



THE LANGUAGE RESOURCE SPECTRUM: A PERSPECTIVE FROM GOOGLE

Ryan McDonald

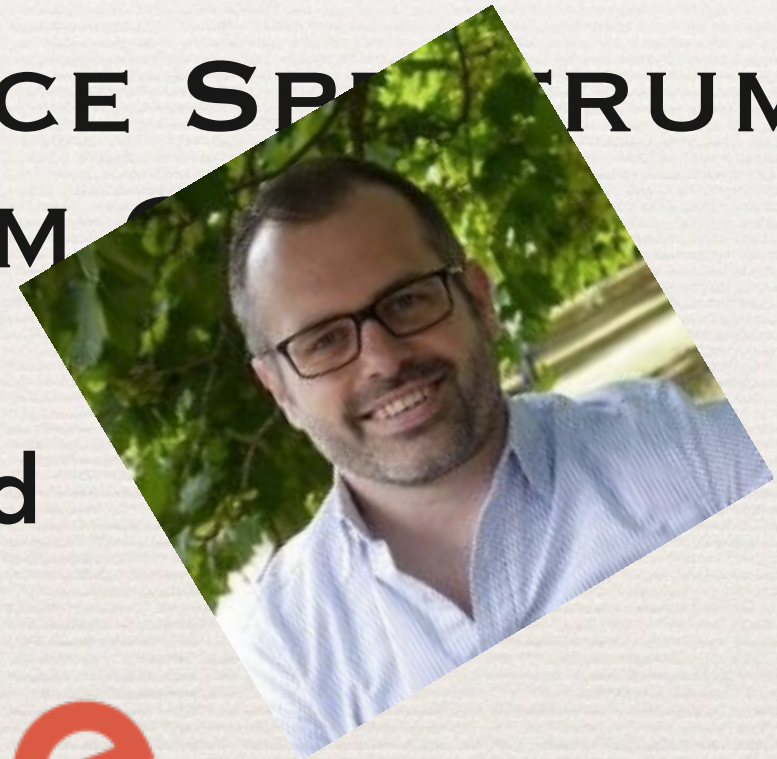


**Google NLU team
Google Linguistics team**



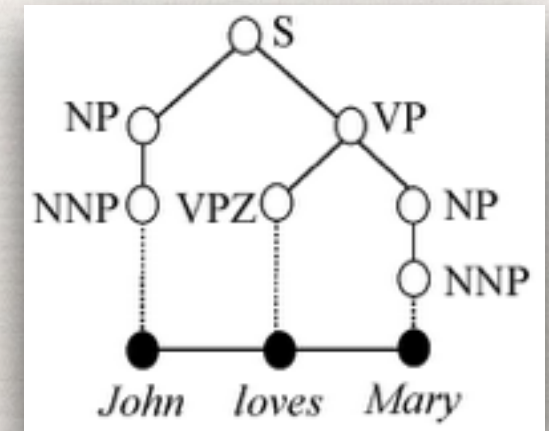
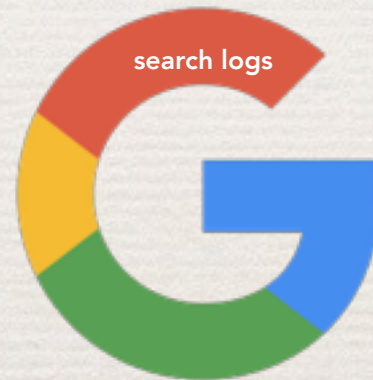
THE LANGUAGE RESOURCE SPECTRUM: A PERSPECTIVE FROM GOOGLE

Ryan McDonald



Google NLU team
Google Linguistics team

LANGUAGE RESOURCE SPECTRUM



LANGUAGE RESOURCE SPECTRUM



unsupervised

weakly supervised

fully supervised

amount
of
data

supervision

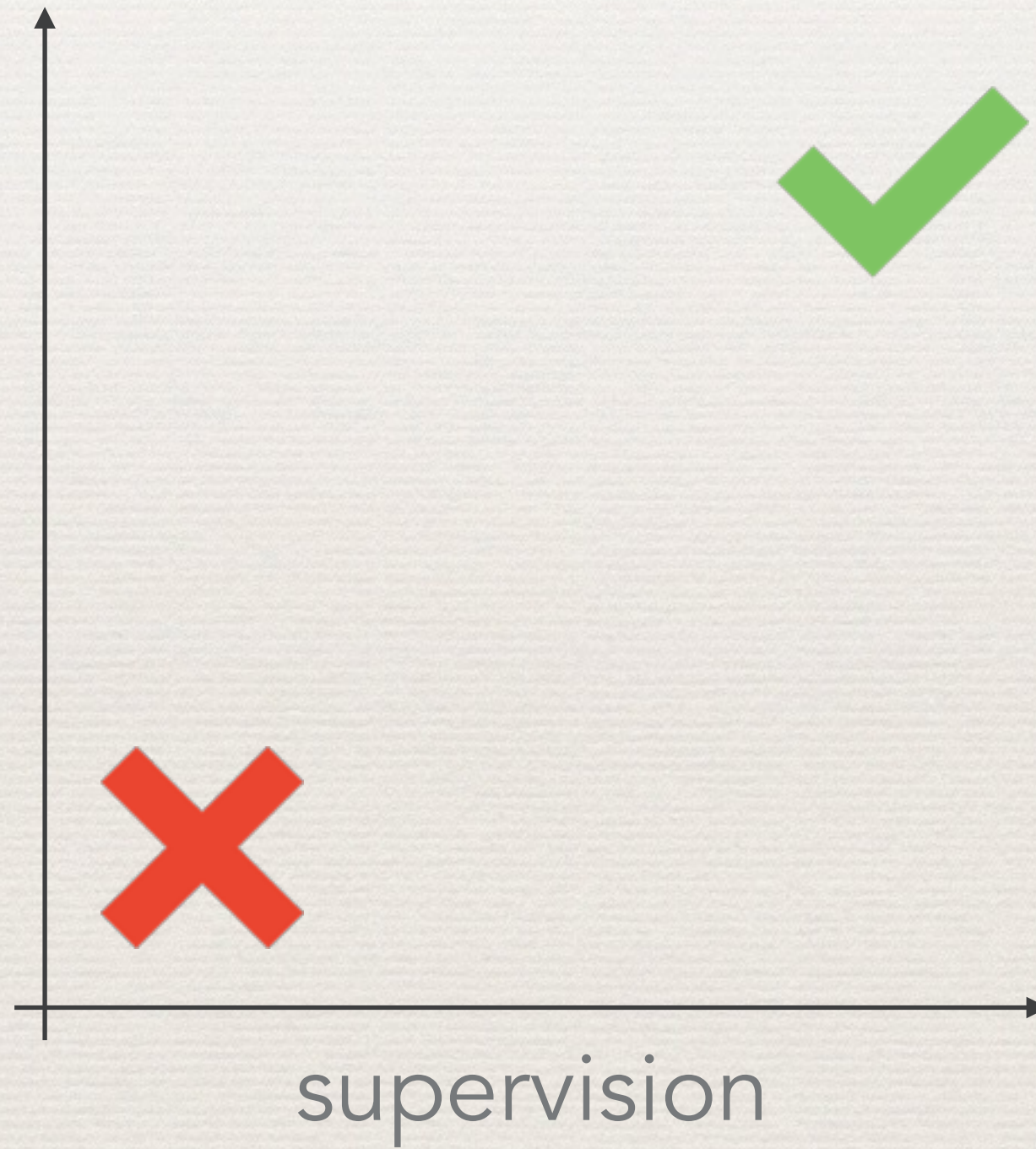


amount
of
data

supervision



amount
of
data



amount
of
data





High quality
annotations



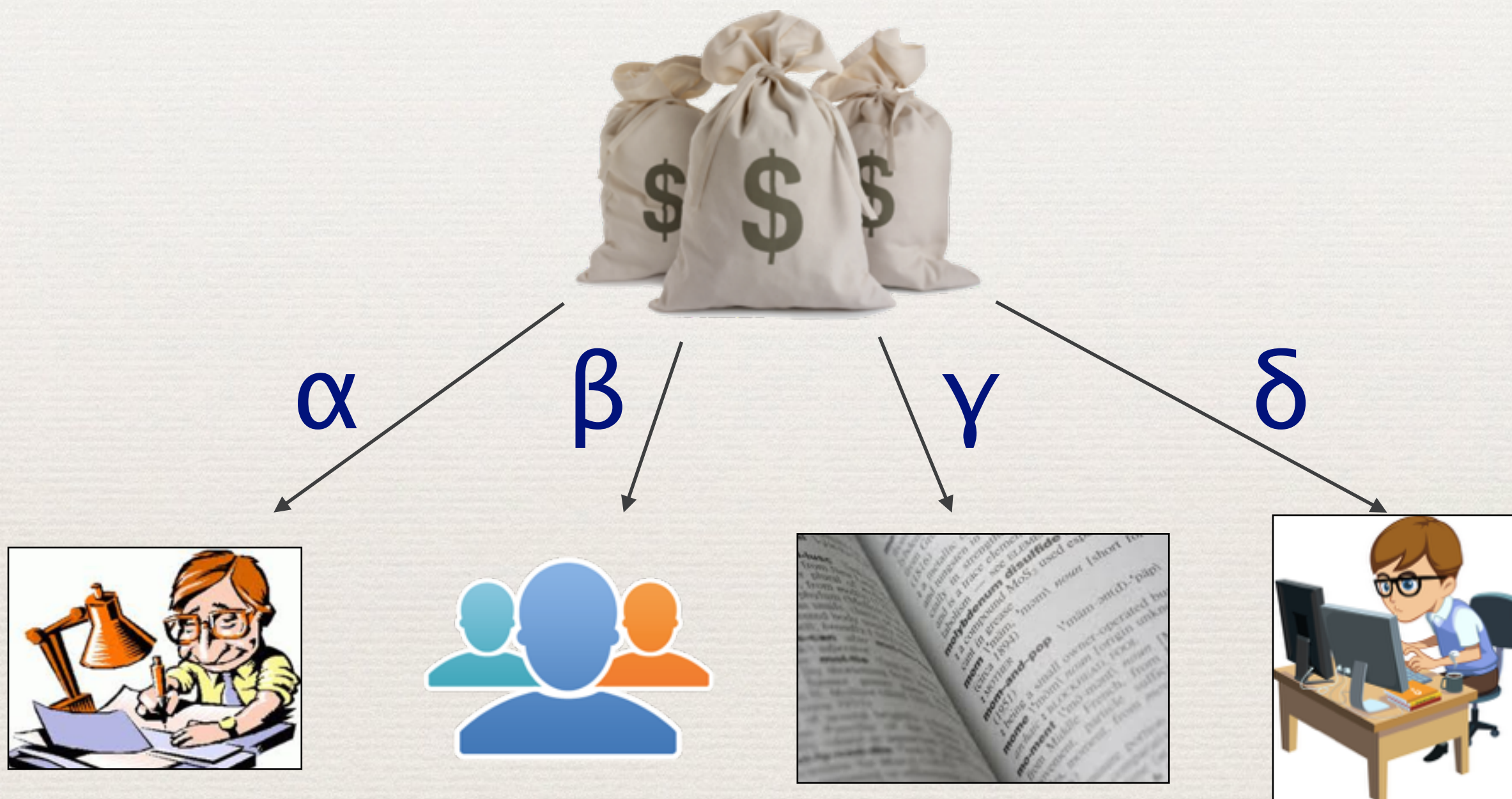
Crowd-sourced



Pre-existing resources



Software Engineer
Auto resources
Models
Active Learning



High quality
annotations

Crowd-sourced

Pre-existing resources

Software Engineer
Auto resources
Models
Active Learning



**High quality
annotations**

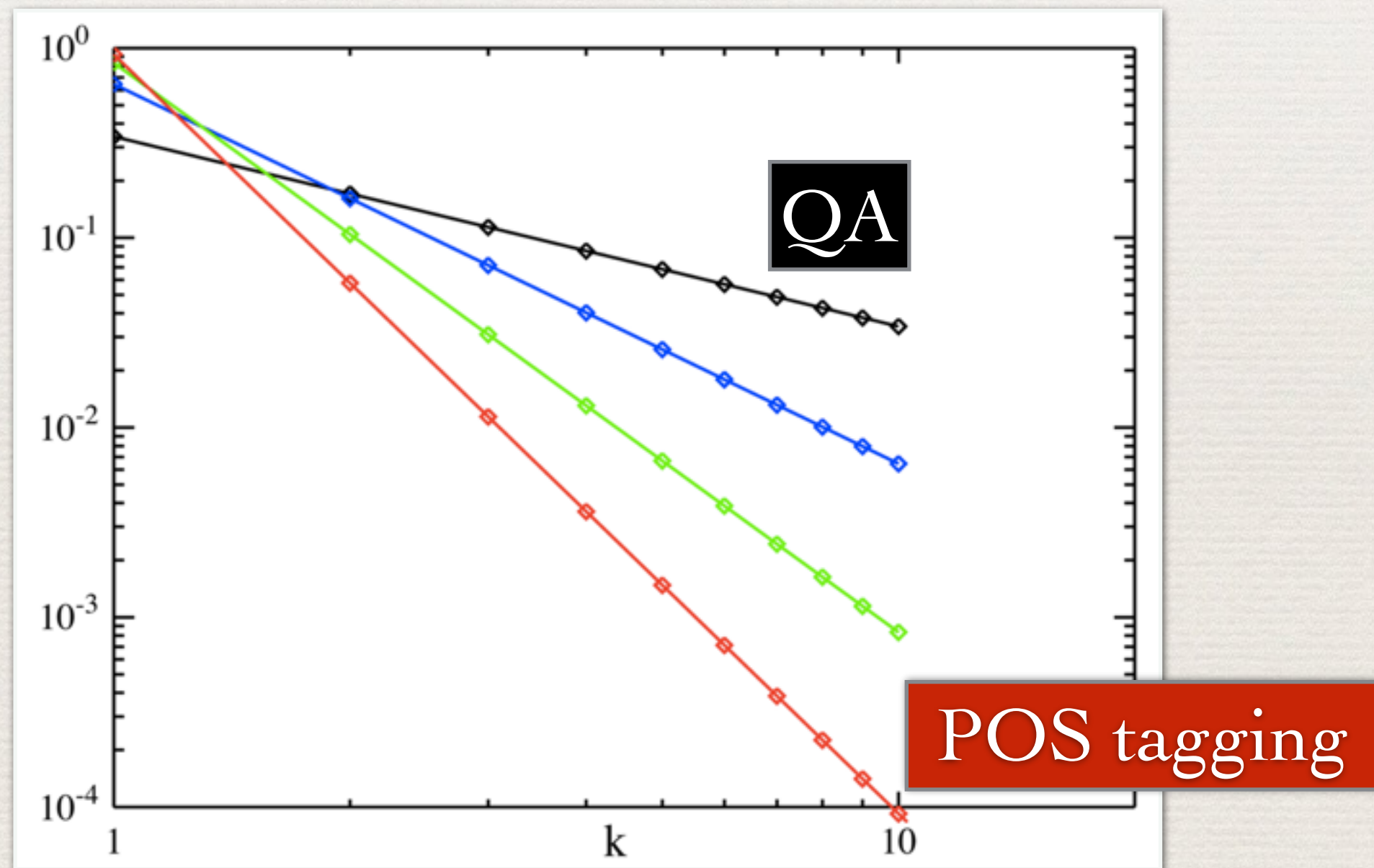
Crowd-sourced

Pre-existing resources

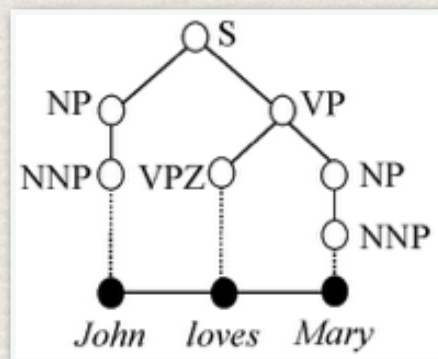
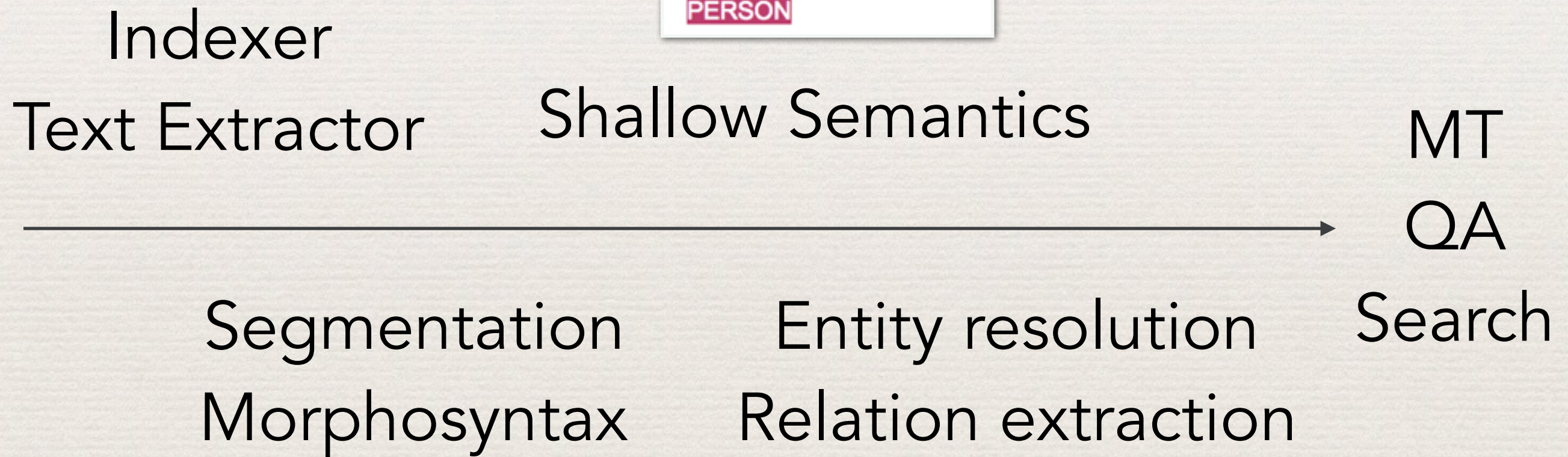
**Software Engineer
Auto resources
Models
Active Learning**

THE TASK MATTERS

- ❖ ML is really good at the head



PIPELINED / MULTI-COMPONENT SYSTEMS



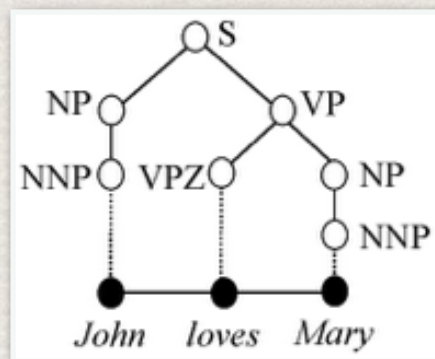
PIPELINED / MULTI-COMPONENT SYSTEMS

Indexer
Text Extractor



Shallow Semantics

Segmentation
Morphosyntax



Entity resolution
Relation extraction



End User Task

MT

QA

Search

PIPELINED / MULTI-COMPONENT SYSTEMS

Indexer
Text Extractor

Shallow Semantics

Bill Gates founded Microsoft

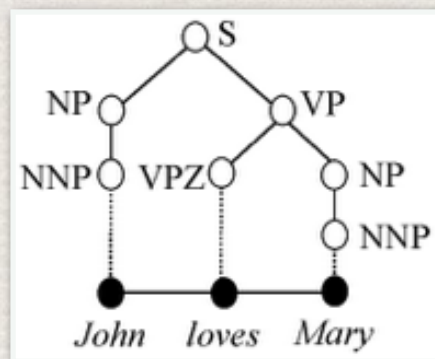
Potential tags:
ORGANIZATION
LOCATION
PERSON

End User Task

MT
QA

Upstream Task

Segmentation
Morphosyntax



Entity resolution
Relation extraction



Search

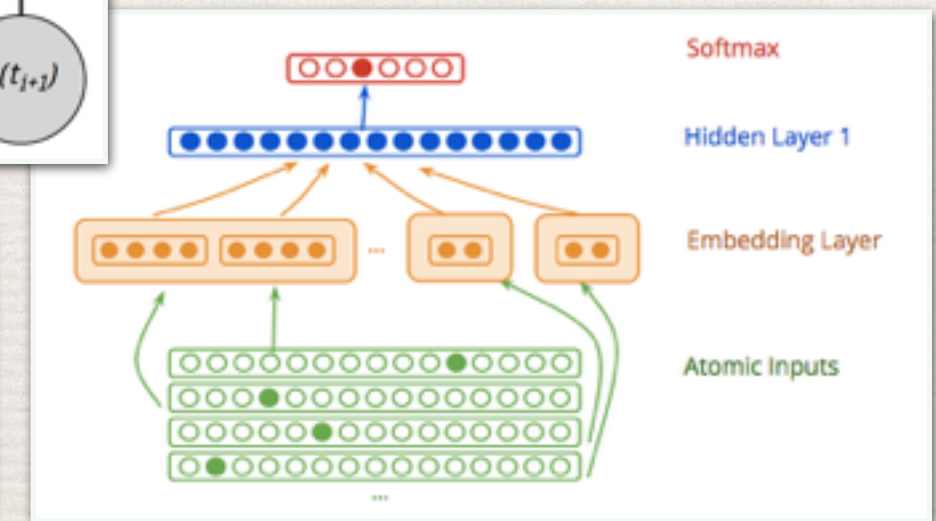
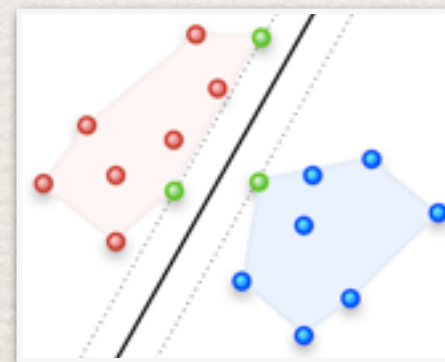
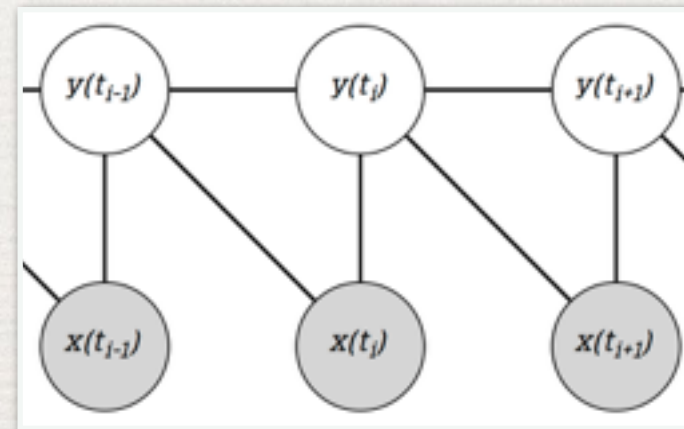
Upstream: Morphosyntactic Tagging

