

# Universal Proposition Bank 2.0



Ishan Jindal, Alexandre Rademaker, Michał Ulewicz,  
Huyen Nguyen, Linh Ha, Khoi-Nguyen Tran, Huaiyu Zhu,  
Yunyao Li

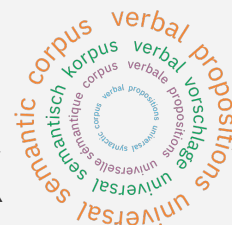


Scalable Knowledge Intelligence  
IBM Research – Almaden  
IBM Research – Brazil



Vietnam National University

## Universal PropBank



<https://universalpropositions.github.io/>

LREC 2022  
Marseille

# Outline

## **Introduction**

## **Multilingual Propbanks**

- Problems
- Challenges
- Solution

## **Universal Representations**

- UP1.0
  - Problems
- Revamp

## **Results**

- Quality
- Analysis

## **Known Issues**

# Semantic Role Labeling (SRL)

Who did what to whom, when, where and how?

(Gildea and Jurafsky, 2000; Màrquez et al., 2008)

# Semantic Role Labeling (SRL)

Derik

broke

the window

with a hammer

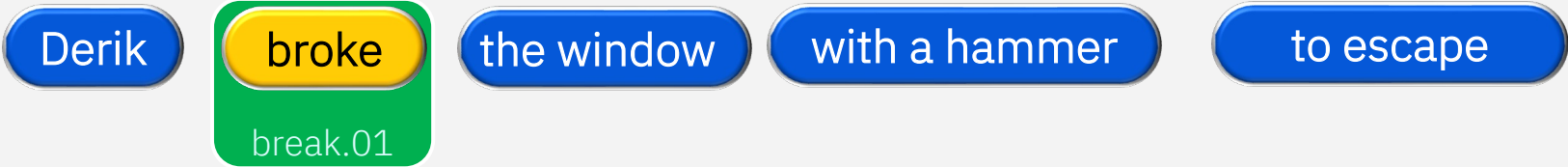
to escape



Predicate Identification

Identify all predicates in the sentence

# Semantic Role Labeling (SRL)



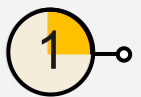


- 1 Predicate Identification Identify all predicates in the sentence
- 2 Sense Disambiguation Classify sense of each predicate

break.01, break  
A0: breaker  
A1: thing broken  
A2: instrument  
A3: pieces  
A4: arg1 broken  
away from what?

[English Propbank](#)

# Semantic Role Labeling (SRL)

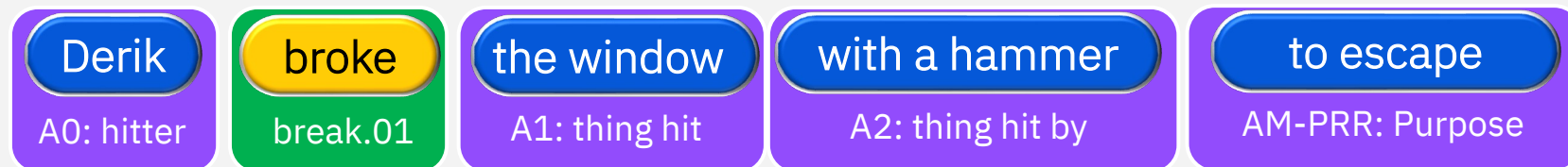


-  **Predicate Identification** Identify all predicates in the sentence
-  **Sense Disambiguation** Classify sense of each predicate
-  **Argument Identification** Find all roles of each predicate

Argument identification can either be

- Identification of span, (span SRL) OR
- Identification of head (dependency SRL)

# Semantic Role Labeling (SRL)

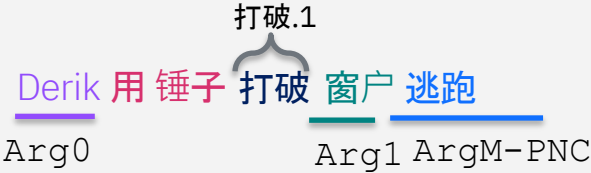


- 1** Predicate Identification Identify all predicates in the sentence
- 2** Sense Disambiguation Classify sense of each predicate
- 3** Argument Identification Find all roles of each predicate
- 4** Argument Classification Assign semantic label to each role

# Multilingual SRL

**Fact:**

- 7,102 known languages
- 23 most spoken languages



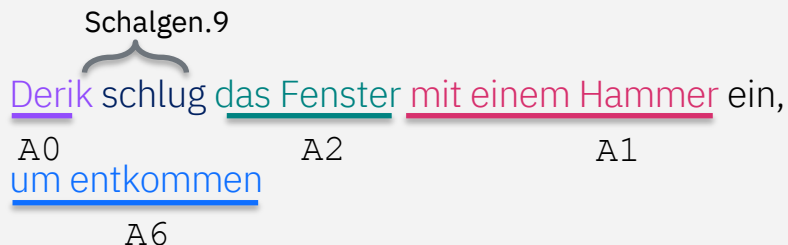
Same color represents same semantic meaning



# Multilingual SRL

## Problem:

- Different meaning Representations
  - for the same language
    - Propbank
    - FrameNet
  - for different languages
    - English Propbank
    - Chinese Propbank
    - German Propbank
    - Hindi Propbank
    - French? Spanish? Arabic?.....
- Applications must be language specific
- Separate text analytics



Same color represents same semantic meaning

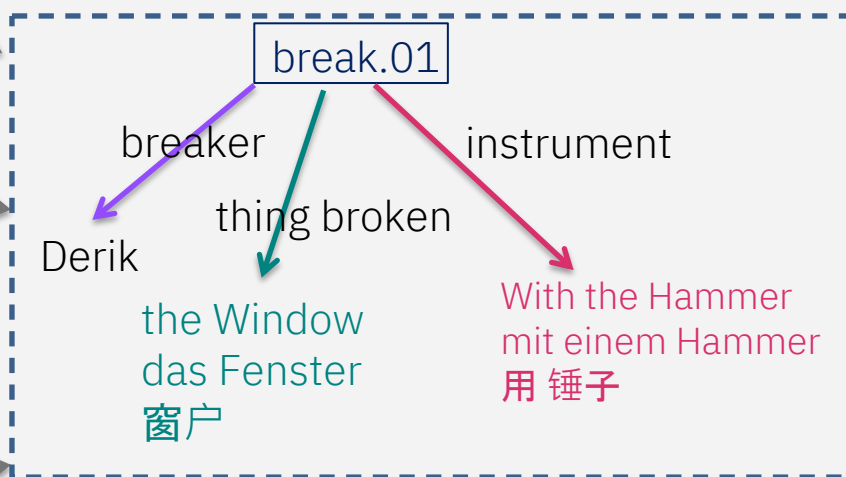
# Potential Solution

break.01  
Derik broke the window with a hammer to escape  
 A0 A1 A2 AM-PRR

Schlagen.9 break.01  
Derik schlug das Fenster mit einem Hammer ein,  
 A0 A2 A1 A1 A2  
um entkommen  
 A6 AM-PRR

