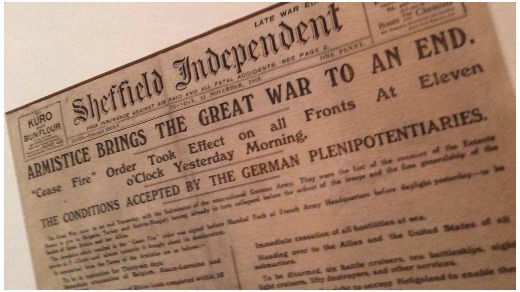
NLP in the Wild

From Akkadian to Biochemistry

Gabriel Stanovsky

THE HEBREW UNIVERSITY OF JERUSALEM



JACOB BORNSTEIN, M.D.
HAROLD N. BORNSTEIN, M.D., P.C.
101 EAST 78TH STREET
NEW YORK, NY 10075-0301
TELE: (212) 988-6600 FAX: (212) 988-6602
E-mail: hbornst1@gmail.com
www.haroldbornsteinmd.com

Over the past twelve months, he has lost at least fifteen pounds. Mr. Trump takes 81 mg of aspirin daily and a low dose of a statin. His PSA test score is 0.15 (very low). His physical strength and stamina are extraordinary.

Mr. Trump has suffered no form of cancer, has never had a hip, knee or shoulder replacement or any other orthopedic surgery. His only surgery was an appendectomy at age ten. His cardiovascular status is excellent. He has no history of ever using alcohol or tobacco products.

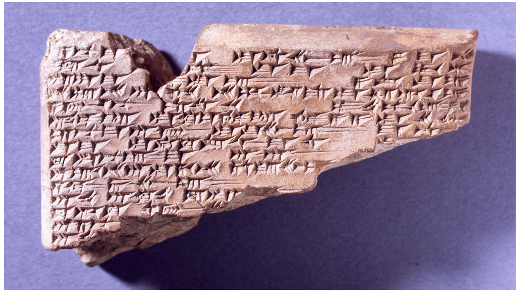
UNITED STATES DISTRICT COURT
DISTRICT OF IDAHO

Plaintiff(s),	)	Case No.: CV-08-999-RHW
v.	)	**Transcript Redaction Request**
Defendant(s).	)	

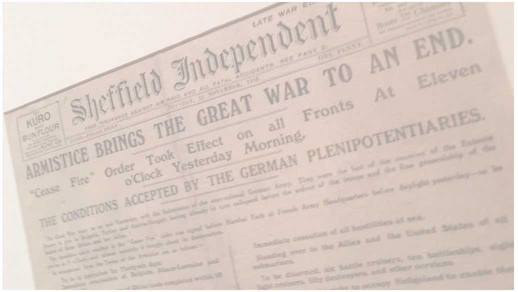
Pursuant to Fed.R.Civ.P. 5.2/Fed.R.Crim.P. 49.1, Plaintiff/Defendant requests that the following personal identifiers be redacted from the transcript filed on __________:

- Redact the Social Security number on page 12, line 8 to read xxx-xx-1111;
- Redact the Taxpayer identification number on page 32, line 5, to read xxxxxxx2233;

Language is Everywhere



```
1: Combine in a vial 50 ng of vector with molar excess of insert.
2: Adjust with dH2O.
3: Add 10 µl of Ligation Buffer and mix.
4: Add 1µl of T4 DNA Ligase and mix thoroughly.
5: Centrifuge briefly and incubate for 5 minutes.
6: Chill ligation mixture on ice.
```


Sheffield Independent
ARMISTICE BRINGS THE GREAT WAR TO AN END.
"Cease Fire" Order Took Effect on all Fronts At Eleven o'Clock Yesterday Morning.
THE CONDITIONS ACCEPTED BY THE GERMAN PLENIPOTENTIARIES.

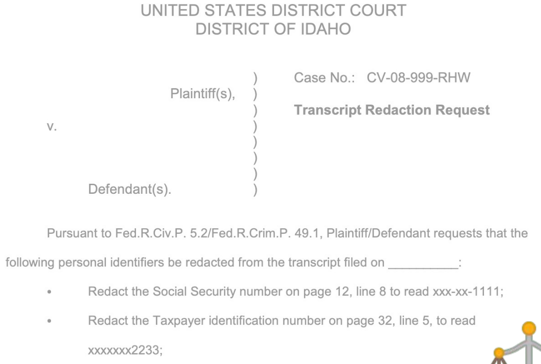
JACOB BORNSTEIN, M.D.
HAROLD N. BORNSTEIN, M.D., P.C.
101 EAST 78TH STREET
NEW YORK, NY 10075-0301
TELE: (212) 988-6600 FAX: (212) 988-6602
E-mail: hbornst1@gmail.com
www.haroldbornsteinmd.com

December 4, 2015

Over the past twelve months, he has lost at least fifteen pounds. Mr. Trump takes 81 mg of aspirin daily and a low dose of a statin. His PSA test score is 0.15 (very low). His physical strength and stamina are extraordinary.

To Whom My Concern:

Mr. Trump has suffered no form of cancer, has never had a hip, knee or shoulder replacement or any other orthopedic surgery. His only surgery was an appendectomy at age ten. His cardiovascular status is excellent. He has no history of ever using alcohol or tobacco products.


UNITED STATES DISTRICT COURT
DISTRICT OF IDAHO

Plaintiff(s), Case No.: CV-08-999-RHW

v. Transcript Redaction Request

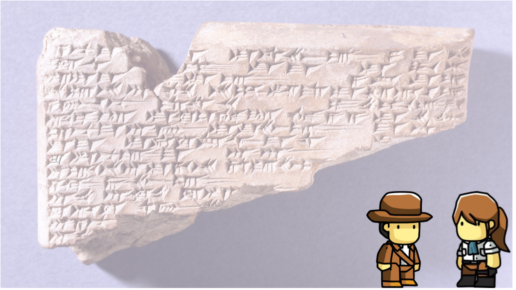
Defendant(s).

