#NLP4CALL2019

Predicting learner's knowledge of single words using machine learning

Klara Ptacinova Klimcikova — Ludwig Maximilians University
Drilon Avdiu — Technical University Munich
Vanessa Bui — Technical University Munich

CONTENTS

1. **Motivation**
2. **Related Work**

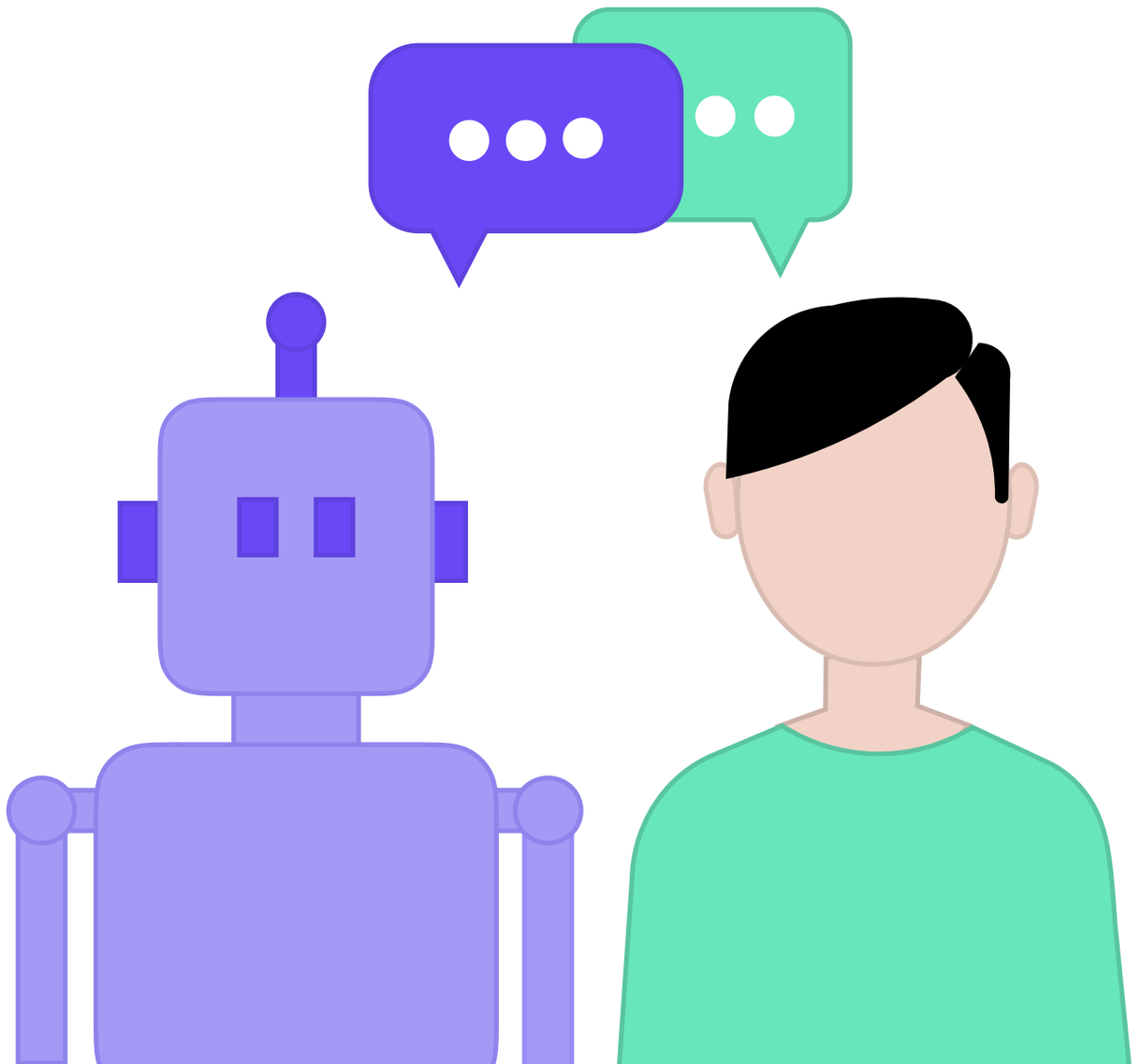
3. **Dataset**
4. **Features**
   - Word-specific
   - Learner-specific
4. **Models**
   - Complete Features Space Dependant
   - Neural Network Based
5. **Results**

6. **Future Work**

MOTIVATION

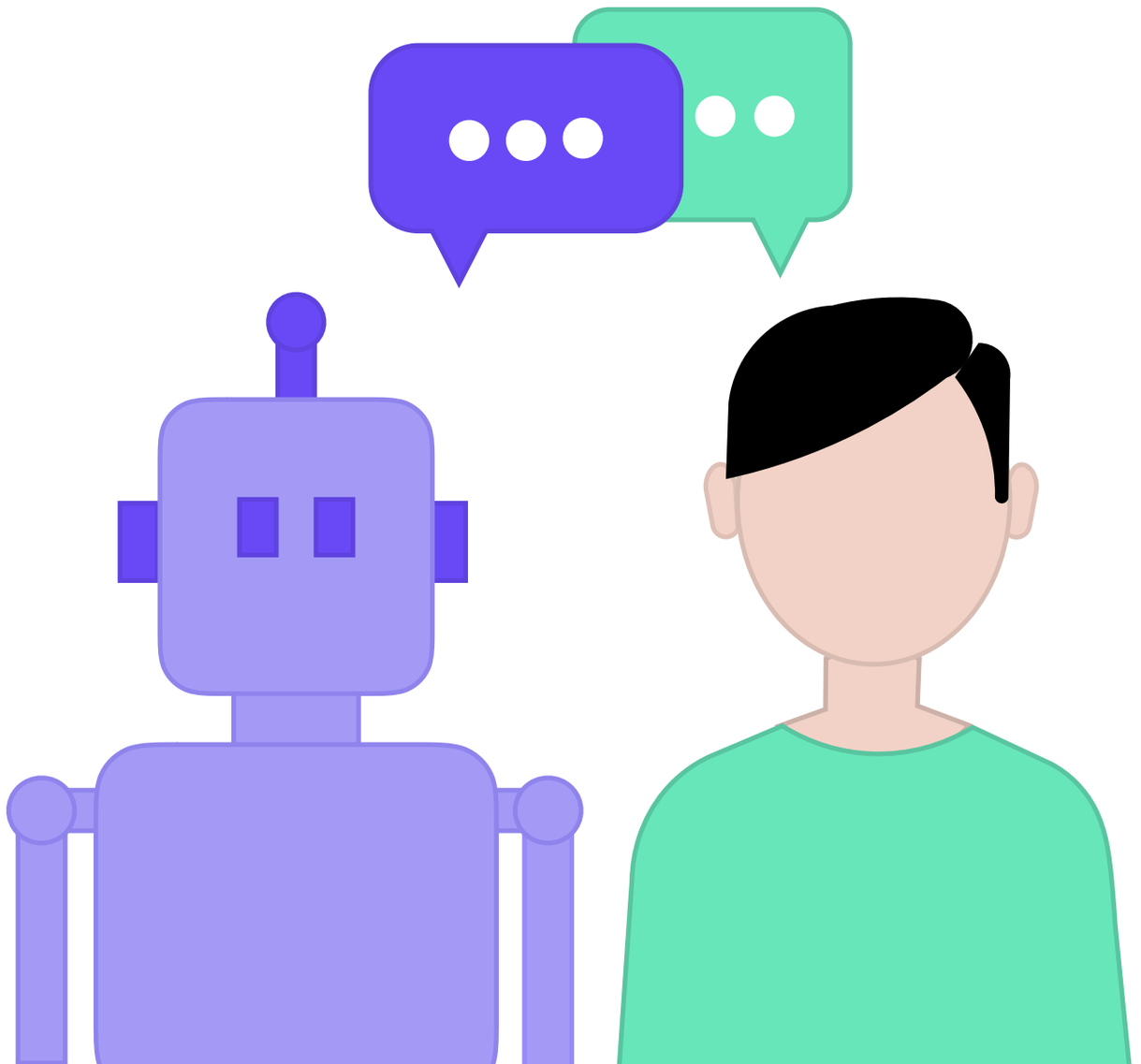
MOTIVATION

for personalized interaction
of intelligent tutoring systems
with language learners



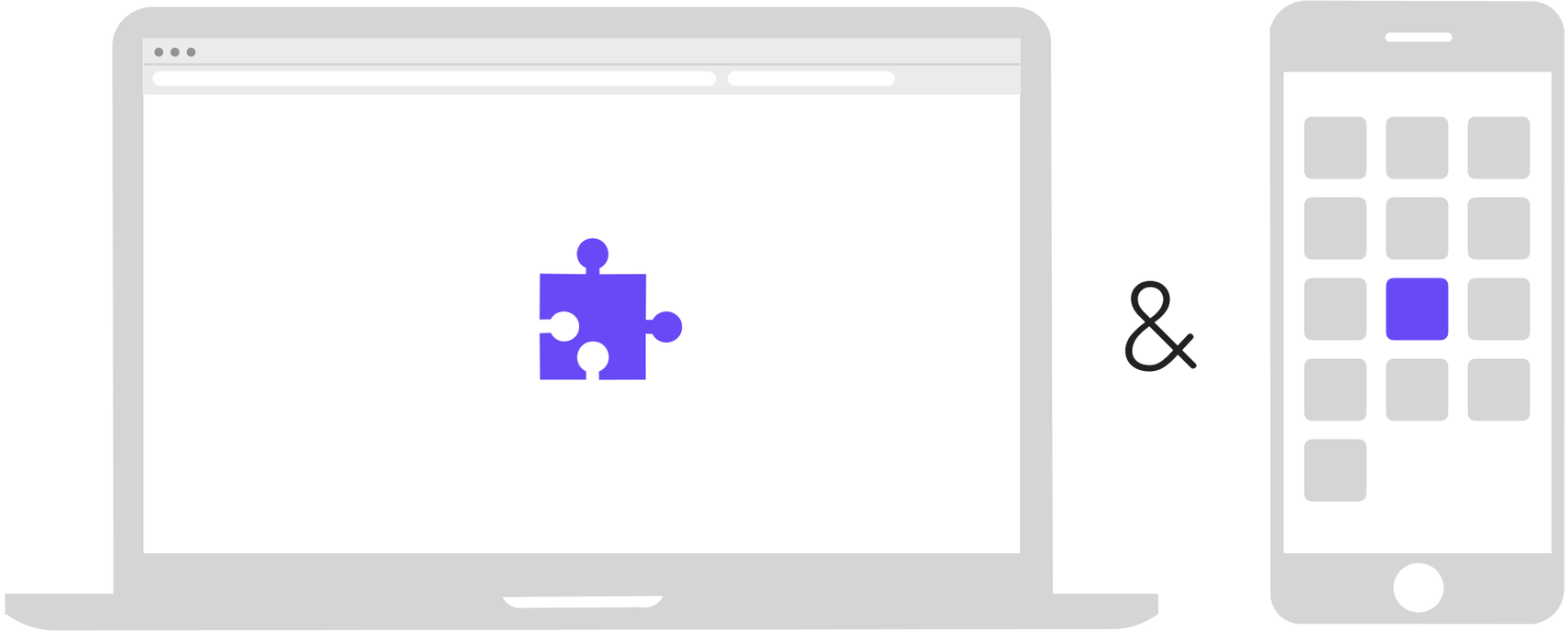
MOTIVATION

for personalized interaction
of the intelligent tutoring system **Elia**
with language learners

