Processing non-canonical data: challenges & opportunities

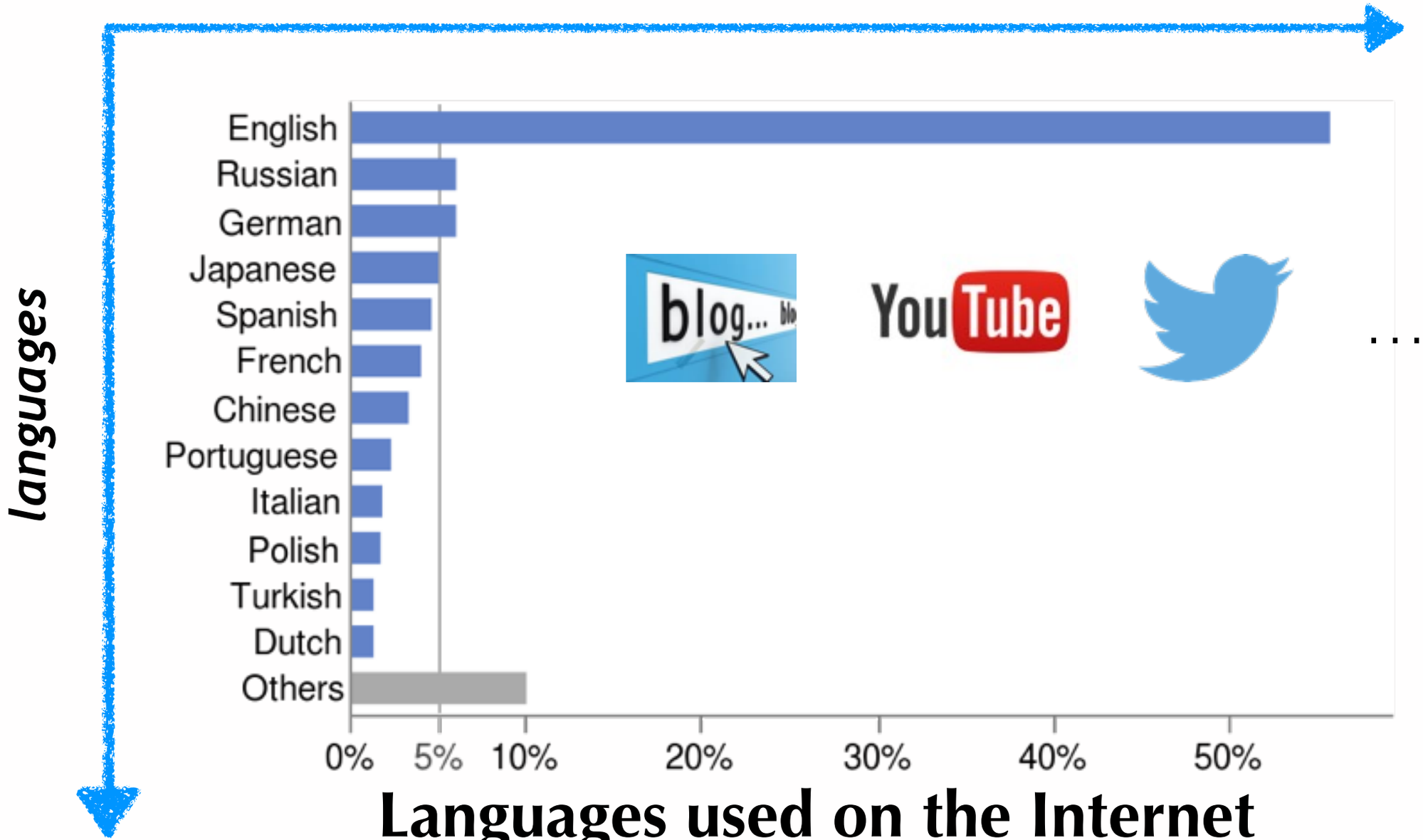
@barbara_plank
University of Groningen, NL

