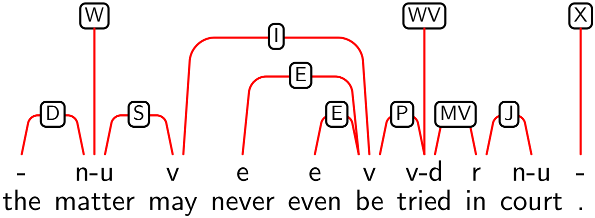
Deriving Multi-Headed Planar Dependency Parses from Link Grammar Parses

Juneki Hong and Jason Eisner

Introduction

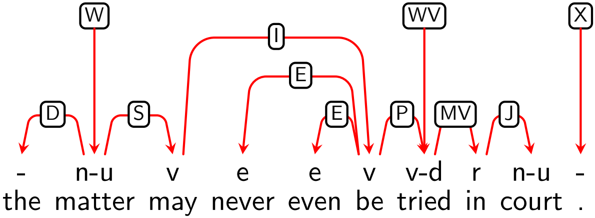
- This talk is about converting from one annotation style to another.
- The conversion could be hard, where information is fragmented, missing, or ambiguous.
- We use a general technique, Integer Linear Programming to help us do this conversion.

In Our Case: What We Started With



Link Grammar: Parse with undirected edges

What We Wanted:



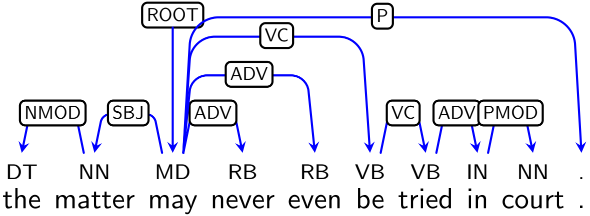
Multiheaded parse with directionalized edges

Why We Wanted That

- We want to develop parsing algorithms for parses that look like this
- We couldn't figure out where to get the data to test them.

Single-headedness

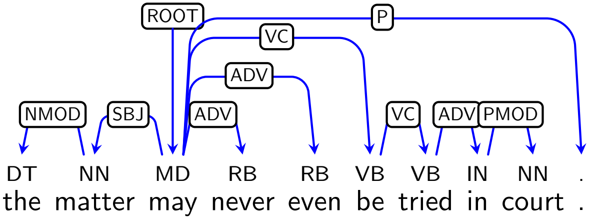
- Dependency parse treebanks today are either single-headed or not planar.
- Stanford Dependencies are multiheaded but not planar



Some example dependency parse.

Single-headedness

- Dependency parse treebanks today are either single-headed or not planar.
- Stanford Dependencies are multiheaded but not planar



Some example dependency parse.

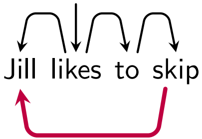
Link Grammar is almost a multiheaded planar corpora! We just need to directionalize the links.

Why Multi-headedness?

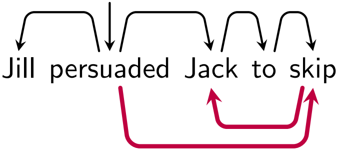
Multi-headedness Can Capture Additional Linguistic Phenomenon

- Control
- Relativization
- Conjunction

Control

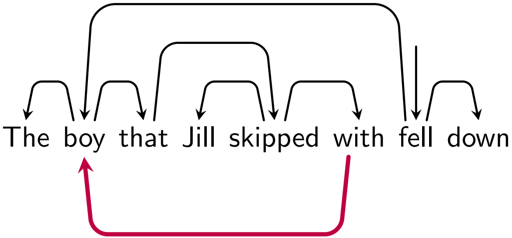


Jill is the subject of two verbs



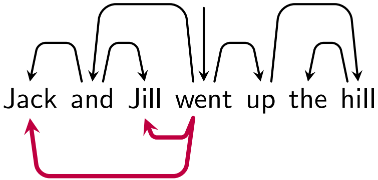
Jack is the object of one verb and the subject of another

Relativization



The boy is the object of *with* as well as the subject of *fell*.

