English Understanding: From Annotations to AMRs

Nathan Schneider

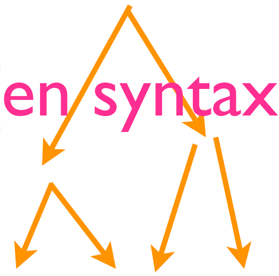
August 28, 2012 :: ISI NLP Group :: Summer Internship Project Presentation

Current state of the art: syntax-based MT

- Hierarchical/syntactic structures on source and/or target side
- Learn string-to-tree, tree-to-string, or tree-to-tree mappings for a language pair
- Syntax good for linguistic well-formedness

美国产妇产下12斤巨婴 选择不麻醉分娩

string-to-tree



[en syntax]

(read off yield of target tree)

U.S. maternal birth to 12 kg giant baby choose not to anesthesia delivery

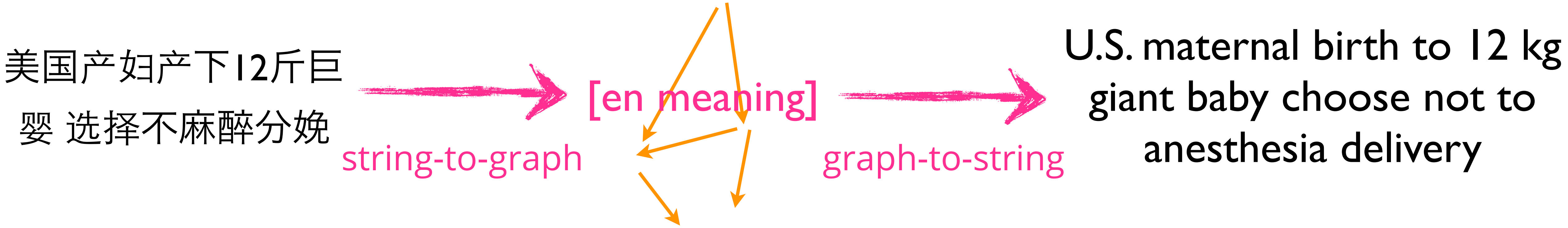
Why go deeper than syntax?

FRAGMENTATION

I lied to her.
She was lied to.
I told her a lie.
I told a lie to her.
She was told a lie.
A lie was told to her.
Lies were told to her by me.
What she was told was a lie.

CONFLATION

She lies all the time
...to her boss.
...on the couch.



美国产妇产下12斤巨婴 选择不麻醉分娩

string-to-graph

[en meaning]

graph-to-string

U.S. maternal birth to 12 kg giant baby choose not to anesthesia delivery

- How to get from the source sentence to target meaning, and from target meaning to target sentence?
  - graph transducer formalisms & rule extraction algorithms *(previous talk!)*
  - **designing English meaning representation & obtaining data**
  - English generation from meaning representation *(next talk!)*

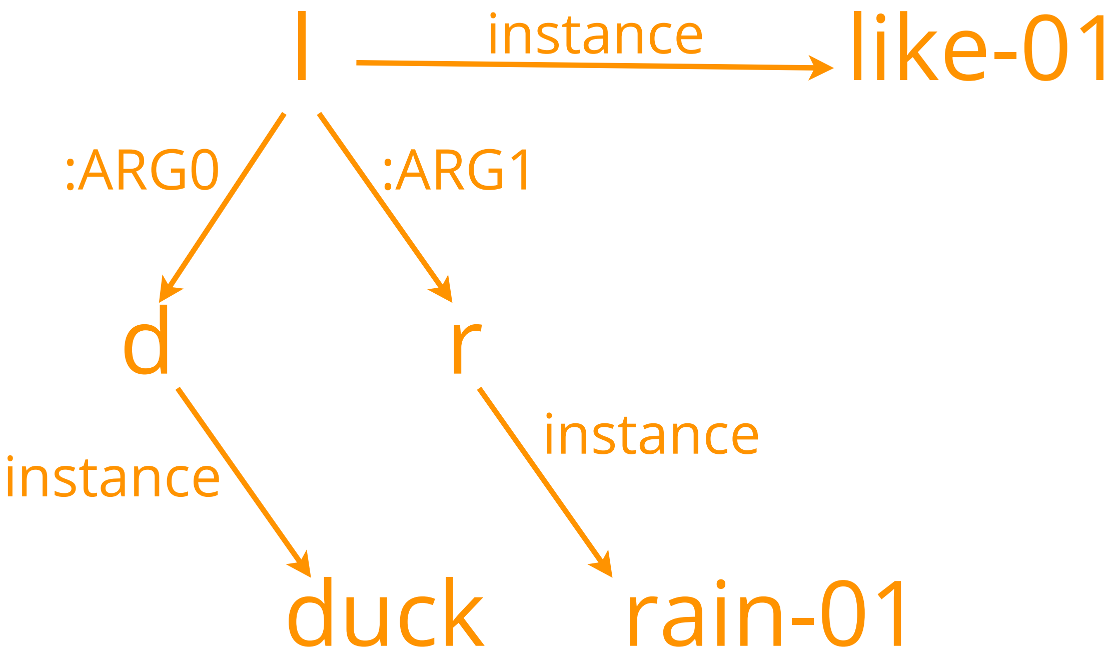
AMR Goals

- Meaning representation for English which is “more logical than syntax,” yet close enough to the surface form to support consistent annotation (*not* an interlingua)
    - Principally: **PropBank event structures with variables** (allowing entity and event coreference)
    - + special conventions for named entities, numeric and time expressions, modality, negation, questions, morphological simplification, etc.
    - in a unified graph structure

AMR Working Group

- ISI, U Colorado, LDC, SDL Language Weaver
- This summer: fine-tuning the AMR specification to the point where we can train annotators and expect decent inter-annotator agreement
  - Practice annotations, heated arguments!
  - Expanding to genres besides news

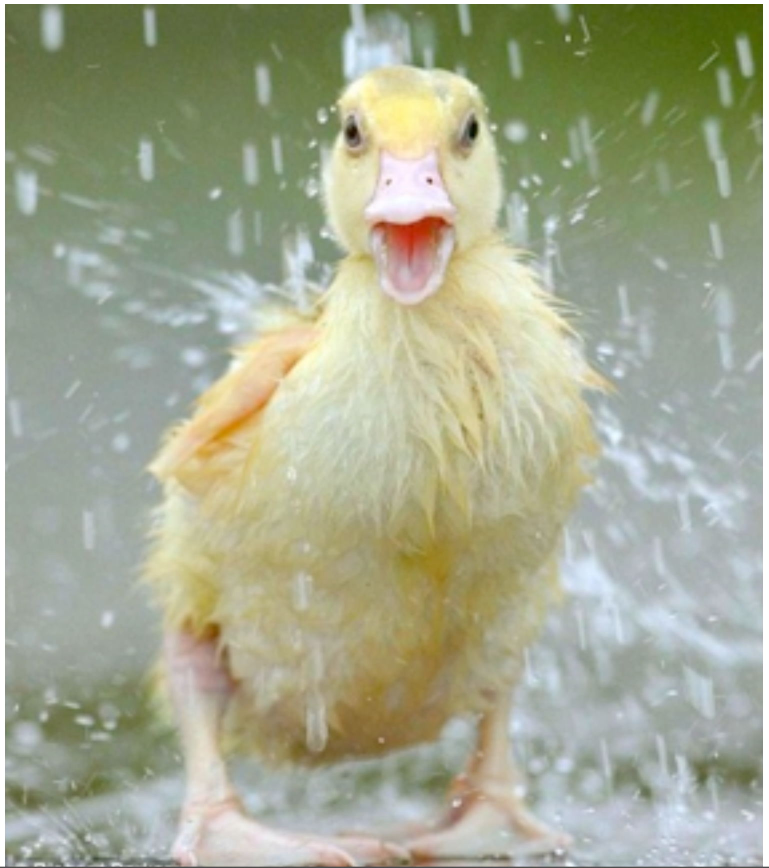
AMRs



l
instance
like-01
:ARG0
:ARG1
d
r
instance
instance
duck
rain-01

```
(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))
```

