THE 8TH INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING (IJCNLP 2017)

CWIG3G2 – COMPLEX WORD IDENTIFICATION TASK ACROSS THREE TEXT GENRES AND TWO USER GROUPS

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

SEID MUHIE YIMAM, SANJA ŠTAJNER
MARTIN RIEDL, AND CHRIS BIEMANN

UNIVERSITY OF MANNHEIM

Language Technology Group (LT)

Data and Web Science Group

NOVEMBER 29, 2017

Introduction

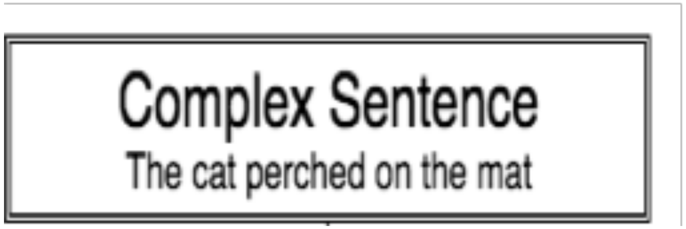
- Complex Word Identification (CWI) is important in lexical simplification
- Lexical simplification (LS):
  - Replace **infrequent** and **difficult phrases**
  - Target readers:

language learners

children

reading impairments

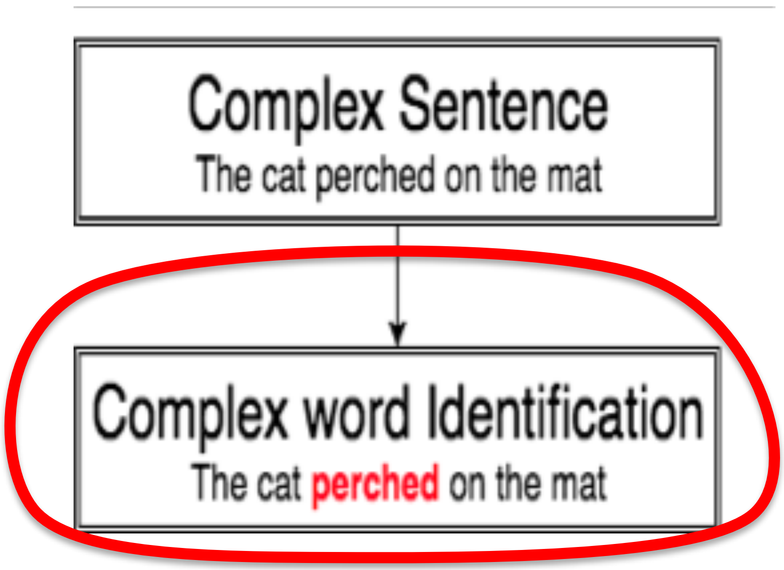
Components of LS (Paetzold, Gustavo (2015)



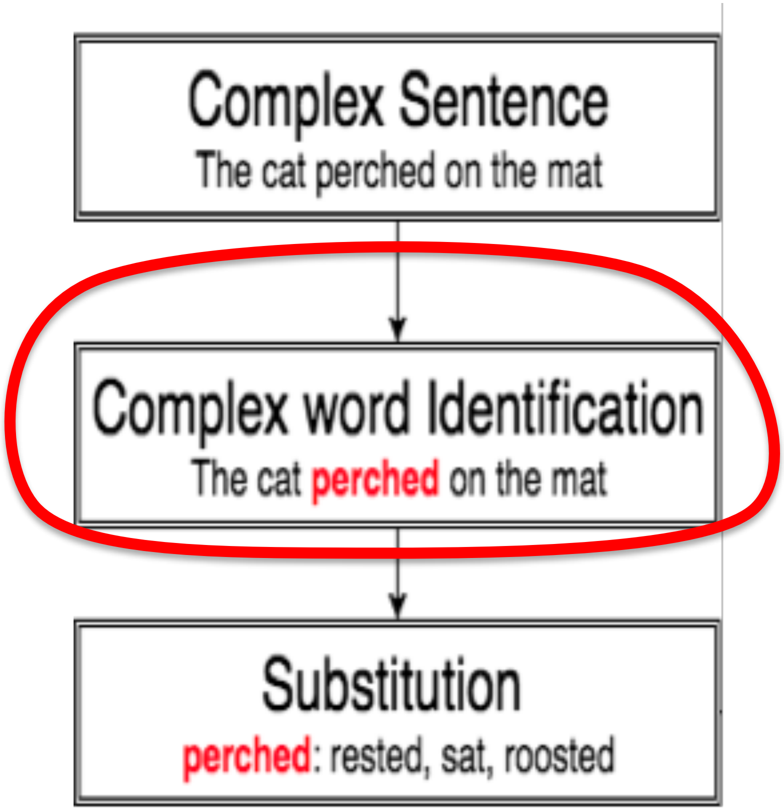
Complex Sentence

The cat perched on the mat

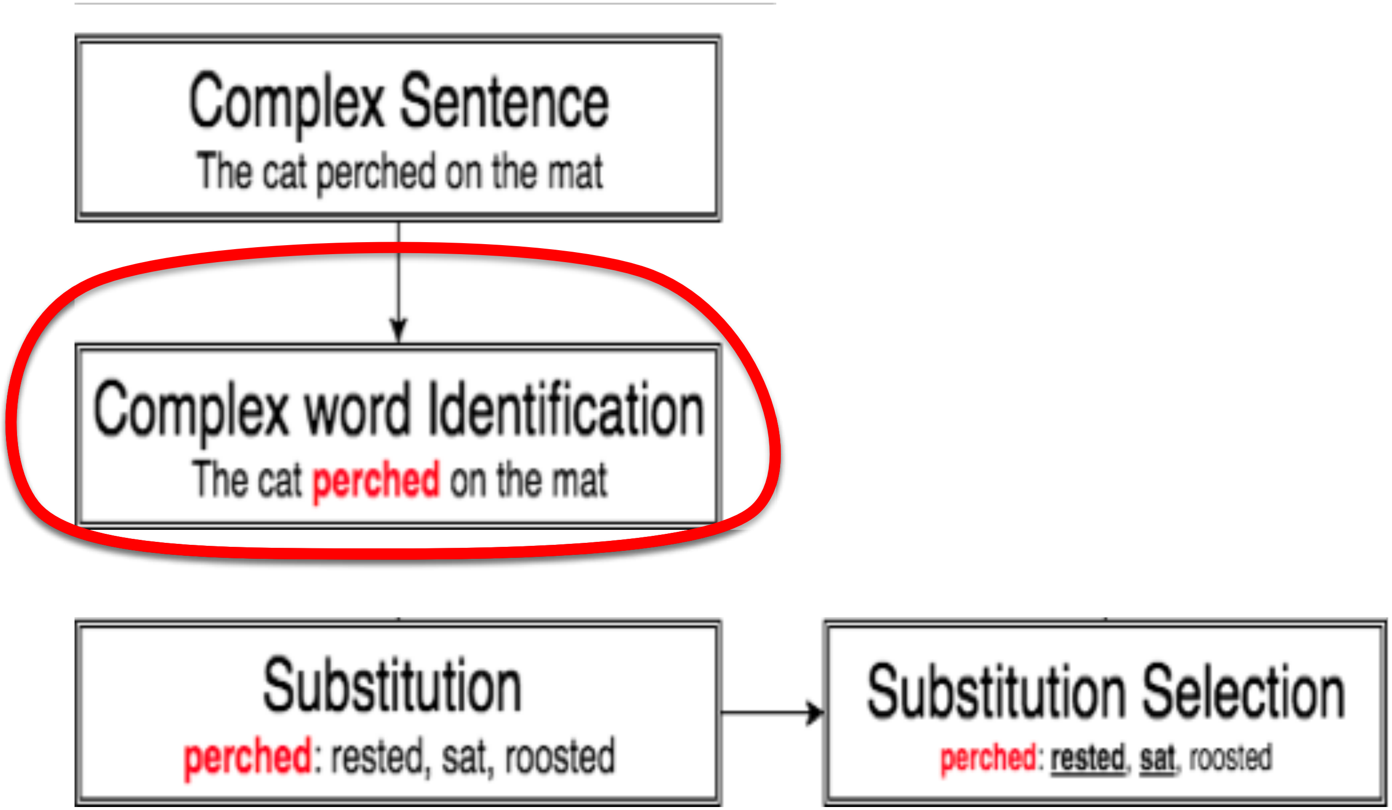
Components of LS (Paetzold, Gustavo (2015)


