Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MULTILINGUAL AND CROSS-LINGUAL COMPLEX WORD IDENTIFICATION

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

SEPTEMBER 5, 2017

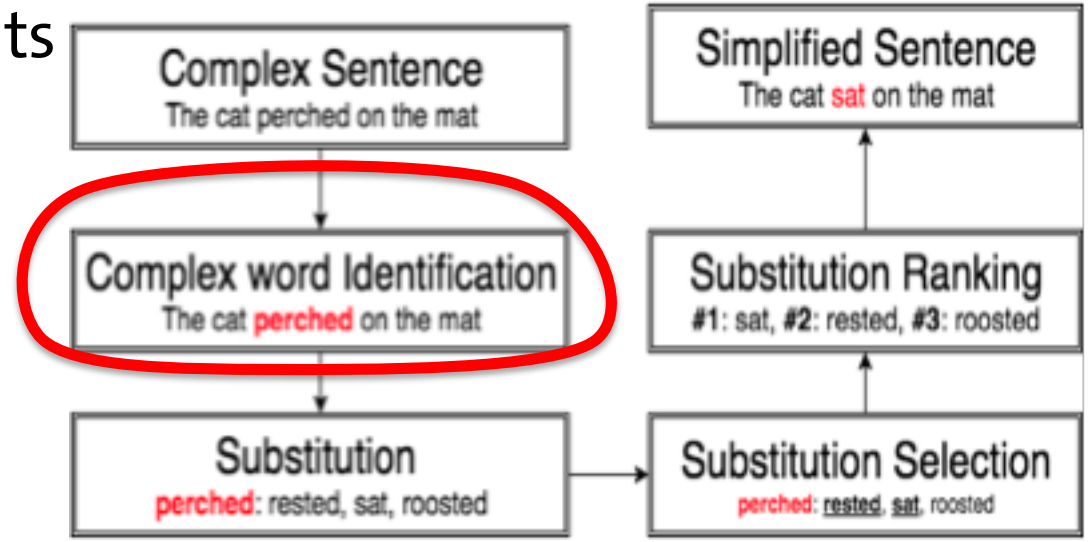
Outline

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Introduction to complex word identification (CWI)
- State of the art
  - CWI datasets
  - CWI Systems
- Approaches
  - New dataset collections
  - Annotation procedure
  - Analysis of collected results
- Classification experiments
- Conclusion

Introduction

- Complex Word Identification (CWI) is important in lexical simplification
- Lexical simplification (LS):
  - Replace infrequent and difficult phrases
  - Target readers: language learners, children, people with reading impairments
- Components of LS:



Introduction[2]

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Gold standard CWI should be compiled using human annotation
- For English - **Semeval-2016 shared task dataset**
  - annotated by non-native English speakers
- No CWI datasets for other languages such as German and Spanish

Introduction[3]

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Our objectives
  - Collect CWI annotations for **English, German** and **Spanish**
  - Propose **language-independent** set of ML features
  - Develop and experiment with **cross-language CWI ML models**

CWI Datasets

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Previous CWI datasets relied on **Simple Wikipedia** and edit histories as a '**gold standard**' annotation (Shardlow (2013), Horn et al., (2014), Kauchak, (2013).)
- The SemEval2016 shared task dataset
  - 9,200 sentences (200 training and 9000 test)
    - Training – annotated by 20 people
    - Test – annotated only by 1 person
  - From LexMTurk corpus and Simple Wikipedia
  - Human annotation (non-native speakers)
  - Content words are annotated (VERB, NOUN, ADJ, ADV)

CWI Systems

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- The systems of the SemEval-2016 shared task
  - Best system by F-score scores 35.30%
- The problem of those best performing systems:
  - The **lexicons** used and **Simple Wikipedia** do not exist for other languages
  - Features can not be obtained for other languages

New dataset collections

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Collect annotations of complex words and phrases (multi-word expressions)
- Use Amazon mechanical turk (**Mturk**) crowdsourcing platform
- For English, German and Spanish
- Using **native** and **non-native** speakers

