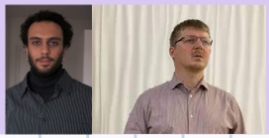
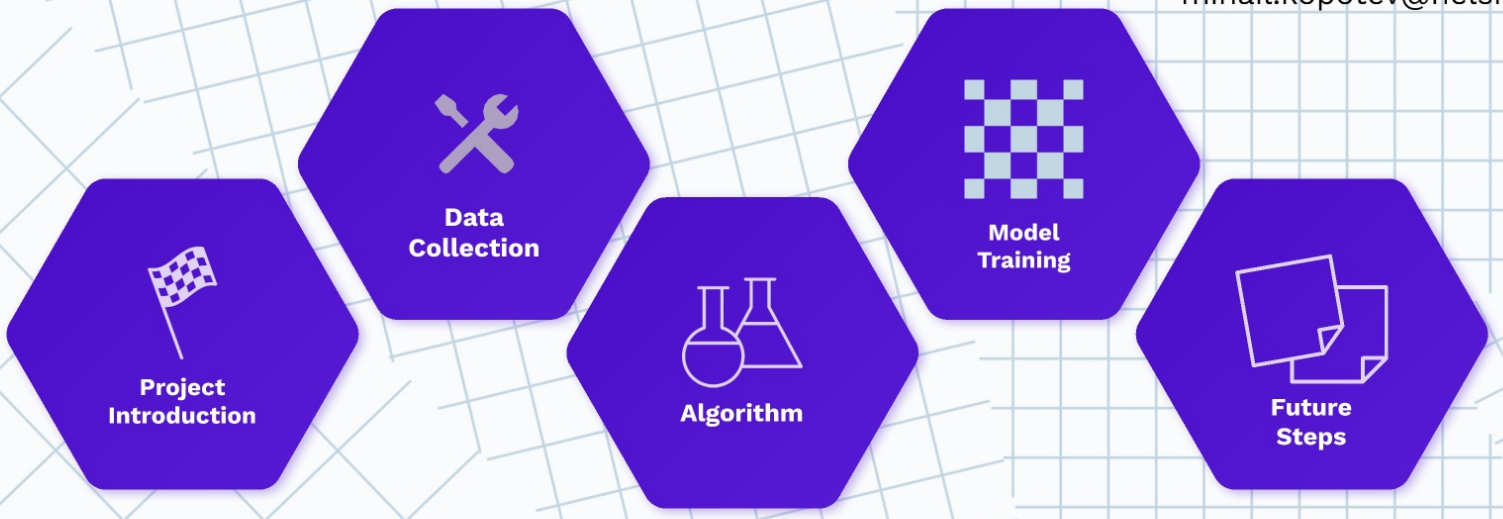
AI in Computer Assisted Language Learning for the Russian Language

Automatic Collocation Suggestion

Andrea Grillandi, Mihail Kopotev
U. of Helsinki



andrea.grillandi@live.it
mihail.kopotev@helsinki.fi




Introduction
• Automatic collocation suggestion
• CAT&kittens
An assistant for Russian academic writing
CAT
kittens
Collocation suggestion

Introduction

- Automatic collocation suggestion
- **CAT&kittens**
  An assistant for Russian academic writing

CAT

- CAT stands for Corpus of Academic Texts
- Standard used for comparison

CAT

- CAT stands for Corpus of Academic Texts
- Standard used for comparison


Introduction
• Automatic collocation suggestion
• CAT&kittens
An assistant for Russian academic writing
CAT
kittens
Collocation suggestion


kittens
• learner texts
• written by L2 students or native speakers

kittens

- learner texts
- written by L2 students or native speakers


Introduction
• Automatic collocation suggestion
• CAT&kittens
An assistant for Russian academic writing
CAT
kittens
Collocation suggestion

Collocation suggestion

- Collocations - series of words which co-occur with a certain frequency
- Given a learner collocation, suggest semantically similar but more appropriate collocation
  e.g. "zone of research" -> "area of research"

Collocation suggestion

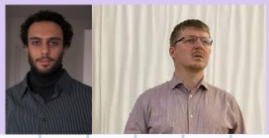
- Collocations - series of words which co-occur with a certain frequency
- Given a learner collocation, suggest semantically similar but more appropriate collocation
  e.g. "zone of research" -> "area of research"


Introduction
• Automatic collocation suggestion
• CAT&kittens
An assistant for Russian academic writing
CAT
kittens
Collocation suggestion