Complex Sentence
The cat perched on the mat
Complex word Identification
The cat perched on the mat

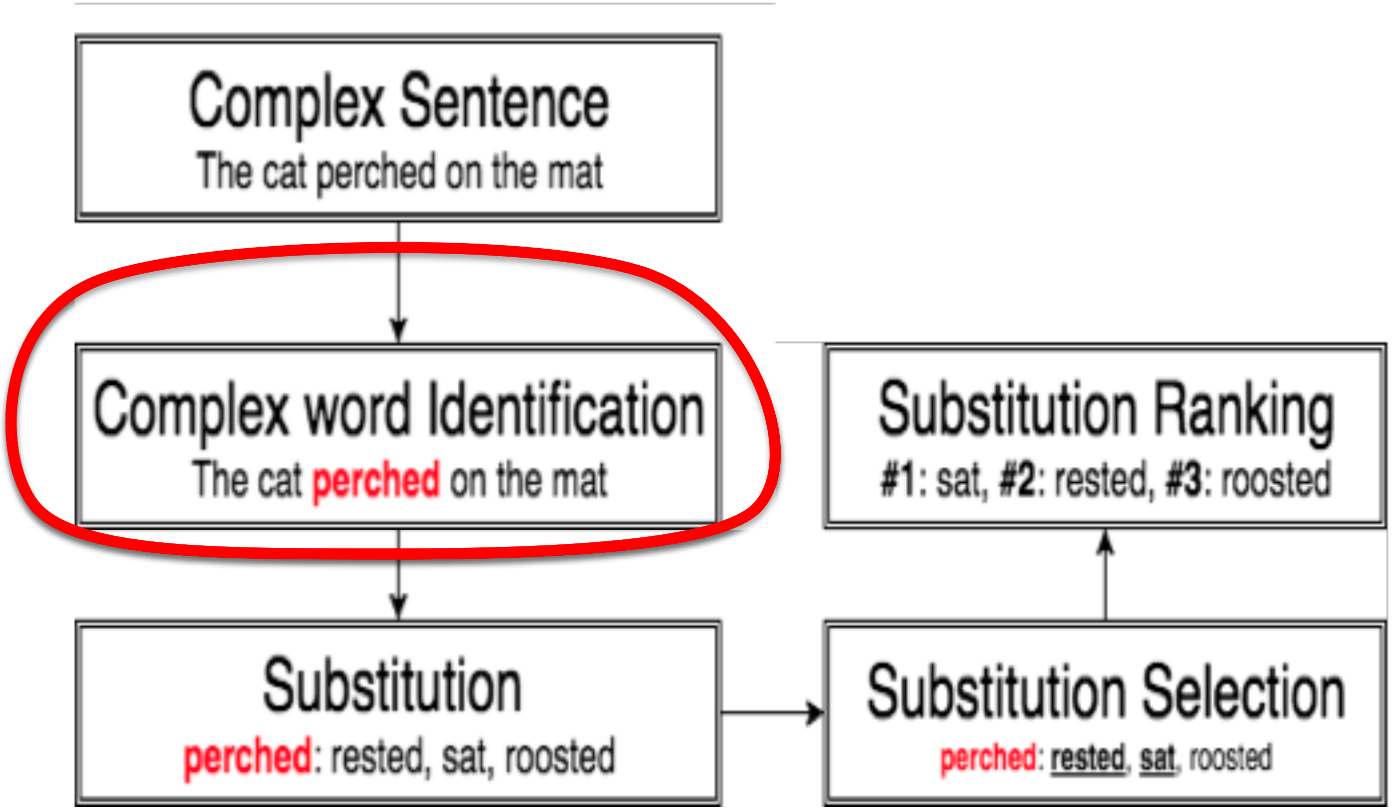
Components of LS (Paetzold, Gustavo (2015)


Complex Sentence
The cat perched on the mat
Complex word Identification
The cat perched on the mat
Substitution
perched: rested, sat, roosted

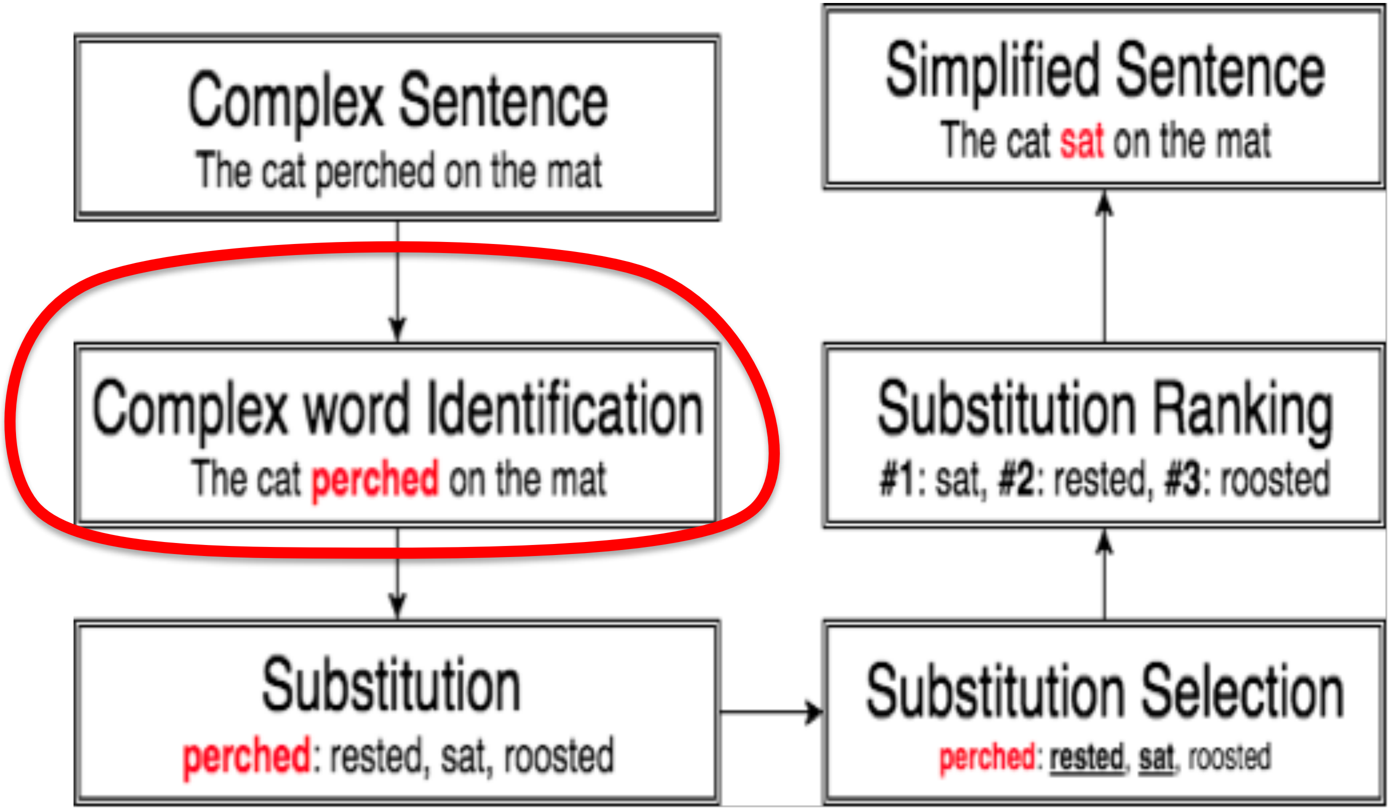
Components of LS (Paetzold, Gustavo (2015)