BAULT 2016
Helsinki, 02-12

Goal: NLP for everyone

domains

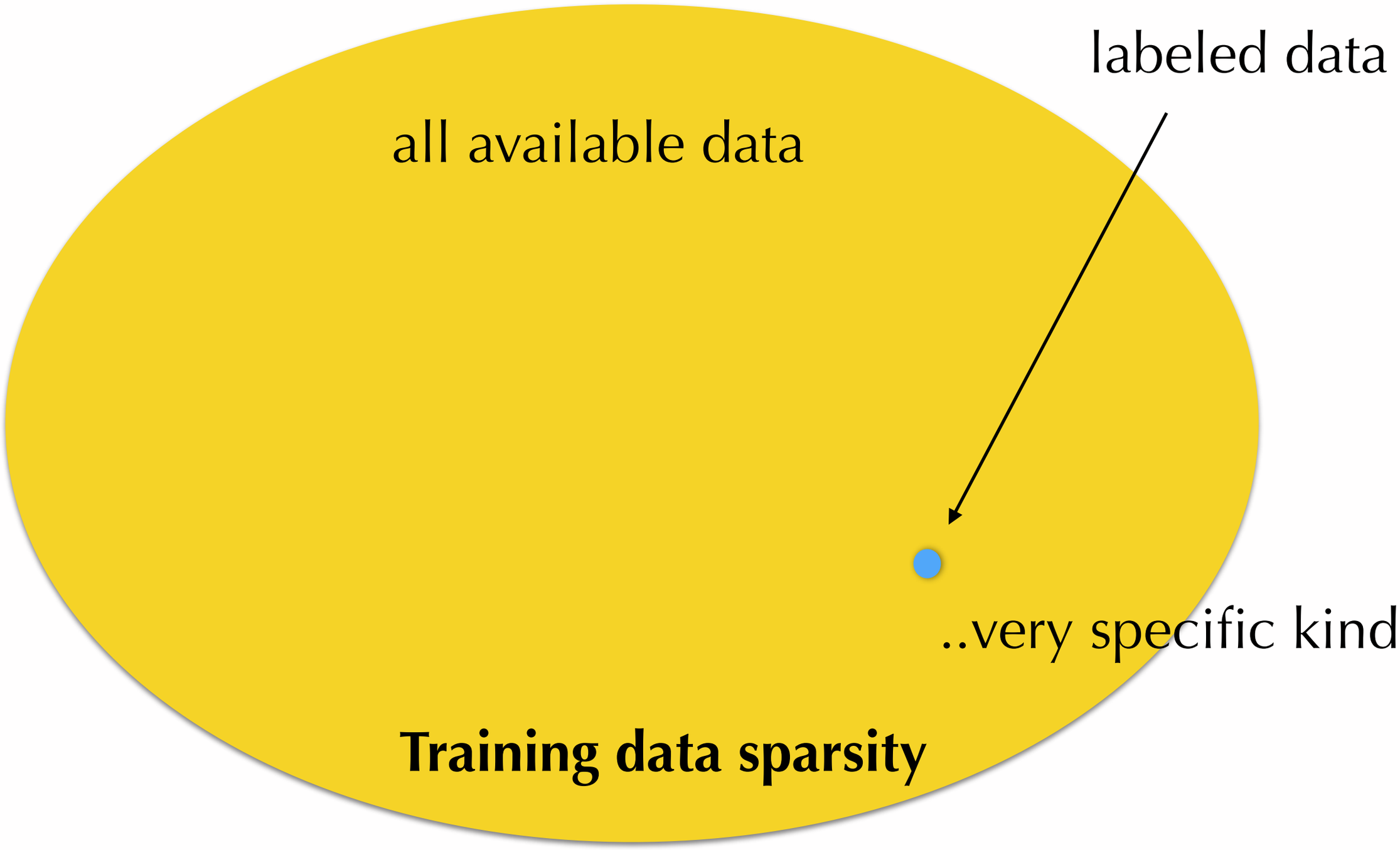
languages



Languages used on the Internet

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

DATA:
NEED: a lot.. — HAVE: a little



The problem

- NLP models are trained on samples from a **limited set** of **canonical data**
  - Mainly: English **newswire**

CROSS-DOMAIN GULF

The consequence (1/2)

