The Linguistic Core: Linguistics for Machine Translation and Textual Analysis

Presented by

Lori Levin

Outline

- Why do we need more linguistics?
- What kind of linguistic knowledge is useful
- How to represent linguistic knowledge so that TA and MT can use it.
- Where will linguistic knowledge come from?

Errors that persist in mature MT systems

Input:	*wo3*	*qian2*	*bei4*	*ta1*	*men*	*tou1*	*le*
Gloss:	I	money	<bei4>	they	<plural>	steal	<completed>
Correct translation:	My money was stolen by them.						
Automatic translation:	I stole the money by them. (incorrect)						

Input:	*wo3*	*qian2*	*bei4*	*tou1*	*le*
Gloss:	I	money	<bei4>	steal	<completed>
Correct translation:	My money was stolen.				
Automatic translation:	I have the money stolen. (incorrect)				

There is an easily identifiable pattern that is not explicitly modeled by mature MT systems.

Missing the pattern is not adequately penalized by automatic metrics for MT evaluation.

Errors that persist in Mature MT systems

- E: This is the company that __ bought the bank.
- J: これは、**銀行を**買った会社です
- E: This is the company bought the banks

- E: This is the company that the bank bought __.
- J: これは、**銀行が**購入した会社です
- E: This is the company that bought the bank

- Google Translate: 10-26-10

- Steedman (2008) showed that Google Translate did not model filler-gap constructions in Arabic.
- In the intermediate Japanese translations, the grammatical role of 銀行 (*the bank)* is clearly indicated by a case marker.

Errors that persist in mature MT systems (Definiteness in Chinese-English MT)

MT Output

Ref	The	a	NONE
The	18%	1%	9%
a	1%	3%	1%
NONE	11%	1%	55%

% of total noun phrases

MT output disagrees with ref for 24% of noun phrases
Results similar for phrase-based and syntax-based
Multiple reference translations agree for 95% of noun phrases

Errors that persist in mature MT systems

English source	Correct Swahili translation	Automatic translation
I am reading a book.	*Ninasoma kitabu.* I-PRESENT-read book	I am kusoma kitabu.
You are reading a book.	*Unasoma kitabu.* you-PRESENT-read book	Wewe ni kusoma kitabu.
He is reading a book.	*Anasoma kitabu.* he-PRESENT-read book	Yeye ni kusoma kitabu.

What is the problem?

- Ignoring readily available knowledge
  - universals
  - linguists and reference grammars
  - other languages
- Syntax is not enough

Syntax is not enough

Example: English rule NP of NP

English	Hmong	Hmong Rule
some of the students	cov tub-kawm-ntawv ib txha	CLF N ib CLF
a book of mine	kuv ib phau ntawv	NP ib CLF NP
A house of bamboo	ib lub tsev-xyoob	ib CLF N N
the top of the tree	tsob ntoo saab sau	CLF NP CLF NP
the mother of that student	tug tub-kawm-ntawv hov leej nam	CLF NP CLF NP
a bottle of liquor	ib fwj cawv	ib CLF NP

(Hmong examples from David Mortensen)

Now what?