Conjunction



Jack and Jill serve as the two arguments to *and*, but are also subjects of *went*.

Motivation

- A multiheaded dependency corpus would be useful for testing new parsing algorithms

Motivation

- A multiheaded dependency corpus would be useful for testing new parsing algorithms
- Such a corpus could be automatically annotated using Integer Linear Programming

Motivation

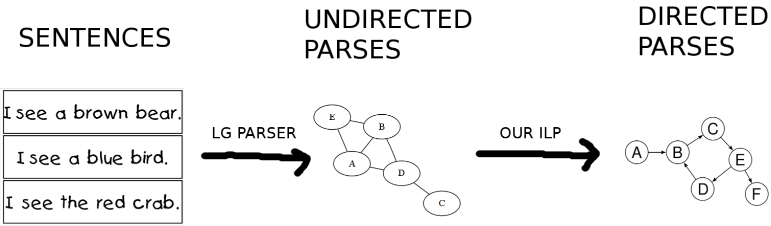
- A multiheaded dependency corpus would be useful for testing new parsing algorithms
- Such a corpus could be automatically annotated using Integer Linear Programming
- We explored whether the Link Grammar could be adapted for this purpose.

Motivation

- A multiheaded dependency corpus would be useful for testing new parsing algorithms
- Such a corpus could be automatically annotated using Integer Linear Programming
- We explored whether the Link Grammar could be adapted for this purpose.
- The results of this are mixed, but provides a good case study.

Corpus Building Strategy

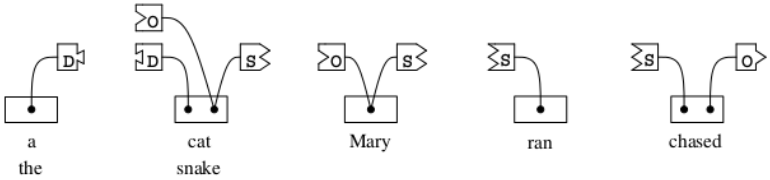
- We start with some sentences and parse them with LG Parser
- We take the undirected parses and try to directionalize them.
- We use an ILP to assign consistent directions for each link type.



Link Grammars

- Grammar-based formalism for projective dependency parsing with undirected links
- Original formalism and English Link Grammar created by Davy Temperley, Daniel Sleator, and John Lafferty (1991)

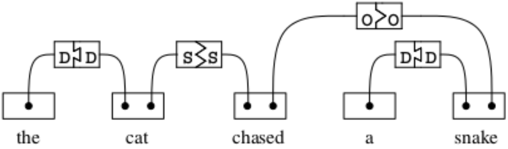
Link Grammars: How They Work



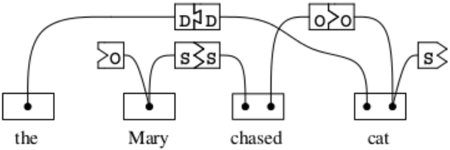
[1]

[1]These figures were clipped from the original Link Grammar paper: "Parsing English with a Link Grammar" by Sleator and Temperley

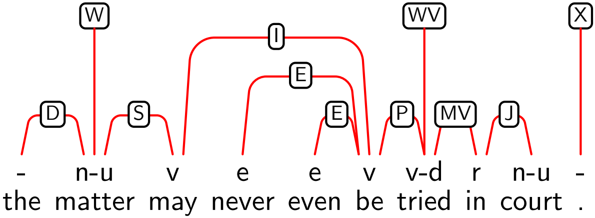
Link Grammars: How They Work



Link Grammars: How They Work



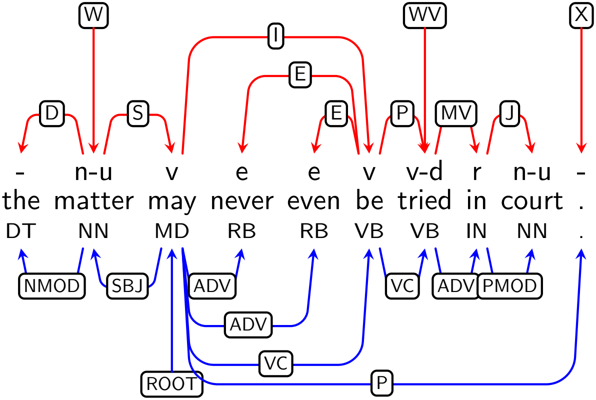
Link Grammars: Same Example Parse From Before Again