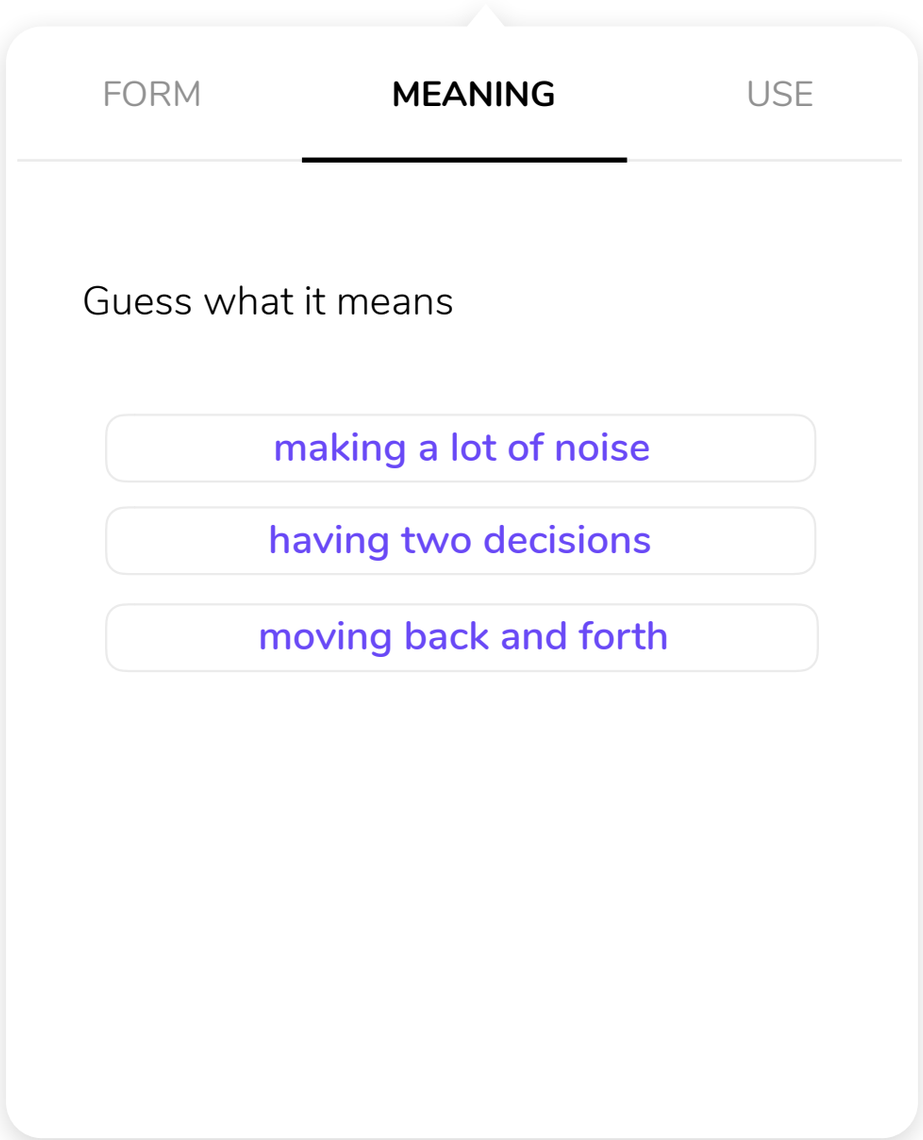


MOTIVATION

for personalized interaction
of the intelligent tutoring system **Elia**
with language learners



MOTIVATION


FORM
MEANING
USE
Guess what it means
making a lot of noise
having two decisions
moving back and forth

MOTIVATION

FORM | **MEANING** | USE

You might know this word!

It comes from the same word family as:

attend VERB

NEXT

MOTIVATION

FORM | **MEANING** | USE

You have seen this word before in these sentence(s)!

This was a nice a nice evening.
They gave me a very nice present.

NEXT

MOTIVATION

Fill the gap with a correct word!

The plan gained rapid media

shortly afterwards.

NEXT

MOTIVATION

Choose the closest synonym of
the word **-attention**!