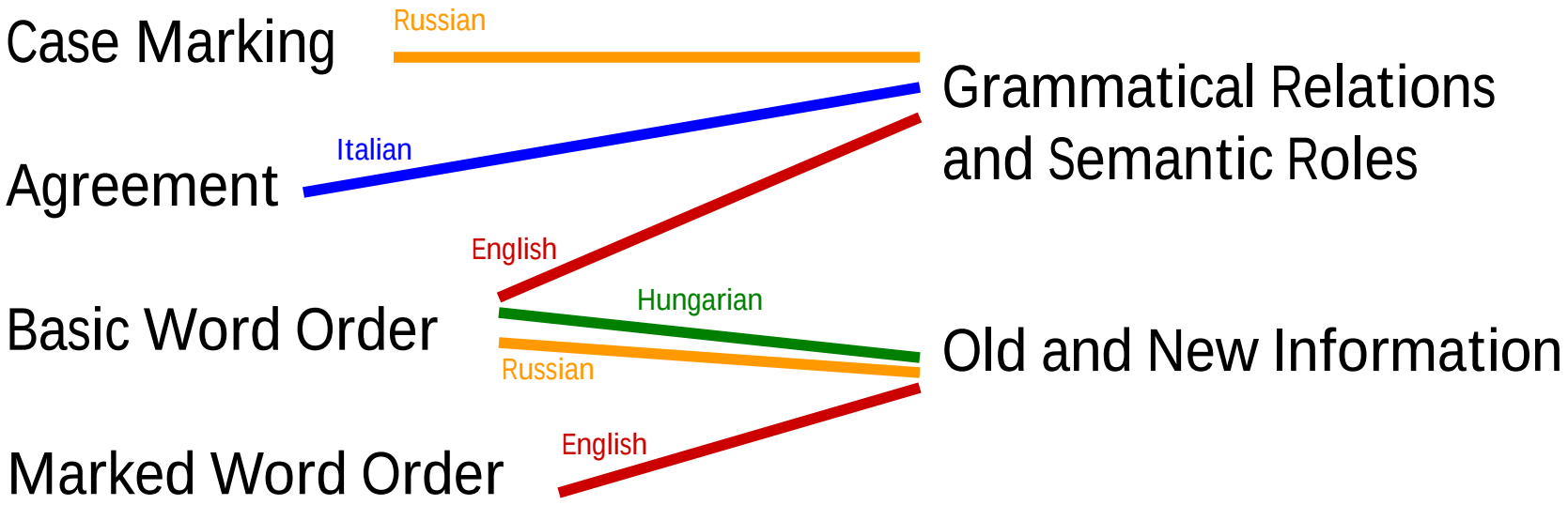
- What type of linguistic knowledge are we talking about?
- What form will it take?
- Where will it come from?

What type of linguistic knowledge?

- Each sentence expresses some communicative functions.
  - The child was hungry so *mother gave him a cookie*.
- Make a statement (vs question)
- Express certainty
- Refer to an event that took place before the time of speech
- Refer to three entities
- Quantity of each entity
- Roles and information status of each entity: an agent (evoked) acting on a patient (new, inanimate) for the benefit of a recipient (in focus, animate)

- Each language has a number of morpho-syntactic mechanisms:
  - affixation
  - compounding
  - word order
- Each language has a number of morpho-syntactic systems
  - agreement
  - case
  - voice
  - determiners
  - classifiers
  - numerals