打破.1 break.01  
Derik 用 锤子 打破 窗户 逃跑  
 Arg0 Arg1 ArgM-PNC  
 A0 A2 A1 AM-PRR

## Universal Meaning Representation



# Multilingual SRL

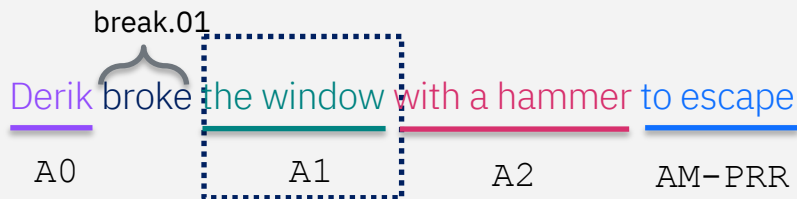
Challenge:

Language-specific formalisms

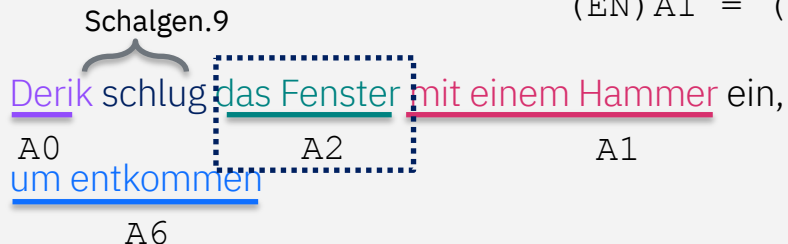
<u>break.01</u>		<u>Schalgen.9</u>
A0: breaker	→	A0: Agent
A1: thing broken	✗	A1: Impactor
A2: instrument	✗	A2: Impactee
A3: pieces	✗	A3: Impactors
A4: arg1 broken away	✗	A4: Result
from what?		A5: Subregion
AM-MNR: Manner	→	A6: Purpose
AM-PRR: Purpose		A7: Period of iteration

Limited Coverage

<u>break.01</u>		<u>打破.1</u>
A0: breaker	→	Arg0: Agent, causer
A1: thing broken	→	Arg1: Theme
A2: instrument	✗	ArgM-PNC: Purpose
A3: pieces	✗	
A4: arg1 broken away	✗	
from what?		
AM-MNR: Manner	→	
AM-PRR: Purpose		



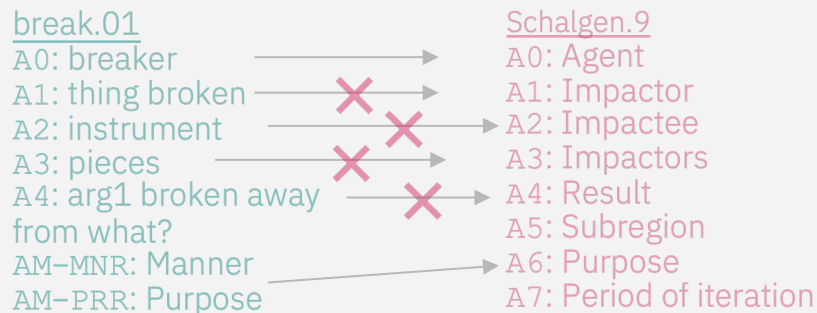
(EN) A1 = (DE) A2



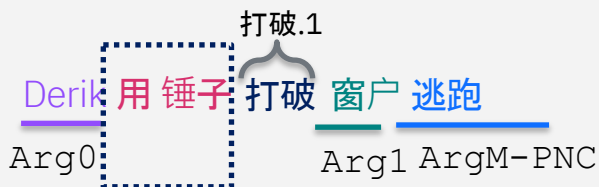
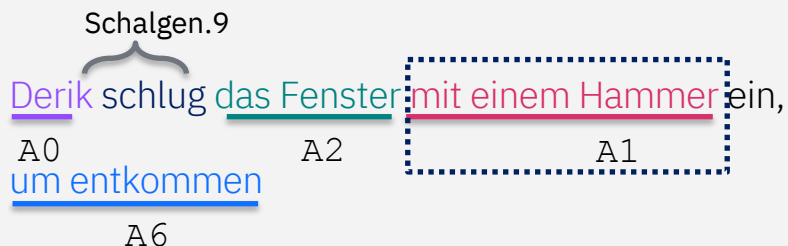
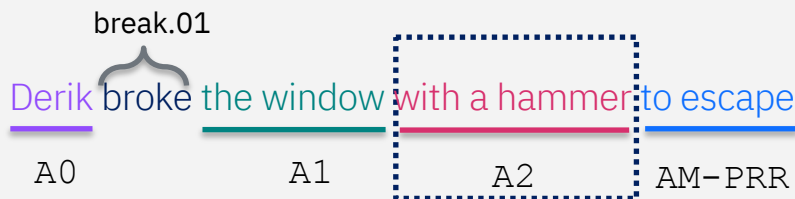
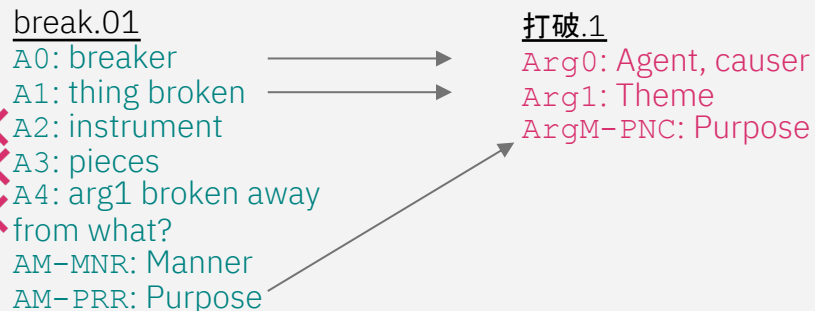
# Multilingual SRL

Challenge:

Language-specific formalisms



Limited Coverage



# Universal Representation via Annotation Projection

## UP1.0

- Project English Propbank SRL labels to target languages
- [ACL 2015] [Generating high quality proposition banks for multilingual semantic role labeling](#)
- <https://github.com/System-T/UniversalPropositions>
  - 7 languages: DE, ES, FI, FR, IT, PT, ZH
  - Released in 2016 and 2017.

