Crowd Prefers the Middle Path: *A New IAA Metric Reveals Turker Biases in Query Segmentation*

Rohan Ramanath, *R. V. College of Engineering, Bangalore*

Monojit Choudhury, Kalika Bali, *Microsoft Research India*

Rishiraj Saha Roy, *IIT Kharagpur*

monojitc@microsoft.com

Query segmentation

new york times square dance

scottish country dancing clubs melbourne

tony hawk american wasteland ps2 cheats

what causes swollen lymph nodes

Query segmentation

new york times | square dance

scottish country | dancing clubs | melbourne

tony hawk american wasteland | ps2 | cheats

what causes | swollen lymph nodes

Similar to CHUNKING of NL Text

Poor Inter-Annotator Agreement

Query Accuracy:	0.58 – 0.61
Segment F-score:	0.69 – 0.72
Segment Accuracy:	0.84 – 0.85

(*Tan and Peng, 2008*)

Sources of Ambiguity

new york times | square dance

new york | times square | dance

scottish country | dancing clubs | Melbourne

scottish country dancing clubs | Melbourne

tony hawk american wasteland | ps2 | cheats

tony hawk | american wasteland | ps2 cheats

what causes | swollen lymph nodes

what causes | swollen | lymph nodes

Issue of Granularity

- Maximal vs. Minimal segments
- Also observed for Text Chunking

A series of happy thoughts | came to mind

A series of | happy thoughts | came to mind

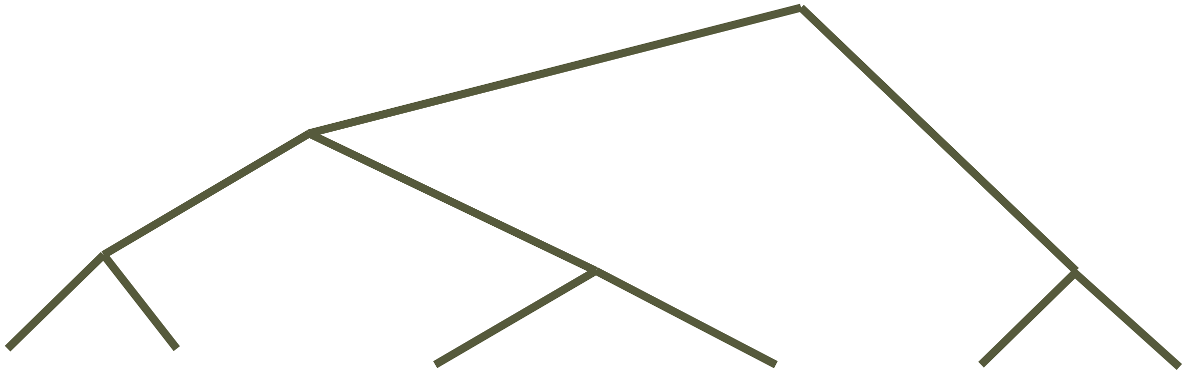
Annotators agree on major (clause or phrase) boundaries, but not on minor ones.

(*Abney, 1992,1995; Bali et al., 2009*)

Hierarchical Segmentation

tony hawk american wasteland | ps2 | cheats

tony hawk | american wasteland | ps2 cheats



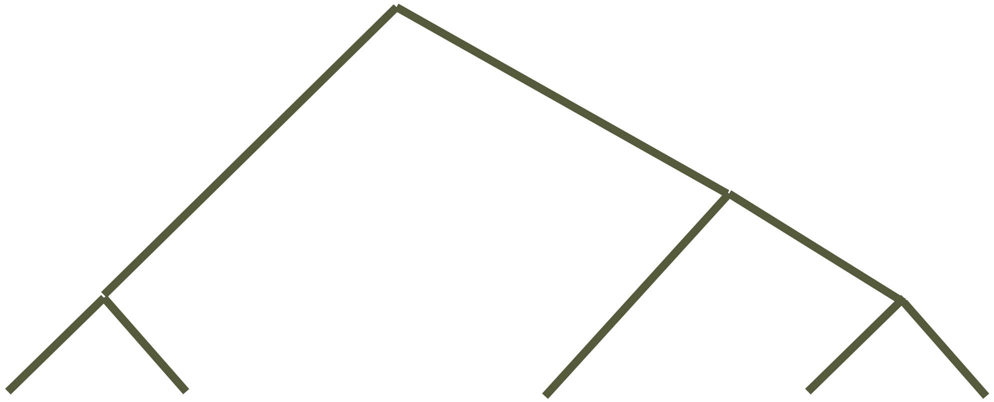
(((tony hawk) (american wasteland))(ps2 cheats))

Flat & Nested Segmentation

what causes | swollen lymph nodes

what causes | swollen | lymph nodes

Flat Segmentation



((what causes) (swollen (lymph nodes)))

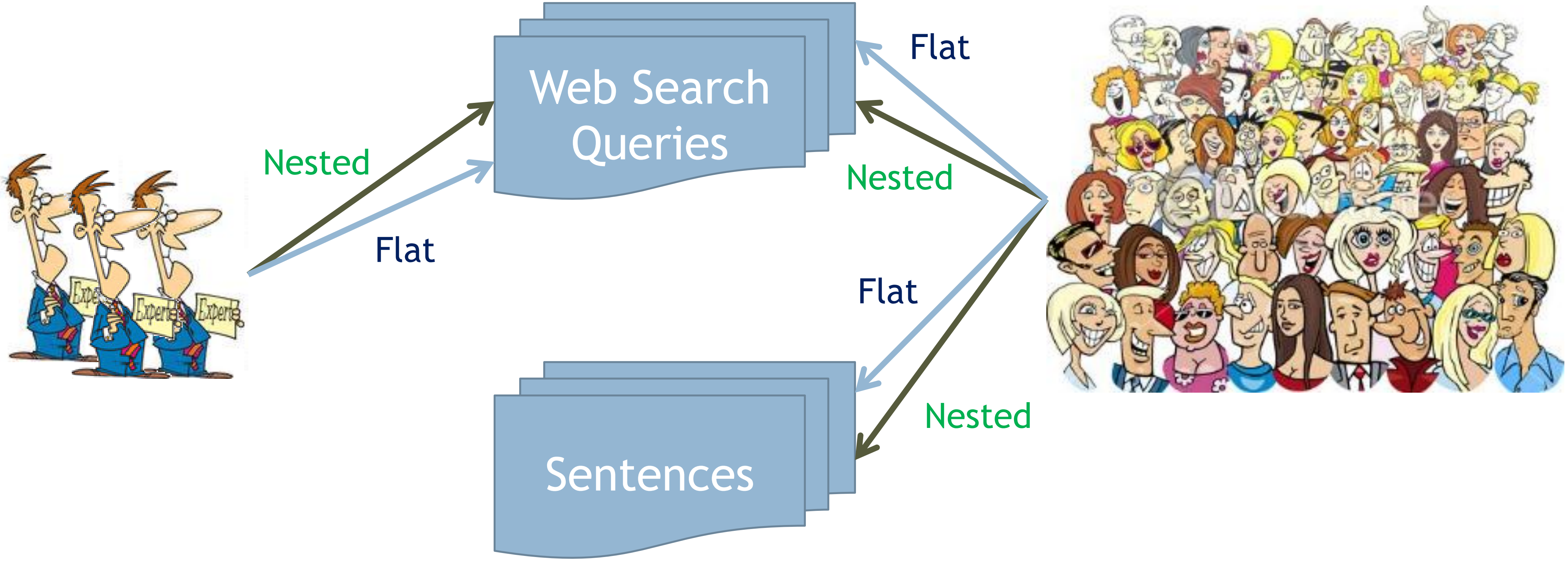
Binary Nested Segmentation

Research Questions

Does *Nested Segmentation* of Queries (& NL texts) lead to better agreement amongst expert annotators?

Can *crowdsourcing* be used for obtaining reliable high quality annotations of this kind?