- ducks like rain
- the duck liked that it was raining

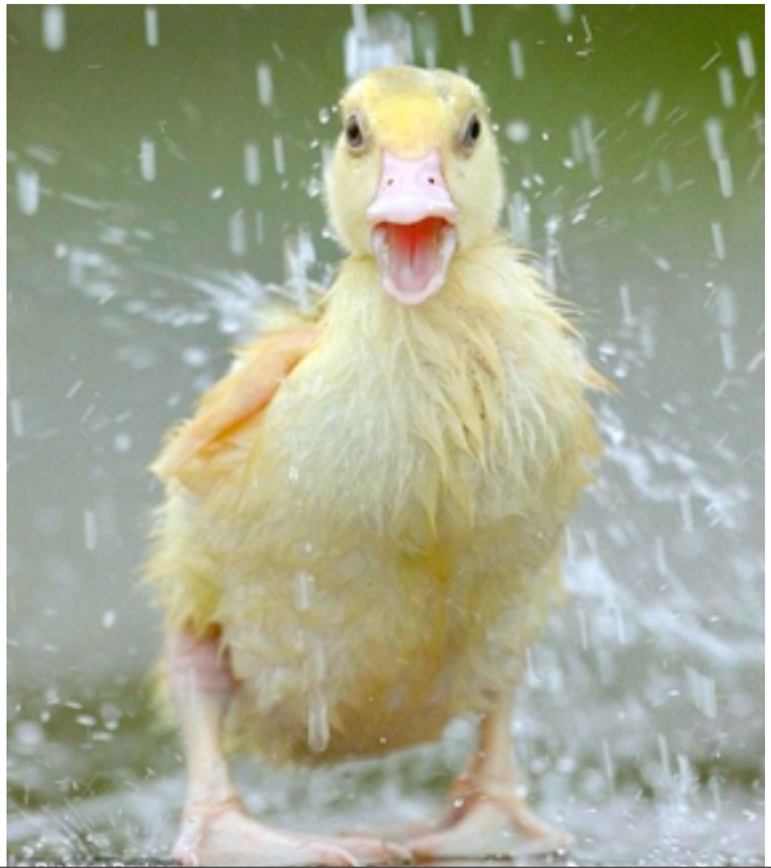


AMRs

```
(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))
```

```
(s2 / see-01
   :ARG0 (i / i)
   :ARG1 (d / duck
          :poss (s / she)))
```

- I saw her duck



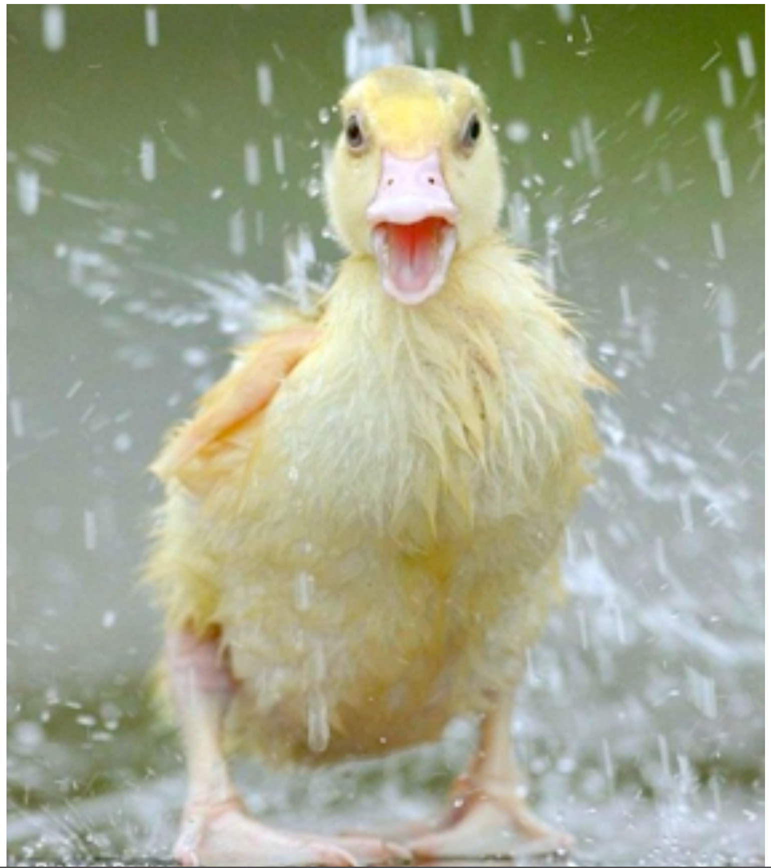
```
(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))
```

AMRs

```
(s2 / see-01
   :ARG0 (i / i)
   :ARG1 (d / duck
         :poss (s / she)))
```

```
(s2 / see-01
   :ARG0 (i / i)
   :ARG1 (d / duck-01
         :ARG0 (s / she)))
```

- I saw her duck [alternate interpretation]