Pursuant to Fed.R.Civ.P. 5.2/Fed.R.Crim.P. 49.1, Plaintiff/Defendant requests that the following personal identifiers be redacted from the transcript filed on __________:

- Redact the Social Security number on page 12, line 8 to read xxx-xx-1111;
- Redact the Taxpayer identification number on page 32, line 5, to read xxxxxxx2233;

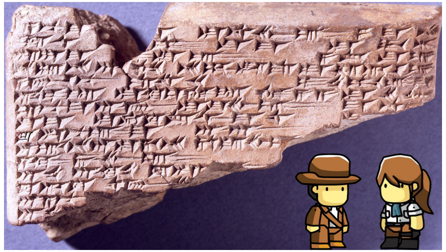
Language is Everywhere

Many ***Interdisciplinary*** *research questions can be addressed with NLP*

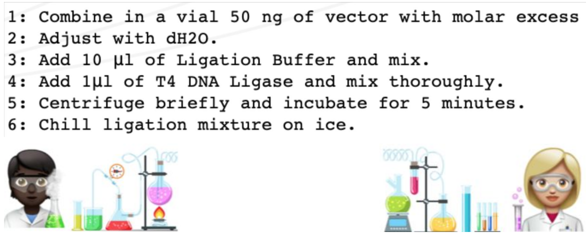


```
1: Combine in a vial 50 ng of vector with molar excess of insert.
2: Adjust with dH2O.
3: Add 10 µl of Ligation Buffer and mix.
4: Add 1µl of T4 DNA Ligase and mix thoroughly.
5: Centrifuge briefly and incubate for 5 minutes.
6: Chill ligation mixture on ice.
```

This Talk



Filling gaps in cuneiform tablets



Understanding scientific protocols

This Talk

Filling gaps in cuneiform tablets

```
1: Combine in a vial 50 ng of vector with molar excess
2: Adjust with dH2O.
3: Add 10 µl of Ligation Buffer and mix.
4: Add 1µl of T4 DNA Ligase and mix thoroughly.
5: Centrifuge briefly and incubate for 5 minutes.
6: Chill ligation mixture on ice.
```

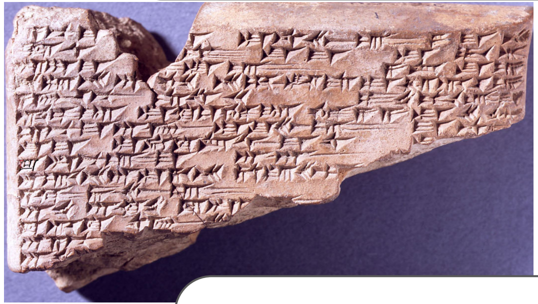

Understanding scientific protocols

When transcribing ancient tablets found in archeological sites we need to **“fill in”** gaps formed in the stone due to erosion over 1000s of years

Prof. Wayne Horowitz

Prof. Nathan Wasserman

The Institute of Archaeology, Hebrew University



and how do you know how to fill in those missing parts?

Koren Lazar
Grad student

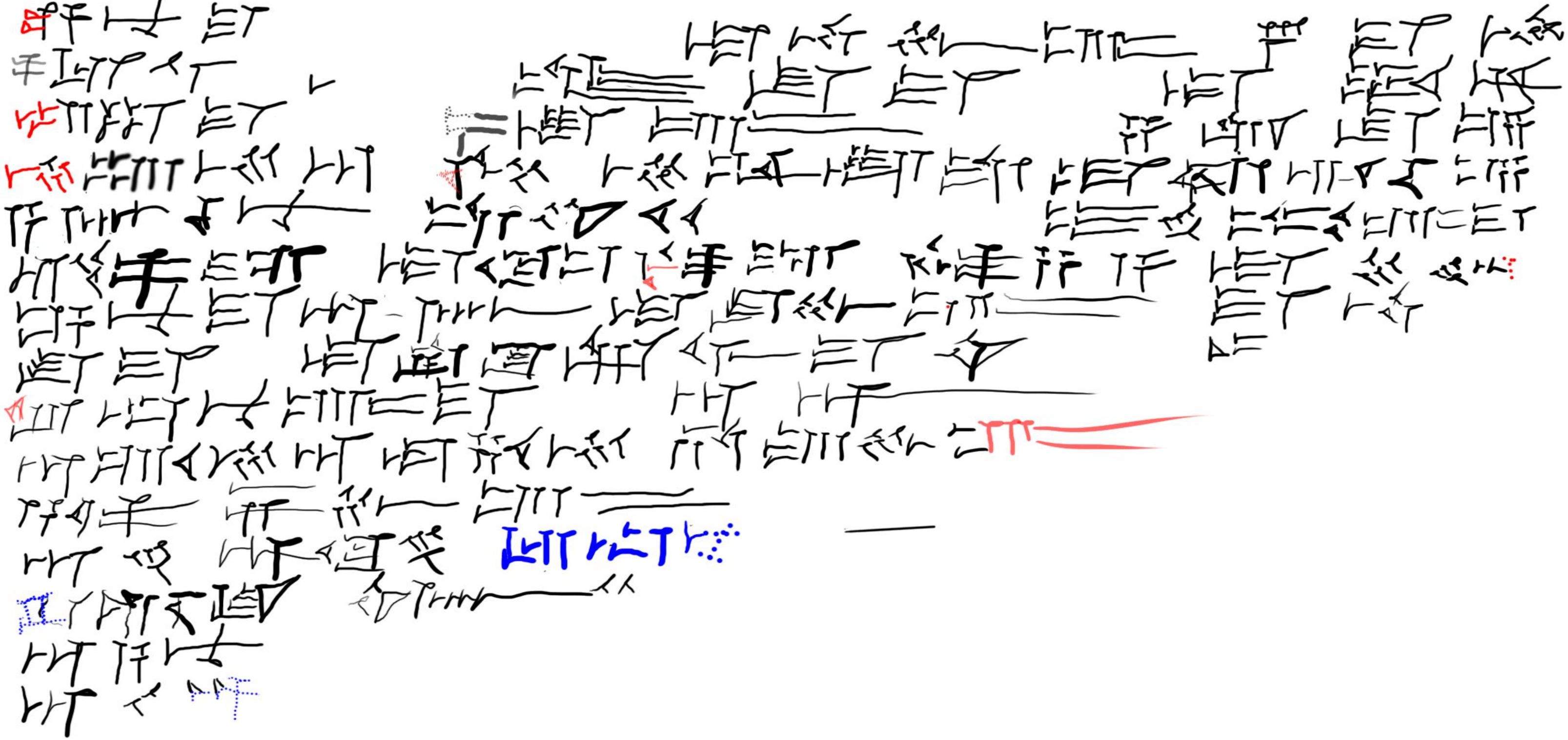
me

Well, we look at the symbols we recognize in the surrounding context, and try to **guess** the most **probable sequence**

That sounds awfully familiar...

The Akkadian language

- Spoken in Mesopotamia (2500 BCE - 100 AD)
- Earliest attested Semitic language
- Lingua Franca of the ancient world