Link Parse of a sentence from Penn Tree Bank

Link Grammars

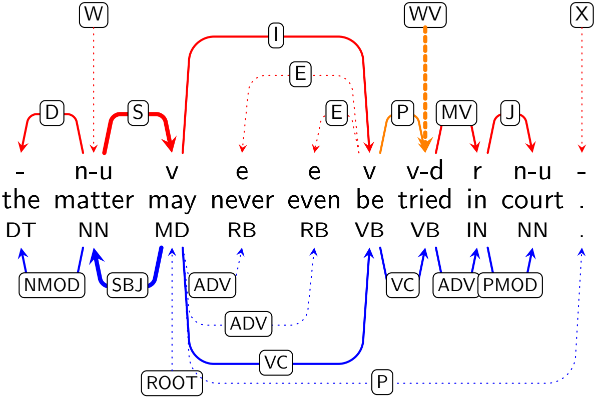
Compare resulting dependency parse with CoNLL 2007 shared task.



Bottom half is CoNLL. Top half is the directionalized link parse.

Link Grammars

Compare resulting dependency parse with CoNLL 2007 shared task.



Bottom half is CoNLL. Top half is the directionalized link parse.

What is Integer Linear Programming?

- An optimization problem where some or all of the variables are integers.
- The objective function and constraints are linear.
- In general, it's NP-Hard! But good solvers exist that work well most of the time.
- Our ILP is encoded as a ZIMPL program and solved using the SCIP Optimization Suite[2]

[2] http://scip.zib.de/

Integer Linear Programming Model

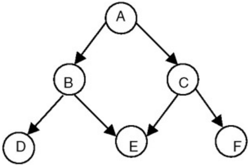
Encoded Constraints:

- Acyclicity
- Connectedness
- Consistency of Directionalized Links

Integer Linear Programming Model

Encoded Constraints:

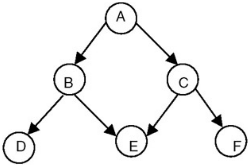
- Acyclicity: (**No cycles!**)
- Connectedness
- Consistency of Directionalized Links

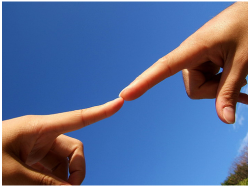


Integer Linear Programming Model

Encoded Constraints:

- Acyclicity: (**No cycles!**)
- Connectedness: (**Every word is reachable from a root**)
- Consistency of Directionalized Links

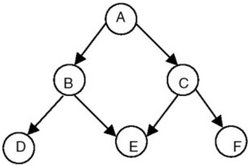


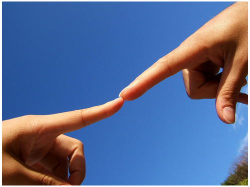


Integer Linear Programming Model

Encoded Constraints:

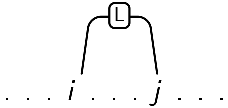
- Acyclicity: (**No cycles!**)
- Connectedness: (**Every word is reachable from a root**)
- Consistency of Directionalized Links:
  (**Similar links oriented the same way**)





Integer Linear Programming Model

For each sentence, for each edge i, j, where $i < j$



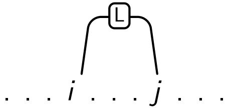
Variables:

$x_{ij}, x_{ji} \in \mathbb{Z} \geq 0$: orientation of each link

$x_{ij} + x_{ji} = 1$

Integer Linear Programming Model

For each sentence, for each edge i, j, where $i < j$



Variables:

$x_{ij}, x_{ji} \in \mathbb{Z} \geq 0$: orientation of each link

$x_{ij} + x_{ji} = 1$

An individual link token can either be oriented left or oriented right

Acyclicity, Connectedness

Acyclicity

Given that node u is the parent of v

n_v: length of the sentence containing node v

$d_v \in [0, n_v]$: depth of the node from the root of the sentence

$$(\forall_u)\ d_v + (1 + n_v) \cdot (1 - x_{uv}) \geq 1 + d_u \tag{1}$$

Connectedness

$$\sum_u x_{uv} \geq 1 \tag{2}$$

Acyclicity, Connectedness

Acyclicity

Given that node u is the parent of v

n_v: length of the sentence containing node v

$d_v \in [0, n_v]$: depth of the node from the root of the sentence

$$(\forall_u)\ d_v + (1 + n_v) \cdot (1 - x_{uv}) \geq 1 + d_u \qquad (1)$$

The depth of a child is greater than the depth of the parent

Connectedness

$$\sum_u x_{uv} \geq 1 \qquad (2)$$

Acyclicity, Connectedness

Acyclicity

Given that node u is the parent of v
n_v: length of the sentence containing node v
$d_v \in [0, n_v]$: depth of the node from the root of the sentence

$$(\forall_u)\ d_v + (1 + n_v) \cdot (1 - x_{uv}) \geq 1 + d_u \qquad (1)$$

The depth of a child is greater than the depth of the parent

Connectedness

$$\sum_u x_{uv} \geq 1 \qquad (2)$$

A word has at least 1 parent

Consistency of Directionalized Links

Consistency of Directionalized Links

$r_L, \ell_L \in \{0, 1\}$: whether all links with label L allowed left/right

$$x_{ij} \leq r_L \qquad\qquad x_{ji} \leq \ell_L \tag{3}$$

Objective Function:

$$\min \left(\sum_L r_L + \ell_L \right) \tag{4}$$

Consistency of Directionalized Links with Slack

Consistency of Directionalized Links

$r_L, \ell_L \in \{0,1\}$: whether all links with label L allowed left/right

$$x_{ij} \leq r_L + s_{ij} \qquad x_{ji} \leq \ell_L + s_{ij} \tag{3}$$

Objective Function:

$$\min \left(\sum_L r_L + \ell_L \right) \cdot \frac{N_L}{4} + \sum_{ij} s_{ij} \tag{4}$$

$s_{ij} \in \mathbb{R} \geq 0$: slack variable

N_L: Number of link tokens with label L

Consistency of Directionalized Links with Slack

Consistency of Directionalized Links

$r_L, \ell_L \in \{0,1\}$: whether all links with label L allowed left/right

$$x_{ij} \leq r_L + s_{ij} \qquad x_{ji} \leq \ell_L + s_{ij} \tag{3}$$

Objective Function:

$$\min \left(\sum_L r_L + \ell_L \right) \cdot \frac{N_L}{4} + \sum_{ij} s_{ij} \tag{4}$$

$s_{ij} \in \mathbb{R} \geq 0$: slack variable
N_L: Number of link tokens with label L

Slack allows a few links with label L in disallowed directions

Data Sets

Data Sets taken from:

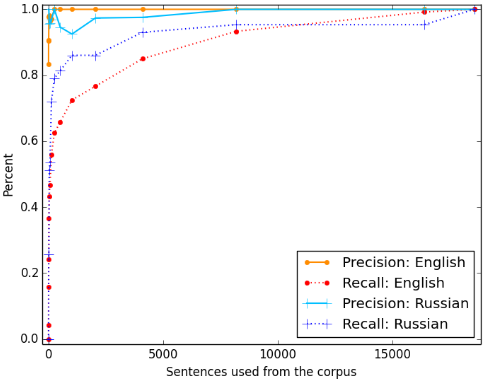
CoNLL 2007 Shared Task (English)

ACL 2013 Shared Task of Machine Translation (Russian)

	Input Sentences	Output Connected Parses
English	18,577	10,960
Russian	18,577	4,913

Stability of Results

- We were worried that the recovered direction mapping might be unstable and sensitive to the input corpus.
- We compared the results of increasing runs of sentences.



