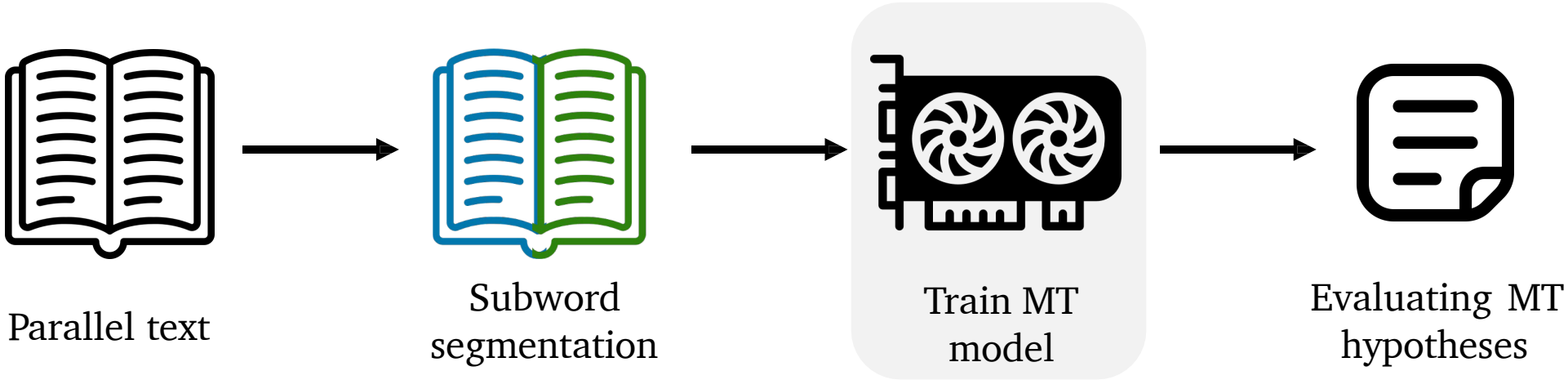
# Our Experiment



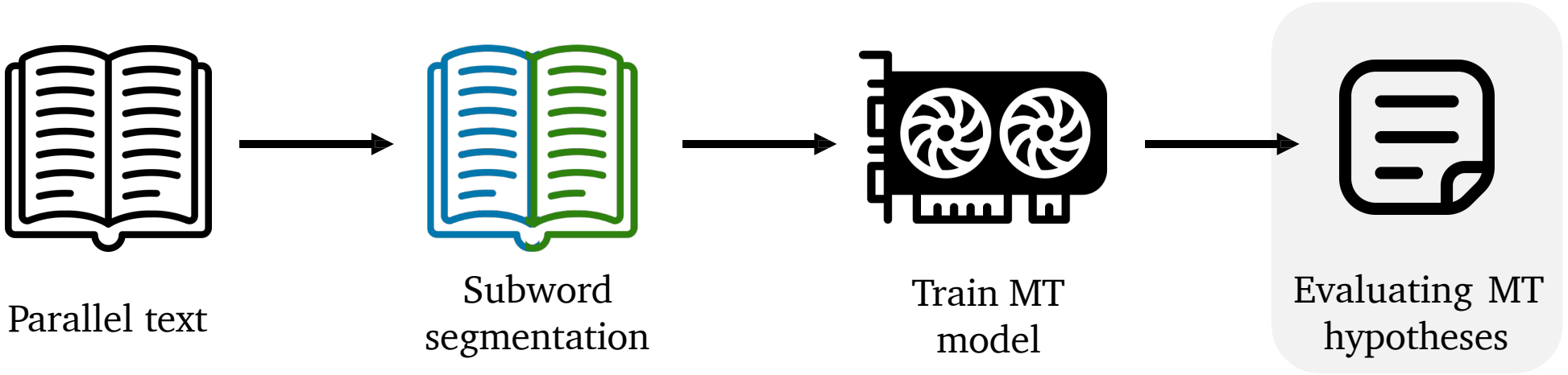
# Our Experiment



# Our Experiment



# Our Experiment



# Results: Morphological Boundaries

# Results: Morphological Boundaries


| | | | | |
|---|---|---|---|---|
| qimme | qa | nngik | ku | ni |

# Results: Morphological Boundaries


qimme qa nngik ku ni

GPT-4: qimmeqanngikkuni

# Results: Morphological Boundaries

$$P = \frac{|\{gold\ morphemes\} \cap \{subwords\}|}{|\{subwords\}|}$$

$$R = \frac{|\{gold\ morphemes\} \cap \{subwords\}|}{|\{gold\ morphemes\}|}$$

*Equations adapted from Nouri and Yangarber (2016)*

# Results: Morphological Boundaries

| Method | Prec. (%) | Rec. (%) | F1 (%) |
|---|---|---|---|
| BPE | 10.81 | 12.42 | 11.56 |
| Unigram | 30.88 | **37.68** | **33.94** |
| Morfessor | **31.08** | 31.61 | 31.34 |
| FlatCat | 29.58 | 29.40 | 29.49 |

