What are the limitations on the flux of syntactic dependencies? Evidence from UD treebanks

Marie-Amélie Botalla (Paris 3 Sorbonne Nouvelle)
Chunxiao Yan (Paris Nanterre)
Sylvain Kahane (Paris Nanterre)

Depling, September 19, 2017, Pisa

Summary

- State of the art
- Dependency flux
- Project UD and method
- Results
- Conclusion

Summary

- State of the art
- Dependency flux
- Project UD and method
- Results
- Conclusion

State of the art

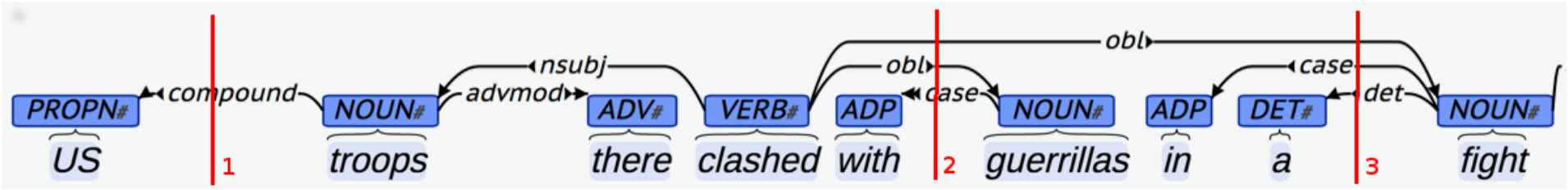
- Speaker performance is limited by several factors and especially by short-term memory (STM)
- The boundaries on STM are apparent in the sentences in spoken language (Yngve 1960, Gibson 1998)
- In Japanese, the number of words on the left of a position that can have a dependent on the right is bounded by 10 (Murata et al. 2001)

Summary

- State of the art
- Dependency flux
- Project UD and method
- Results
- Conclusion

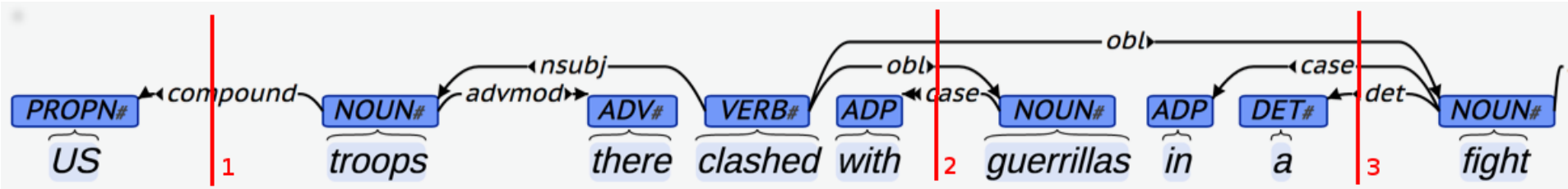
Dependency flux

- The *dependency flux* in a given inter-word position is the set of dependencies crossing this position, that is, linking a word on the left with a word on the right.



Size of the flux

- The *size* of the flux is the number of dependencies belonging to it



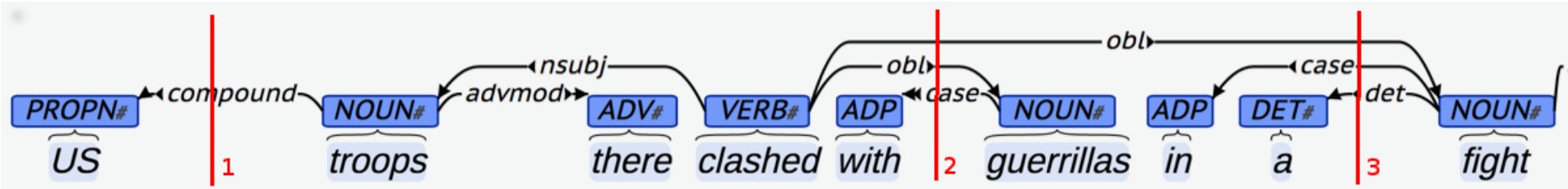
Position 1: size = 1

Position 2: size = 3

Position 3: size = 3

Left and right spans

- The *left span* and *right span* are the numbers of vertices on each side of the studied position



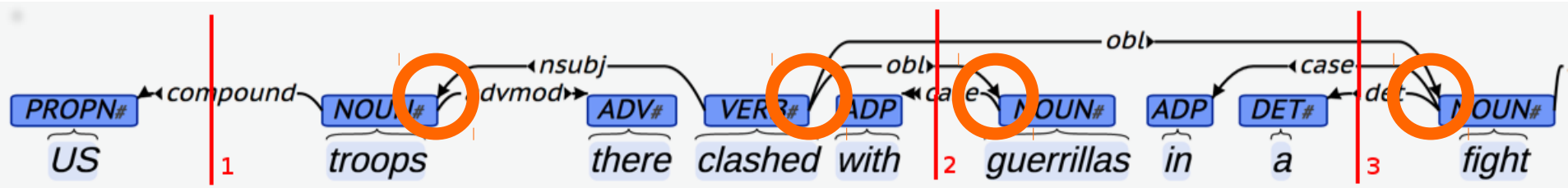
Position 1: left span = 1, right span = 1