Experiments



Nested

Flat

Web Search Queries

Sentences

Flat

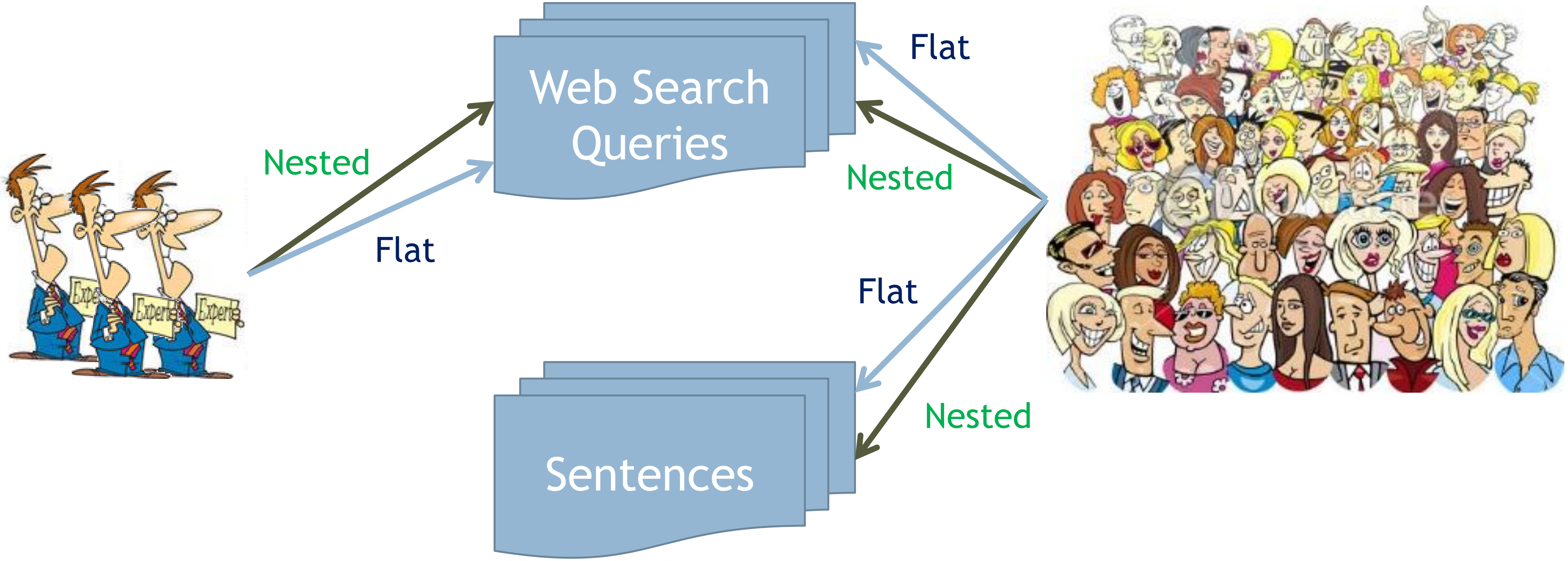
Nested

Flat

Nested

Experiments

1200 queries from Bing
4-8 words long



Web Search Queries

Sentences

Nested

Flat

Flat

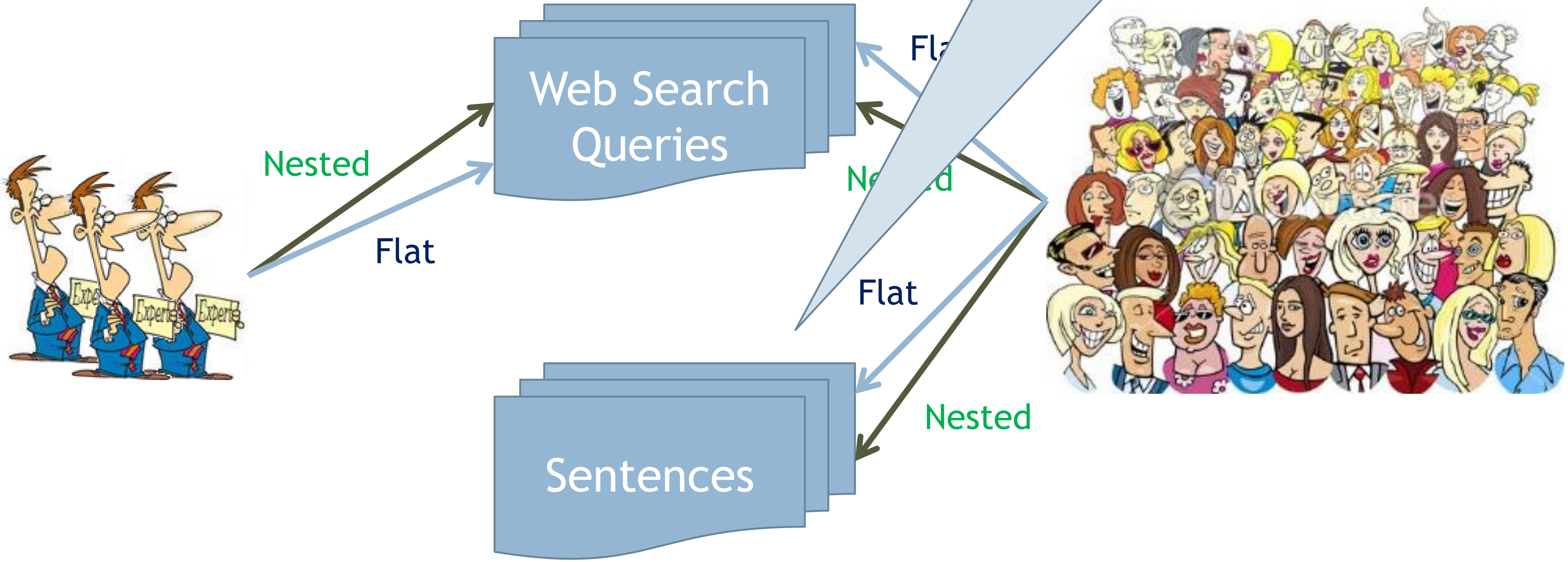
Nested

Flat

Nested

Experiments

300 English sentences
5-15 words long



Web Search Queries

Sentences

Nested

Flat

Fla

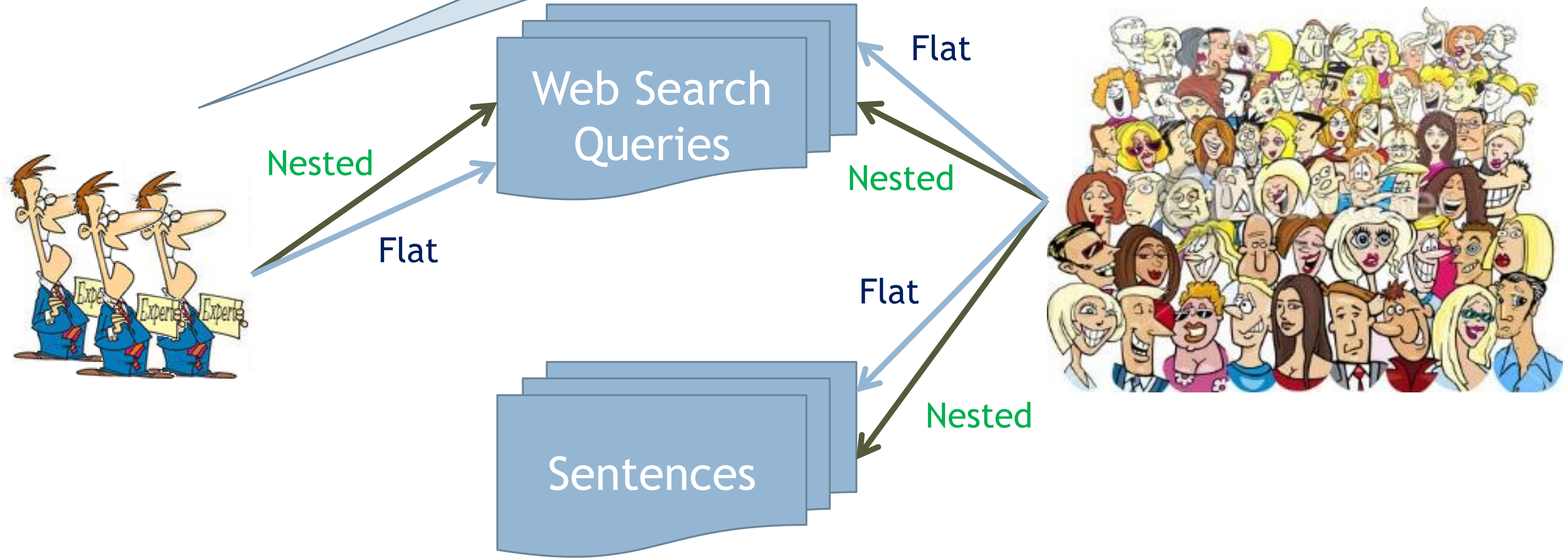
Ne d

Flat

Nested

Experiments