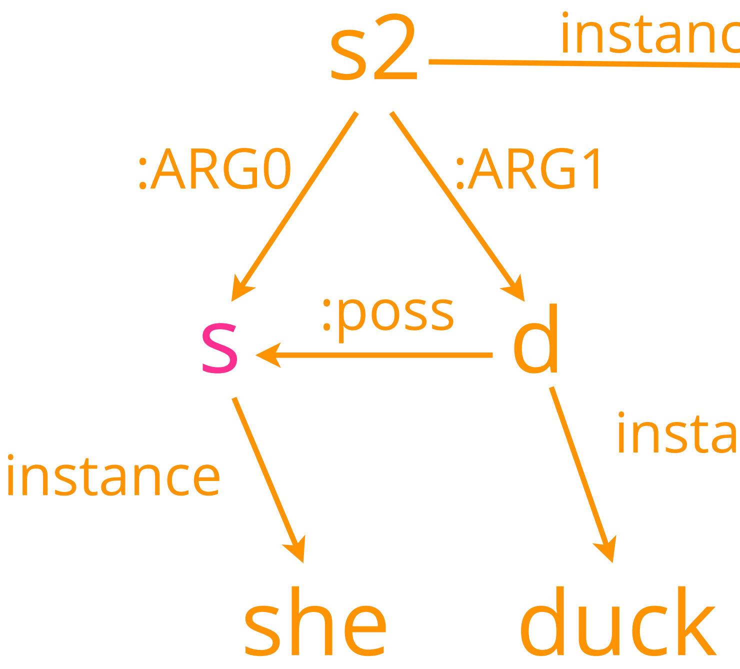


AMRs

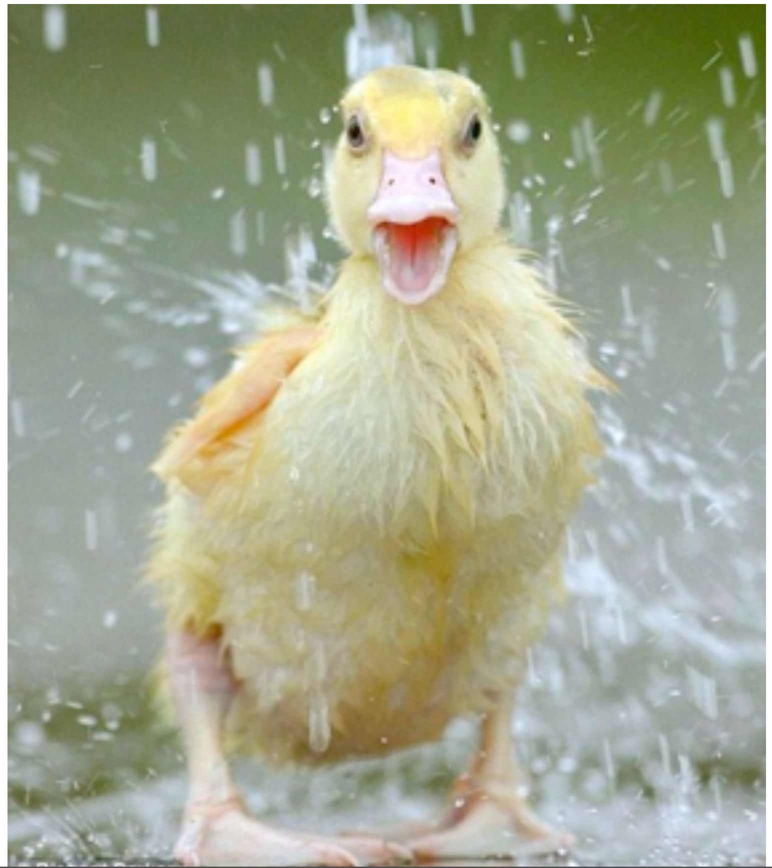
```
(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))
```

```
(s2 / see-01
   :ARG0 (i / i)
   :ARG1 (d / duck
          :poss (s / she)))
```

```
(s2 / see-01
   :ARG0 (s / she)
   :ARG1 (d / duck
          :poss s))
```



- She saw her (own) duck

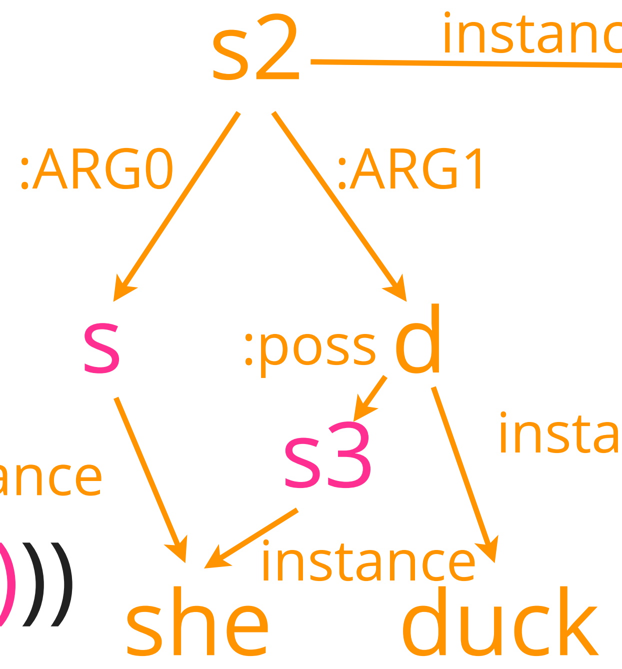


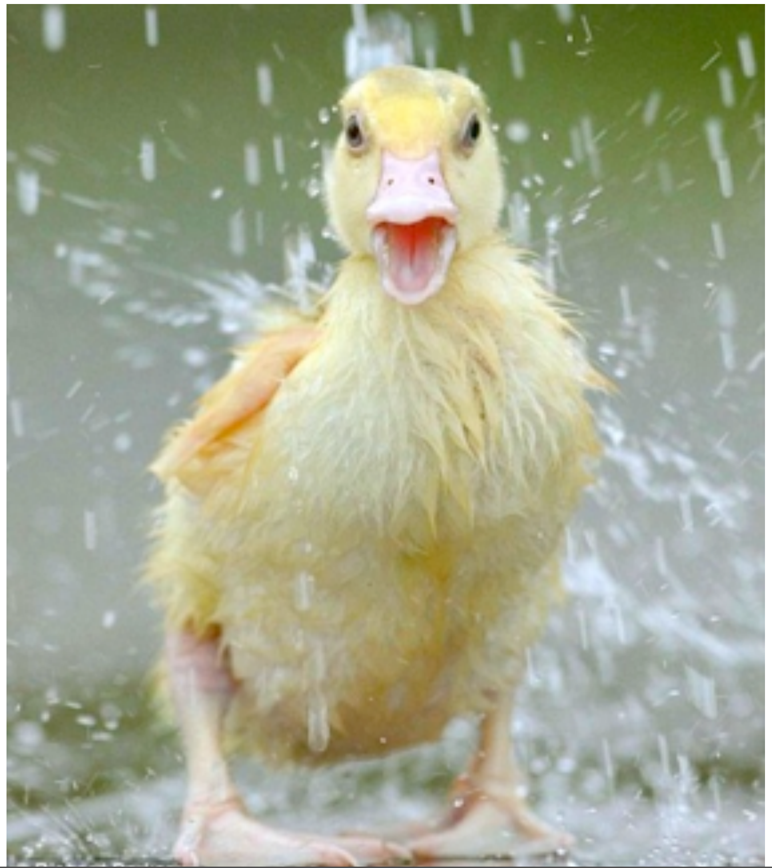
AMRs

```
(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))
```

```
(s2 / see-01
   :ARG0 (i / i)
   :ARG1 (d / duck
          :poss (s / she)))
```

```
(s2 / see-01
   :ARG0 (s / she)
   :ARG1 (d / duck
          :poss (s3 / she)))
```





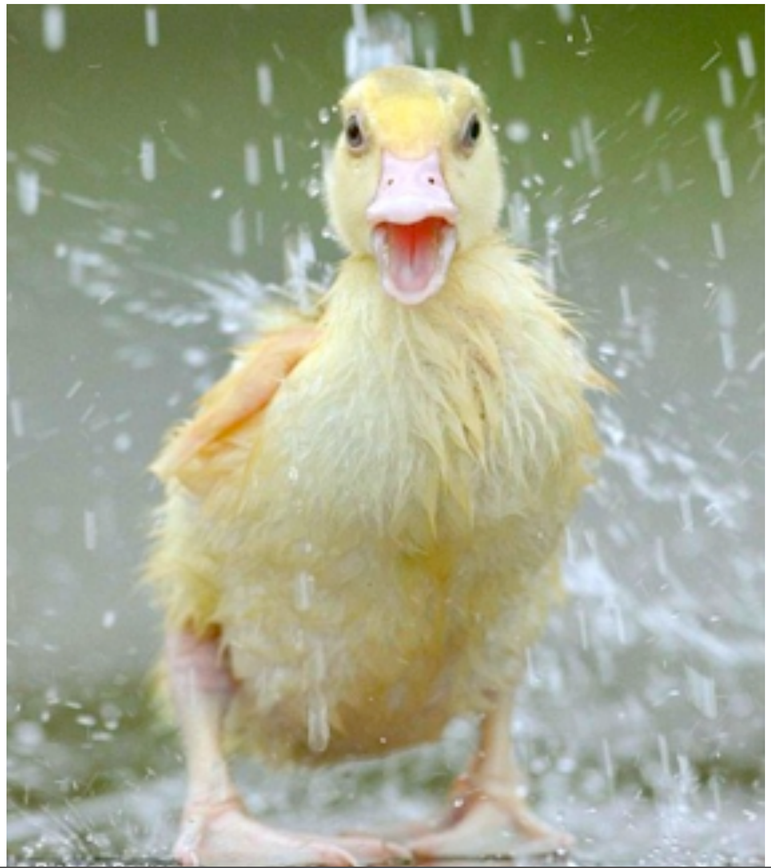
- She saw her (someone else's) duck

AMRs

```
(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))
```

```
(h / happy
   :domain (d / duck
              :ARG0-of (l / like-01
                          :ARG1 (r / rain-01))))
```