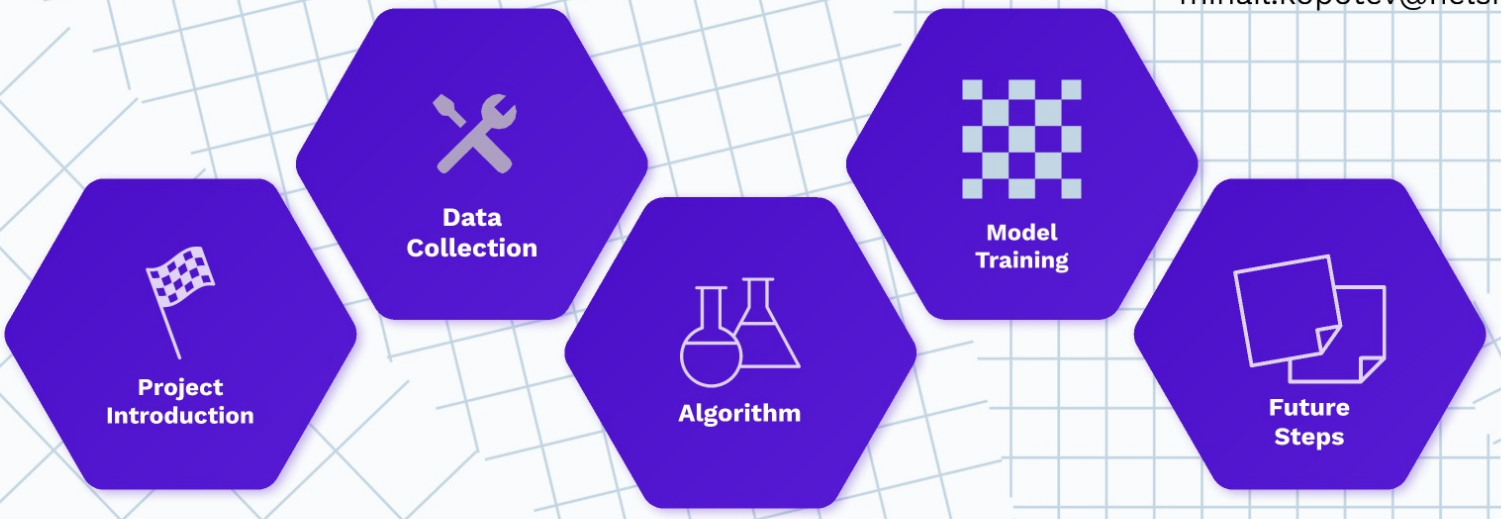
AI in Computer Assisted Language Learning for the Russian Language

Automatic Collocation Suggestion

Andrea Grillandi, Mihail Kopotev
U. of Helsinki



andrea.grillandi@live.it
mihail.kopotev@helsinki.fi




Data Collection and Preprocessing
• Academic texts from Cyberleninka
• Corpus composition
• Text preprocessing
Cyberleninka
Corpus
Preprocessing

Data Collection and Preprocessing

- Academic texts from Cyberleninka
- Corpus composition
- Text preprocessing


Cyberleninka
• scientific electronic library
• Russian academic articles

Cyberleninka

- scientific electronic library
- Russian academic articles


Data Collection and Preprocessing
• Academic texts from Cyberleninka
• Corpus composition
• Text preprocessing
Cyberleninka
Corpus
Preprocessing

Corpus composition

- ~ 60 mln tokens
- ~ 10 mln per domain
- 6 domains: Economics, Education and Psychology, Linguistics, Law, History, Sociology

Corpus composition

- ~ 60 mln tokens
- ~ 10 mln per domain
- 6 domains: Economics, Education and Psychology, Linguistics, Law, History, Sociology


Data Collection and Preprocessing
• Academic texts from Cyberleninka
• Corpus composition
• Text preprocessing
Cyberleninka
Corpus
Preprocessing

Text preprocessing

- Extraction from PDF
- Cleaning through regex
- Tokenization, Lemmatization & POS tagging with Treetagger

Text preprocessing

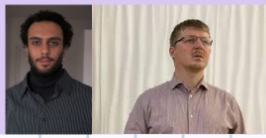
- Extraction from PDF
- Cleaning through regex
- Tokenization, Lemmatization & POS tagging with Treetagger


Data Collection and Preprocessing
• Academic texts from Cyberleninka
• Corpus composition
• Text preprocessing
Cyberleninka
Corpus
Preprocessing

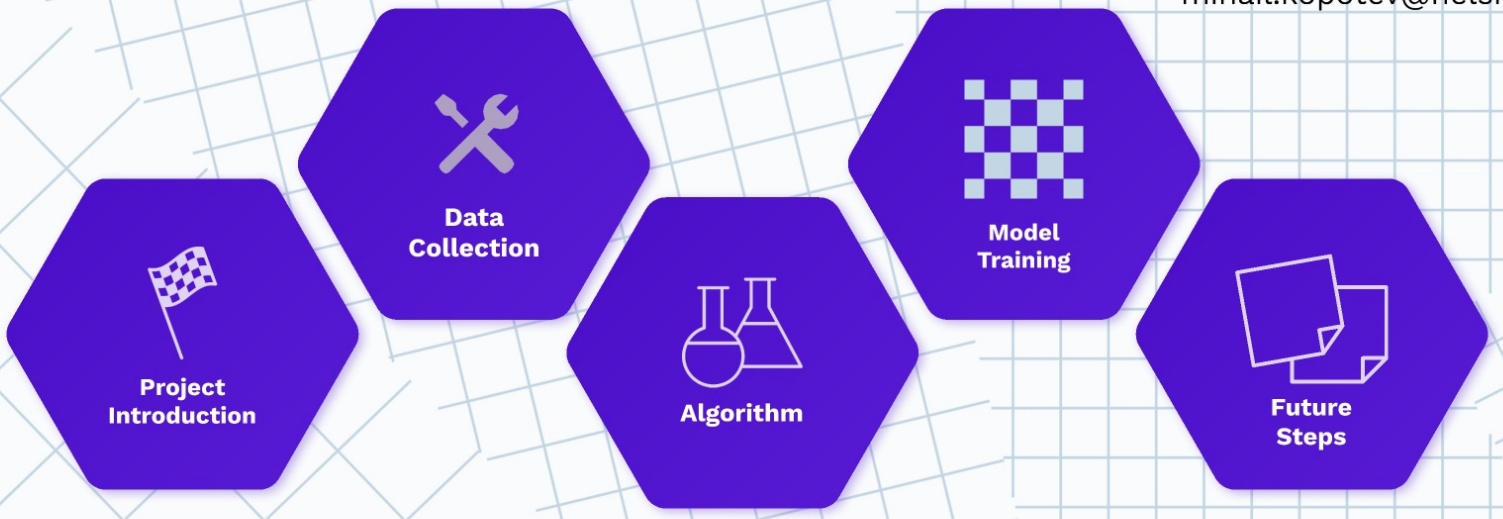
AI in Computer Assisted Language Learning for the Russian Language

