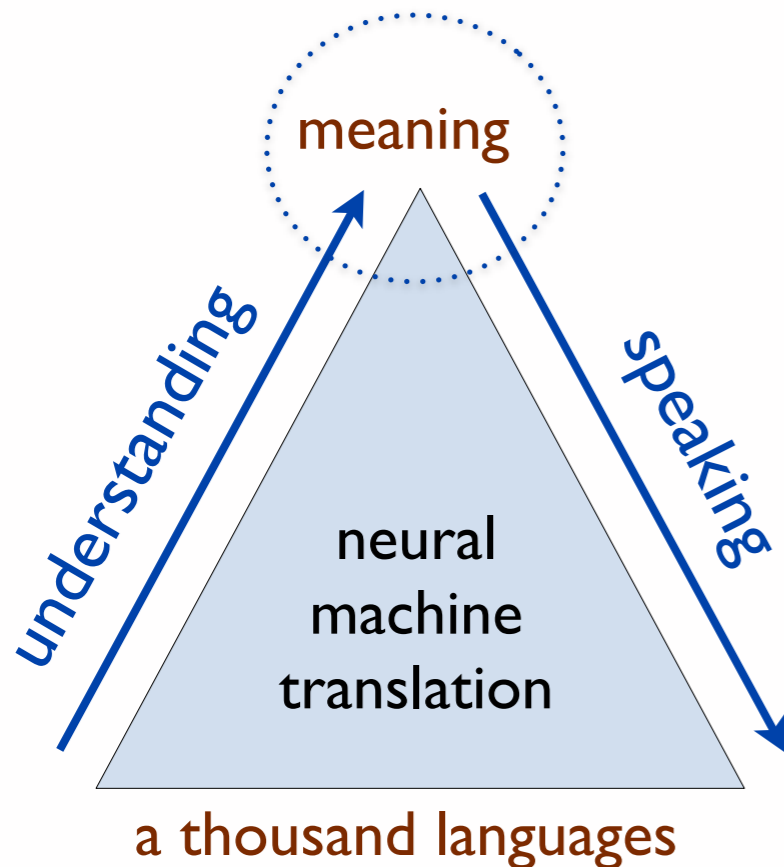




**MeMAD**

Methods for Managing  
Audiovisual Data

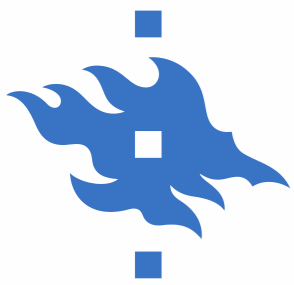
## ... and all that Jazz



**Jörg Tiedemann**  
*Department of Digital Humanities*  
*University of Helsinki*  
*[jorg.tiedemann@helsinki.fi](mailto:jorg.tiedemann@helsinki.fi)*

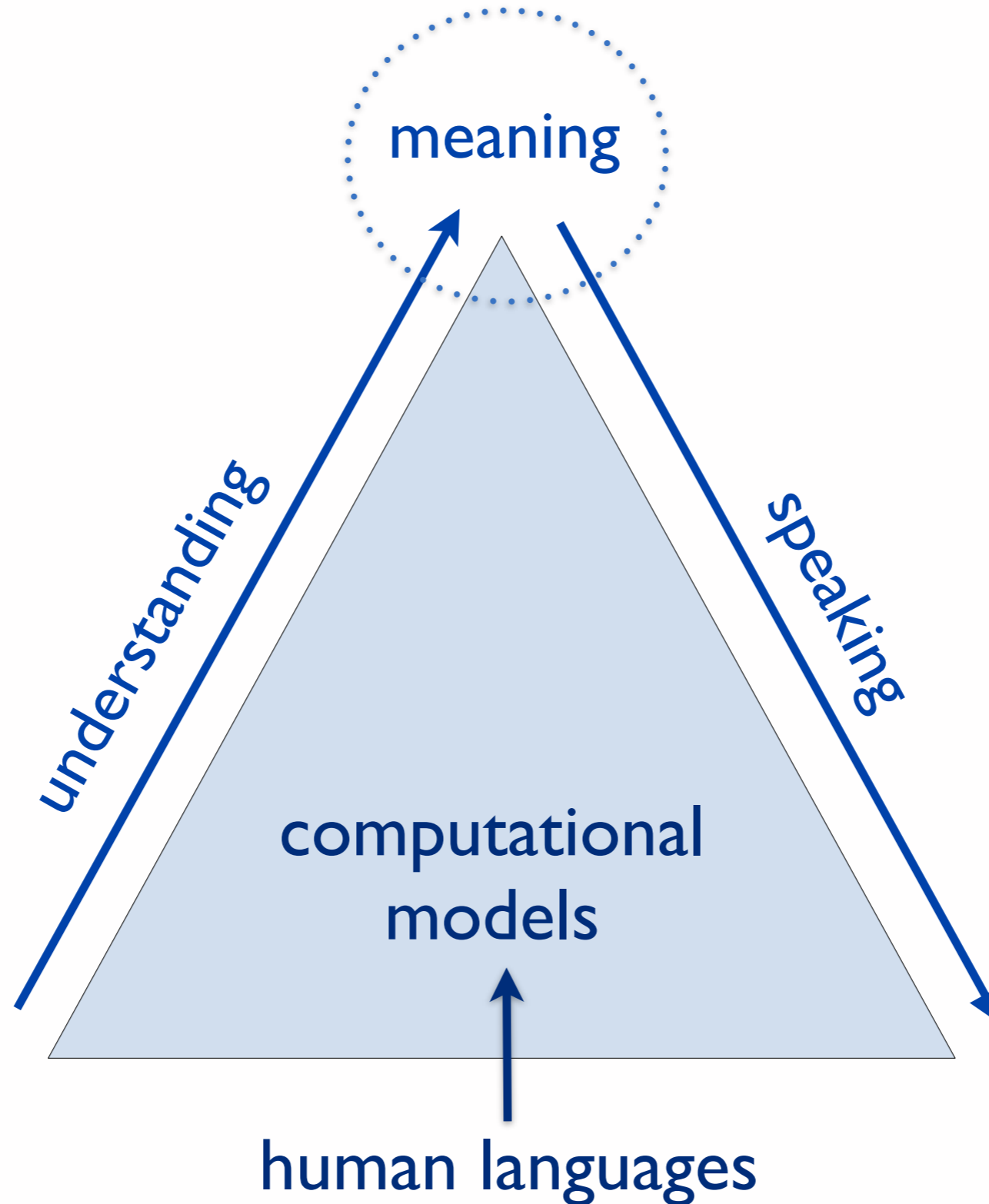
# Why are we here?

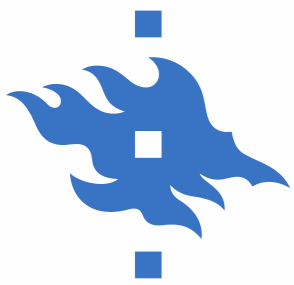




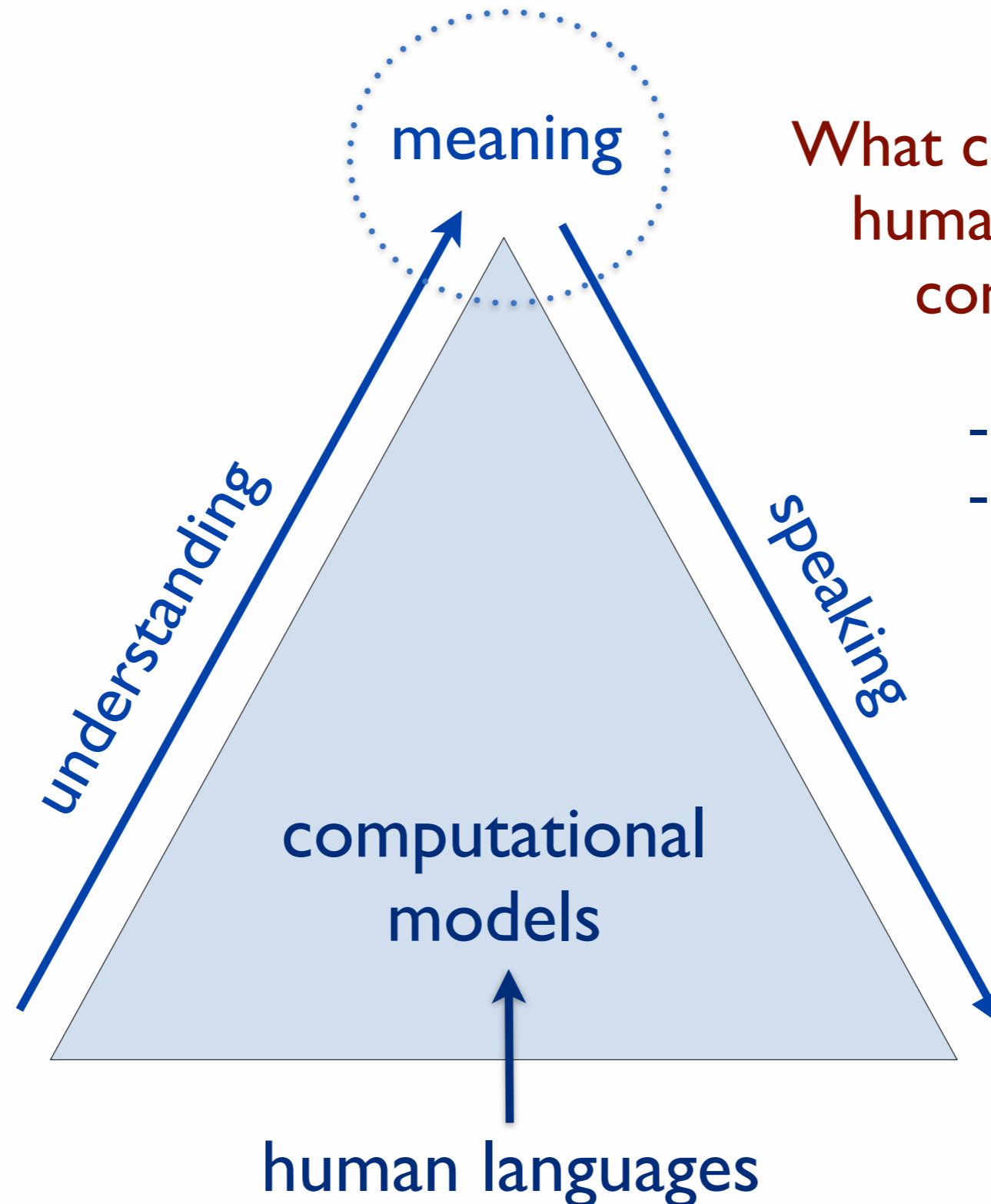
# Language Technology and AI

---



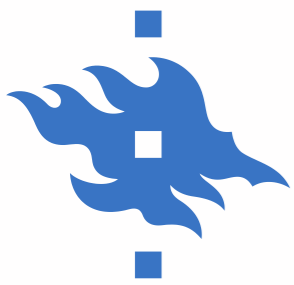


# Language Technology and AI



What can we learn about human languages and communication?

- linguistics
- cognitive science



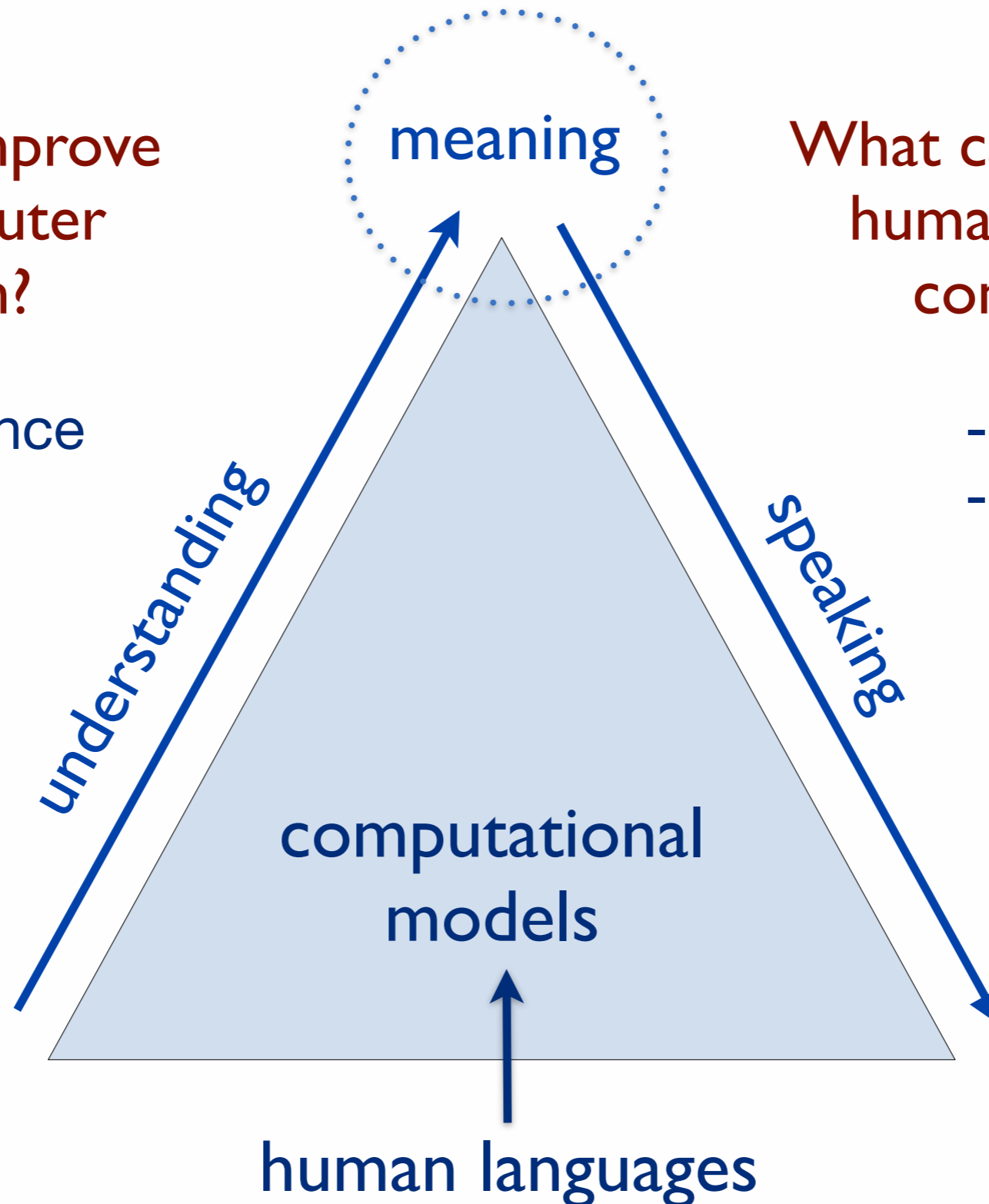
# Language Technology and AI

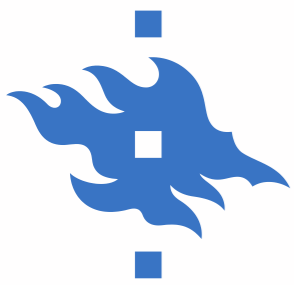
How can we improve human-computer interaction?

- computer science
- data science

What can we learn about human languages and communication?