3 very frequent search engine users, special training provided



Web Search Queries

Sentences

Nested

Flat

Flat

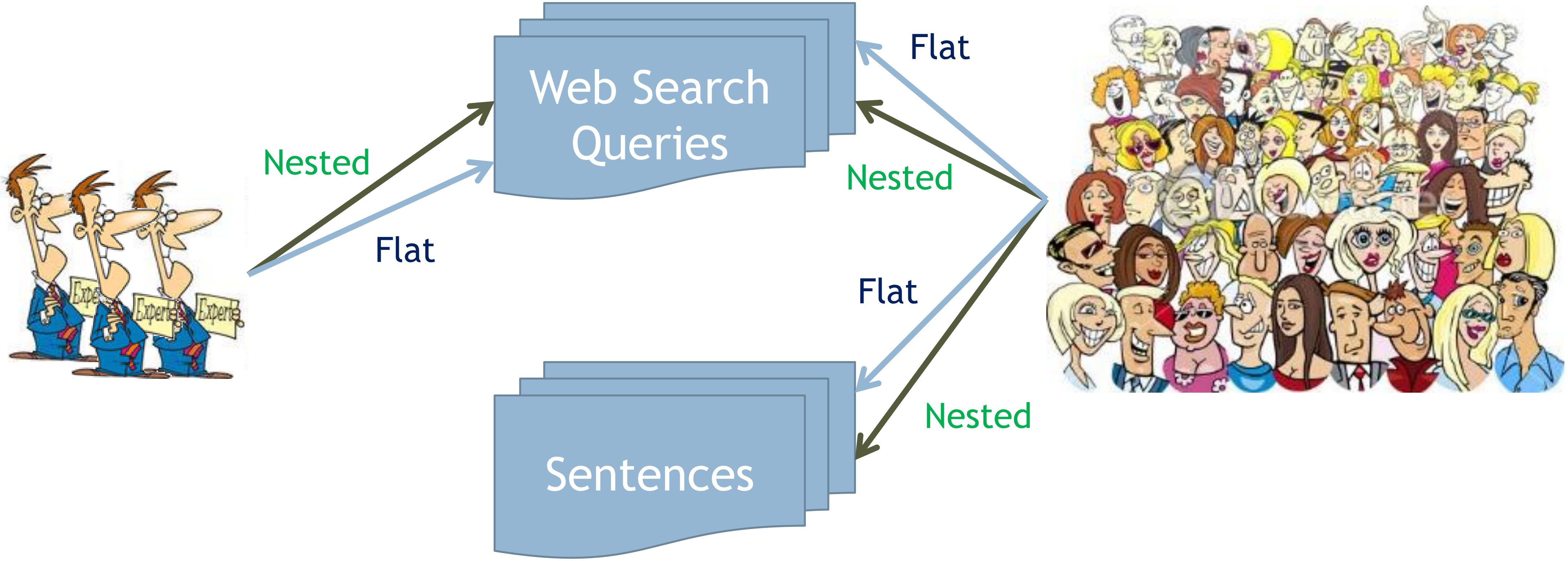
Nested

Flat

Nested

Experiments

Amazon Mechanical Turk
10 annotations per item
1 min video for training



Web Search Queries

Sentences

Nested

Flat

Flat

Nested

Flat

Nested

Inter Annotator Agreement

Challenge 1

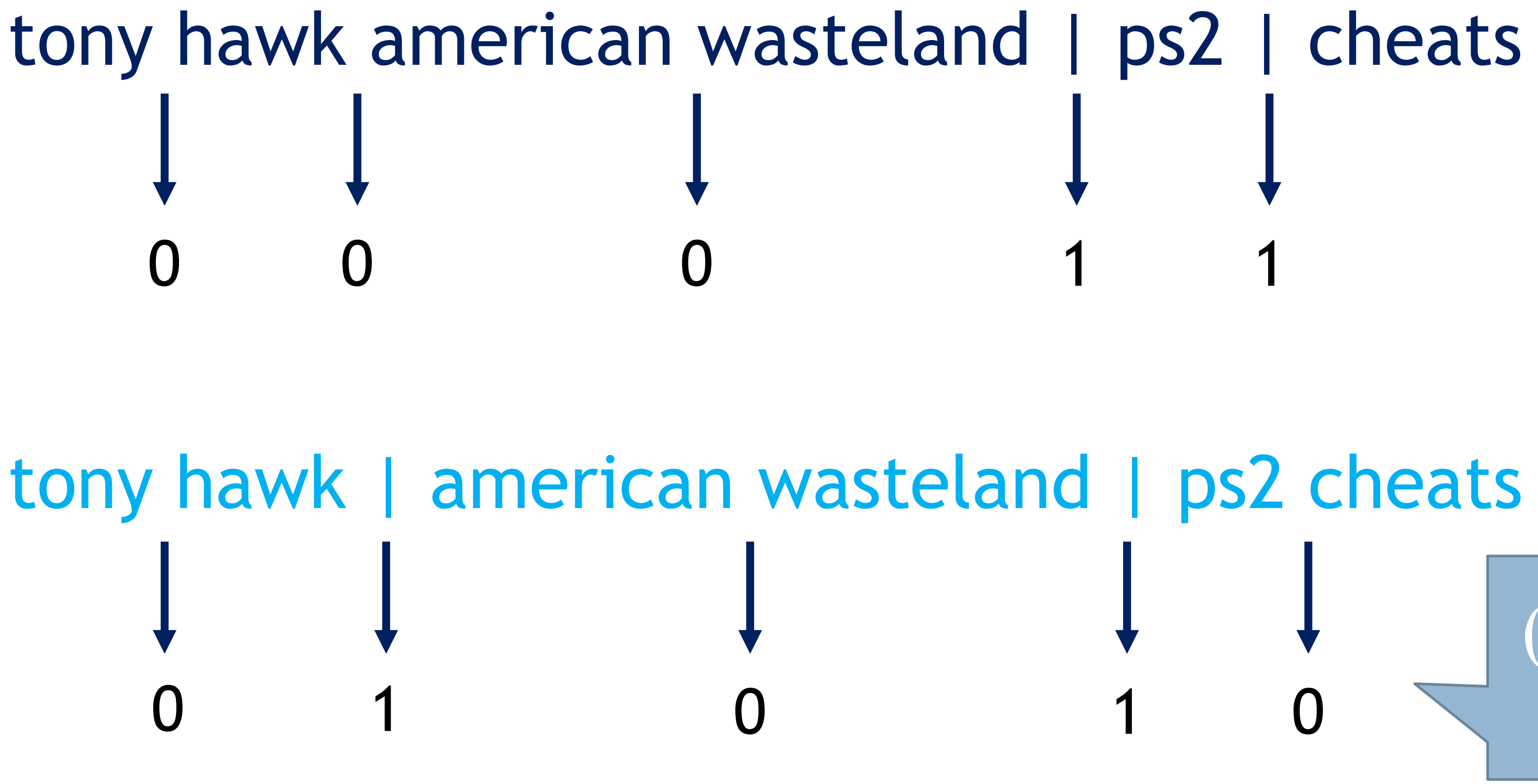
Given two flat/nested annotations, how to define the similarity?

Challenge 2

What is the chance agreement?