- Ducks who like rain are happy

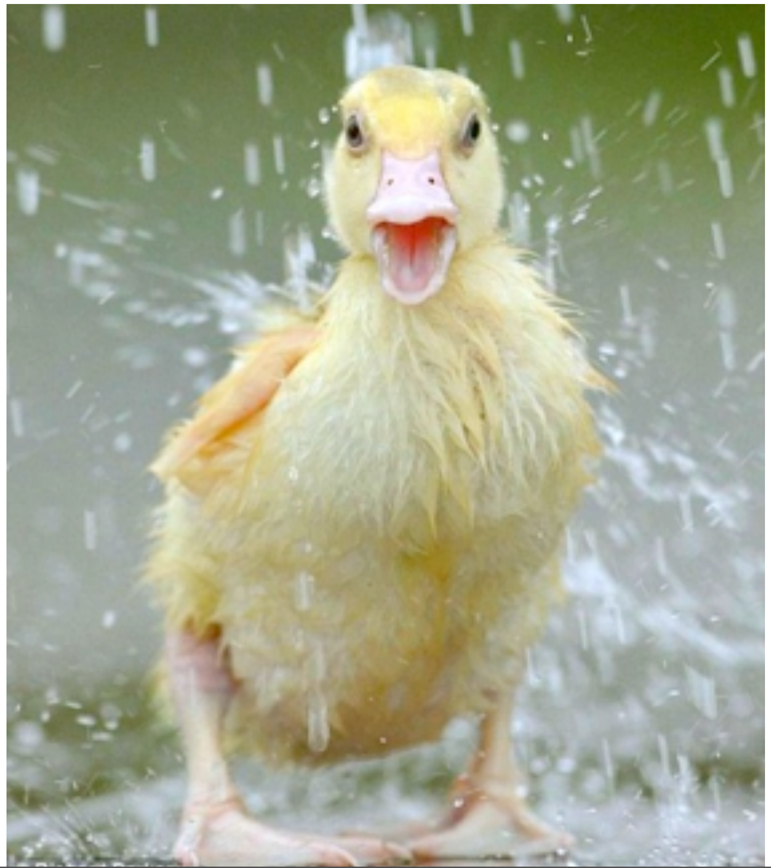


AMRs

```
(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))
```

```
(h / happy
   :domain (d / duck
             :ARG0-of (l / like-01
                        :ARG1 (r / rain-01))))
```

- Ducks who like rain are happy



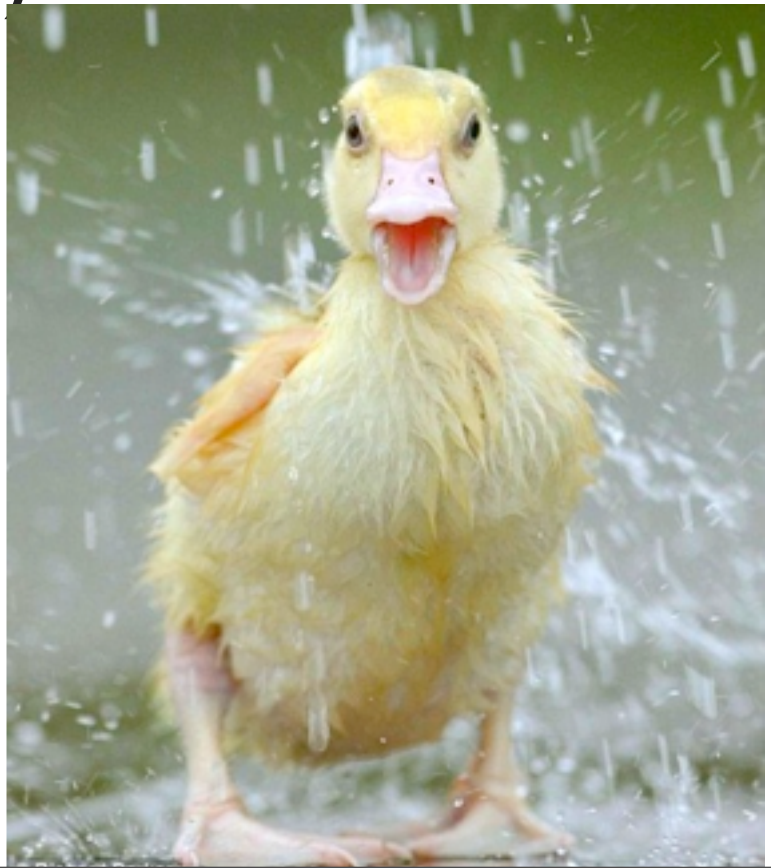
AMRs

```
(l / like-01
  :ARG0 (d / duck)
  :ARG1 (r / rain-01))
```

```
(h / happy
   :domain (d / duck
              :ARG0-of (l / like-01
                          :ARG1 (r / rain-01))))
```

```
(l / like-01
   :ARG0 (d / duck
           :domain-of/:mod (h / happy))
   :ARG1 (r / rain-01))
```

- Happy ducks like rain



Getting the AMRs we want

- **Ideal goal:** Learn a string-to-graph transducer using parallel data with Chinese string and gold-standard AMRs

Getting the AMRs we want

- **Ideal goal:** Learn a string-to-graph transducer using parallel data with Chinese string and ~~gold-standard AMRs~~
  predictions of an English semantic analyzer that was trained on gold standard AMRs

Getting the AMRs we want

- **Ideal goal:** Learn a string-to-graph transducer using parallel data with Chinese string and ~~gold-standard AMRs~~
  predictions of an English semantic analyzer that was ~~trained on gold standard AMRs~~
  hand-coded (rule-based)
- **Intermediate goal:** Build a rule-based English semantic analyzer for data that already has some gold-standard semantic representations
- **Next:** Fully automate so an AMR can be generated for any sentence (with existing tools and/or bootstrapping off of gold-standard annotations)

Combining Representations

```
(TOP
  (S
    (NP-SBJ
      (NP (NNP Pierre) (NNP Vinken))
      (, ,)
      (ADJP (NML (CD 61) (NNS years)) (JJ old))
      (, ,))
    (VP
      (MD will)
      (VP
        (VB join)
        (NP (DT the) (NN board))
        (PP-CLR (IN as) (NP (DT a) (JJ nonexecutive)
                                       (NN director)))

        (NP-TMP (NNP Nov.) (CD 29))))
    (. .)))
```