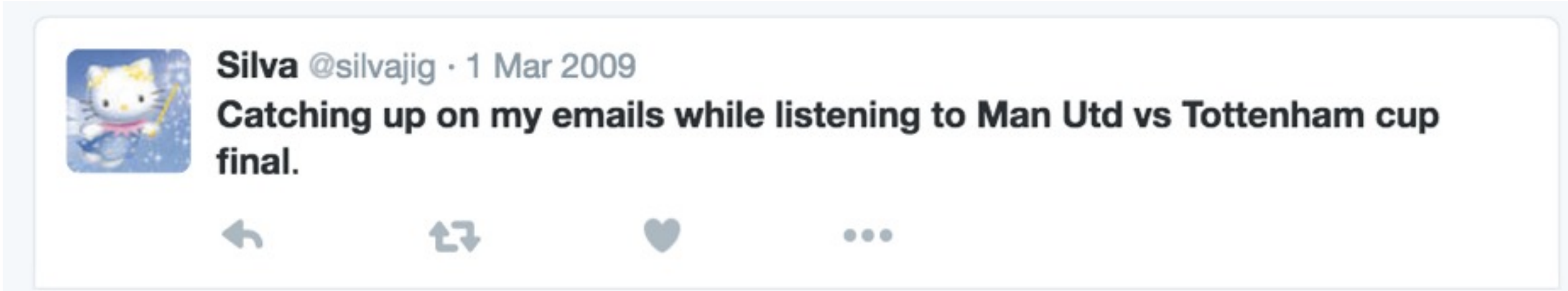
Sanchit Vir Gogia @s_v_g · Apr 19
#**INTJ** via @PersonalityHack youtu.be/gzDAaK1WeB4 >> **IMHO**, this pretty much nails it. #personalitytypes

#/ # NNP/ INTJ IN/ via IN/ @ NNS/ PersonalityHacks NN/ youtube.be/gzDAaK1WeB4 NN/ >> NNP/ IMHO , , DT/ this RB/ pretty JJ/ much NNS/ nails NN/ it. #/ # NNS/ personalitytypes

http://cogcomp.cs.illinois.edu/page/demo_view/POS

Example: Tagging Twitter #hard

The consequence (2/2)




```
          [Catching]_VP   [up]_PRT  ...
[listening]_VP   [to]/PP  [Man]/NP  [Utd]/VP
                    [vs]/NP
      [tottenham cup]/NP   [final]/ADJP
```

Example: Chunking Twitter #hard

Still newswire?

Training data sparsity

subset of treebanks from Universal Dependencies v1.3 (Nivre et al. 2016) for which domain/genre info is available

	news	*fiction*	*nonfict.*	*blog*	*bible*	*legal*	*medical*	*social*	*spoken*	*wiki*	*web*	*reviews*
Anc. Greek		✓	✓		✓							
Arabic	✓											
Basque	✓	✓										
Bulgarian	✓	✓				✓						
Catalan	✓											
Chinese										✓		
Croatian	✓									✓		
Czech	✓		✓			✓	✓					✓
Danish	✓	✓	✓						✓			
Dutch	✓						✓			✓		
English		✓	✓	✓				✓	✓		✓	✓
Estonian	✓	✓										
Finnish	✓	✓		✓		✓				✓		
French	✓			✓						✓		✓
Galician	✓		✓			✓	✓					
German	✓									✓		✓
Gothic					✓							
Greek	✓								✓	✓		
Hebrew	✓											
Hindi	✓											
Hungarian	✓											
Indonesian	✓			✓								
Irish	✓	✓				✓					✓	
Italian	✓					✓				✓		
Kazakh		✓								✓		
Latin		✓	✓		✓							
Latvian	✓											
Norwegian	✓		✓	✓								
O.Slavonic					✓							
Persian	✓	✓	✓			✓	✓	✓	✓			
Polish	✓	✓	✓									
Portuguese	✓			✓								
Romanian	✓	✓	✓			✓	✓			✓		
Russian	✓	✓	✓							✓		
Slovenian	✓	✓	✓						✓			
Spanish	✓			✓						✓		✓
Swedish	✓	✓	✓						✓			
Tamil	✓											
Turkish	✓		✓									

(Plank, 2016 KONVENS)

Where are we? How good is it?