On the English Data Set:

On the English Data Set:

Multiheadedness

Link Data has 8% additional edges over the CoNLL.
(average about 2 multiheaded words per sentence)

On the English Data Set:

Multiheadedness

Link Data has 8% additional edges over the CoNLL.
(average about 2 multiheaded words per sentence)

CoNLL Matches

52% of links match CoNLL arcs

57% of CoNLL arcs match links

On the English Data Set:

- Multiheadedness
  - Link Data has 8% additional edges over the CoNLL. (average about 2 multiheaded words per sentence)
- CoNLL Matches
  - 52% of links match CoNLL arcs
  - 57% of CoNLL arcs match links
- Directionality
  - 6.19% of link types allowed both directions
  - 2.07% of link tokens required disallowed direction via slack

ILP Results: Top 25 Most Occurring Labels

Label	Rightward	Multiheaded	CoNLL Match	CoNLL Dir Match
A	0% (0/8501)	0% (0/8501)	84% (7148/8501)	98% (7002/7148)
AN	0% (0/9401)	0% (0/9401)	83% (7825/9401)	98% (7639/7825)
B	100% (1514/1515)	61% (919/1515)	53% (806/1515)	84% (678/806)
C	100% (3272/3272)	0% (0/3272)	3% (85/3272)	53% (45/85)
CO	0% (0/2478)	1% (32/2478)	5% (114/2478)	68% (78/114)
CV	100% (3237/3237)	100% (3237/3237)	56% (1827/3237)	28% (512/1827)
D	0% (56/19535)	0% (71/19535)	85% (16656/19535)	100% (16608/16656)
E	0% (0/1897)	0% (2/1897)	67% (1279/1897)	99% (1263/1279)
G	0% (0/6061)	0% (0/6061)	70% (4258/6061)	96% (4070/4258)
I	100% (5405/5424)	60% (3247/5424)	95% (5168/5424)	47% (2408/5168)
IV	100% (1626/1627)	100% (1626/1627)	85% (1389/1627)	97% (1353/1389)
J	98% (16400/16673)	2% (280/16673)	87% (14522/16673)	97% (14069/14522)
M	100% (9594/9596)	0% (16/9596)	74% (7124/9596)	92% (6583/7124)
MV	100% (13375/13376)	0% (61/13376)	51% (6797/13376)	98% (6681/6797)
MX	100% (1999/1999)	4% (83/1999)	42% (836/1999)	91% (763/836)
O	100% (11027/11028)	0% (0/11028)	81% (8932/11028)	96% (8535/8932)
P	100% (3755/3756)	31% (1167/3756)	94% (3528/3756)	100% (3523/3528)
S	97% (13138/13520)	57% (7662/13520)	92% (12476/13520)	5% (586/12476)
SJ	50% (2736/5468)	0% (0/5468)	69% (3778/5468)	93% (3502/3778)
TO	100% (1733/1734)	0% (1/1734)	0% (5/1734)	100% (5/5)
VJ	51% (765/1500)	1% (8/1500)	71% (1059/1500)	89% (939/1059)
W	100% (10528/10528)	0% (5/10528)	5% (504/10528)	46% (232/504)
WV	100% (7563/7563)	100% (7557/7563)	57% (4345/7563)	97% (4214/4345)
X	80% (13132/16406)	5% (806/16406)	8% (1364/16406)	95% (1300/1364)
YS	0% (0/1645)	0% (0/1645)	98% (1619/1645)	0% (0/1619)