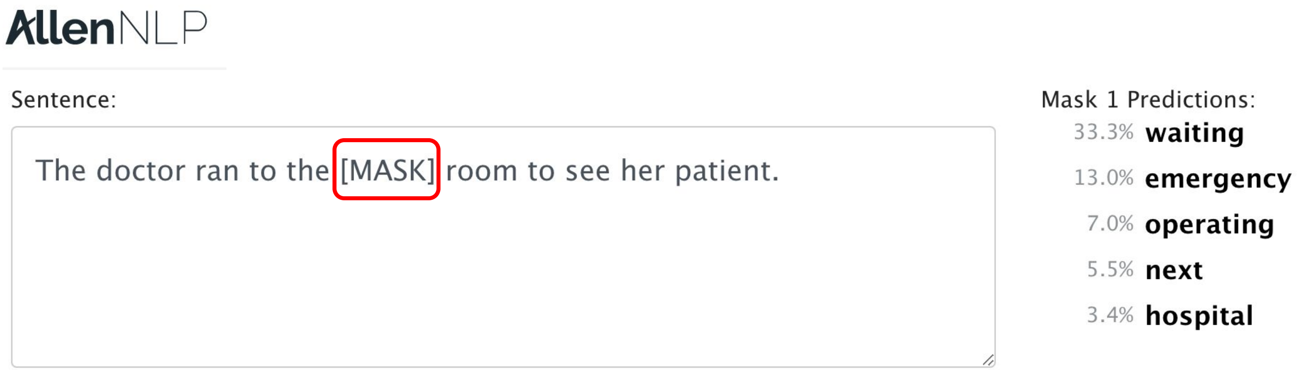


[e]-nu-ma [e-li]š la na-bu-ú ša-ma-mu
[šap]-liš [am]-[ma]-tum šu-ma la zak-rat
[ZU].AB-ma [re]š-tu-ú za-ru-šu-un
[mu]-um-mu D [ti]-amat mu-al-li-da-at gim-ri-šu-un
A.MEŠ-šu-nu iš-te-niš i-hi-qu-ú-ma
gi-pa-ra la ki-iṣ-ṣu-ra ṣu-ṣa-a la še-'u
e-nu-ma D MEŠ la šu-pu-ú ma-na-[ma]
šu-ma la zuk-ku-ru ši-ma-tu [la]
šub-ba-nu-ú-ma D D
D lah-mu D la-ha-mu uš-ta-pu-[ú]
a-di ir-bu-ú
AN.ŠAR D KI.ŠAR ib-ba-nu
[u]r-ri-ku u4.MEŠ
D a-nu
AN.[ŠAR] D

Filling in the gaps

- Tablets deteriorate creating gaps, blurred signs

- Can contextual language models predict the missing parts?
  - downstream task == pretraining task!



Limited available data: ORACC

- The ORACC corpus collects transliterations

- 1M words <<< 3B words in English BERT

Language or Dialect (abbreviation in the CLI dataset)	Texts	Lines	Signs
Sumerian (SUX)	5,000	107,345	*c.* 400,000
Old Babylonian (OLB)	527	7,605	*c.* 65,000
Middle Babylonian peripheral (MPB)	365	11,015	*c.* 95,000
Standard Babylonian (STB)	1,661	35,633	*c.* 390,000
Neo-Babylonian (NEB)	1,212	19,414	*c.* 200,000
Late Babylonian (LTB)	671	31,893	*c.* 260,000
Neo-Assyrian (NEA)	3,570	65,932	*c.* 490,000

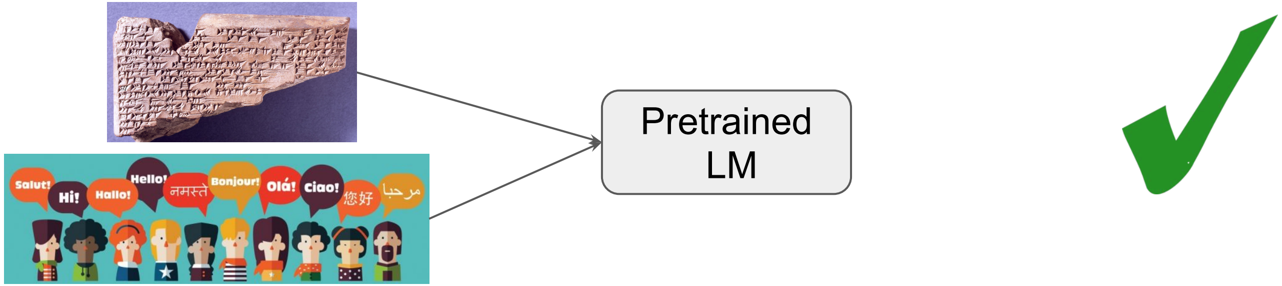
Method

- **Failed attempt:** Train BERT *from scratch* on ORACC

Method

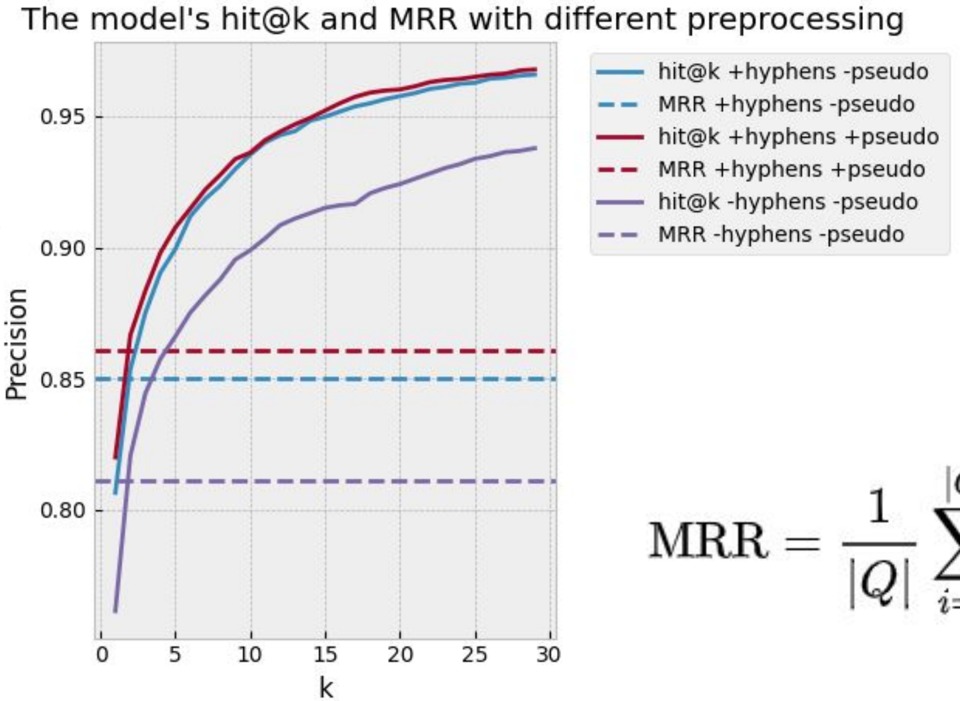
- **Failed attempt:** Train BERT *from scratch* on ORACC

- **(much) better results:** Finetune M-BERT on ORACC



Method

- **(much) better results:** Finetune M-BERT on ORACC



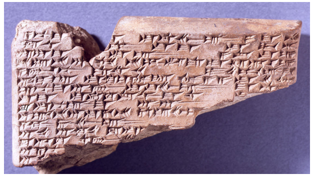
Akkadian benefits from pretraining of modern languages!