# Does this actually help machine translation?

# Results: Machine Translation

| | ***Machine Translation*** | | | |
|---|---|---|---|---|
| | Danish→Kalaallisut | | Kalaallisut→Danish | |
| **Method** | **chrF2** | **BLEU** | **chrF2** | **BLEU** |
| *None* | *44.6* | *3.4* | *61.5* | *15.4* |
| BPE | 56.4 | 8.7 | **64.2** | **21.5** |
| Unigram | 61.4 | **10.1** | 58.9 | 17.1 |

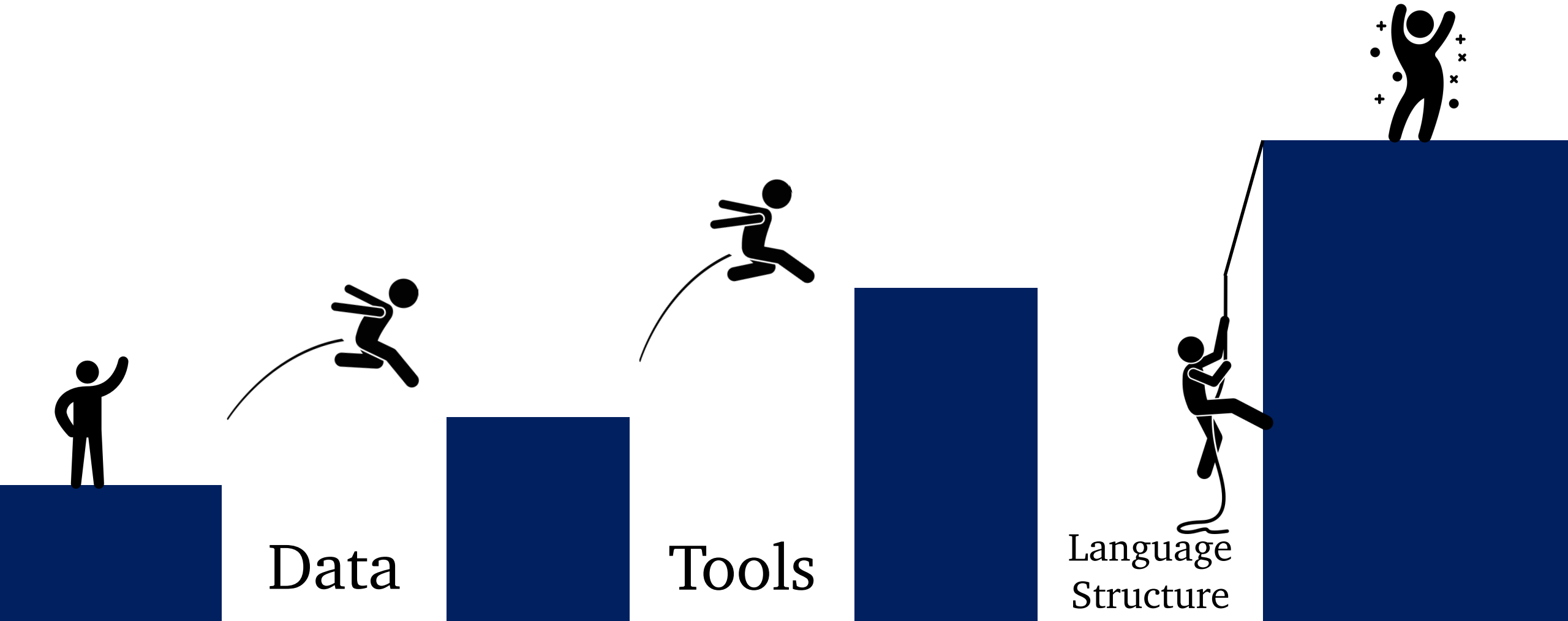
# Results: Machine Translation

| | *Machine Translation* | | | |
|---|---|---|---|---|
| | Danish→Kalaallisut | | Kalaallisut→Danish | |
| **Method** | **chrF2** | **BLEU** | **chrF2** | **BLEU** |
| *None* | *44.6* | *3.4* | *61.5* | *15.4* |
| BPE | 56.4 | 8.7 | **64.2** | **21.5** |
| Unigram | 61.4 | **10.1** | 58.9 | 17.1 |