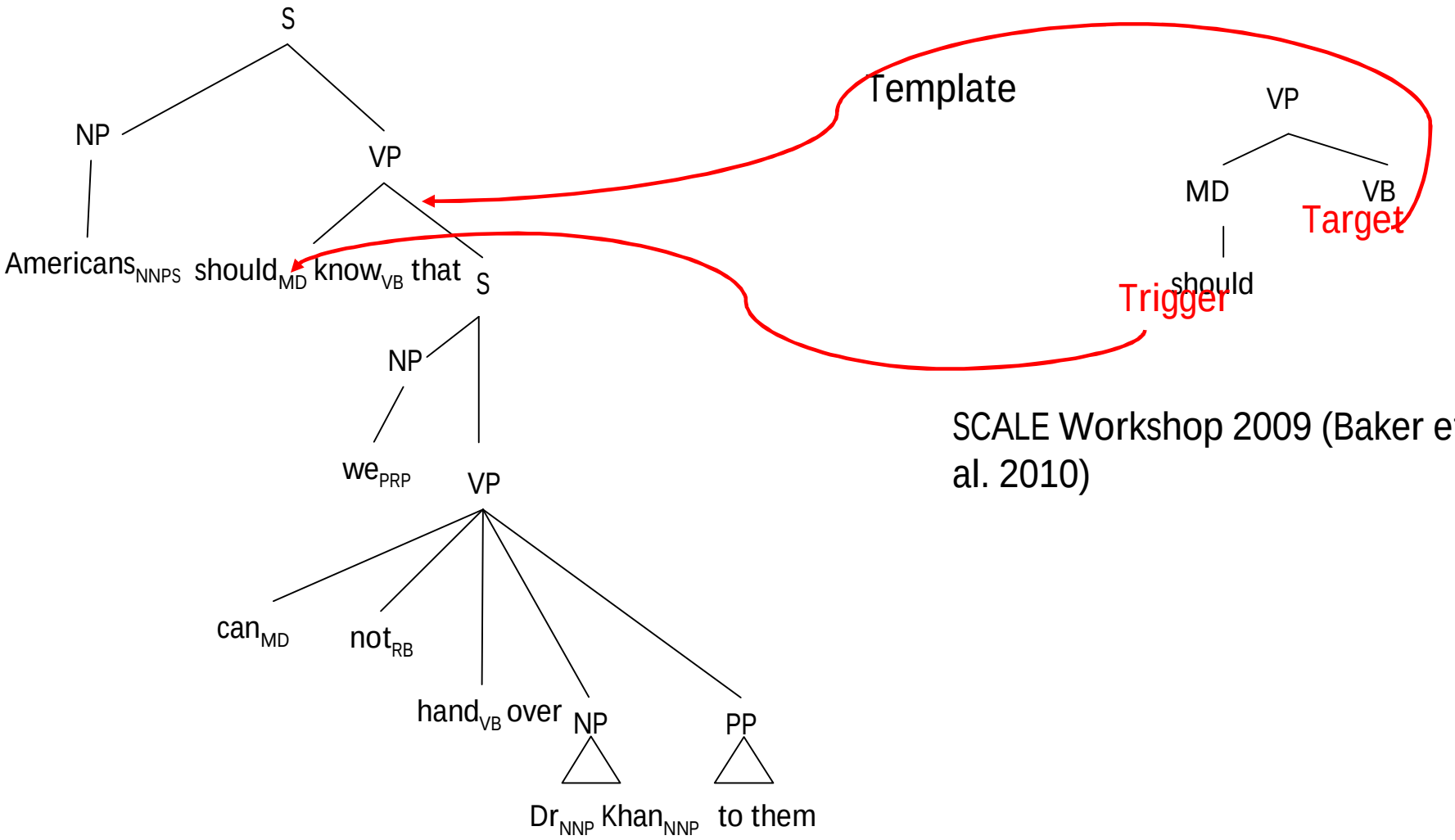
Cross-Linguistic Variation in the expression of communicative functions



What form will it take?