Complex Sentence
The cat perched on the mat
Complex word Identification
The cat perched on the mat
Substitution
perched: rested, sat, roosted
Substitution Selection
perched: rested, sat, roosted

Components of LS (Paetzold, Gustavo (2015)


Complex Sentence
The cat perched on the mat
Complex word Identification
The cat perched on the mat
Substitution
perched: rested, sat, roosted
Substitution Selection
perched: rested, sat, roosted
Substitution Ranking
#1: sat, #2: rested, #3: roosted

Components of LS (Paetzold, Gustavo (2015)


Complex Sentence
The cat perched on the mat
Complex word Identification
The cat perched on the mat
Substitution
perched: rested, sat, roosted
Substitution Selection
perched: rested, sat, roosted
Substitution Ranking
#1: sat, #2: rested, #3: roosted
Simplified Sentence
The cat sat on the mat

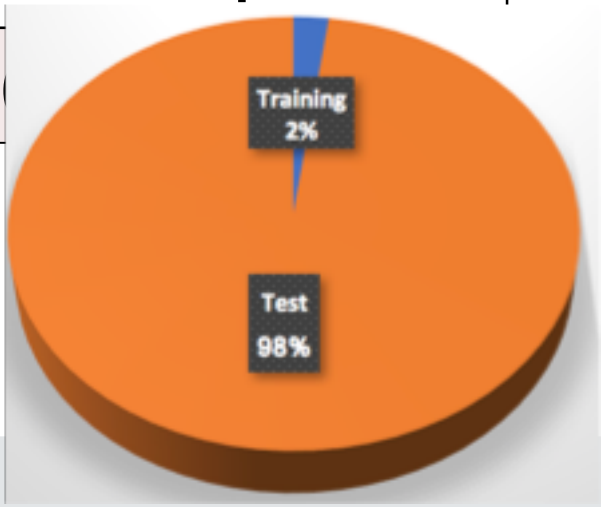
Objectives

- Collect CWI annotations (**CWIG3G2**)
  - For **three** genres
  - For **two** user groups
- Develop CWI systems
  - Across **genre**
  - Across **user**
  - Across **user-gerne**

Related works

- Previous CWI datasets relied on **Simple Wikipedia** and edit histories as a ‘**gold standard**’ annotation
- The SemEval2016 shared task dataset

	# Sentence	# Annotators
Training	200	20
Test	9,000	1
# Annotators = 40		



CWIG3G2 Dataset Collections

EMM NewsBrief

1200 sentences

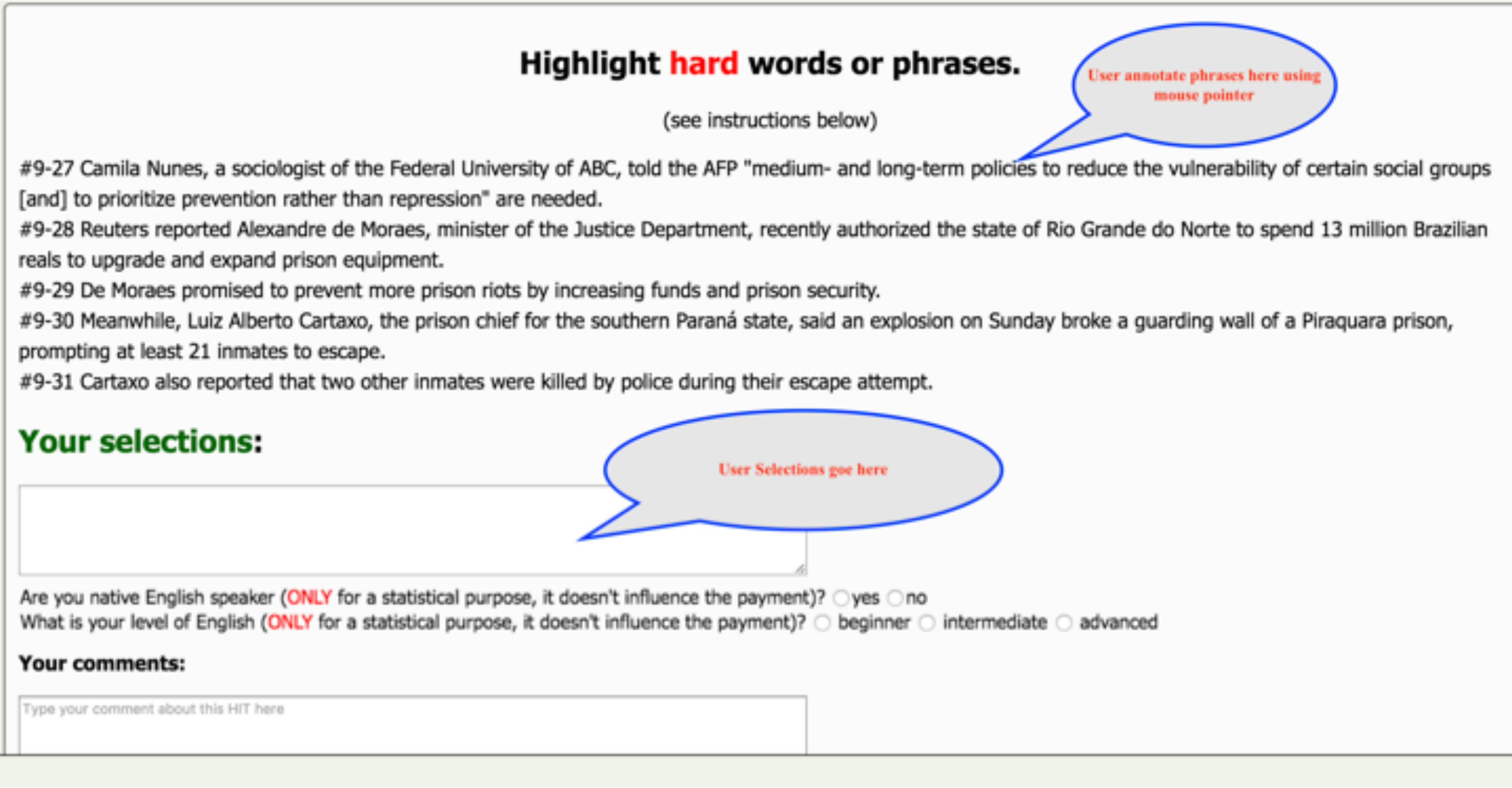
820 sentences

500 sentences

10 **native** Annotators

10 **non-native** Annotators

Mturk UI



Highlight hard words or phrases.