Automatic Collocation Suggestion

Andrea Grillandi, Mihail Kopotev
U. of Helsinki



andrea.grillandi@live.it
mihail.kopotev@helsinki.fi




Collocation suggestion
1. search learner collocations on the CAT
2. if collocation is not found, suggest new collocation based on static or dynamic ML models
3. keep if found in corpus, else discard
4. sort suggested collocations according to cosine (=semantic) similarity to the original learner collocation
ngram search
Static models
Dynamic models
Filter & Sort

Collocation suggestion

1. search learner collocations on the CAT

2. if collocation is not found, suggest new collocation based on static or dynamic ML models

3. keep if found in corpus, else discard

4. sort suggested collocations according to cosine (=semantic) similarity to the original learner collocation

Find collocations

- learner text is split into 2-grams, 3-grams and 4-grams
- "NLP is a key area of research" -> ["NLP is a", "is a key", "a key area", "key area of", "area of research"]
- "zone of research" -> "area of research"

Find collocations

- learner text is split into 2-grams, 3-grams and 4-grams
- "NLP is a key area of research" -> ["NLP is a", "is a key", "a key area", "key area of", "area of research"]
- "zone of research" -> "area of research"


Collocation suggestion
1. search learner collocations on the CAT
2. if collocation is not found, suggest new collocation based on static or dynamic ML models
3. keep if found in corpus, else discard
4. sort suggested collocations according to cosine (=semantic) similarity to the original learner collocation
ngram search
Static models
Dynamic models
Filter & Sort

Static Models

- Static because non-contextual
- w2v, fastText, GloVe
- *zone of research
  area of research
  field of research
  zone of study
  area of study

Static Models

- Static because non-contextual
- w2v, fastText, GloVe
- *zone of research
  area of research
  field of research
  zone of study
  area of study


Collocation suggestion
1. search learner collocations on the CAT
2. if collocation is not found, suggest new collocation based on static or dynamic ML models
3. keep if found in corpus, else discard
4. sort suggested collocations according to cosine (=semantic) similarity to the original learner collocation
ngram search
Static models
Dynamic models
Filter & Sort

Dynamic Models

- Dynamic because word representations are contextual
- BERT and GPT

- NLP is a key ***zone of research** ->
  NLP is a key [MASK] of research
  NLP is a key zone of [MASK]
  NLP is a key [MASK] of [MASK]

Dynamic Models

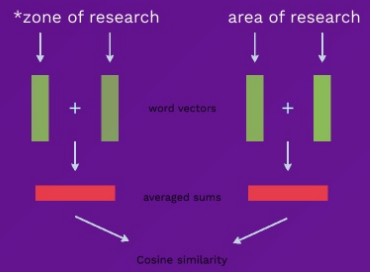
- Dynamic because word representations are contextual
- BERT and GPT

- NLP is a key ***zone of research** ->
  NLP is a key [MASK] of research
  NLP is a key zone of [MASK]
  NLP is a key [MASK] of [MASK]


Collocation suggestion
1. search learner collocations on the CAT
2. if collocation is not found, suggest new collocation based on static or dynamic ML models
3. keep if found in corpus, else discard
4. sort suggested collocations according to cosine (=semantic) similarity to the original learner collocation
ngram search
Static models
Dynamic models
Filter & Sort