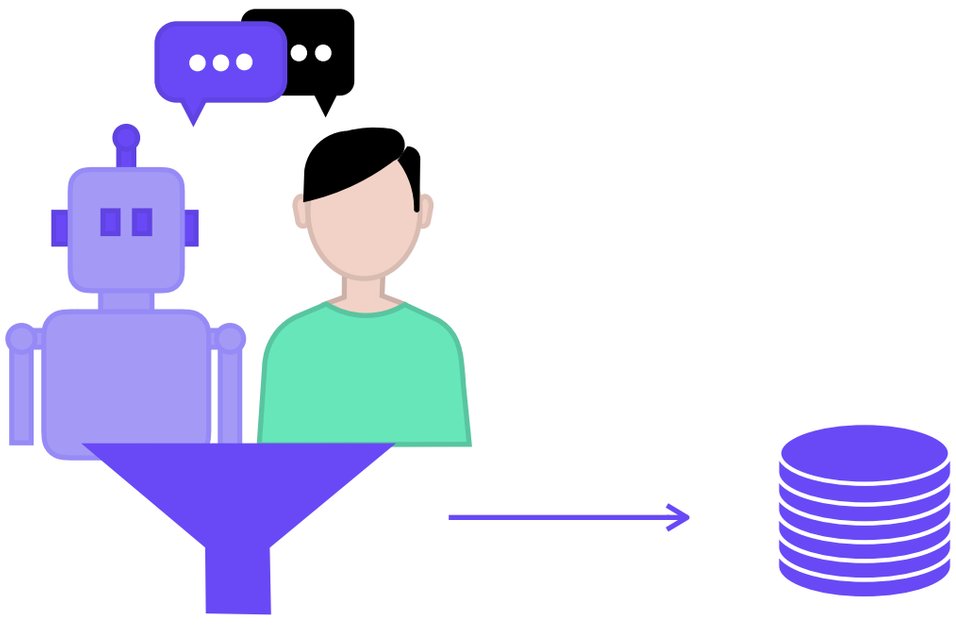
focus

characteristic

uniqueness

NEXT

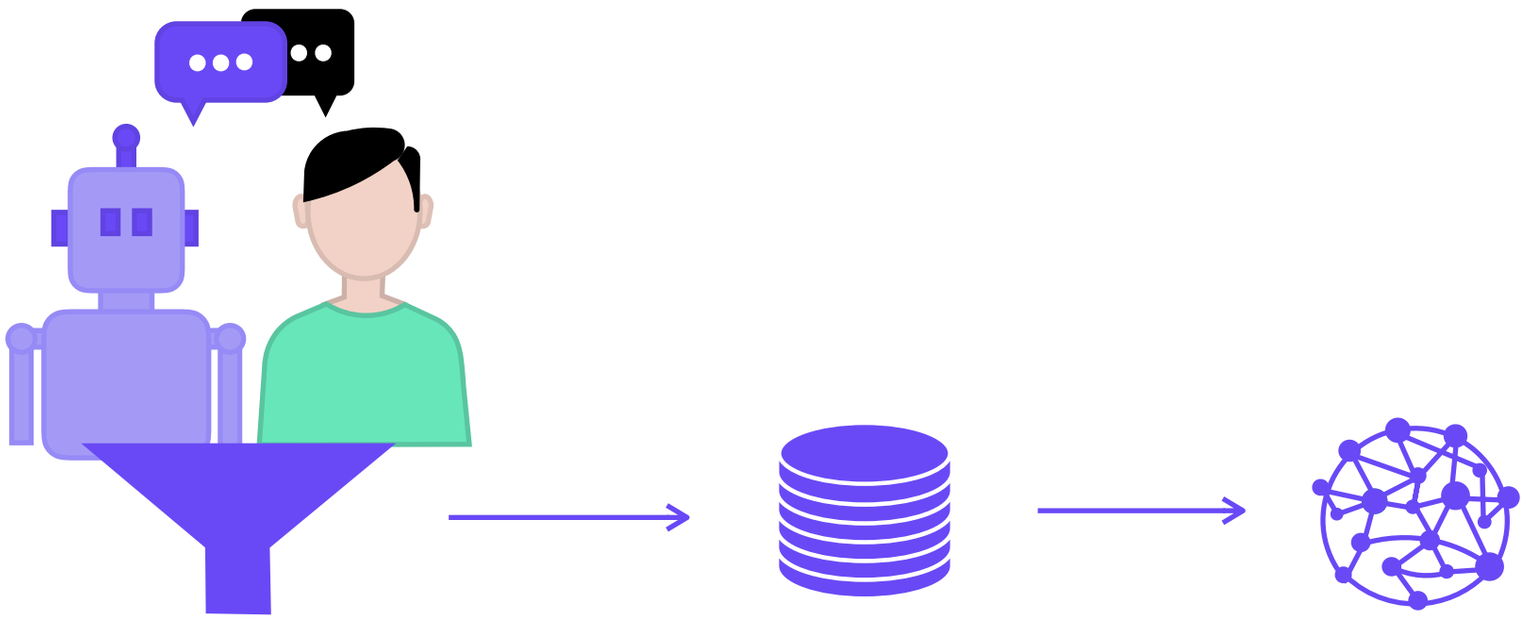
MOTIVATION



INTERACTION

DATA

MOTIVATION

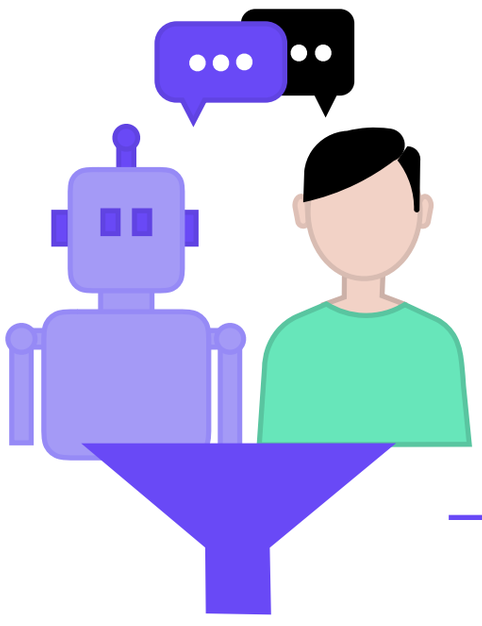


INTERACTION

DATA

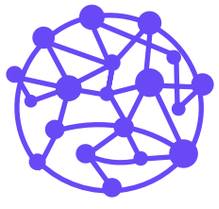
LEARNER MODEL

MOTIVATION

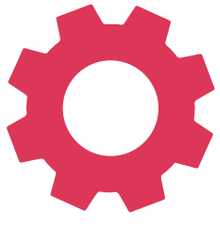


INTERACTION

DATA

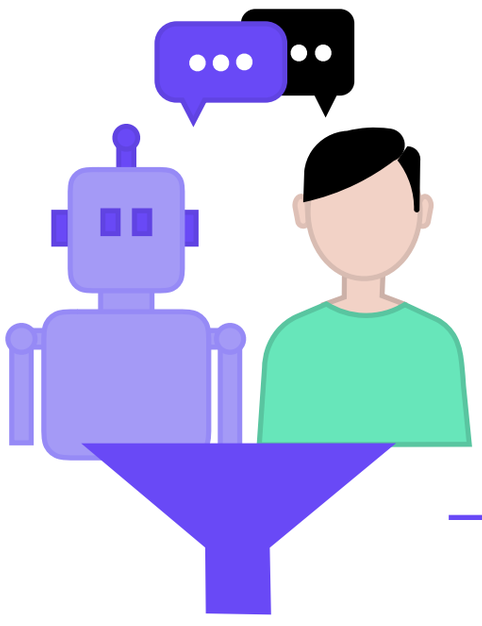


LEARNER MODEL



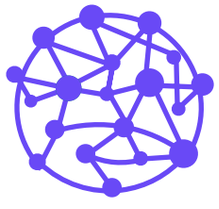
KNOWLEDGE PREDICTION MODEL

MOTIVATION

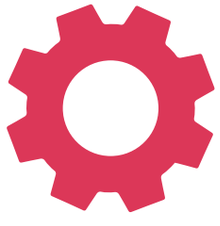


INTERACTION

DATA

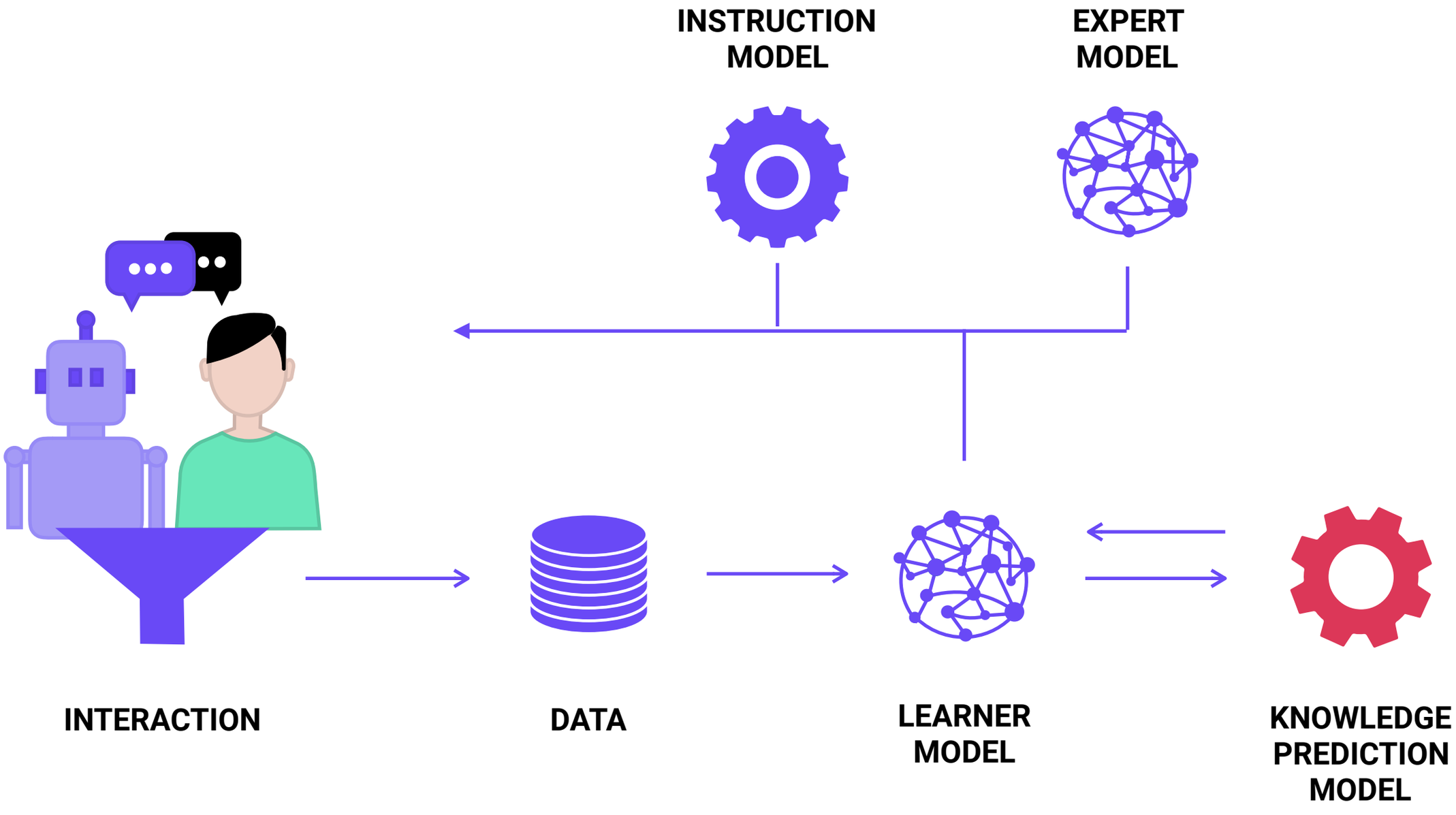


LEARNER MODEL



KNOWLEDGE PREDICTION MODEL

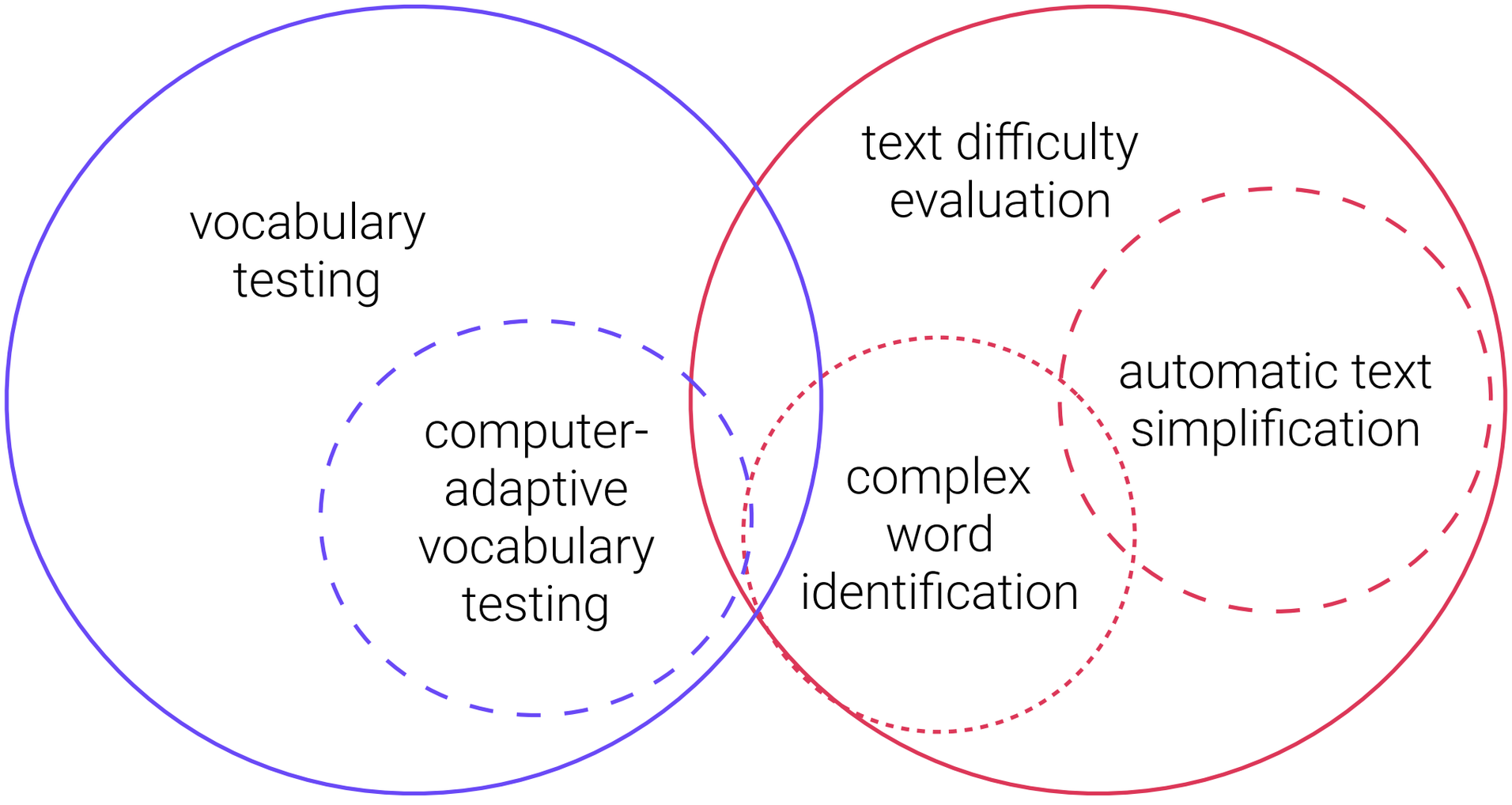
MOTIVATION


INSTRUCTION MODEL
EXPERT MODEL
INTERACTION
DATA
LEARNER MODEL
KNOWLEDGE PREDICTION MODEL

SINGLE-WORD KNOWLEDGE PREDICTION

	Word 1	Word 2	Word 3	Word 4	Word 5	Word …
Learner 1	???	KNOWN	UNKNOWN	???	UNKNOWN	KNOWN
Learner 2	UNKNOWN	???	???	KNOWN	KNOWN	KNOWN
Learner 3	KNOWN	KNOWN	UNKNOWN	UNKNOWN	???	KNOWN
Learner 4	???	KNOWN	???	KNOWN	???	UNKNOWN
Learner 5	UNKNOWN	???	???	???	UNKNOWN	???
Learner 6	KNOWN	???	KNOWN	UNKNOWN	KNOWN	KNOWN
Learner …	UNKNOWN	???	KNOWN	KNOWN	???	???