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}$$

Human Evaluation: Interface

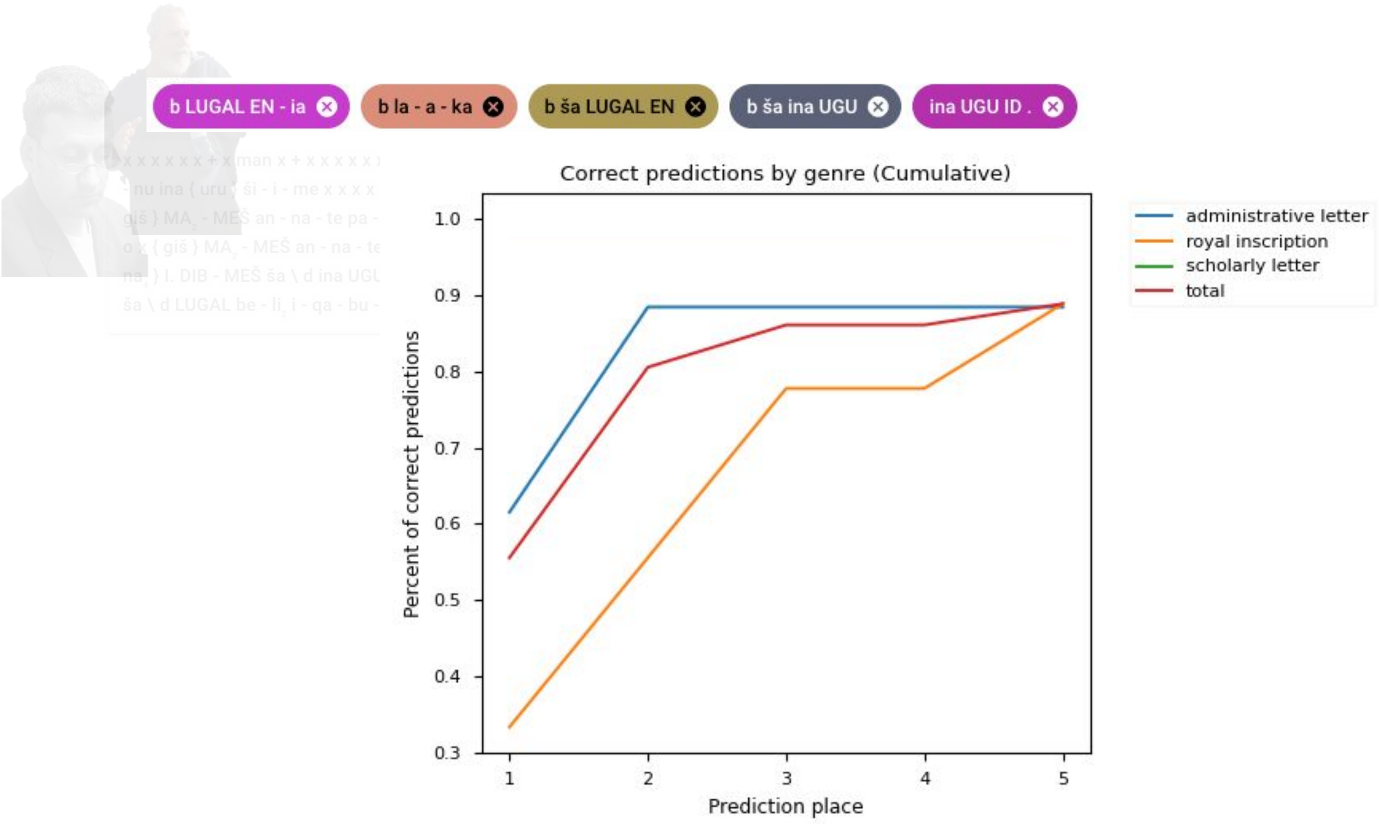
b LUGAL EN - ia | b la - a - ka | b ša LUGAL EN | b ša ina UGU | ina UGU ID .

x x x x x x + x man x + x x x x x x x + x - ni ši - na x x x x + x KUR - e ša \ d - ni - u x x x x x + x - ha - ni i - sa - hu - ra x x x x x + x dan - nu ina { uru } ši - i - me x x x x - hi ša \ d { giš } MA$_2$ - MEŠ KALAG - MEŠ x x x ta - ha - ni - šu$_2$ - ni mar ša \ d i - ba - šu - ni x x x x { giš } MA$_2$ - MEŠ an - na - te pa - a - ṣa x x x - da - du i - sa - hu - ra x x x x - ha - ni - šu$_2$ - nu i - su - ri LUGAL be - li$_2$ i - qa - bi ma - a o x { giš } MA$_2$ - MEŠ an - na - te o x x x x x x x x x i x x x x x x x x x { na$_4$ } I. DIB - MEŠ x x x x x x x na - me - ri x x x x x x x x - ni { na$_4$ } I. DIB - MEŠ ša \ d ina UGU ID$_2$ kar - ra - a - ni u$_2$ - še - ba - ra XXXX dul$_6$ - lu ša \ td { giš } MA$_2$ - MEŠ e - pa - aš$_2$ mi - i - nu ša \ d LUGAL be - li$_2$ i - qa - bu - ni



Key	Value
id_text	P313445
genre	administrative letter
period	Neo-Assyrian
language	**Akkadian**
provenience	Nineveh
project_name	saao/saa05
url	http://oracc.iaas.upenn.edu/saao/saa05/

Human Evaluation: Initial Results



Open Questions

- What kind of errors does the model make?
- What is the inter-annotator agreement?
- Pretraining with some languages helps more than others?
  - E.g., semitic languages

This Talk

Filling gaps in cuneiform tablets

```
1: Combine in a vial 50 ng of vector with molar excess
2: Adjust with dH2O.
3: Add 10 µl of Ligation Buffer and mix.
4: Add 1µl of T4 DNA Ligase and mix thoroughly.
5: Centrifuge briefly and incubate for 5 minutes.
6: Chill ligation mixture on ice.
```

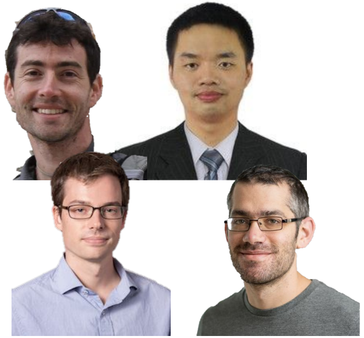
Understanding scientific protocols

Protocols in scientific experiments describe **executable actions** in a lab, but they're so **hard to reproduce reliably**

Why is it hard? The texts seem objective & precise

If you look closely, you can see that many details aren't specified:
What's "thoroughly" ?
What's "briefly" ?

Ronen Tamari   Fan Bai



Alan Ritter   me

```
1: Combine in a vial 50 ng of vector with molar excess of insert.
2: Adjust with dH2O.
3: Add 10 µl of Ligation Buffer and mix.
4: Add 1µl of T4 DNA Ligase and mix thoroughly.
5: Centrifuge briefly and incubate to 25°C for 5 minutes.
6: Chill ligation mixture on ice.
```

Can we design a representation that is both **lenient & executable**?