(see instructions below)

#9-27 Camila Nunes, a sociologist of the Federal University of ABC, told the AFP "medium- and long-term policies to reduce the vulnerability of certain social groups [and] to prioritize prevention rather than repression" are needed.
#9-28 Reuters reported Alexandre de Moraes, minister of the Justice Department, recently authorized the state of Rio Grande do Norte to spend 13 million Brazilian reals to upgrade and expand prison equipment.
#9-29 De Moraes promised to prevent more prison riots by increasing funds and prison security.
#9-30 Meanwhile, Luiz Alberto Cartaxo, the prison chief for the southern Paraná state, said an explosion on Sunday broke a guarding wall of a Piraquara prison, prompting at least 21 inmates to escape.
#9-31 Cartaxo also reported that two other inmates were killed by police during their escape attempt.

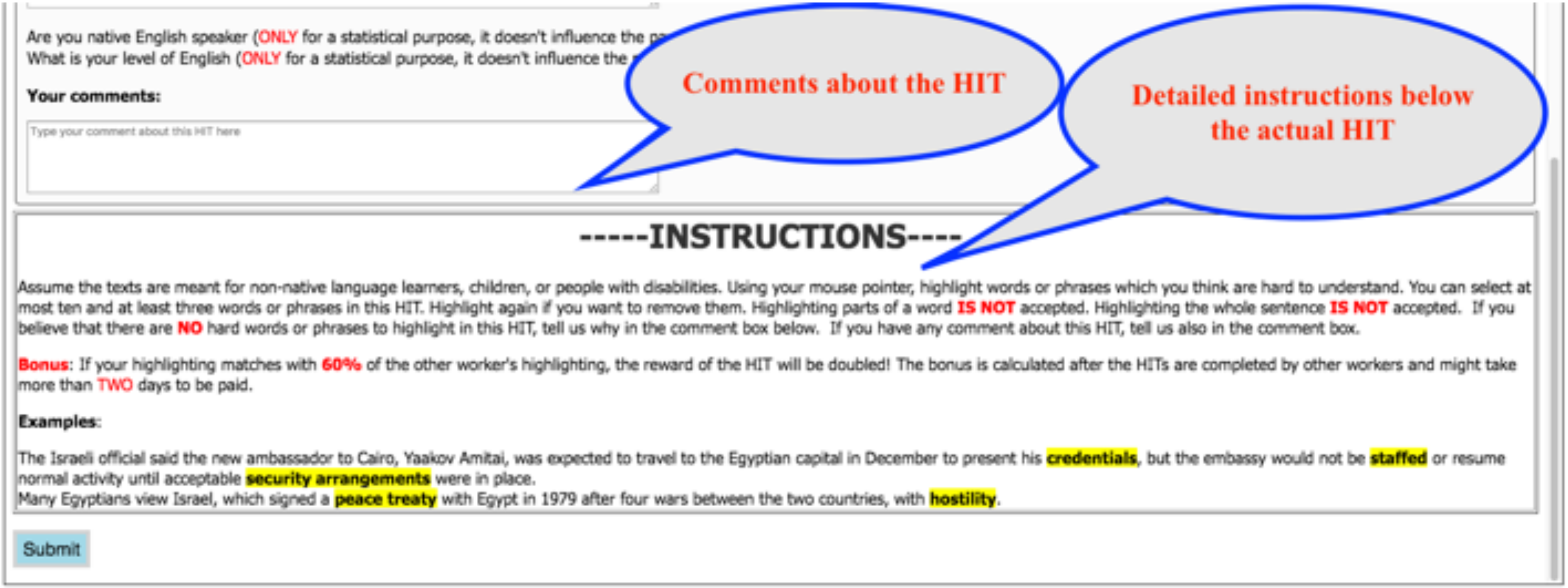
Your selections:

Are you native English speaker (ONLY for a statistical purpose, it doesn't influence the payment)? ○ yes ○ no
What is your level of English (ONLY for a statistical purpose, it doesn't influence the payment)? ○ beginner ○ intermediate ○ advanced

Your comments:

Type your comment about this HIT here

Mturk UI



Are you native English speaker (ONLY for a statistical purpose, it doesn't influence the pa

What is your level of English (ONLY for a statistical purpose, it doesn't influence the r

Your comments:

Type your comment about this HIT here

Comments about the HIT

Detailed instructions below the actual HIT

-----INSTRUCTIONS----

Assume the texts are meant for non-native language learners, children, or people with disabilities. Using your mouse pointer, highlight words or phrases which you think are hard to understand. You can select at most ten and at least three words or phrases in this HIT. Highlight again if you want to remove them. Highlighting parts of a word **IS NOT** accepted. Highlighting the whole sentence **IS NOT** accepted. If you believe that there are **NO** hard words or phrases to highlight in this HIT, tell us why in the comment box below. If you have any comment about this HIT, tell us also in the comment box.

Bonus: If your highlighting matches with **60%** of the other worker's highlighting, the reward of the HIT will be doubled! The bonus is calculated after the HITs are completed by other workers and might take more than TWO days to be paid.

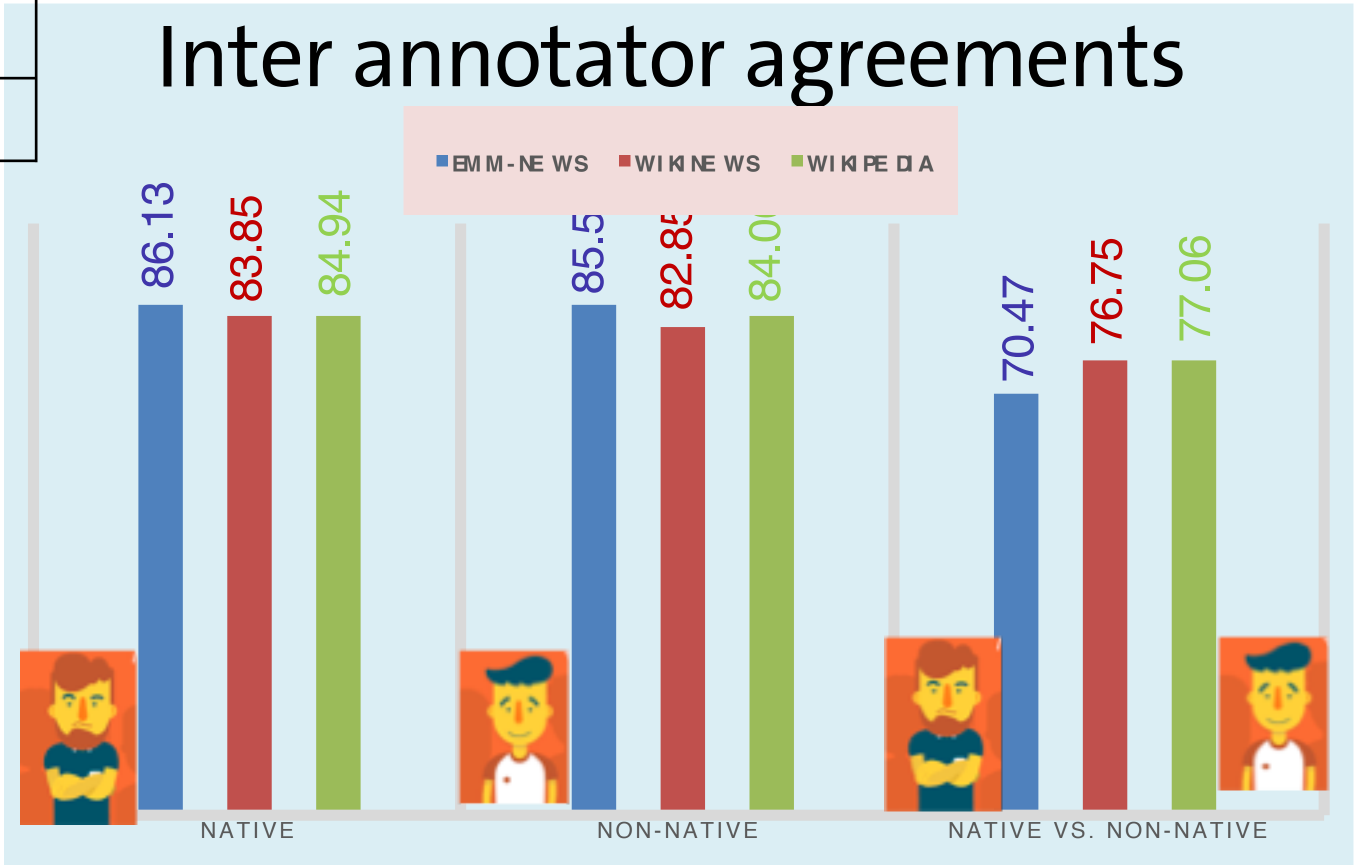
Examples:

The Israeli official said the new ambassador to Cairo, Yaakov Amitai, was expected to travel to the Egyptian capital in December to present his **credentials**, but the embassy would not be **staffed** or resume normal activity until acceptable **security arrangements** were in place.

Many Egyptians view Israel, which signed a **peace treaty** with Egypt in 1979 after four wars between the two countries, with **hostility**.

Analysis of collected results

	# Annotators
Native	134
Non-native	49



Inter annotator agreements

Classification Experiment

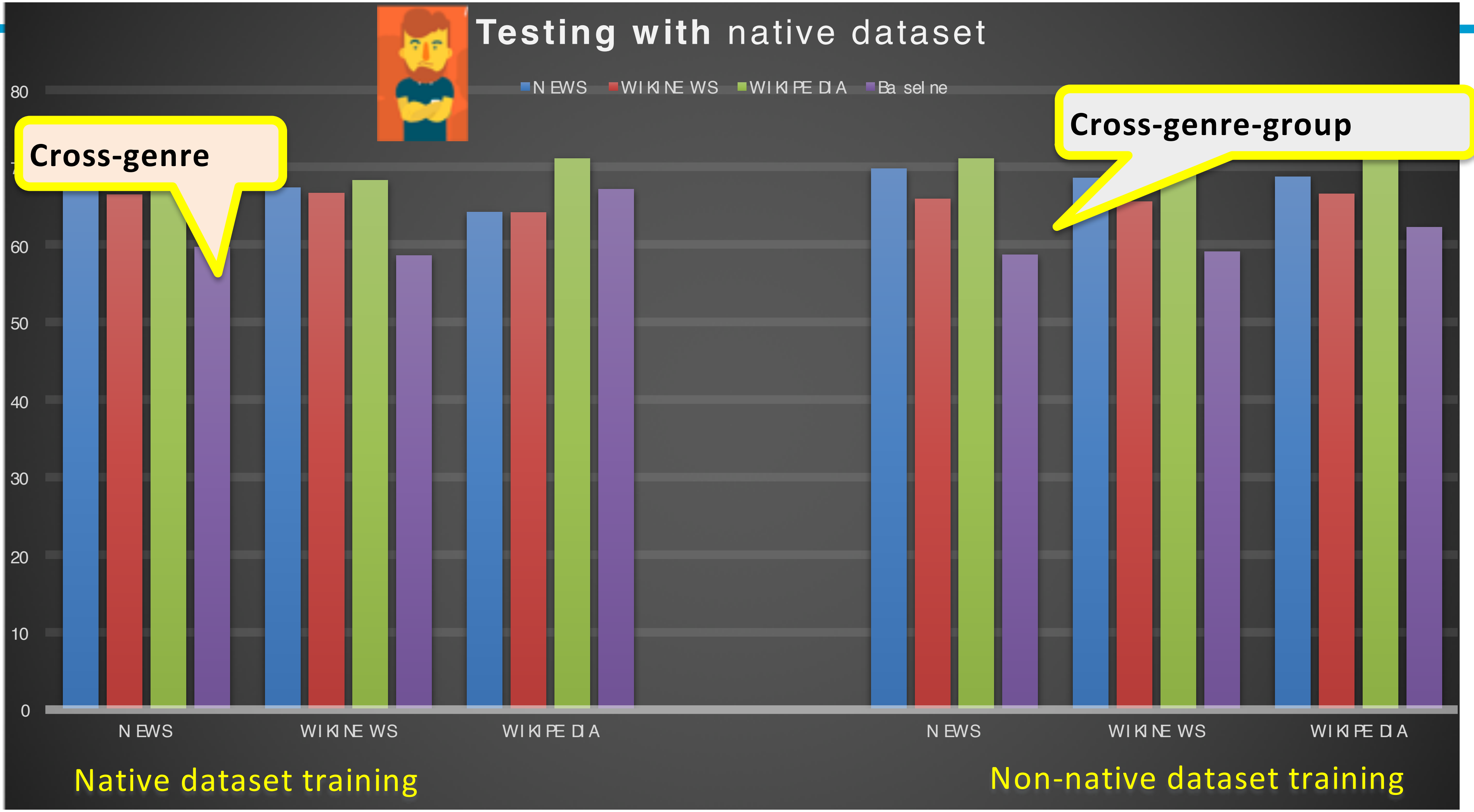
- Can we use datasets collected for one genre to predict CPs for another genre?
- Can we use datasets annotated by native speakers to predict CPs annotated by non-native speakers and vice versa?

Features

Feature groups	Descriptions
Length	number of vowels, number of syllables, number of characters
Frequency	Wikipedia, Google Web 1T 5-Grams, HIT/paragraph
Syntactic features	POS tags
Word Embedding	Word2vec using Wikipedia
Topic Features	LDA using Wikipedia