SINGLE-WORD KNOWLEDGE PREDICTION



RELATED STUDIES

Studies which tried to predict the knowledge of **single words** for **language learners**.

	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study (2019)

RELATED STUDIES

Studies which tried to predict the knowledge of **single words** for **language learners**.

application	text simplification /adaptation	exercise/ material generation	search for reading materials	text simplification	reading support	intelligent tutoring system
Tack et al. (2016)						
Alfter & Volodina (2018)						
Lee & Yeung (2018)						
Yancey & Lepage (2018)						
Ehara et al. (2018)						
This study (2019)						

RELATED STUDIES

Studies which tried to predict the knowledge of **single words** for **language learners**.

	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study (2019)
application	**text simplification /adaptation**	**exercise/ material generation**	**search for reading materials**	**text simplification**	**reading support**	**intelligent tutoring system**
L2 learners	**French**	**Swedish**	**Chinese**	**Korean**	**English**	**English**

RELATED STUDIES

Studies which tried to predict the knowledge of **single words** for **language learners**.

	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study (2019)
application	**text simplification /adaptation**	**exercise/ material generation**	**search for reading materials**	**text simplification**	**reading support**	**intelligent tutoring system**
L2 learners	**French**	**Swedish**	**Chinese**	**Korean**	**English**	**English**
personalized	**NO**	**NO**	**YES**	**NO**	**YES**	**YES**

RELATED STUDIES

Studies which tried to predict the knowledge of **single words** for **language learners**.

	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study (2019)
application	**text simplification /adaptation**	**exercise/ material generation**	**search for reading materials**	**text simplification**	**reading support**	**intelligent tutoring system**
L2 learners	**French**	**Swedish**	**Chinese**	**Korean**	**English**	**English**
personalized	**NO**	**NO**	**YES**	**NO**	**YES**	**YES**
machine learning	**NO**	**YES**	**YES**	**YES**	**NO**	**YES**

RELATED STUDIES

Studies which tried to predict the knowledge of **single words** for **language learners**.

application	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study (2019)
L2 learners	French	Swedish	Chinese	Korean	English	English
personalized	NO	NO	YES	NO	YES	YES
machine learning	NO	YES	YES	YES	NO	YES
unseen words	NO	NO	YES	YES	YES	YES

RELATED STUDIES

Studies which tried to predict the knowledge of **single words** for **language learners**.

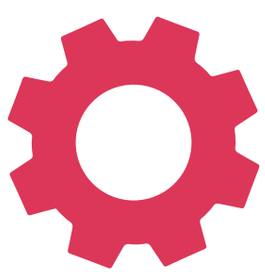
application	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study (2019)
L2 learners	French	Swedish	Chinese	Korean	English	English
personalized	NO	NO	YES	NO	YES	YES
machine learning	NO	YES	YES	YES	NO	YES
unseen words	NO	NO	YES	YES	YES	YES
all-learner-together training	-	-	NO	-	-	YES

OUR STUDY

OUR STUDY

datatset

features



models

DATASET

Ehara (2012)

	Word 1	Word 2	Word 3	Word 4	Word 5	Word …
Learner 1						
Learner 2						
Learner 3						
Learner 4						
Learner 5						
Learner …						

DATASET

Ehara (2012)

16 Japanese learners x **12,000 words** from the SVL list

	Word 1	Word 2	Word 3	Word 4	Word 5	Word …
Learner 1						
Learner 2						
Learner 3						
Learner 4						
Learner 5						
Learner …						

DATASET

Ehara (2012)

16 Japanese learners x **12,000 words** from the SVL list

Knowledge scale from “Never seen the word” to “Know the word’s meaning”

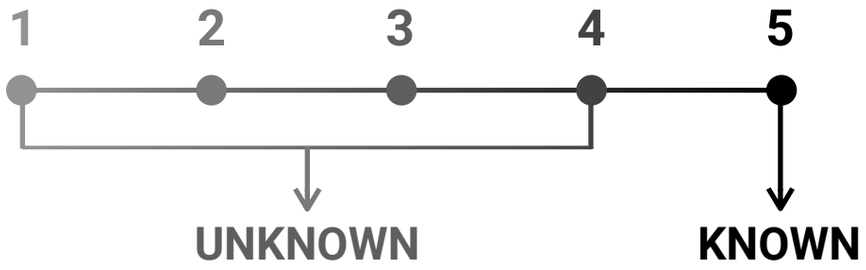
	Word 1	Word 2	Word 3	Word 4	Word 5	Word …
Learner 1	5	3	5	1	5	3
Learner 2	2	5	2	2	4	5
Learner 3	3	5	1	4	1	1
Learner 4	4	5	5	1	2	2
Learner 5	3	4	5	3	5	5
Learner …	2	4	2	4	2	5

DATASET

Ehara (2012)

16 Japanese learners x **12,000 words** from the SVL list

Knowledge scale from “Never seen the word” to “Know the word’s meaning”



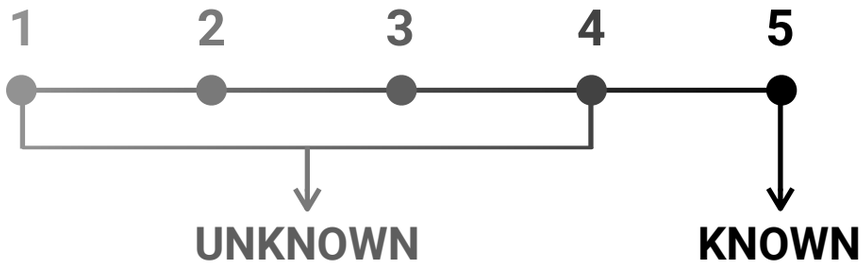
	Word 1	Word 2	Word 3	Word 4	Word 5	Word …
Learner 1	5	3	5	1	5	3
Learner 2	2	5	2	2	4	5
Learner 3	3	5	1	4	1	1
Learner 4	4	5	5	1	2	2
Learner 5	3	4	5	3	5	5
Learner …	2	4	2	4	2	5

DATASET

Ehara (2012)

16 Japanese learners x **12,000 words** from the SVL list

Knowledge scale from “Never seen the word” to “Know the word’s meaning”



	Word 1	Word 2	Word 3	Word 4	Word 5	Word …
Learner 1	KNOWN	UNKNOWN	KNOWN	UNKNOWN	KNOWN	UNKNOWN
Learner 2	UNKNOWN	KNOWN	UNKNOWN	UNKNOWN	UNKNOWN	KNOWN
Learner 3	UNKNOWN	KNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN
Learner 4	UNKNOWN	KNOWN	KNOWN	UNKNOWN	UNKNOWN	UNKNOWN
Learner 5	UNKNOWN	UNKNOWN	KNOWN	UNKNOWN	KNOWN	KNOWN
Learner …	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN	KNOWN

FEATURES

	Feature 1	Feature 2	...	Feature 23	Feature 24	Feature 25	...	Feature 78	Label
Learner 1 x **Word 1**									**KNOWN**
Learner 1 x **Word 2**									**KNOWN**
Learner 1 x **Word 3**									**UNKNOWN**
Learner 2 x **Word 1**									**KNOWN**
Learner 2 x **Word 2**									**UNKNOWN**
Learner 2 x **Word ...**									**UNKNOWN**

FEATURES

Word-specific

	Feature 1	Feature 2	...	Feature 23	Feature 24	Feature 25	...	Feature 78	Label
Learner 1 x **Word 1**									**KNOWN**
Learner 1 x **Word 2**									**KNOWN**
Learner 1 x **Word 3**									**UNKNOWN**
Learner 2 x **Word 1**									**KNOWN**
Learner 2 x **Word 2**									**UNKNOWN**
Learner 2 x **Word ...**									**UNKNOWN**

FEATURES

	Word-specific				Learner-specific				
	Feature 1	Feature 2	...	Feature 23	Feature 24	Feature 25	...	Feature 78	Label
Learner 1 x **Word 1**									KNOWN
Learner 1 x **Word 2**									KNOWN
Learner 1 x **Word 3**									UNKNOWN
Learner 2 x **Word 1**									KNOWN
Learner 2 x **Word 2**									UNKNOWN
Learner 2 x **Word ...**									UNKNOWN

FEATURES

Word-specific

	History	Movies	...	Feature 23	Feature 24	Feature 25	...	Feature 78	Label
Learner 1 x **Word 1**									**KNOWN**
Learner 1 x **Word 2**									**KNOWN**
Learner 1 x **Word 3**									**UNKNOWN**
Learner 2 x **Word 1**									**KNOWN**
Learner 2 x **Word 2**									**UNKNOWN**
Learner 2 x **Word ...**									**UNKNOWN**

FEATURES

normalized frequency of the target word in different genres

Word-specific

	History	Movies	...	Feature 23	Feature 24	Feature 25	...	Feature 78	Label
Learner 1 x **Word 1**	345261	568432							KNOWN
Learner 1 x **Word 2**	7432	6930							KNOWN
Learner 1 x **Word 3**	90234112	89996010							UNKNOWN
Learner 2 x **Word 1**	345261	568432							KNOWN
Learner 2 x **Word 2**	7432	6930							UNKNOWN
Learner 2 x **Word ...**	90234112	89996010							UNKNOWN

FEATURES

normalized frequency of the target word in different genres ⟶ COCA corpus (Davies, 2008)

Word-specific

	History	Movies	...	Feature 23	Feature 24	Feature 25	...	Feature 78	Label
Learner 1 x **Word 1**	345261	568432							KNOWN
Learner 1 x **Word 2**	7432	6930							KNOWN
Learner 1 x **Word 3**	90234112	89996010							UNKNOWN
Learner 2 x **Word 1**	345261	568432							KNOWN
Learner 2 x **Word 2**	7432	6930							UNKNOWN
Learner 2 x **Word ...**	90234112	89996010							UNKNOWN