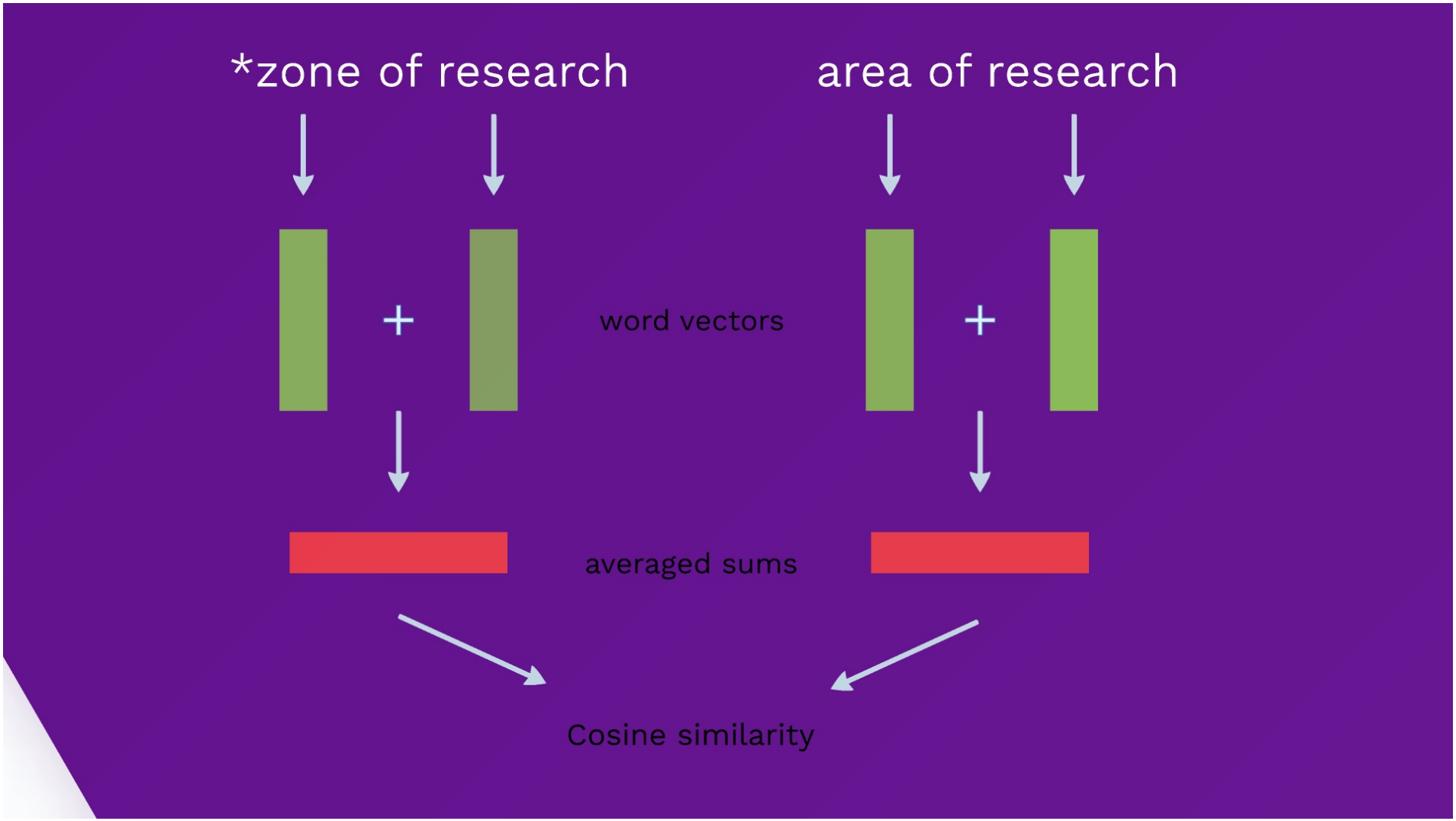
Filtering and Sorting

- Keep or discard
- Sum up static semantic representations of all words in collocation (both for target and suggested collocation)
- Compute cosine similarity with the target collocation
- Sorting by highest semantic similarity



Filtering and Sorting

- Keep or discard
- Sum up static semantic representations of all words in collocation (both for target and suggested collocation)
- Compute cosine similarity with the target collocation
- Sorting by highest semantic similarity

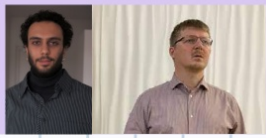

*zone of research
area of research
word vectors
averaged sums
Cosine similarity


Collocation suggestion
1. search learner collocations on the CAT
2. if collocation is not found, suggest new collocation based on static or dynamic ML models
3. keep if found in corpus, else discard
4. sort suggested collocations according to cosine (=semantic) similarity to the original learner collocation
ngram search
Static models
Dynamic models
Filter & Sort

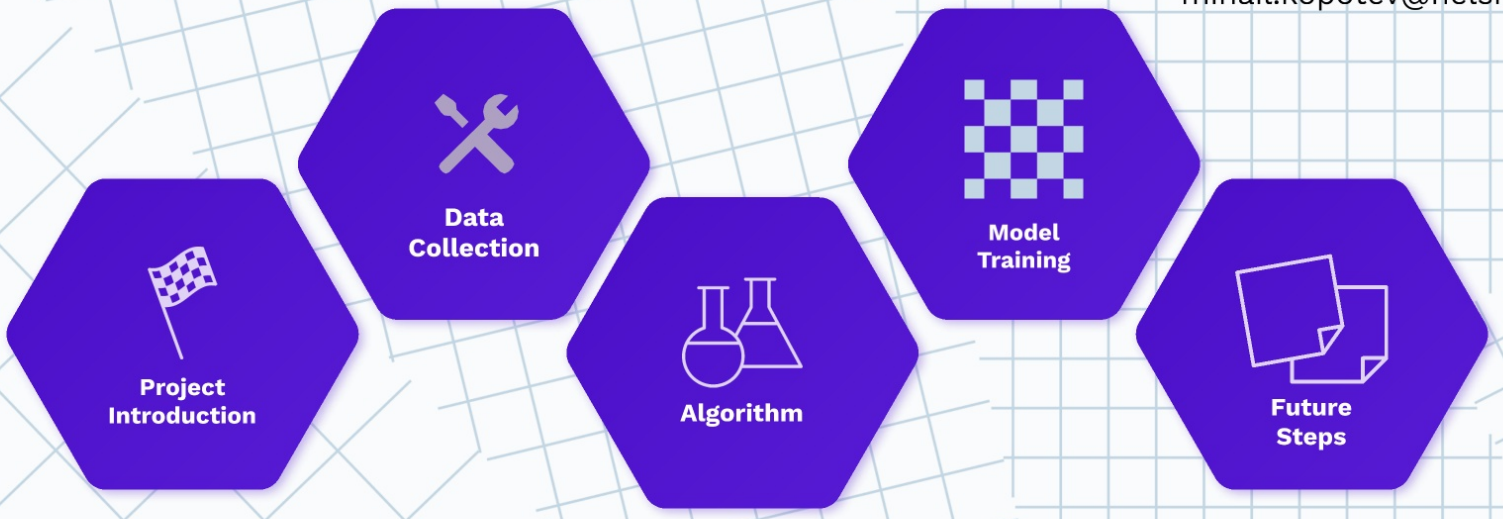
AI in Computer Assisted Language Learning for the Russian Language