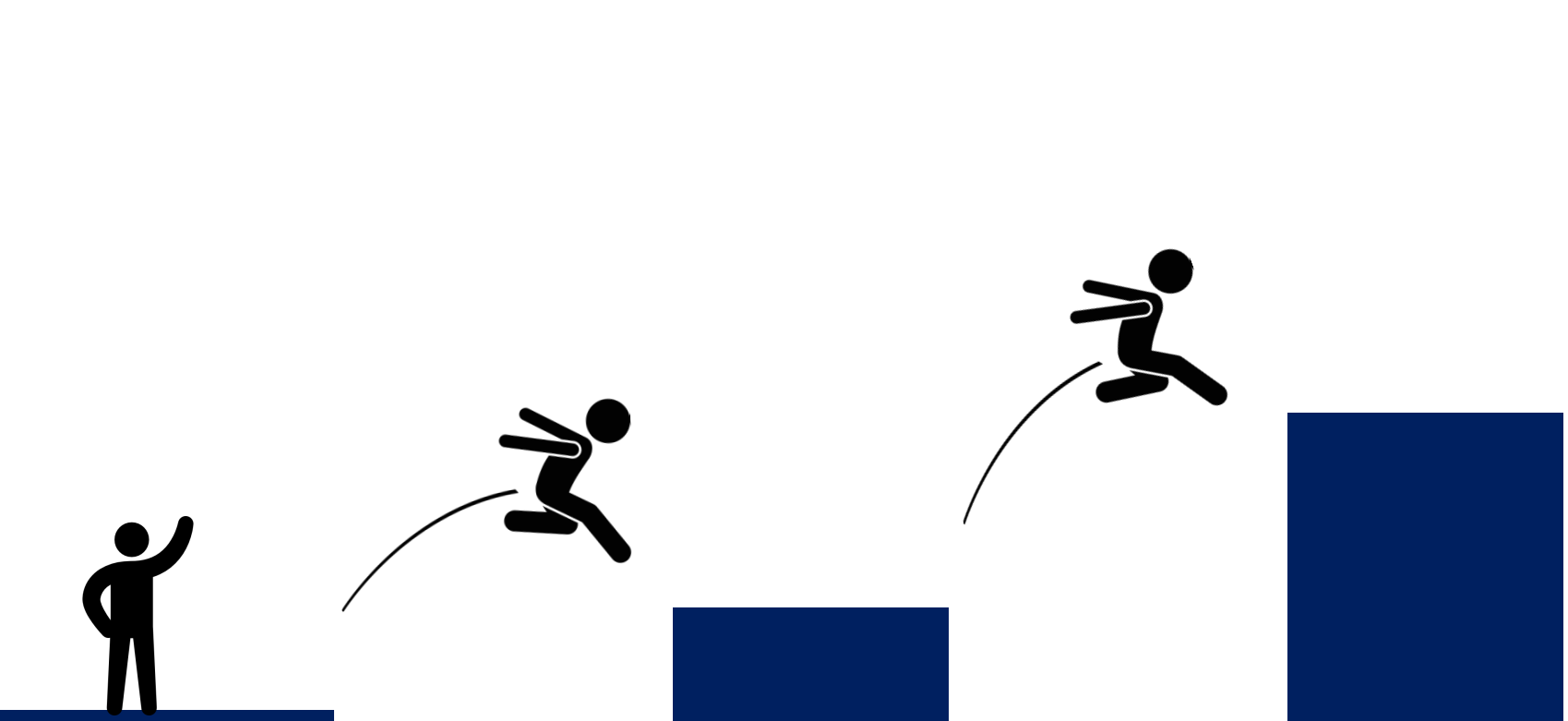
# Results: Machine Translation

| | *Machine Translation* | | | |
|---|---|---|---|---|
| | Danish→Kalaallisut | | Kalaallisut→Danish | |
| **Method** | **chrF2** | **BLEU** | **chrF2** | **BLEU** |
| *None* | *44.6* | *3.4* | *61.5* | *15.4* |
| BPE | 56.4 | 8.7 | **64.2** | **21.5** |
| Unigram | 61.4 | **10.1** | 58.9 | 17.1 |