FEATURES

normalized frequency of the target word in different genres ⟶ COCA corpus (Davies, 2008)
psycholinguistic characteristics of the word

Word-specific

	History	Movies	…	Meaningful	Feature 24	Feature 25	…	Feature 78	Label
Learner 1 x **Word 1**	345261	568432	…	3					KNOWN
Learner 1 x **Word 2**	7432	6930	…	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3					KNOWN
Learner 2 x **Word 2**	7432	6930	…	5					UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2					UNKNOWN

FEATURES

normalized frequency of the target word in different genres ⟶ COCA corpus (Davies, 2008)
psycholinguistic characteristics of the word ⟶ MRC psycholinguistic database

Word-specific

	History	Movies	…	Meaningful	Feature 24	Feature 25	…	Feature 78	Label
Learner 1 x **Word 1**	345261	568432	…	3					KNOWN
Learner 1 x **Word 2**	7432	6930	…	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3					KNOWN
Learner 2 x **Word 2**	7432	6930	…	5					UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2					UNKNOWN

FEATURES

normalized frequency of the target word in different genres ⟶ COCA corpus (Davies, 2008)
psycholinguistic characteristics of the word ⟶ MRC psycholinguistic database

Word-specific

	History	Movies	…	Meaningful	Feature 24	Feature 25	…	Feature 78	Label
Learner 1 x **Word 1**	345261	568432	…	3					KNOWN
Learner 1 x **Word 2**	7432	6930	…	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3					KNOWN
Learner 2 x **Word 2**	7432	6930	…	5					UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2					UNKNOWN

FEATURES

normalized frequency of the target word in different genres ⟶ COCA corpus (Davies, 2008)
psycholinguistic characteristics of the word ⟶ MRC psycholinguistic database

Word-specific

	History	Movies	…	Meaningful	Feature 24	Feature 25	…	Feature 78	Label
Learner 1 x **Word 1**	345261	568432	…	3					KNOWN
Learner 1 x **Word 2**	7432	6930	…	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3					KNOWN
Learner 2 x **Word 2**	7432	6930	…	5					UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2					UNKNOWN

FEATURES

	Word-specific				Learner-specific				
	History	Movies	...	Meaningful	Feature 24	Feature 25	...	Feature 78	Label
Learner 1 x **Word 1**	345261	568432	...	3					KNOWN
Learner 1 x **Word 2**	7432	6930	...	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	...	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	...	3					KNOWN
Learner 2 x **Word 2**	7432	6930	...	5					UNKNOWN
Learner 2 x **Word ...**	90234112	89996010	...	2					UNKNOWN

FEATURES

learner's knowledge of vocabulary in different genres

	Word-specific				Learner-specific				
	History	Movies	...	Meaningful	History	Movies	...	Sports	Label
Learner 1 x **Word 1**	345261	568432	...	3					KNOWN
Learner 1 x **Word 2**	7432	6930	...	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	...	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	...	3					KNOWN
Learner 2 x **Word 2**	7432	6930	...	5					UNKNOWN
Learner 2 x **Word ...**	90234112	89996010	...	2					UNKNOWN

FEATURES

learner's knowledge of vocabulary in different genres and frequencies

	Word-specific				Learner-specific				
	History	Movies	...	Meaningful	History	Movies	...	Sports	Label
Learner 1 x **Word 1**	345261	568432	...	3					KNOWN
Learner 1 x **Word 2**	7432	6930	...	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	...	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	...	3					KNOWN
Learner 2 x **Word 2**	7432	6930	...	5					UNKNOWN
Learner 2 x **Word ...**	90234112	89996010	...	2					UNKNOWN

FEATURES

learner's knowledge of vocabulary in different genres and frequencies

⟶ COCA corpus (Davies, 2008)

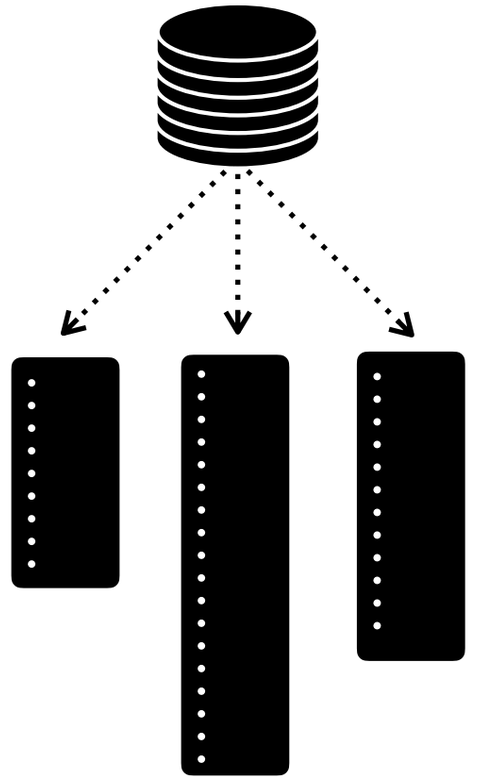
	Word-specific				Learner-specific				
	History	Movies	...	Meaningful	History	Movies	...	Sports	Label
Learner 1 x **Word 1**	345261	568432	...	3					KNOWN
Learner 1 x **Word 2**	7432	6930	...	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	...	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	...	3					KNOWN
Learner 2 x **Word 2**	7432	6930	...	5					UNKNOWN
Learner 2 x **Word ...**	90234112	89996010	...	2					UNKNOWN

FEATURES

learner's knowledge of vocabulary in different genres and frequencies

⟶ COCA corpus (Davies, 2008)

1. *For each COCA subcorpus, keywordlist created*



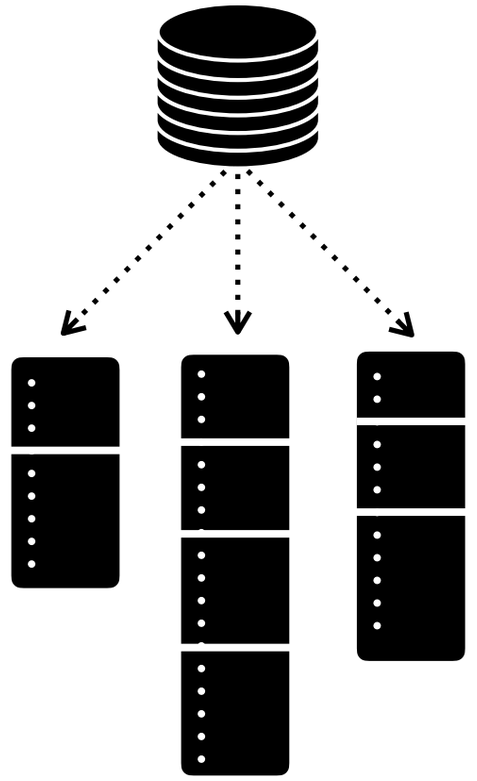
	Word-specific				Learner-specific				
	History	Movies	…	Meaningful	History	Movies	…	Sports	Label
Learner 1 x **Word 1**	345261	568432	…	3					KNOWN
Learner 1 x **Word 2**	7432	6930	…	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3					KNOWN
Learner 2 x **Word 2**	7432	6930	…	5					UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2					UNKNOWN

FEATURES

learner's knowledge of vocabulary in different genres and frequencies

⟶ COCA corpus (Davies, 2008)

1. *For each COCA subcorpus, keywordlist created*
2. *For each keywordlist, frequency bands created*



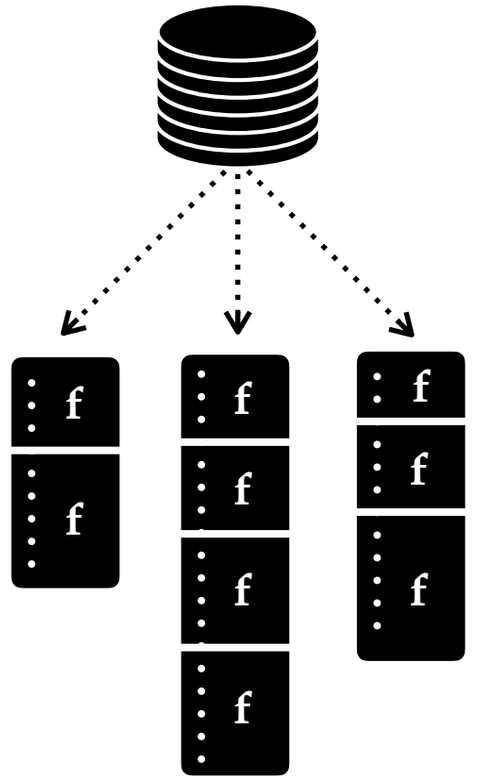
	Word-specific				Learner-specific				
	History	Movies	…	Meaningful	History	Movies	…	Sports	Label
Learner 1 x **Word 1**	345261	568432	…	3					KNOWN
Learner 1 x **Word 2**	7432	6930	…	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3					KNOWN
Learner 2 x **Word 2**	7432	6930	…	5					UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2					UNKNOWN

FEATURES

learner's knowledge of vocabulary in different genres and frequencies

⟶ COCA corpus (Davies, 2008)

1. *For each COCA subcorpus, keywordlist created*
2. *For each keywordlist, frequency bands created*
3. *For each freqeuncy band, "frequency power" calculated*



	Word-specific				Learner-specific				
	History	Movies	…	Meaningful	History	Movies	…	Sports	Label
Learner 1 x **Word 1**	345261	568432	…	3					KNOWN
Learner 1 x **Word 2**	7432	6930	…	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3					KNOWN
Learner 2 x **Word 2**	7432	6930	…	5					UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2					UNKNOWN