Similarity betw
Flat Annotation

$$d_1(q_m, q'_n) = \frac{1}{|q|-1} \sum_{i=1}^{|q|-1} |b_{m,i} - b'_{n,i}|$$



tony hawk american wasteland | ps2 | cheats

0 0 0 1 1

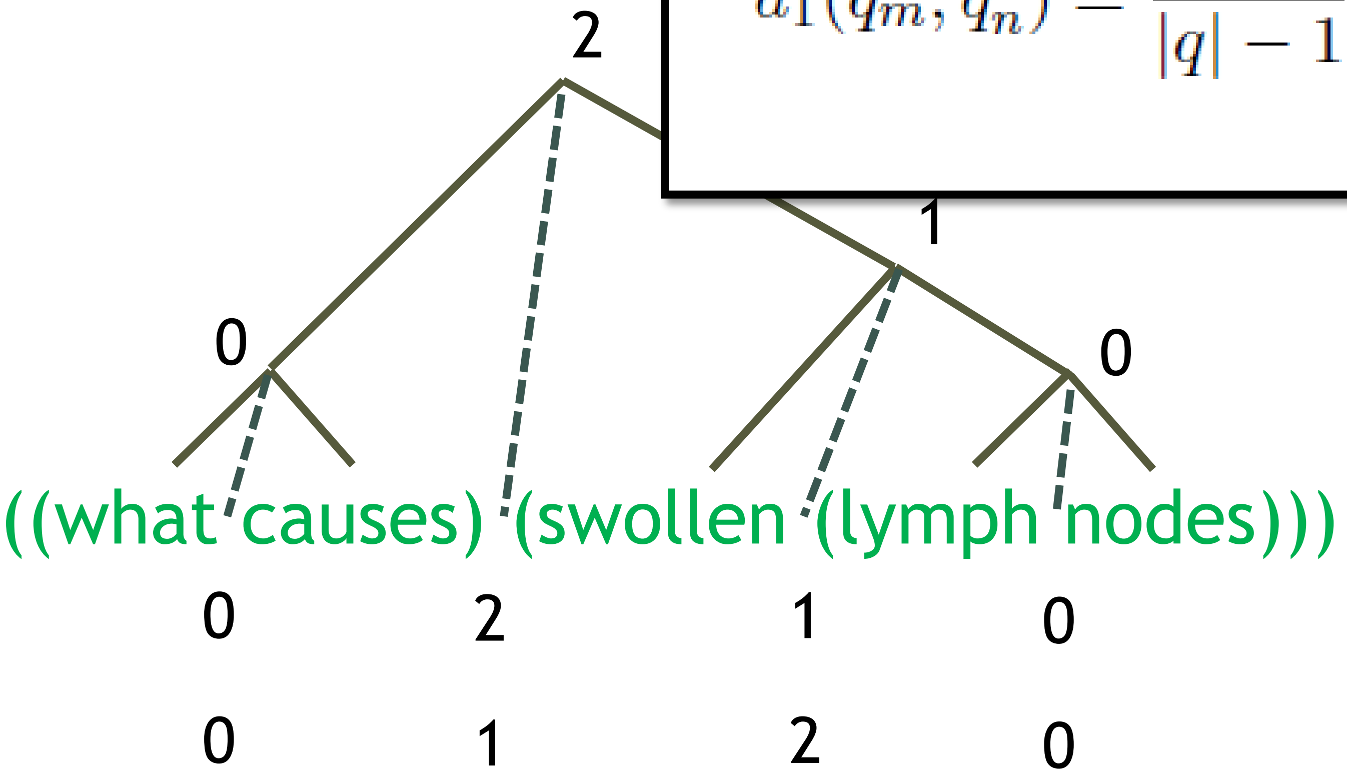
tony hawk | american wasteland | ps2 cheats

0 1 0 1 0

(0+1+0+0+1)/5
= 2/5

Similarity between Nested Annotation

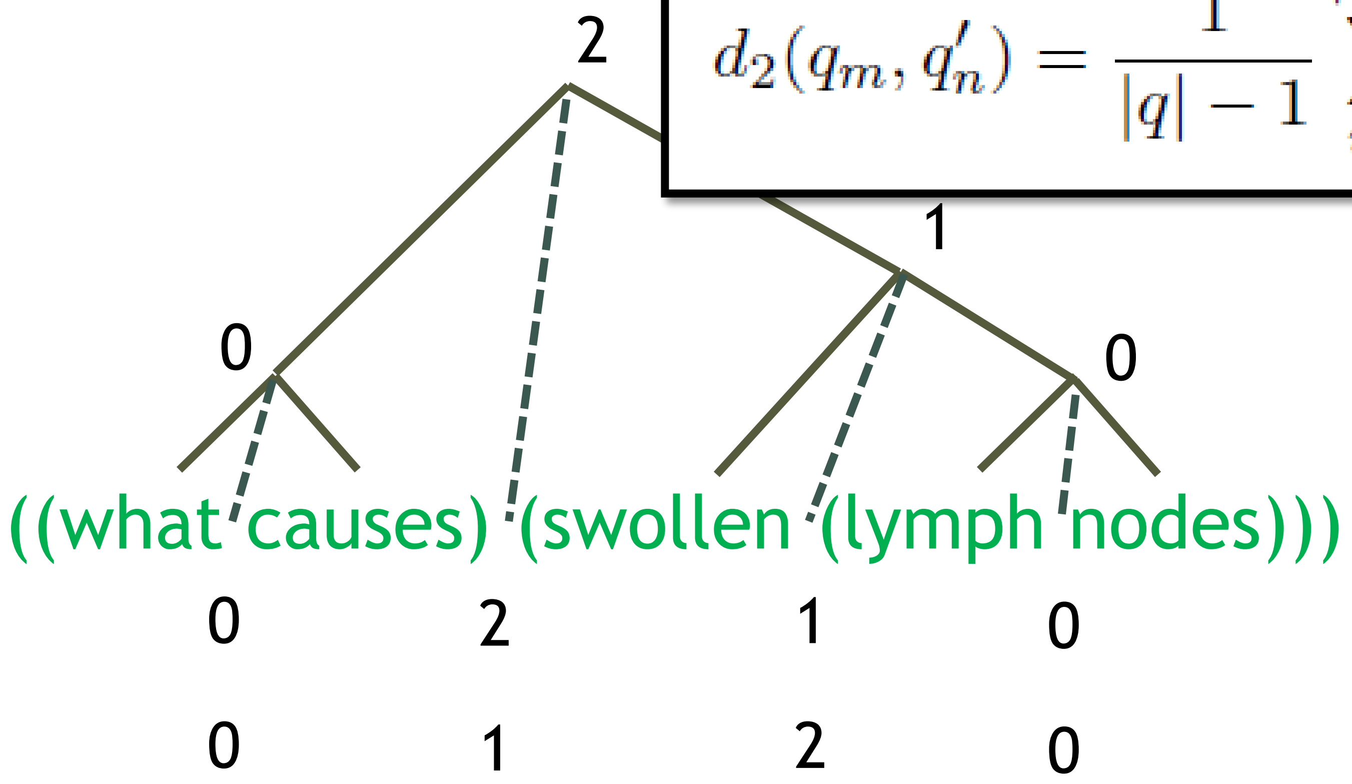
$$d_1(q_m, q'_n) = \frac{1}{|q|-1} \sum_{i=1}^{|q|-1} |b_{m,i} - b'_{n,i}|$$



(0+1+1+0)/4
= 2/4

Similarity between Nested Annotation

$$d_2(q_m, q'_n) = \frac{1}{|q|-1} \sum_{i=1}^{|q|-1} |b^2_{m,i} - (b'_{n,i})^2|$$