FEATURE-BASED CLASSIFICATION

ce	que	vous	n	entendrez	jamais	dans	un	débat	politique	français
PDEM PRON	IN ADP	lui PRP PRON	RB ADV	entendre VBC VERB	RB ADV	IN ADP	DT DET	NN NOUN	JJ ADJ	JJ ADJ
POS=PRON number=sing	POS=ADP	POS=PRON number=plur person=2	POS=ADV	POS=VERB number=plur person=2 tense=fut mood=ind	POS=ADV	POS=ADP	POS=DET gender=masc number=sing	POS=NOUN gender=masc number=sing	POS=ADJ gender=masc number=sing	POS=ADJ gender=masc number=sing

FEATURE-BASED CLASSIFICATION

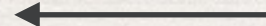
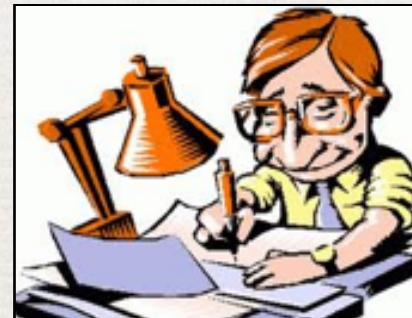
word=entendrez
 suffix3=rez
 word-1=n
 word+1=jamais
 cluster=124
 cluster-1=53
 cluster+1=210



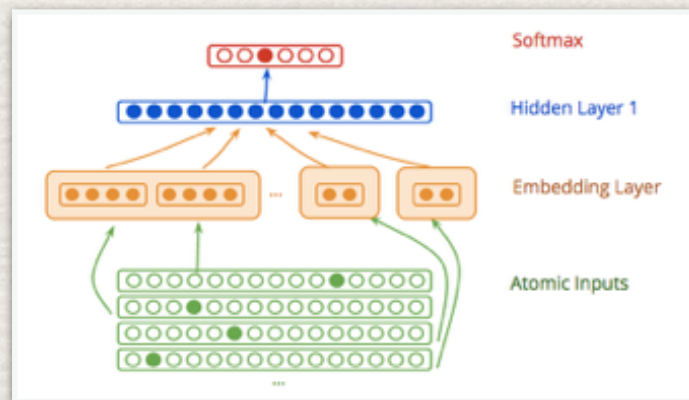
ce	que	vous	n	entendrez	jamais	dans	un	débat	politique	français
PDEM PRON	IN ADP	lui PRP PRON	RB ADV	entendre VBC VERB	RB ADV	IN ADP	DT DET	NN NOUN	JJ ADJ	JJ ADJ
POS=PRON	POS=ADP	POS=PRON	POS=ADV	POS=VERB	POS=ADV	POS=ADP	POS=DET	POS=NOUN	POS=ADJ	POS=ADJ
number=sing		number=plur person=2		number=plur person=2 tense=fut mood=ind			gender=masc number=sing	gender=masc number=sing	gender=masc number=sing	gender=masc number=sing