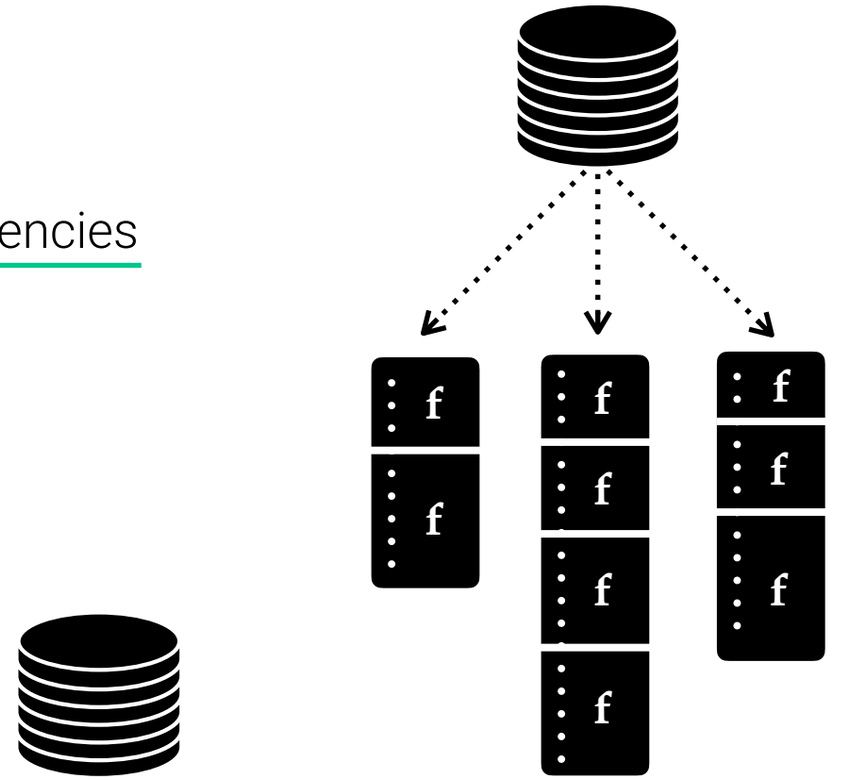
FEATURES

learner's knowledge of vocabulary in different genres and frequencies

⟶ COCA corpus (Davies, 2008)

1. *For each COCA subcorpus, keywordlist created*
2. *For each keywordlist, frequency bands created*
3. *For each freqeuncy band, "frequency power" calculated*

⟶ Vocabulary Knowledge Dataset (Ehara, 2012)



	Word-specific				Learner-specific				
	History	Movies	…	Meaningful	History	Movies	…	Sports	Label
Learner 1 x **Word 1**	345261	568432	…	3					KNOWN
Learner 1 x **Word 2**	7432	6930	…	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3					KNOWN
Learner 2 x **Word 2**	7432	6930	…	5					UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2					UNKNOWN

FEATURES

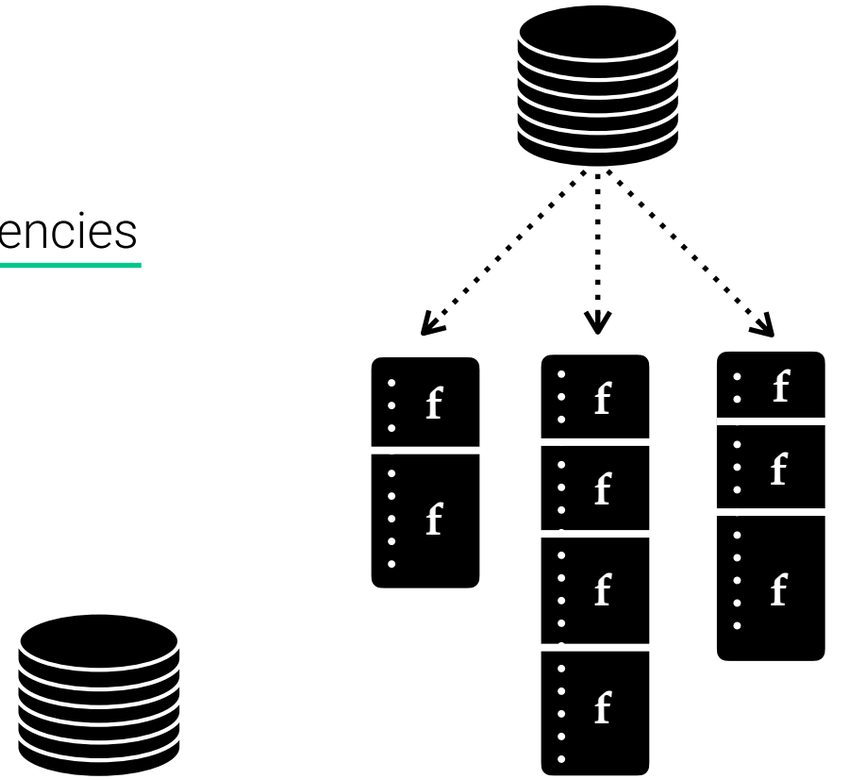
learner's knowledge of vocabulary in different genres and frequencies

⟶ COCA corpus (Davies, 2008)

1. *For each COCA subcorpus, keywordlist created*
2. *For each keywordlist, frequency bands created*
3. *For each freqeuncy band, "frequency power" calculated*

⟶ Vocabulary Knowledge Dataset (Ehara, 2012)

4. *For each keywordlist, learner's knowledge calculated*



	Word-specific				Learner-specific				
	History	Movies	…	Meaningful	History	Movies	…	Sports	Label
Learner 1 x **Word 1**	345261	568432	…	3					KNOWN
Learner 1 x **Word 2**	7432	6930	…	5					KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2					UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3					KNOWN
Learner 2 x **Word 2**	7432	6930	…	5					UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2					UNKNOWN

FEATURES

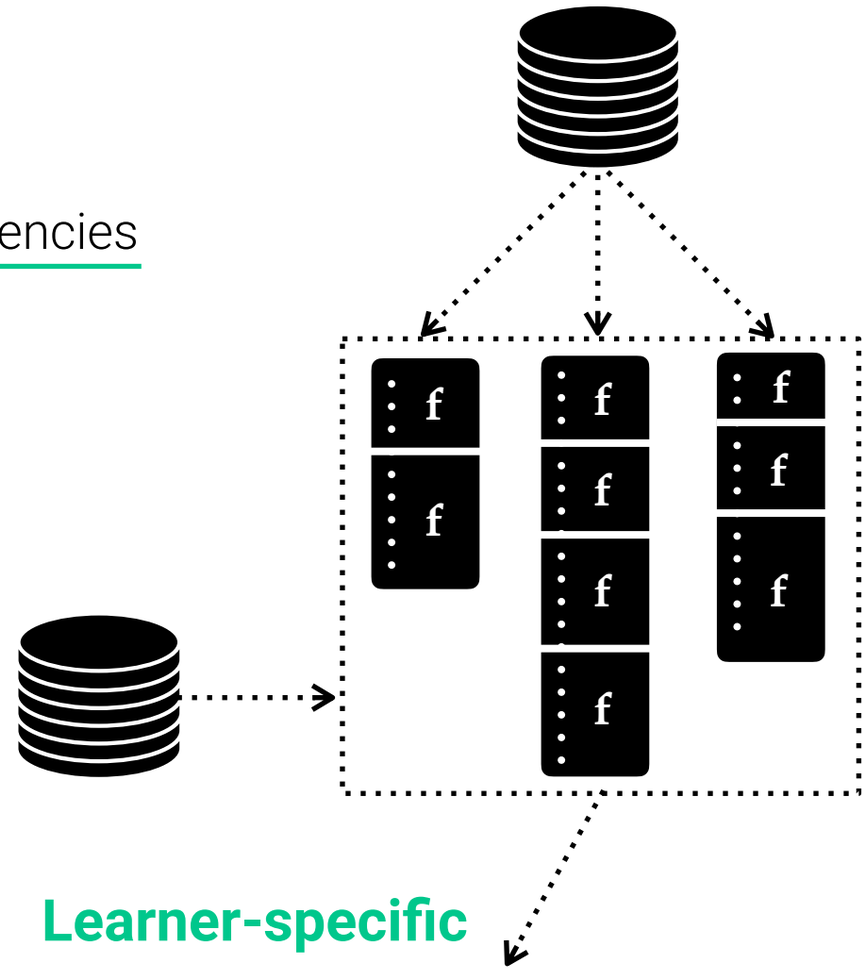
learner's knowledge of vocabulary in different genres and frequencies

⟶ COCA corpus (Davies, 2008)

1. *For each COCA subcorpus, keywordlist created*
2. *For each keywordlist, frequency bands created*
3. *For each freqeuncy band, "frequency power" calculated*

⟶ Vocabulary Knowledge Dataset (Ehara, 2012)

4. *For each keywordlist, learner's knowledge calculated*



	Word-specific				Learner-specific				
	History	Movies	…	Meaningful	History	Movies	…	Sports	Label
Learner 1 x **Word 1**	345261	568432	…	3	40852	2526	…	1679	KNOWN
Learner 1 x **Word 2**	7432	6930	…	5	40852	2526	…	1679	KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2	40852	2526	…	1679	UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3	75679	99781	…	87540	KNOWN
Learner 2 x **Word 2**	7432	6930	…	5	75679	99781	…	87540	UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2	75679	99781	…	87540	UNKNOWN

FEATURES

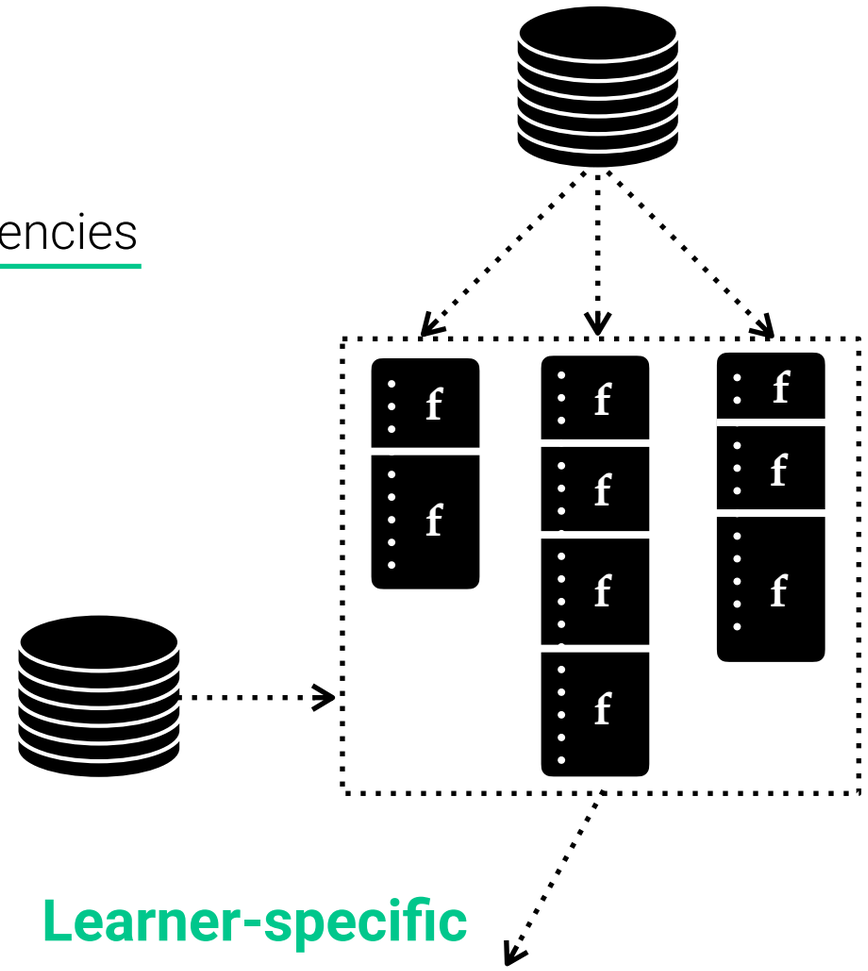
learner's knowledge of vocabulary in different genres and frequencies