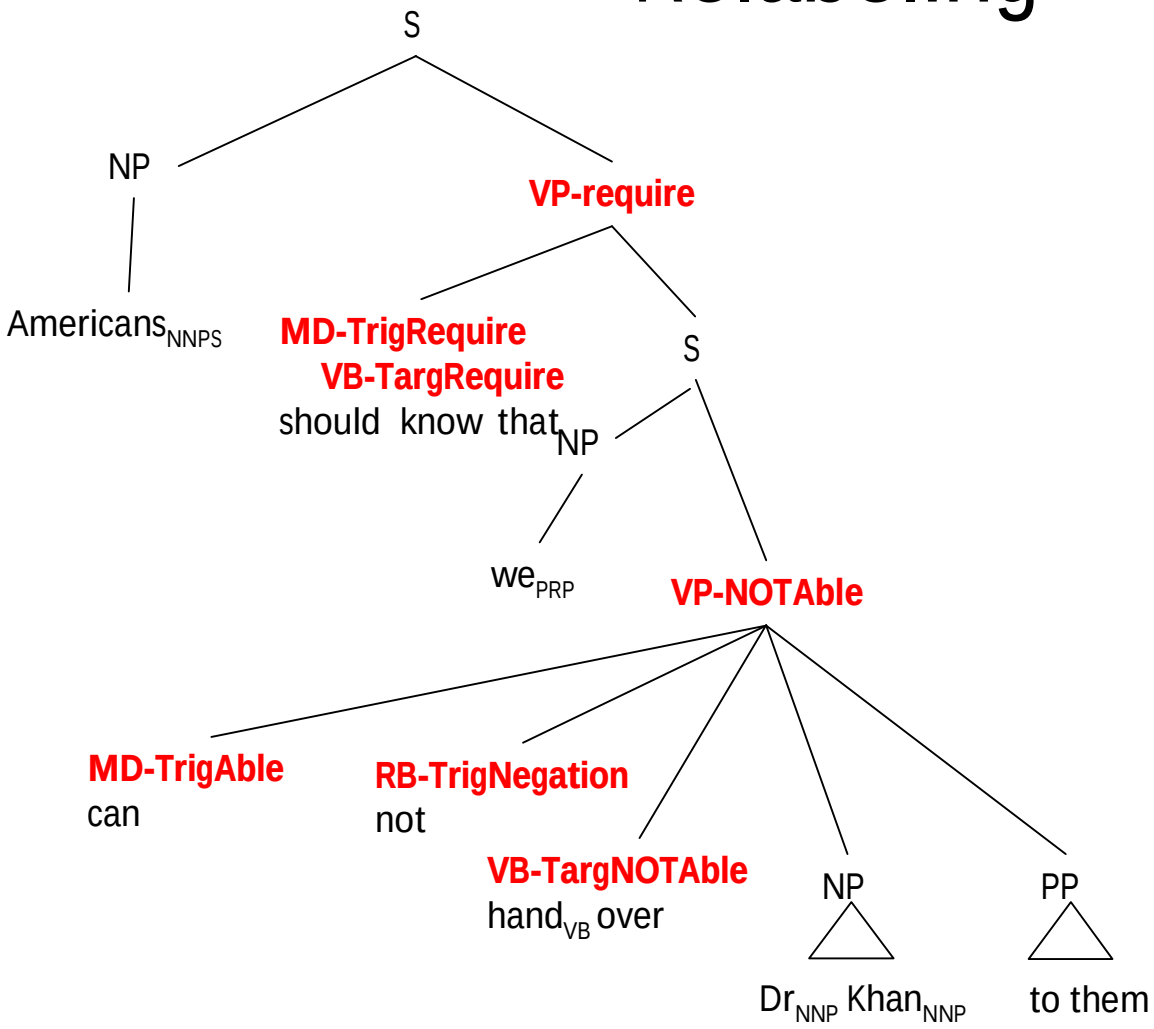
```
RBMT RULE BASE AND RESULTS

  qjo.red     <-> r o j          qje.green     <-> v e r d e
  qnsmasc.car <-> c o c h e      qnsmasc.cat   <-> g a t o
  qnsfem.moon <-> l u n a        qnesfem.light <-> l u z
  qdmasc.a    <-> u n            qdfem.a       <-> u n a

  qmasc.JJ(x0:) <-> qjo.x0 o     qmasc.N(x0:) <-> qnsmasc.x0
  qmasc.JJ(x0:) <-> qje.x0       qmasc.N(x0:) <-> qnesmasc.x0
  qplmasc.x0:JJ <-> qmasc.x0 s   qplmasc.N(x0: x1:pl) <-> qnsmasc.x0 s
  qplmasc.N(x0: x1:pl) <-> qnesmasc.x0 e s
    ...
  q.NP(DT(x0:) x1:JJ x2:N) <-> qdmasc.x0 _ qmasc.x2 _ qmasc.x1

  % echo 'NP(DT(a) JJ(red) N(car))' | tiburon -l -k 1 - rbmt.xlnts

  OUTPUTS:
    u n _ c o c h e _ r o j o # 1.0        (no other outputs)
```

How can we add semantics to tree transduction?