- linguistics
- cognitive science



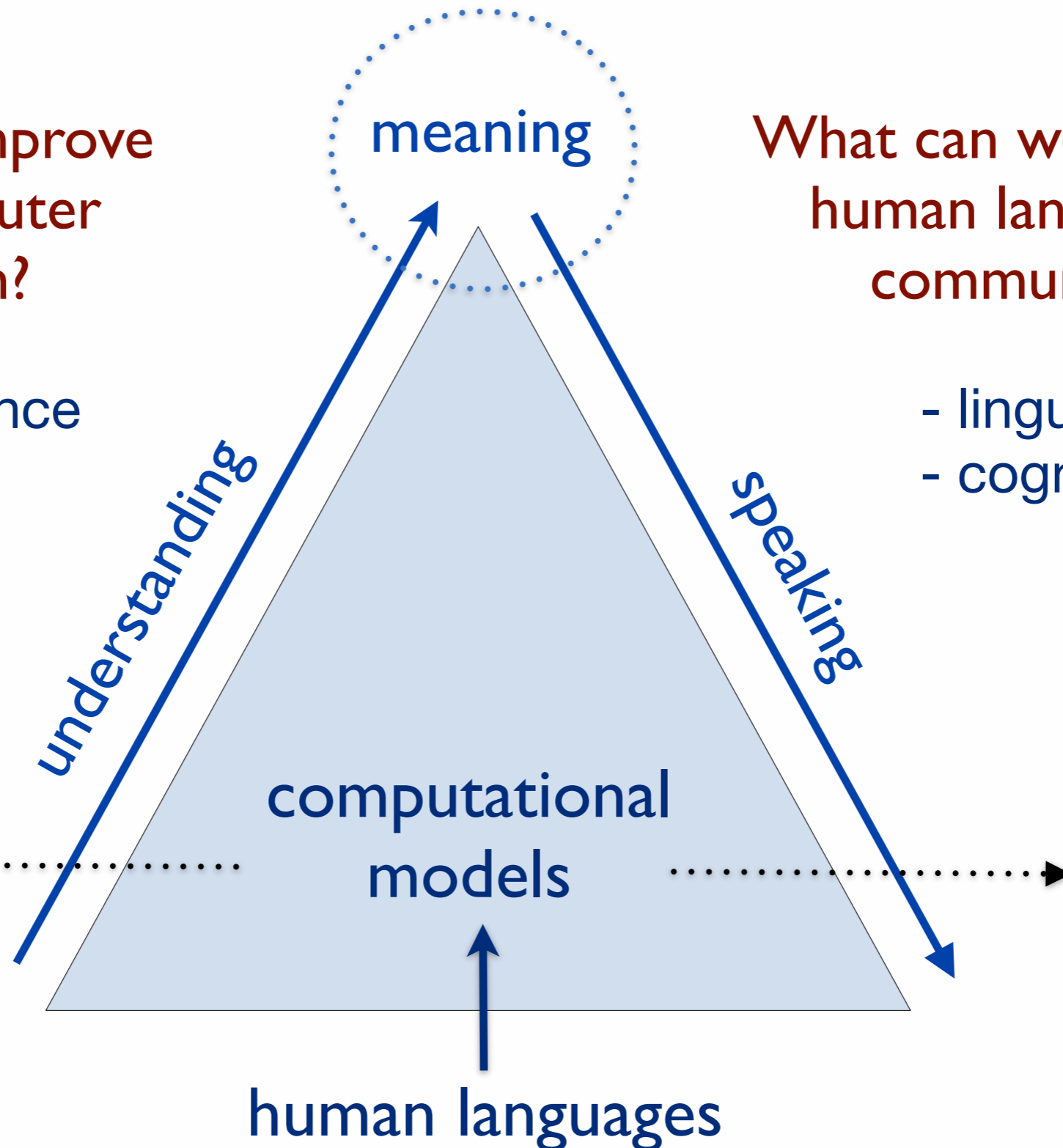


# Language Technology and AI

How can we improve human-computer interaction?

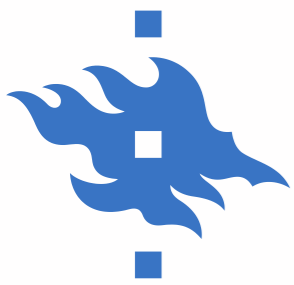
- computer science
- data science

tools & applications



What can we learn about human languages and communication?

- linguistics
- cognitive science



# Language Technology and AI

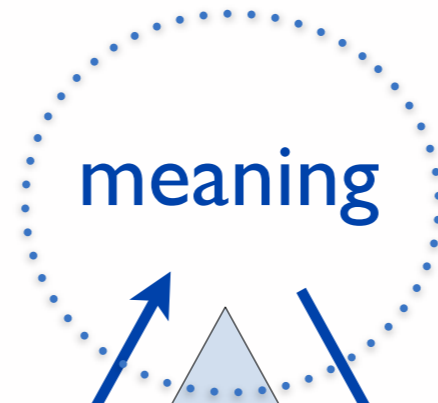
How can we improve human-computer interaction?

- computer science
- data science

tools & applications



**MeMAD**  
Methods for Managing  
Audiovisual Data



What can we learn about human languages and communication?

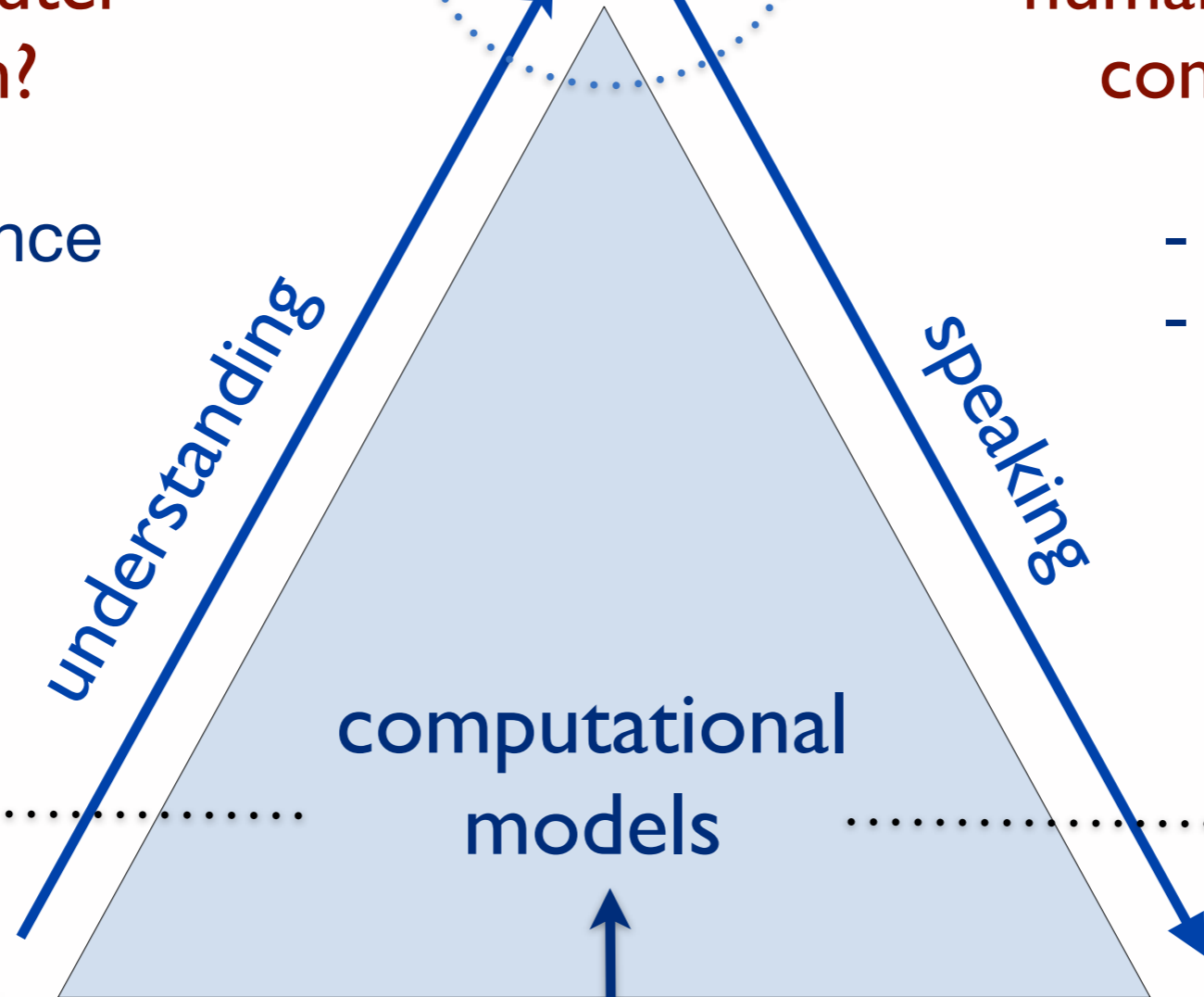
- linguistics
- cognitive science

academic research



**FOTRAN**  
Found in Translation

HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI



human languages



# MeMAD: Our Work Package

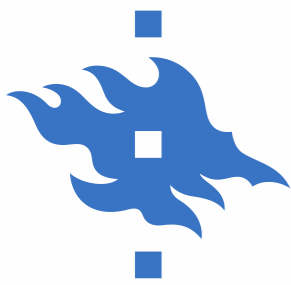


MeMAD

Methods for Managing  
Audiovisual Data

- Development of **multilingual** machine translation with **multimodal** input.
- Implementation and training of neural machine translation models with **non-symbolic interlingual representations** covering at least six EU languages, both minor and major.
- Development of **discourse-oriented machine translation** optimised for the dynamics of the narrative in the audiovisual data streams.
- Providing support for **cross-lingual content retrieval** based on automatic content analysis.





# MeMAD: Our Work Package



MeMAD

Methods for Managing  
Audiovisual Data

- Development of **multilingual** machine translation with **multimodal** input.
- Implementation and training of neural machine translation models with **non-symbolic interlingual representations** covering at least six EU languages, both minor and major.
- Development of **discourse-oriented machine translation** optimised for the dynamics of the narrative in the audiovisual data streams.
- Providing support for **cross-lingual content retrieval** based on automatic content analysis.

**We must be mad!**



# What Have We Done So Far?

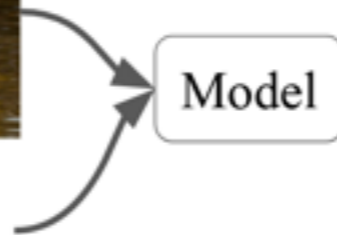


**MeMAD**

Methods for Managing  
Audiovisual Data



A bird flies  
over the water



Ein Vogel fliegt  
über das Wasser

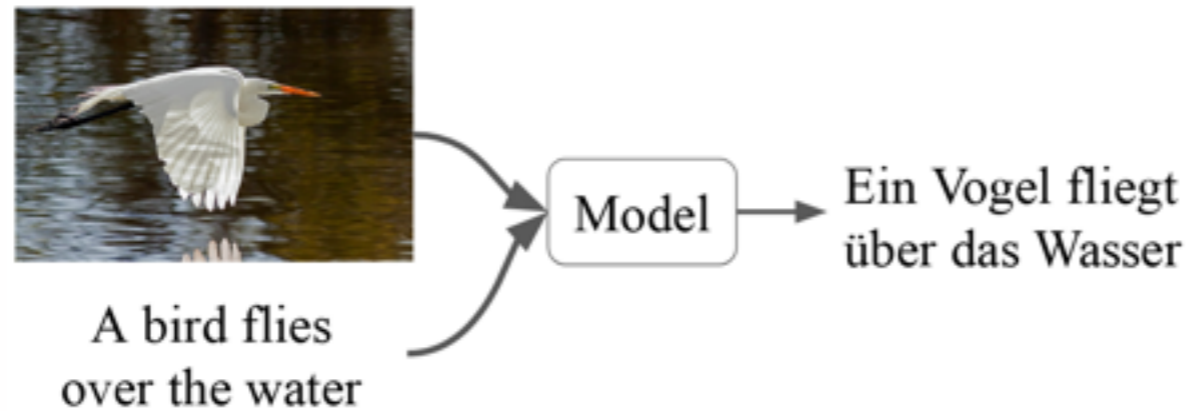


# What Have We Done So Far?



MeMAD

Methods for Managing  
Audiovisual Data



## WMT 2018 multimodal translation task

- best system for English - French (+ 3.5 BLEU)
- best system for English - German (+ 6.0 BLEU)
- **Wow!** ... but is the system really multimodel?

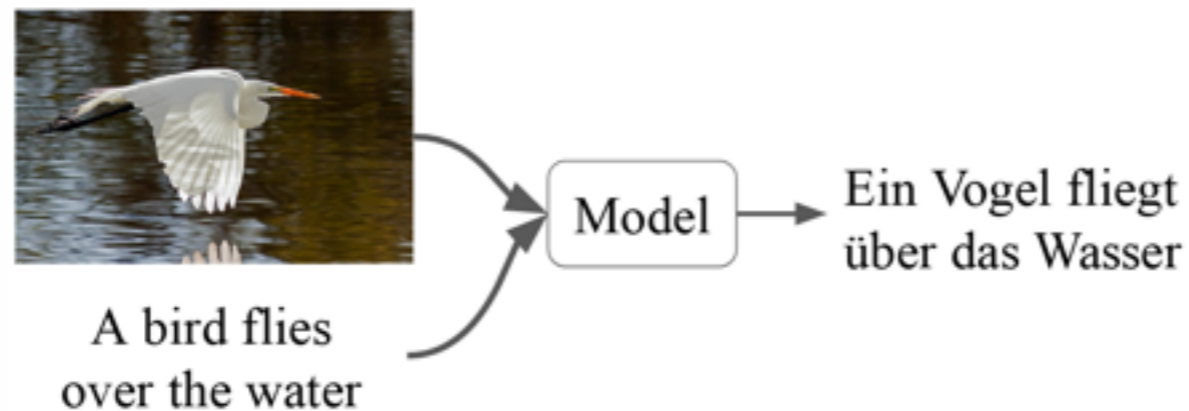


# What Have We Done So Far?



MeMAD

Methods for Managing  
Audiovisual Data



## WMT 2018 multimodal translation task

- best system for English - French (+ 3.5 BLEU)
- best system for English - German (+ 6.0 BLEU)
- **Wow!** ... but is the system really multimodal?

## IWSLT 2018 speech-to-text translation

- pipeline approach: ASR + text-based MT
- end-to-end model not yet useful



# Translate English to ASR-English

---

**Original:** Because in the summer of 2006, the E.U. Commission tabled a directive.

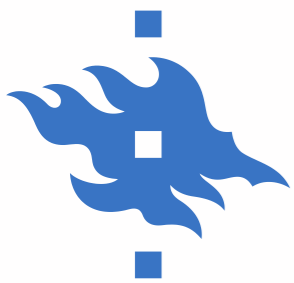
**ASR-like:** because in the summer of two thousand and six you commission tabled a directive

**Original:** I'm a child of 1984,

**ASR-like:** i am a child of nineteen eighty four

**Original:** Stasi was the secret police in East Germany.

**ASR-like:** stars he was the secret police in east germany



# Language Grounding

---

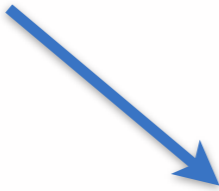


- En:** A *wall* divided the city.
- De 1:** Eine *Wand* teilte die Stadt. ×
- De 2:** Eine *Mauer* teilte die Stadt. ✓

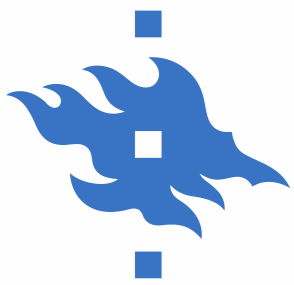


# Language Grounding

visual  
grounding



- En:** *A wall* divided the city.  
**De 1:** *Eine Wand* teilte die Stadt. ×  
**De 2:** *Eine Mauer* teilte die Stadt. ✓

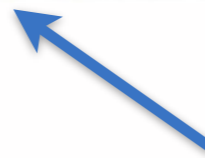


# Language Grounding

visual  
grounding

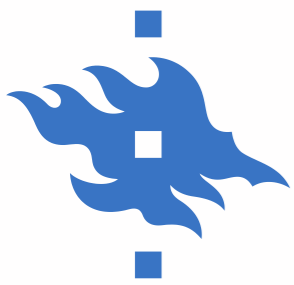


- En:** *A wall* divided the city.  
**De 1:** Eine *Wand* teilte die Stadt. ×  
**De 2:** Eine *Mauer* teilte die Stadt. ✓