Wet lab protocols

14 word / sent

13 sents / doc

Complex coref &
cross-sent relations

Temporally-dependant actions

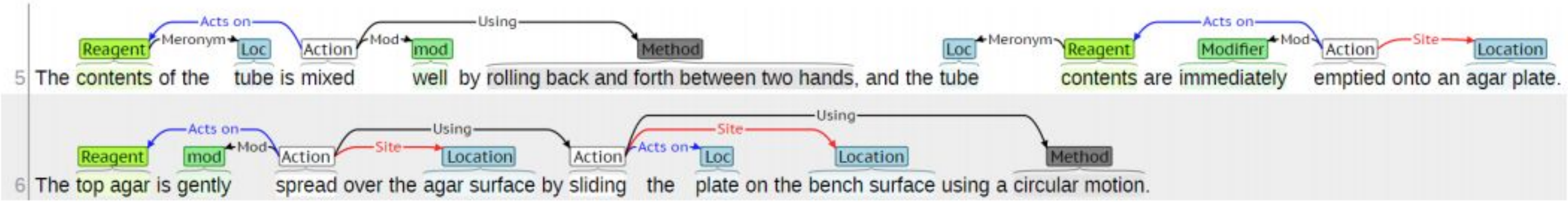
```
1: Combine in a vial 50 ng of vector with molar excess of insert.
2: Adjust with dH2O.
3: Add 10 µl of Ligation Buffer and mix.
4: Add 1µl of T4 DNA Ligase and mix thoroughly.
5: Centrifuge briefly and incubate to 25°C for 5 minutes.
6: Chill ligation mixture on ice.
```

Executable semantic parsing

- **Lab protocols as an executable program?**
- Benefits lab technicians when reproducing experiments
- Similar to other procedural text understanding (e.g., recipes)

Existing work

WLP (Kulkarni et al., 2018)

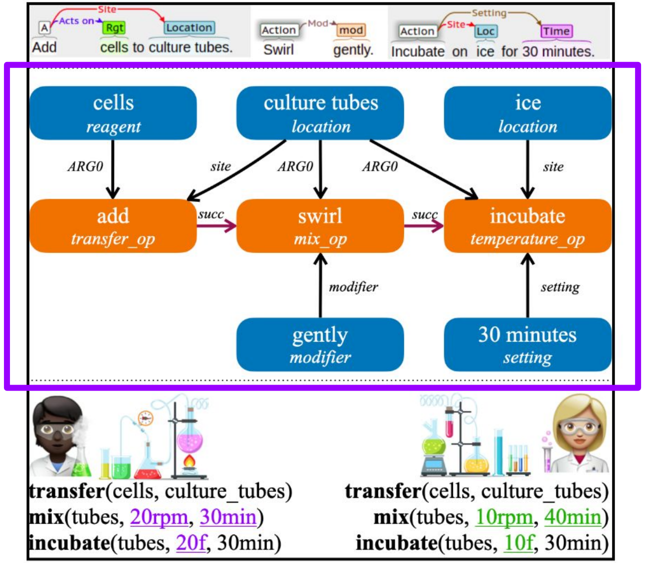


- SRL-like **Sentence-level** predicate-argument annotation
- Doesn’t capture cross-sentence relations
- No notion of execution

Our proposal: Process Execution Graphs (PEG)

Tamari et al., EACL 2021

- Process-level abstract executable representation
- Bridges between procedural text and automated execution



PEG: Definitions

- Directed, a-cyclic labeled graph
- Ontology based on Autoprotocol

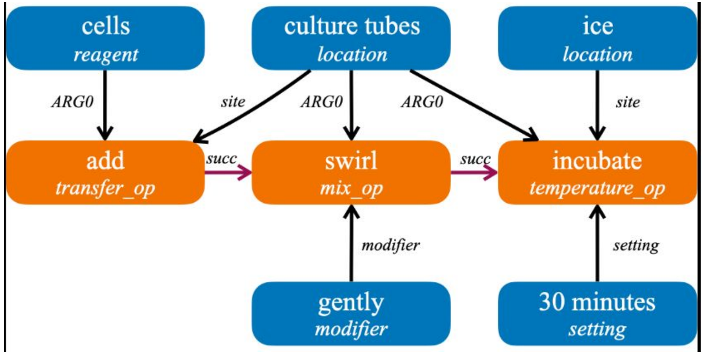
seal

Containers must be covered or sealed for storage, incubation, and centrifugation operations (among others). Seal `type`s have useful properties ranging from optical clarity to gas permeability. Seals can be applied by either `thermal` or `adhesive` sealers which result in different seal integrity. `thermal` seals can be applied with a range of temperatures and durations that can be optimized for different plate types. Many instructions including liquid handling operations require that a container be uncovered before use.

```
{
  "op": "seal",
  "object": Container,
  "type": String,
  "mode": Option<Enum("thermal", "adhesive")>,
  "mode_params": Option<{
    "temperature": Option<Temperature>,
    "duration": Option<Time>
  }>
}
```

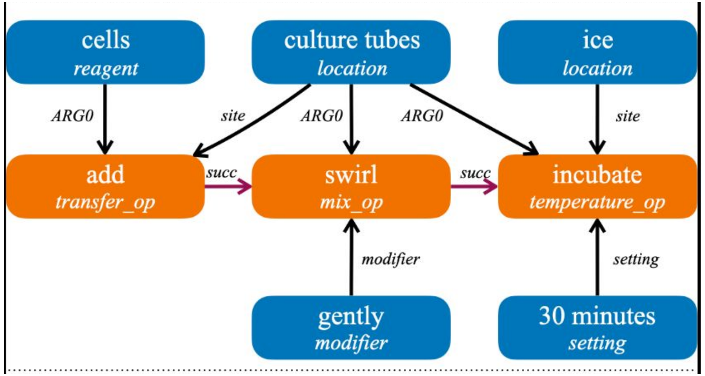
PEG: Definitions

- Directed, a-cyclic labeled graph
- Ontology based on Autoprotocol
- Nodes
  - Predicates (**`mix`**, **`transfer`**)
  - Arguments
    - Physical lab entities (device, reagent)
    - Abstract entities (amounts, modifiers)



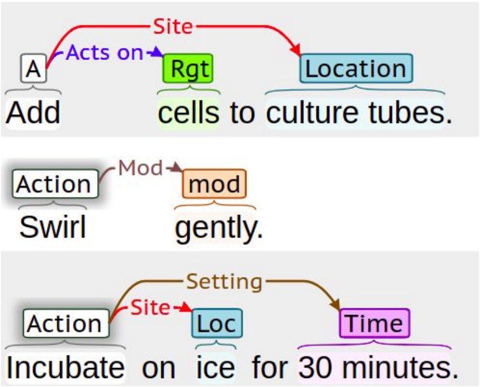
PEG: Definitions

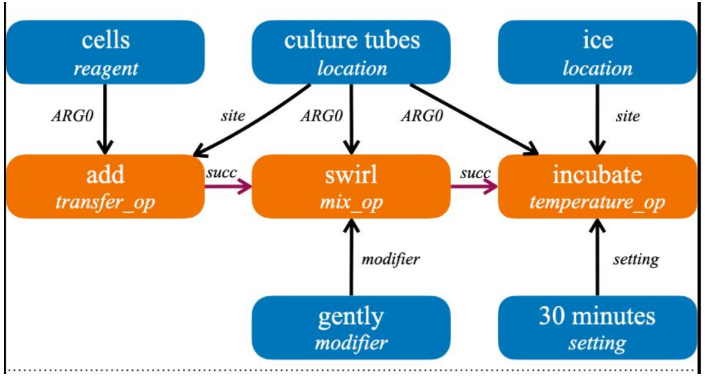
- Directed, a-cyclic labeled graph
- Ontology based on Autoprotocol
- Edges
  - Core-roles (~positional arguments)
  - Non-core roles (predicate agnostic)
  - Temporal dependency relation