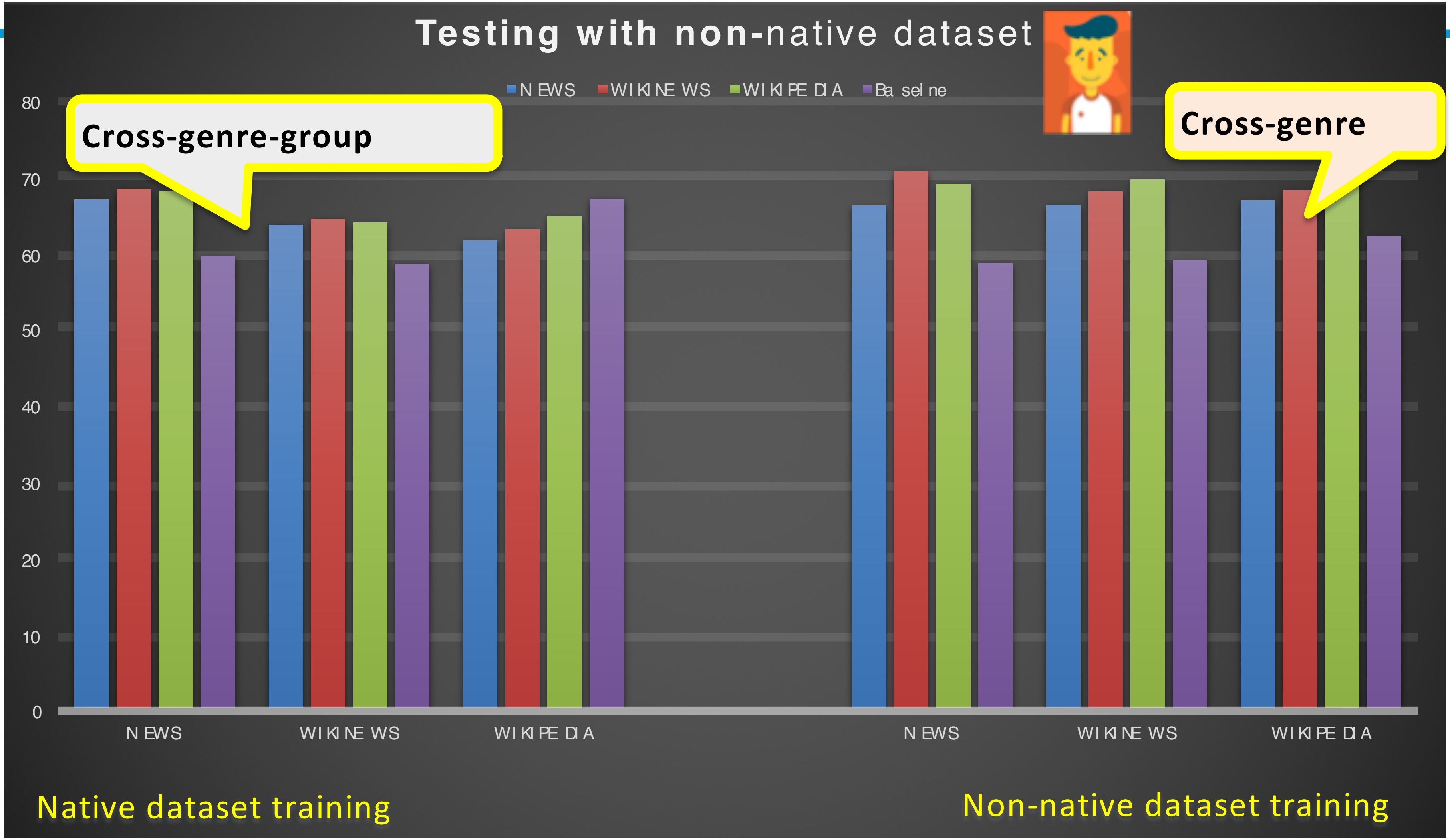
Results – F-score in %

Baseline → frequency thresholds based on **Simple English Wikipedia**



Results – F-score in %

Baseline → document **frequency thresholds** based on **Simple English Wikipedia**



Discussions

- **Within-group-genre**: better results on native datasets.
- **Cross-genre**: Slight differences in performance.
- **Cross-group**: training on non-native and testing on native gives better results.
- F-score is influenced by the IAA on the test set, and native annotators have higher IAA.

Conclusion

- **Six new datasets** for CWI tasks.
- Native speakers have higher inter-annotator agreement than the non-native speakers regardless of the text genre.
- Build different CWI systems.
- Within-genre CWI leads to better classification performances.
- CWI systems trained on native datasets can be used to predict CWs for non-native speakers and vice versa.
- We recommend to take language proficiency levels into account.

Datasets

https://www.inf.uni-hamburg.de/en/inst/ab/lt/resources/data/complex-word-identification-dataset.html

Announcement

BEA Workshop - NAACL 2018

Shared Task on Complex Word Identification

https://sites.google.com/view/cwisharedtask2018

Thank you!
Questions ?

Multilingual and Cross-Lingual Complex Word Identification

Yimam S.M, Štajner S., Riedl Martin, Biemann C. (2017): Multilingual and Cross-Lingual Complex Word Identification. In Proceedings of The 2017 International Conference on Recent Advances in Natural Language Processing (RANLP). Varna, Bulgaria

Experimental Setups

- **Six datasets** (three different genres times two different groups of annotators)
- The baseline is based on document **frequency thresholds** of Wikipedia corpora in the respective languages
- The **Nearest Centroid(NC)** ML algorithm from **scikit-learn** is used to build the CWI systems

Analysis of collected results

- English → 25623, German → 7403, Spanish → 14280
  Complex phrase annotations

	Native (%)		Non-native(%)	
Dataset	One	Mult.	One	Mult.
NewsBrief	25.36	74.64	38.42	61.58
WikiNews	23.62	76.38	59.07	40.93
Wikipedia	26.97	73.03	45.94	54.06
German	41.50	58.5	29.34	70.66
Spanish	28.16	71.84	95.16	~~4.84~~

Analysis of collected results

Distribution of collected CP (lengths in %)

dataset	uni-gram	bi-gram	tri-gram+
NewsBrief	83.50	12.50	3.99
WikiNews	86.00	10.02	3.98
Wikipedia	84.77	11.73	3.50
German	92.29	4.81	2.90
Spanish	77.03	13.83	9.14

Analysis of collected results[3]

- **German**:
    - fewer annotators (23 in total, 12 native and 11 non-native)
    - More **non-native** than **native** annotators per HIT (6.1 non-native and 3.9 native on average per HIT
    - In contrast to English and Spanish CP annotations, in the German task, more than 92% of the annotations are **single words**
    - Higher IAA among **non-native** German annotators (70.66%) than native German annotators (58.5%).

Analysis of collected results[4]

- **Spanish**
  - 54 annotators, 48 native speakers and 6 non-native speakers
  - Very low number of non-native speakers – excluded from our analysis and experiments
  - Lower IAA among Spanish native speakers than among English native speakers
  - Annotators highlighted mostly **multi-word expressions** (23% of the annotations)

Classification Experiments

Language independent features:

- **Length and frequency features**:
  - Length: the number of vowels, the number of syllables, and the number of characters in the word
  - Frequency: frequency of the word in Wikipedia, frequency of the word in the Google Web 1T 5-Grams, and frequency of the word in the HIT/paragraph
- **Syntactic features**: POS tags → tags transformed into universal POS tags
- **Word Embedding**: A single shared embedding space for more than fifty languages (from work of **Ammar et al. (2016)**)
- **Topic Features**: topic-relatedness feature that is extracted based on LDA model

Experimental Setups

- **Nine datasets** (three different genres times two different groups of annotators for English, native and non-native datasets for German and the native dataset for Spanish)

- ~~**Set I**: Monolingual experiments on nine datasets (for all three languages)~~.
- **Set II**: Cross-language experiments.
- The baseline is based on document **frequency thresholds** of Wikipedia corpora in the respective languages

Monolingual Results – F-score in %

Dataset	Native		Non-native	
	System (NC)	Baseline	System (NC)	Baseline
NewsBrief	69.97	66.01	62.35	60.28
WIKINEWS	69.25	66.56	57.89	51.5
WIKIPEDIA	70.79	67.2	58.31	53.53
GERMAN	54.92	51.37	58.5	56.57
SPANISH	45.83	44.04	–	–

- **Baseline** → is based on document **frequency thresholds** of Wikipedia corpora in the respective languages

Monolingual Results[2]

- All systems preform better than the baseline
- For **English**, CWI systems based on native speakers preform better than datasets from non-native speakers
- For **German**, CWI systems based on non-native speakers preform better than datasets from native speakers