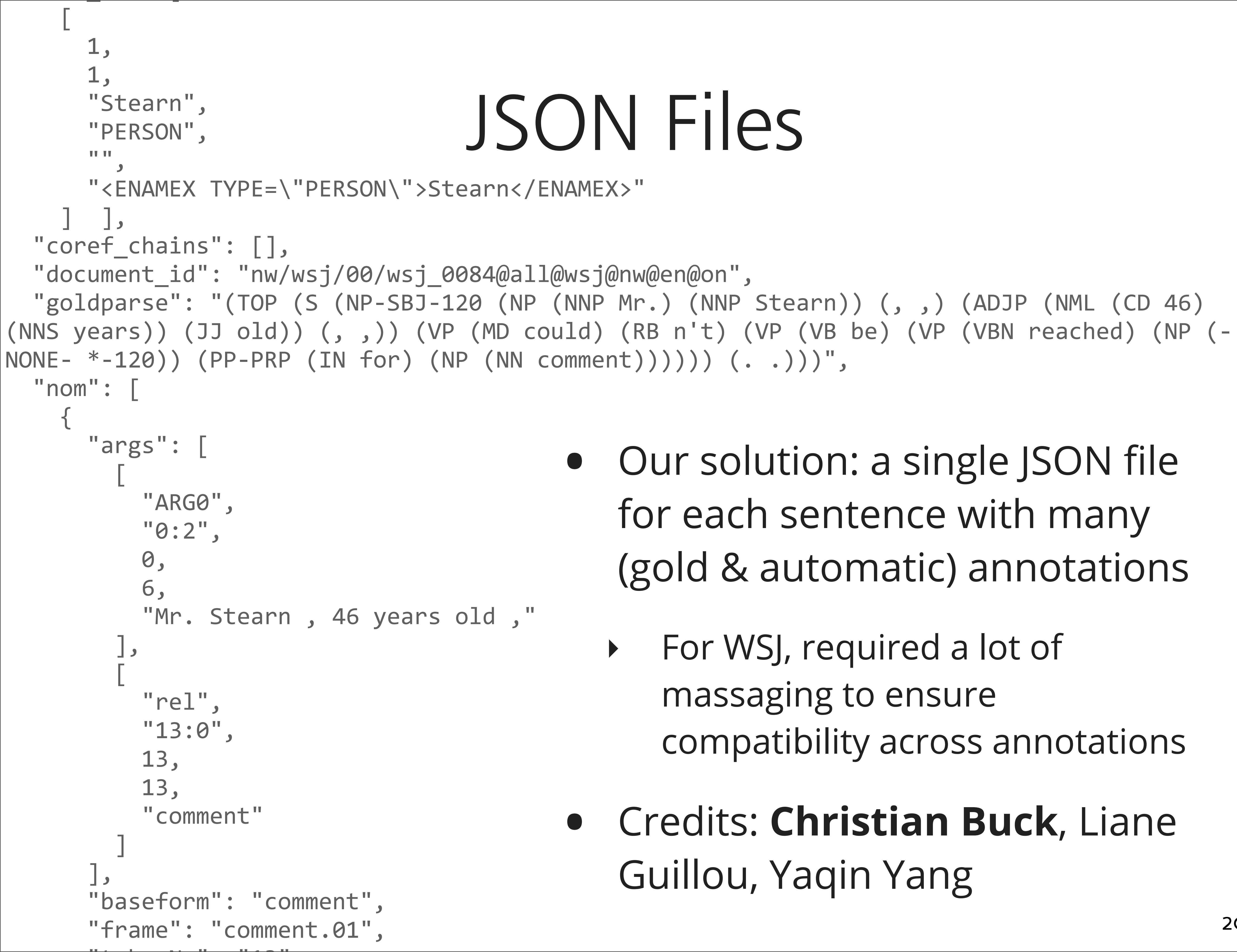
```
nn(Vinken-2, Pierre-1)
nsubj(join-9, Vinken-2)
num(years-5, 61-4)
dep(old-6, years-5)
amod(Vinken-2, old-6)
aux(join-9, will-8)
root(ROOT-0, join-9)
det(board-11, the-10)
dobj(join-9, board-11)
det(director-15, a-13)
amod(director-15, nonexecutive-14)
prep_as(join-9, director-15)
tmod(join-9, Nov.-16)
num(Nov.-16, 29-17)
```

```
nw/wsj/00/wsj_0001@0001@wsj@nw@en@on 0 8 gold join-v join.01 ----- 8:0-rel 0:2-ARG0 7:0-ARGM-MOD 9:1-ARG1 11:1-ARGM-PRD 15:1-ARGM-TMP
nw/wsj/00/wsj_0001@0001@wsj@nw@en@on 1 10 gold publish-v publish.01 ----- 10:0-rel 11:0-ARG0
```

```
<DOCNO> WSJ0001 </DOCNO>
     <ENAMEX TYPE="PERSON">Pierre Vinken</ENAMEX> , <TIMEX TYPE="DATE:AGE">61 years old</TIMEX> , will join the <ENAMEX TYPE="ORG_DESC:OTHER">board</ENAMEX> as a nonexecutive <ENAMEX TYPE="PER_DESC">director</ENAMEX> <TIMEX TYPE="DATE:DATE">Nov. 29</TIMEX> .
```

- In practice, working with the many different file formats and representational details is very tedious

JSON Files



```
    [
      1,
      1,
      "Stearn",
      "PERSON",
      "",
      "<ENAMEX TYPE=\"PERSON\">Stearn</ENAMEX>"
    ]  ],
  "coref_chains": [],
  "document_id": "nw/wsj/00/wsj_0084@all@wsj@nw@en@on",
  "goldparse": "(TOP (S (NP-SBJ-120 (NP (NNP Mr.) (NNP Stearn)) (, ,) (ADJP (NML (CD 46)
(NNS years)) (JJ old)) (, ,)) (VP (MD could) (RB n't) (VP (VB be) (VP (VBN reached) (NP (-
NONE- *-120)) (PP-PRP (IN for) (NP (NN comment)))))) (. .)))",
  "nom": [
    {
      "args": [
        [
          "ARG0",
          "0:2",
          0,
          6,
          "Mr. Stearn , 46 years old ,"
        ],
        [
          "rel",
          "13:0",
          13,
          13,
          "comment"
        ]
      ],
      "baseform": "comment",
      "frame": "comment.01",
```

- Our solution: a single JSON file for each sentence with many (gold & automatic) annotations
  - For WSJ, required a lot of massaging to ensure compatibility across annotations
- Credits: **Christian Buck**, Liane Guillou, Yaqin Yang

AMR Generation

- Rule-based integration of OntoNotes annotations
  (+ some output of existing tools)
- The sentence below will illustrate the pipeline and the kinds of annotations it exploits
  - The AMR is built up incrementally as each new piece of annotation is considered
  - This is the actual system behavior
    ...albeit on a short and easy example!