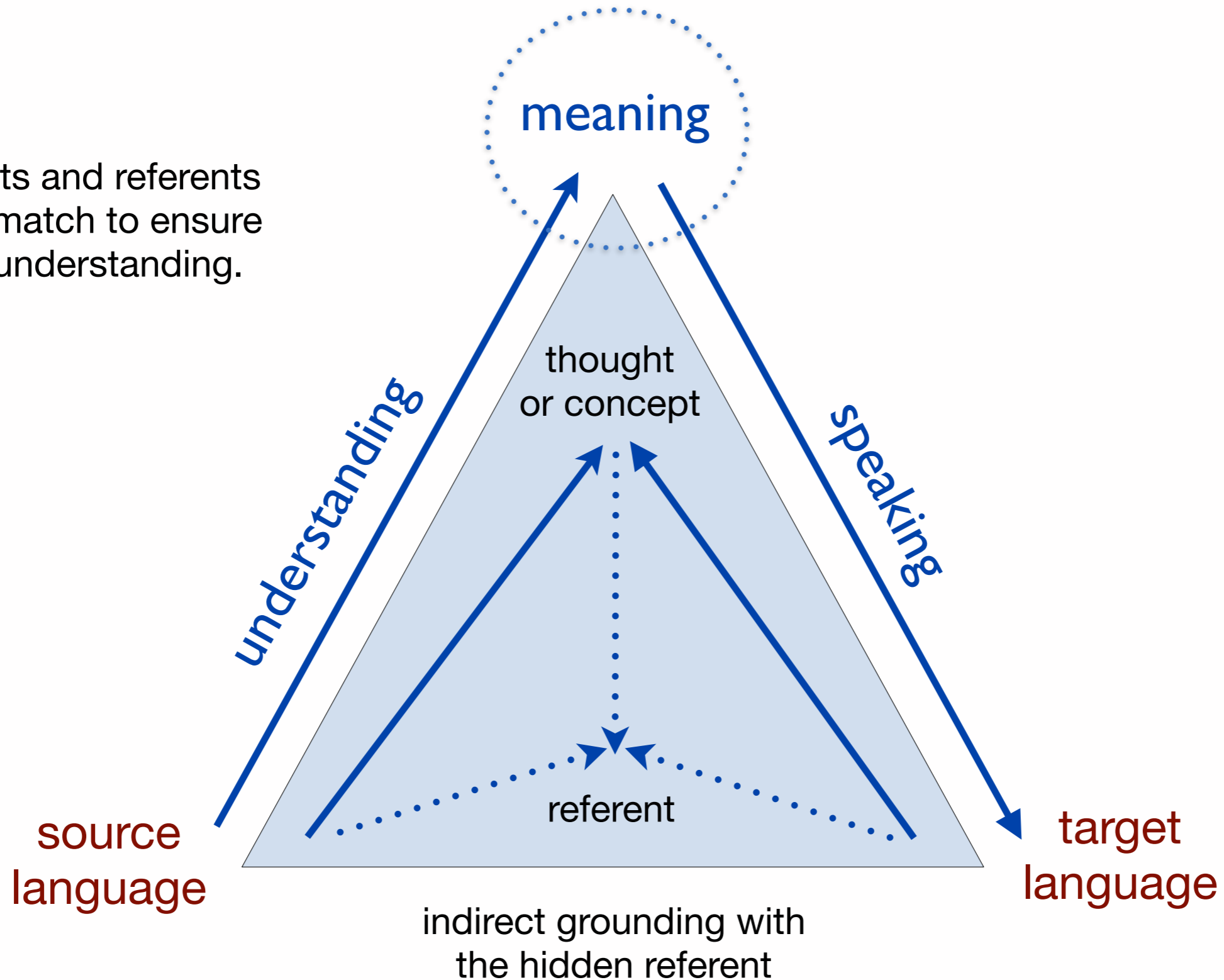
translational  
grounding





# MT and Meaning Representations

Concepts and referents should match to ensure mutual understanding.

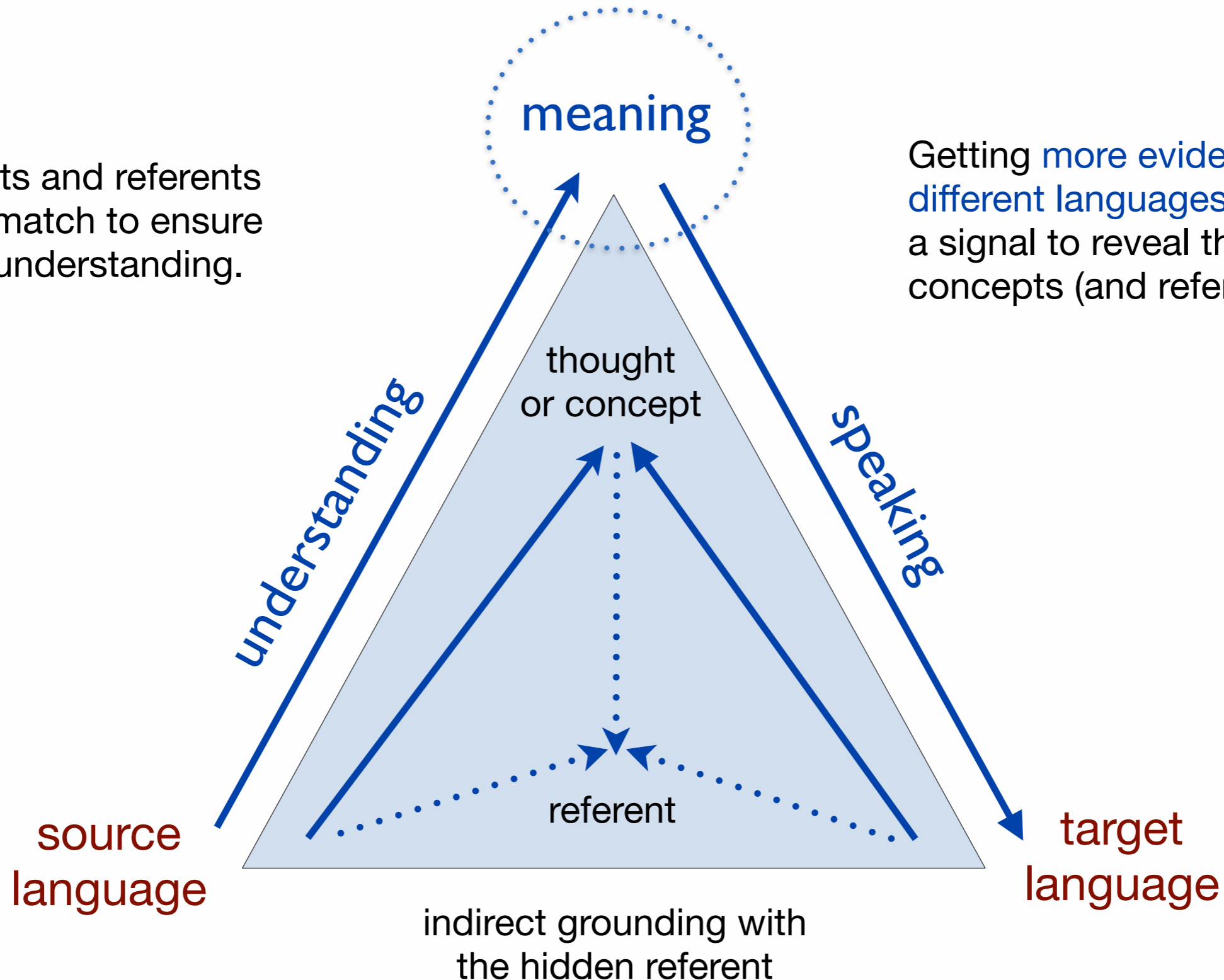




# MT and Meaning Representations

Concepts and referents should match to ensure mutual understanding.

Getting more evidence from different languages provides a signal to reveal the hidden concepts (and referents).





# Bitexts as Semantic Mirrors

---

Bitexts as Semantic Mirrors

## Research Question:

Can we use **translations** as **implicit supervision** for learning abstract meaning representations?

## Hypothesis:

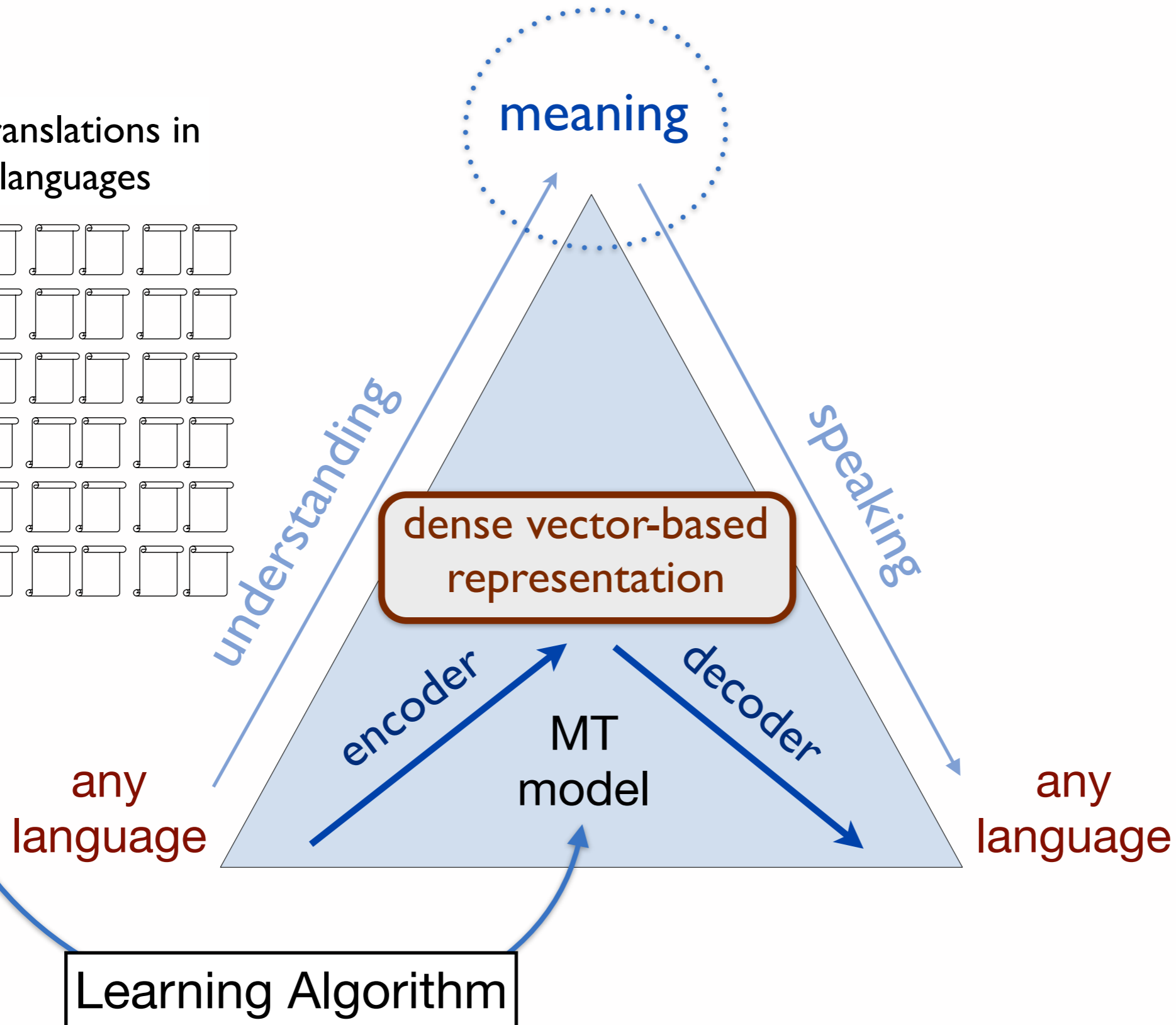
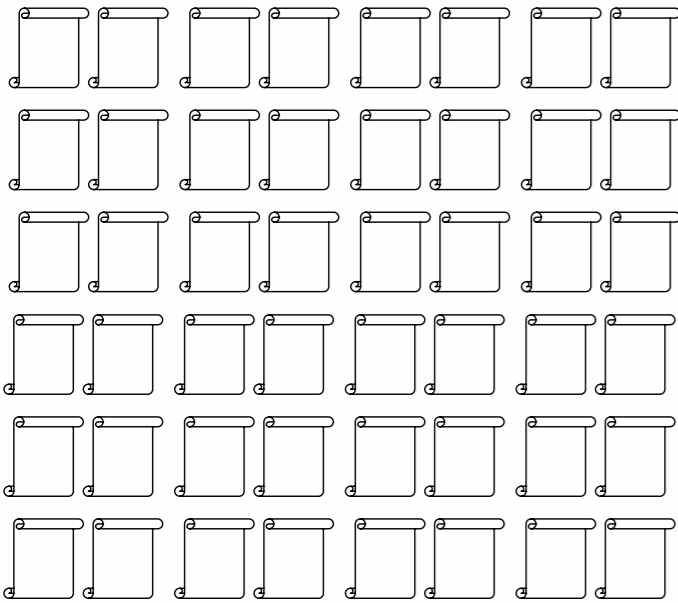
Increasing **linguistic diversity**  
improves **abstraction**





# Multilingual Translation Models

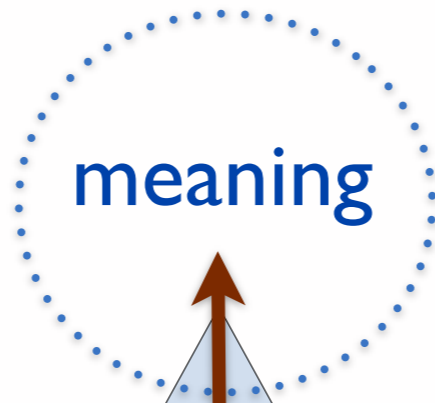
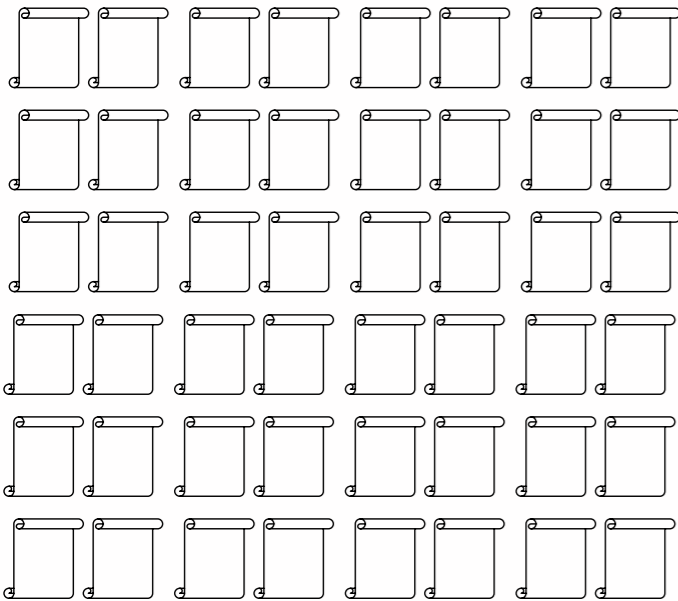
human translations in many languages