(0+3+3+0)/4
= 6/4

Chance Agreement

Model 1: *S*

All annotations are equally likely

Model 2: *Cohen's κ*

Every annotator has a different bias

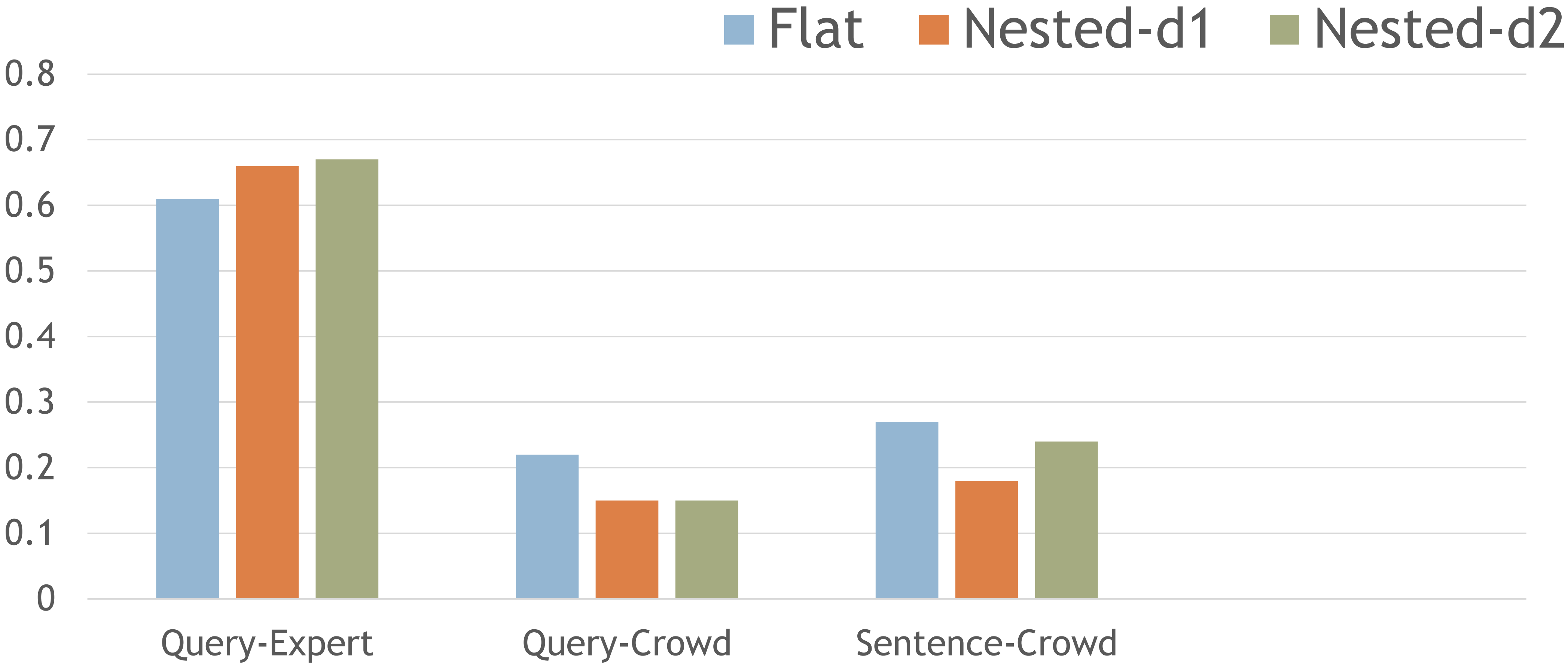
[doesn't apply to crowdsourcing]

Model 3: *Krippendorff's a*

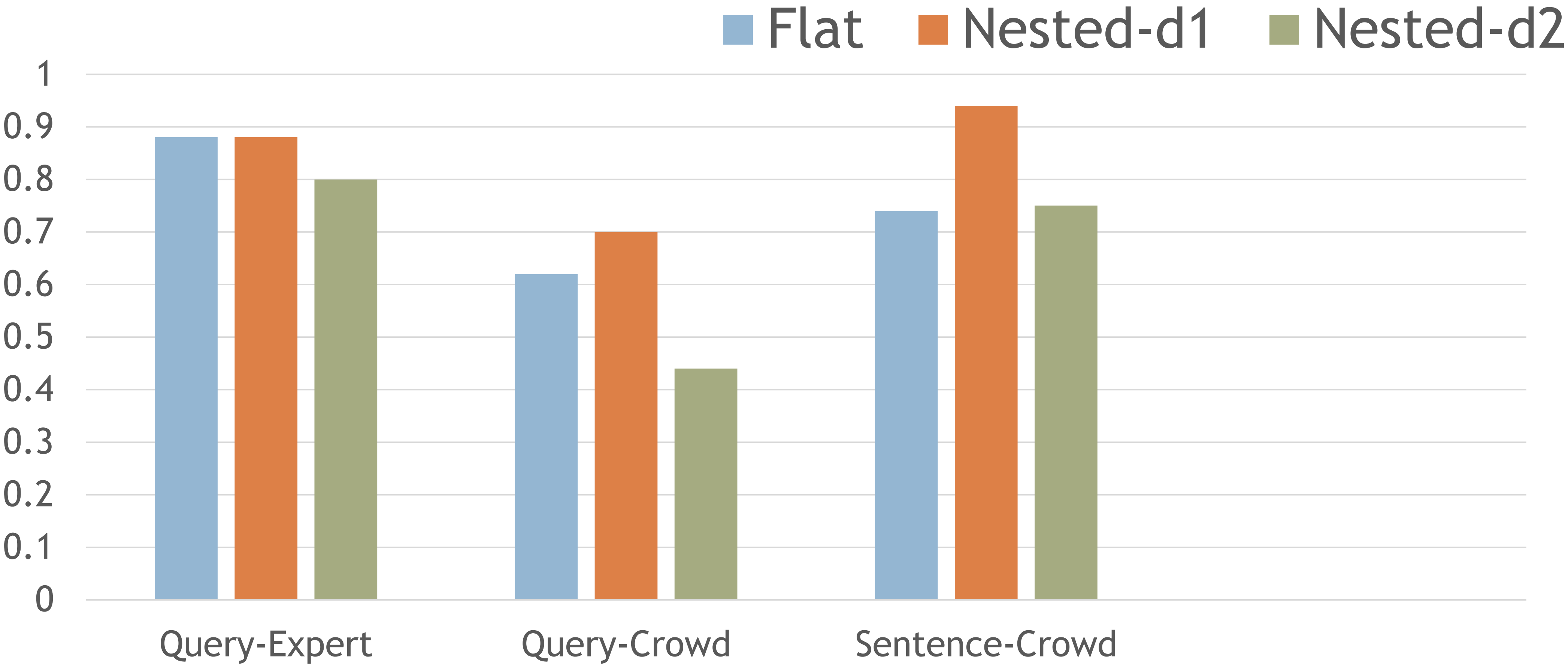
The population has a bias

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{s^2_{within}}{s^2_{total}}$$

IAA Statistics - α



IAA Statistics - *S*



Turker Bias 1: Two segments of equal length

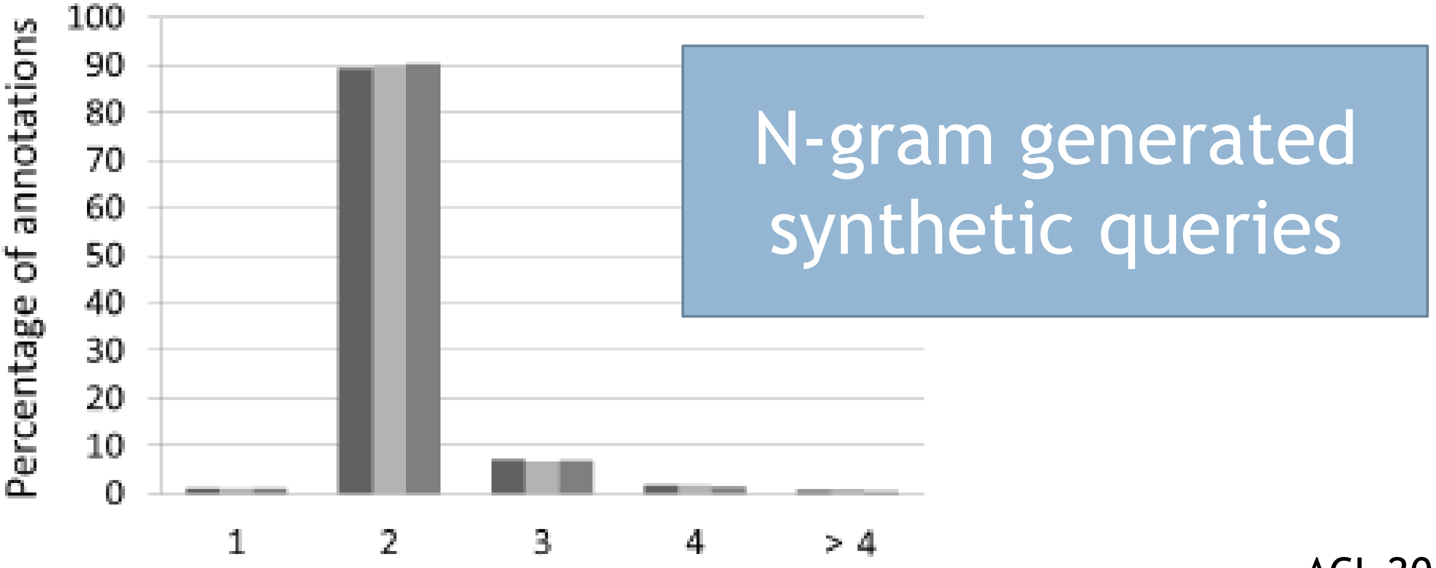
- 80% queries and 60% sentences have 2 segments
- The length of the two segments differ by 0 or 1 words