RESOURCE TRADE-OFF

Annotated data

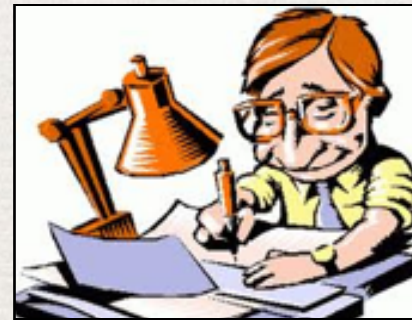


Model

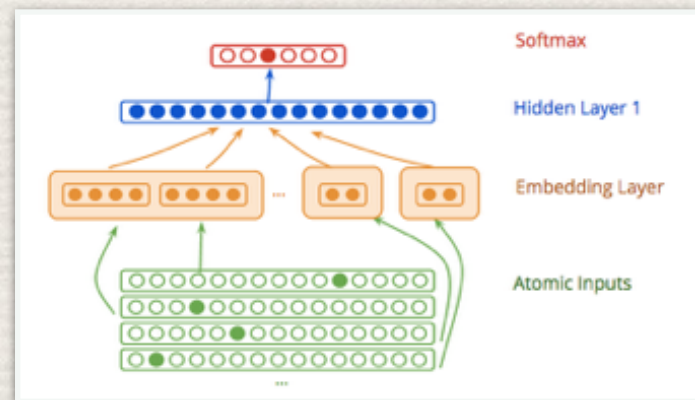


RESOURCE TRADE-OFF

Annotated data



Model

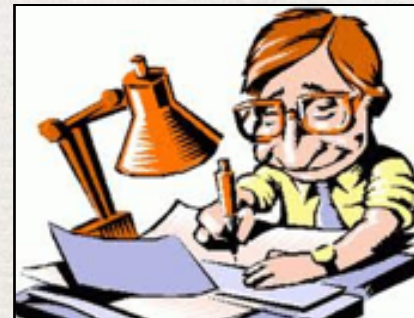


Dictionaries / Lexicons



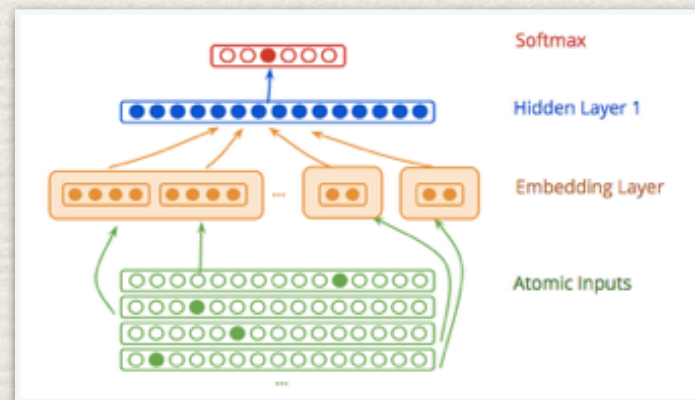
RESOURCE TRADE-OFF

Annotated data

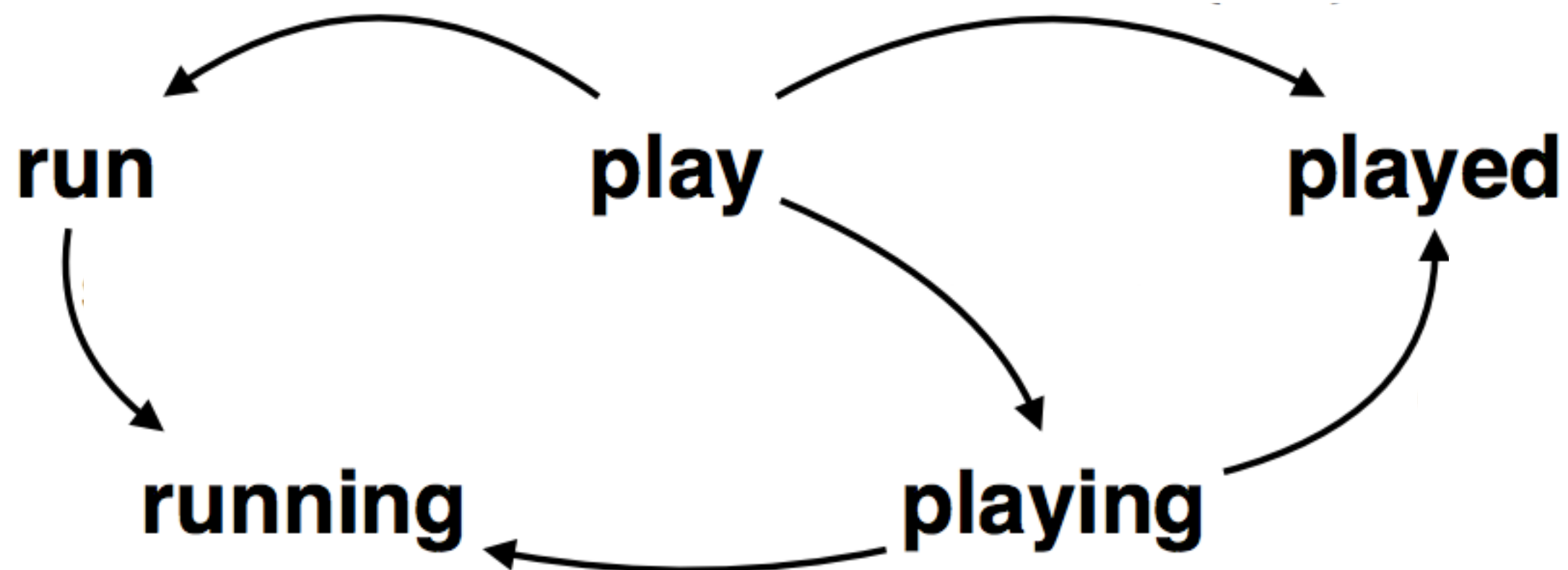


Dictionaries / Lexicons

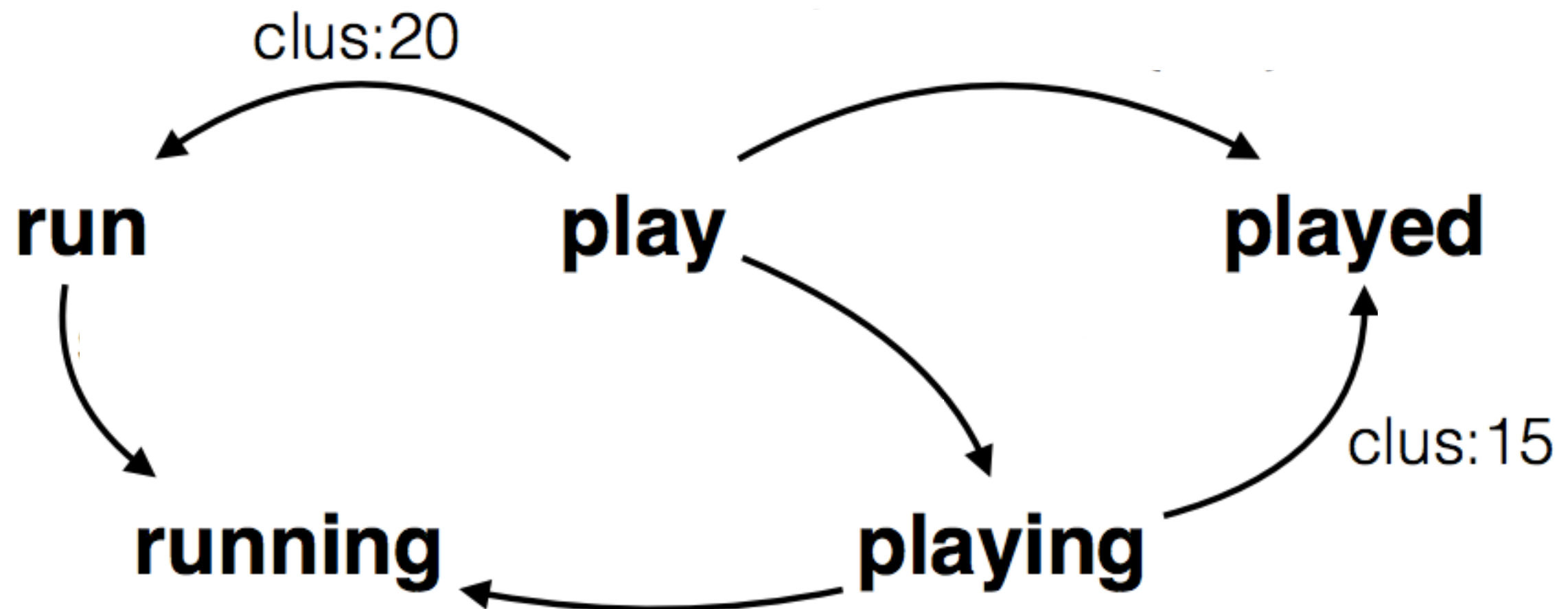
Model



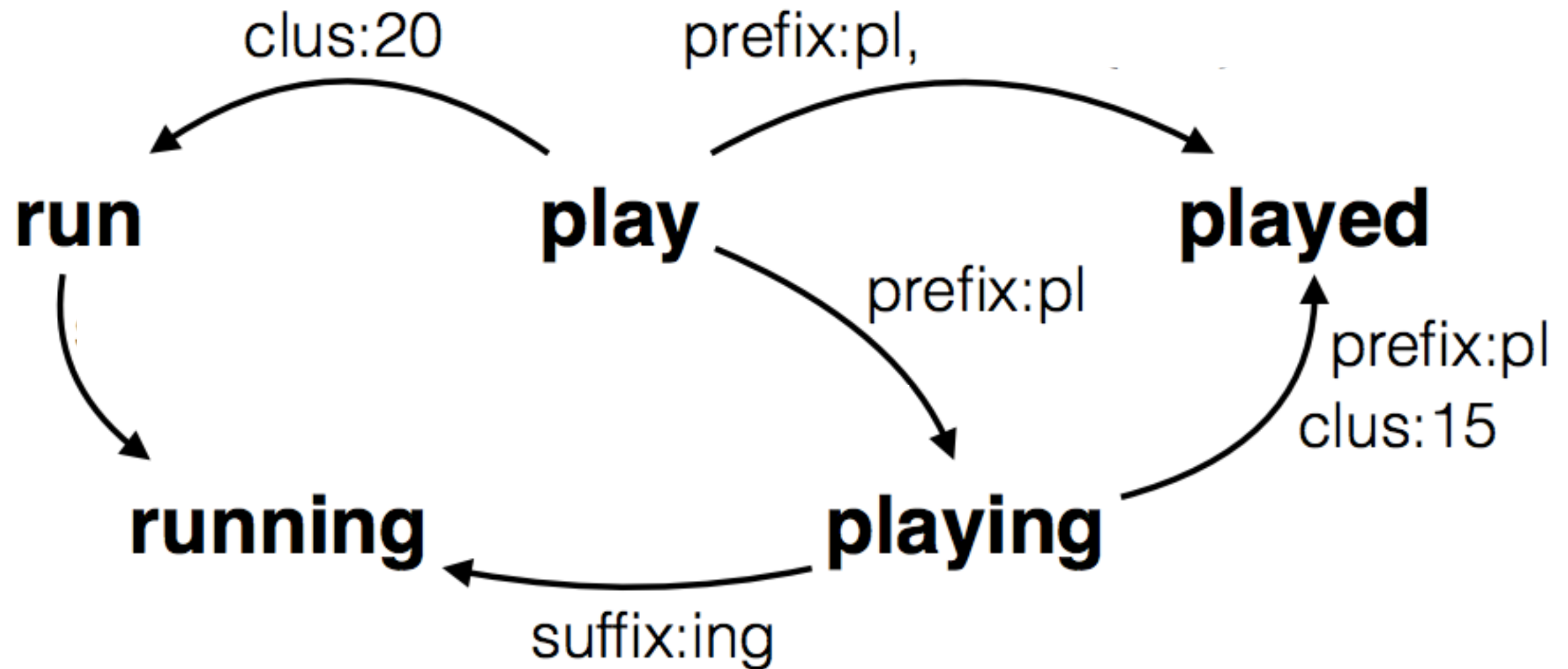
MORPHOSYNTACTIC LEXICONS VIA GRAPH-PROPAGATION



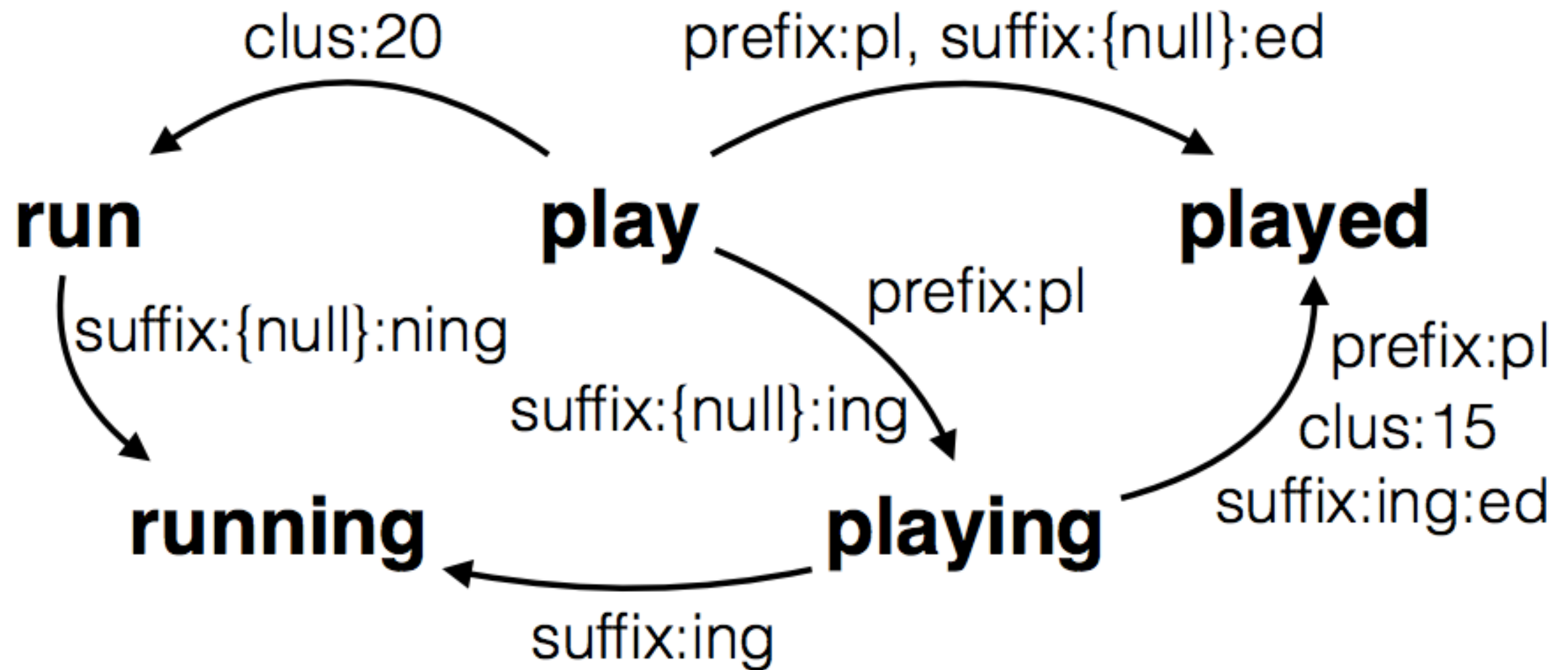
MORPHOSYNTACTIC LEXICONS VIA GRAPH-PROPAGATION



MORPHOSYNTACTIC LEXICONS VIA GRAPH-PROPAGATION

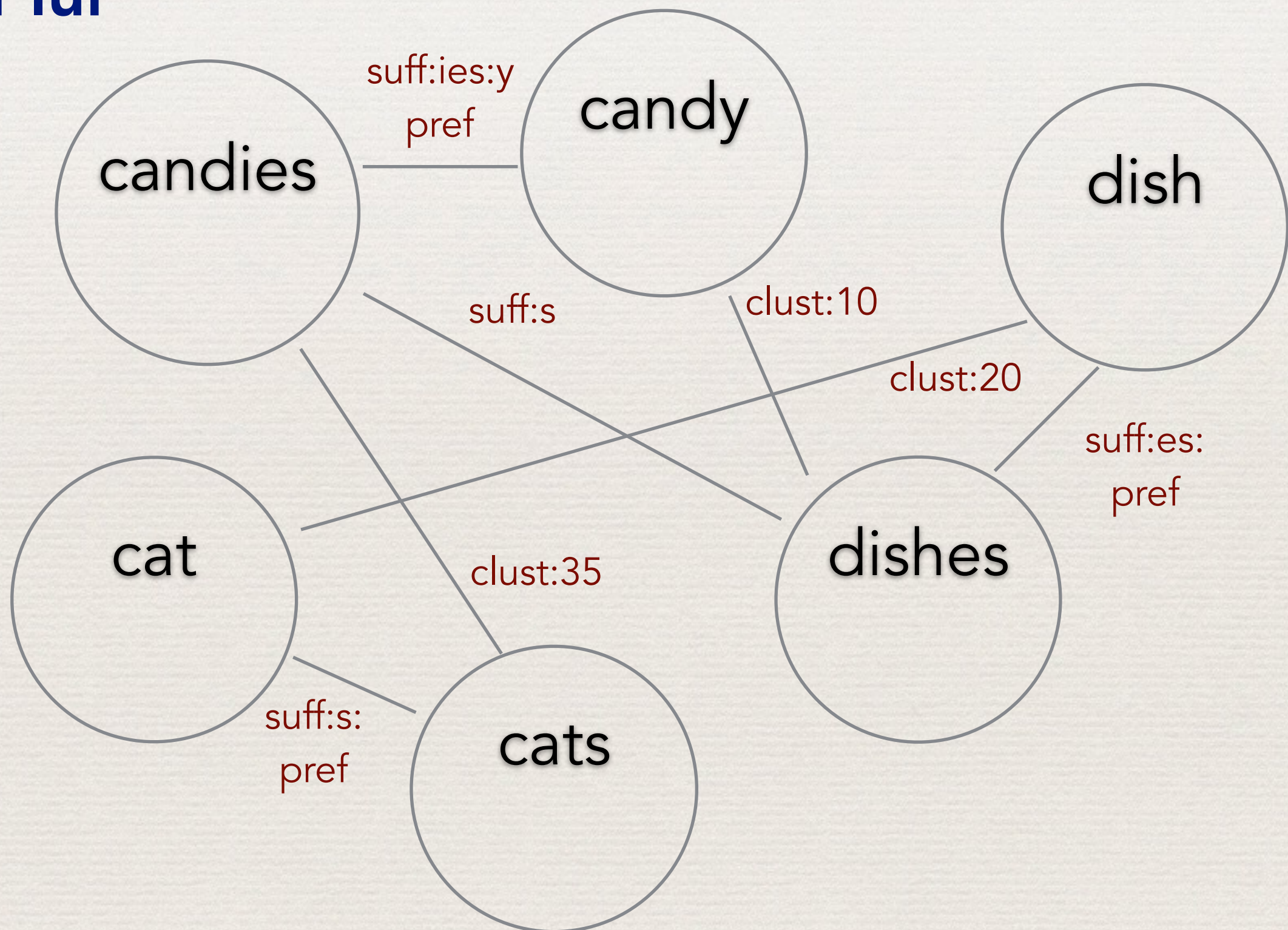


MORPHOSYNTACTIC LEXICONS VIA GRAPH-PROPAGATION



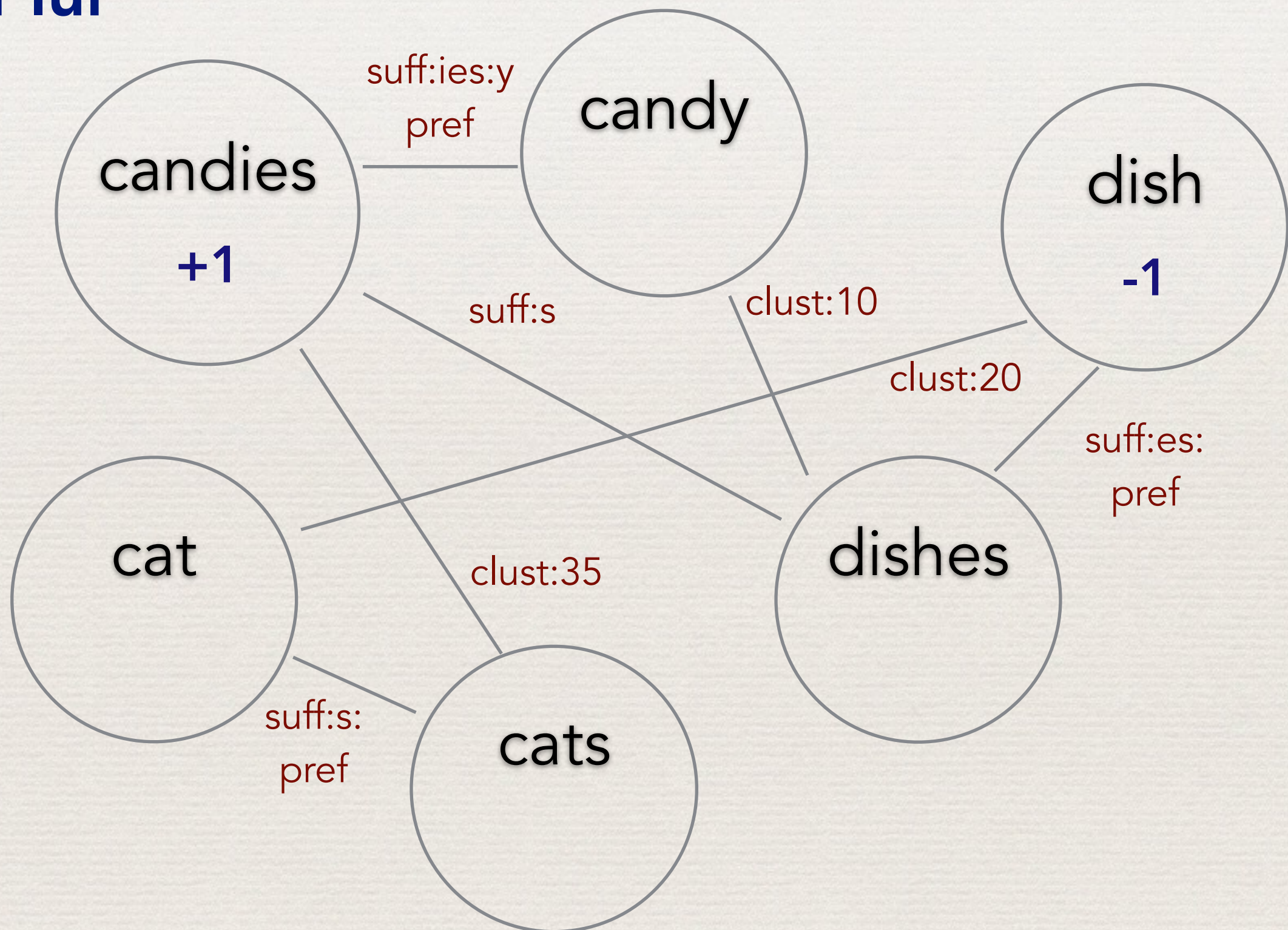
MORPHOSYNTACTIC LEXICONS VIA GRAPH-PROPAGATION

Number=Plur



MORPHOSYNTACTIC LEXICONS VIA GRAPH-PROPAGATION

Number=Plur



MORPHOSYNTACTIC LEXICONS VIA GRAPH-PROPAGATION

Number=Plur

Reinforce (1)

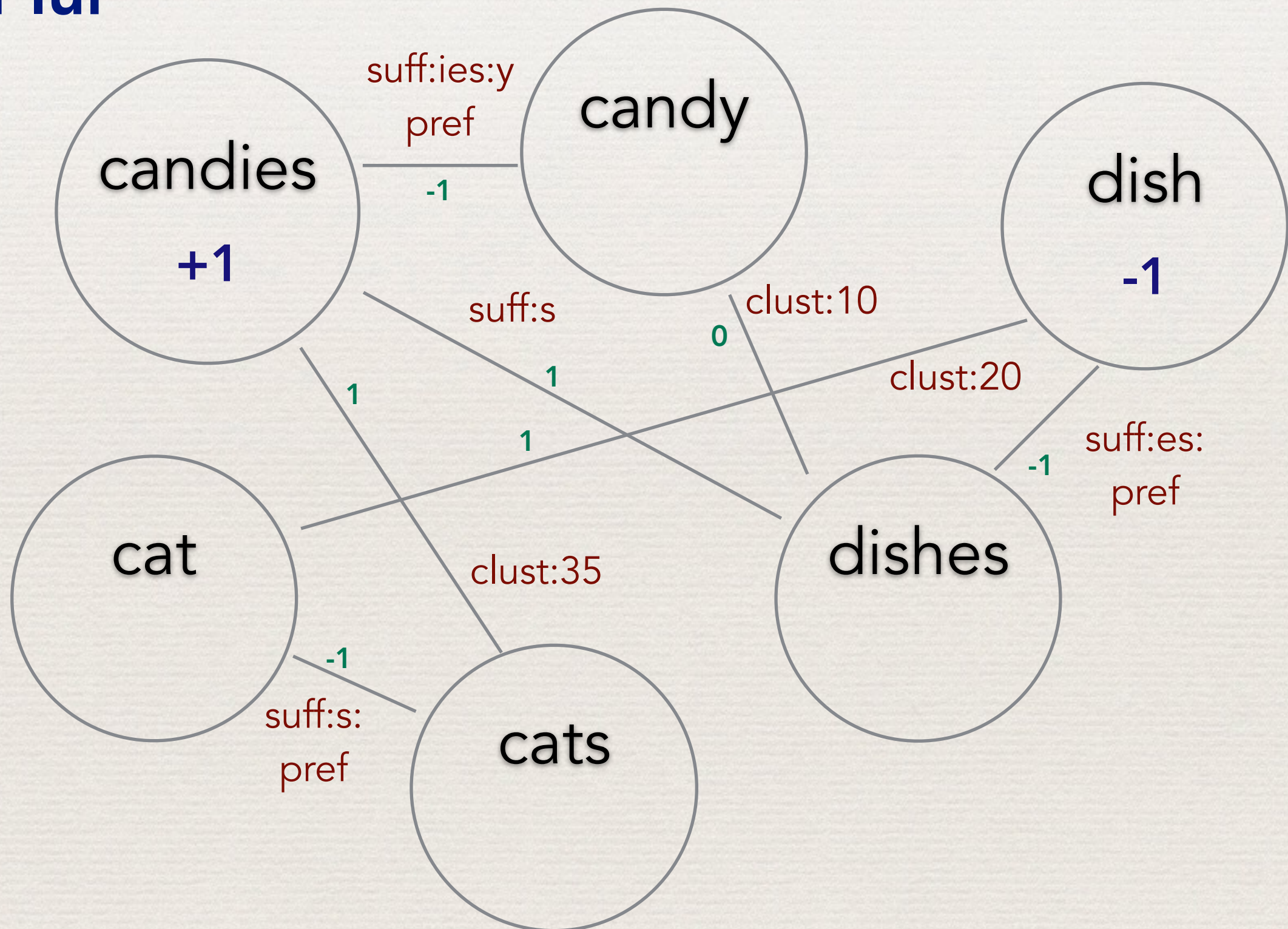
suff:s
clust:35
clust:20

Flip (-1)

suff:ies:y
suff:es:
suff:s:

Neutral (0)

pref
clust:10



MORPHOSYNTACTIC LEXICONS VIA GRAPH-PROPAGATION

Number=Plur

Reinforce (1)

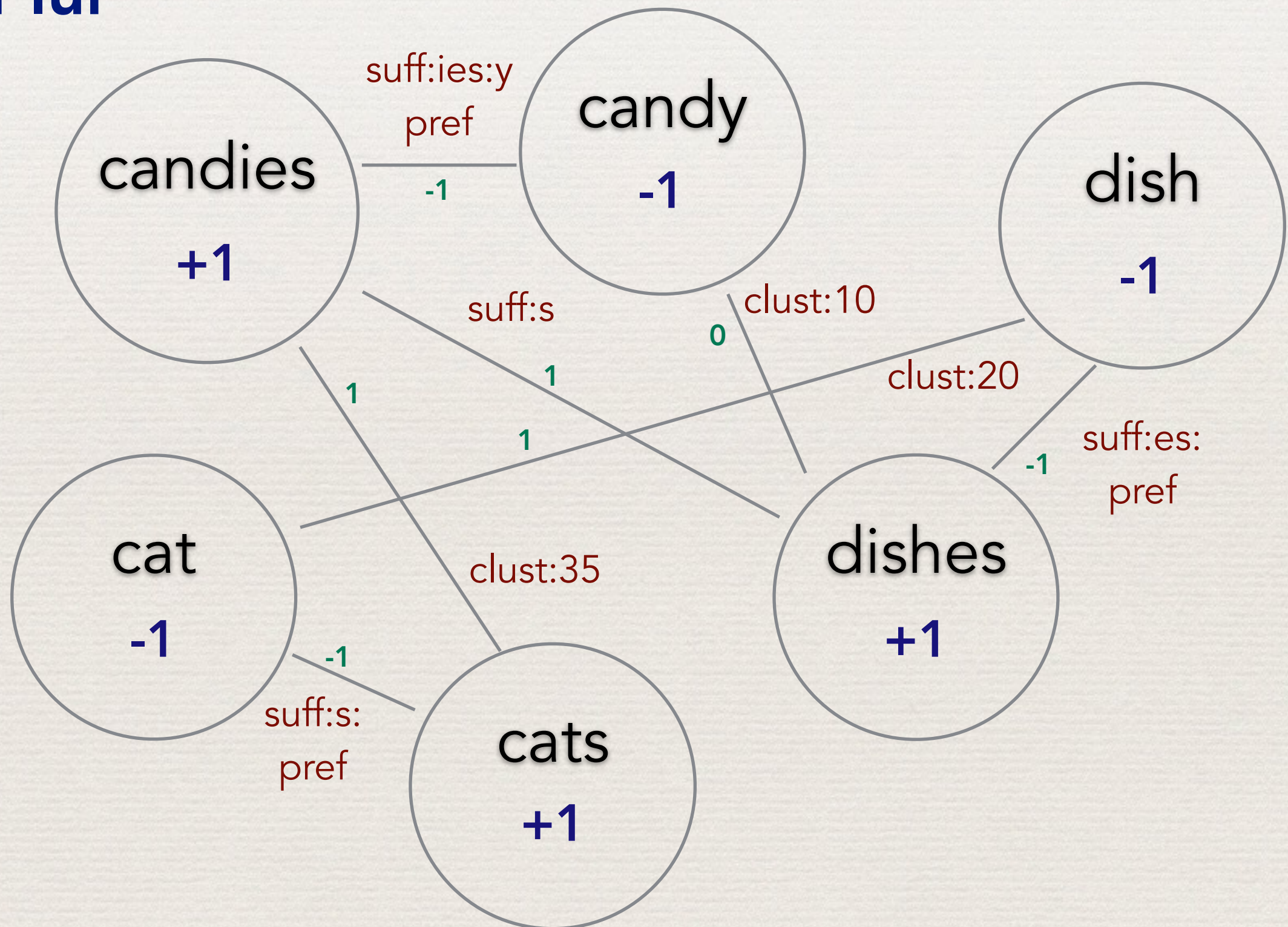
suff:s
clust:35
clust:20

Flip (-1)

suff:ies:y
suff:es:
suff:s:

Neutral (0)

pref
clust:10



MORPHOSYNTACTIC LEXICONS VIA GRAPH-PROPAGATION

Number=Plur

Reinforce (1)

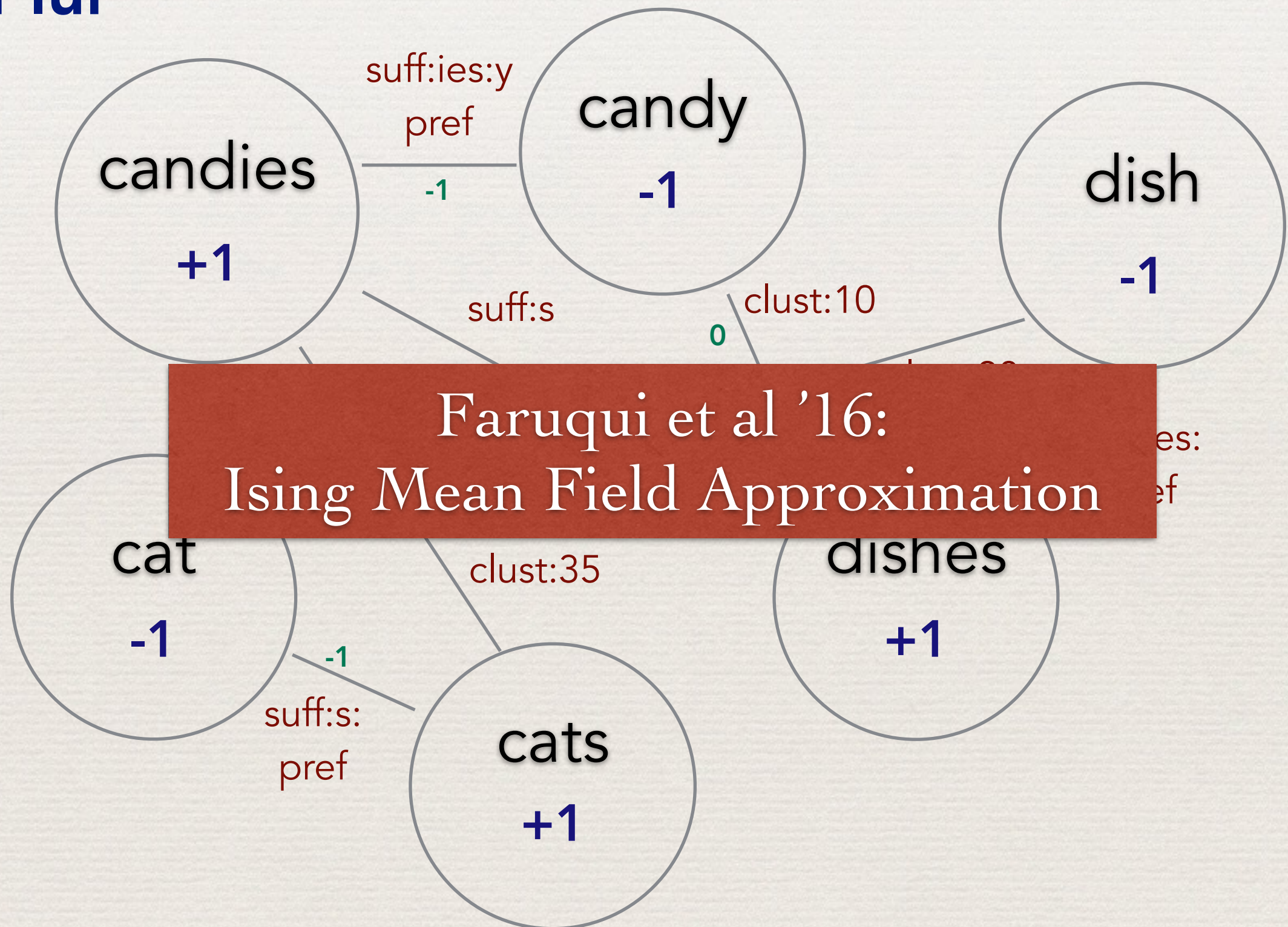
suff:s
clust:35
clust:20

Flip (-1)

suff:ies:y
suff:es:
suff:s:

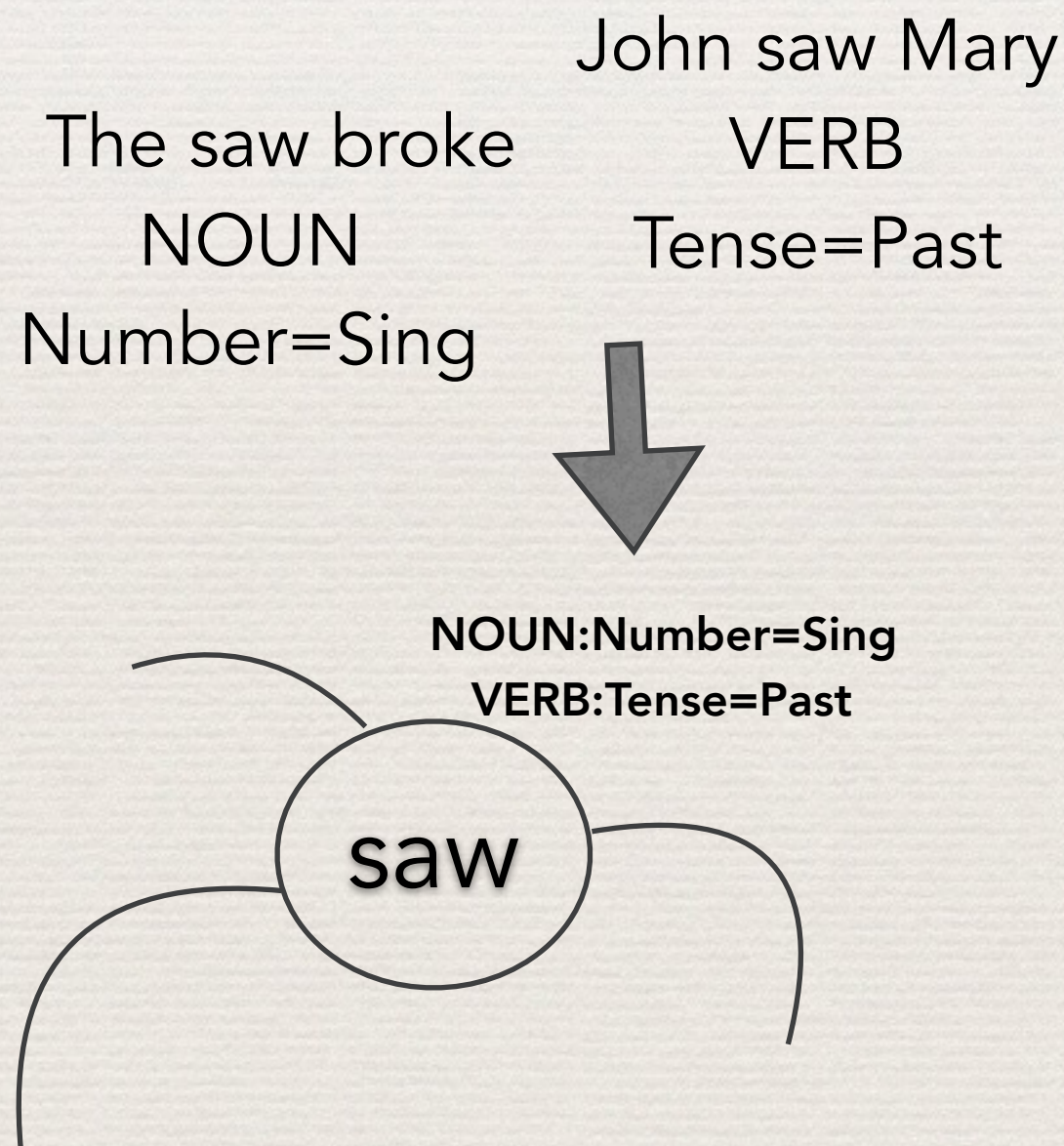
Neutral (0)

pref
clust:10

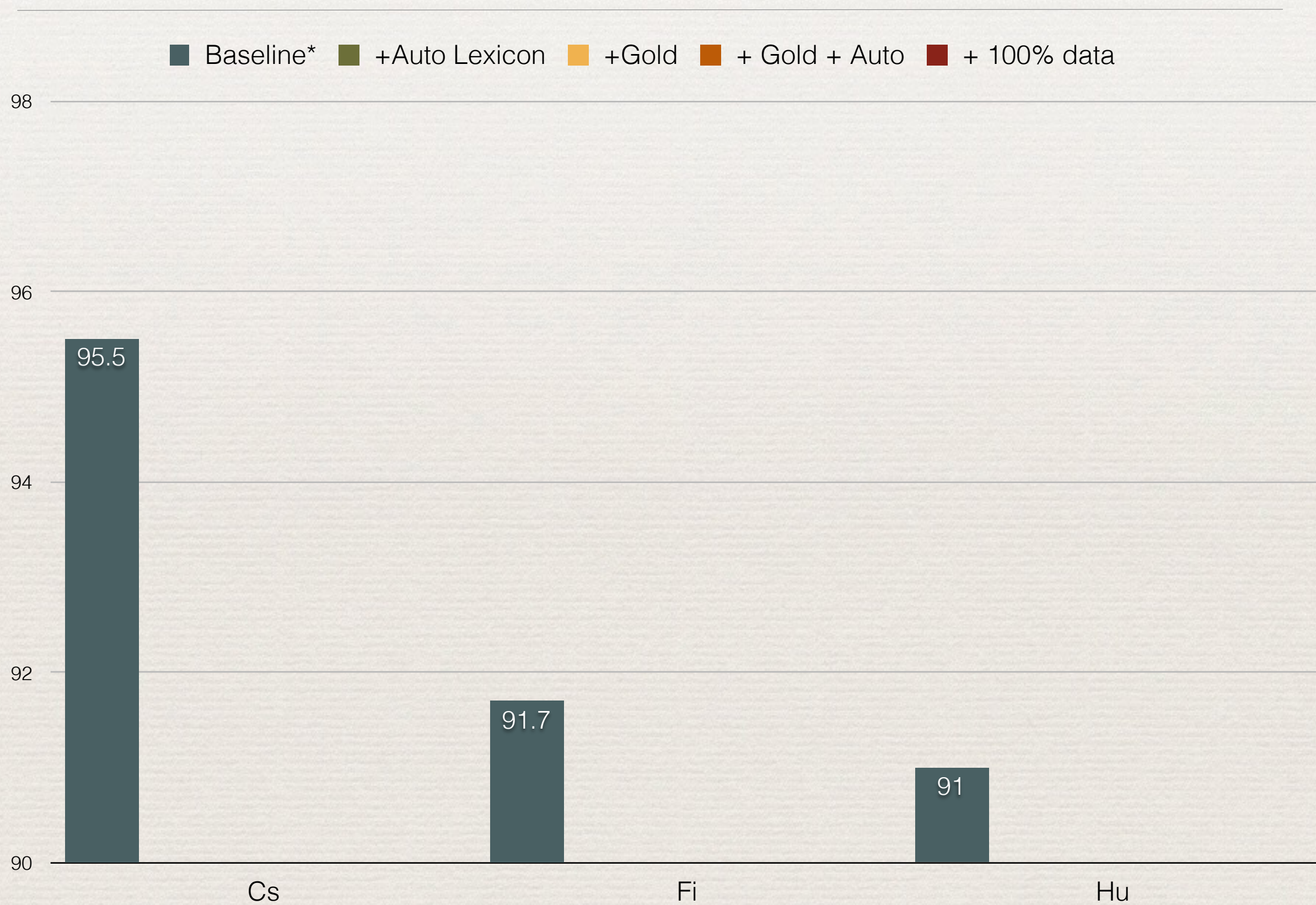


UNIVERSAL LEXICONS

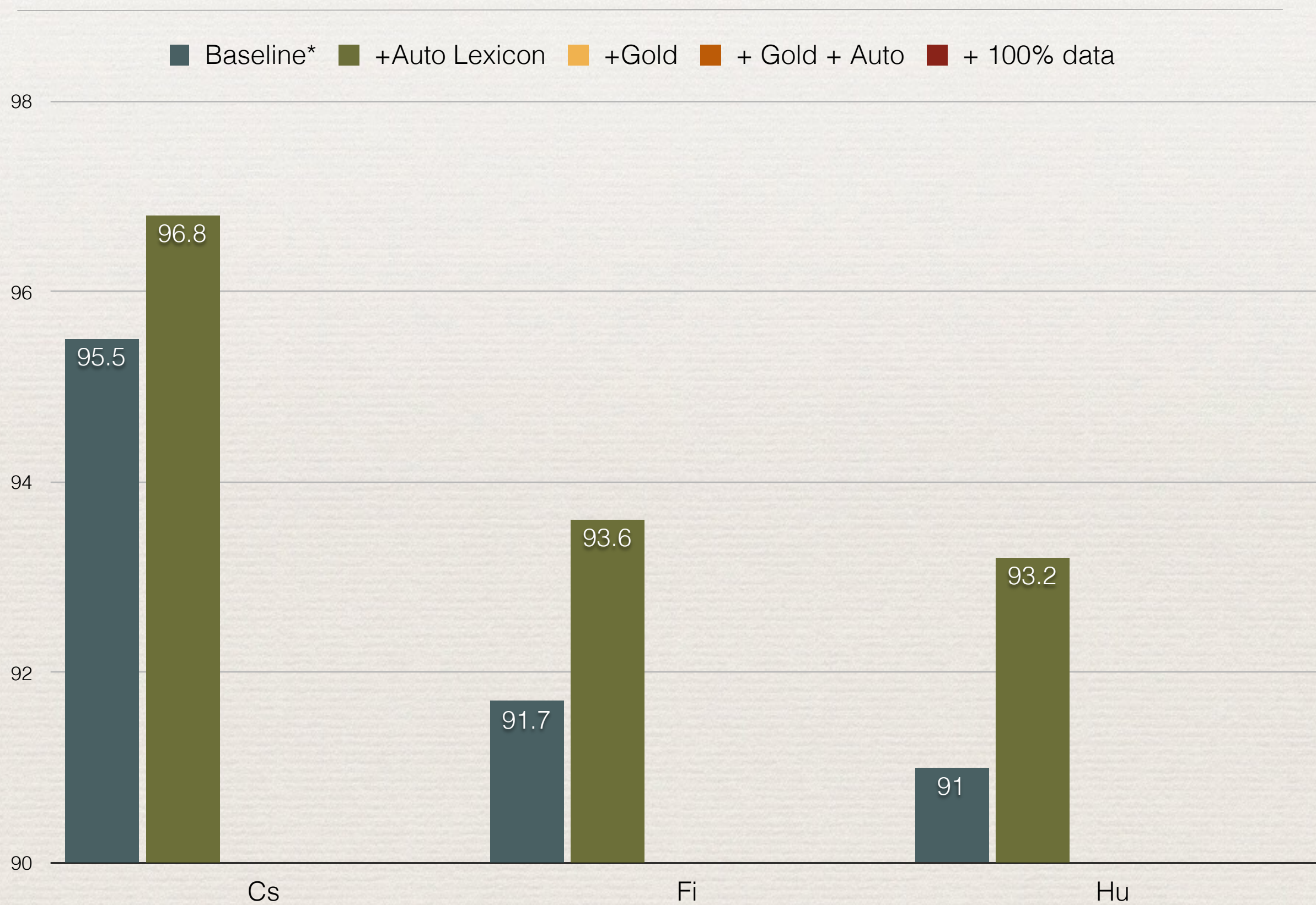
- ❖ Seed with Universal Dependencies (Nivre et al. '16)



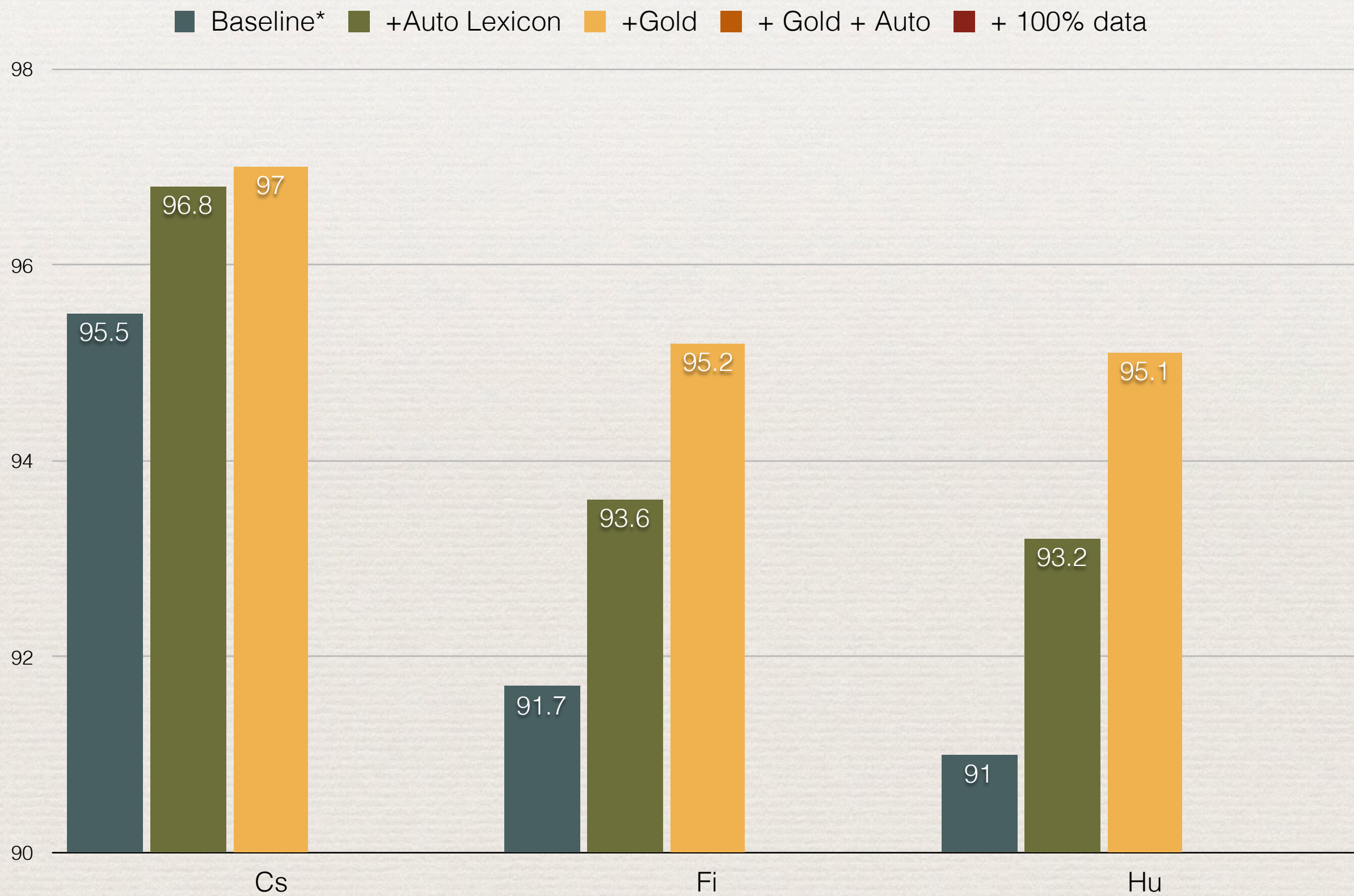
RESOURCE TRADE-OFF



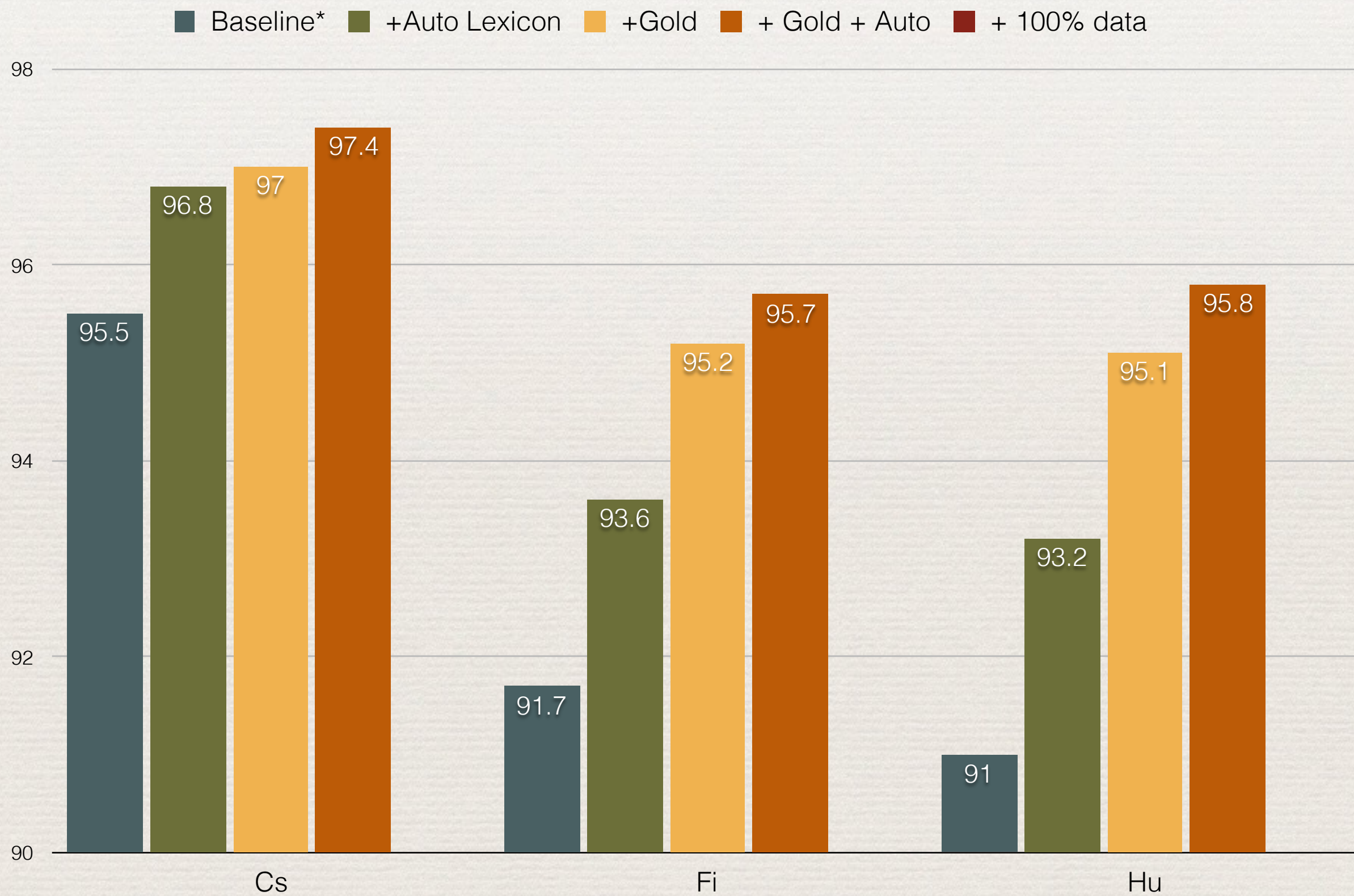
RESOURCE TRADE-OFF



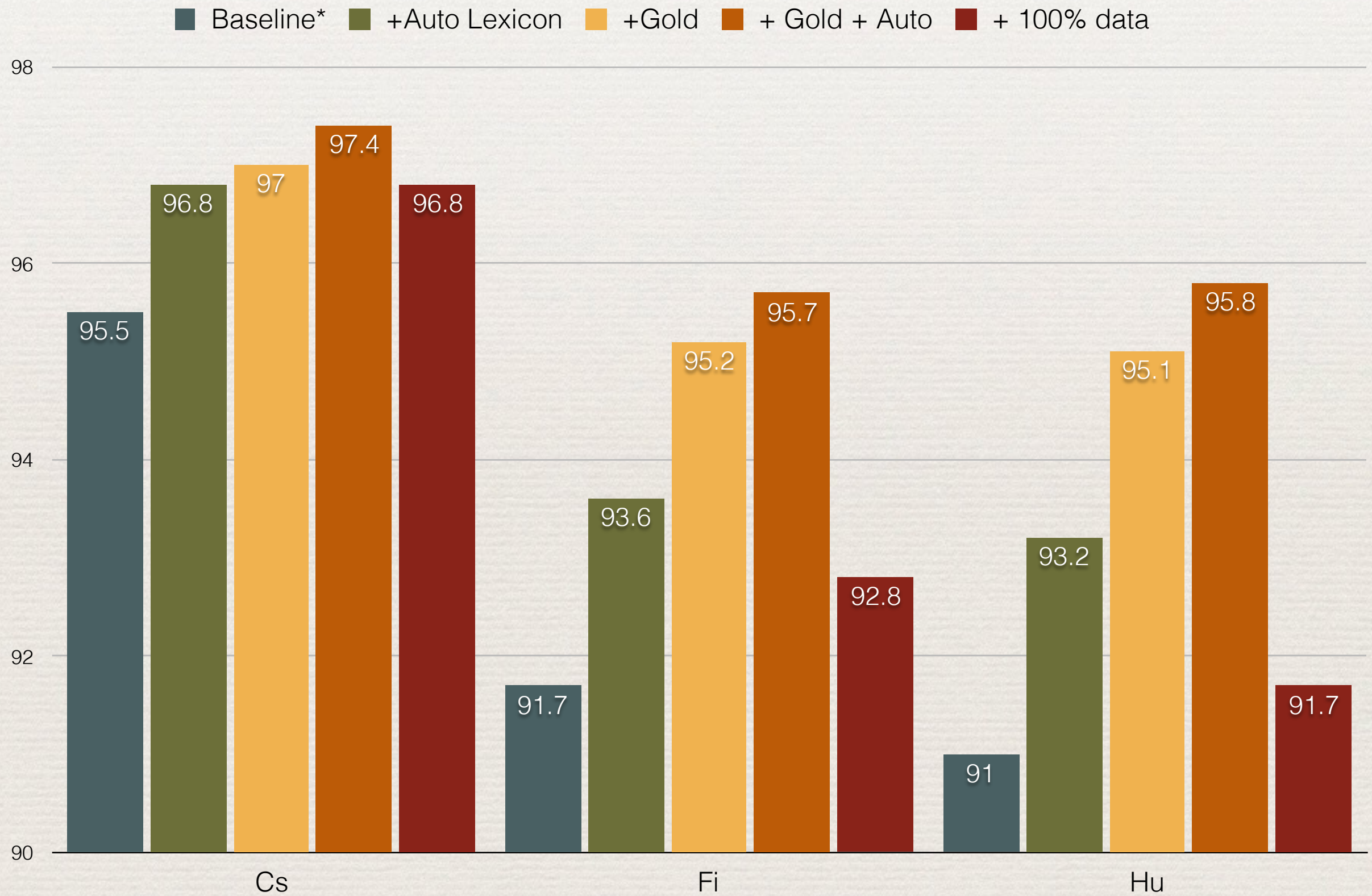
RESOURCE TRADE-OFF



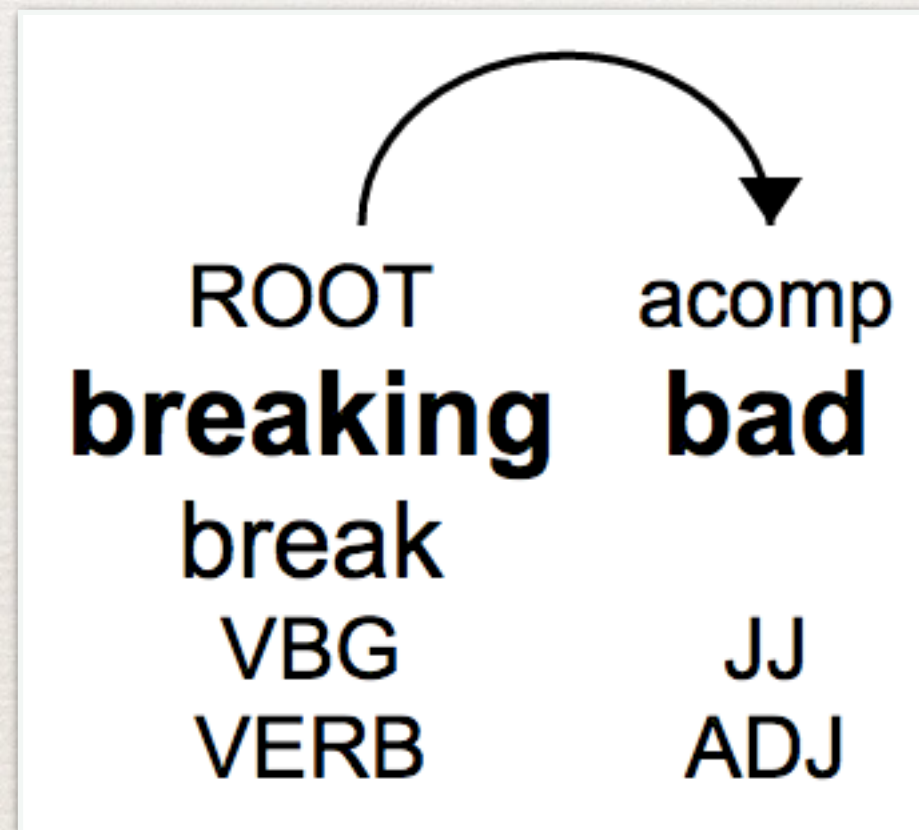
RESOURCE TRADE-OFF



RESOURCE TRADE-OFF



PART-OF-SPEECH TAGGING: QUERIES



SEARCH LOGS



breaking bad



Breaking Bad – AMC - AMC.com

www.amc.com/shows/breaking-bad

The official site for AMC's critically-acclaimed series **Breaking Bad**: Get full episodes, games, videos, plus episode & character guides.