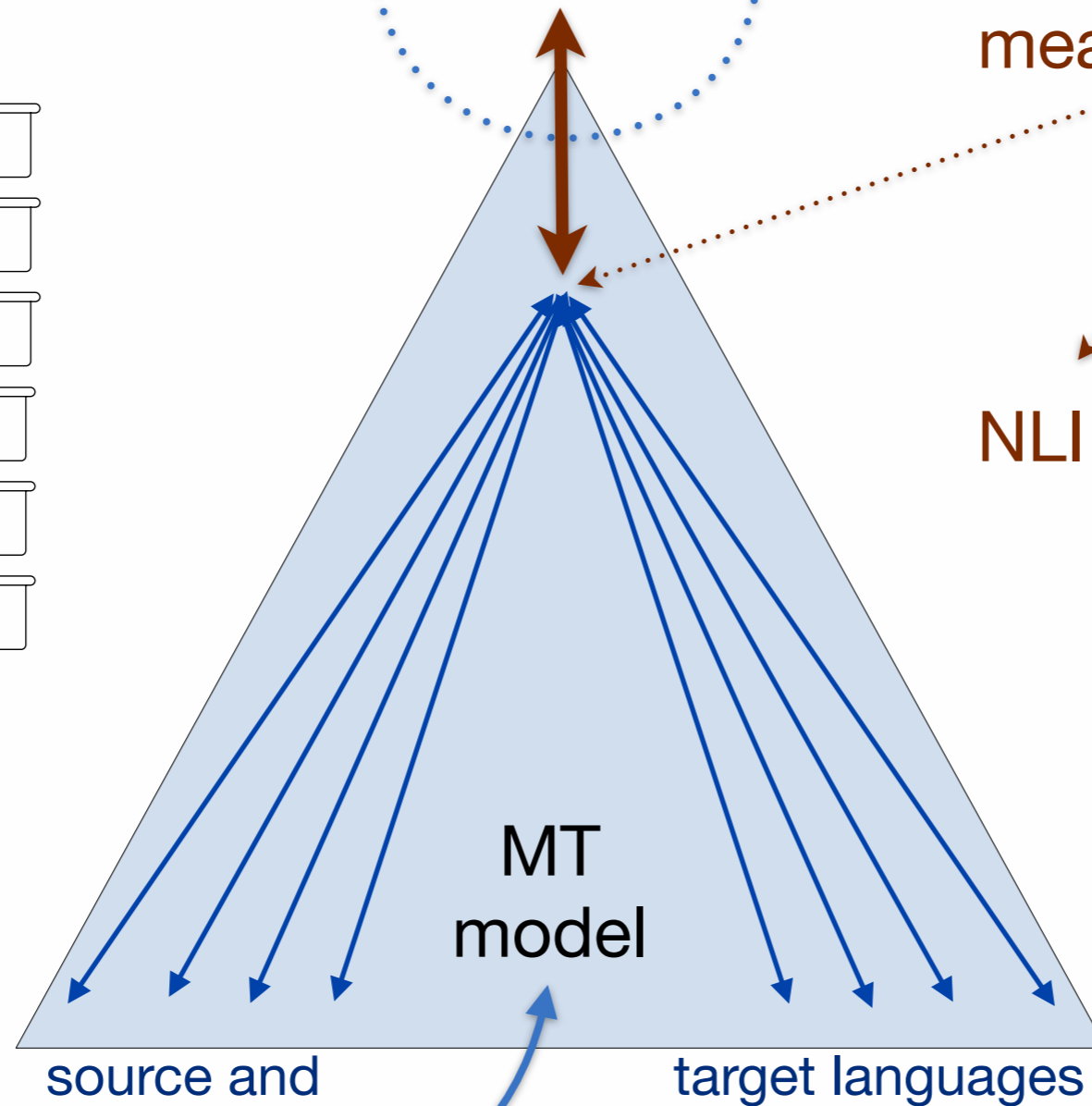
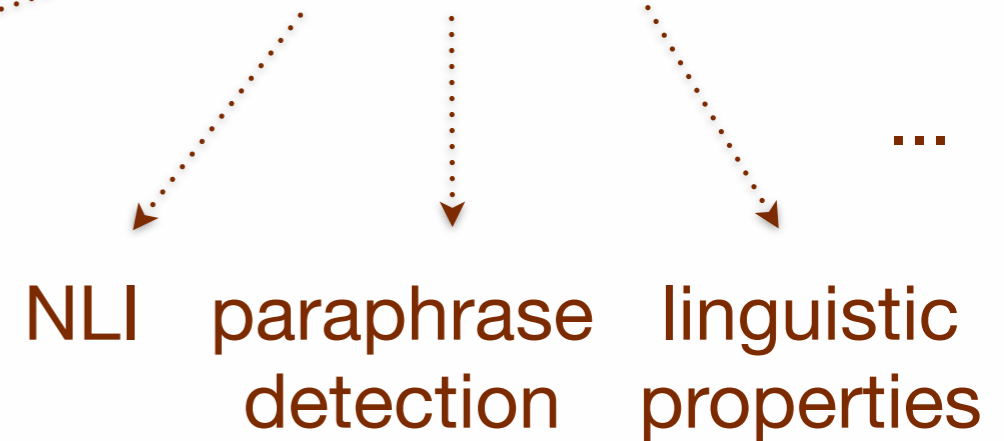


# Multilingual Translation Models

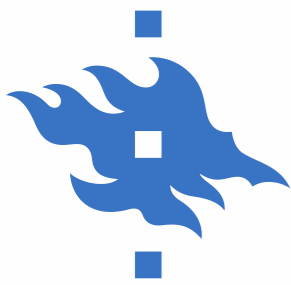
human translations in many languages



**LIMR:** language-independent meaning representation?



Learning Algorithm



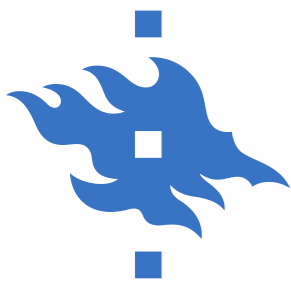
# Found in Translation

Natural Language Understanding with cross-lingual grounding



sub-project 1:  
modeling/development

What is the  
best model and how  
can we optimize  
learning?



# Found in Translation

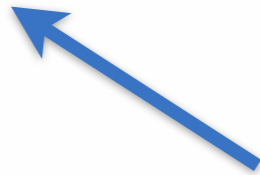
Natural Language Understanding with cross-lingual grounding

sub-project 2:  
interpretation

What do the  
representations  
cover and how?

sub-project 1:  
modeling/development

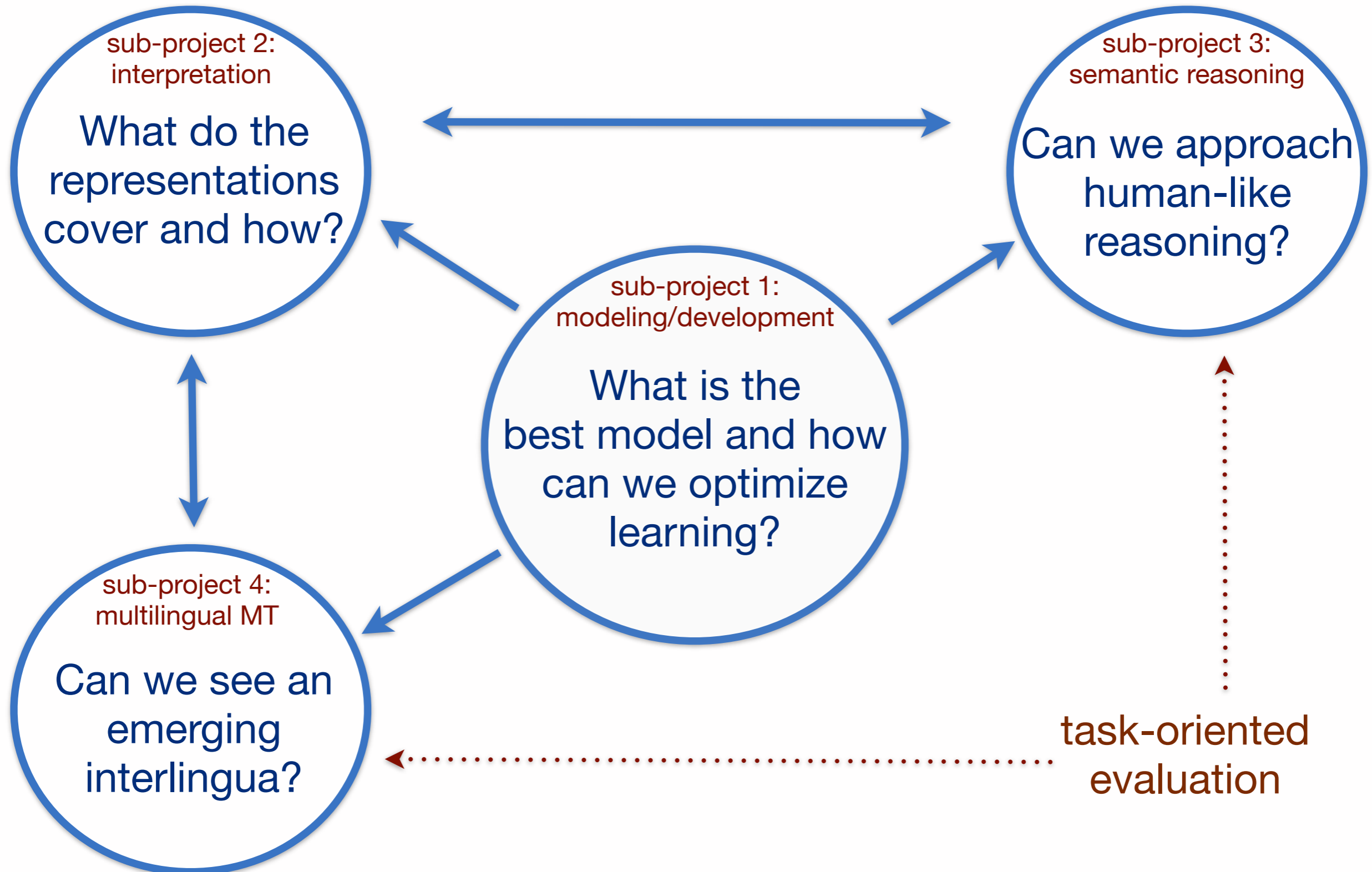
What is the  
best model and how  
can we optimize  
learning?





# Found in Translation

## Natural Language Understanding with cross-lingual grounding

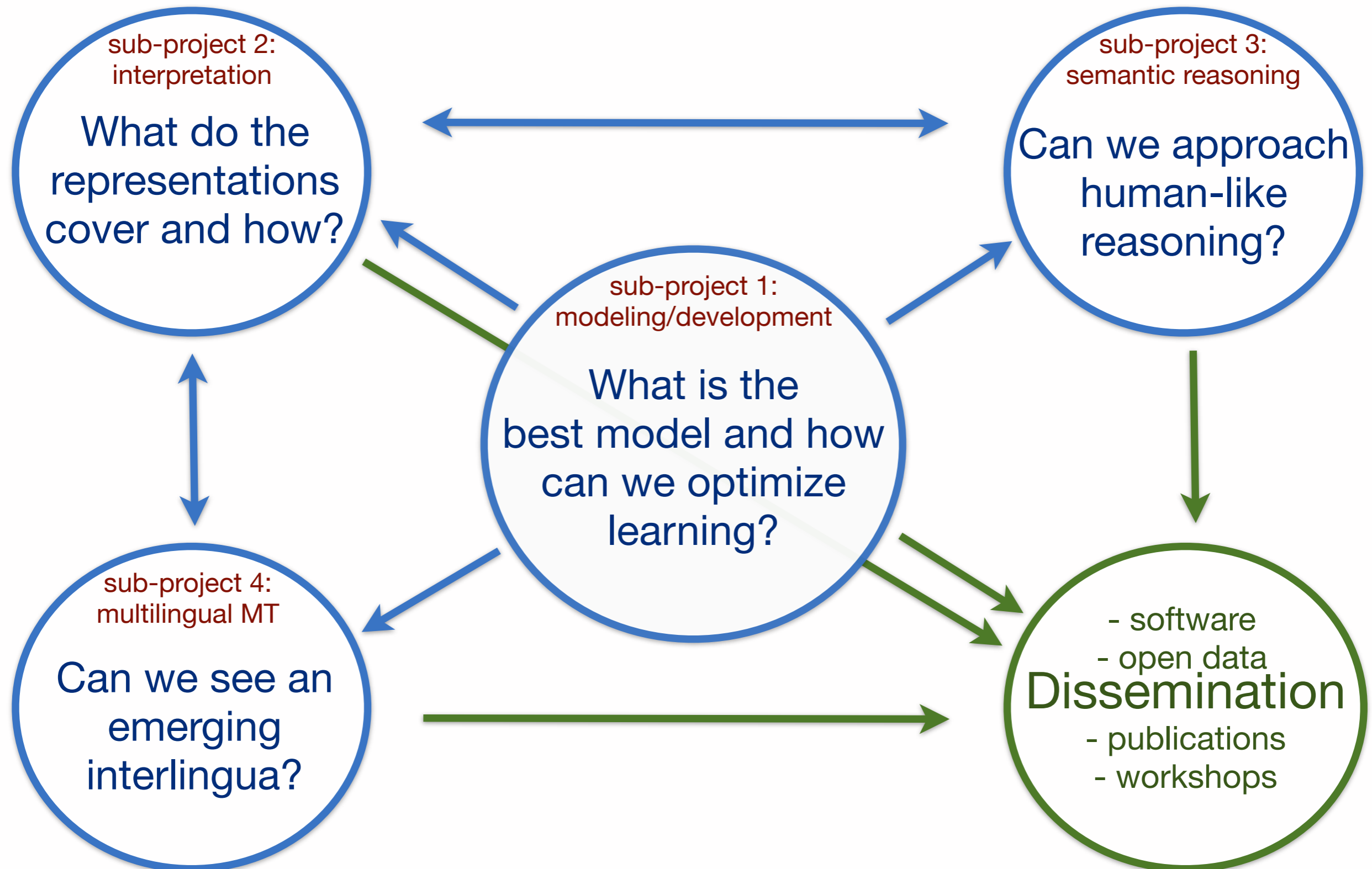






# Found in Translation

## Natural Language Understanding with cross-lingual grounding





# Found in Translation

Natural Language Understanding with cross-lingual grounding



Goals:





# Found in Translation

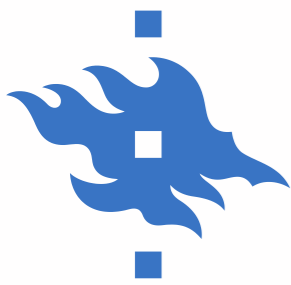
Natural Language Understanding with cross-lingual grounding



Goals:



**multilingual machine translation**

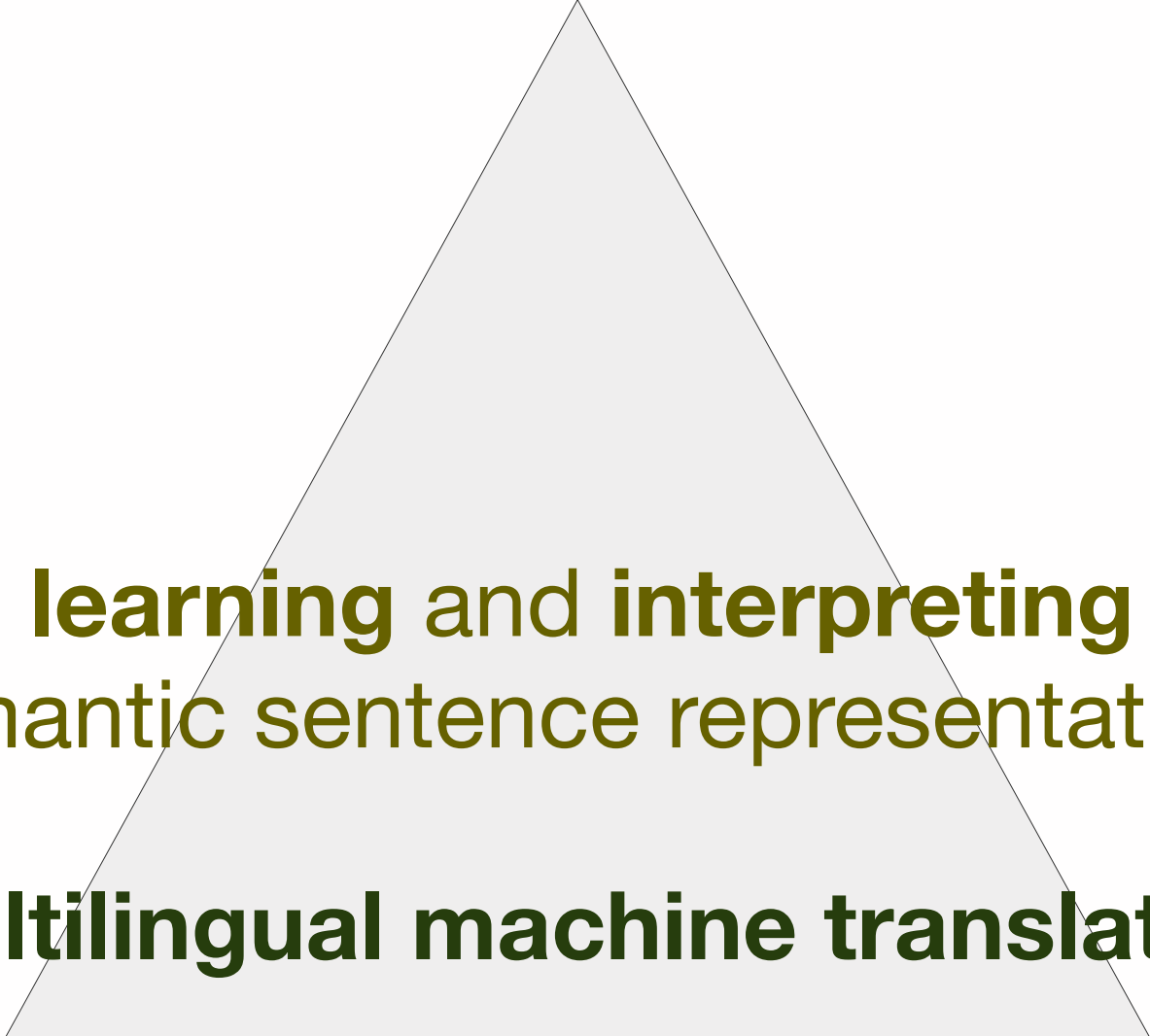


# Found in Translation

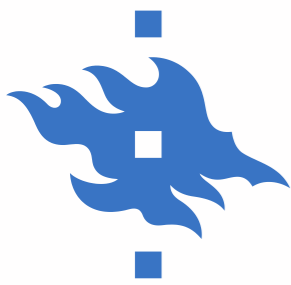
Natural Language Understanding with cross-lingual grounding