- Mr. Stearn, 46 years old, couldn't be reached for comment.

nes: BBN Corpus

- BBN Pronoun Coreference & Entity Type Corpus: fine-grained named entity labels and anaphoric coreference for WSJ
  - Entity categories include refinements of the standard PERSON/ORG/LOCATION (e.g. LOCATION:CITY) as well as other categories (LAW, CHEMICAL, DISEASE, ...)
  - BBN IdentiFinder tagger

```
(0 / person-FALLBACK
 :name (1 / name
        :op1 "Stearn"))
```

PERSON

- Mr. Stearn, 46 years old, couldn't be reached for comment.

timex: Stanford sutime

- TIMEX3 is a markup format for time expressions (*last Tuesday*, *several years from now*, *7:00 pm*, *Tuesday, Aug. 28*)
  - Stanford sutime tagger produces XML, e.g.: `<TIMEX3 tid="t1" value="P46Y" type="DURATION">46 years old</TIMEX3>`
  - We implemented rules to handle different kinds of normalized time expressions

```
(0 / person-FALLBACK
 :name (1 / name
        :op1 "Stearn"))
(2 / temporal-quantity-AGE
 :quant 46
 :unit (3 / year) )
```

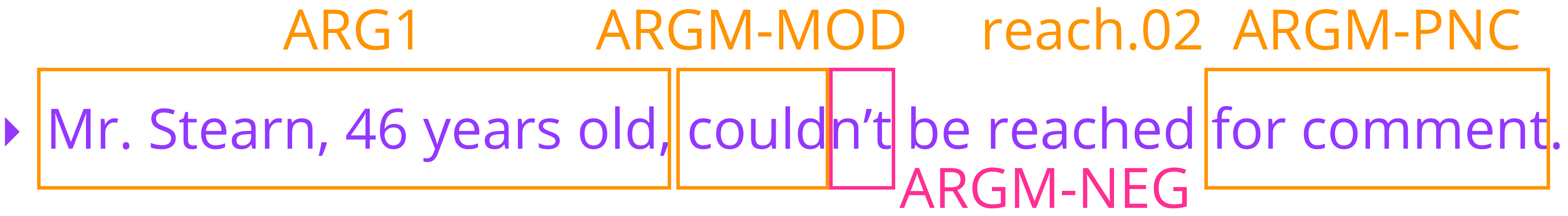
DURATION:P46Y

- Mr. Stearn, 46 years old, couldn't be reached for comment.

vprop: PropBank (verbs)

- PropBank annotations from OntoNotes provide the main skeleton of the sentence
  - AMR has a somewhat different set of non-core roles; here **:ARGM-PNC** ought to be replaced with **:purpose**
  - Note that the **:ARG1** is a fragment from a previous module. Done with variable-to-token alignments and head finding for phrases.

```
(2 / temporal-quantity-AGE
 :quant 46
 :unit (3 / year) )
(4 / reach-02
 :ARG1 (0 / person-FALLBACK
        :name (1 / name
               :op1 "Stearn"))
 :ARGM-PNC (5 / comment)
 :polarity -)
```



- Mr. Stearn, 46 years old, couldn't be reached for comment.

nprop: NomBank

(argument-taking nouns)

- NomBank annotations not included in OntoNotes but available for all of WSJ
  - AMR does not use NomBank predicates directly, but they are inserted as an intermediate step
  - Because the token *Stearn* is already associated with a variable, the **:ARG0** of **comment-n-01** is reentrant

```
(2 / temporal-quantity-AGE
 :quant 46
 :unit (3 / year) )
(4 / reach-02
 :ARG1 (0 / person-FALLBACK
        :name (1 / name
               :op1 "Stearn"))
 :ARGM-PNC (5 / comment
        :-PRED (6 / comment-n-01
               :ARG0 0)
 :polarity -)
```

ARG0

comment.01

- Mr. Stearn, 46 years old, couldn't be reached for comment.

verbalize: NomBank nouns to PropBank verbs

- AMR uses only verbal predicates, so mappings in the NomBank lexicon are used to convert nouns to verbs where possible
  - Here, we know **comment.n.01** corresponds to **comment.v.01**
  - Some nouns refer to a verb's argument: a *filter* in AMR essentially becomes a *thing that filters*
  - Deciding when to convert a noun to a verb is often tricky, even for humans!

```
(2 / temporal-quantity-AGE
 :quant 46
 :unit (3 / year) )
(4 / reach-02
 :ARG1 (0 / person-FALLBACK
        :name (1 / name
               :op1 "Stearn"))
 :ARGM-PNC (5 / comment-01
        :-COREF (6 / comment-01
               :ARG0 0)
 :polarity -)
```

- Mr. Stearn, 46 years old, couldn't be reached for comment.

conjunctions

- Identify coordinate structures based on the dependency parse (Stanford dependency converter). No coordination in this sentence.

copulas

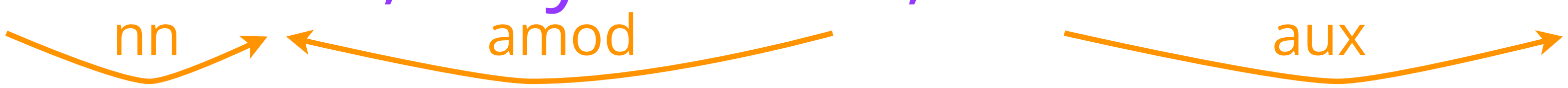
- Predicate nominals/adjectives and nominal appositives. None in this sentence.

- Mr. Stearn, 46 years old, couldn't be reached for comment.

- adjsAndAdverbs: Modifiers: adjectives, adverbs, quantities
  - Special detection of __ *years old* as an **:age**; attaches the time expression to the person concept
  - The AMR is now connected
- auxes: Maps modal auxiliaries to modal concepts (here, uncertainty about the meaning of ***could***)
- misc: Noun-noun modifiers and remaining prepositional phrases

```
(7 / possible-or-permit-01
 :domain (4 / reach-02
        :ARG1 (0 / person-FALLBACK
               :age (2 / temporal-quantity
                      :quant 46
                      :unit (3 / year) )
               :mod-NN (8 / mr)
               :name (1 / name
                      :op1 "Stearn"))
        :ARGM-PNC (5 / comment-01
               :-COREF (6 / comment-01
                      :ARG0 0))
        :polarity -))
```

- Mr. Stearn, 46 years old, couldn’t be reached for comment.



coref

- Coreferent expressions (typically, pronouns and their antecedents) are marked. N/A here.

top

- Heuristically designates a main concept based on the dependency parse (here, incorrectly).

```
(7 / possible-or-permit-01
 :domain (4 / reach-02-ROOT
        ...))
```

beautify

- Produces the final version of the AMR for human eyes…

‣ Mr. Stearn, 46 years old, couldn't be reached for comment.

Generated AMR

```
(r / reach-02
 :ARG1 (p / person
        :age (t / temporal-quantity
               :quant 46
               :unit (y / year) )
        :mod (m / mr)
        :name (n / name
               :op1 "Stearn"))
 :ARGM-PNC (c / comment-01
        :ARG0 p)
 :domain-of (p1 / possible-or-permit-01)
 :polarity -)
```

- Mr. Stearn, 46 years old, couldn't be reached for comment.

Generated AMR: Flaws

```
(r / reach-02
 :ARG1 (p / person
        :age (t / temporal-quantity
               :quant 46
               :unit (y / year) )
        :mod (m / mr)
        :name (n / name
               :op1 "Stearn"))
 :ARGM-PNC (c / comment-01
        :ARG0 p)
 :domain-of (p1 / possible-or-permit-01)
 :polarity -)
```

- Mr. Stearn, 46 years old, couldn’t be reached for comment.

AMR Generation

- 13 modules, each addressing some part of the meaning by consulting annotations and updating the working AMR
    - Ulf has built a similar pipeline; ours uses more preexisting semantic representations (e.g. NomBank), Ulf’s is more fine-tuned and relies more heavily on lexical lists and specialized rules
- The system produces something reasonable for a cherry-picked example. But overall?
    - Do we gain anything from NomBank?