Position 2: left span = 2, right span = 2

Position 3: left span = 3, right span = 1

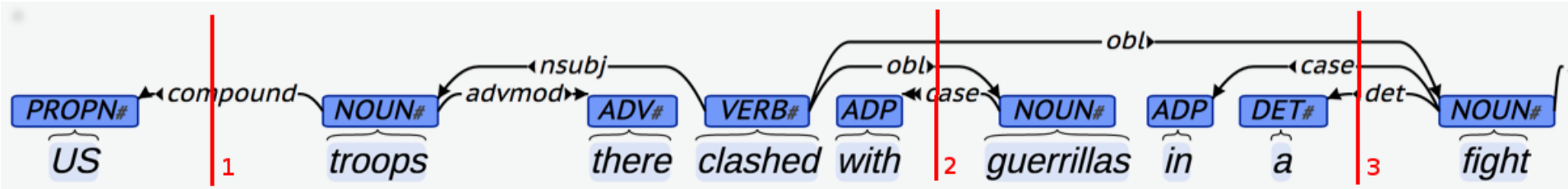
Bouquets and disjoint dependencies

- If dependencies share a vertex, they form a *bouquet*
- A set of dependencies is said *disjoint* if its dependencies do not share any vertex



Right/Left ratio

- The *right/left (or R/L) ratio* is the ratio of the right span to the left span.



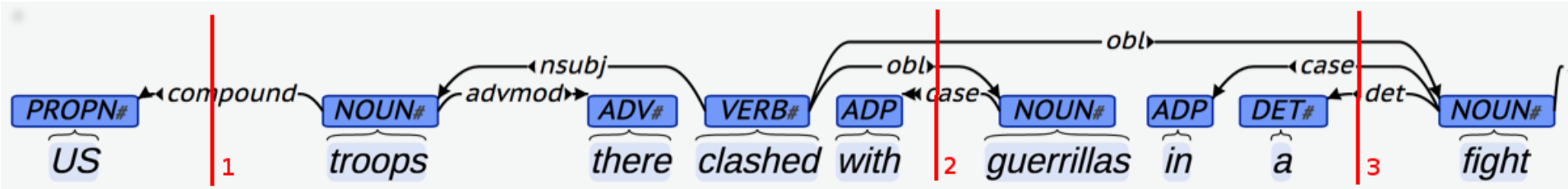
Position 1: R/L ratio = 1

Position 2: R/L ratio = 1

Position 3: R/L ratio = 1/3

Weight

- The *weight* of the flux is the size of the largest disjoint subset of the flux



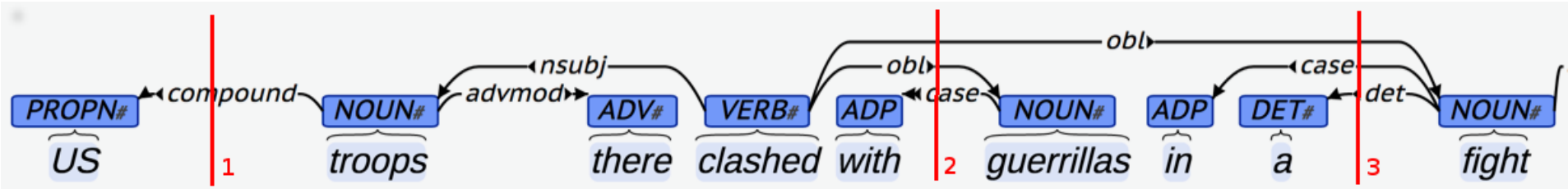
Position 1: weight = 1

Position 2: weight = 2

Position 3: weight = 1

Density

- The *density* of the flux is the weight/size ratio



Position 1: density = 1

Position 2: density = 2/3

Position 3: density = 1/3

Hypotheses

- The size of the flux should be limited and not exceed 7 ± 2
- The weight of the flux is a good measure of the complexity of the sentence
  - does it have a limitation?
  - what is its relation with the 7 ± 2 boundary?
- Initial-headed languages have a R/L ratio higher than 1, and vice versa

Summary

- State of the art
- Dependency flux
- Project UD and method
- Results
- Conclusion