power rangers operation | overdrive multiplayer online game

st francis of | assisi primary school

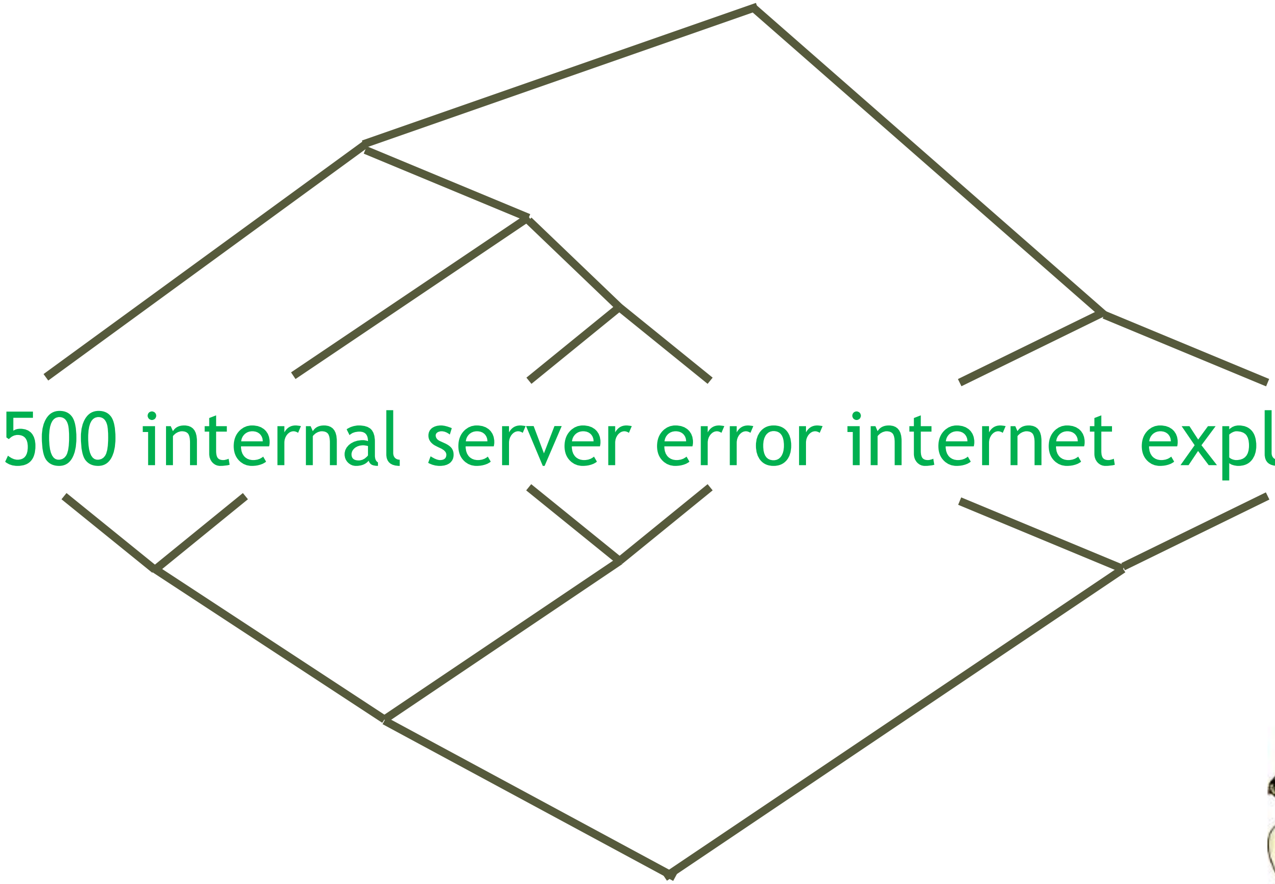
Turker Bias 1: Two segments of equal length

- 80% queries and 60% sentences have 2 segments
- The length of the two segments differ by 0 or 1 words



N-gram generated synthetic queries

Turker Bias 2: Balanced Trees



500 internal server error internet explorer

Linguistic Features

- Phrase structure drives segmentation only if reconcilable with Biases 1 and 2.

- Prepositions grouped with following word in NL sentences, but no such dominant trends in queries

flights to, ideas for

Conclusions

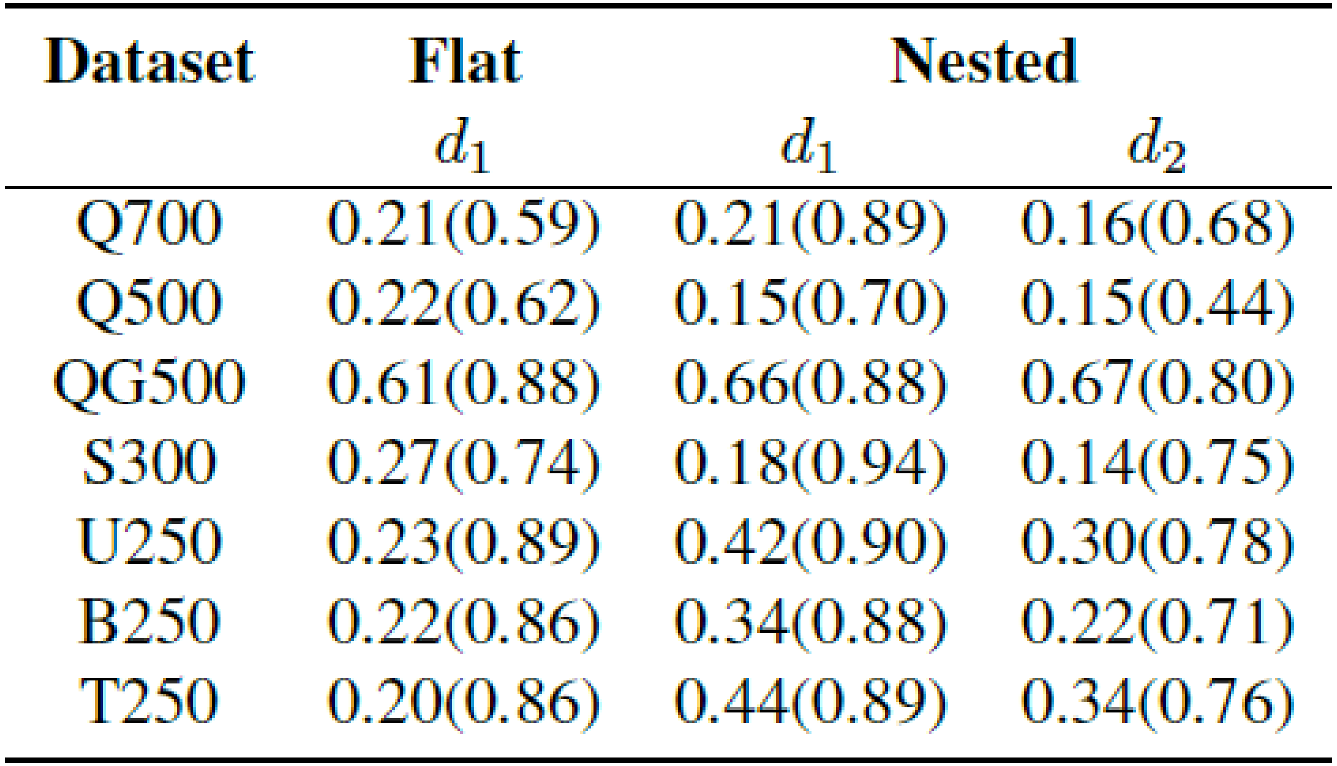
- Crowdsourcing unreliable for query segmentation
- Nested segmentation improves IAA for experts, but degrades it for the crowd (due to higher cognitive load)
- Crowd has strong bias towards balanced structures leading to apparently high IAA, but unreliable annotations
- The proposed IAA metric can correct for annotator biases in crowdsourcing

Thank you!