A1	be.01			A2	AM-TMP	
Muiriel	is	20	years	old	now	.

Bây giờ	Muiriell	được	20	tuổi	.
now	Muiriell	is	20	years old	
AM-TMP	A1	be.01		A2	

# Universal PropBank via Annotation Projection

## Issues with UP1.0

- **Quality:** Released in 2016 and 2017.
  - SRL model
  - Parser model
  - Word aligner model
- **Language coverage:** Only 7 languages from 2 language families.
- **SRL representations:** Only dependency based SRL
- **Gold labels:** no language with gold SRL labels

Better models became available

2016 → 2022

- Feature based SRL → NNSRL
  - ~ 7F1 points better is performance
- Statistical Parser → Neural Parser
  - >10F1 points better
- Statistical Aligner → Neural Aligner
  - >10F1 points better

# Universal PropBank via Annotation Projection

## Issues with UP1.0

- **Quality:** Released in 2016 and 2017.
  - SRL model
  - Parser model
  - Word aligner model
- **Language coverage:** Only 7 languages from 2 language families.
- **SRL representations:** Only dependency based SRL
- **Gold labels:** no language with gold SRL labels

Deep learning models cover more languages

- Multilingual Language Models
  - Trained on 100+ languages.
- 7,102 known languages
  - 23 most spoken languages

# Universal PropBank via Annotation Projection

## Issues with UP1.0

- **Quality:** Released in 2016 and 2017.
  - SRL model
  - Parser model
  - Word aligner model
- **Language coverage:** Only 7 languages from 2 language families.
- **SRL representations:** Only dependency based SRL
- **Gold labels:** no language with gold SRL labels

## Need for SPAN based SRL

- Head annotation is insufficient to compute the span annotation.
  - Needs high quality syntactic parser
  - If exists, Not always correct
- Decoupling of syntactic analysis
  - Train syntax-agnostic SRL models



# Universal PropBank via Annotation Projection

## Issues with UP1.0

- **Quality:** Released in 2016 and 2017.
  - SRL model
  - Parser model
  - Word aligner model
- **Language coverage:** Only 7 languages from 2 language families.
- **SRL representations:** Only dependency based SRL
- **Gold labels:** no language with gold SRL labels

➤ Enabling the research community to perform fair evaluation of their multilingual and cross-lingual SRL systems

# Universal PropBank Revamp

## UP2.0

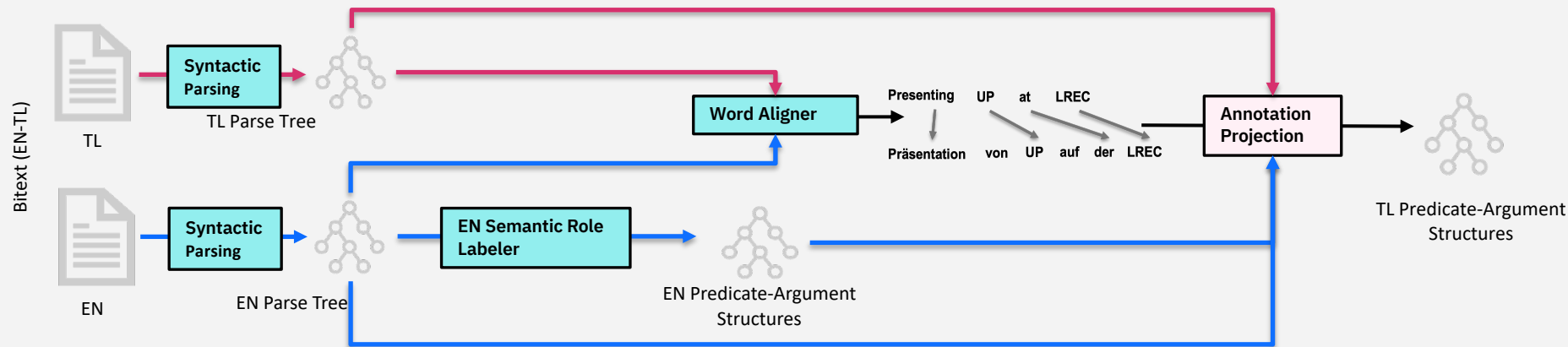
Significant update to UP-1.0

- **[High Quality]** ~10 F1 points quality improvement over UP-1.0
- **[Language Expansion]** 7 → 23 languages from 8 language families:
  - CS, DE, EL, EN, ES, FI, FR, HI, HU, ID, IT, JA, KO, MR, NL, PL, PT, RO, RU, TA, TE, UK, VI, ZH
- **[Span Annotations]** Head only → both span- and head-based SRL
- **[Gold Data]** Gold data in EN, PT, PL, VI languages

A1	be.01			A2	AM-TMP	
Muiriel	is	20	years	old	now	.