ILP Results: Top 25 Most Occurring Labels

Label	Rightward	Multiheaded	CoNLL Match	CoNLL Dir Match
A	0% (0/8501)	0% (0/8501)	84% (7148/8501)	98% (7002/7148)
AN	0% (0/9401)	0% (0/9401)	83% (7825/9401)	98% (7639/7825)
B	**100% (1514/1515)**	**61% (919/1515)**	**53% (806/1515)**	**84% (678/806)**
C	100% (3272/3272)	0% (0/3272)	3% (85/3272)	53% (45/85)
CO	0% (0/2478)	1% (32/2478)	5% (114/2478)	68% (78/114)
CV	**100% (3237/3237)**	**100% (3237/3237)**	**56% (1827/3237)**	**28% (512/1827)**
D	0% (56/19535)	0% (71/19535)	85% (16656/19535)	100% (16608/16656)
E	0% (0/1897)	0% (2/1897)	67% (1279/1897)	99% (1263/1279)
G	0% (0/6061)	0% (0/6061)	70% (4258/6061)	96% (4070/4258)
I	100% (5405/5424)	60% (3247/5424)	95% (5168/5424)	47% (2408/5168)
IV	100% (1626/1627)	100% (1626/1627)	85% (1389/1627)	97% (1353/1389)
J	98% (16400/16673)	2% (280/16673)	87% (14522/16673)	97% (14069/14522)
M	100% (9594/9596)	0% (16/9596)	74% (7124/9596)	92% (6583/7124)
MV	100% (13375/13376)	0% (61/13376)	51% (6797/13376)	98% (6681/6797)
MX	100% (1999/1999)	4% (83/1999)	42% (836/1999)	91% (763/836)
O	100% (11027/11028)	0% (0/11028)	81% (8932/11028)	96% (8535/8932)
P	100% (3755/3756)	31% (1167/3756)	94% (3528/3756)	100% (3523/3528)
S	97% (13138/13520)	57% (7662/13520)	92% (12476/13520)	5% (586/12476)
SJ	50% (2736/5468)	0% (0/5468)	69% (3778/5468)	93% (3502/3778)
TO	100% (1733/1734)	0% (1/1734)	0% (5/1734)	100% (5/5)
VJ	51% (765/1500)	1% (8/1500)	71% (1059/1500)	89% (939/1059)
W	100% (10528/10528)	0% (5/10528)	5% (504/10528)	46% (232/504)
WV	100% (7563/7563)	100% (7557/7563)	57% (4345/7563)	97% (4214/4345)
X	80% (13132/16406)	5% (806/16406)	8% (1364/16406)	95% (1300/1364)
YS	0% (0/1645)	0% (0/1645)	98% (1619/1645)	0% (0/1619)

ILP Results: Top 25 Most Occurring Labels

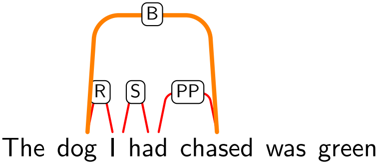
Label	Rightward	Multiheaded	CoNLL Match	CoNLL Dir Match
B	**100% (1514/1515)**	**61% (919/1515)**	**53% (806/1515)**	**84% (678/806)**

Label	Rightward	Multiheaded	CoNLL Match	CoNLL Dir Match
CV	**100% (3237/3237)**	**100% (3237/3237)**	**56% (1827/3237)**	**28% (512/1827)**

ILP Results: Top 25 Most Occurring Labels

Label	Rightward	Multiheaded	CoNLL Match	CoNLL Dir Match
B	**100% (1514/1515)**	**61% (919/1515)**	**53% (806/1515)**	**84% (678/806)**

“B” link relative clauses

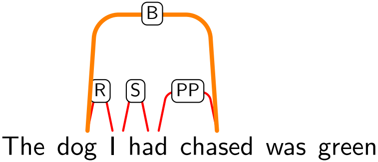


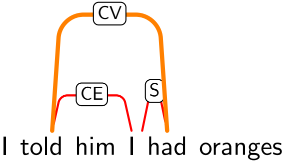
Label	Rightward	Multiheaded	CoNLL Match	CoNLL Dir Match
CV	**100% (3237/3237)**	**100% (3237/3237)**	**56% (1827/3237)**	**28% (512/1827)**

ILP Results: Top 25 Most Occurring Labels

Label	Rightward	Multiheaded	CoNLL Match	CoNLL Dir Match
B	**100% (1514/1515)**	**61% (919/1515)**	**53% (806/1515)**	**84% (678/806)**