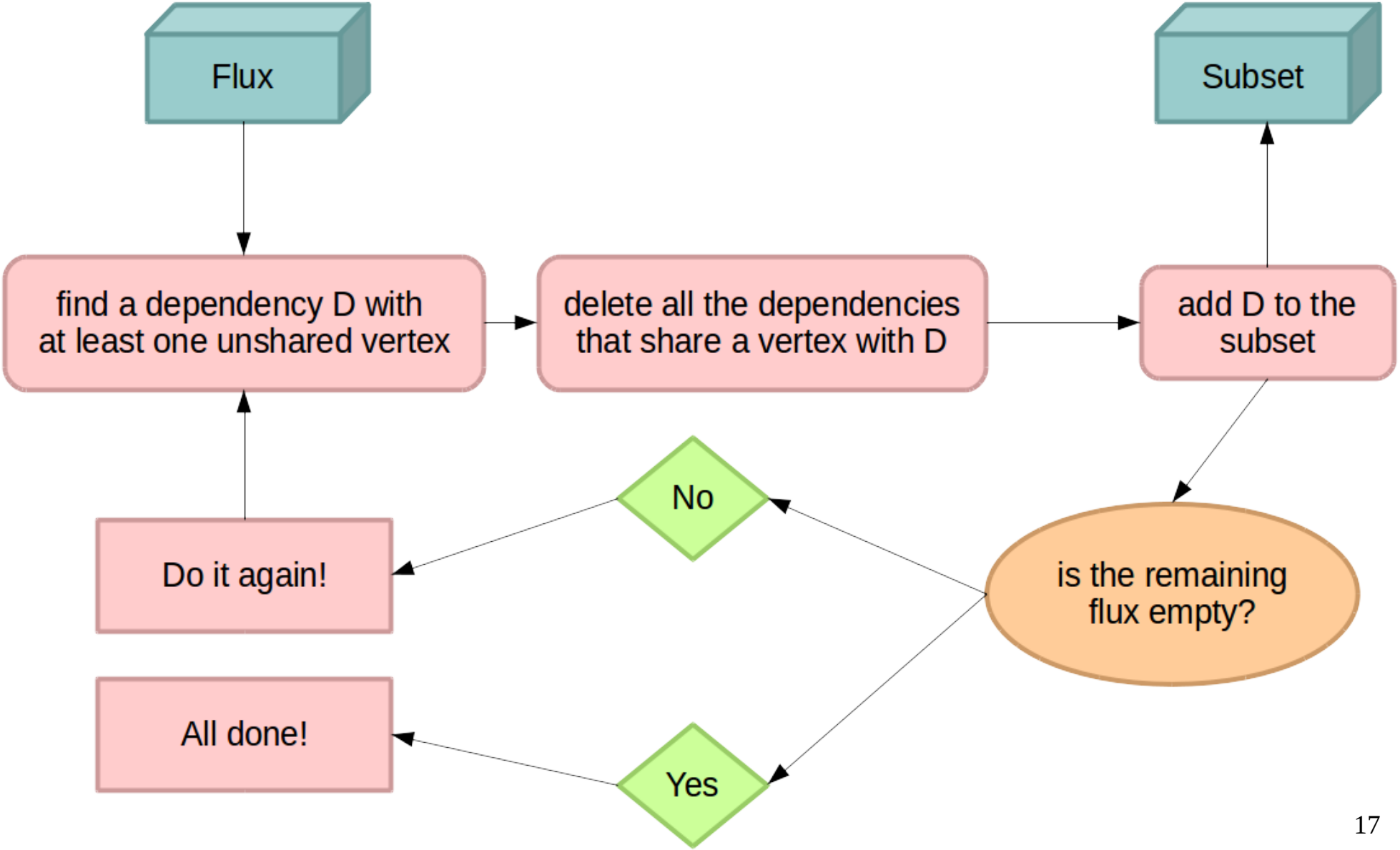
Project UD and method

- The UD scheme is mainly based on an evolution of (universal) Stanford Dependencies (de Marneffe et al., 2014) and Google universal part-of-speech tags (Petrov et al., 2012)
- We used the 70 treebanks in 50 languages distributed by the UD version 2.0 project

Project UD and method

- We calculated
  - flux size
  - flux weight
  - flux density
  - left span and right span
  - R/L ratio

Algorithm for calculating weight



Project UD and method

- All *punct* relations were eliminated
- Punctuation is a specificity of written language
- The annotation of *punct* is not unified in each treebank

Summary

- State of the art
- Dependency flux
- Project UD and method
- Results
- Conclusion

Results

- S-max: maximum size
- S-av: average size
- W-max: maximum weight
- W-av: average weight
- L-max: maximum left span
- R-max: maximum right span
- L-av: average left span
- R-av: average right span
- R/L-av: average R/L ratio
- D-av: density = average W/S ratio