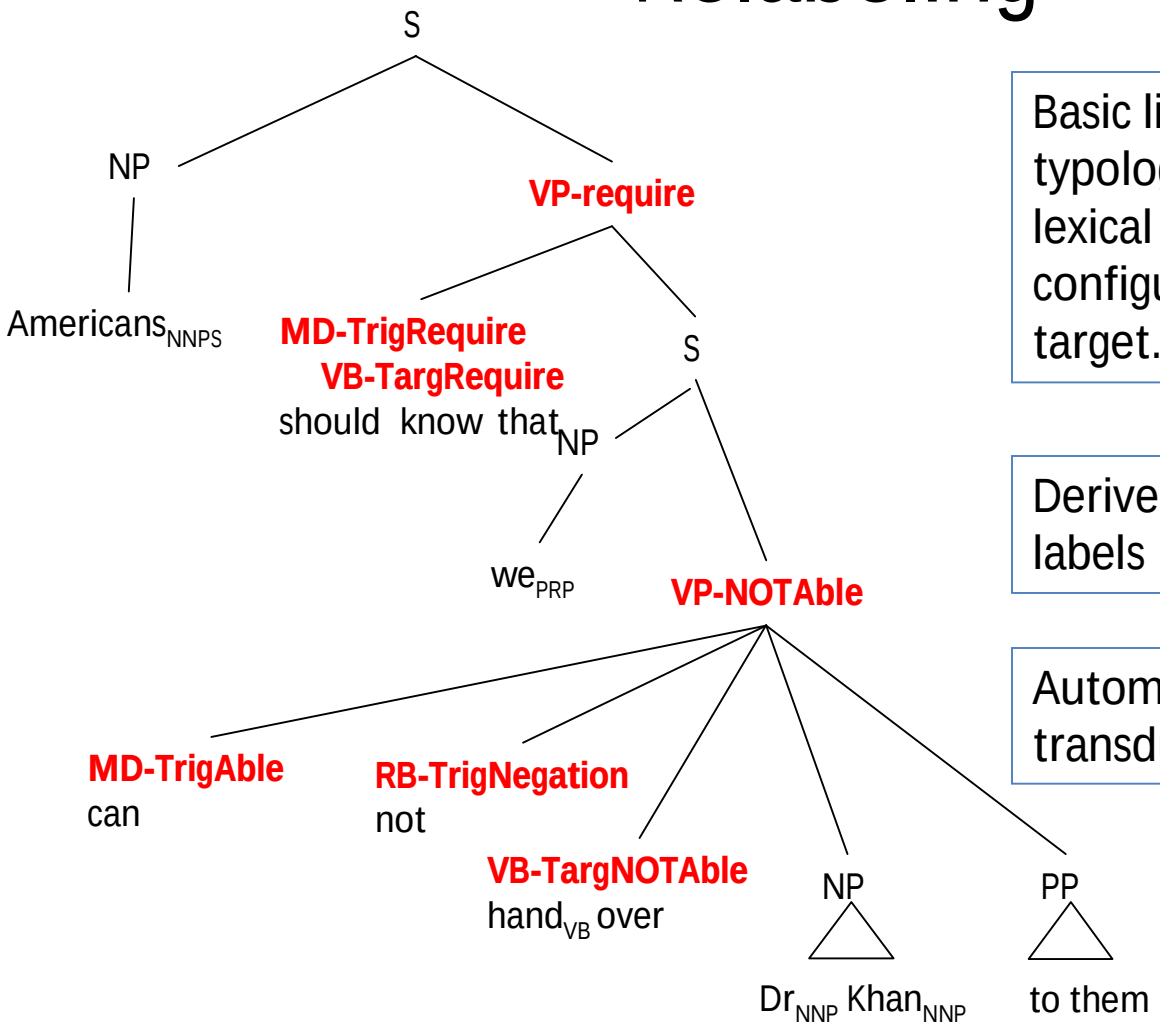


SCALE Workshop 2009 (Baker et al. 2010)

Introducing Semantics through Node Relabeling



Introducing Semantics through Node Relabeling



Basic linguistic knowledge: typological inventory of modality; lexical items; structural configuration of the trigger and target.