Breaking Bad - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Breaking_Bad

Breaking Bad is an American crime drama television series created and produced by Vince Gilligan. The show originally aired on the AMC network for five ...

[List of Breaking Bad episodes](#) · [Bryan Cranston](#) · [Breaking Bad \(season 5\)](#) · [RJ Mitte](#)

Breaking Bad (@BreakingBad_AMC) | Twitter

https://twitter.com/BreakingBad_AMC

1 day ago - [View on Twitter](#)

Looking at you, Jesse. Watch @NightManagerAMC now before Tuesday's Finale: bit.ly/1SZdnc8
pic.twitter.com/rE5sRL3sJ...

2 days ago - [View on Twitter](#)

Watch new show @PreacherAMC's open, from one of the minds behind @BreakingBad_AMC. bit.ly/1Tit3YH #Preacher
pic.twitter.com/D7WddMwM9...

Breaking Bad (TV Series 2008–2013) - IMDb

www.imdb.com/title/tt0903747/

★★★★★ Rating: 9.5/10 - 847,508 votes

Breaking Bad — Events set in motion long ago move toward a conclusion. Photos. Betsy Brandt at **Breaking Bad** (2008) Luis Moncada and Daniel Moncada in ...

[Full Cast & Crew](#) · [62 Episodes](#) · [Season 2](#) · [Season 4](#)



Breaking Bad

American drama series

9.5/10
IMDb

95%
Rotten Tomatoes

Mild-mannered high school chemistry teacher Walter White thinks his life can't get much worse. His salary barely makes ends meet, a situation not likely to improve once his pregnant wife gives birth, and their teenage son is battling cerebral palsy. But Walter is dumbstruck when he learns he has ter... [More](#)

Final episode date: September 29, 2013

Spin-off: [Better Call Saul](#)

Awards: [Primetime Emmy Award for Outstanding Drama Series](#), [More](#)

SEARCH LOGS



breaking bad



Click

Breaking Bad – AMC - AMC.com

www.amc.com/shows/breaking-bad

The official site for AMC's critically-acclaimed series **Breaking Bad**: Get full episodes, games, videos, plus episode & character guides.

Breaking Bad - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/Breaking_Bad

Breaking Bad is an American crime drama television series created and produced by Vince Gilligan. The show originally aired on the AMC network for five ...

[List of Breaking Bad episodes](#) · [Bryan Cranston](#) · [Breaking Bad \(season 5\)](#) · [RJ Mitte](#)

Breaking Bad (@BreakingBad_AMC) | Twitter

https://twitter.com/BreakingBad_AMC

1 day ago - [View on Twitter](#)

Looking at you, Jesse. Watch @NightManagerAMC now before Tuesday's Finale: bit.ly/1SZdnc8
pic.twitter.com/rE5sRL3sJ...

2 days ago - [View on Twitter](#)

Watch new show @PreacherAMC's open, from one of the minds behind @BreakingBad_AMC. bit.ly/1Tit3YH #Preacher
pic.twitter.com/D7WddMwM9...

Breaking Bad (TV Series 2008–2013) - IMDb

www.imdb.com/title/tt0903747/

★★★★★ Rating: 9.5/10 - 847,508 votes

Breaking Bad — Events set in motion long ago move toward a conclusion. Photos. Betsy Brandt at **Breaking Bad** (2008) Luis Moncada and Daniel Moncada in ...

[Full Cast & Crew](#) · [62 Episodes](#) · [Season 2](#) · [Season 4](#)



Breaking Bad

American drama series

9.5/10
IMDb

95%
Rotten Tomatoes

Mild-mannered high school chemistry teacher Walter White thinks his life can't get much worse. His salary barely makes ends meet, a situation not likely to improve once his pregnant wife gives birth, and their teenage son is battling cerebral palsy. But Walter is dumbstruck when he learns he has ter... [More](#)

Final episode date: September 29, 2013

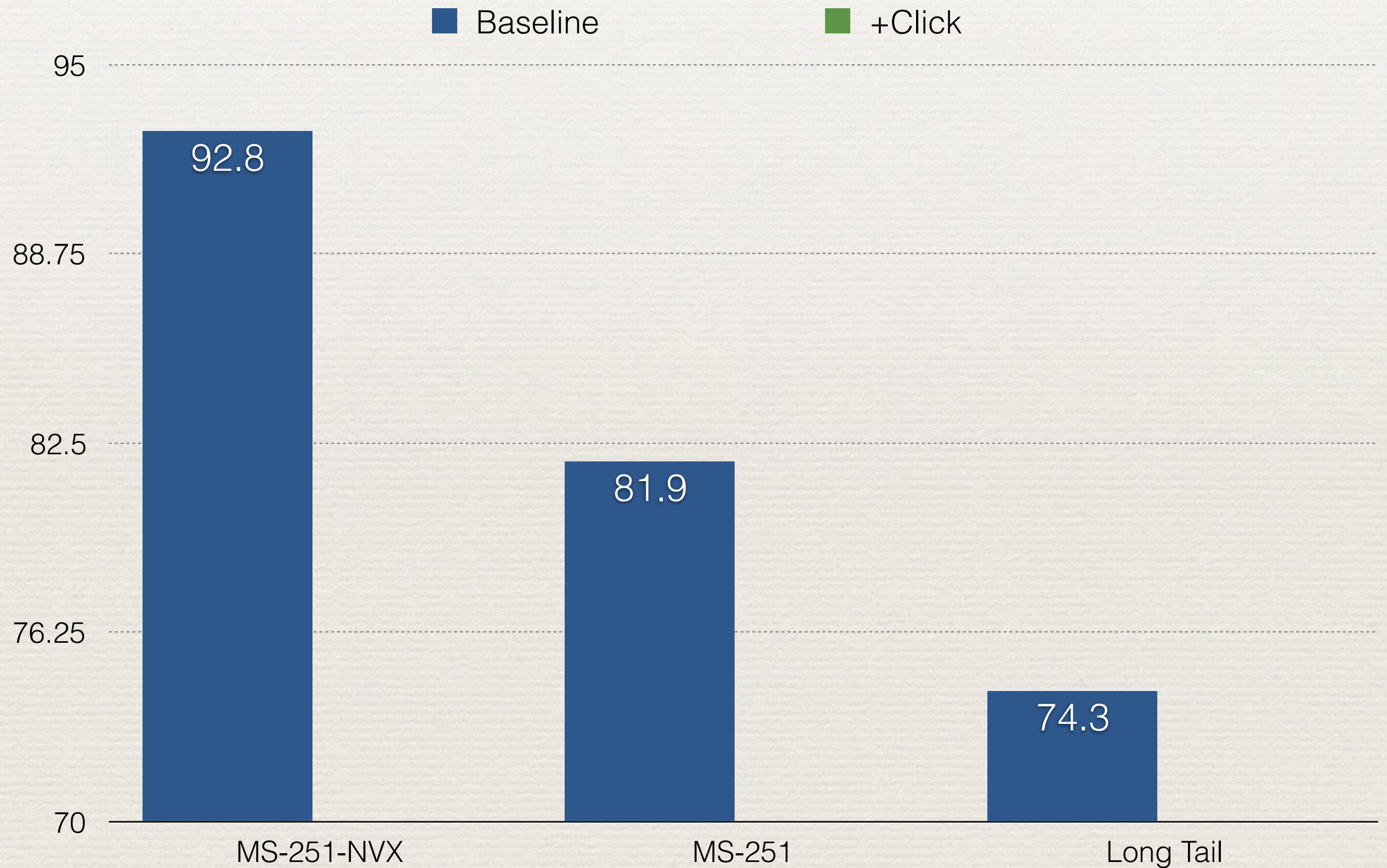
Spin-off: [Better Call Saul](#)

Awards: [Primetime Emmy Award for Outstanding Drama Series](#), [More](#)



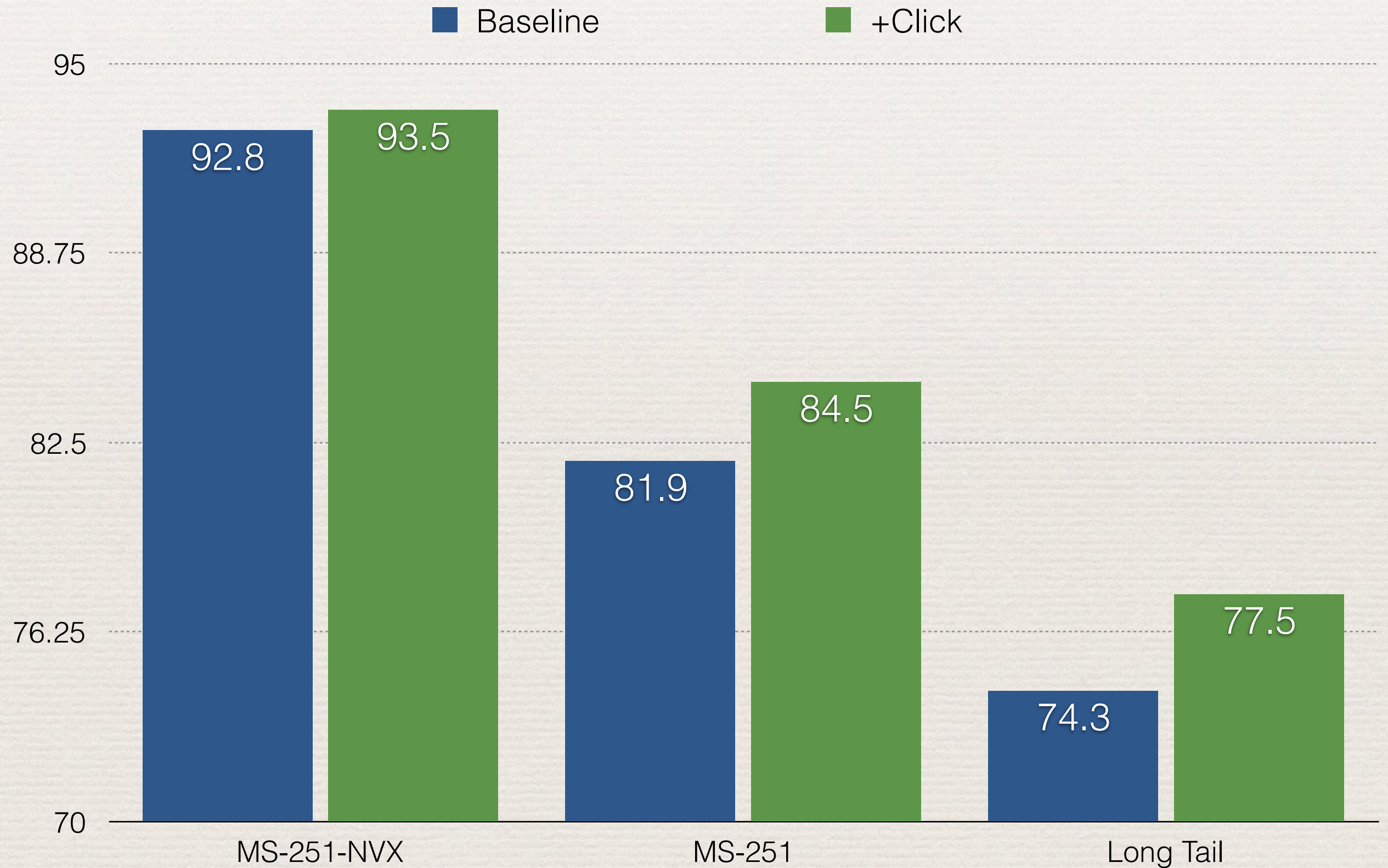
POS TAGGERS FROM CLICKS

GANCHEV ET AL. (2012)



POS TAGGERS FROM CLICKS

GANCHEV ET AL. (2012)



MORPHOSYNTAX CONCLUSIONS

- ❖ Money on more supervised data not necessarily optimal
- ❖ Better alternative: lexical resources (auto, manual & both)
- ❖ Better alternative: correlate usage statistics (click logs)



End User: Machine Translation

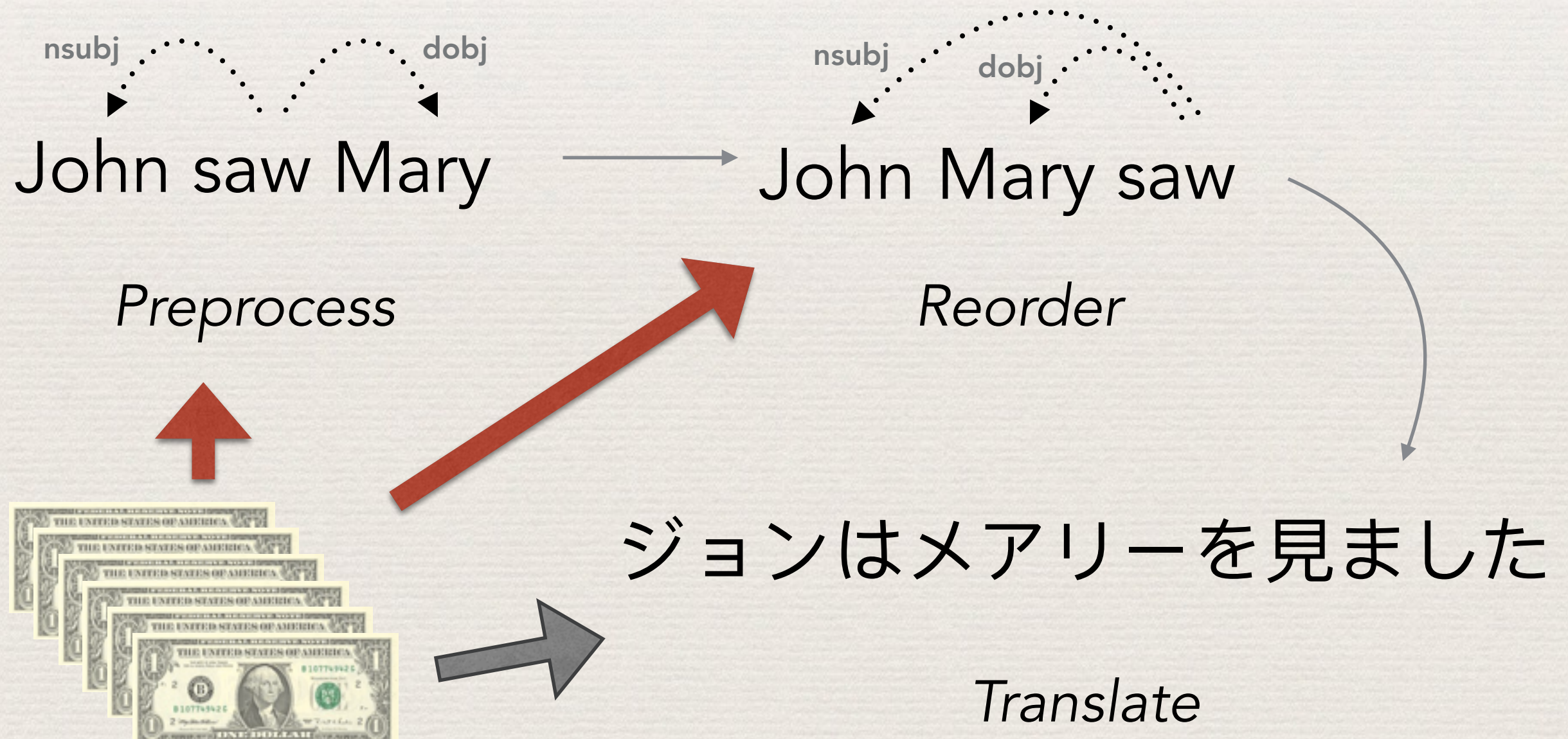
PIPELINED MACHINE TRANSLATION

Preprocess → Reorder → Translate → Postorder → Postprocess



PIPELINED MACHINE TRANSLATION

Preprocess → Reorder → Translate → Postorder → Postprocess



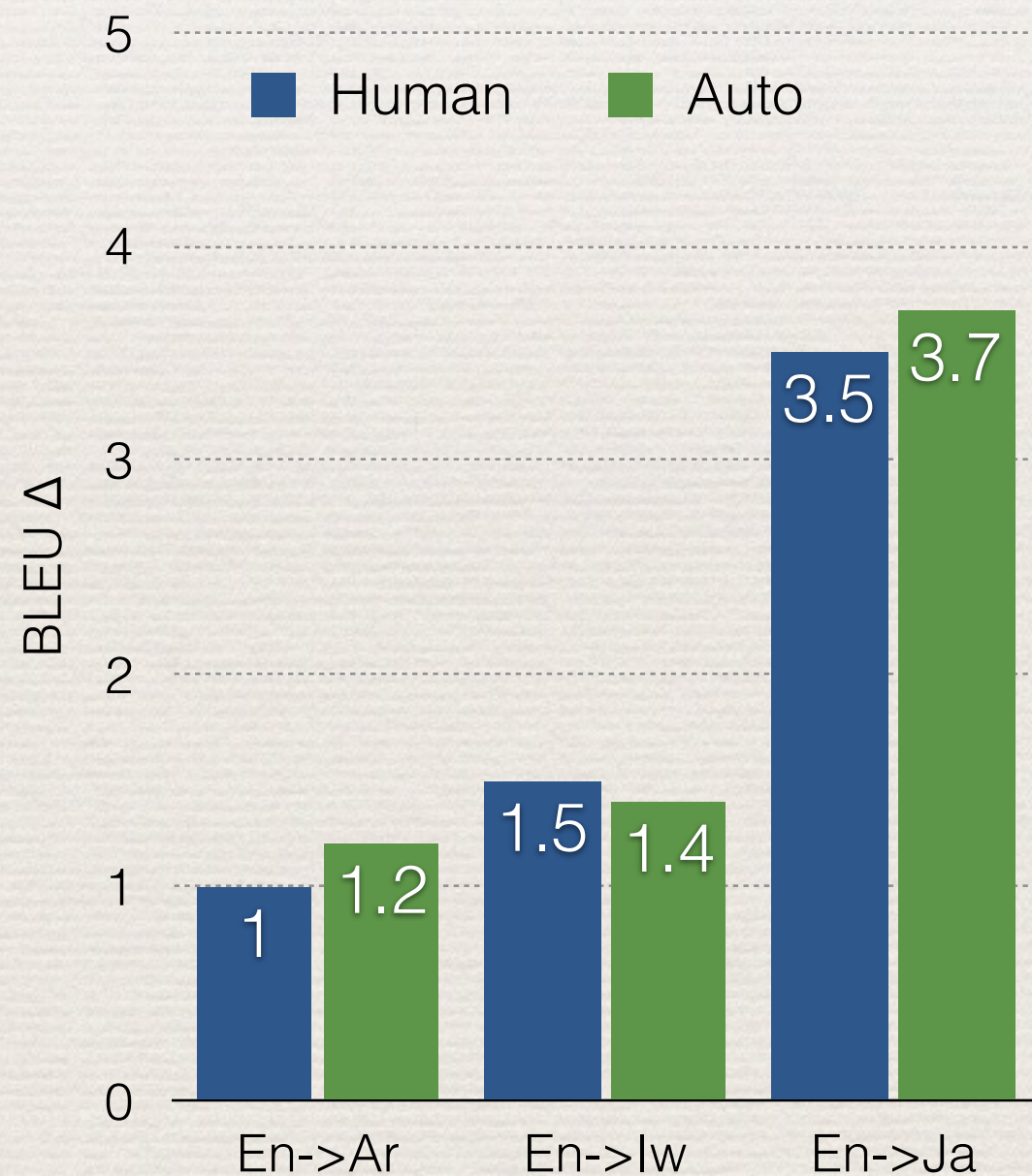
REORDERING DATA



VS.