Comparison with action-graphs

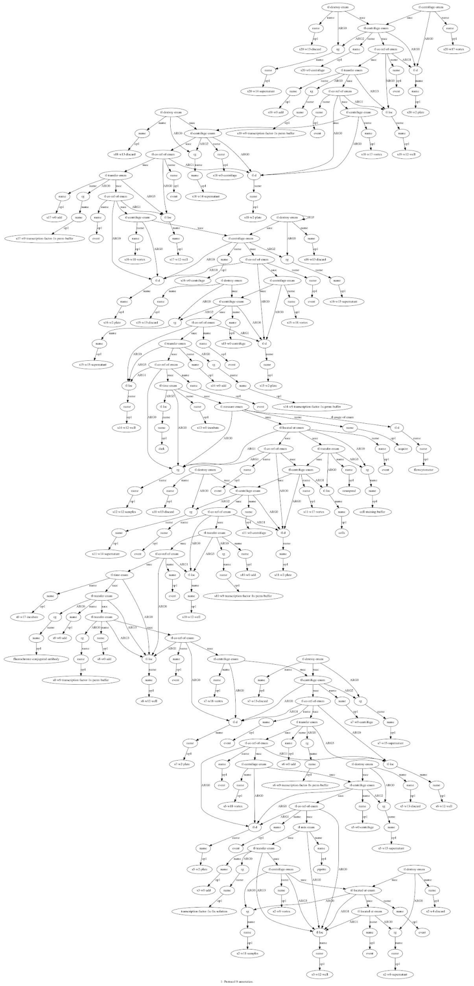
- Fine-grained operation types
- Cross-sentence relations
- Argument re-use: arguments can be persistent objects
- Enforcing required arguments





Annotation interface

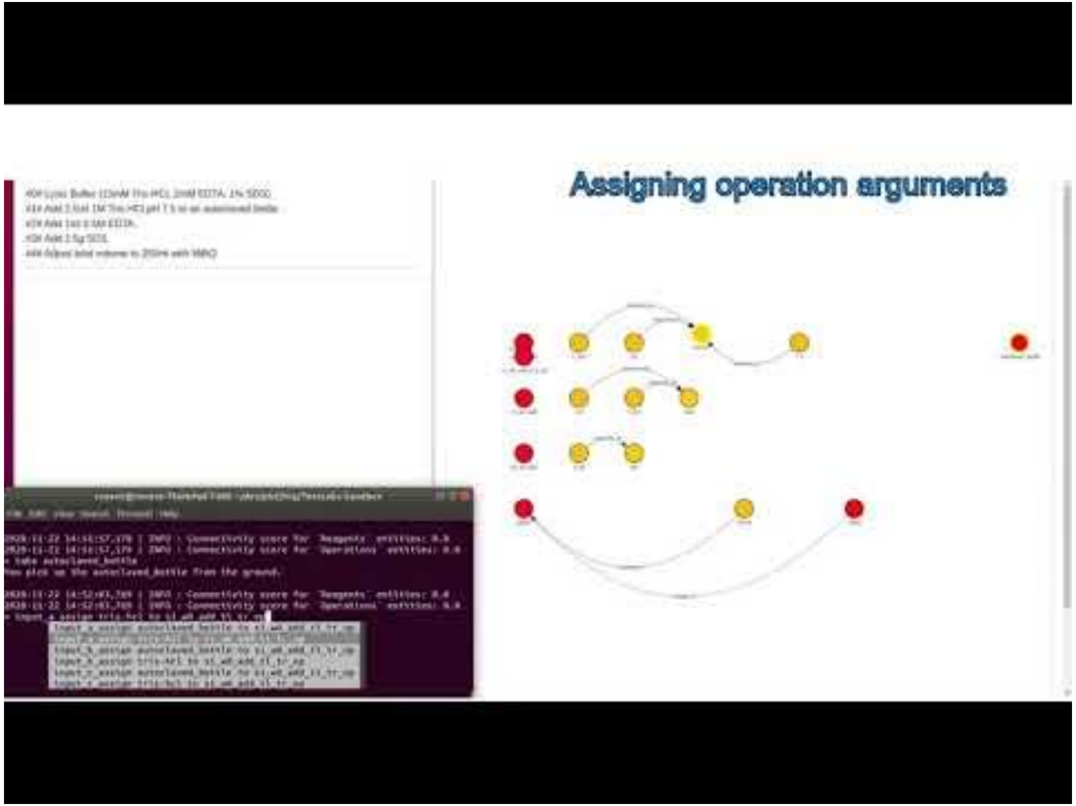
- Predicate specific execution semantics
  - (container moves -> containee moves)
- Tracking temporal dependencies and entity states over long texts
- Argument validation



Too complex for span-based annotation!

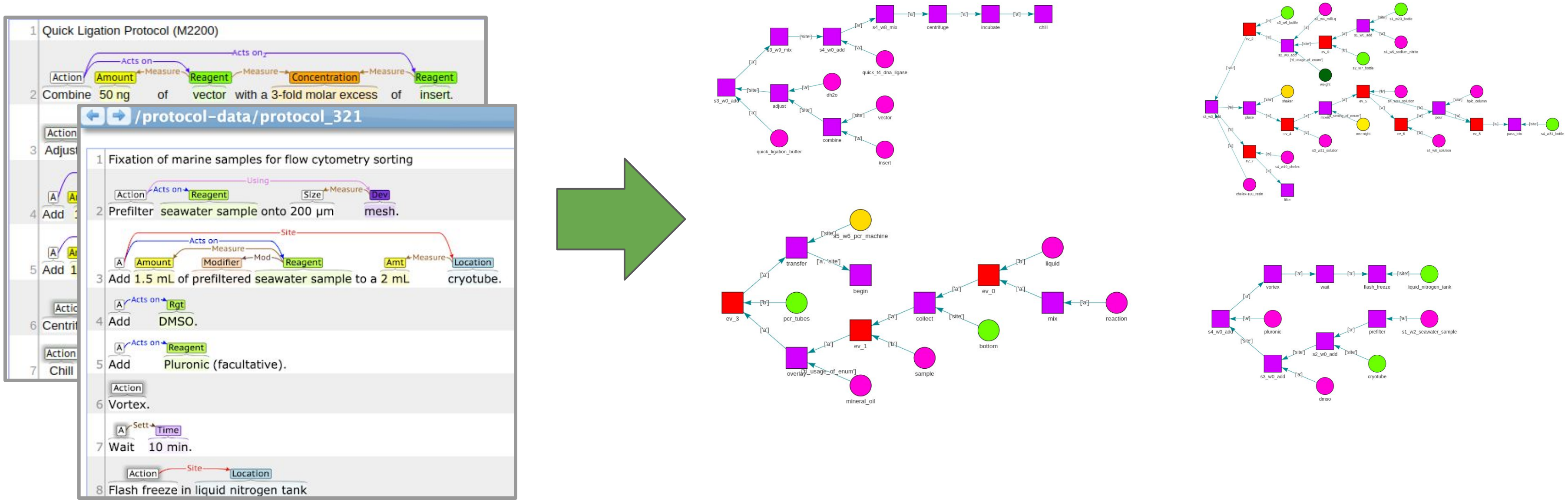
PEG visualization in AMR using AMRICA (Safra & Lopez, 2015)