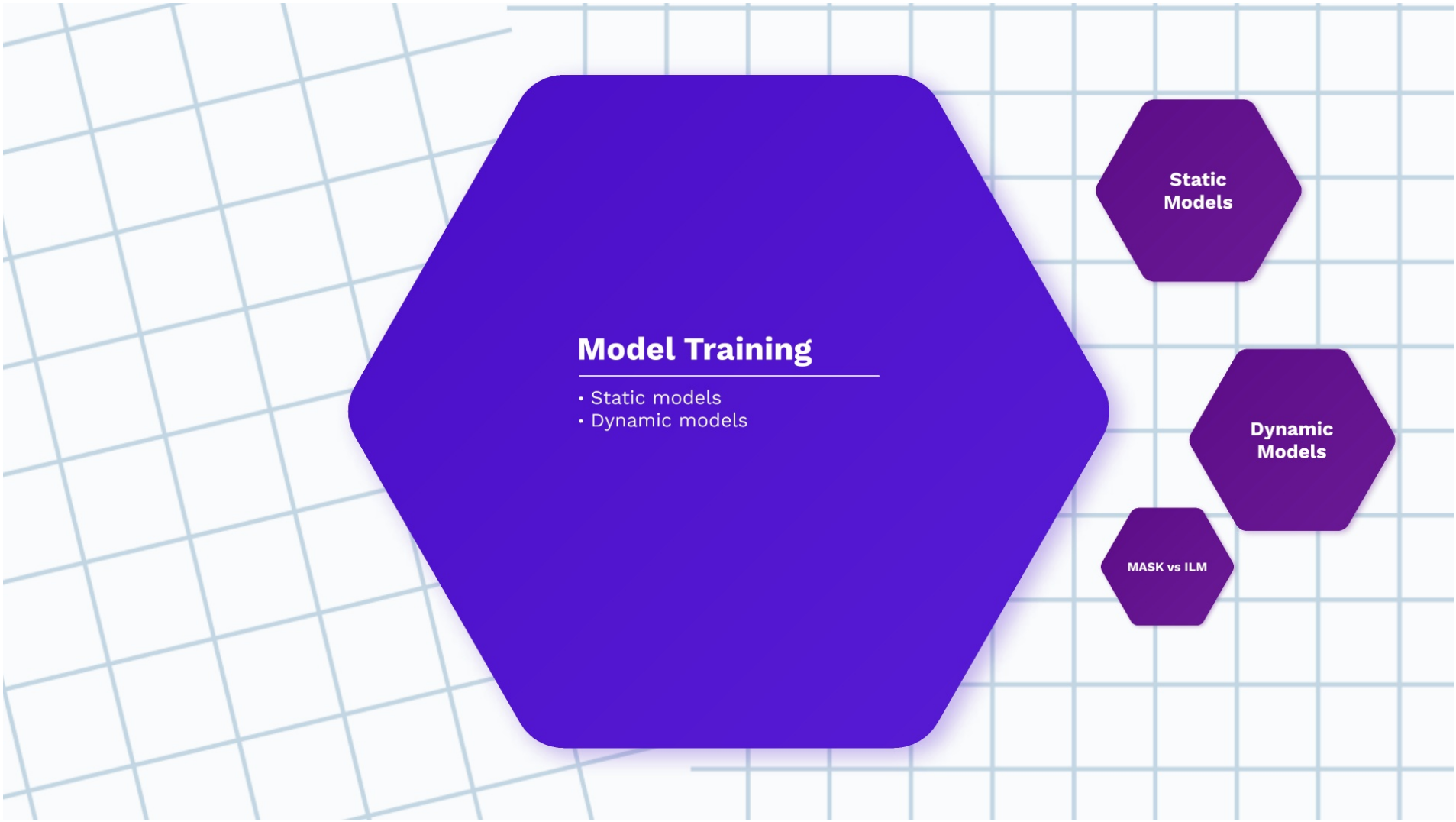
Automatic Collocation Suggestion

Andrea Grillandi, Mihail Kopotev
U. of Helsinki



andrea.grillandi@live.it
mihail.kopotev@helsinki.fi



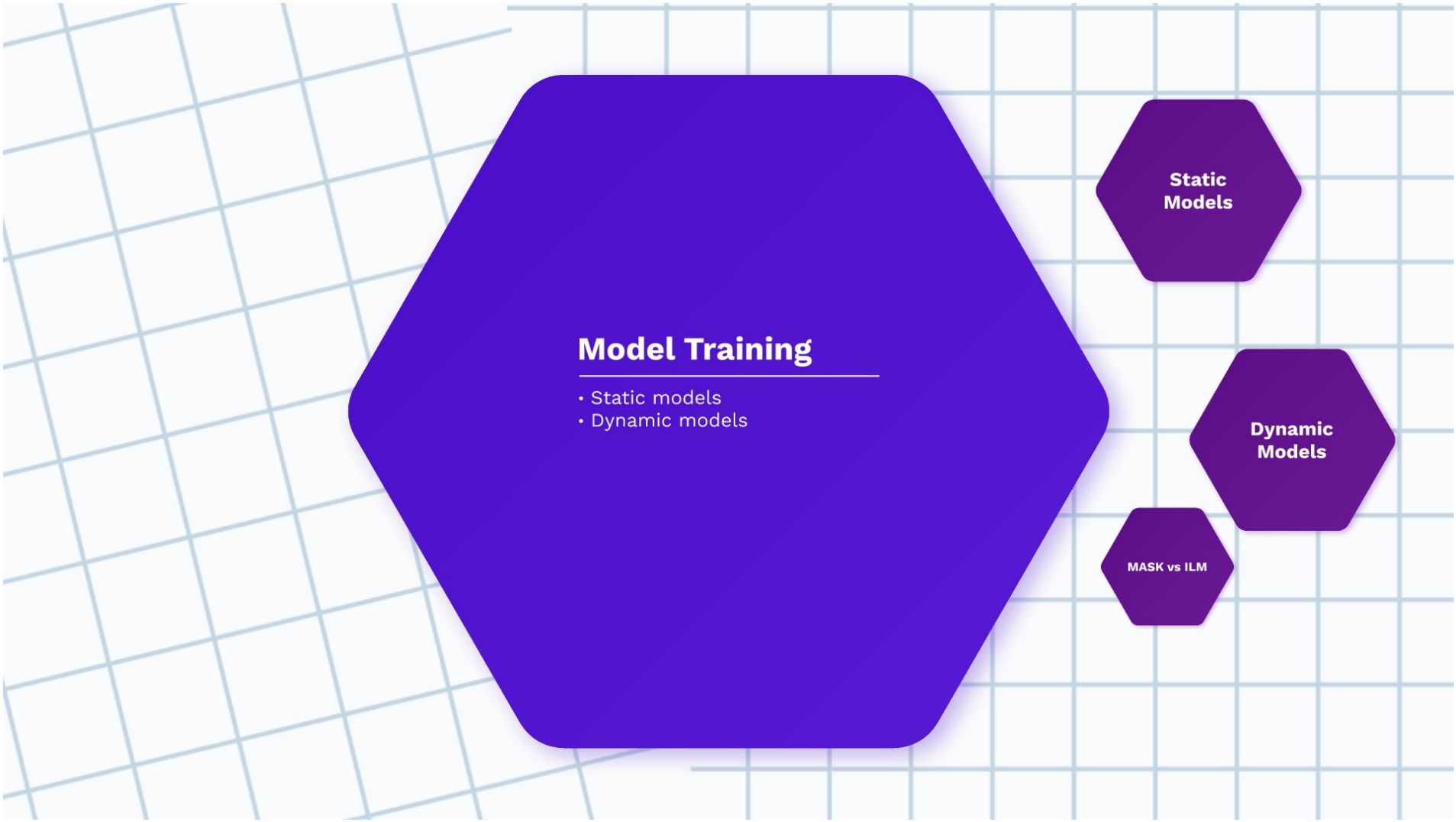

Model Training
• Static models
• Dynamic models
Static Models
Dynamic Models
MASK vs ILM

Static models

- Lemmas + POS: "zone_N of_P research_N"
- Gensim for w2v and fastText
- Stanford GloVe repository
- One model per domain and one for full data

Static models

- Lemmas + POS: "zone_N of_P research_N"
- Gensim for w2v and fastText