Results

	Tokens	Trees	S-max	S-av	W-max	W-av	L-max	R-max	L-av	R-av	R/L-av	D-av
UD_Ancient_Greek	182030	12613	97	3.01	6	1.49	12	97	2.31	1.99	1.13	60.32%
UD_Arabic	254120	6984	36	2.93	5	1.66	9	35	2.06	2.41	1.32	66.47%
UD_Arabic-NYUAD	738889	19738	78	3.12	6	1.66	12	78	1.95	2.74	1.55	64.65%
UD_Chinese	111271	4497	27	3.24	6	1.65	14	25	2.77	1.86	0.84	61.28%
UD_Croatian	183816	8289	13	2.52	5	1.40	11	13	2.13	1.65	0.98	65.74%
UD_Czech-CLTT	26781	814	28	3.61	7	1.77	10	24	2.36	2.83	1.36	62.24%
UD_English	229733	14545	18	2.58	6	1.35	13	17	2.19	1.58	0.92	63.01%
UD_Finnish	180911	13581	33	2.31	6	1.31	10	33	1.89	1.63	1.06	68.53%
UD_Finnish-FTB	143326	16856	14	2.06	5	1.19	11	14	1.77	1.39	0.98	70.29%
UD_Galician	109106	3139	15	2.56	5	1.41	11	15	2.04	1.80	1.08	64.54%
UD_Gothic	45138	4372	21	2.53	4	1.38	10	20	1.87	1.91	1.23	65.96%
UD_Irish	13826	566	18	2.88	5	1.56	7	18	1.94	2.34	1.37	64.95%
UD_Japanese	173458	7675	15	2.79	5	1.55	15	11	2.17	2.03	1.17	64.52%
UD_Kazakh	529	31	8	2.67	4	1.52	6	5	2.21	1.82	1.00	67.07%
UD_Korean	63426	5350	23	2.73	5	1.62	9	20	2.25	1.93	0.99	68.80%
UD_Latin	18184	1334	17	2.86	5	1.52	8	16	2.31	1.87	1.02	63.32%
UD_Old_Church_Slavonic	47532	5196	20	2.48	5	1.34	8	19	1.76	1.93	1.31	66.03%
UD_Persian	136896	5397	14	3.45	6	1.64	13	10	3.03	1.81	0.77	57.00%
UD_Polish	72763	7127	10	1.92	4	1.18	8	7	1.62	1.40	1.04	72.20%
UD_Sanskrit	1206	190	8	2.23	3	1.29	6	5	2.05	1.39	0.82	68.76%
UD_Slovak	93015	9543	10	2.00	4	1.18	9	8	1.74	1.36	0.96	70.22%
UD_Uyghur	1662	100	8	2.93	5	1.73	7	6	2.75	1.80	0.77	67.31%
UD_Vietnamese	31799	2200	8	2.09	4	1.25	7	8	1.68	1.57	1.12	70.45%

Table 1 (Extract)

Size

Maximum size

- 8 for Kazakh, Sanskrit, Uyghur, and Vietnamese,
- 97 for Ancient-Greek

Average size

- Between 1.92 for Polish and 3.61 for Czech-CLTT.

Left and right span

- Left span is between 7 and 17
- Right span is between 5 and 97

R/L ratio

- Minimum: 0.77 for UD_Persian
- Maximum: 1.55 for UD_Arabic-NYUAD

R/L ratio

- Head-initial languages have the highest R/L ratio
  - 1.31 for Old Church Slavonic
  - 1.37 for Irish
  - 1.55 and 1.32 for Arabic
- Results for head-final languages are not relevant
  - 1.17 for Japanese
  - 1.04 for Turkish
  - 0.99 for Korean .

→Problem with UD scheme, for coordination and MWEs in particular

Weight

Maximum weight

- Between 3 (only for Sanskrit) and 6.
- Most of the flux with a maximum weight we have checked were due to erroneous analysis.

Average weight

- 1.18 for Polish and Slovak, 1.77 for Czech-CLTT.

Results

	Tokens	Trees	1	2	3	4	5	6
UD_Ancient_Greek	182030	12613	57.77%	35.79%	5.96%	0.45%	0.02%	0.00%
UD_Arabic	254120	6984	47.15%	41.10%	10.60%	1.10%	0.05%	0.00%
UD_Arabic-NYUAD	738889	19738	47.16%	40.86%	10.67%	1.23%	0.08%	0.00%
UD_Chinese	111271	4497	49.73%	37.35%	10.91%	1.78%	0.22%	0.01%
UD_Croatian	183816	8289	64.46%	31.70%	3.66%	0.17%	0.01%	0.00%
UD_Czech-CLTT	26781	814	42.78%	41.74%	12.19%	2.71%	0.53%	0.05%
UD_English	229733	14545	68.45%	28.08%	3.29%	0.17%	0.00%	0.00%
UD_Finnish	180911	13581	72.60%	23.79%	3.28%	0.30%	0.03%	0.00%
UD_Finnish-FTB	143326	16856	82.77%	15.91%	1.22%	0.10%	0.00%	0.00%
UD_Galician	109106	3139	62.78%	33.73%	3.34%	0.14%	0.00%	0.00%
UD_Gothic	45138	4372	65.83%	30.51%	3.46%	0.21%	0.00%	0.00%
UD_Irish	13826	566	53.11%	38.57%	7.47%	0.82%	0.03%	0.00%
UD_Japanese	173458	7675	50.98%	42.82%	6.03%	0.17%	0.00%	0.00%
UD_Kazakh	529	31	55.27%	37.42%	6.88%	0.43%	0.00%	0.00%
UD_Korean	63426	5350	51.30%	37.10%	10.24%	1.28%	0.09%	0.00%
UD_Latin	18184	1334	56.34%	35.90%	7.01%	0.69%	0.06%	0.00%
UD_Old_Church_Slavonic	47532	5196	69.31%	27.41%	3.15%	0.13%	0.00%	0.00%
UD_Persian	136896	5397	45.75%	45.14%	8.52%	0.59%	0.01%	0.00%
UD_Polish	72763	7127	82.46%	16.96%	0.57%	0.00%	0.00%	0.00%
UD_Sanskrit	1206	190	71.95%	27.07%	0.98%	0.00%	0.00%	0.00%
UD_Slovak	93015	9543	83.09%	16.24%	0.66%	0.01%	0.00%	0.00%
UD_Uyghur	1662	100	43.87%	42.16%	11.50%	2.12%	0.34%	0.00%
UD_Vietnamese	31799	2200	76.10%	22.71%	1.18%	0.01%	0.00%	0.00%