Syntax-based reordered
(Lerner & Petrov '13)



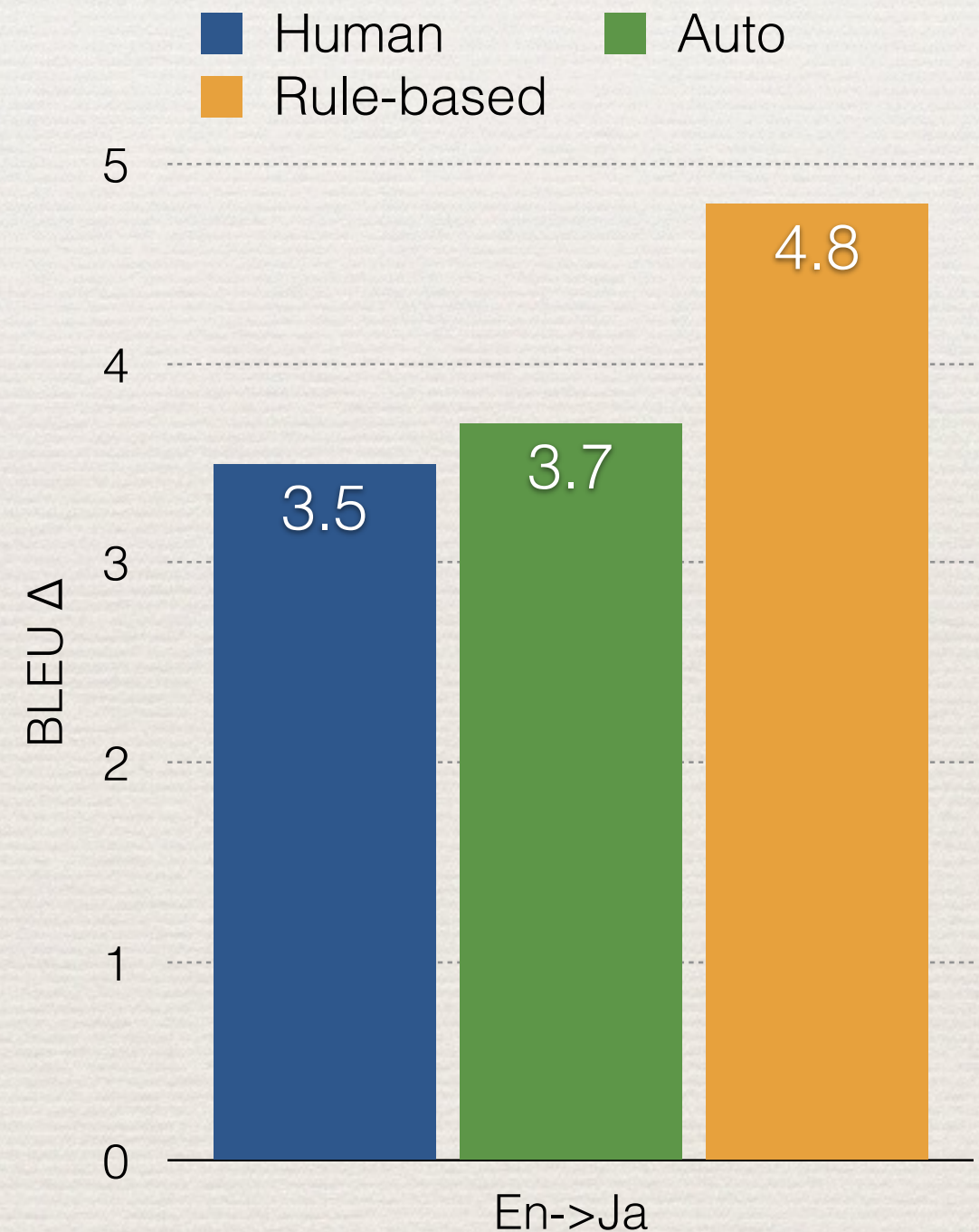
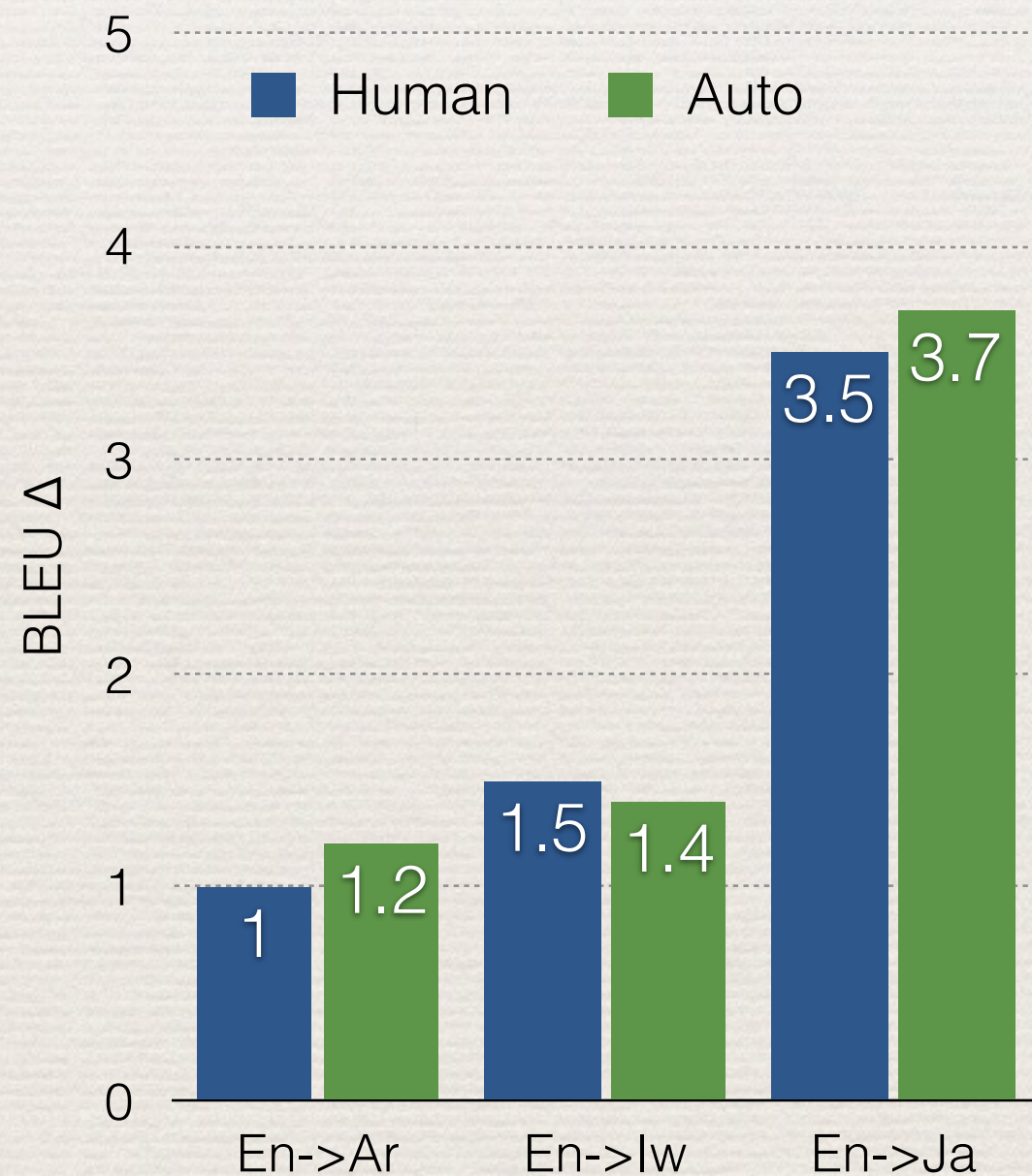
REORDERING DATA



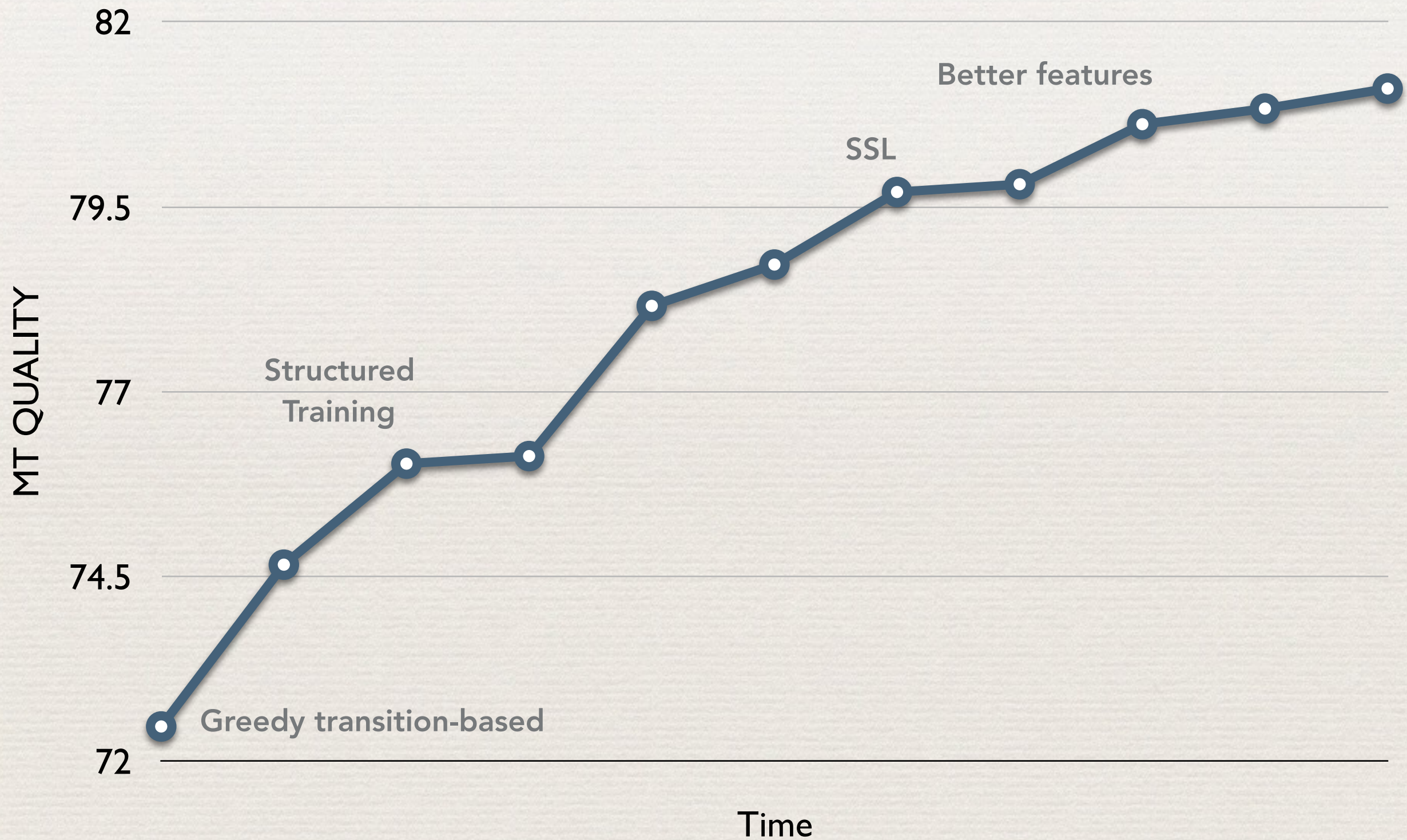
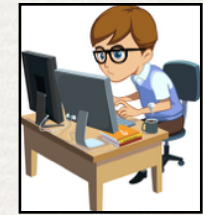
VS.



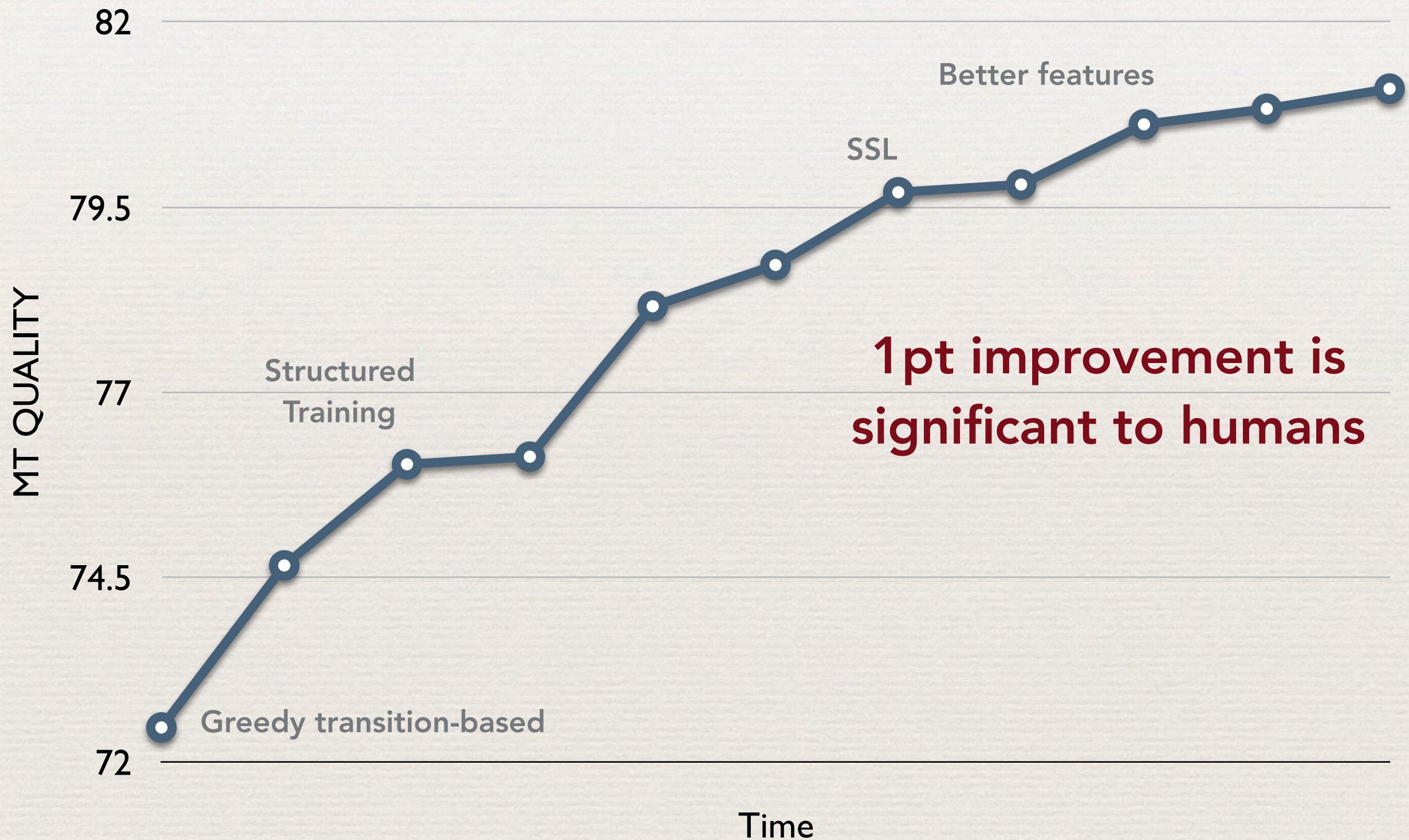
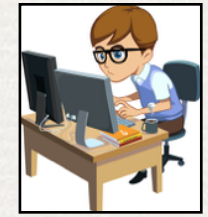
Syntax-based reordered
(Lerner & Petrov '13)



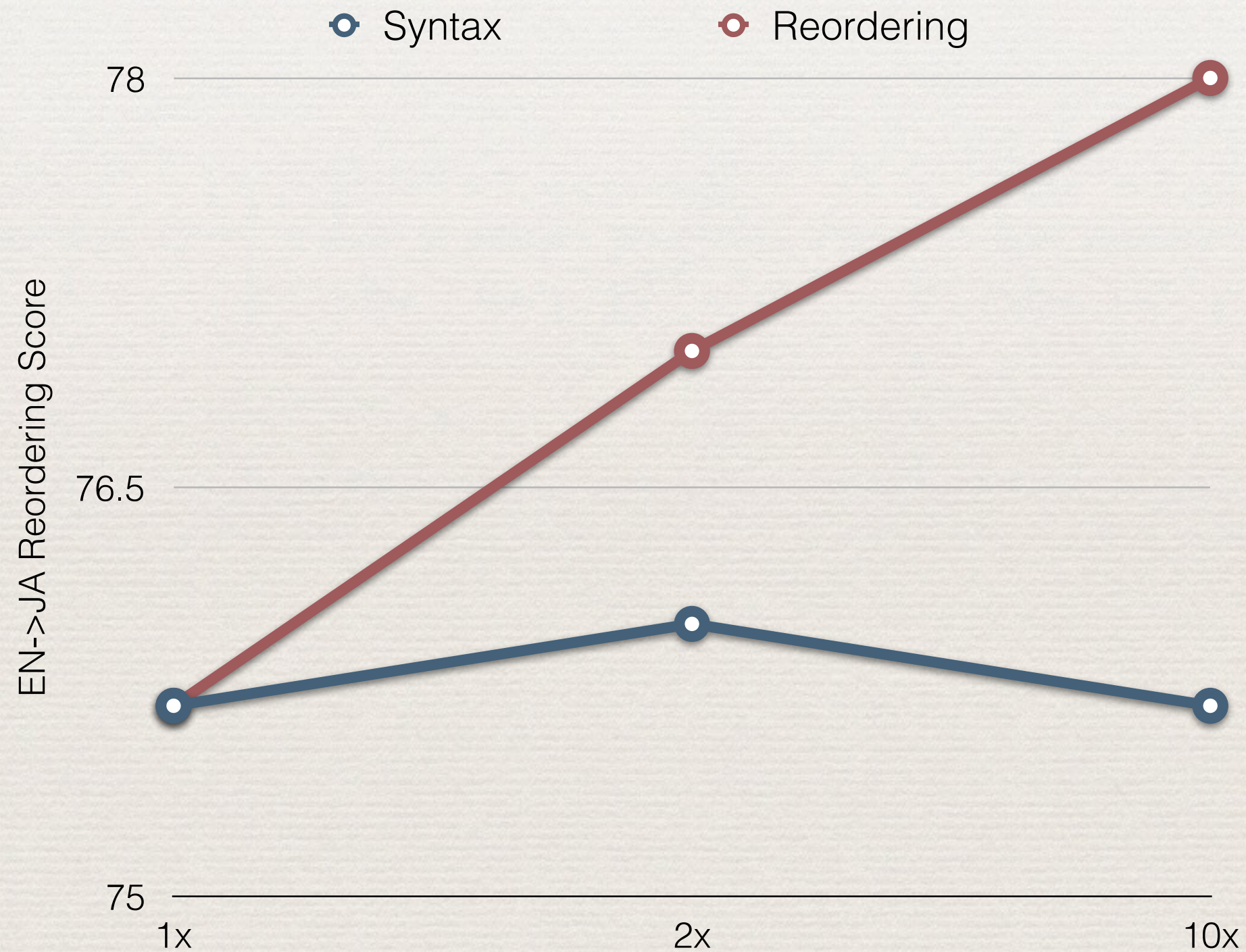
BETTER PARSERS



BETTER PARSERS



MORE DATA



MACHINE TRANSLATION



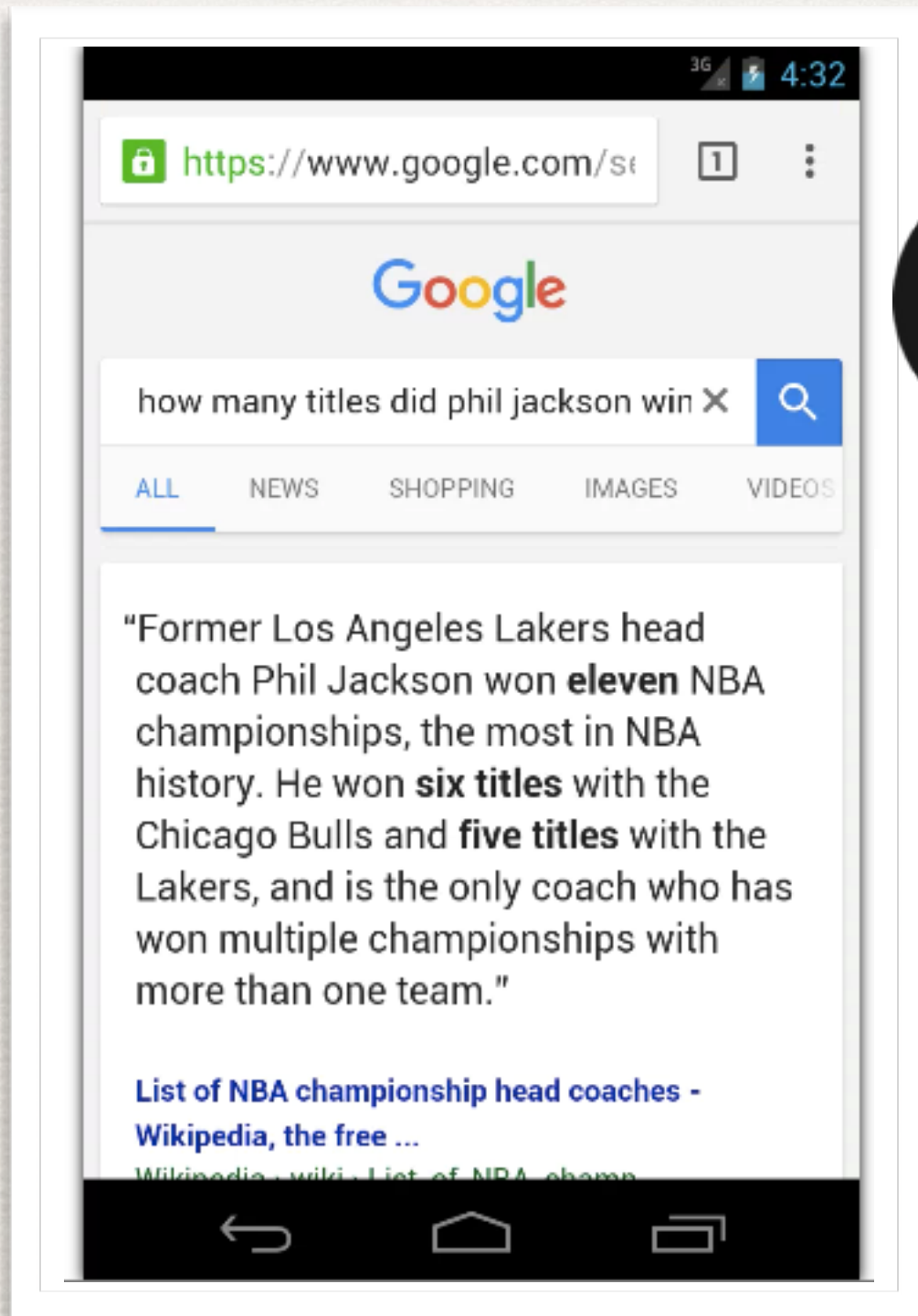
vs.