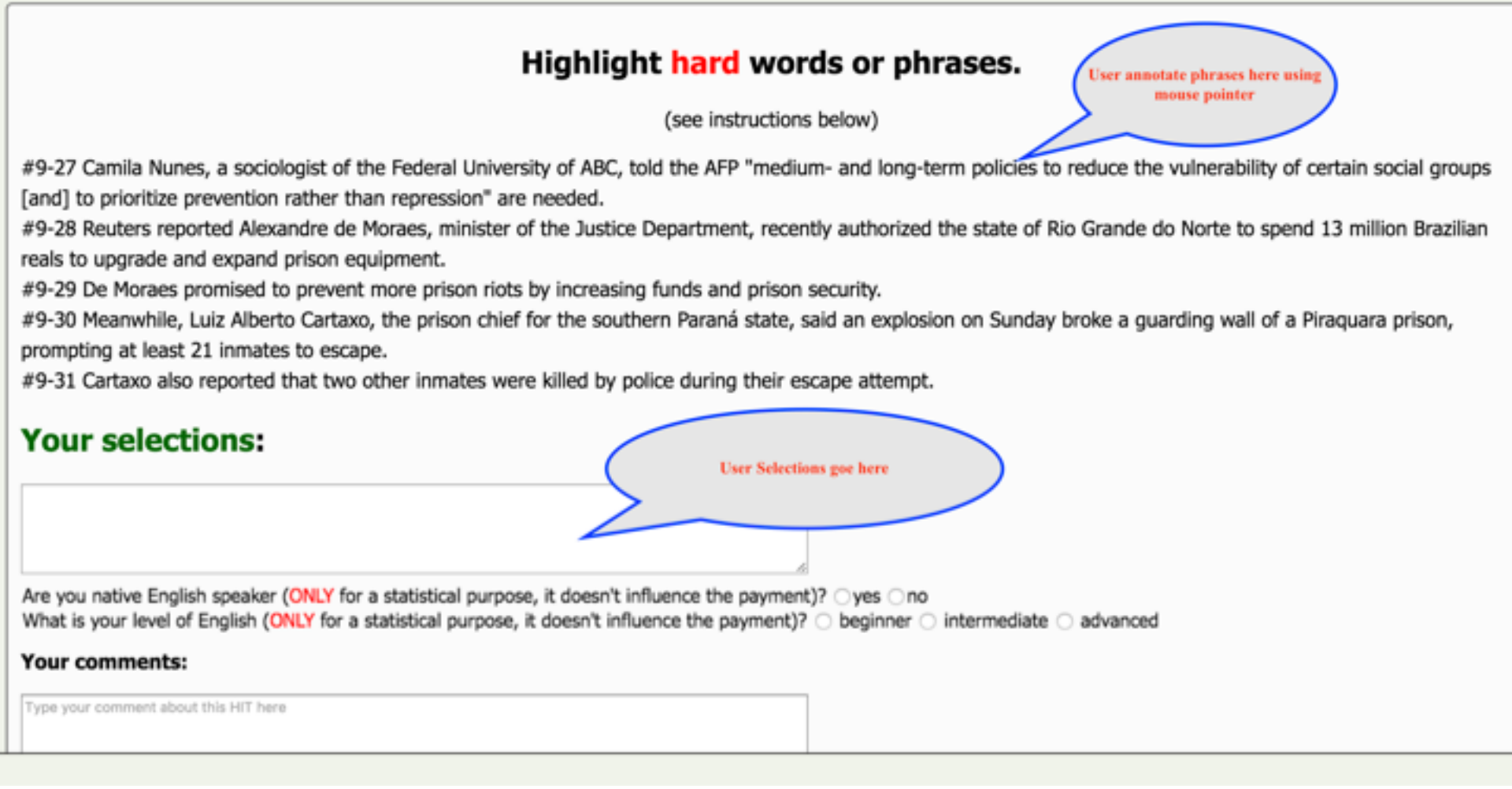
Data Selection

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- **English**:
  - professionally written news (100 news articles from **EMM NewsBrief**)
  - **Wikinews** (42 articles)
  - **Wikipedia** articles (500 sentences)
- **German**: 978 sentences from German Wikipedia articles
- **Spanish**: 1,387 sentences from Spanish Wikipedia articles