⟶ COCA corpus (Davies, 2008)

1. *For each COCA subcorpus, keywordlist created*
2. *For each keywordlist, frequency bands created*
3. *For each freqeuncy band, "frequency power" calculated*

⟶ Vocabulary Knowledge Dataset (Ehara, 2012)

4. *For each keywordlist, learner's knowledge calculated*



	Word-specific				Learner-specific				
	History	Movies	…	Meaningful	History	Movies	…	Sports	Label
Learner 1 x **Word 1**	345261	568432	…	3	40852	2526	…	1679	KNOWN
Learner 1 x **Word 2**	7432	6930	…	5	40852	2526	…	1679	KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2	40852	2526	…	1679	UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3	75679	99781	…	87540	KNOWN
Learner 2 x **Word 2**	7432	6930	…	5	75679	99781	…	87540	UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2	75679	99781	…	87540	UNKNOWN

FEATURES

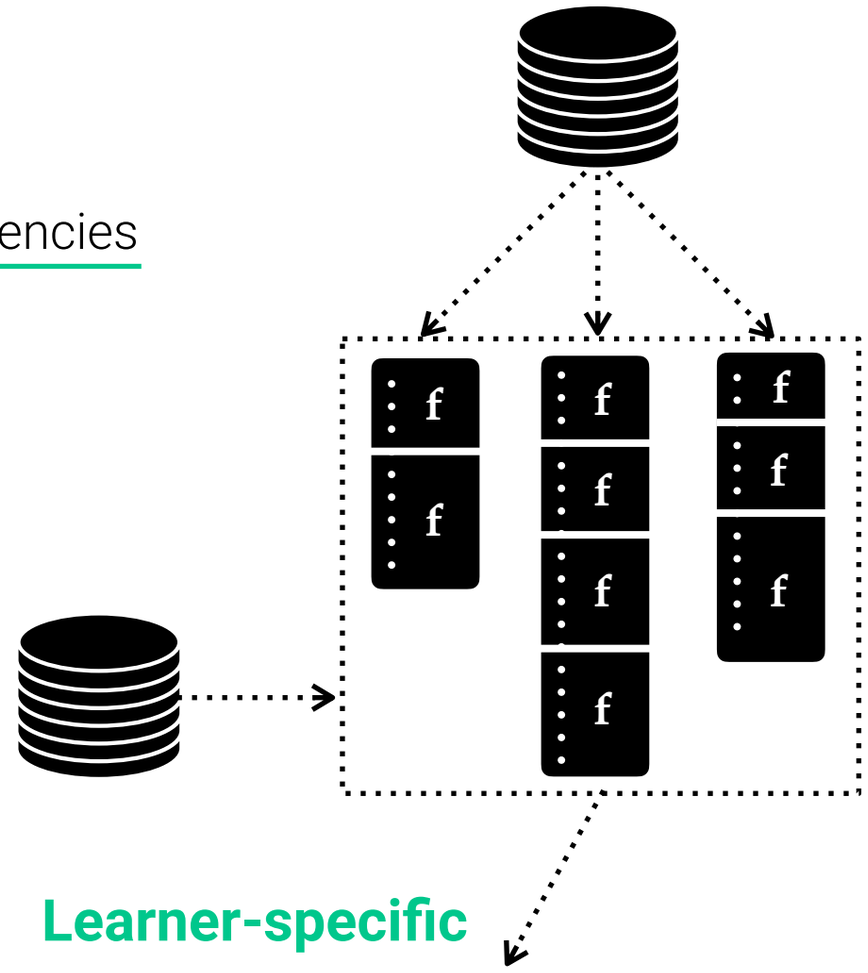
learner's knowledge of vocabulary in different genres and frequencies

⟶ COCA corpus (Davies, 2008)

1. *For each COCA subcorpus, keywordlist created*
2. *For each keywordlist, frequency bands created*
3. *For each freqeuncy band, "frequency power" calculated*

⟶ Vocabulary Knowledge Dataset (Ehara, 2012)

4. *For each keywordlist, learner's knowledge calculated*



	Word-specific				Learner-specific				
	History	Movies	…	Meaningful	History	Movies	…	Sports	Label
Learner 1 x **Word 1**	345261	568432	…	3	40852	2526	…	1679	KNOWN
Learner 1 x **Word 2**	7432	6930	…	5	40852	2526	…	1679	KNOWN
Learner 1 x **Word 3**	90234112	89996010	…	2	40852	2526	…	1679	UNKNOWN
Learner 2 x **Word 1**	345261	568432	…	3	75679	99781	…	87540	KNOWN
Learner 2 x **Word 2**	7432	6930	…	5	75679	99781	…	87540	UNKNOWN
Learner 2 x **Word …**	90234112	89996010	…	2	75679	99781	…	87540	UNKNOWN

FEATURES

105 features in total

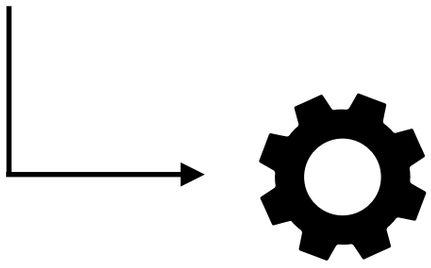
	Word-specific			Learner-specific				
	History	Movies	...	Meaningful	History	Movies	...	Sports	Label
Learner 1 x **Word 1**	345261	568432	...	3	40852	2526	...	1679	KNOWN
Learner 1 x **Word 2**	7432	6930	...	5	40852	2526	...	1679	KNOWN
Learner 1 x **Word 3**	90234112	89996010	...	2	40852	2526	...	1679	UNKNOWN
Learner 2 x **Word 1**	345261	568432	...	3	75679	99781	...	87540	KNOWN
Learner 2 x **Word 2**	7432	6930	...	5	75679	99781	...	87540	UNKNOWN
Learner 2 x **Word ...**	90234112	89996010	...	2	75679	99781	...	87540	UNKNOWN

FEATURE SELECTION

105 features in total

FEATURE SELECTION

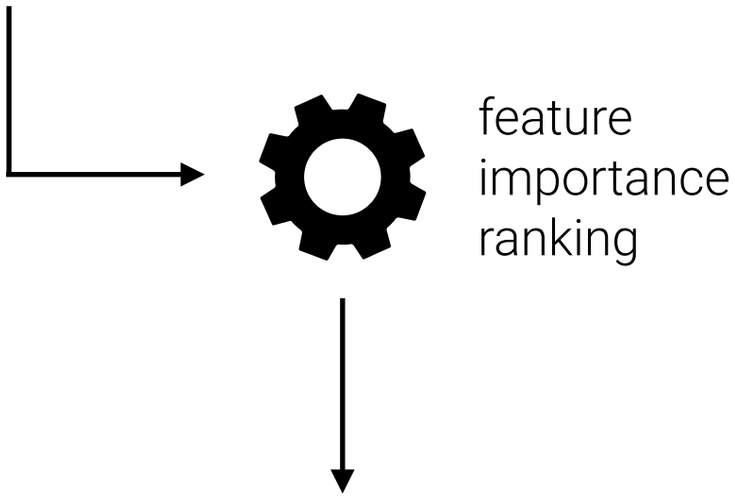
105 features in total



feature
importance
ranking

FEATURE SELECTION

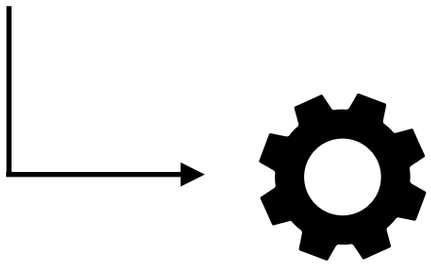
105 features in total



RANK	TYPE	NAME
1	**word**	total
2	**word**	spoken (national public radio)
3	**word**	academic
4	**word**	magazine (women/men)
5	**word**	fiction (journals)
...	...	...
101	**learner**	magazine (African American)
102	**learner**	magazine (home/health)
103	**learner**	magazine (science/tech)
104	**learner**	news (editorial)
105	**learner**	academic (history)

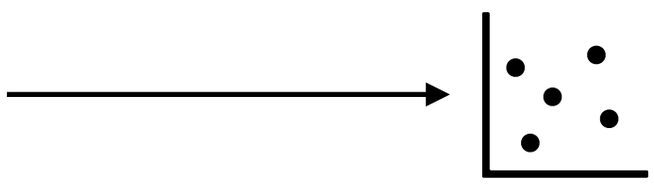
FEATURE SELECTION

105 features in total



feature importance ranking

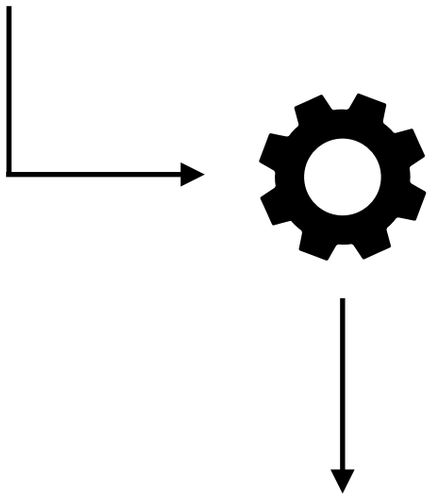
RANK	TYPE	NAME
1	**word**	total
2	**word**	spoken (national public radio)
3	**word**	academic
4	**word**	magazine (women/men)
5	**word**	fiction (journals)
...	...	...
101	**learner**	magazine (African American)
102	**learner**	magazine (home/health)
103	**learner**	magazine (science/tech)
104	**learner**	news (editorial)
105	**learner**	academic (history)