- ❖ Human vs. auto data: about the same
- ❖ Human models sometimes better than learned
- ❖ Better parsing models = better translation
- ❖ Better to **spend on targeted resources** — reordering

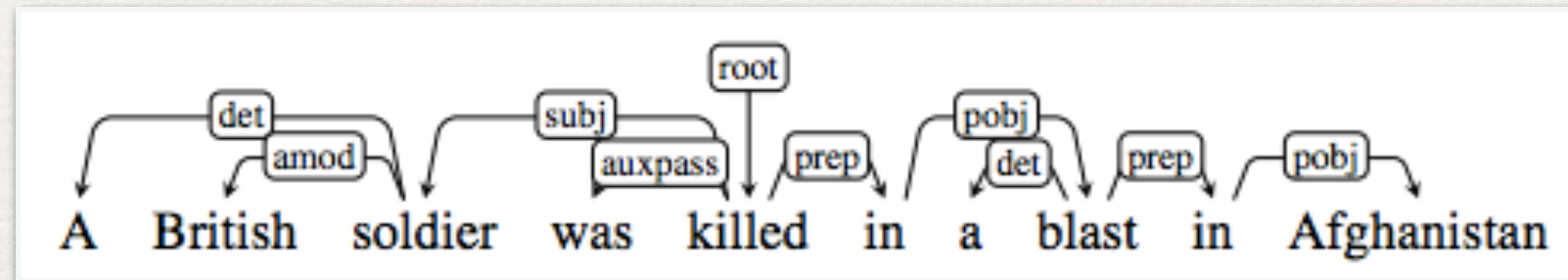
End User: Sentence Compression

SENTENCE COMPRESSION @GOOGLE



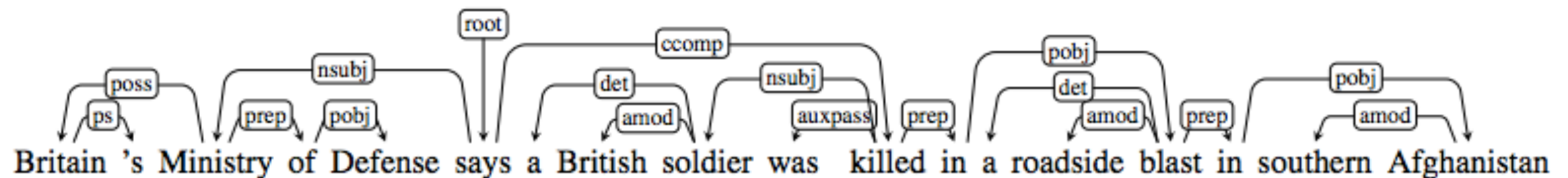
Former Los Angeles Lakers head coach Phil Jackson won eleven NBA championships. He won six titles with the Chicago Bulls and five titles with the Lakers.

GOOGLE NEWS



Headline

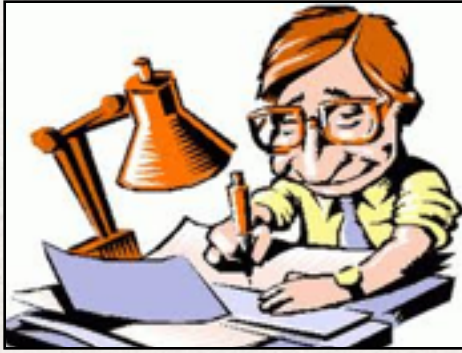
First sentence



Filippova & Altun '13

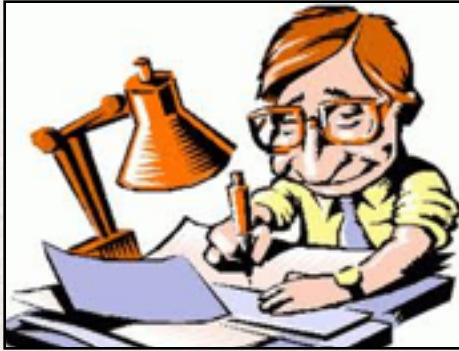
- ❖ Can extract millions of pairs
- ❖ Quality \sim expert annotations
- ❖ 81.4 \rightarrow 84.3 F1 (10% \rightarrow 100% data)

NEED FOR HIGH QUALITY ANNOTATIONS?



?

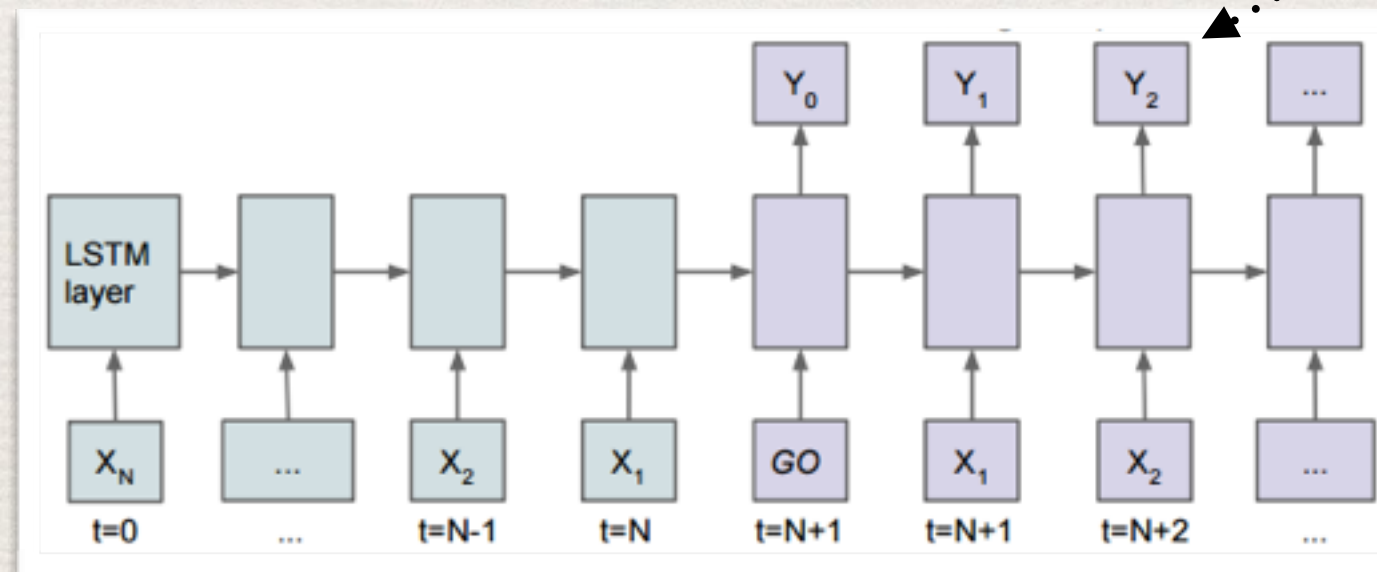
NEED FOR HIGH QUALITY ANNOTATIONS?



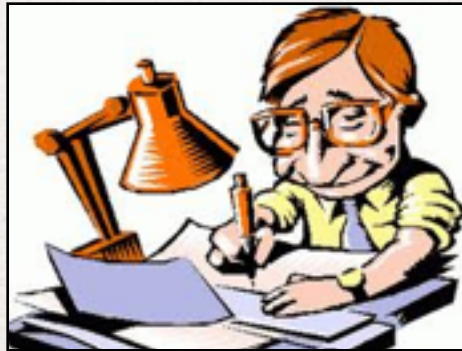
?

Filippova et al. '15:
LSTM compression
by deletion

1/0



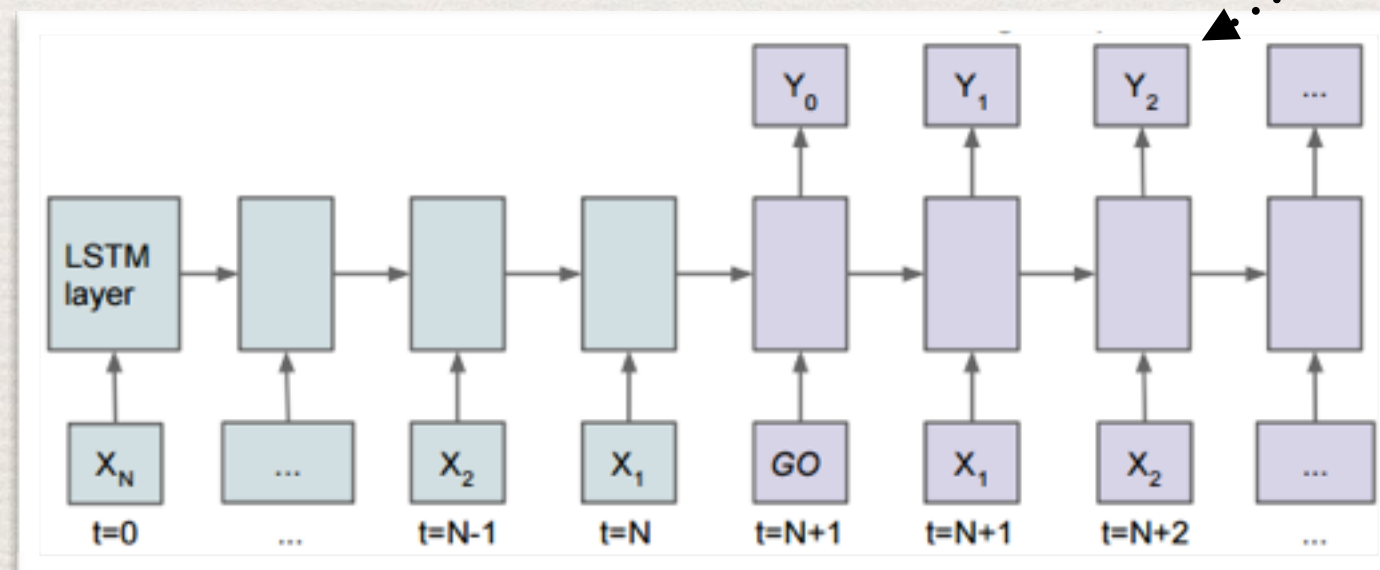
NEED FOR HIGH QUALITY ANNOTATIONS?



?

Filippova et al. '15:
LSTM compression
by deletion

1/0



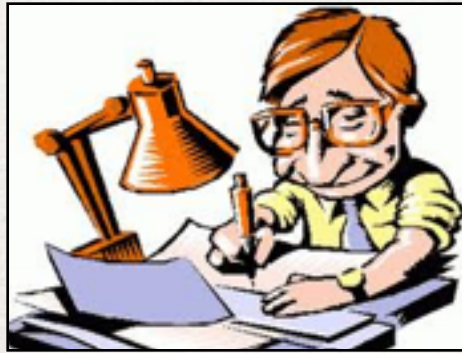
word

[] [] [] []

30

Accuracy

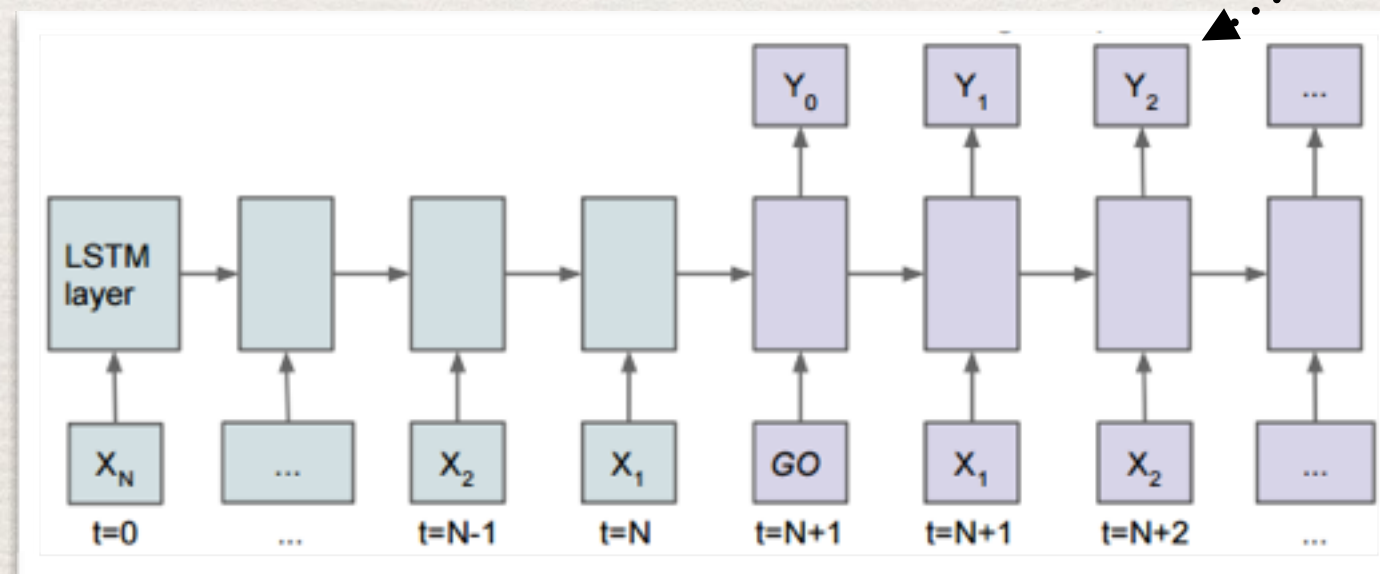
NEED FOR HIGH QUALITY ANNOTATIONS?



?

Filippova et al. '15:
LSTM compression
by deletion

1/0



word
parent

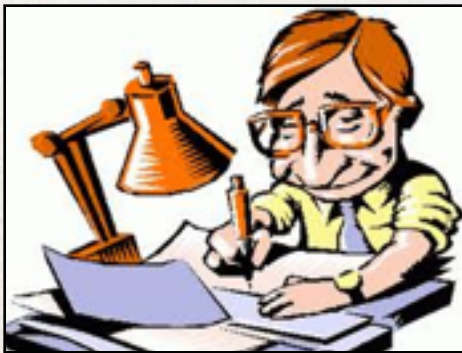
$\begin{bmatrix} \text{ } \\ \text{ } \end{bmatrix}$ $\begin{bmatrix} \text{ } \\ \text{ } \end{bmatrix}$ $\begin{bmatrix} \text{ } \\ \text{ } \end{bmatrix}$ $\begin{bmatrix} \text{ } \\ \text{ } \end{bmatrix}$

30

31

Accuracy

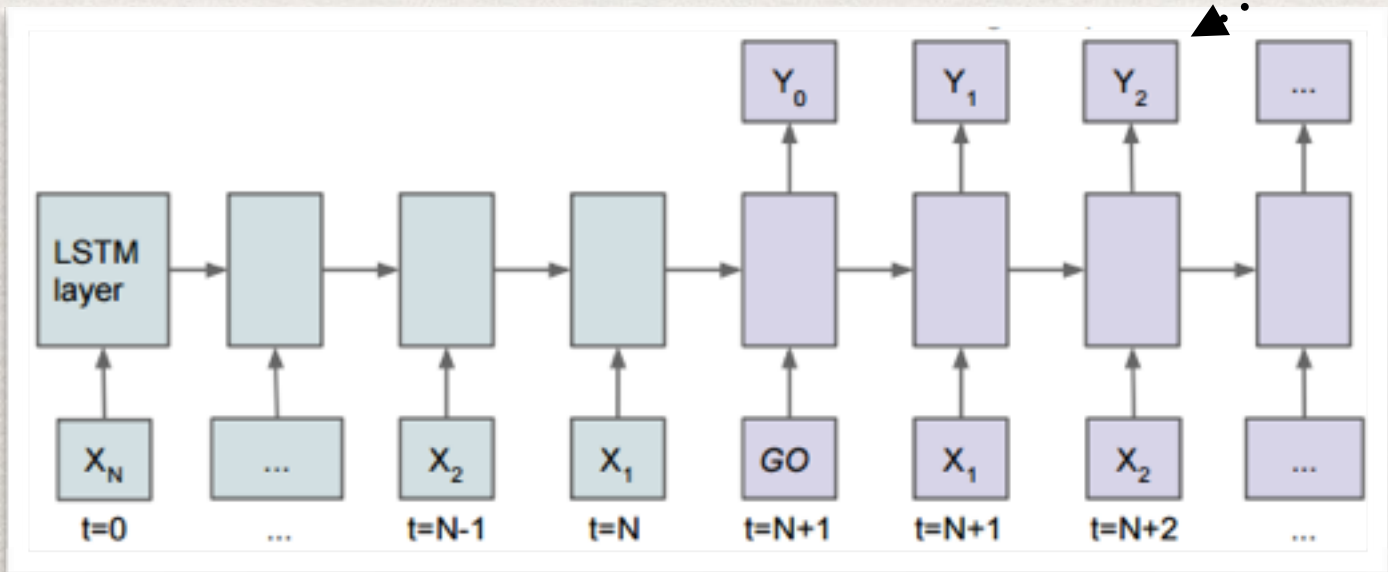
NEED FOR HIGH QUALITY ANNOTATIONS?



?

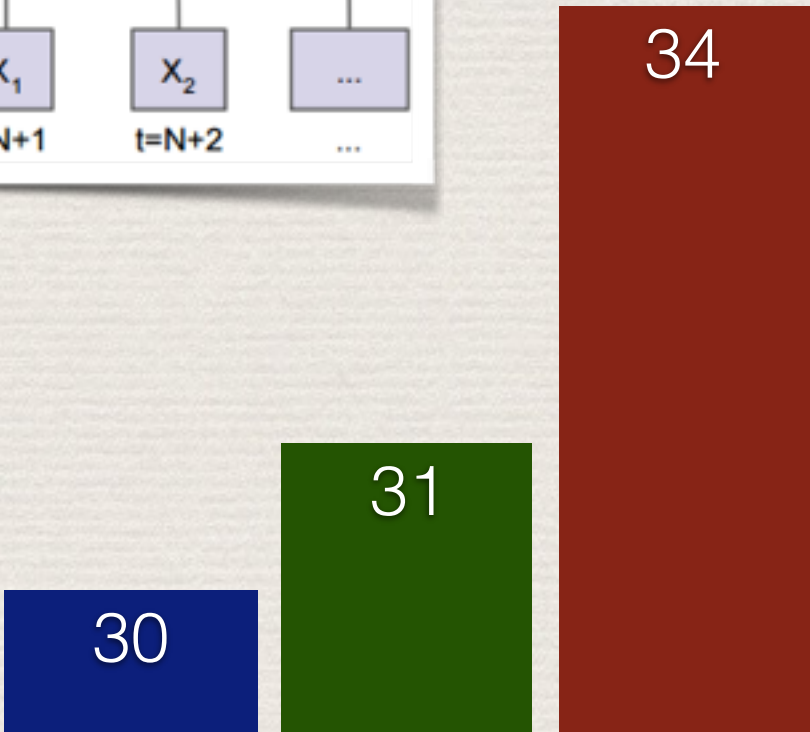
Filippova et al. '15:
LSTM compression
by deletion

1/0



word
parent
syn struct

[]	[]	[]	[]
[]	[]	[]	[]
[]	[]	[]	[]



Accuracy

THE RESOURCE TRADE-OFF

Data + model

Data + model



High quality
annotations



Crowd-sourced




Auto resources



End User: QA & Knowledge Extraction

QA & KNOWLEDGE EXTRACTION






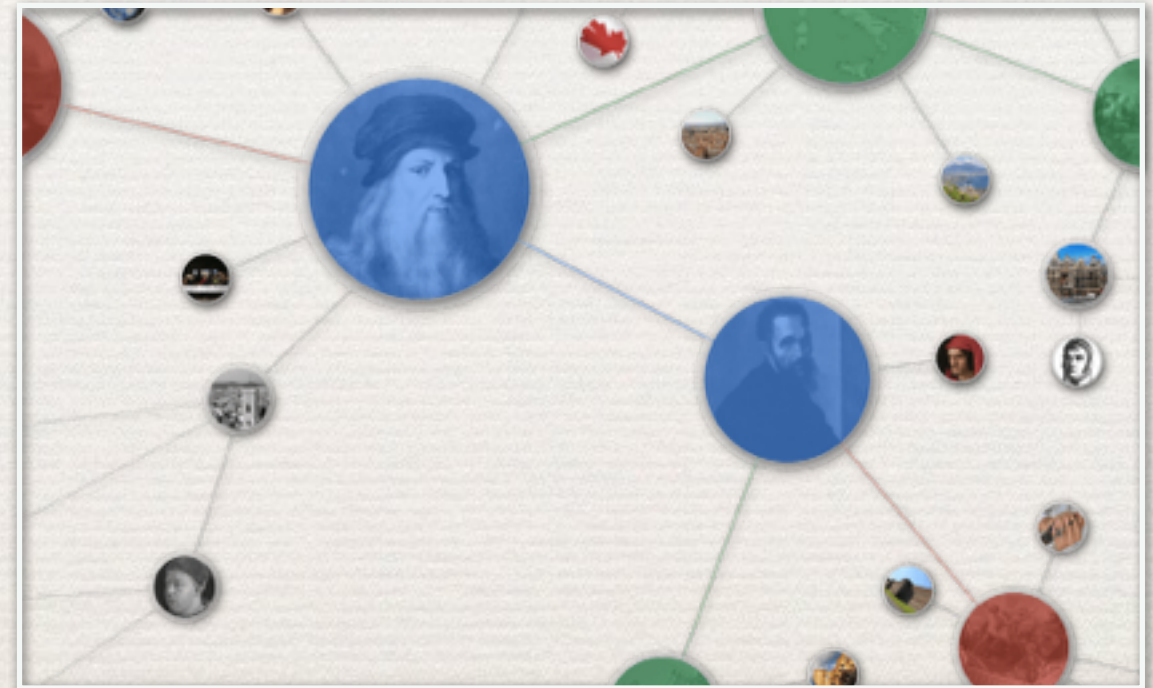
Jacob Bernoulli
Mathematician

Jacob Bernoulli was one of the many prominent mathematicians in the Bernoulli family. He was an early proponent of Leibnizian calculus and had sided with Leibniz during the Leibniz–Newton calculus controversy.
[Wikipedia](#)