Bây giờ	Muiriel	được	20	tuổi	
now	Muiriel	is	20	years	old
AM-TMP	A1	be.01			A2
AM-TMP	A1	be.01	A2		

# Automatic Data Generation



## Bitext (EN-TL)

- Europarl
- Tatoeba
- Opensubtitles
- UN

## Syntactic Parser

- MATE
- STANFORD
- Stanza
- Spacy
- UDPipe

## Word Aligner

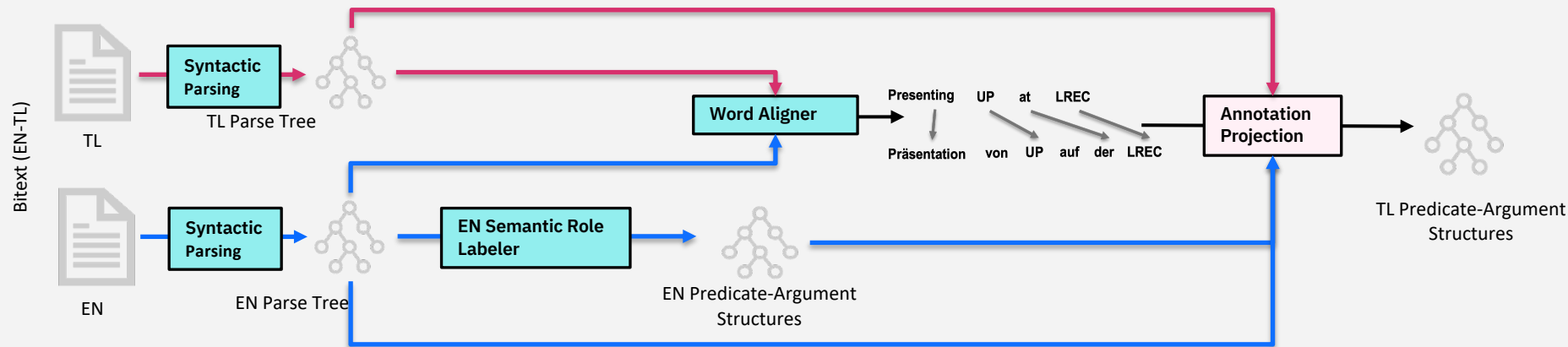
- Berkeley Aligner
- Simalign
- Awesome Align

## EN SRL Model

- ClearNLP
- MATE-SRL
- Neural SRL
- Span- and dependency-based SRL

# Automatic Data Generation

How we revamp?



## R1) Component Selection

1	UP1.0
2	UP2.0

### Bitext (EN-TL)

- Europarl **2**
- Tatoeba **2**
- Opensubtitles **2** **1**

### Syntactic Parser

- MATE **1**
- STANFORD
- Stanza **2**
- Spacy
- UDPipe

### Word Aligner

- Berkeley Aligner **1**
- Simalign **2**
- Awesome Align

### EN SRL Model

- ClearNLP **1**
- MATE-SRL
- Neural SRL
- Span- and **2** dependency-based SRL

# R1) Component Selection

How we revamp?

## Bitext Selection

Select the bitext which improves the domain adaptability of the Target languages SRL model.



Experimented with Europarl, Tatoeba and Opensubtitles

## Parser Selection

Select the best quality parser on UD 2.9 data.  
**STANZA<sup>1</sup>**



Lang.	Parser	POS	UAS	LAS
FR	spaCy/fr_core_news_lg	88.33	78.07	69.11
	spaCy/fr_core_news_sm	85.87	73.11	63.90
	spaCy/fr_dep_news_trf	96.62	91.50	85.16
	Stanza	<b>97.99</b>	<b>93.15</b>	<b>88.68</b>
	UDPipe	94.05	87.42	81.03

## Word Aligner Selection

Select the best quality word aligner on word alignment gold data. **Simalign<sup>2</sup>**



Lang.	Aligner	AER	P	R	F1
EN-FR	Awesome-align	24.54	63.05	93.96	75.46
	BerkeleyAligner	34.71	52.99	85.04	65.29
	SimAlign/argmax	<b>21.85</b>	<b>68.28</b>	91.36	<b>78.15</b>
	SimAlign/itermax	27.43	58.67	<b>95.10</b>	72.57
	SimAlign/match	30.48	55.54	92.92	69.52

[1] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020.

[2] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online.

# R1) Component Selection

How we revamp?

## EN-NNSRL Model

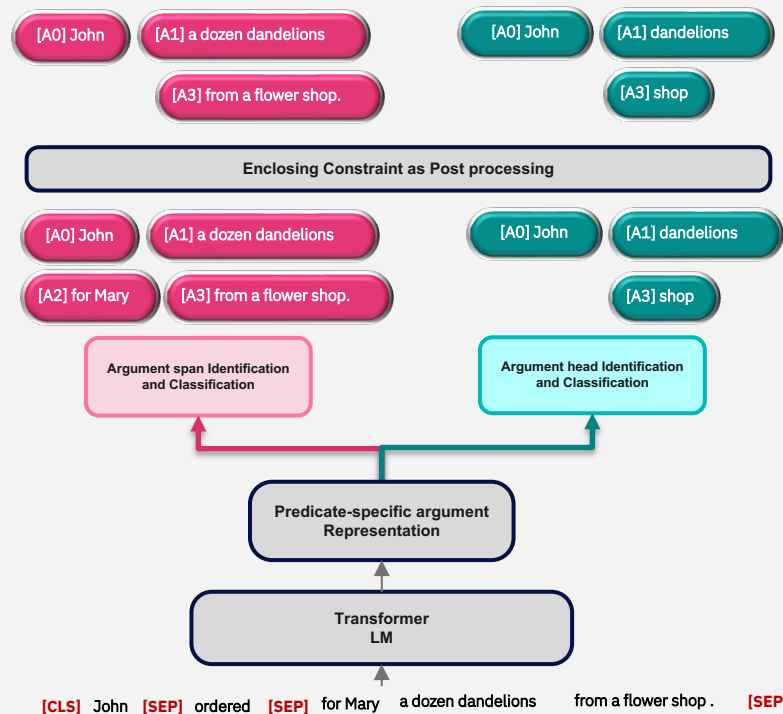
- A novel neural SRL (NNSRL) architecture to predict both span- and dependency-based SRL
- Trained on EN OntoNotes

OntoNotes argument **spans**:

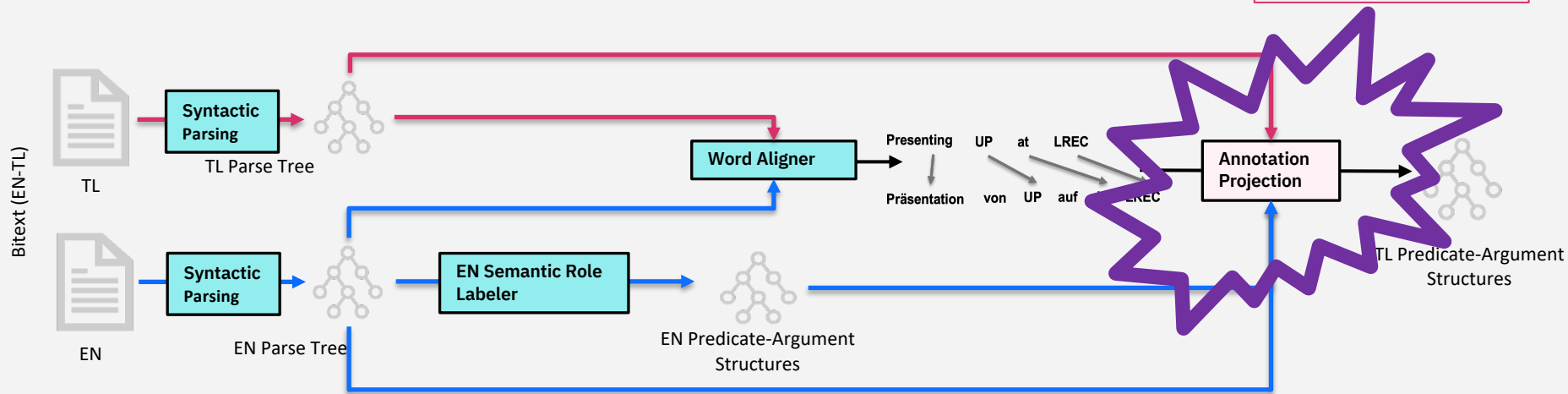
are available via LDC catalog LDC2013T19  
<https://catalog.ldc.upenn.edu/LDC2013T19>

OntoNotes argument **heads**:

obtained via transforming the constituent analysis to dependency tree using CoreNLP Library with additional postprocessing to adapt the UD syntactic analysis to the most current UD guidelines.



# Automatic Data Generation



## R1) Component Selection

1	UP1.0
2	UP2.0

### Bitext (EN-TL)

- Europarl <sup>2</sup>
- Tatoeba <sup>2</sup>
- Opensubtitles <sup>2</sup> <sup>1</sup>

### Syntactic Parser

- MATE <sup>1</sup>
- STANFORD
- Stanza <sup>2</sup>
- Spacy
- UDPipe

### Word Aligner

- Berkeley Aligner <sup>1</sup>
- Simalign <sup>2</sup>
- Awesome Align

### EN SRL Model

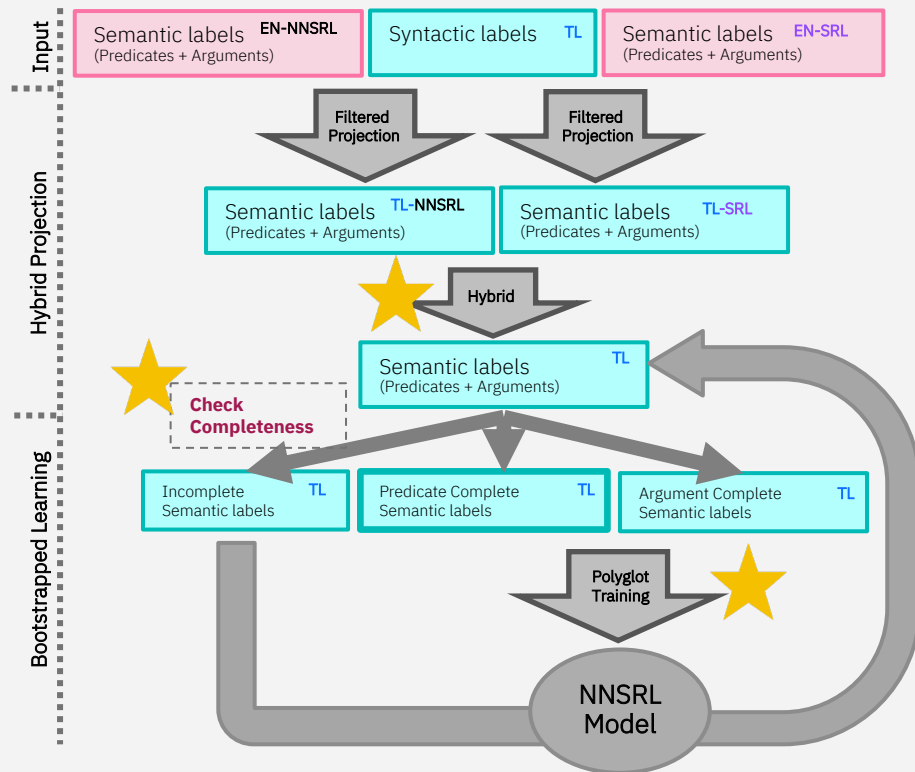
- ClearNLP <sup>1</sup>
- MATE-SRL
- Neural SRL
- Span- and <sup>2</sup> dependency-based SRL

# Annotation Projection

How we revamp?

We follow the same two-stage process in UP-1.0.

- Apply a **filtered annotation projection** to parallel corpora to achieve annotations with high precision for target-language.
- Then we **bootstrap and retrain** the TL SRL to iteratively improve recall without reducing precision.