- Stanford GloVe repository
- One model per domain and one for full data


Model Training
• Static models
• Dynamic models
Static Models
Dynamic Models
MASK vs ILM

Dynamic Models

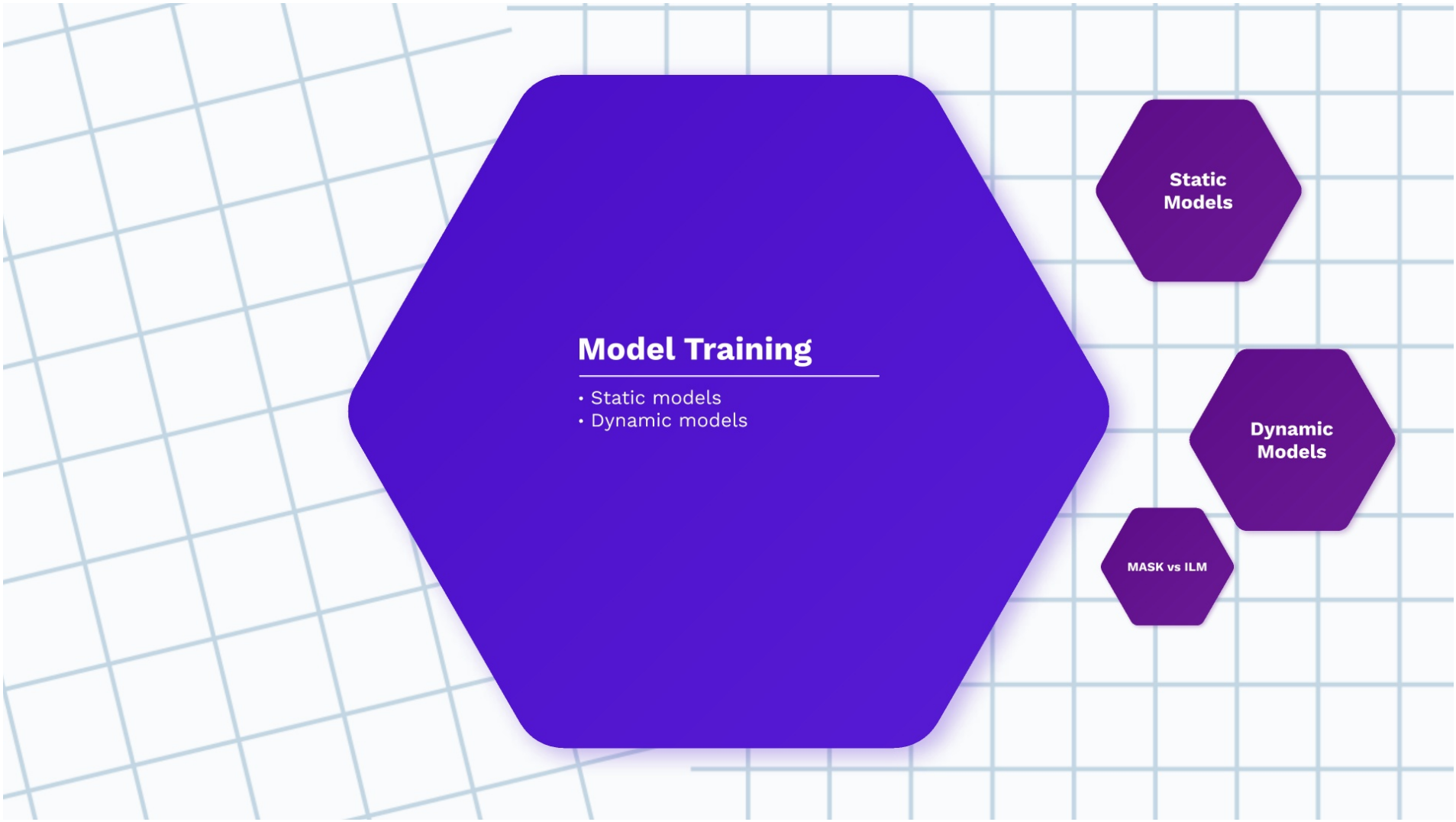
- Huggingface transformers library (py)
- BERT large:
  - 340 mln. parameters, trained on NL inference dataset
  - and fine-tuned on our lemmatized data
  - "[MASK]**_N** of**_P** research**_N**"
- ruGPT3xl:
  - 1.3 bln parameters, gpt-2 based model integrating ideas from gpt-3 paper
  - and fine-tuned on our tokenized data