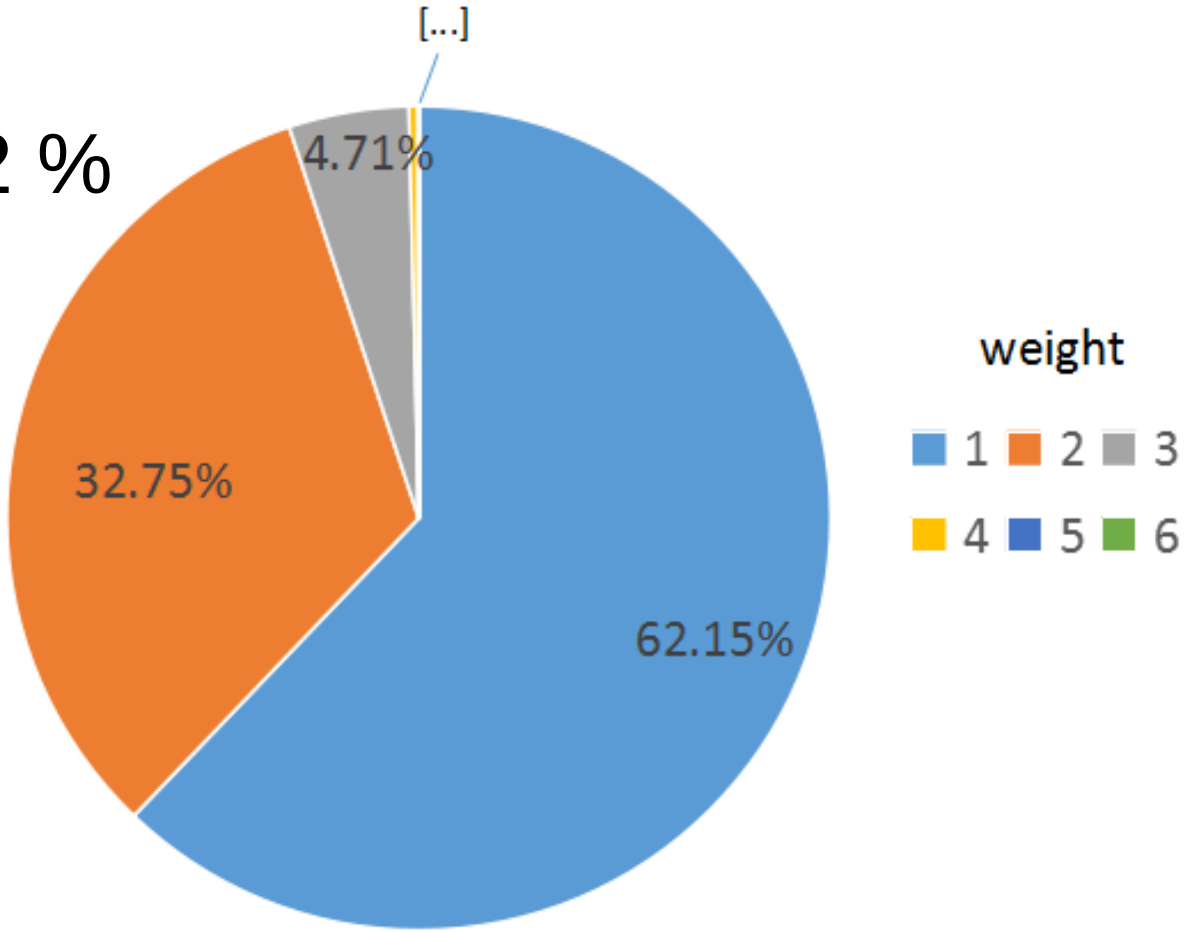
Table 2 (Extract)

Weight

- 99.9% of inter-words positions have a flux weight that is less than 3 in Polish, Sanskrit, Slovak, and Vietnamese
- More than 10% of inter-words positions have a flux weight that equals 3 in Arabic, Chinese, and Korean
- More than 80 % of inter-words positions have a flux weight that equals 1 in Finnish-FTB, Polish, and Slovak