Data and supplementary material available from
http://research.microsoft.com/apps/pubs/default.aspx?id=192002

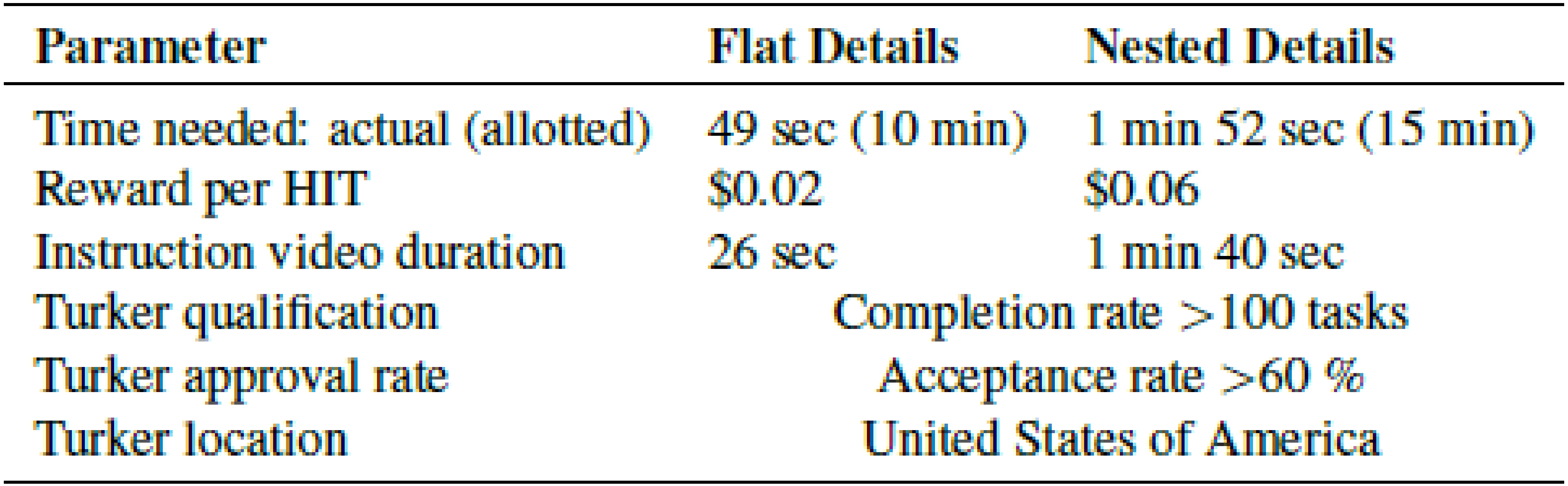
Entailment: An Effective Metric for Comparing and Evaluating Hierarchical and Non-hierarchical Annotation Schemes, *Linguistic Annotation Workshop*
(8th August, 11:40am)

Detailed IAA Stats



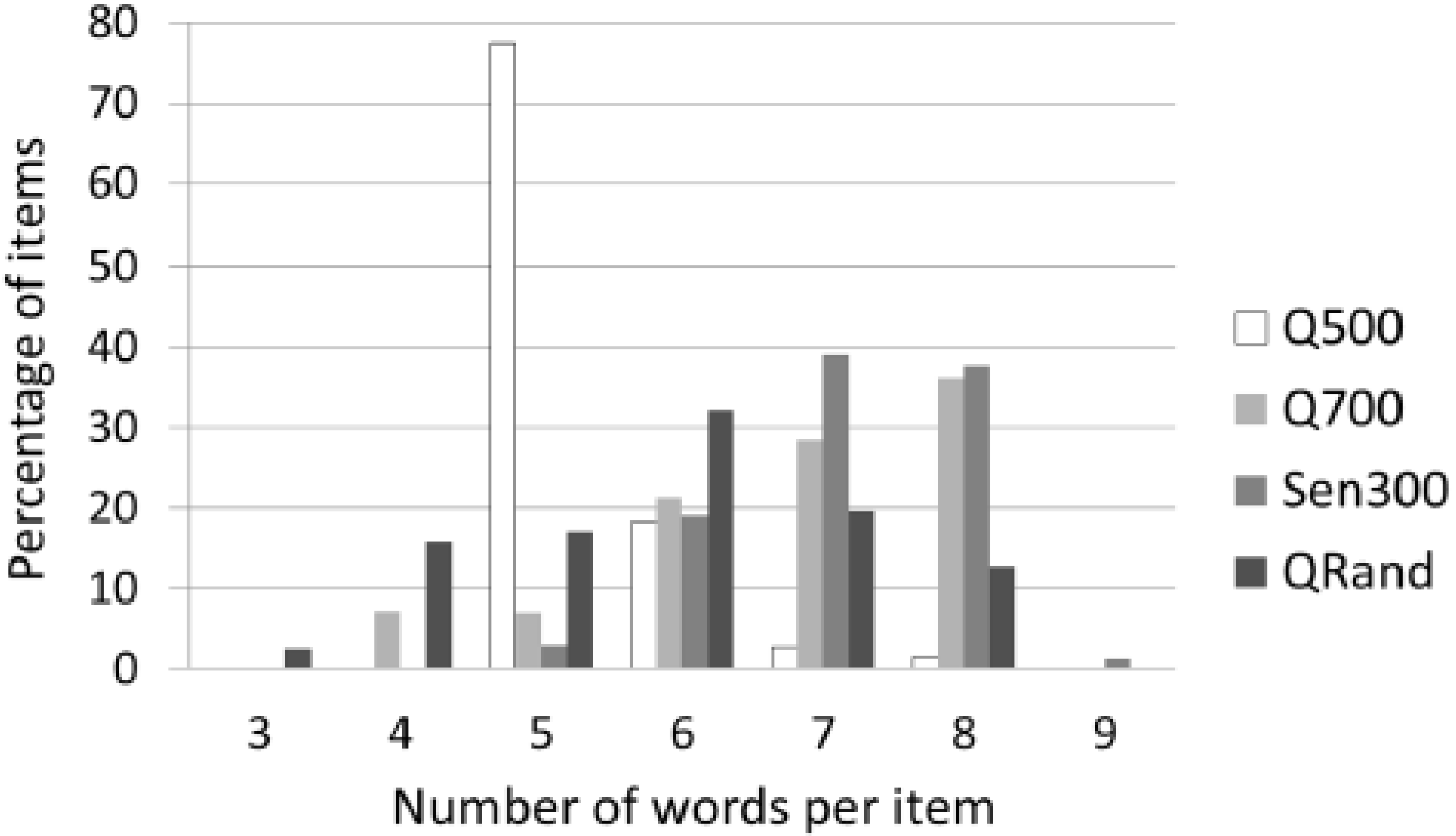
Dataset	Flat	Nested	
	d_1	d_1	d_2
Q700	0.21(0.59)	0.21(0.89)	0.16(0.68)
Q500	0.22(0.62)	0.15(0.70)	0.15(0.44)
QG500	0.61(0.88)	0.66(0.88)	0.67(0.80)
S300	0.27(0.74)	0.18(0.94)	0.14(0.75)
U250	0.23(0.89)	0.42(0.90)	0.30(0.78)
B250	0.22(0.86)	0.34(0.88)	0.22(0.71)
T250	0.20(0.86)	0.44(0.89)	0.34(0.76)