- ILM framework to deploy both models

Dynamic Models

- Huggingface transformers library (py)
- BERT large:
  - 340 mln. parameters, trained on NL inference dataset
  - and fine-tuned on our lemmatized data
  - "[MASK]**_N** of**_P** research**_N**"
- ruGPT3xl:
  - 1.3 bln parameters, gpt-2 based model integrating ideas from gpt-3 paper
  - and fine-tuned on our tokenized data

- ILM framework to deploy both models

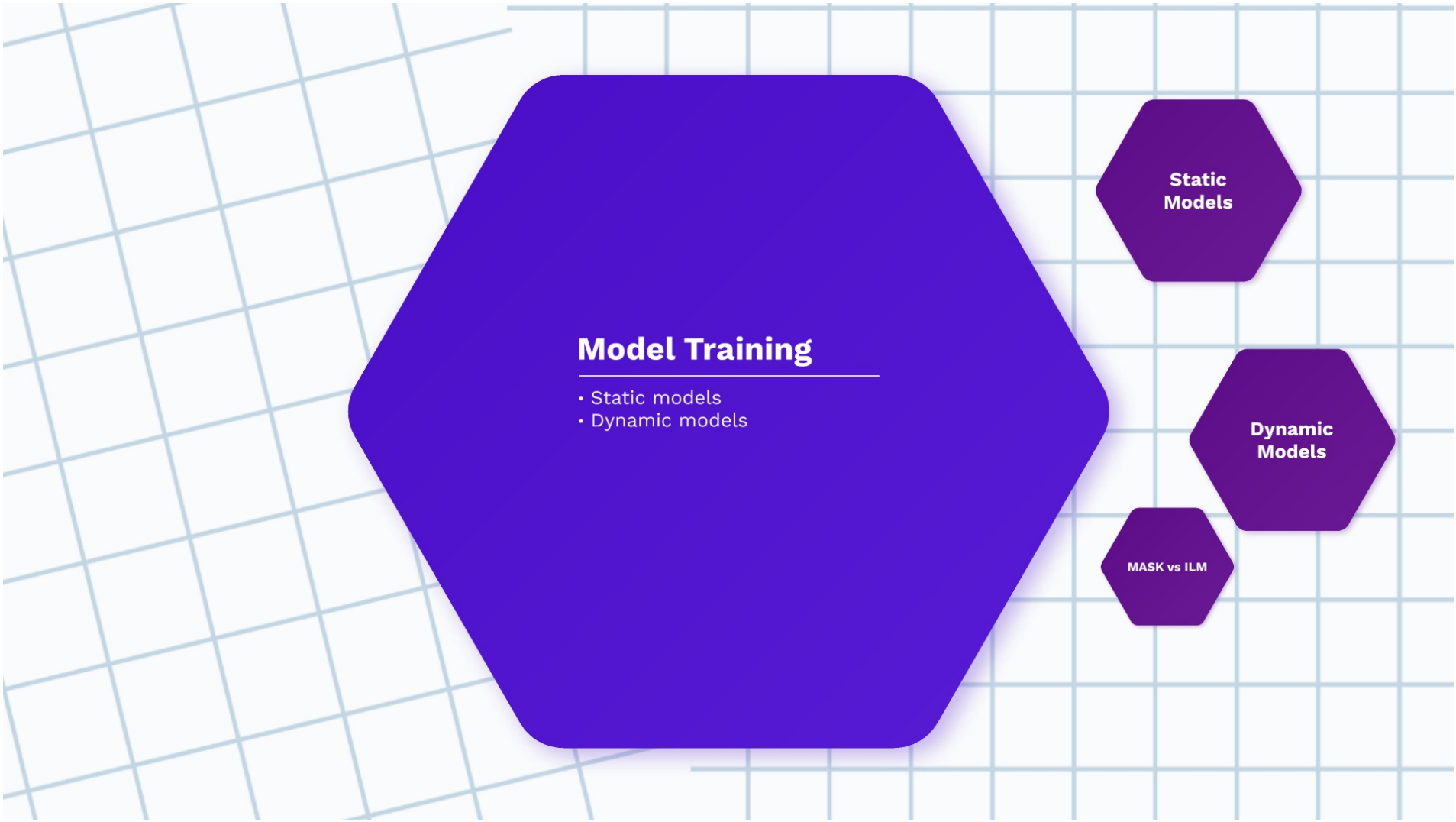

Model Training
• Static models
• Dynamic models
Static Models
Dynamic Models
MASK vs ILM

Masked Models vs. Infill model (ILM)

- "NLP is a key [MASK] of research -> ["area", "field", ...]
- "NLP is a key [infill] -> ["area", "research field", "area of research"]

Masked Models vs. Infill model (ILM)

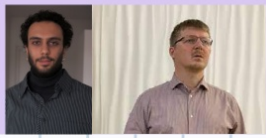
- "NLP is a key [MASK] of research -> ["area", "field", ...]
- "NLP is a key [infill] -> ["area", "research field", "area of research"]


Model Training
• Static models
• Dynamic models
Static Models
Dynamic Models
MASK vs ILM

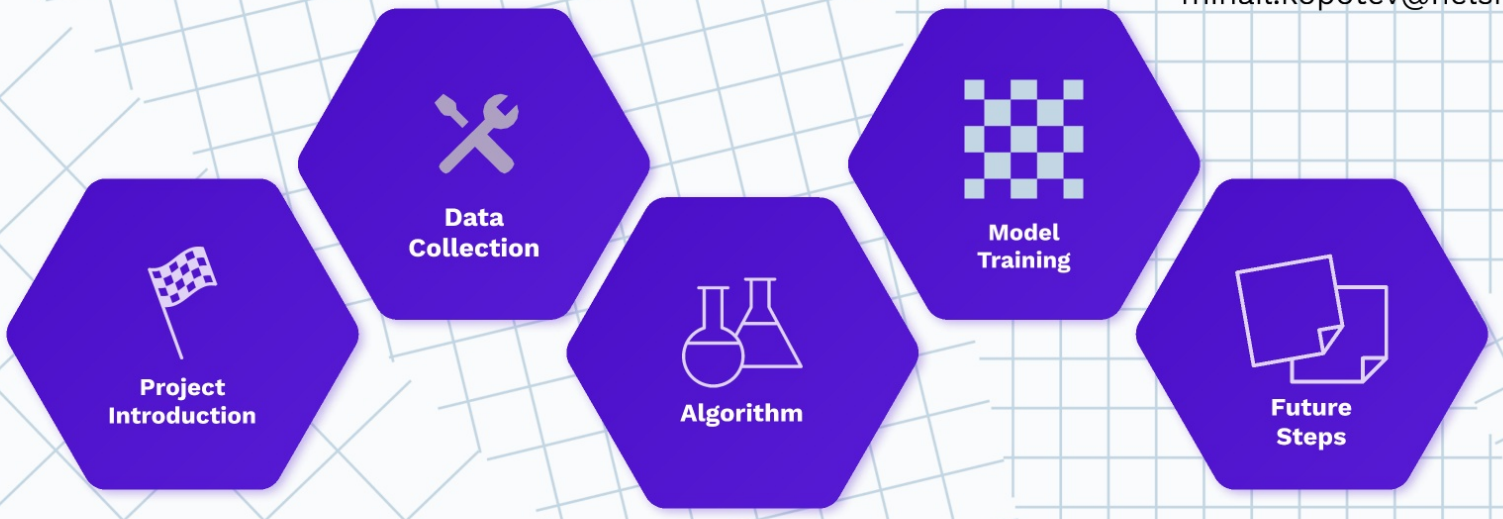