Goals:



**learning and interpreting**  
semantic sentence representations  
**multilingual machine translation**



# Found in Translation

Natural Language Understanding with cross-lingual grounding

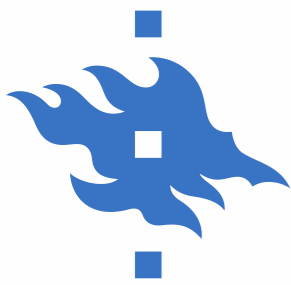


## Goals:

abstract continuous  
**meaning spaces of language**

**learning and interpreting**  
semantic sentence representations

**multilingual machine translation**

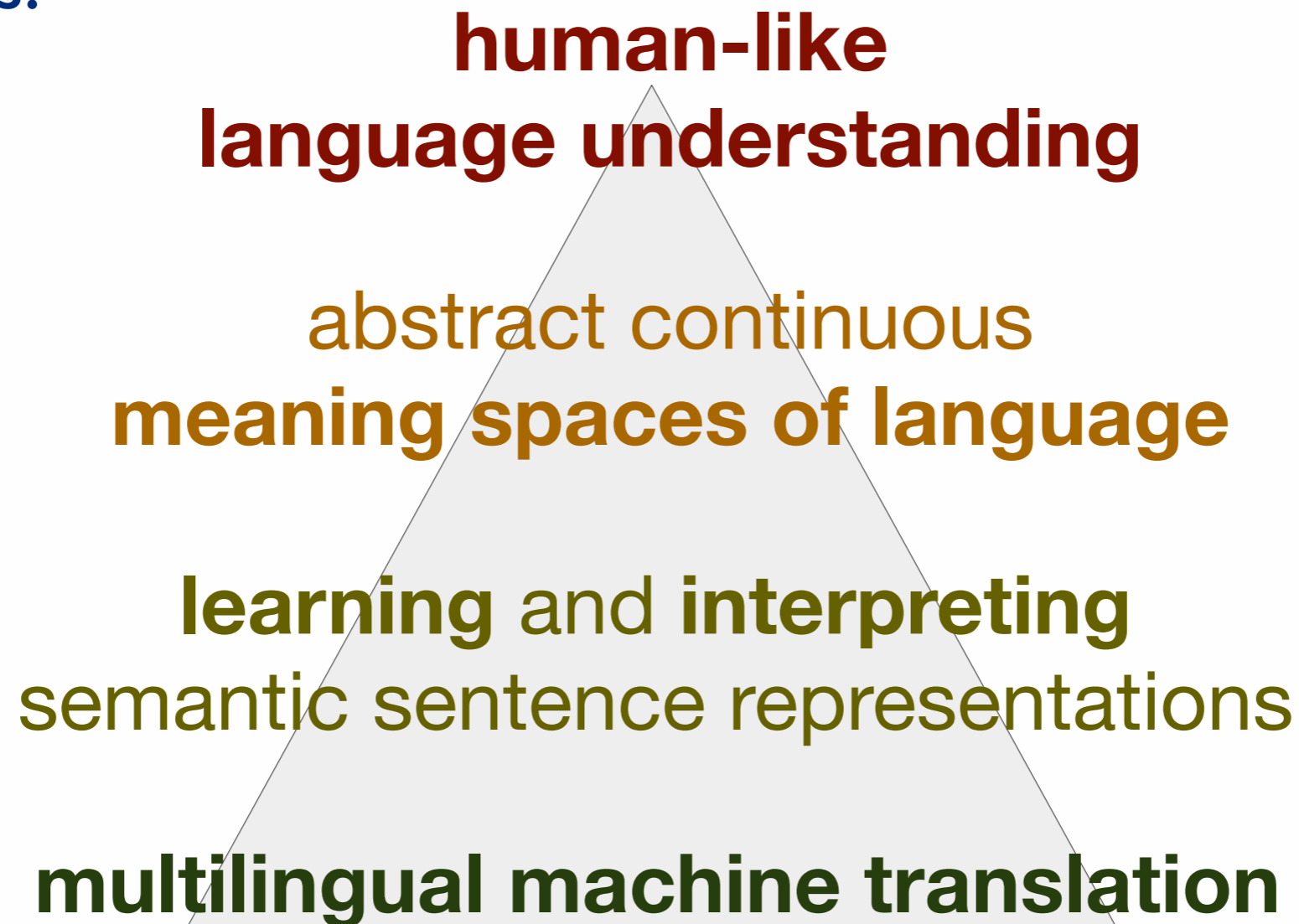


# Found in Translation

Natural Language Understanding with cross-lingual grounding

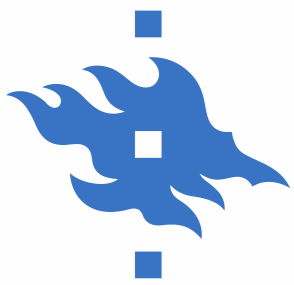


Goals:



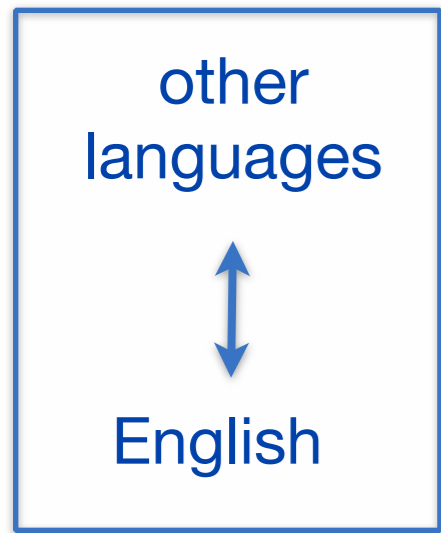
**... is there more time?**



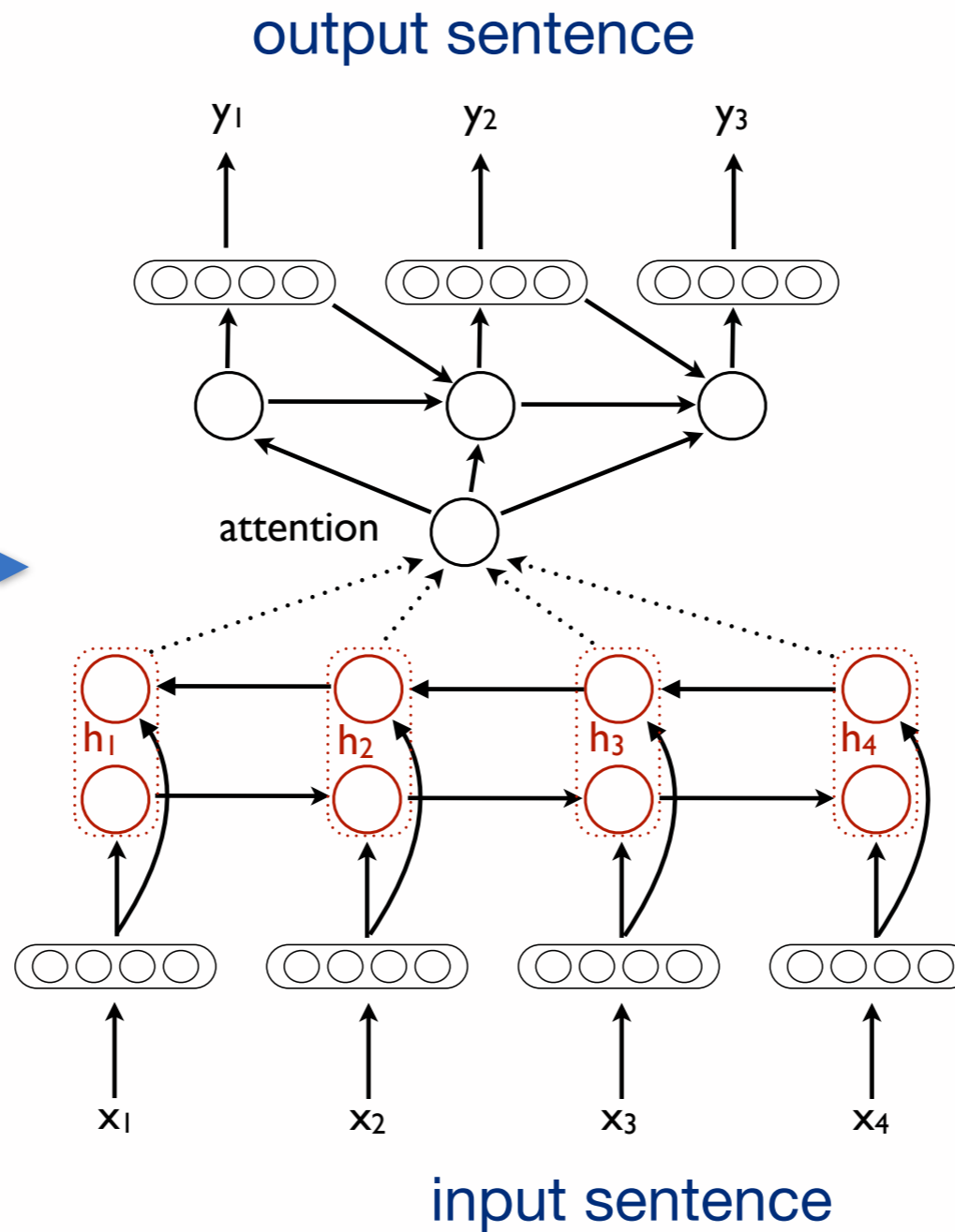


# Emerging Language Spaces

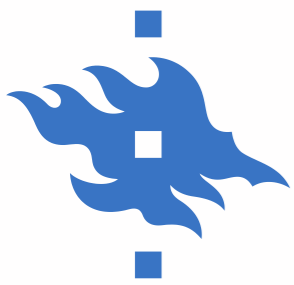
translations in  
> 900 languages



train

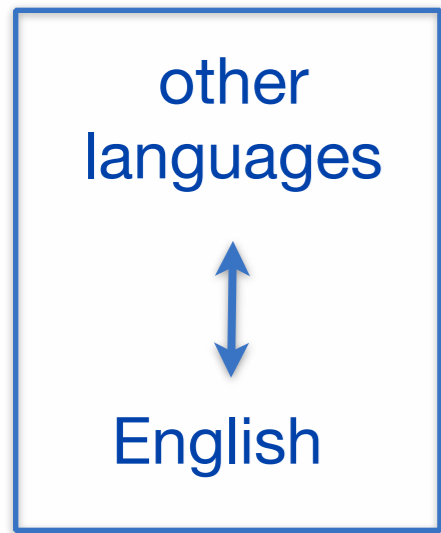




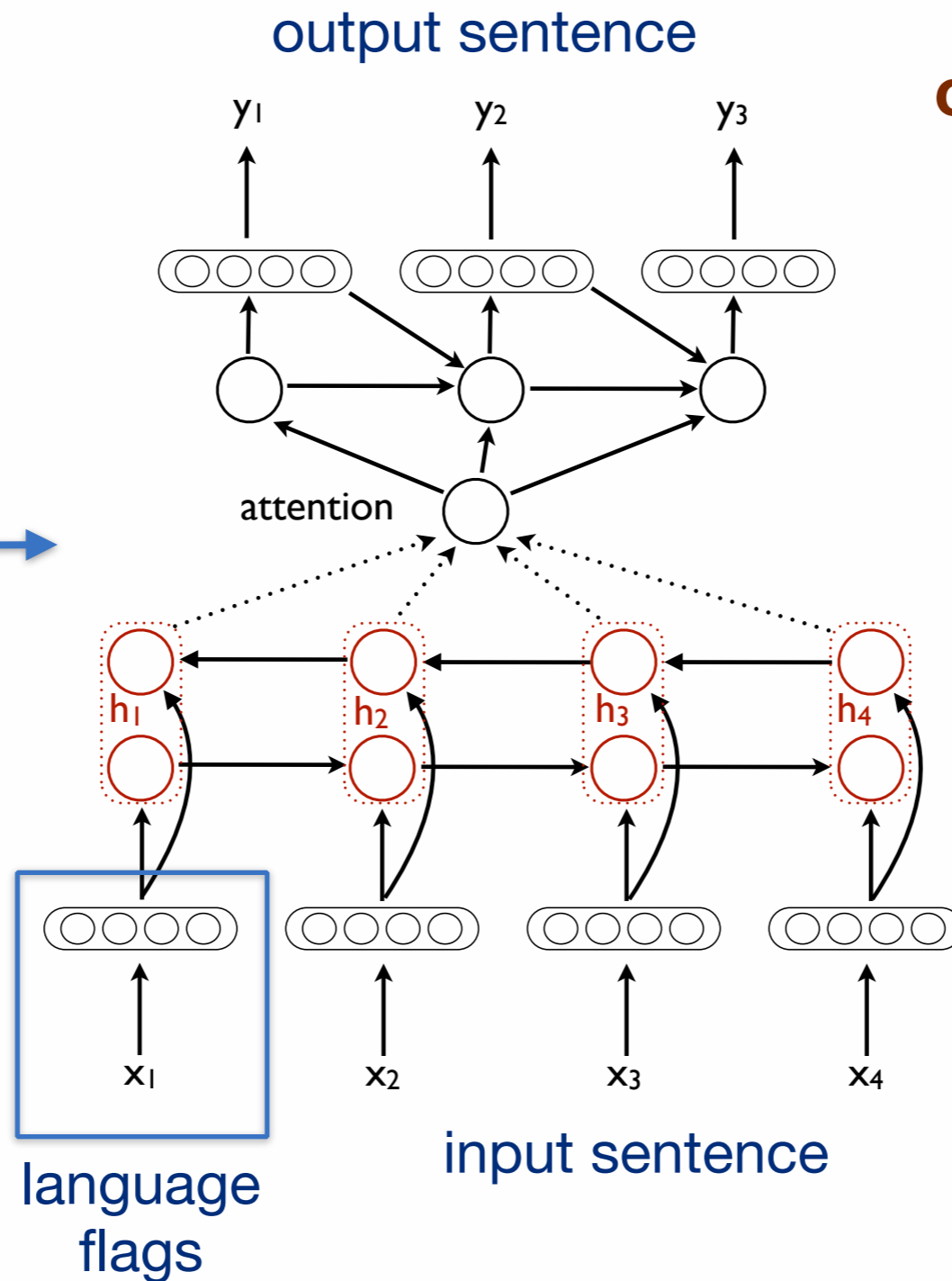


# Emerging Language Spaces

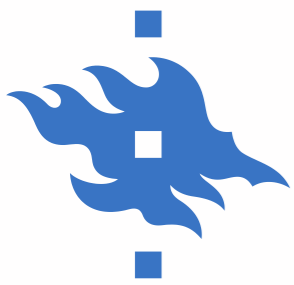
translations in  
> 900 languages



train

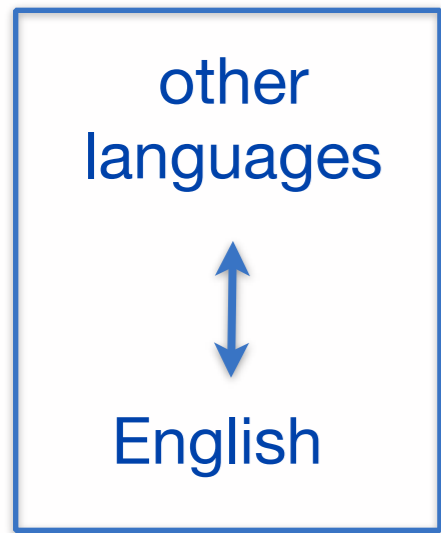


The network needs to **compress** information!