Weight

- Weight ≤ 3: 99.62 %



Summary

- State of the art
- Dependency flux
- Project UD and method
- Results
- Conclusion

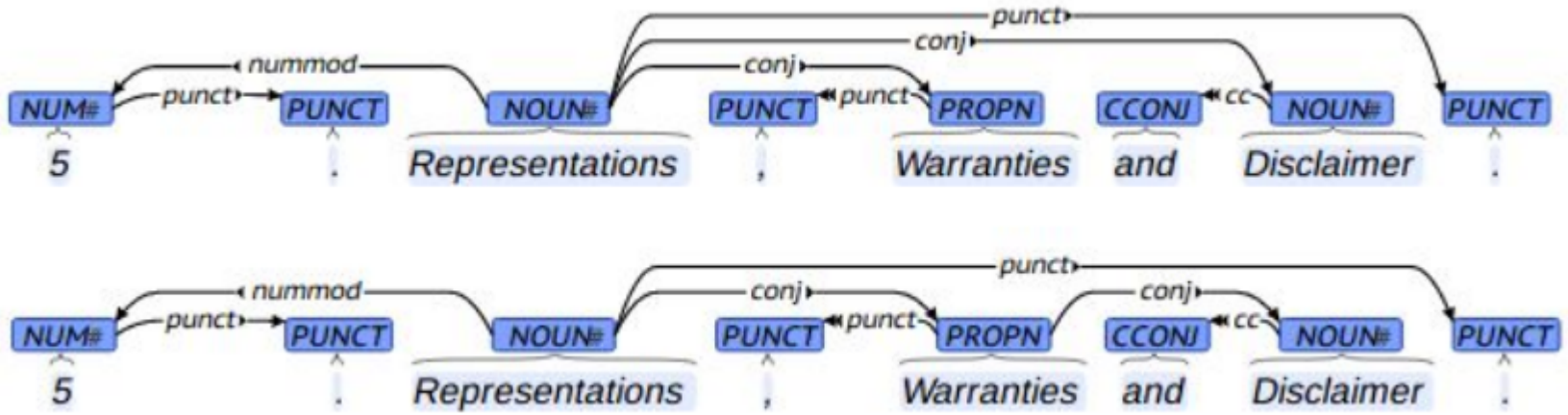
Conclusion

- The size, as well as the left and right spans, of the flux can vary a lot according to the corpus and its language and are not clearly bounded.
- The values of dependency flux weight appear to be more homogeneous.
- The weight is bounded by 5 except for a very few positions, related to short-term memory limitations.

Thank you!

In the UD schema, coordination is annotated in a bouquet. This choice obviously influences the flux size.

Bouquet vs chain annotation (Gerdes & Kahane, 2009)