AI in Computer Assisted Language Learning for the Russian Language

Automatic Collocation Suggestion

Andrea Grillandi, Mihail Kopotev
U. of Helsinki



andrea.grillandi@live.it
mihail.kopotev@helsinki.fi



Future steps

- Data collection -- DONE
- Data processing -- DONE
- Training the models -- DONE
- ILM framework tuning -- DONE
- Evaluation of the results, choosing the best model(s) and parameters -- IN PROGRESS

- Integration the algorithm into a AI-writing-assistant similar to Grammarly, but for the Russian language -- TO BE DONE

GRAMOTA.RU

Future steps

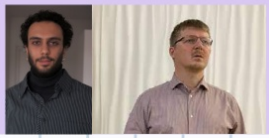
- Data collection -- DONE
- Data processing -- DONE
- Training the models -- DONE
- ILM framework tuning -- DONE
- Evaluation of the results, choosing the best model(s) and parameters -- IN PROGRESS

- Integration the algorithm into a AI-writing-assistant similar to Grammarly, but for the Russian language -- TO BE DONE

AI in Computer Assisted Language Learning for the Russian Language

Automatic Collocation Suggestion

Andrea Grillandi, Mihail Kopotev
U. of Helsinki



andrea.grillandi@live.it
mihail.kopotev@helsinki.fi