Effect of NomBank

- Shu Cai's smatch metric applied to compare 73 generated vs. gold-standard AMRs
  - Precision, recall, F_1 of graph edges under best matching of nodes
  - Daniel Bauer's implementation
- Baseline: Pipeline − NomBank (no predicates for *comment*, *filter*, *president*)
- Full NomBank with verbalization: comment-01, (thing :ARG0-of filter-01), president-n-01
- Only NomBank predicates that are verbalized: comment-01, (thing :ARG0-of filter-01), president

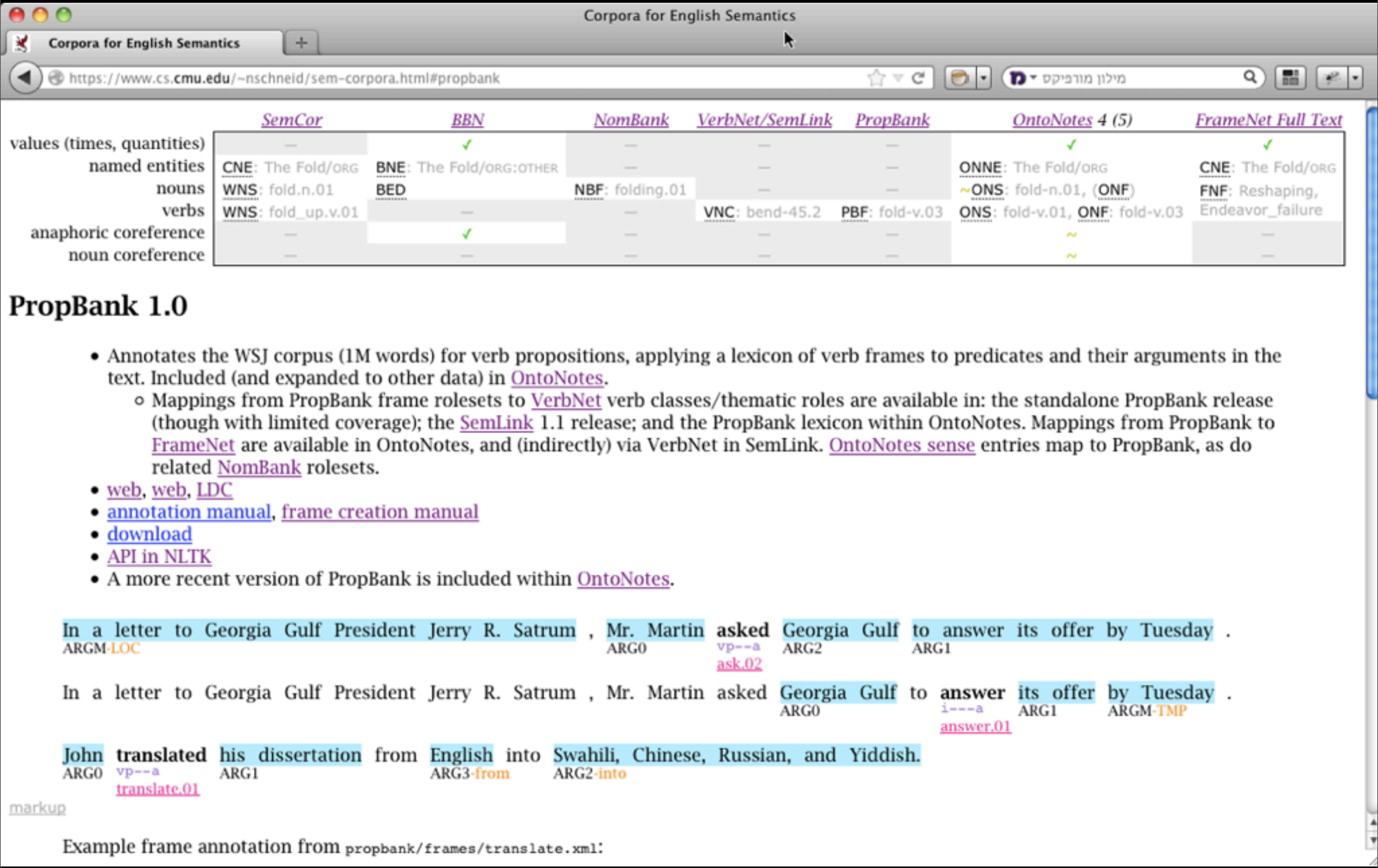
P	R	F_1
58	57	57
57	53	55
60	**58**	**59**

Taking stock

- We get decent AMRs given gold annotations, but there is room to grow
  - Lots of obvious tweaks that can be made
  - Interesting NLP subproblems: prepositions/relations between nominals, modality/negation, etc.
  - Complementary techniques from Ulf's approach
- Automating the process so we can get AMRs for non-OntoNotes parallel data
- End-to-end MT!

Contributions

- Understanding of English semantic annotation schemes, corpora, and tools


Corpora for English Semantics
Corpora for English Semantics
https://www.cs.cmu.edu/~nschneid/sem-corpora.html#propbank
מילון מורפיקס

	SemCor	BBN	NomBank	VerbNet/SemLink	PropBank	OntoNotes 4 (5)	FrameNet Full Text
values (times, quantities)	—	✓	—	—	—	✓	✓
named entities	CNE: The Fold/ORG	BNE: The Fold/ORG:OTHER	—	—	—	ONNE: The Fold/ORG	CNE: The Fold/ORG
nouns	WNS: fold.n.01	BED	NBF: folding.01	—	—	~ONS: fold-n.01, (ONF)	FNF: Reshaping, Endeavor_failure
verbs	WNS: fold_up.v.01	—	—	VNC: bend-45.2	PBF: fold-v.03	ONS: fold-v.01, ONF: fold-v.03	
anaphoric coreference	—	✓	—	—	—	~	—
noun coreference	—	—	—	—	—	~	—

PropBank 1.0

- Annotates the WSJ corpus (1M words) for verb propositions, applying a lexicon of verb frames to predicates and their arguments in the text. Included (and expanded to other data) in OntoNotes.
    - Mappings from PropBank frame rolesets to VerbNet verb classes/thematic roles are available in: the standalone PropBank release (though with limited coverage); the SemLink 1.1 release; and the PropBank lexicon within OntoNotes. Mappings from PropBank to FrameNet are available in OntoNotes, and (indirectly) via VerbNet in SemLink. OntoNotes sense entries map to PropBank, as do related NomBank rolesets.
- web, web, LDC
- annotation manual, frame creation manual
- download
- API in NLTK
- A more recent version of PropBank is included within OntoNotes.