AMT Parameters

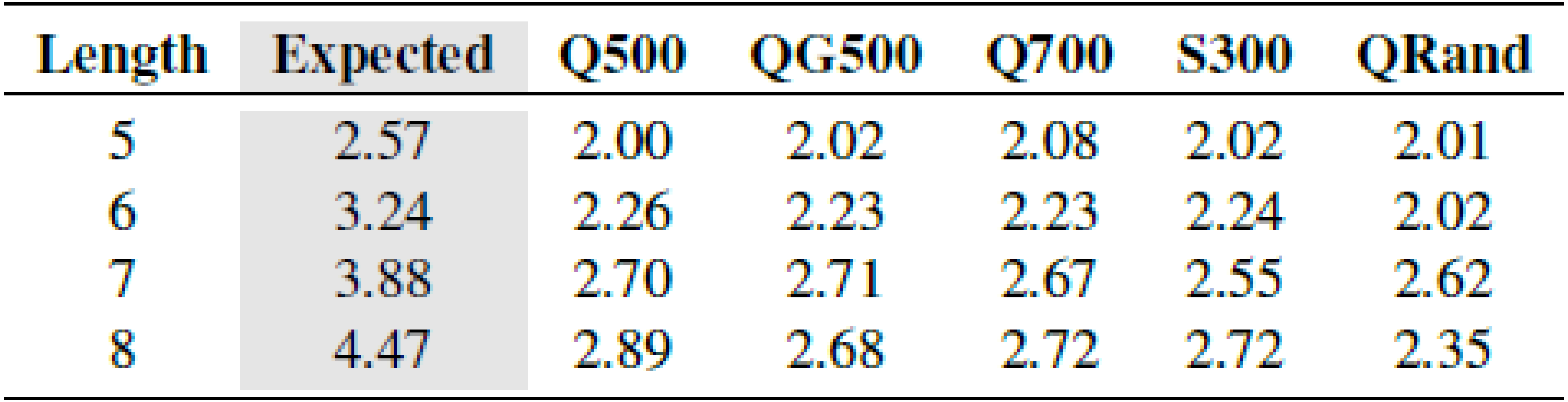


Parameter	Flat Details	Nested Details
Time needed: actual (allotted)	49 sec (10 min)	1 min 52 sec (15 min)
Reward per HIT	$0.02	$0.06
Instruction video duration	26 sec	1 min 40 sec
Turker qualification	Completion rate >100 tasks	
Turker approval rate	Acceptance rate >60 %	
Turker location	United States of America	

Text Length Distribution

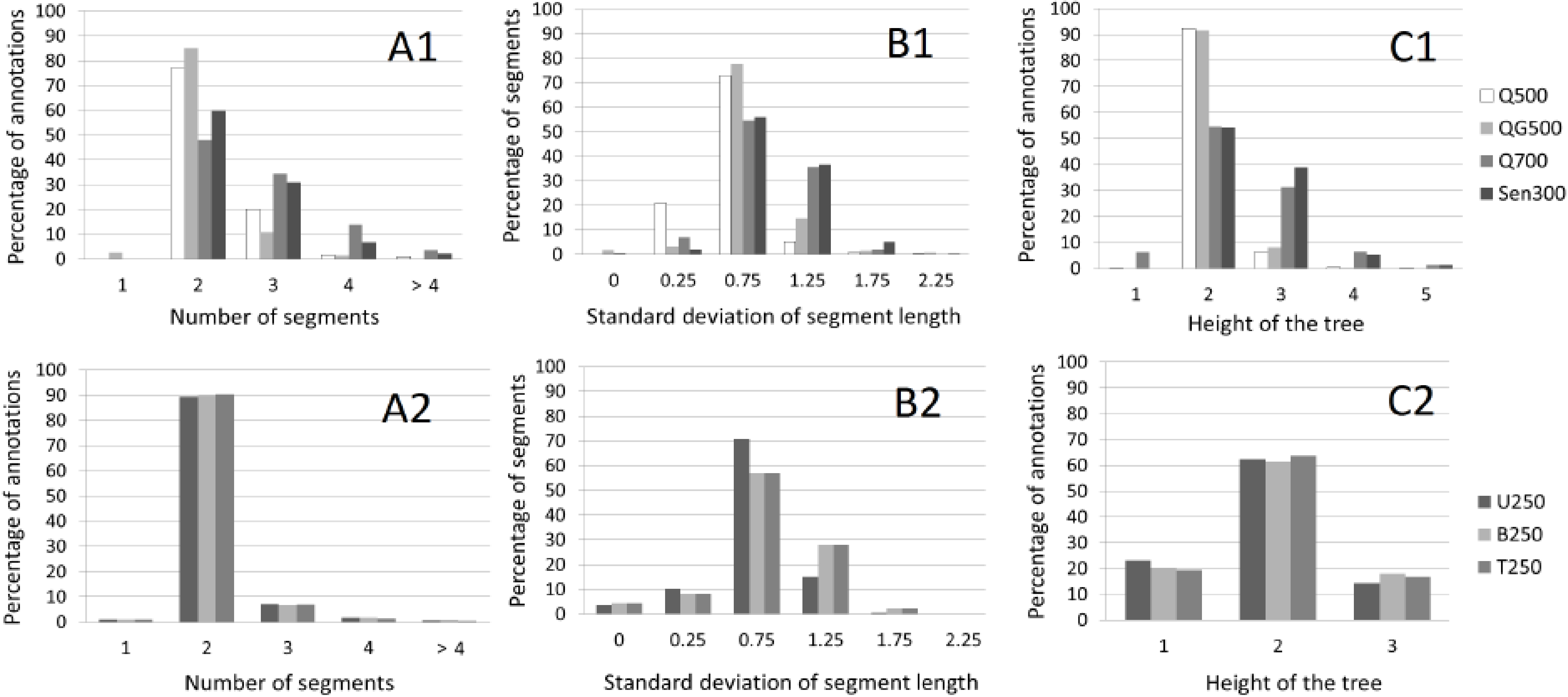


Height of Nested Segmentation

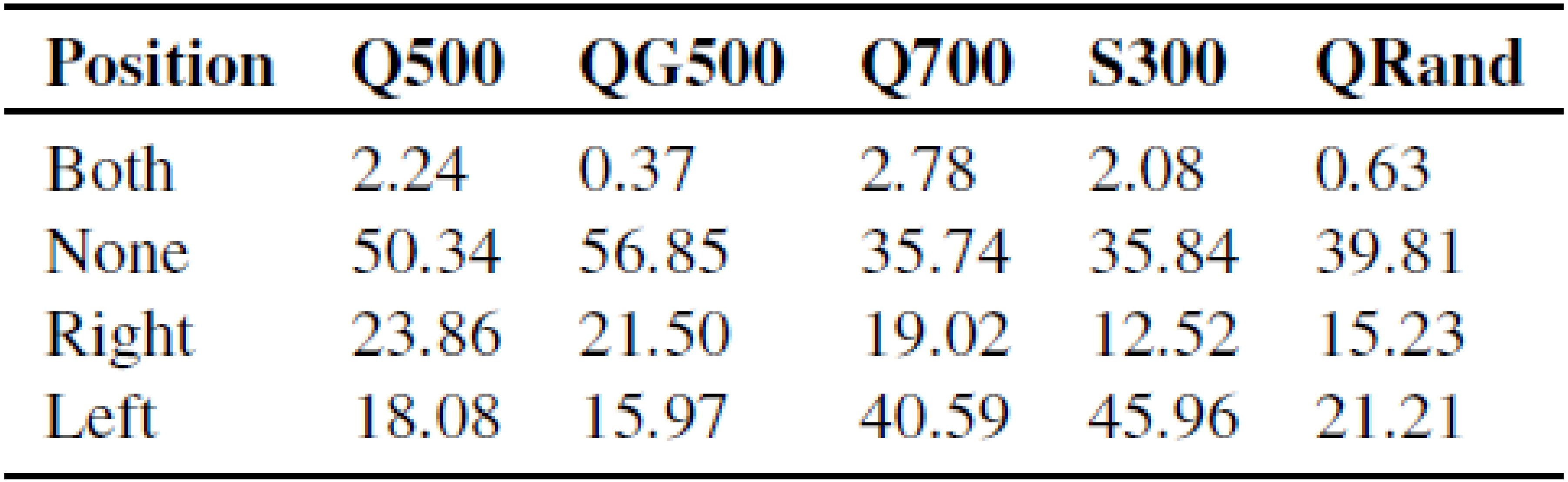


Length	Expected	Q500	QG500	Q700	S300	QRand
5	2.57	2.00	2.02	2.08	2.02	2.01
6	3.24	2.26	2.23	2.23	2.24	2.02
7	3.88	2.70	2.71	2.67	2.55	2.62
8	4.47	2.89	2.68	2.72	2.72	2.35

Segments Length Distribution



Preposition Statistics



Position	Q500	QG500	Q700	S300	QRand
Both	2.24	0.37	2.78	2.08	0.63
None	50.34	56.85	35.74	35.84	39.81
Right	23.86	21.50	19.02	12.52	15.23
Left	18.08	15.97	40.59	45.96	21.21