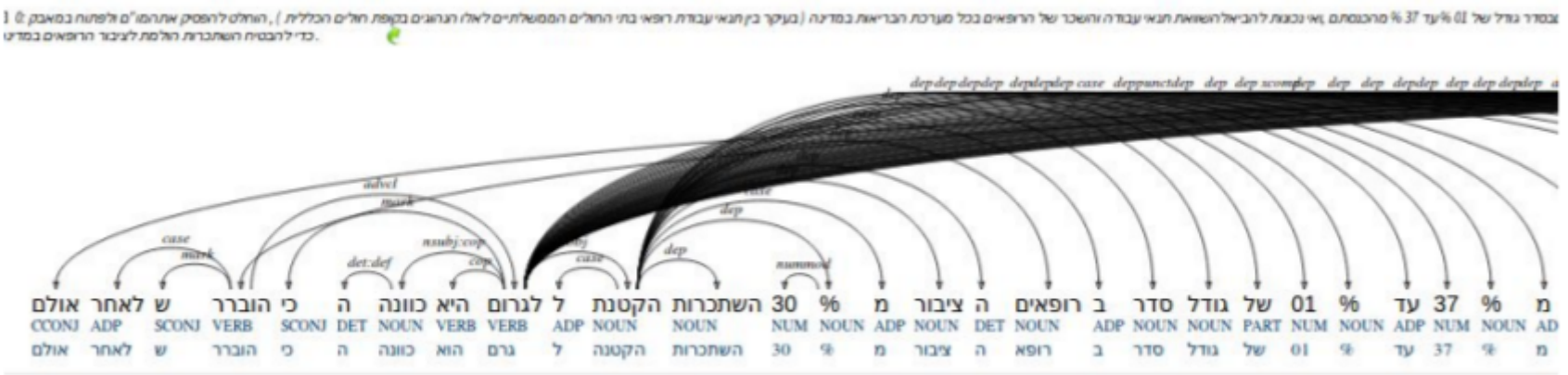


The chain representation => narrower flux

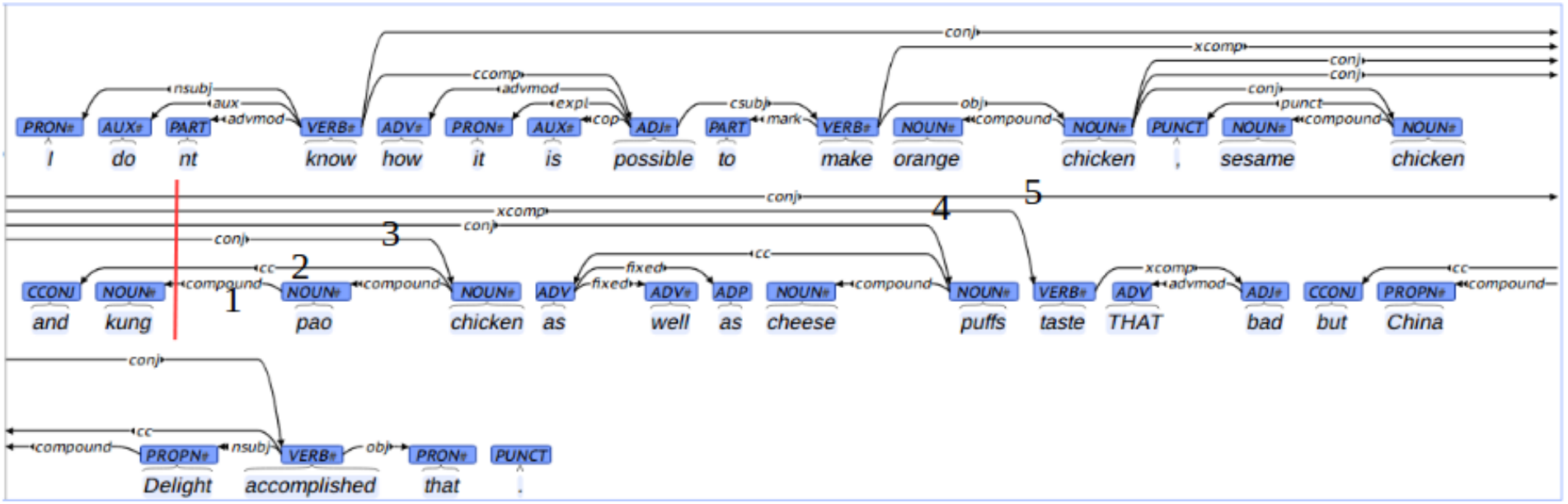
The maximum flux size decreases to 6 (Sanskrit) and 77 (Arabic-NYUAD)

1.89 (Polish) and 3.44 (Persian) for the size averages.

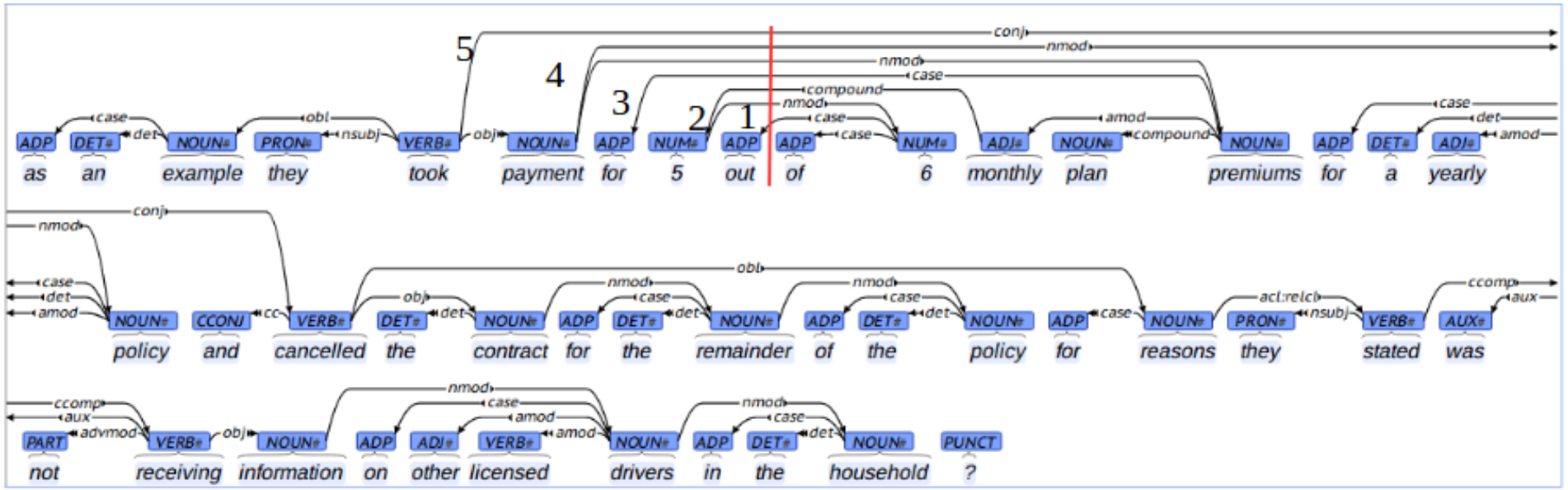
Difficult to treat some relations like the relation dep or the relation nmod.