Born: January 6, 1655, Basel, Switzerland
Died: August 16, 1705, Old Swiss Confederacy
Education: [University of Basel](#)
Siblings: [Johann Bernoulli](#)
Doctoral students: [Johann Bernoulli](#), [Jakob Hermann](#), [Nicolaus I Bernoulli](#)

Nephews

		
Daniel Bernoulli through Johann Bernoulli	Nicolaus II Bernoulli through Johann Bernoulli	Johann II Bernoulli through Johann Bernoulli



QA & KNOWLEDGE EXTRACTION



Jacob Bernoulli
Mathematician


Jacob Bernoulli was one of the many prominent mathematicians in the Bernoulli family. He was an early proponent of Leibnizian calculus and had sided with Leibniz during the Leibniz–Newton calculus controversy. [Wikipedia](#)

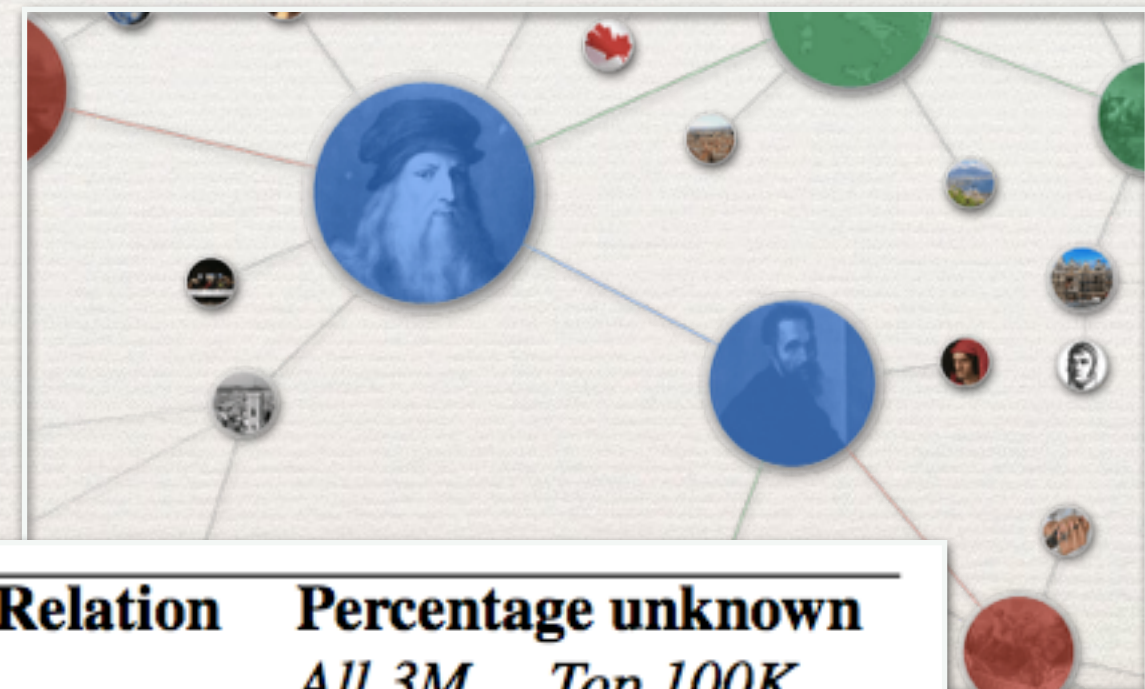
Born: January 6, 1655, [Basel, Switzerland](#)
Died: August 16, 1705, [Old Swiss Confederacy](#)
Education: [University of Basel](#)
Siblings: [Johann Bernoulli](#)
Doctoral students: [Johann Bernoulli](#), [Jakob Hermann](#), [Nicolaus Bernoulli](#)

Nephews


Daniel Bernoulli
through Johann Bernoulli

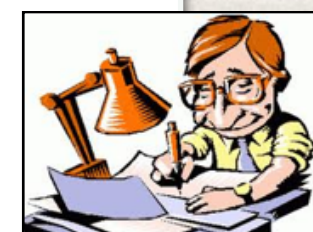

Nicolaus II Bernoulli
through Johann Bernoulli


Johann II Bernoulli
through Johann Bernoulli



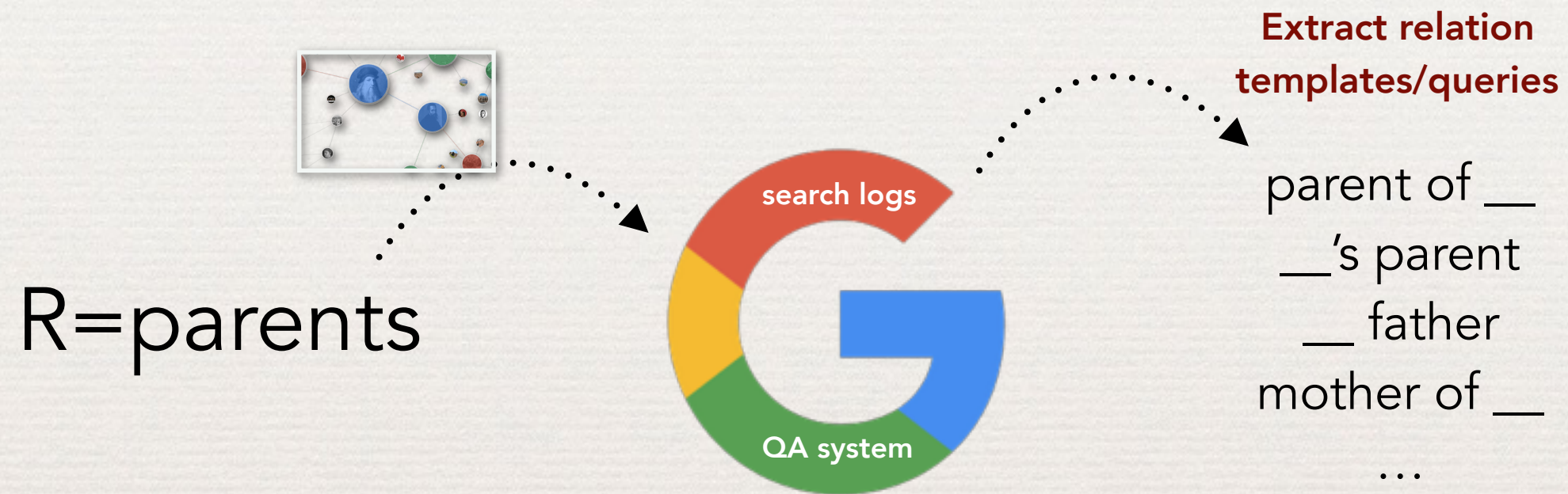
Relation	Percentage unknown	
	<i>All 3M</i>	<i>Top 100K</i>
PROFESSION	68%	24%
PLACE OF BIRTH	71%	13%
NATIONALITY	75%	21%
EDUCATION	91%	63%
SPOUSES	92%	68%
PARENTS	94%	77%
CHILDREN	94%	80%
SIBLINGS	96%	83%
ETHNICITY	99%	86%

as of 2014



WEAKLY SUP. KNOWLEDGE EXTRACTION

(WEST ET AL. 2014)



WEAKLY SUP. KNOWLEDGE EXTRACTION

(WEST ET AL. 2014)



R=parents



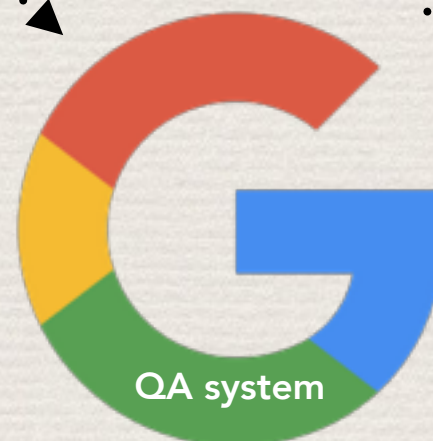
Extract relation
templates/queries

parent of __
__'s parent
__ father
mother of __
...

Q=Frank Zappa

parent of Frank Zappa
Frank Zappa's parent
Frank Zappa father
mother of Frank Zappa
...

Issue queries

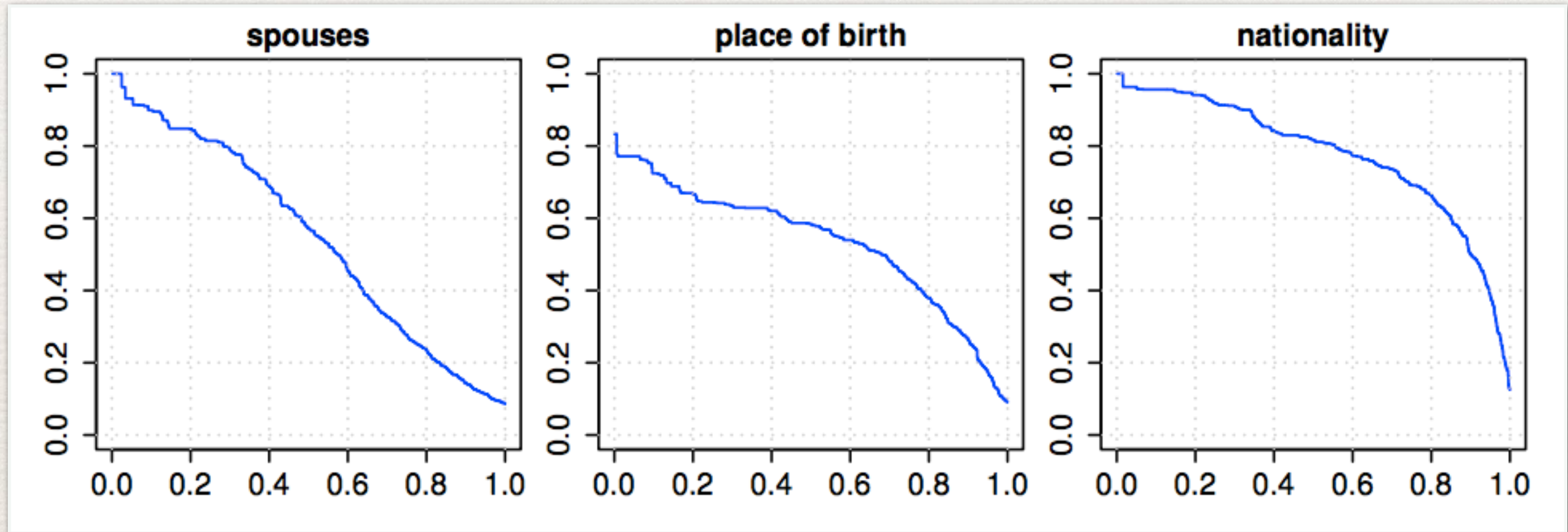


Score entities in
result snippets
& aggregate

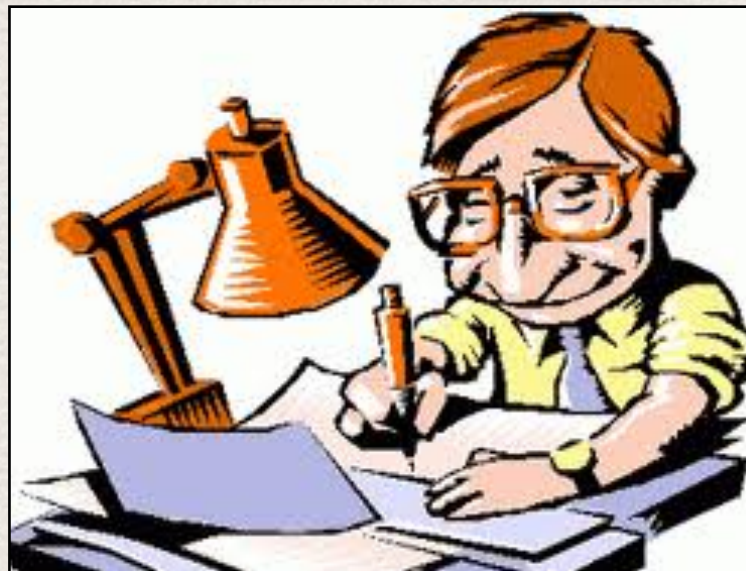
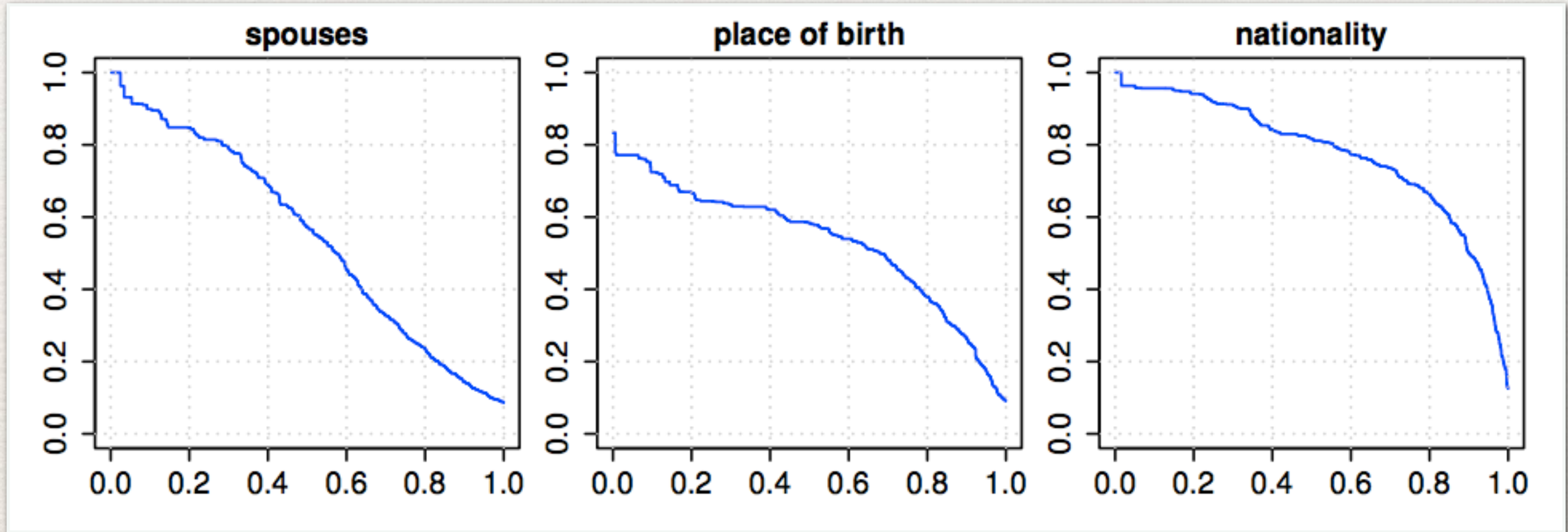
Mothers of Inversion
Ray Collins
Rose Marie Colimore
Francis Zappa
Gail Zappa
Rose Marie



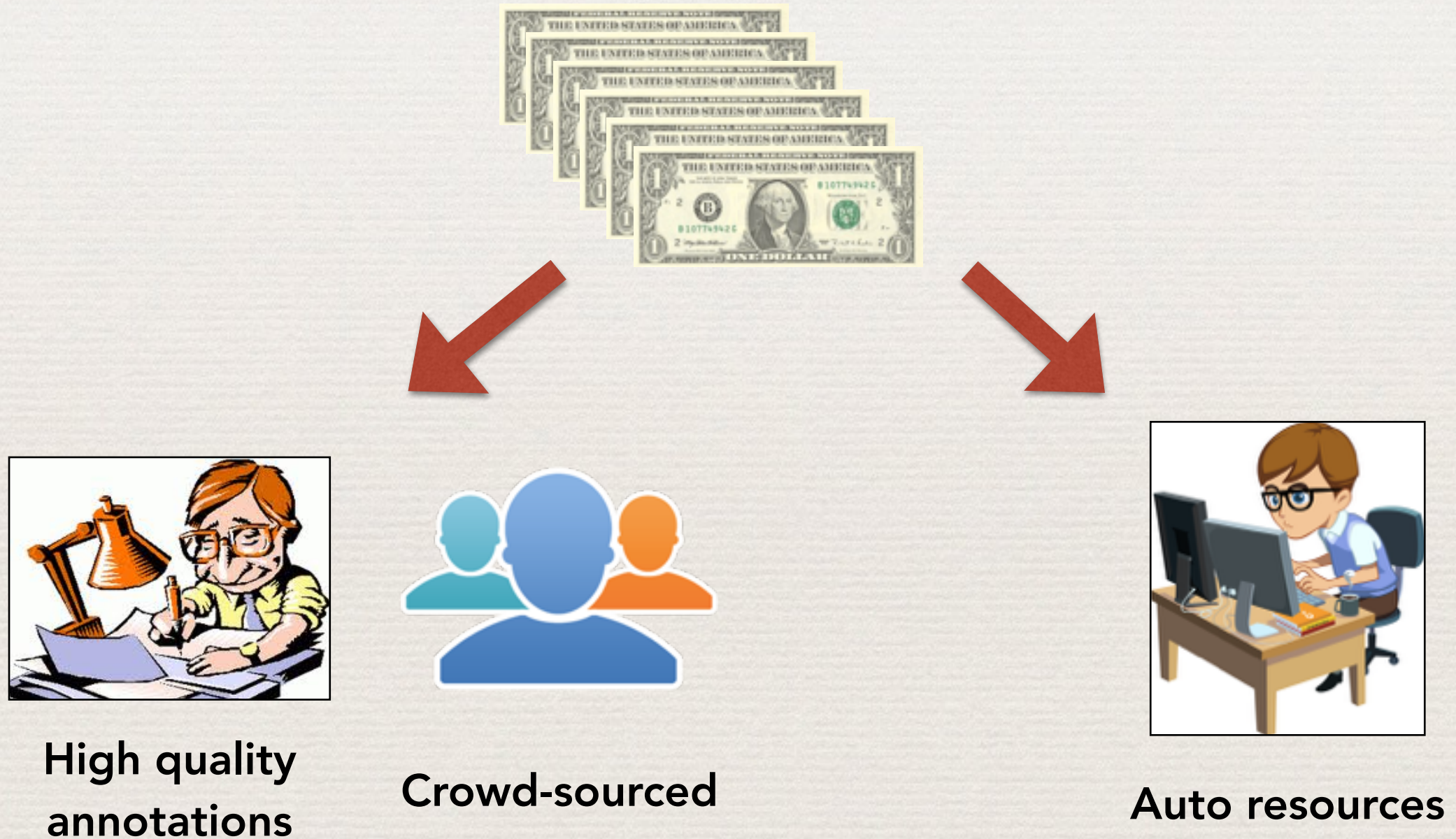
DOES IT WORK?



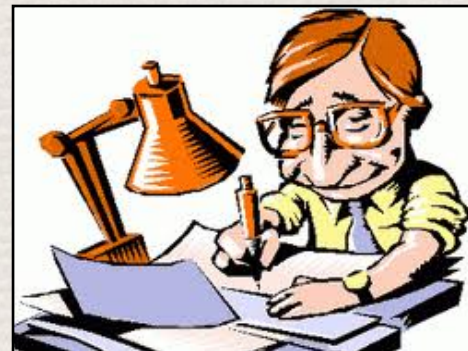
DOES IT WORK?



THE RESOURCE TRADE-OFF



THE RESOURCE TRADE-OFF



High quality
annotations

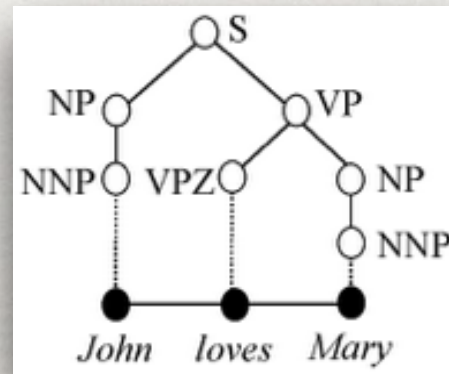
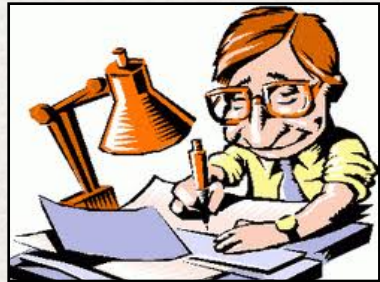


Crowd-sourced

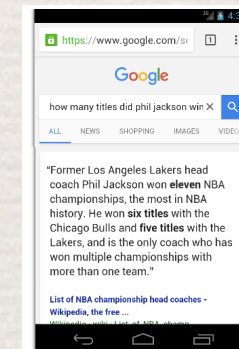
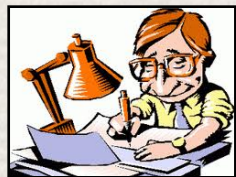


Auto resources

TOP-LEVEL CONCLUSION



Syntax



Semantics



Thanks