Demo



X-WLP stats

- 3 annotators, enriched 45% ofWLP protocols to PEG format



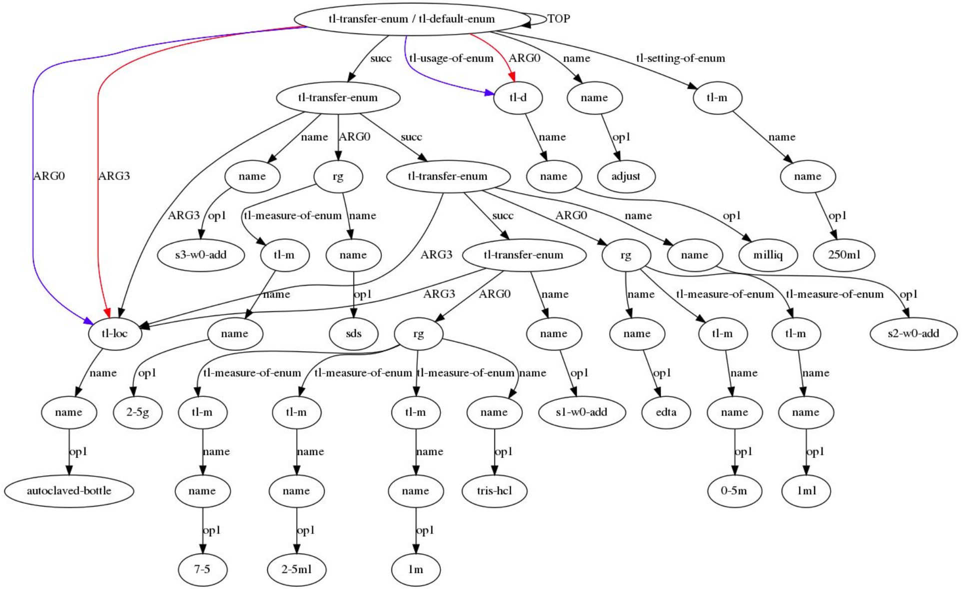
X-WLP stats

- 3 annotators, enriched 279/622 (45%) WLP protocols to PEG format
- Comparable with other procedural text datasets

	X-WLP (ours)	MSPTC	CSP	ProPara
# words	54k	56k	45k	29k
# words / sent.	12.7	26	25.8	9
# sentences	4,262	2,113	1,764	3,300
# sentences / docs.	15.28	9	N/A	6.8
# docs.	279	230	N/A	488

Quantitative analysis: annotator agreement

- Use Abstract Meaning Representation (AMR) format for established graph agreement metrics (Smatch, Cai & Knight, 2013)



1- Protocol 604 annotation.

Quantitative analysis: annotator agreement

- Use Abstract Meaning Representation (AMR) format for established graph agreement metrics Mean 84.99
- F1 Smatch comparable to AMR datasets (69 - 89 F1)

Benefits from underlying WLP annotations

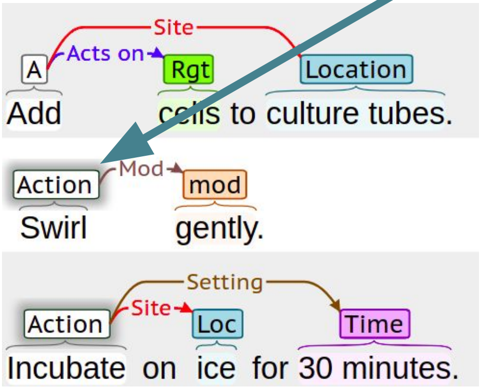
Longer-range, often cross sentence relations

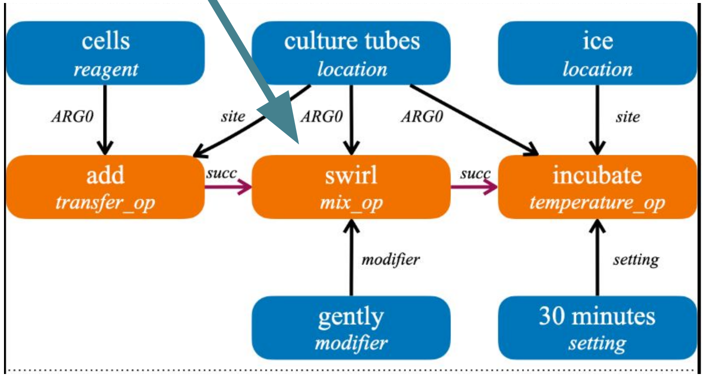
Agreement Metric	F1
Smatch	84.99
Argument identification	89.72
Predicate identification	86.68
Core roles	80.52
Re-entrancies	73.12

Quantitative analysis: operation arguments

- Simulator input validation prevents semantic underspecification, increases overall argument count per op.

Dataset	Avg. #args/op	#Ops. w/o core arg.	#Ops.	Pct.
WLP	1.87	3297	17485	18.9
X-WLP	3.01	0	3915	0.0





Quantitative analysis: relation types

- Significant proportion of arguments are re-entrancies (>30%)
- Many cross-sentence coreference relations (>90%)
  - provide process-level structure

Relation	# Intra.	# Inter.	Total	# Re-entrancy
Core				
• ARG0	2962	952	3914	1645
• ARG1	560	127	687	3
• ARG2	84	123	207	77
Total (core)	3606	1202	4808	1725
Non-Core				
• site	1306	325	1631	360
• setting	3499	2	3501	-
• usage	1114	24	1138	-
• co-ref	129	1575	1704	-
• located-at	199	72	271	-
• measure	2936	18	2954	-
• modifier	1861	2	1863	-
• part-of	72	65	137	-
Total (non-core)	11116	2083	13199	360

