Parsing Performance across Domains and Genres

Profile	SA
Linux_Blog	29.2
Linux_Wiki	31.8
NLP_Wiki	37.5
NLP_Blog	39.2

Performance over NLP is better than over Linux

Parsing Performance across Domains and Genres

Profile	SA	PE
Linux_Blog	29.2	81.8
Linux_Wiki	31.8	83.8
NLP_Wiki	37.5	84.1
NLP_Blog	39.2	83.3

Performance over NLP is better than over Linux or maybe not.

Parsing Performance across Domains and Genres

Profile	SA	PE	Length
Linux_Blog	29.2	81.8	10.8
Linux_Wiki	31.8	83.8	18.1
NLP_Wiki	37.5	84.1	16.1
NLP_Blog	39.2	83.3	13.7

Performance over NLP is better than over Linux or maybe not.

Average item length has a strong impact on exact match.

Variation across profiles

Profile	PE	SA
WS01	86.1	40.1
WS02	84.5	37.7
WS03	85.6	38.5
WS04	83.2	34.2
WS05	82.0	36.0
WS06	82.8	30.5
WS07	83.2	33.4
WS08	84.7	35.6
WS09	83.7	31.1
WS10	85.5	34.1
WS11	82.8	32.3
WS12	84.4	36.3
WS13	83.5	36.8

Absolute difference in parseval over the derivation tree: 4.1, **for the same domain and genre**.

For exact match, it is even larger: 9.0.

For comparison:
DeepBank morsels: 87.2 – 90.6
PTB ParseEval: 89.9 (91.1) – 92.6

Discussion Questions

Is the variation in parser performance across profiles because of:

- uneven parser performance?
- variation in the data?
- overly sensitive evaluation metrics?
- ...

and

- can we measure any of these things independently?
- does it matter?
- what does it mean for those of us working on parse ranking?