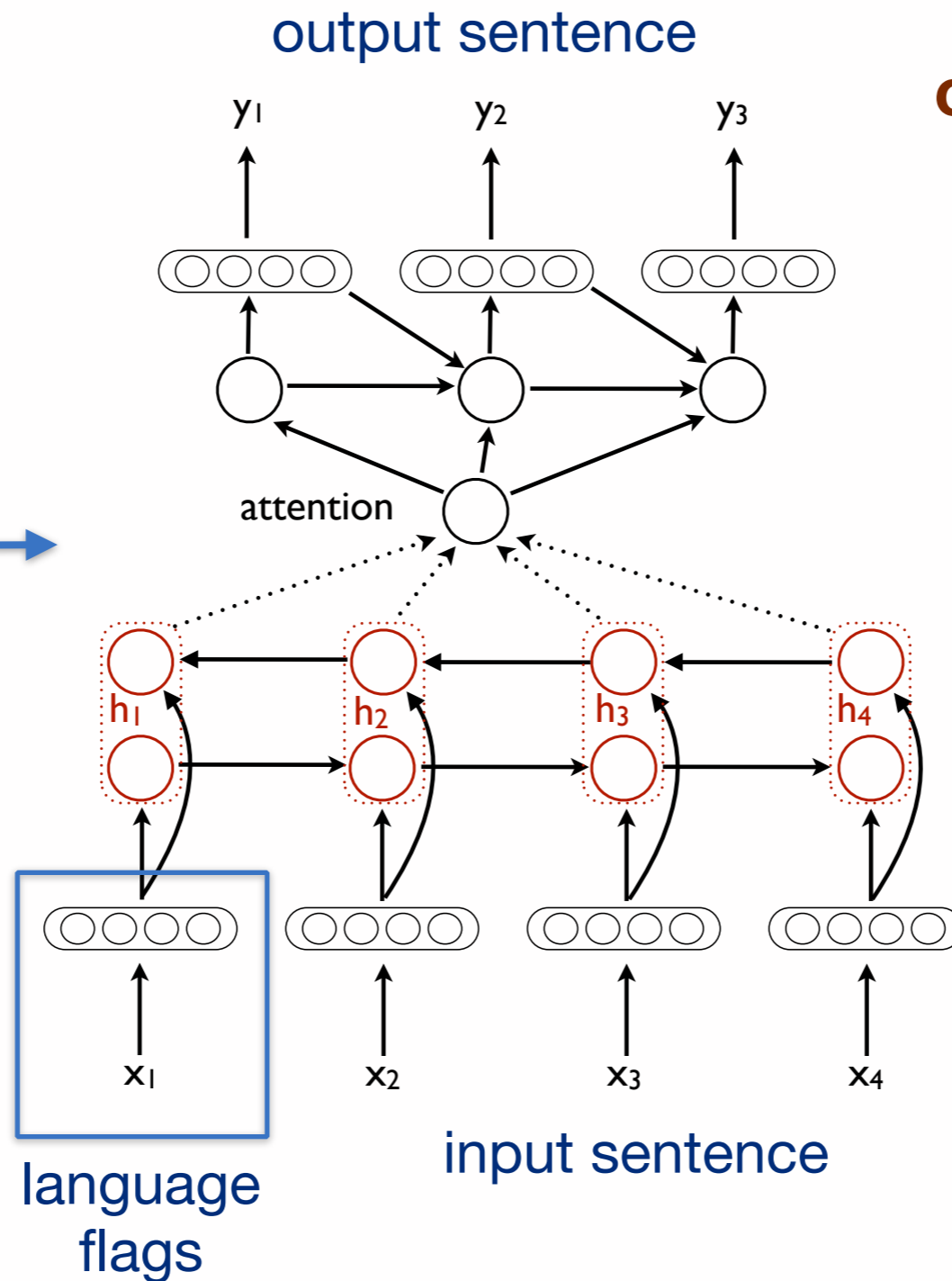
# Emerging Language Spaces

translations in  
> 900 languages



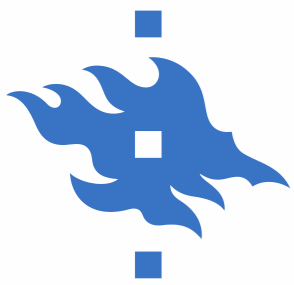
train

vector space  
of language  
representations

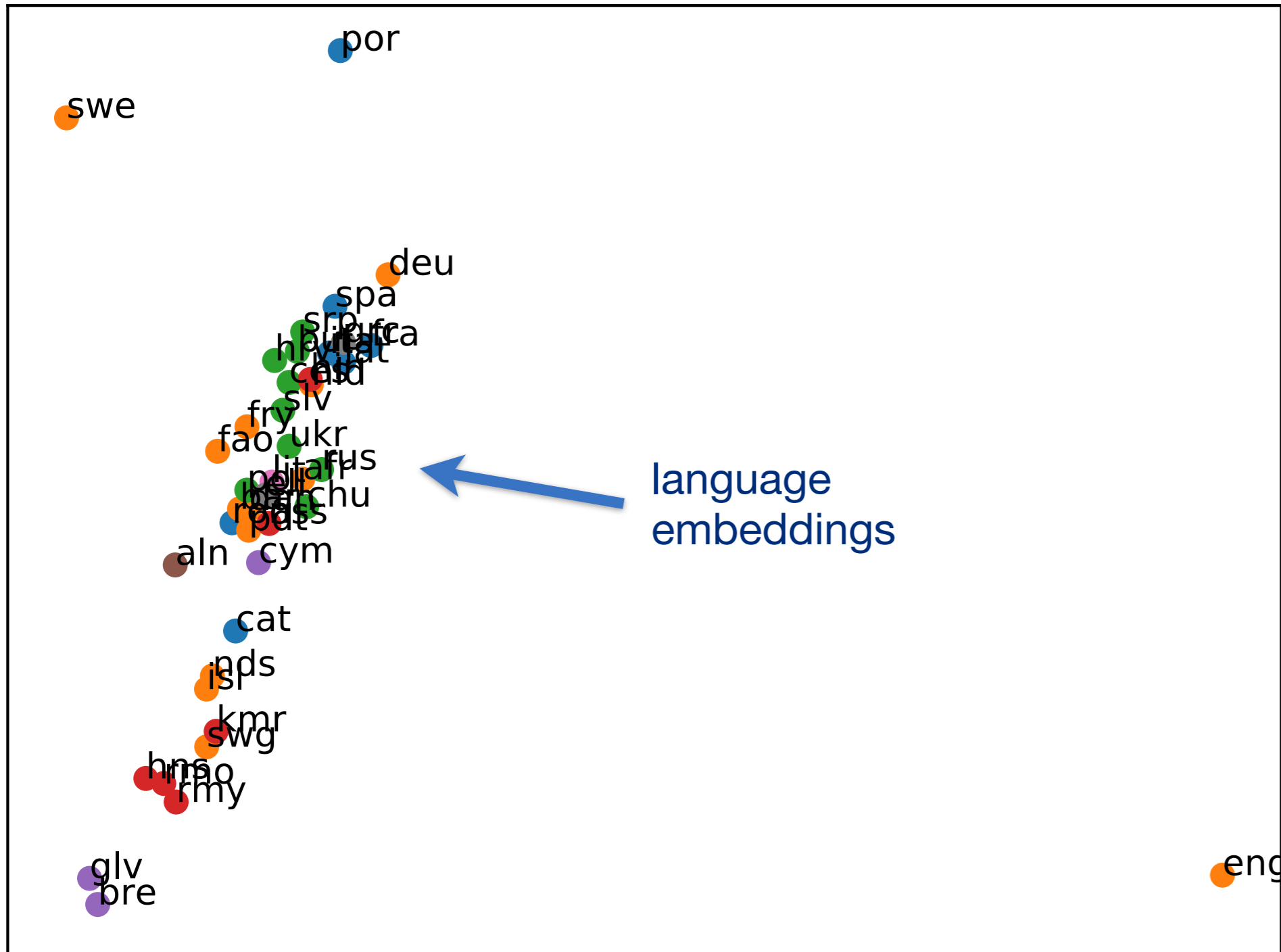


The network needs to  
**compress** information!

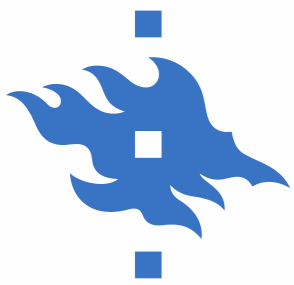
Learns to  
**re-use parameters**  
for languages with  
**common properties**



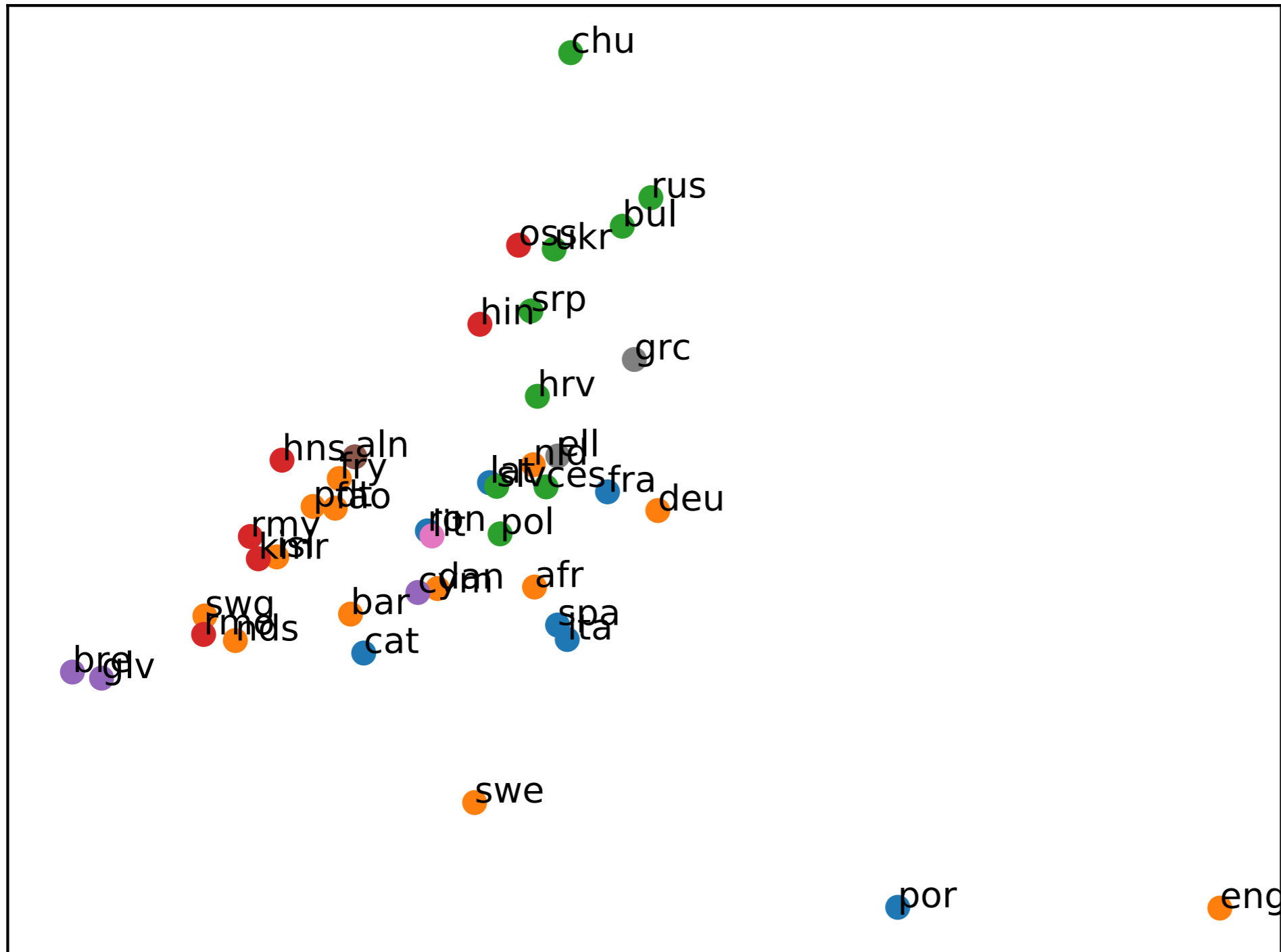
# Training with Indo-European Languages



(PCA)

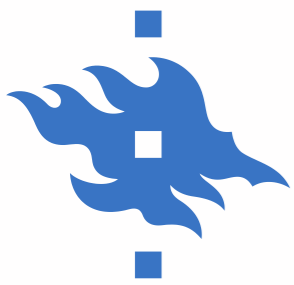


# Training with Indo-European Languages

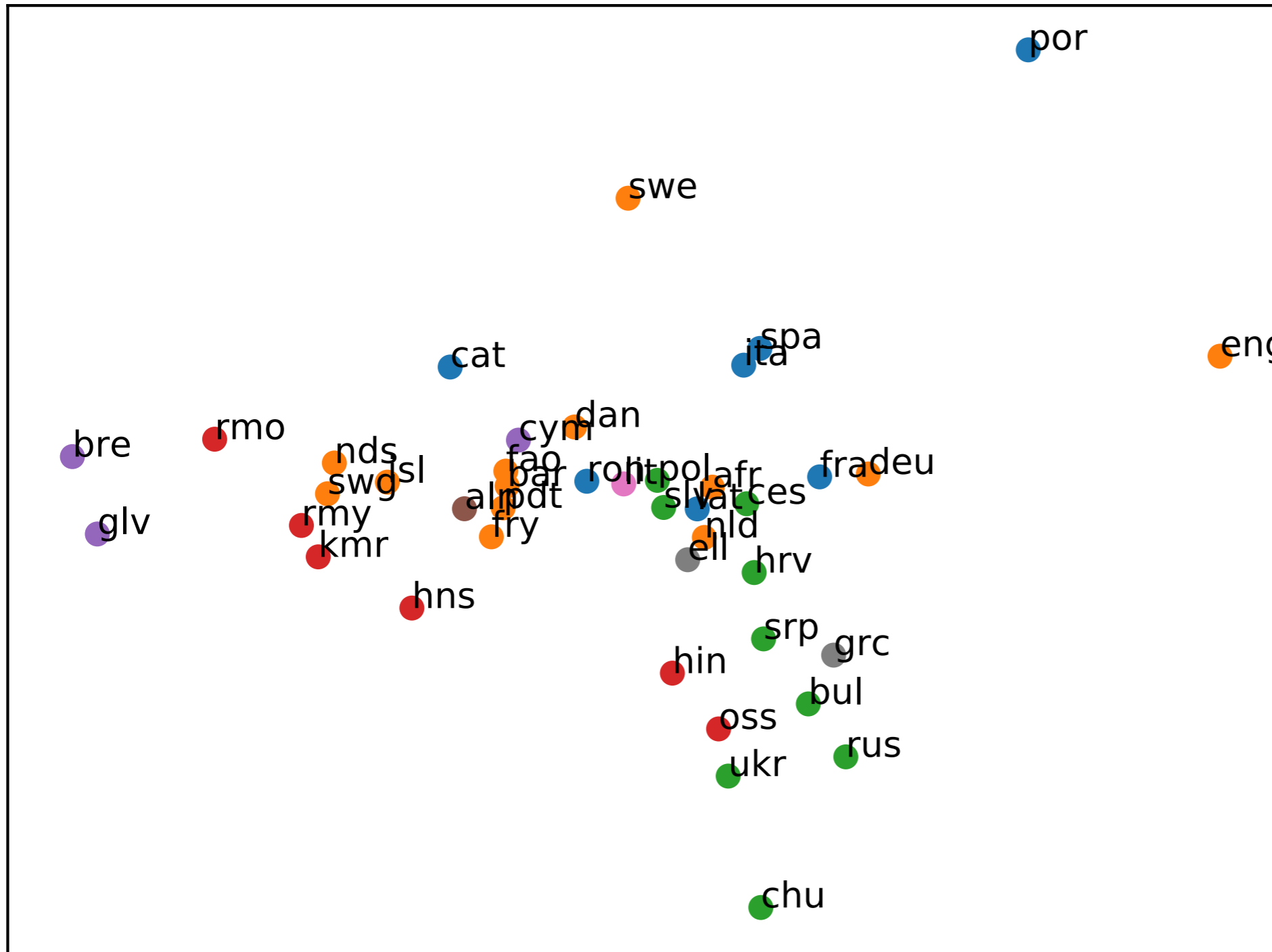


- Italic
- Germanic
- Slavic
- Indo-Iranian
- Celtic
- Albanian
- Baltic
- Greek

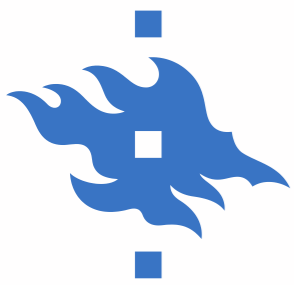
(PCA)