Derived Linguistic Knowledge: node labels

Automatically learned transductions.

How can we add semantics to tree transductions?

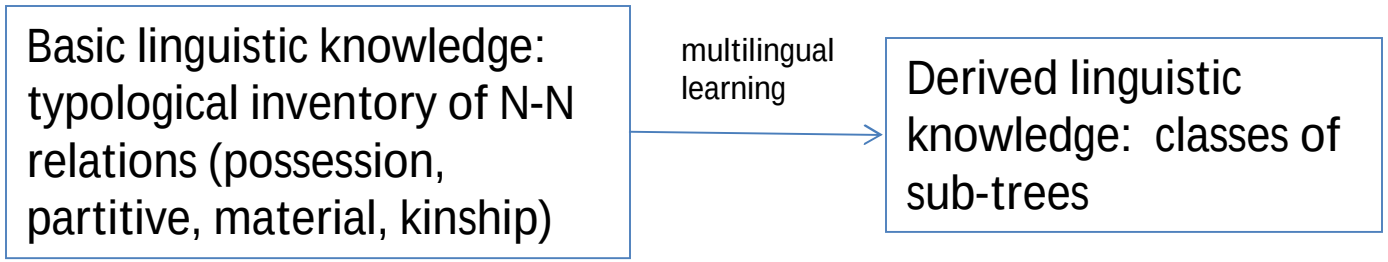
- Combine the classifier and the transducer
  - make the classifier's features available when the transducer is learned
    - The mother (kinship) of that student
    - Some (quantifier) of the students
    - A house of bamboo (material)

Basic linguistic knowledge: semantic properties of lexical items; typological inventory of N-N semantics (possession, material, kinship, partitive)

Automatically learned transductions

How can we add semantics to tree transductions

- Use multi-lingual data to learn classes
  - See Barzilay's presentation
    - the mother of that student/the brother of my friend
    - some of the students/all of the students
    - a house of bamboo/a table of exotic wood



Where will it come from?

- Reference grammars
- Linguists
- Native speakers
  - active learning
  - elicitation
- Studies of typology and universals
- Corpus annotation

Elicitation of Communicative Functions

srcsent: A man sang.
tgtsent: Aŋun atuq+tuaq .
aligned: ((2,1),(3,2))
context: the speaker is talking about a specific man

srcsent: A man sang.
tgtsent: Aŋun atuq+tuaq .
aligned: ((2,1),(3,2))
context: the speaker is not speaking about a specific man

srcsent: Men sang.
tgtsent: Aŋuti+k atuq+tuak .
aligned: ((1,1),(2,3))
context: Men = two people

srcsent: Men sang.
tgtsent: Aŋuti+t atuq+tuat .
aligned: ((1,1),(1,2),(2,3))
context: Men = five people

srcsent: We sang.
tgtsent: Atuq+tuagut .
aligned: ((1,2),(2,1))
context: 'We' does not include the listener; we = five men

srcsent: The man should sing.
tgtsent: Aŋun atuġ+li .
aligned: ((2,1),(4,2))

srcsent: The man sings.
tgtsent: Aŋun atuq+tuq .
aligned: ((2,1),(3,2))

srcsent: The man may sing.
tgtsent: Aŋun atuġ+li .
aligned: ((2,1),(4,2))

Current Elicitation Corpus

- 3000 sentences
- Based on typological inventories and fieldwork checklists
  - tense, person, number, formality, type of causation, mood, evidentiality, etc.
- Available in LDC language packs in several languages

Future work on elicitation

- Dynamic
  - A different set of sentences is relevant for each language
- Targeted
  - Active learning determines which data will be the most valuable
- Larger vocabulary

Corpus Annotation: Triple Gold