EN-SRL: EN SRL labels using SRL model from UP1.0

EN-NNSRL: EN SRL labels using SRL model from UP2.0

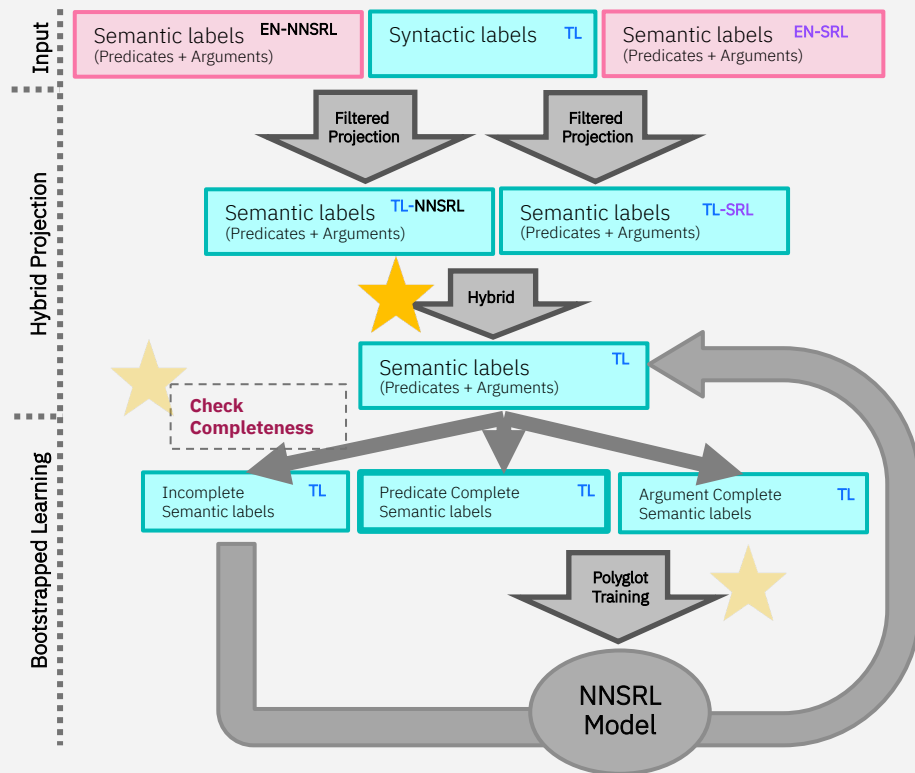


# Annotation Projection

How we revamp?

## R2) Hybrid Projection

- High-quality EN NNSRL model results in **higher projection precision**, but at the expense of low recall for predicate identification.
- One of the major advantages of syntax-based EN SRL model is its very **high recall for predicate identification**.
- Augment the predicates of TL-NNSRL with TL-SRL.



EN-SRL: EN SRL labels using SRL model from UP1.0

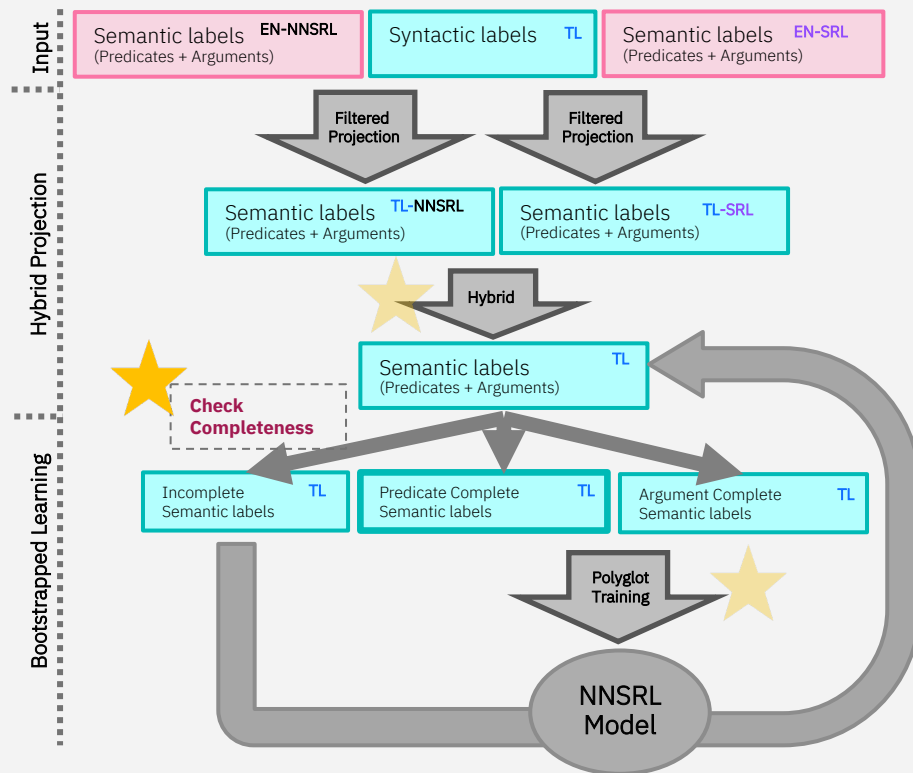
EN-NNSRL: EN SRL labels using SRL model from UP2.0

# Annotation Projection

How we revamp?

## R3) Completeness Measure

- **Predicate Completeness:** A sentence in TL is deemed predicate-complete if it has the *same number of predicates as its corresponding EN sentence*, where the predicates of the EN sentence are those identified using a trained NNSRL model and the predicates of the TL sentence are obtained via annotation projection.
- **Argument Completeness** (Equivalent to *k*-complete in UP1.0). A direct component of a labeled sentence in TL is either a verb in TL or a syntactic dependent of a verb. Then a sentence in TL is *k*-complete if it contains equal to or fewer than *k* unlabeled direct components. 0-complete is abbreviated as argument-complete.



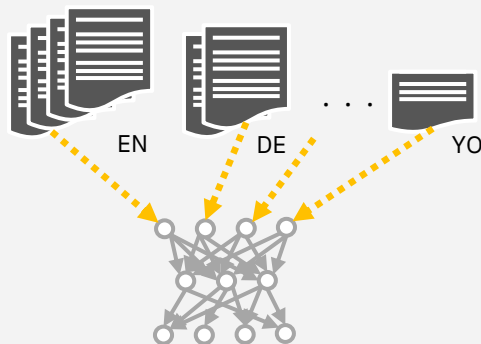
EN-SRL: EN SRL labels using SRL model from UP1.0