"B" link relative clauses

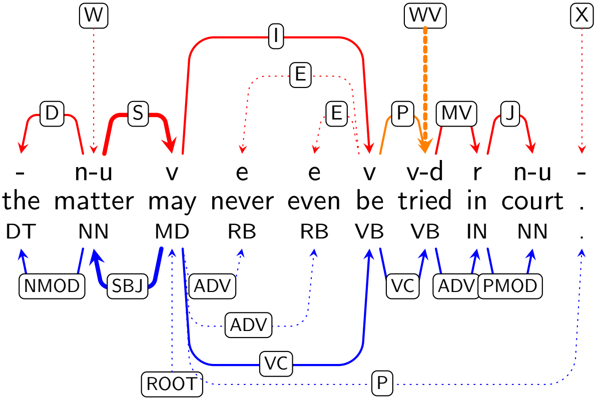




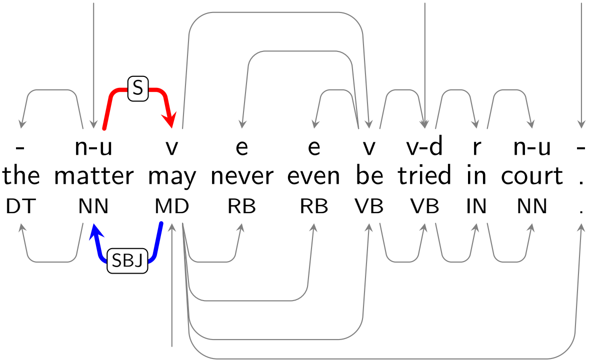
Label	Rightward	Multiheaded	CoNLL Match	CoNLL Dir Match
CV	**100% (3237/3237)**	**100% (3237/3237)**	**56% (1827/3237)**	**28% (512/1827)**

"CV" link conjunctions to main verbs of clauses.

Link Results: Subject-Verb Links are Backwards

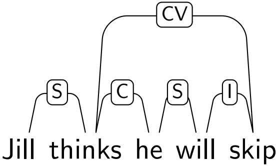


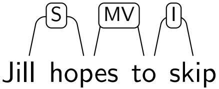
Link Results: Subject-Verb Links are Backwards



Link Results: Subject-Verb Links are Backwards

- This is due to a possible inconsistency of the Link Grammar, discovered by our method.





Link Results: Subject-Verb Links are Backwards

- The Link Grammar seems to be inconsistent about whether the auxiliary verb or the main verb is the head of a clause.

Link Results: Subject-Verb Links are Backwards

- The Link Grammar seems to be inconsistent about whether the auxiliary verb or the main verb is the head of a clause.
- Sometimes the governing verb links to the auxilliary, and sometimes to the main, depending on the type of clause.

Link Results: Subject-Verb Links are Backwards

- The Link Grammar seems to be inconsistent about whether the auxiliary verb or the main verb is the head of a clause.
- Sometimes the governing verb links to the auxilliary, and sometimes to the main, depending on the type of clause.
- But the governing verb usually links to the subject when there is one.

Link Results: Subject-Verb Links are Backwards

- The Link Grammar seems to be inconsistent about whether the auxiliary verb or the main verb is the head of a clause.
- Sometimes the governing verb links to the auxilliary, and sometimes to the main, depending on the type of clause.
- But the governing verb usually links to the subject when there is one.
- So this makes the subject a consistent choice to make the head of a clause.

Link Results: Subject-Verb Links are Backwards

- The Link Grammar seems to be inconsistent about whether the auxiliary verb or the main verb is the head of a clause.
- Sometimes the governing verb links to the auxilliary, and sometimes to the main, depending on the type of clause.
- But the governing verb usually links to the subject when there is one.
- So this makes the subject a consistent choice to make the head of a clause.

To fix this, we could edit the link grammar, link parses, or the ILP.

Conclusions

- Link Grammar parses can be oriented into connected DAGs
- A new corpus available for building multi-headed dependency parsers
- ILP can be used to help annotate incomplete or missing data in corpora.

Questions?