# A Tale of Three Obstacles



# A Tale of Three Obstacles



Data

Tools

Language
Structure

*Better* does not necessarily mean *good*…

# Take-home message

## The good & the bad news

# The Bad News

# The Bad News

- Data is (very) scarce

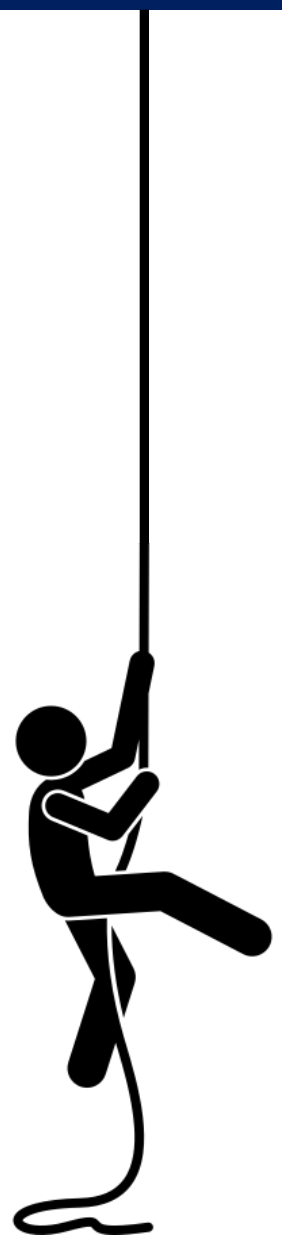
# The Bad News

- Data is (very) scarce
- Out-of-the-box tools (large pre-trained language models, mainstream tokenization algorithms) are not equally applicable to all languages!

# The Bad News

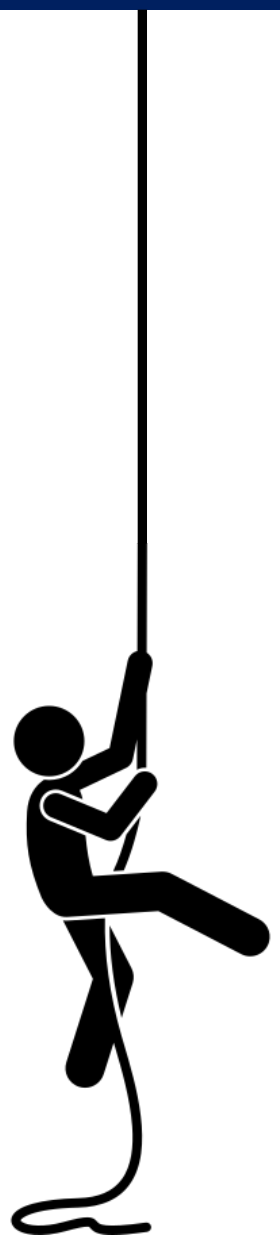
- Data is (very) scarce
- Out-of-the-box tools (large pre-trained language models, mainstream tokenization algorithms) are not equally applicable to all languages!
- Limited commercial (and academic?) interest

# The Good News



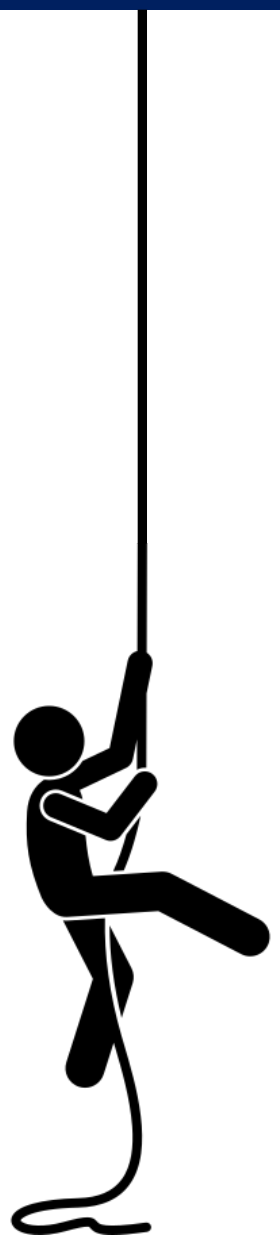
# The Good News

- We can be informed by structural language similarities!



# The Good News

- We can be informed by structural language similarities!
- Small tweaks can bring at least modest improvements!



Who will knock on
a door next?

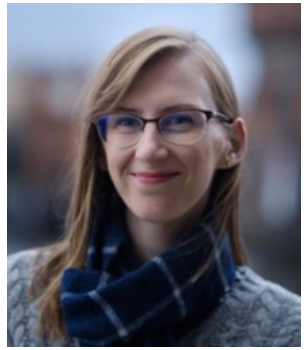
# Qujanaq! Tak! Thank you!

And many thanks to my collaborators
and the Greenlandic Language Secretariat!

Paola Saucedo

Johannes Bjerva

Ross Deans Kristensen-McLachlan



Heather Lent

EP, JB and HL were funded by the Carlsberg Foundation, under the Semper Ardens: Accelerate programme (project nr. CF21-0454).

# Questions?

# Linguistic Disparities in Language Technology
## Lessons from West-Greenlandic

**Esther Ploeger**
Department of Computer Science
Aalborg University, Copenhagen

espl@cs.aau.dk