EN-NNSRL: EN SRL labels using SRL model from UP2.0

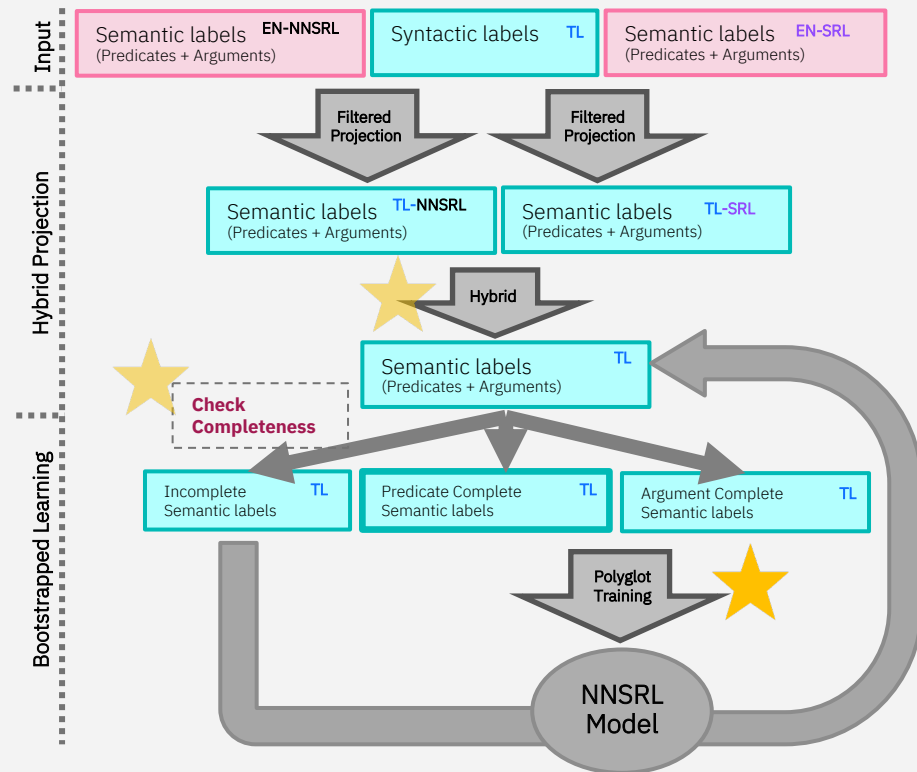
# Annotation Projection

How we revamp?

## R4) Polyglot Training



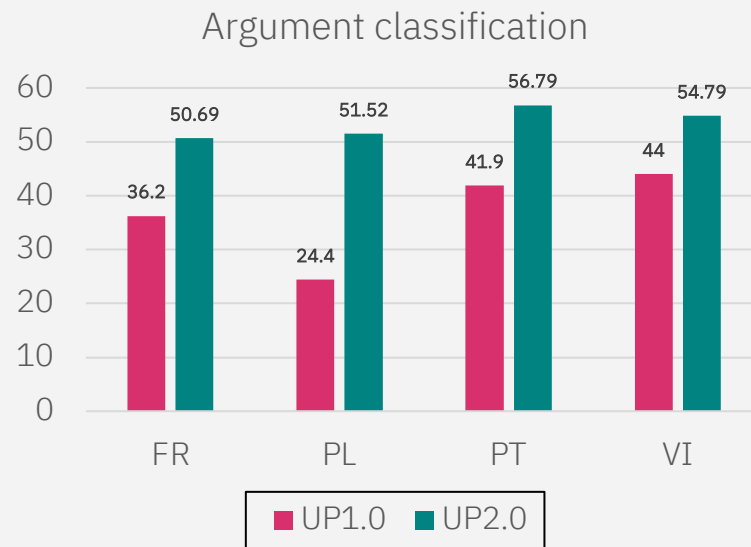
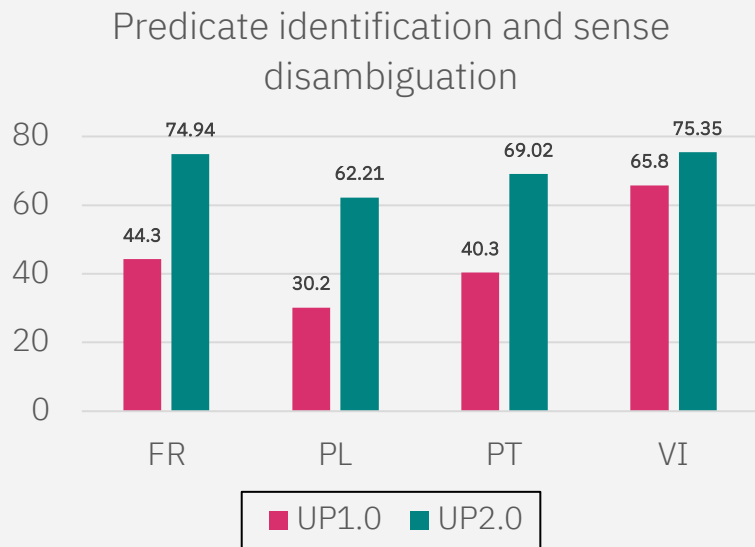
- Models trained jointly on multiple languages with homogeneous annotations generalize better across languages.



EN-SRL: EN SRL labels using SRL model from UP1.0

EN-NNSRL: EN SRL labels using SRL model from UP2.0

# Results: Overall Performance

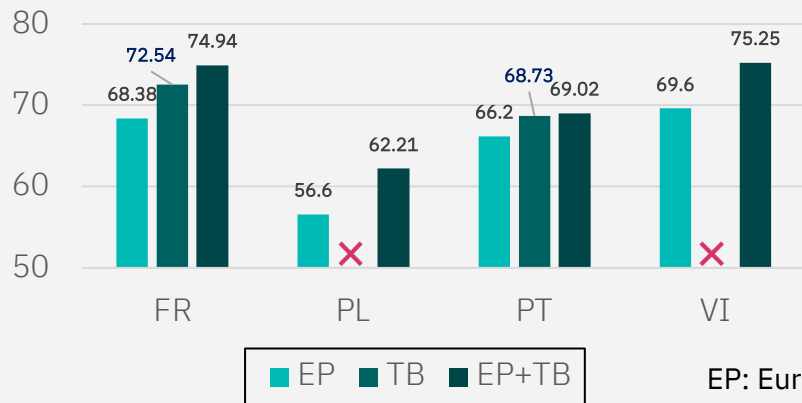


- > 10 points F1 ↑ on predicates and arguments

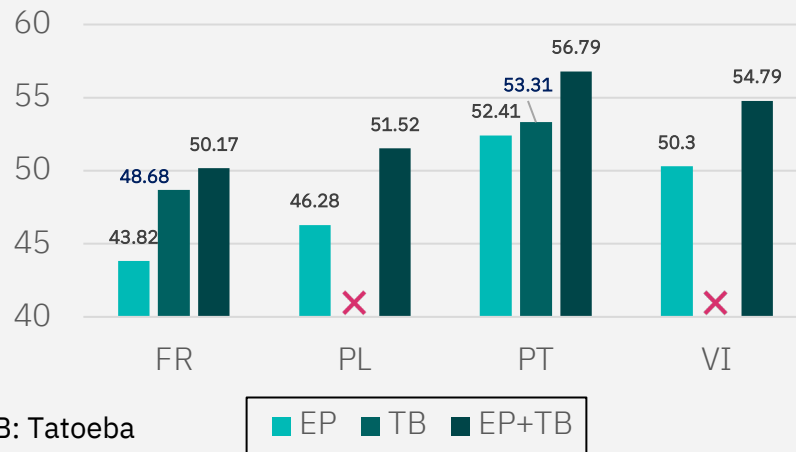
# Results: Bitext Selection

Analysis?