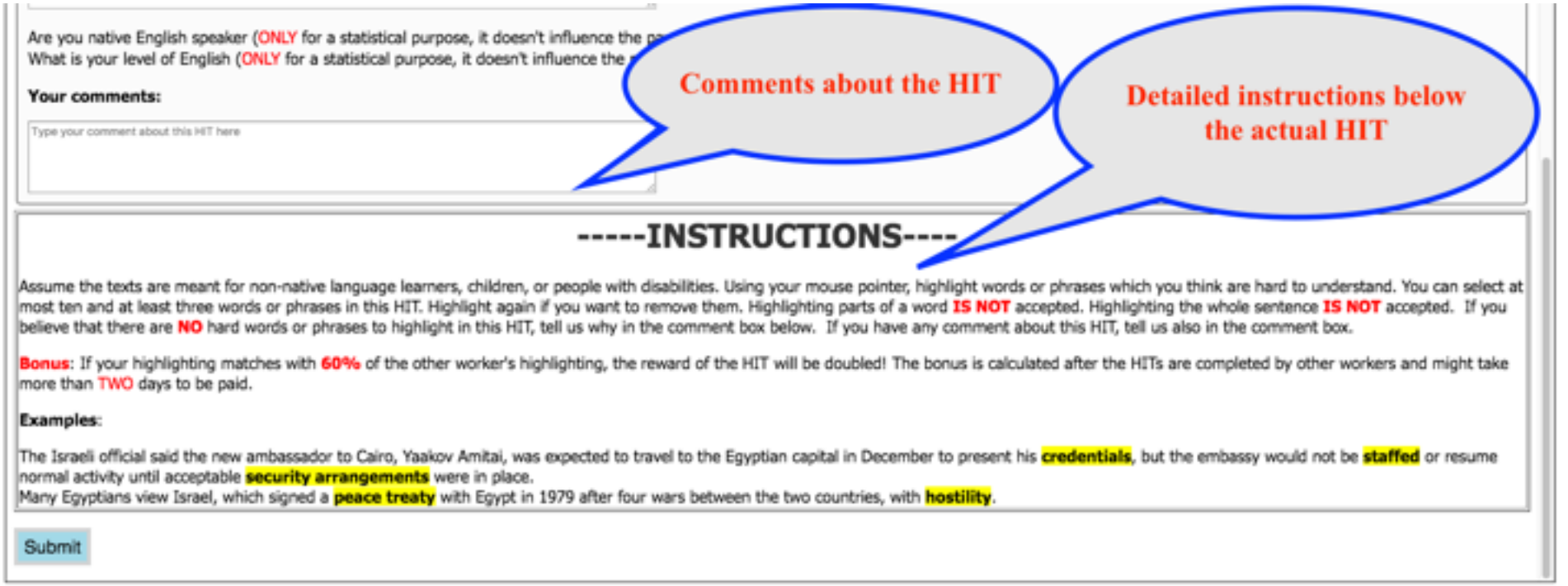
Mturk UI

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Mturk UI

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



Mult. = Annotations selected by two or more annotators

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

Dataset	Native (%)		Non-native(%)	
	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 %)

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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[2]

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- **English**:
  - 87 native and 25 non-native annotators
  - The percentage of multiply-selected CPs by **native** speakers stays stable across genres
  - the percentage of multiply selected CPs by **non-native** speakers is always significantly lower (54%–62%) than the percentage of multiply selected CPs by native speakers (73%– 75%), regardless of the text genre

Analysis of collected results[3]

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- **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]

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- **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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- **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
- The **Nearest Centroid(NC)** ML algorithm from **scikit-learn** is used to build the CWI systems

Monolingual Results – F-score in %

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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]

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- 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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

	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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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

Discussions

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- 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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- 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

Thank you
Questions ?

CP annotation Examples - English

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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**

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

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

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Se ubica exactamente en la falda del cerro Uliachin y al pie de la laguna Patarcocha en la re-gin 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 .