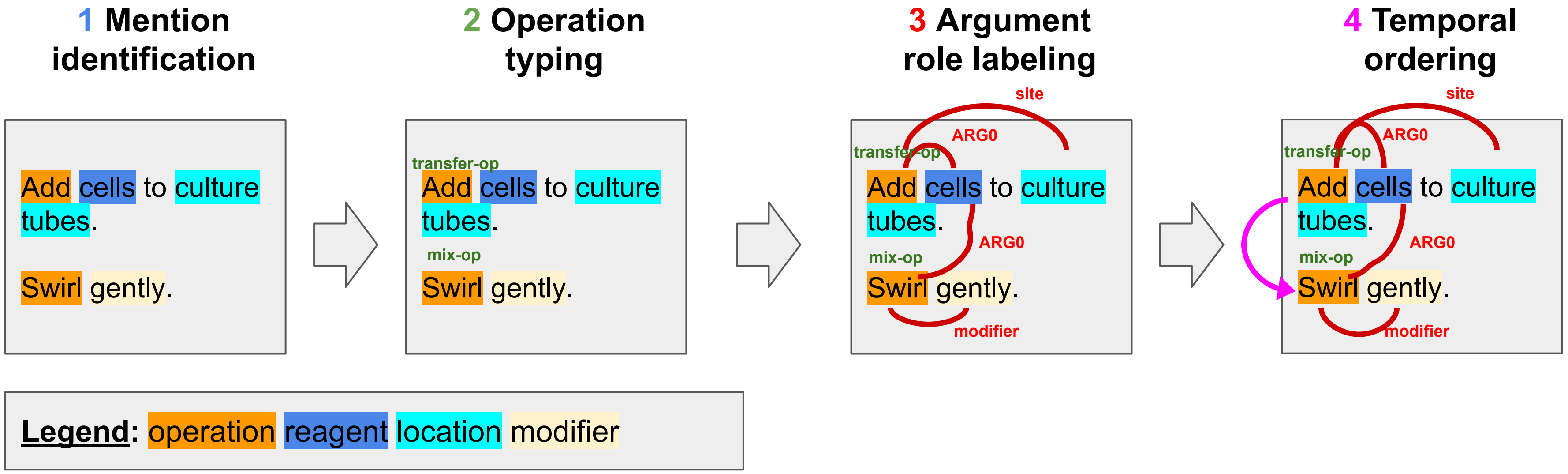
Modelling: Pipeline vs. Joint Learning

- **Pipeline model**
  - Breaks PEG prediction into subtasks
  - Predicts each separately

- **Multi-task**: jointly predicts entire PEG

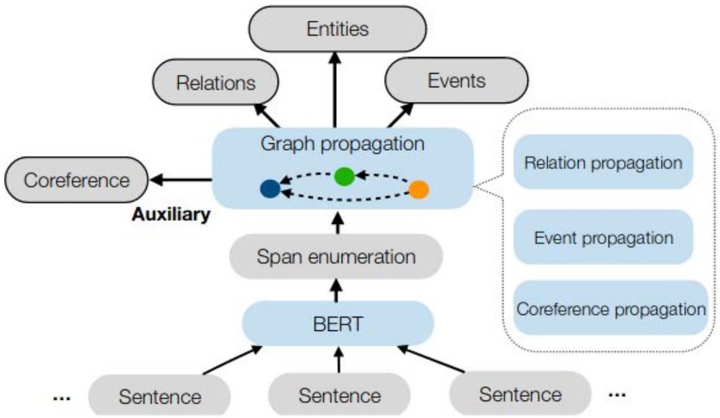
Modelling (1): Pipeline Approach

- Train model for each sub-task, chain together to obtain full PEG



Modelling (2): Multi-task Approach

- Adapted DyGIE++ for our protocols
  - Used sliding window as length exceeded SciBERT 512-token limit



DyGIE++ Framework (Wadden et. al, 2019)

Results

1. Mention Identification

Data Split	System	F_1
original	Kulkarni et al. (2018)	78.0
	Wadden et al. (2019)	**79.7**
	PIPELINE	78.3

2. Fine-grained operation typing

System	P	R	F_1
MULTI-TASK	75.6	68.9	72.1
PIPELINE	69.2	78.4	73.6
• w/ gold mentions	78.6	81.0	79.8

Results

3 + 4: Argument role labeling + temporal ordering (relation classification)

Task	MULTI-TASK	PIPELINE	# gold
Core			
• All roles	**59.2**	49.1	2839
• All roles (gold mentions)	-	70.8	2839
• ARG0	**62.0**	52.2	2313
• ARG1	**39.4**	28.9	412
• ARG2	**70.7**	57.4	114
Non-Core			
• All roles	**55.6**	44.6	4827
• All roles (gold mentions)	-	72.3	4827
• site	**60.5**	52.5	962
• setting	**77.4**	62.7	974
• usage	**35.0**	29.5	297
• co-ref	**41.2**	30.8	1014
• measure	**64.0**	52.6	804
• modifier	**50.0**	42.4	519
• located-at	**13.4**	10.5	179
• part-of	**8.5**	8.5	78
Temporal Ordering	**60.3**	49.0	1200
Temp. Ord. (gold mentions)	-	67.0	1200

Multi-task does better on all relation-classification tasks

Local relations easier to predict than cross sentence relations

Results: intra vs inter sentence relations

- For core-roles:

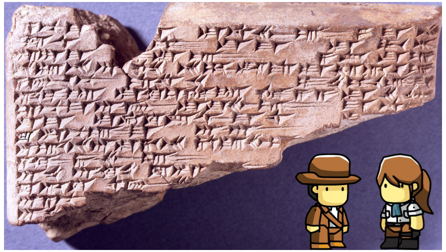
Split	MULTI-TASK	PIPELINE	# gold
Intra-sentence	**63.3**	55.6	2160
Inter-sentence	**42.1**	29.4	679

- For co-reference (92% are inter-sentence):

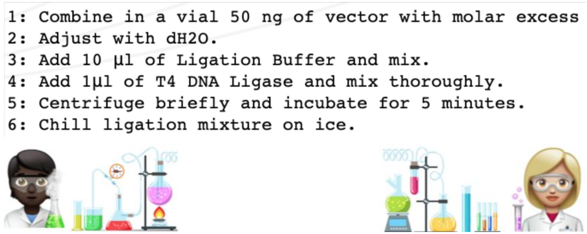
co-reference	**41.2**	30.8	1014

Cross-sentence relations are a key challenge for modelling!

Conclusion: NLP in the Wild



Filling gaps in cuneiform tablets



Understanding scientific protocols

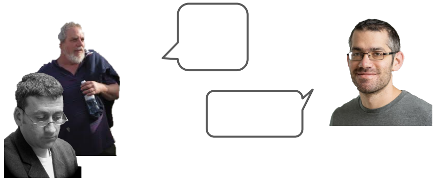
Conclusion: NLP in the Wild

Real-world texts

- Small amounts of data
- Long-range dependencies
- Specialized language

Interdisciplinary research questions

- Filling in the gaps in ancient texts
- Lenient & executable representations



Thanks!