Predicate identification and sense disambiguation



Argument classification



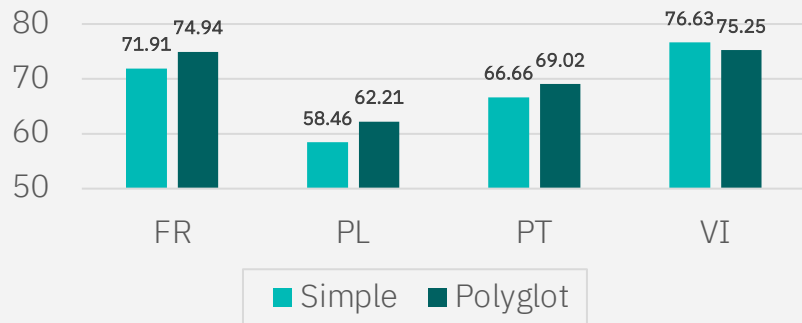
✗ means training data is too low to train NN model

- Combined Bitext from different domain improves the generalizability of TL SRL model

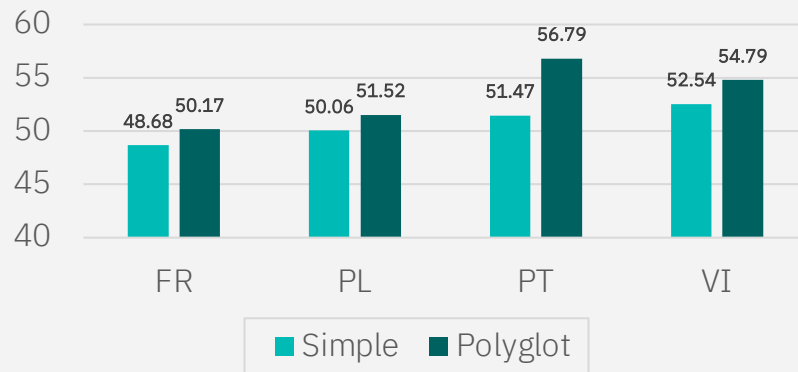
# Results: Polyglot Training

Analysis?

Predicate identification and sense disambiguation



Argument classification



- Supervision from **EN always helps** TL both for predicates and arguments.
- >1 point lower on VI data predicates
  - Due to language dissimilarity.

# Propbank Data

What we generate?

## Propbank data for 23 languages

A UP release contains treebanks of the corresponding UD release.

To differentiate Gold from UP data we use the following conventions:

*UP\_<language>-<corpus>*

*GOLD\_<language>-<corpus>*

In addition, each language has a folder with verb overview files (produced from the frame files) in HTML format. These files can be viewed in a browser and give an overview of all English frames that each target language verb can evoke.

<https://github.com/UniversalPropositions>

Lang.	UP2.0		#Unique Frames
	#Arg comp.	#Pred comp.	
cs	257K	71K	2991
de	453K	262K	2977
el	282K	80K	5044
es	613K	139K	2833
fi	512K	181K	1848
fr	698K	180K	2517
hi	109K	150K	413
hu	162K	47K	2713
id	920K	717K	4972
it	606K	256K	2771
ja	127K	100K	2942
ko	42K	18K	1718
mr	5K	6K	167
nl	457K	136K	2656
pl	223K	40K	2354
pt	788K	152K	2978
ro	147K	55K	1495
ru	641K	417K	4683
ta	22K	24K	458
te	16K	14K	678
uk	128K	81K	2396
vi	359K	420K	1261
zh	389K	314K	4408

# Gold Data

What we generate?

## Polish

- We select 100 English sentences from EN OntoNotes and translate them into Polish.
- Then we manually label all the predicates and arguments according to English PropBank

## Portuguese

- We merge the two resources (Propbank.Br, UD Bosque ), projecting the SRL annotation from Propbank.Br on top of the dependencies from UD Bosque (UD 2.9) solving inconsistencies and fixing annotation errors

## Vietnamese

- We manually labeled 378 examples randomly selected from Tatoeba corpus.

Lang.	#Sentences	#Predicates	#Arguments
EN <sub>GOLD</sub>	16622	50258	101603
PL <sub>GOLD</sub>	100	223	495
PT <sub>GOLD</sub>	3779	6173	15097
VI <sub>GOLD</sub>	378	770	1625

Characteristics of gold data for each language.



## Components quality

- **Parser quality:** underlying parser mistakes in identifying the correct lemma for certain verbs. Therefore, there exists certain frame filenames that do not make sense in that particular language. For example
  - *επέμβουμ* is lemmatized as *επιμβάνω* but it should be *επεμβαίνω*. This unnecessarily cause incorrect projection.

## Language peculiarities

- For the languages where subject/object can be omitted, one may expect to observe incorrect role label transfer. One potential reason for such issues is incorrect word alignment.
- AUX (be, have, do) in EN is likely to be misaligned with other tokens in other languages. In EN, these AUX are used to construct tenses (perfect perfective), polarity etc., but different languages represent tense and polarity differently.
  - I **have been reading** for a week. (aux verb)
  - Ich **lese seit** einer Woche. (prep)
  - 我**已经读了**一个星期. (adv, participle)

# What's Next

Further improve the quality of generated propbanks

- fixing known issues

Generate Gold data for more languages to facilitate benchmarking in multilingual SRL

Community engagement with open-source collaborations.

# Questions?

Dataset



<https://github.com/UniversalPropositions>



For more examples



<https://universalpropositions.github.io/#introduction>