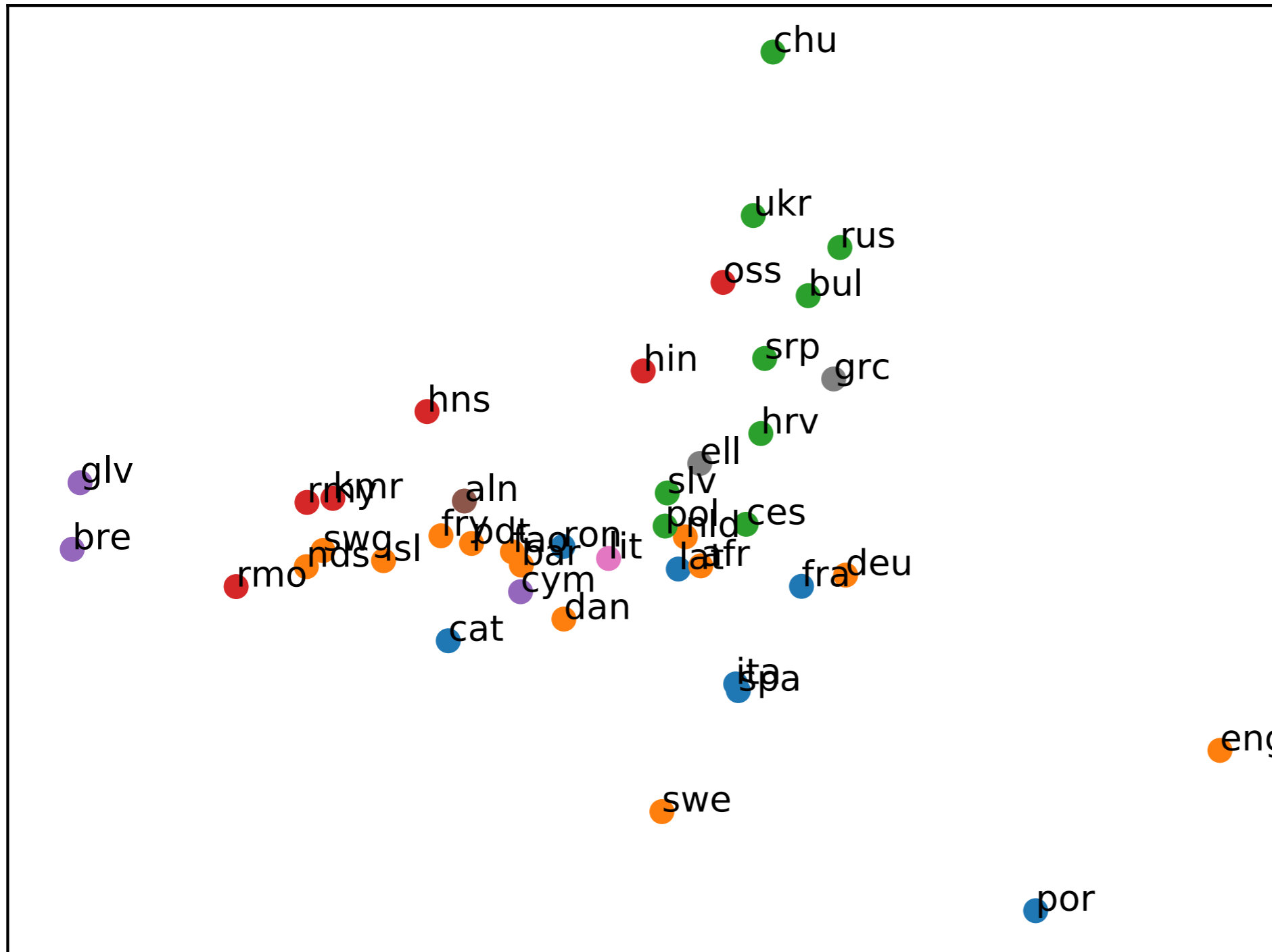
# Training with Indo-European Languages



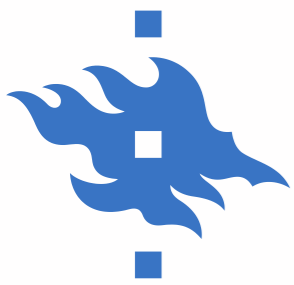
(PCA)



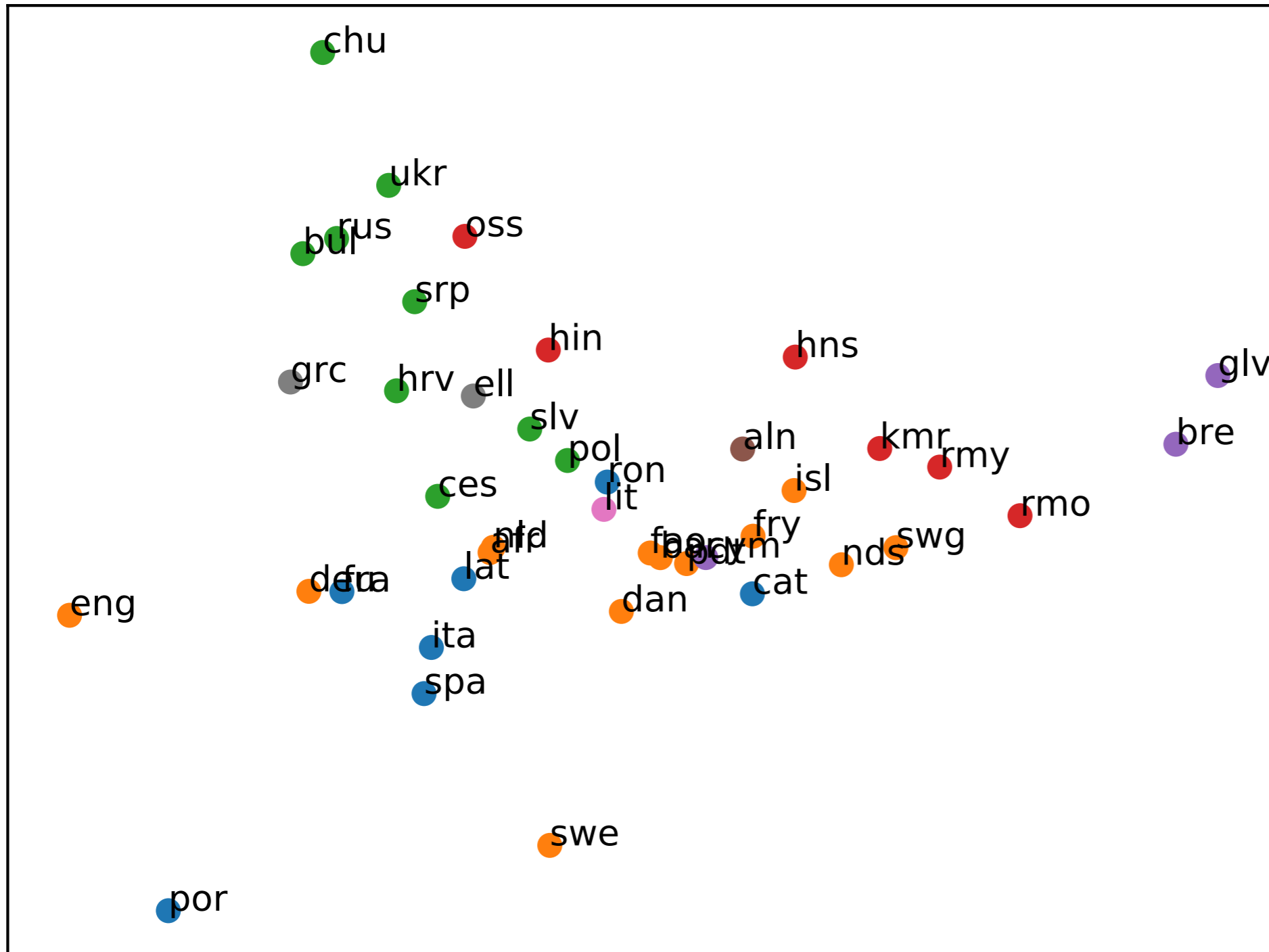
# Training with Indo-European Languages



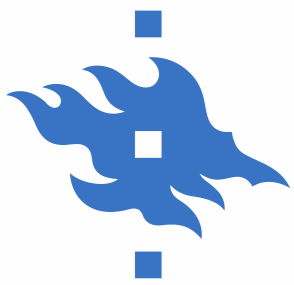
(PCA)



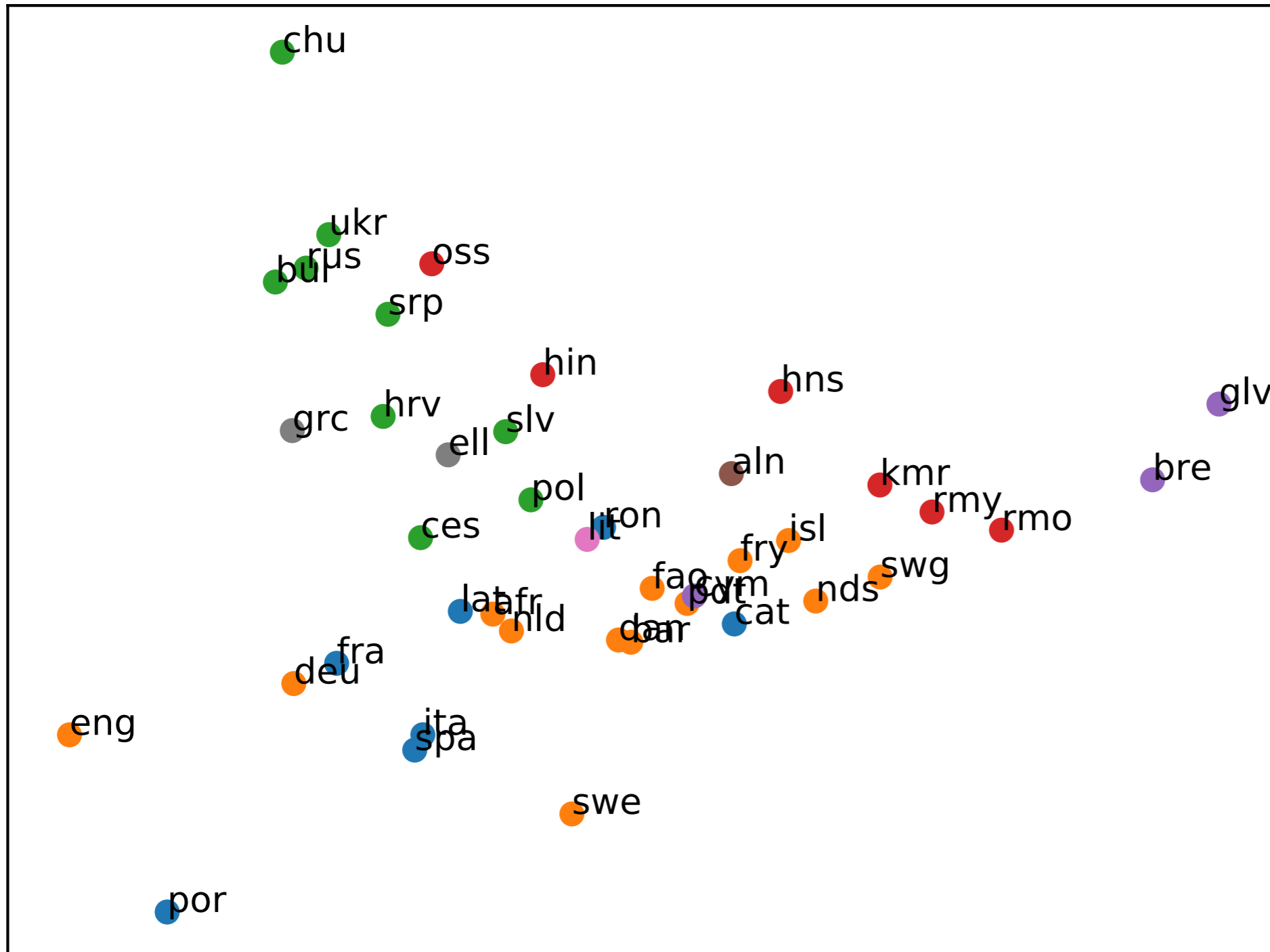
# Training with Indo-European Languages



(PCA)

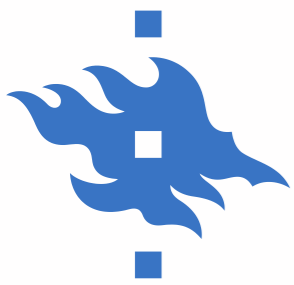


# Training with Indo-European Languages

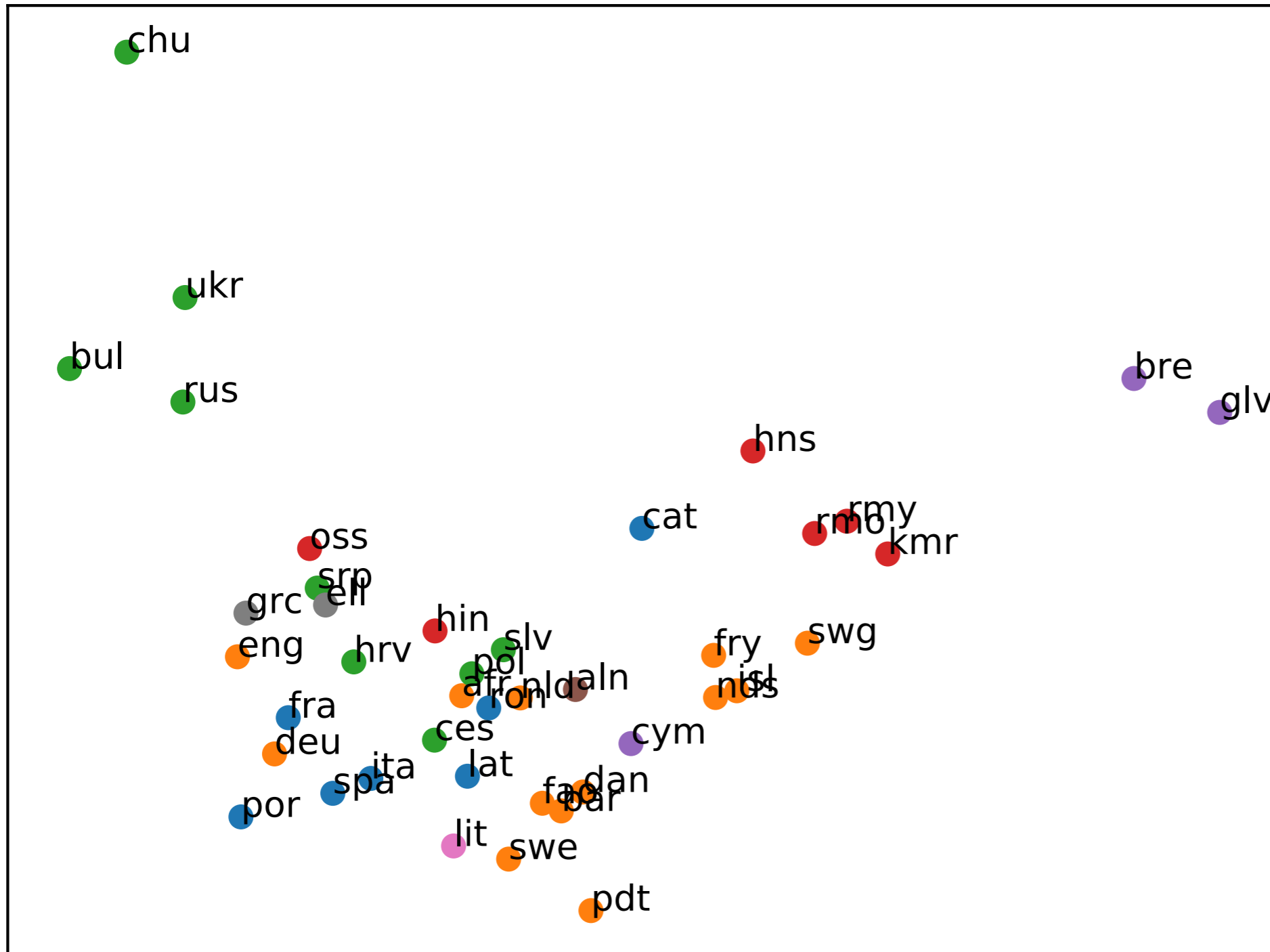


(PCA)