Pearson correlation
If >0.99, keep the top one

FEATURE SELECTION

105 features in total

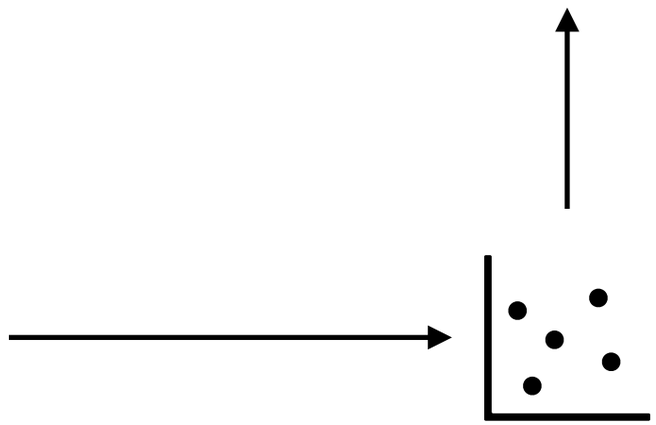


feature importance ranking

RANK	TYPE	NAME
1	**word**	total
2	**word**	spoken (national public radio)
3	**word**	academic
4	**word**	magazine (women/men)
5	**word**	fiction (journals)
...	...	...
101	**learner**	magazine (African American)
102	**learner**	magazine (home/health)
103	**learner**	magazine (science/tech)
104	**learner**	news (editorial)
105	**learner**	academic (history)

39 features in total

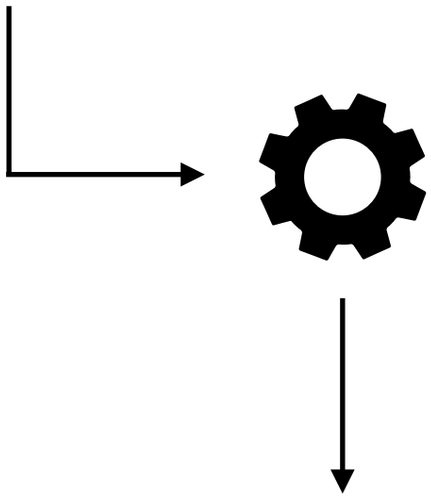
RANK	TYPE	NAME
1	**word**	total
2	**word**	spoken (national public radio)
3	**word**	academic
4	**word**	magazine (women/men)
5	**word**	fiction (journals)
...	...	...
35	**learner**	news (local)
36	**learner**	spoken (independent)
37	**learner**	fiction
38	**learner**	spoken (public broadcasting)
39	**learner**	spoken (national public radio)



Pearson correlation
If >0.99, keep the top one

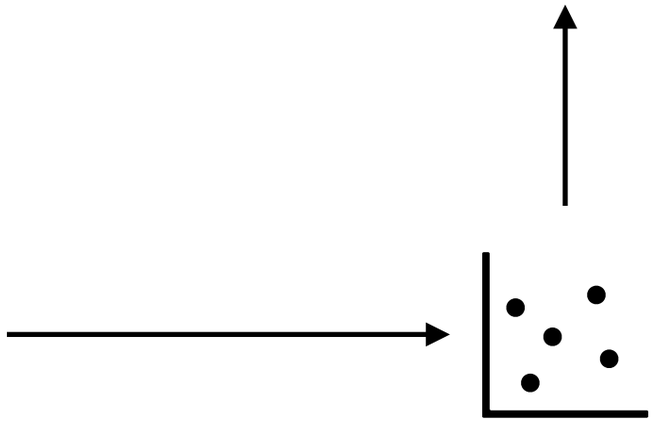
FEATURE SELECTION

105 features in total



feature importance ranking

RANK	TYPE	NAME
1	**word**	total
2	**word**	spoken (national public radio)
3	**word**	academic
4	**word**	magazine (women/men)
5	**word**	fiction (journals)
...	...	...
101	**learner**	magazine (African American)
102	**learner**	magazine (home/health)
103	**learner**	magazine (science/tech)
104	**learner**	news (editorial)
105	**learner**	academic (history)



Pearson correlation
If >0.99, keep the top one

39 features in total

RANK	TYPE	NAME
1	**word**	total
2	**word**	spoken (national public radio)
3	**word**	academic
4	**word**	magazine (women/men)
5	**word**	fiction (journals)
...	...	...
35	**learner**	news (local)
36	**learner**	spoken (independent)
37	**learner**	fiction
38	**learner**	spoken (public broadcasting)
39	**learner**	spoken (national public radio)

MODEL

Complete Feature Space Dependent

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

Random Forest
Support Vector Machine
k Nearest Neighbors
Logistic Regression.

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

Random Forest
~~Support Vector Machine~~
k Nearest Neighbors
~~Logistic Regression.~~

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

Random Forest
~~Support Vector Machine~~
~~k Nearest Neighbors~~
~~Logistic Regression.~~

80% - 3% - 17%
training - validation - testing

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

Random Forest
~~Support Vector Machine~~
~~k Nearest Neighbors~~
~~Logistic Regression.~~

80% - 3% - 17%
training - validation - testing

Neural Network Based

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

Random Forest
~~Support Vector Machine~~
~~k Nearest Neighbors~~
~~Logistic Regression.~~

80% - 3% - 17%
training - validation - testing

Neural Network Based

word-specific

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

Random Forest
~~Support Vector Machine~~
~~k Nearest Neighbors~~
~~Logistic Regression.~~

80% - 3% - 17%
training - validation - testing

Neural Network Based

word-specific

Fully Connected Neural Network
5 hidden layers

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

Random Forest
~~Support Vector Machine~~
~~k Nearest Neighbors~~
~~Logistic Regression.~~

80% - 3% - 17%
training - validation - testing

Neural Network Based

word-specific
learner representation

Fully Connected Neural Network
5 hidden layers

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

Random Forest
~~Support Vector Machine~~
~~k Nearest Neighbors~~
~~Logistic Regression.~~

80% - 3% - 17%
training - validation - testing

Neural Network Based

word-specific
learner representation

Fully Connected Neural Network
5 hidden layers

80% - 3% - 17%
training - validation - testing

MODEL

Complete Feature Space Dependent

word-specific
learner-specific

Random Forest
~~Support Vector Machine~~
~~k Nearest Neighbors~~
~~Logistic Regression.~~

80% - 3% - 17%
training - validation - testing

79.89%

ACC

Neural Network Based

word-specific
learner representation

Fully Connected Neural Network
5 hidden layers

80% - 3% - 17%
training - validation - testing

79.18%

ACC

COMPARISON TO OTHER STUDIES

	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study
L2 learners	French	Swedish	Chinese	Korean	English	English
personalized	NO	NO	YES	NO	YES	YES
machine learning	NO	YES	YES	YES	NO	YES
for unseen words	NO	NO	YES	YES	YES	YES
all-learner-together training	-	-	NO	-	YES	YES
features						
models & results (accuracy)						

COMPARISON TO OTHER STUDIES

	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study
L2 learners	**French**	**Swedish**	**Chinese**	**Korean**	**English**	**English**
personalized	**NO**	**NO**	**YES**	**NO**	**YES**	**YES**
machine learning	**NO**	**YES**	**YES**	**YES**	**NO**	**YES**
for unseen words	**NO**	**NO**	**YES**	**YES**	**YES**	**YES**
all-learner-together training	**-**	**-**	**NO**	**-**	**YES**	**YES**
features	**-**	**count-based, morphological, semantic, context-based**	**difficulty level, frequency, count-based**	**frequency, count-based, proficiency level**	**frequency in 6 corpora, learner ability learnt**	**freq, genre, learner's knowledge in genres**
models & results (accuracy)	**FLELex 92.3%** (only for known words)	**SVM 48% MLPerc 53% ExtraTree 59%**	**SVM 78.0% Oracle 79.1%**	**SVM Linear 83.0% SVM RBF 84.3%**	**Rasch 66.3% Shared 77.7% Proposd 77.8%**	**RForest 79.89% NNet 79.18%**

COMPARISON TO OTHER STUDIES

	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study
L2 learners	French	Swedish	Chinese	Korean	English	English
personalized	NO	NO	YES	NO	YES	YES
machine learning	NO	YES	YES	YES	NO	YES
for unseen words	NO	NO	YES	YES	YES	YES
all-learner-together training	-	-	NO	-	YES	YES
features	-	count-based, morphological, semantic, context-based	difficulty level, frequency, count-based	frequency, count-based, proficiency level	frequency in 6 corpora, learner ability learnt	freq, genre, learner's knowledge in genres
models & results (accuracy)	FLELex 92.3% (only for known words)	SVM 48% MLPerc 53% ExtraTree 59%	SVM 78.0% Oracle 79.1%	SVM Linear 83.0% SVM RBF 84.3%	Rasch 66.3% Shared 77.7% Proposd 77.8%	RForest 79.89% NNet 79.18%

COMPARISON TO OTHER STUDIES

	Tack et al. (2016)	Alfter & Volodina (2018)	Lee & Yeung (2018)	Yancey & Lepage (2018)	Ehara et al. (2018)	This study
L2 learners	French	Swedish	Chinese	Korean	English	English
personalized	NO	NO	YES	NO	YES	YES
machine learning	NO	YES	YES	YES	NO	YES
for unseen words	NO	NO	YES	YES	YES	YES
all-learner-together training	-	-	NO	-	YES	YES
features	-	count-based, morphological, semantic, context-based	difficulty level, frequency, count-based	frequency, count-based, proficiency level	frequency in 6 corpora, learner ability learnt	freq, genre, learner's knowledge in genres
models & results (accuracy)	FLELex 92.3% (only for known words)	SVM 48% MLPerc 53% ExtraTree 59%	SVM 78.0% Oracle 79.1%	SVM Linear 83.0% SVM RBF 84.3%	Rasch 66.3% Shared 77.7% Proposd 77.8%	RForest 79.89% NNet 79.18%

FUTURE WORK

FUTURE WORK

- adaptation & integration of this model to an intelligent system Elia

FUTURE WORK

- adaptation & integration of this model to an intelligent system Elia
- data size effect on the accuracy of the model

FUTURE WORK

- adaptation & integration of this model to an intelligent system Elia
- data size effect on the accuracy of the model

- collection & creation of a new dataset for single-word knowledge prediction
  - more learnes
  - less words
  - different levels of knowledge
  - different aspects of knowledge

FUTURE WORK

- adaptation & integration of this model to an intelligent system Elia
- data size effect on the accuracy of the model

  - collection & creation of a new dataset for single-word knowledge prediction
    - more learnes
    - less words
    - different levels of knowledge
    - different aspects of knowledge
  - more studies
    - different features
    - different algorithms

THANK YOU

Klara Ptacinova Klimcikova (k.klimcikova@lmu.de)
Drilon Avdiu (drilon.avdiu@tum.de)
Vanessa Bui (vanessa.bui@tum.de)

Ludwig Maximilians University
Technical University

Munich, Germany