Cross-Language Results – F-score in %

English genres as training

Training	German Native	German Non-native	Spanish Native
NewsBrief	53.89	58.32	45.19
Wikinews	54.54	58.42	44.48
Wikipedia	52.93	58.64	45.29

(a) Native English genres as training

NewsBrief	53.02	58.92	44.79
Wikinews	56.03	58.31	43.26
Wikipedia	51.53	59.14	44.39

(b) Non-Native English genres as training

Cross-Language Results – F-score in %

German and Spanish as training

	NewsBrief		WikiNews		Wikipedia	
Training	Native	Non-native	Native	Non-native	Native	Non-native
German-Native	67.42	57.55	66.79	57.08	62.14	51.22
German-Non-nat	66.99	58.51	64.17	55.53	63.78	54.09
Spanish	66.05	56.37	62.03	51.89	62.04	56.15

Cross-Language Results – F-score in %

Across German and Spanish datasets

Training	Spanish Native	German Native	German Non-native
German native	42.76		
German non-native	41.52		
Spanish native		53.53	56.82

Discussions

- CWI model trained on one of the English datasets
  - Similar or better result than on monolingual German and Spanish models
- CWI model trained on Spanish native
  - A slight decrease in performance than monolingual German models (still very close)
  - A drop in performance than monolingual English models
- CWI model trained on German dataset
  - A drop in performance than monolingual English models

Conclusion

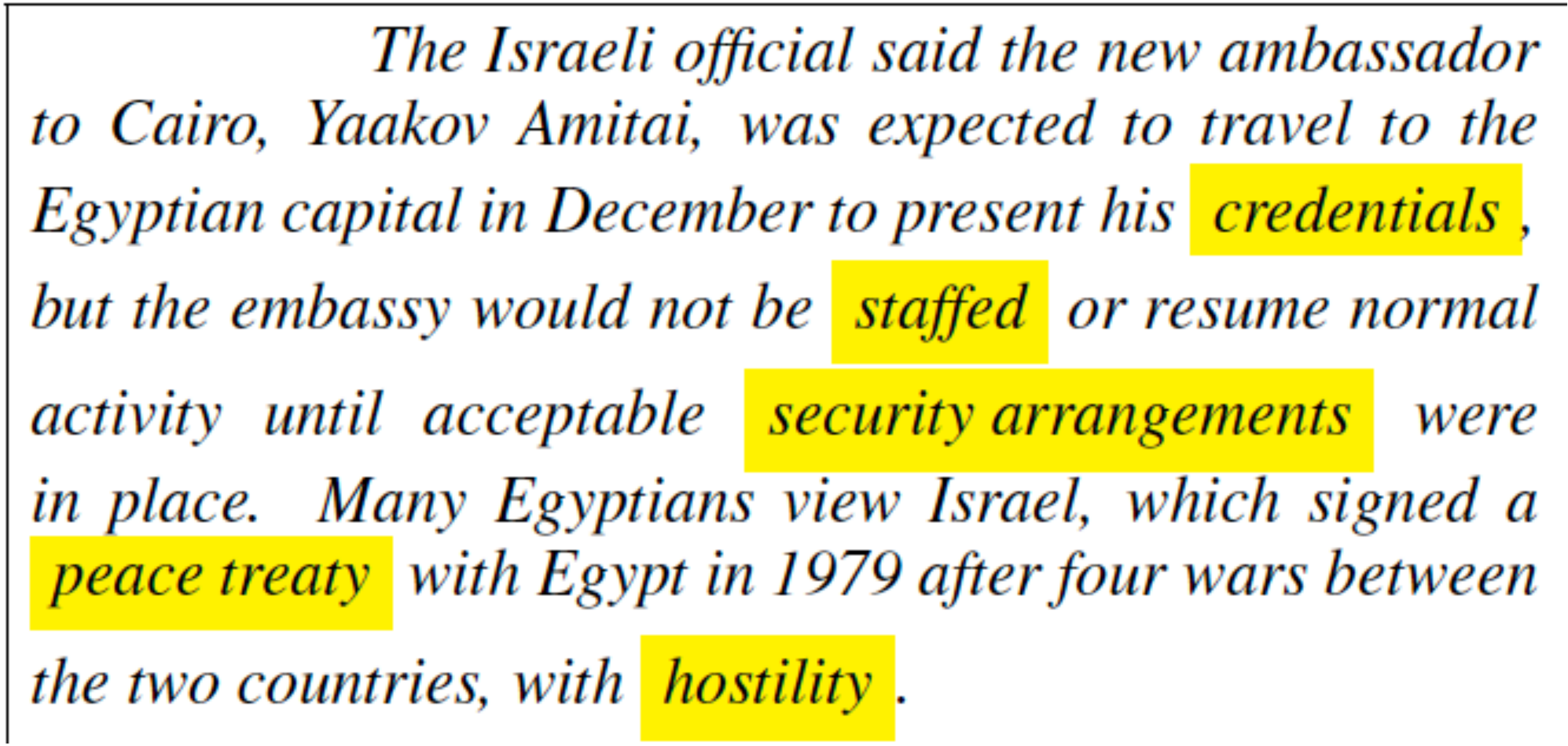
- CWI is important task in text accessibility and text simplification.
- Collected a total of **nine** ‘**gold standard**’ CWI datasets
- Developed a state-of-the- art automated **CWI system** with **language independent** feature representations
- Demonstrate that **cross-lingual CWI systems** work very well
- In the future, **balance** the number of annotators per HIT for native and non-native annotations

- MORE!!!