In a letter to Georgia Gulf President Jerry R. Satrum	,	Mr. Martin	**asked**	Georgia Gulf	to answer its offer by Tuesday	.
ARGM-LOC		ARG0	vp--a ask.02	ARG2	ARG1	

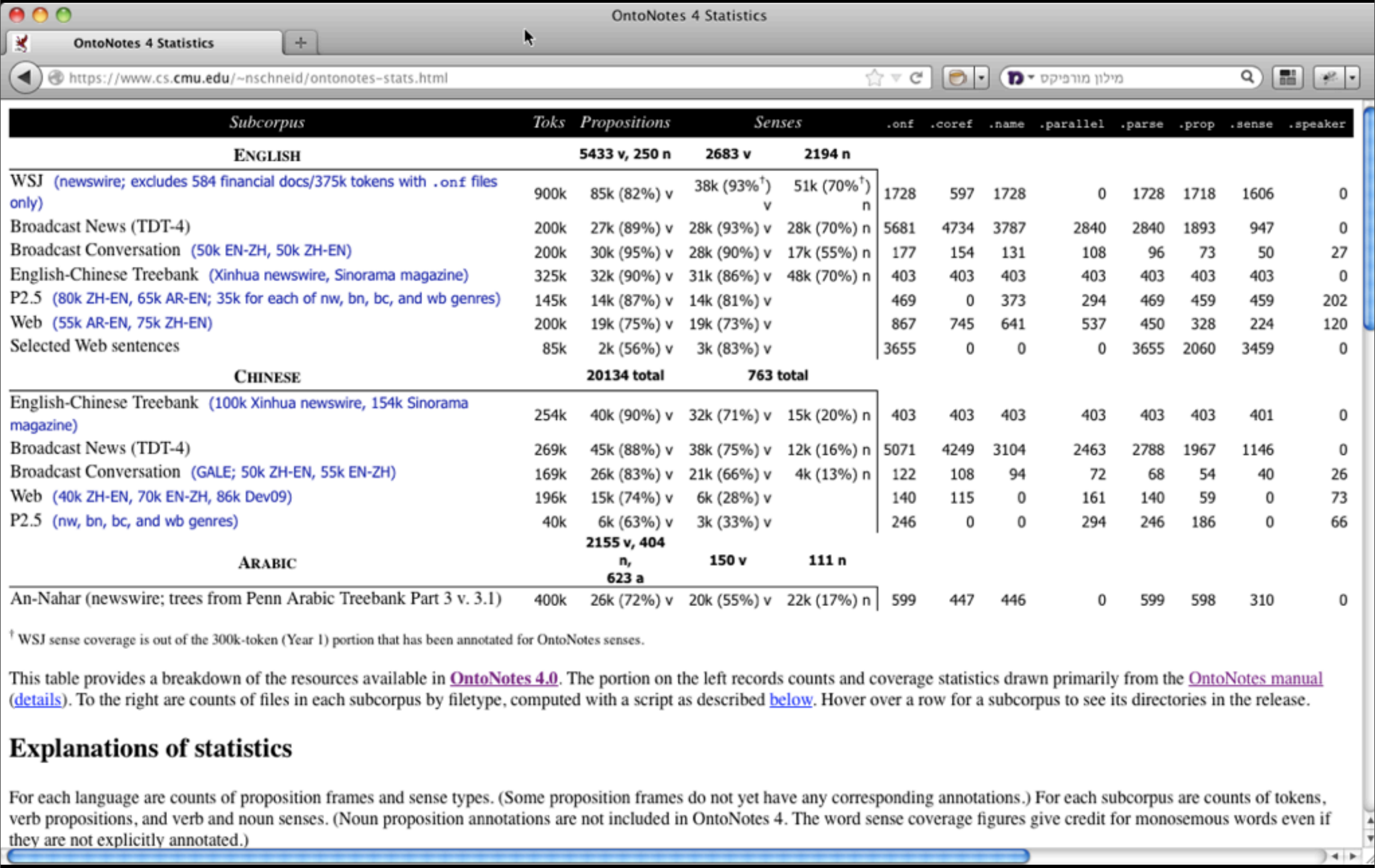
In a letter to Georgia Gulf President Jerry R. Satrum , Mr. Martin asked	Georgia Gulf	to	**answer**	its offer	by Tuesday	.
	ARG0		i---a answer.01	ARG1	ARGM-TMP	

John	**translated**	his dissertation	from	English	into	Swahili, Chinese, Russian, and Yiddish.
ARG0	vp--a translate.01	ARG1		ARG3-from		ARG2-into

markup

Example frame annotation from `propbank/frames/translate.xml`:

http://tinyurl.com/semcorpora



Subcorpus	*Toks*	*Propositions*	*Senses*		.onf	.coref	.name	.parallel	.parse	.prop	.sense	.speaker
ENGLISH		**5433 v, 250 n**	**2683 v**	**2194 n**								
WSJ (newswire; excludes 584 financial docs/375k tokens with .onf files only)	900k	85k (82%) v	38k (93%[†]) v	51k (70%[†]) n	1728	597	1728	0	1728	1718	1606	0
Broadcast News (TDT-4)	200k	27k (89%) v	28k (93%) v	28k (70%) n	5681	4734	3787	2840	2840	1893	947	0
Broadcast Conversation (50k EN-ZH, 50k ZH-EN)	200k	30k (95%) v	28k (90%) v	17k (55%) n	177	154	131	108	96	73	50	27
English-Chinese Treebank (Xinhua newswire, Sinorama magazine)	325k	32k (90%) v	31k (86%) v	48k (70%) n	403	403	403	403	403	403	403	0
P2.5 (80k ZH-EN, 65k AR-EN; 35k for each of nw, bn, bc, and wb genres)	145k	14k (87%) v	14k (81%) v		469	0	373	294	469	459	459	202
Web (55k AR-EN, 75k ZH-EN)	200k	19k (75%) v	19k (73%) v		867	745	641	537	450	328	224	120
Selected Web sentences	85k	2k (56%) v	3k (83%) v		3655	0	0	0	3655	2060	3459	0
CHINESE		**20134 total**	**763 total**									
English-Chinese Treebank (100k Xinhua newswire, 154k Sinorama magazine)	254k	40k (90%) v	32k (71%) v	15k (20%) n	403	403	403	403	403	403	401	0
Broadcast News (TDT-4)	269k	45k (88%) v	38k (75%) v	12k (16%) n	5071	4249	3104	2463	2788	1967	1146	0
Broadcast Conversation (GALE; 50k ZH-EN, 55k EN-ZH)	169k	26k (83%) v	21k (66%) v	4k (13%) n	122	108	94	72	68	54	40	26
Web (40k ZH-EN, 70k EN-ZH, 86k Dev09)	196k	15k (74%) v	6k (28%) v		140	115	0	161	140	59	0	73
P2.5 (nw, bn, bc, and wb genres)	40k	6k (63%) v	3k (33%) v		246	0	0	294	246	186	0	66
ARABIC		**2155 v, 404 n, 623 a**	**150 v**	**111 n**								
An-Nahar (newswire; trees from Penn Arabic Treebank Part 3 v. 3.1)	400k	26k (72%) v	20k (55%) v	22k (17%) n	599	447	446	0	599	598	310	0

[†] WSJ sense coverage is out of the 300k-token (Year 1) portion that has been annotated for OntoNotes senses.

This table provides a breakdown of the resources available in OntoNotes 4.0. The portion on the left records counts and coverage statistics drawn primarily from the OntoNotes manual (details). To the right are counts of files in each subcorpus by filetype, computed with a script as described below. Hover over a row for a subcorpus to see its directories in the release.

Explanations of statistics

For each language are counts of proposition frames and sense types. (Some proposition frames do not yet have any corresponding annotations.) For each subcorpus are counts of tokens, verb propositions, and verb and noun senses. (Noun proposition annotations are not included in OntoNotes 4. The word sense coverage figures give credit for monosemous words even if they are not explicitly annotated.)

http://tinyurl.com/on4stats

Contributions

- Understanding of English semantic annotation schemes, corpora, and tools
- A tool for integrating several kinds of English annotations (OntoNotes, NomBank, automatic) into a single JSON file (with compatible indexing!)
- Manual AMR annotations & improvements to the AMR specification
- A prototype AMR generator that is highly modular and leverages many existing representations
- Understanding of the major challenges that remain for automatic AMR generation

Thanks & Questions?