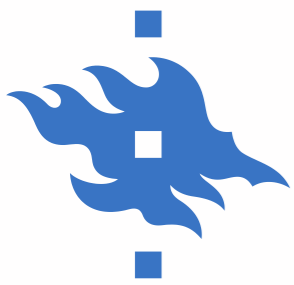




# Training with Indo-European Languages

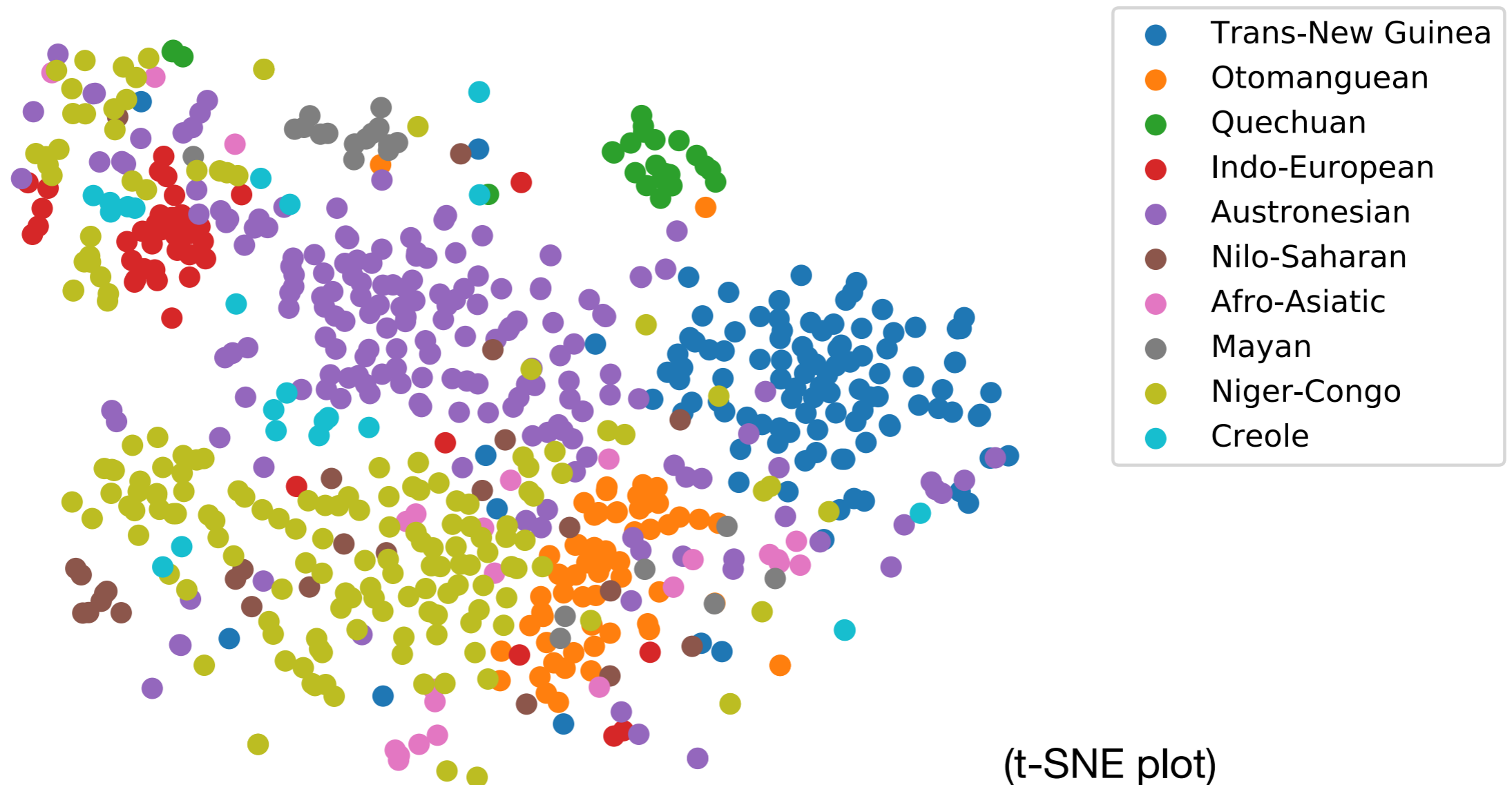


(PCA)



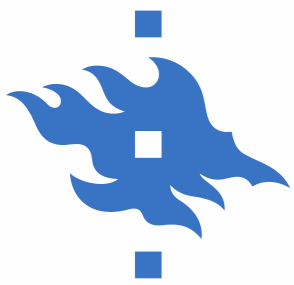
# Training with 972 Languages

Rough clusters of language families

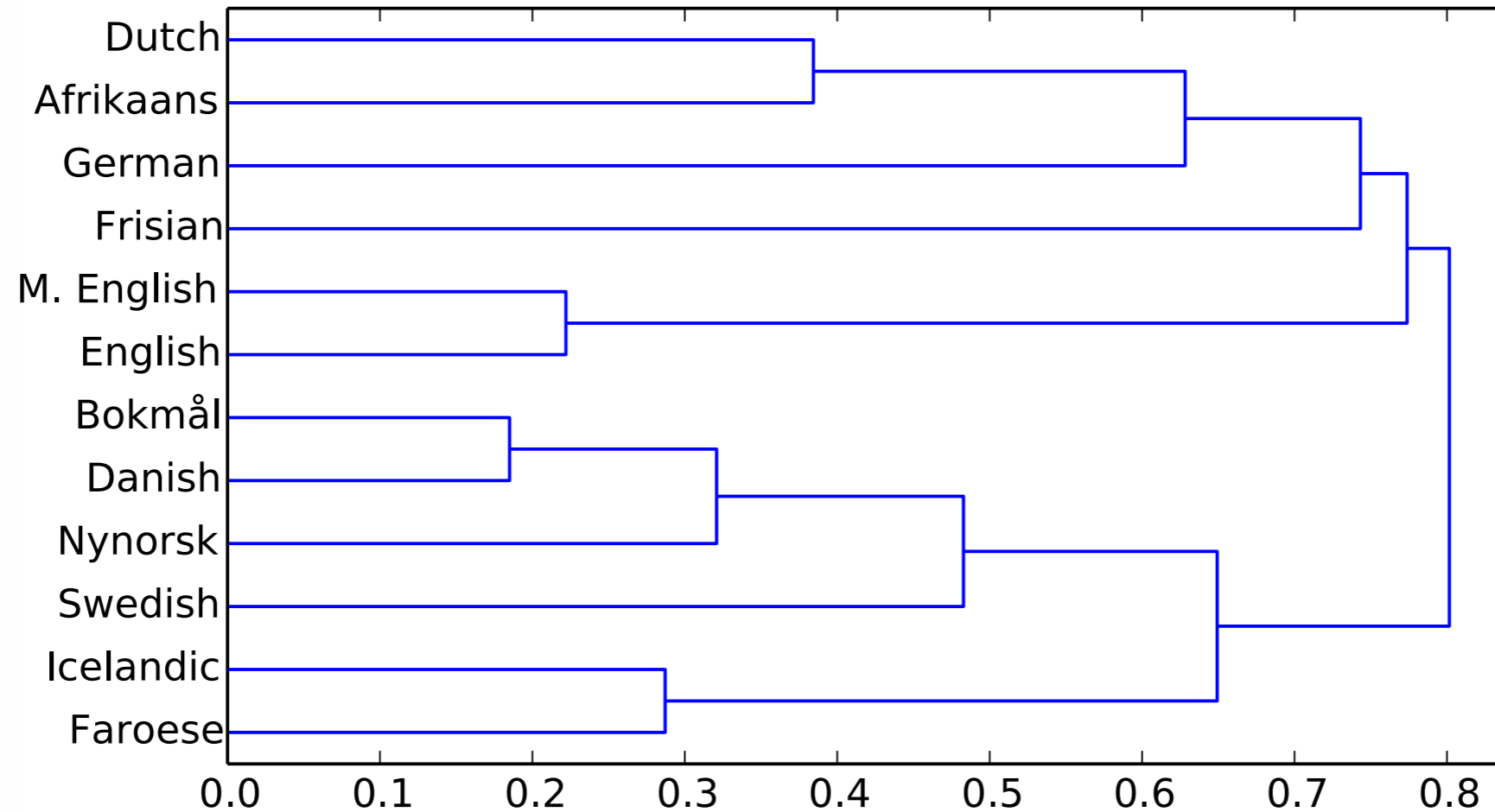


# Questions?





# Data-Driven Language Typology?





# MeMAD@WMT2018

---

## Collaboration between

- Aalto University
- University of Helsinki
- EURECOM

## Paper accepted to Conference for Machine Translation

- Preprint available <https://arxiv.org/abs/1808.10802v2>
- Open-Source, code available on github Waino/OpenNMT-py (branch develop\_mmod)



# MeMAD@WMT2018

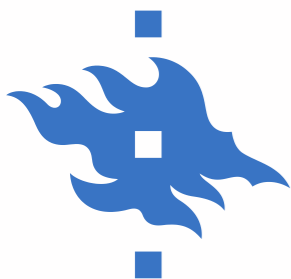
---

## Submitted system:

- Transformer (Vaswani et al., 2017) architecture
- Global image features extracted from Detectron, a pre-trained object detection and localization neural network.
- Multi-lingual training: a single model trained to translate into both languages simultaneously, then finetuned for each language separately.
- Ensemble of 3 independently trained models.

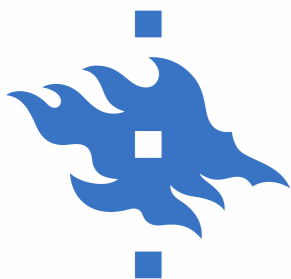
## Data:

- multi30k, filtered OpenSubtitles2018
- automatically translated MS-COCO image captions



# WMT2018: English-French

	<b>Submission Name</b>	<b>BLEU</b>	<b>Meteor</b>	<b>TER</b>
1	MeMAD_1_FLICKR_FR_MeMAD-OpenNMT-mmod_U	44.1	64.3	36.9
2	CUNI_1_FLICKR_FR_NeuralMonkeyTextual_U	40.6	61	40.7
3	CUNI_1_FLICKR_FR_NeuralMonkeyImagination_U	40.4	60.7	40.7
4	UMONS_1_FLICKR_FR_DeepGru_C	39.2	60	41.8
5	LIUMCVC_1_FLICKR_FR_MNMTEnsemble_C	39.5	59.9	41.7
6	LIUMCVC_1_FLICKR_FR_NMTEsemble_C	39.1	59.8	41.9
7	SHEF_1_FR_LT_C	38.8	59.8	41.5
8	SHEF_1_FR_MLT_C	38.9	59.8	41.5
9	SHEF1_1_FR_ENMT_C	38.9	59.8	41.2
10	SHEF1_1_FR_MFS_C	38.8	59.7	41.6
11	OSU-BD_1_FLICKR_FR_RLNMT_C	39	59.5	41.2
12	OSU-BD_1_FLICKR_FR_RLMIX_C	38.6	59.3	41.5
13	LIUMCVC_1_FLICKR_FR_MNMTSingle_C	37.9	58.5	43.4
14	LIUMCVC_1_FLICKR_FR_NMTSingle_C	37.6	58.4	43.2
15	Baseline	36.3	56.9	54.3



# WMT2018: English-German

	<b>Submission Name</b>	<b>BLEU</b>	<b>Meteor</b>	<b>TER</b>
1	MeMAD_1_FLICKR_DE_MeMAD-OpenNMT-mmod_U	38.5	56.6	44.6
2	CUNI_1_FLICKR_DE_NeuralMonkeyTextual_U	32.5	52.3	50.8
3	CUNI_1_FLICKR_DE_NeuralMonkeyImagination_U	32.2	51.7	51.7
4	UMONS_1_FLICKR_DE_DeepGru_C	31.1	51.6	53.4
5	LIUMCVC_1_FLICKR_DE_NMTEnsemble_C	31.1	51.5	52.6
6	LIUMCVC_1_FLICKR_DE_MNMTEnsemble_C	31.4	51.4	52.1
7	OSU-BD_1_FLICKR_DE_RLNMT_C	32.3	50.9	49.9
8	OSU-BD_1_FLICKR_DE_RLMIX_C	32.1	50.7	49.6
9	SHEF_1_DE_LT_C	30.5	50.7	53
10	SHEF_1_DE_MLT_C	30.4	50.7	52.9
11	SHEF1_1_DE_ENMT_C	30.9	50.7	52.4
12	SHEF1_1_DE_MFS_C	30.3	50.7	53.1
13	LIUMCVC_1_FLICKR_DE_MNMTSingle_C	28.8	49.9	55.6
14	LIUMCVC_1_FLICKR_DE_NMTSingle_C	29.5	49.9	54.3
15	Baseline	27.6	47.4	55.2
16	AFRL-OHIO-STATE_1_FLICKR_DE_4COMBO_U	24.3	45.4	58.6
17	AFRL-OHIO-STATE_1_FLICKR_DE_2IMPROVE_U	10	25.4	79.2
18	AFRL-OHIO-STATE_1_FLICKR_DE_CAPONLY_U	5	17.7	80.1





# MeMAD@IWSLT2018

---

## Automatic speech recognition

- Hybrid TDNN-HMM ASR based on Kaldi

## ASR-output translation

- transformer model (MarianNMT)
- ASR of TED data
- English-to-ASR-like English translation model for translating OpenSubtitles2018

## End-to-end model

- work in progress ...



# Translate English to ASR-English

---

**Original:** Because in the summer of 2006, the E.U. Commission tabled a directive.

**ASR-like:** because in the summer of two thousand and six you commission tabled a directive

**Original:** I'm a child of 1984,

**ASR-like:** i am a child of nineteen eighty four

**Original:** Stasi was the secret police in East Germany.

**ASR-like:** stars he was the secret police in east germany



# MeMAD@IWSLT2018

---

## Results on dev set

Training data	BLEU	
	Untuned	Tuned
TED-ASR-TOP-10+SUBS	20.44	20.58
TED-ASR-TOP-10+SUBS-ASR	19.79	20.80

## Results on test set 2018

Training data	BLEU
TED-ASR-TOP-10	14.34
TED-ASR-TOP-10+SUBS	16.45
TED-ASR-TOP-10+SUBS-ASR	15.80