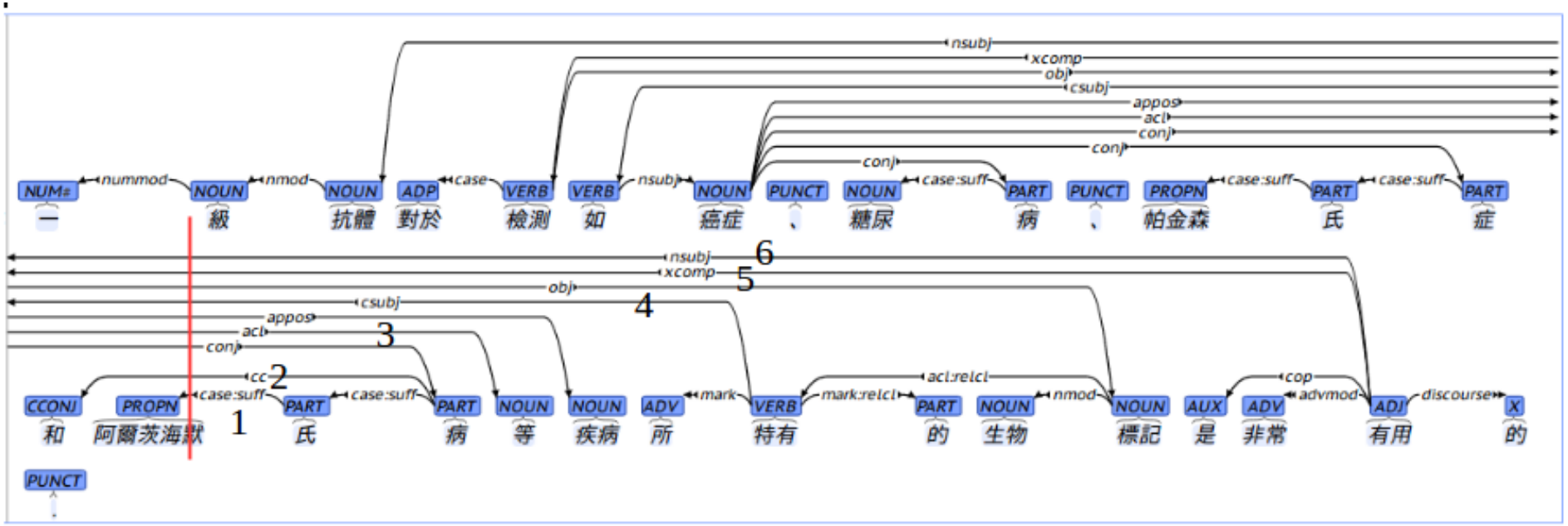
Example 1


PRON AUX PART VERB ADV PRON AUX ADJ PART VERB NOUN NOUN PUNCT NOUN NOUN
I do nt know how it is possible to make orange chicken , sesame chicken
nsubj aux advmod ccomp advmod expl cop csubj mark obj compound conj xcomp conj conj conj punct compound
CCONJ NOUN NOUN NOUN ADV ADV ADP NOUN NOUN VERB ADV ADJ CCONJ PROPN
and kung pao chicken as well as cheese puffs taste THAT bad but China
cc compound fixed cc xcomp advmod
1 2 3 4 5
PROPN VERB PRON PUNCT
Delight accomplished that .
cc compound nsubj obj

Example 2


5
4
3
2
1
ADP DET NOUN PRON VERB NOUN ADP NUM ADP ADP NUM ADJ NOUN NOUN ADP DET ADJ
as an example they took payment for 5 out of 6 monthly plan premiums for a yearly
NOUN CCONJ VERB DET NOUN ADP DET NOUN ADP DET NOUN ADP NOUN PRON VERB AUX
policy and cancelled the contract for the remainder of the policy for reasons they stated was
PART VERB NOUN ADP ADJ VERB NOUN ADP DET NOUN PUNCT
not receiving information on other licensed drivers in the household ?

Example 3


nsubj
xcomp
obj
csubj
appos
acl
conj
conj
NUM# nummod NOUN nmod NOUN ADP case VERB VERB nsubj NOUN PUNCT NOUN case:suff PART PUNCT PROPN case:suff PART case:suff PART
一 級 抗體 對於 檢測 如 癌症 、 糖尿 病 、 帕金森 氏 症
nsubj
xcomp
obj
csubj
appos
acl
conj
cc
CCONJ PROPN case:suff PART case:suff PART NOUN NOUN ADV mark VERB mark:relcl PART NOUN nmod NOUN AUX ADV cop advmod ADJ discourse X
和 阿爾茨海默 氏 病 等 疾病 所 特有 的 生物 標記 是 非常 有用 的
PUNCT
1
2
3
4
5
6