Annotation procedure

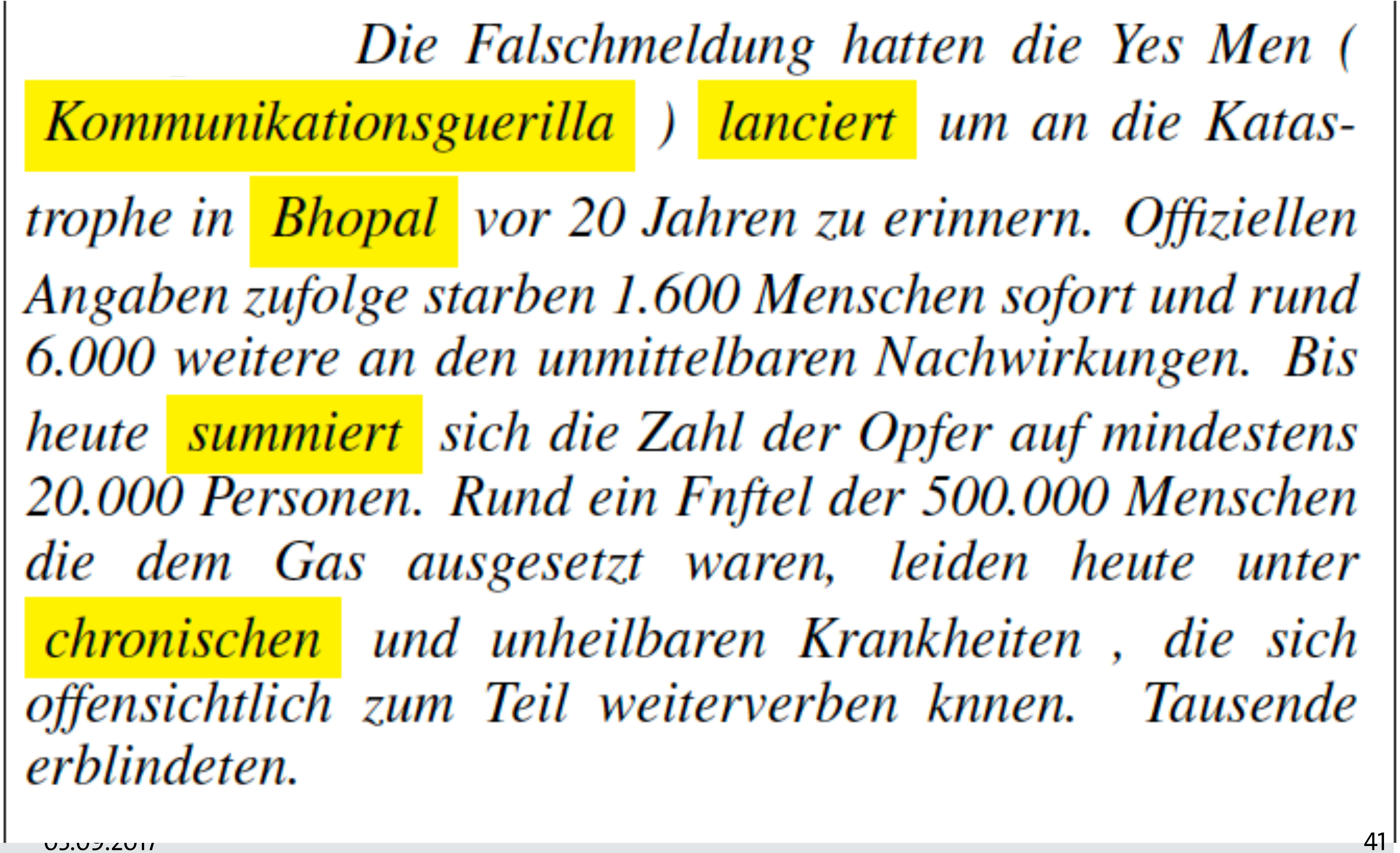
- A HIT (human intelligence task) is:
  - 5—10 sentences
  - Completed by 10 native and 10 non-native speakers each
  - Workers can highlight **single** words or **sequences** of words
  - Maximum of **10 annotations** per HIT
- Annotation selection
  - No simple words
  - No more than 50 characters
- Question about nativeness

CP annotation Examples - English



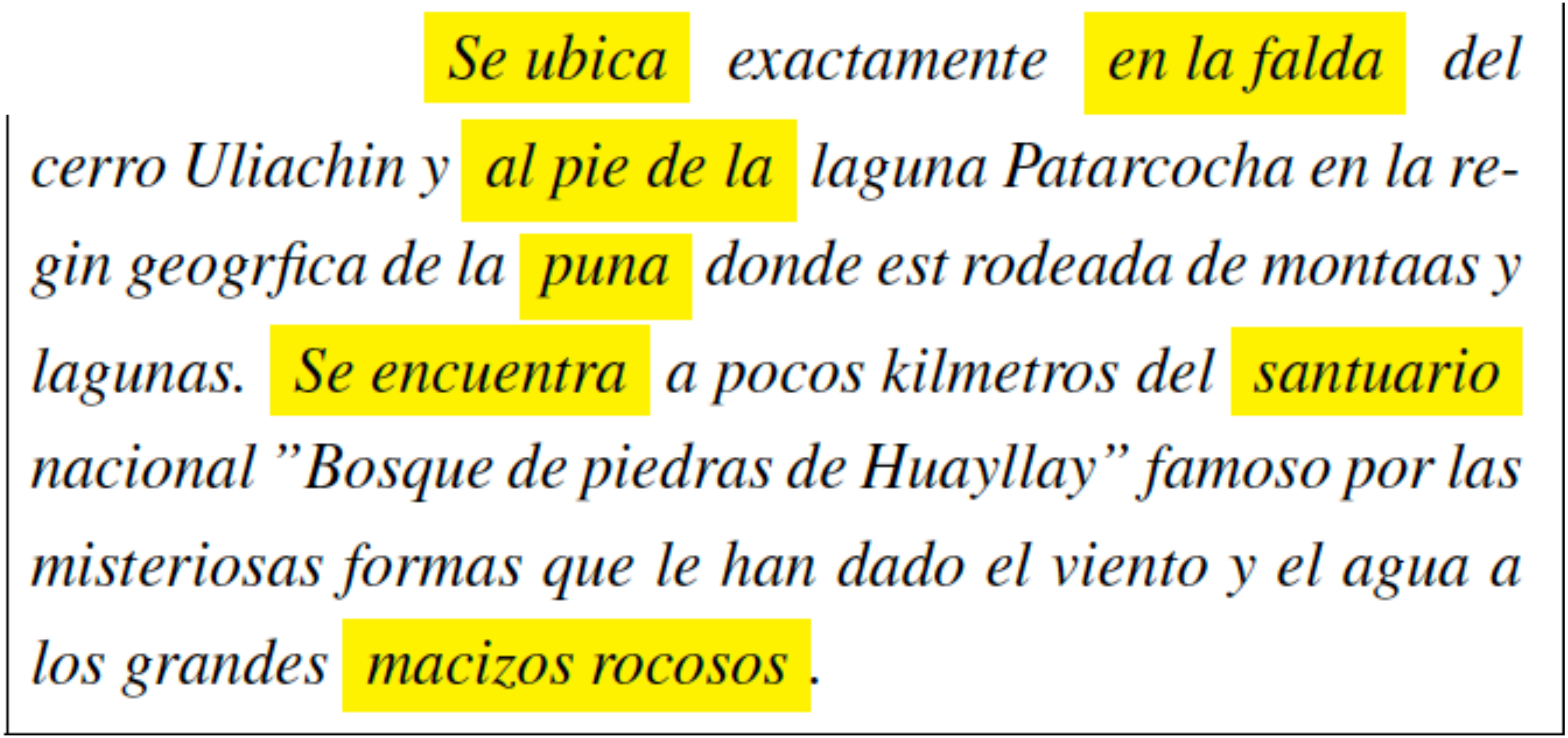
The Israeli official said the new ambassador to Cairo, Yaakov Amitai, was expected to travel to the Egyptian capital in December to present his credentials , but the embassy would not be staffed or resume normal activity until acceptable security arrangements were in place. Many Egyptians view Israel, which signed a peace treaty with Egypt in 1979 after four wars between the two countries, with hostility .

- **CP annotation Examples - German**



Die Falschmeldung hatten die Yes Men (Kommunikationsguerilla) lanciert um an die Katastrophe in Bhopal vor 20 Jahren zu erinnern. Offiziellen Angaben zufolge starben 1.600 Menschen sofort und rund 6.000 weitere an den unmittelbaren Nachwirkungen. Bis heute summiert sich die Zahl der Opfer auf mindestens 20.000 Personen. Rund ein Fnftel der 500.000 Menschen die dem Gas ausgesetzt waren, leiden heute unter chronischen und unheilbaren Krankheiten , die sich offensichtlich zum Teil weiterverben knnen. Tausende erblindeten.

CP annotation Examples - Spanish



Se ubica exactamente en la falda del cerro Uliachin y al pie de la laguna Patarcocha en la regin geogrfica de la puna donde est rodeada de montaas y lagunas. Se encuentra a pocos kilmetros del santuario nacional ”Bosque de piedras de Huayllay” famoso por las misteriosas formas que le han dado el viento y el agua a los grandes macizos rocosos .

- **On the shared task dataset**: almost the same F-score (35.44%) as the best F-scored system (35.30%), but better G-score (75.51%) than the same system (60.80%)