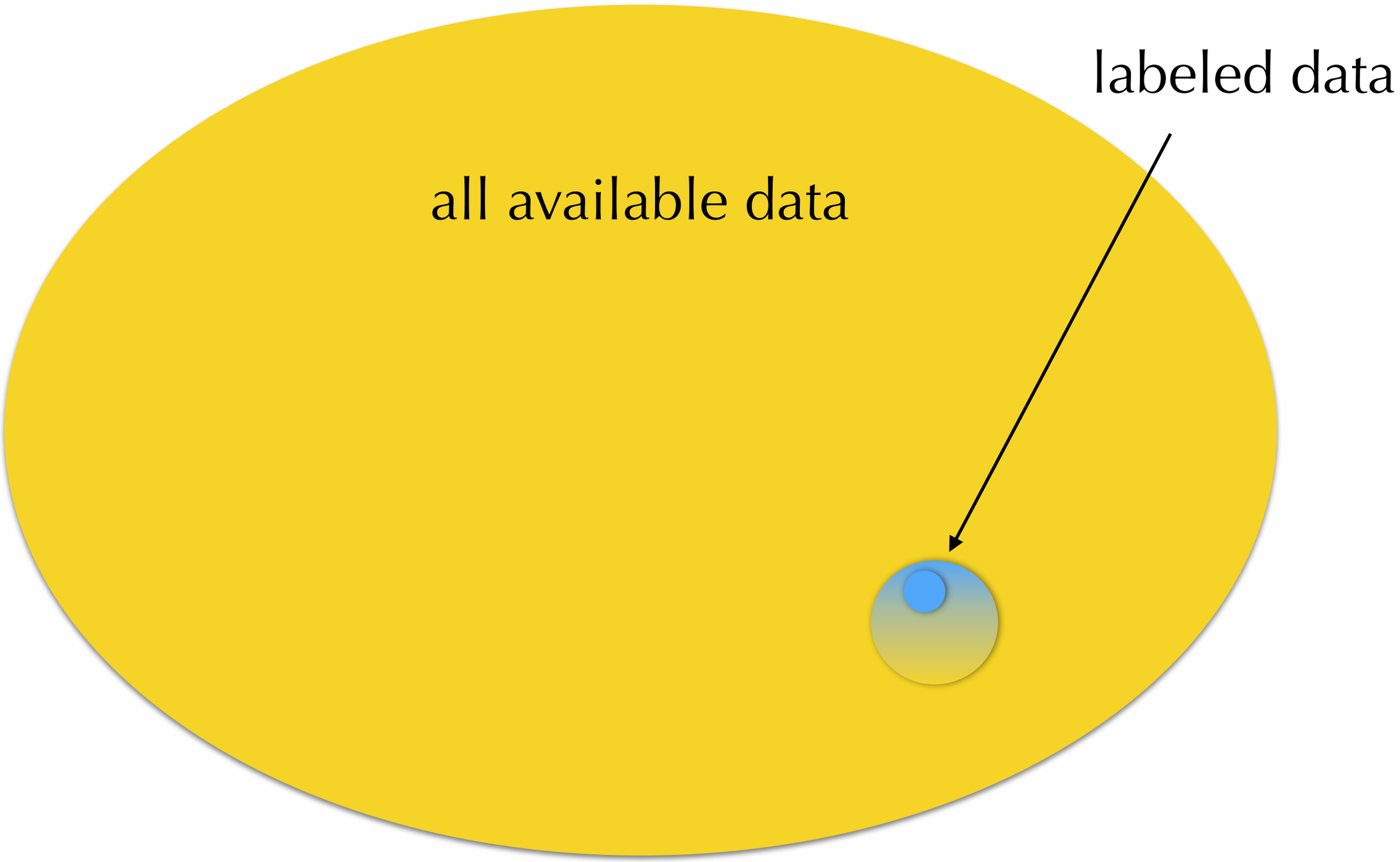
"well, it depends"

Mikko Kurimo yesterday

[for speech] → ***same for language processing***

variability

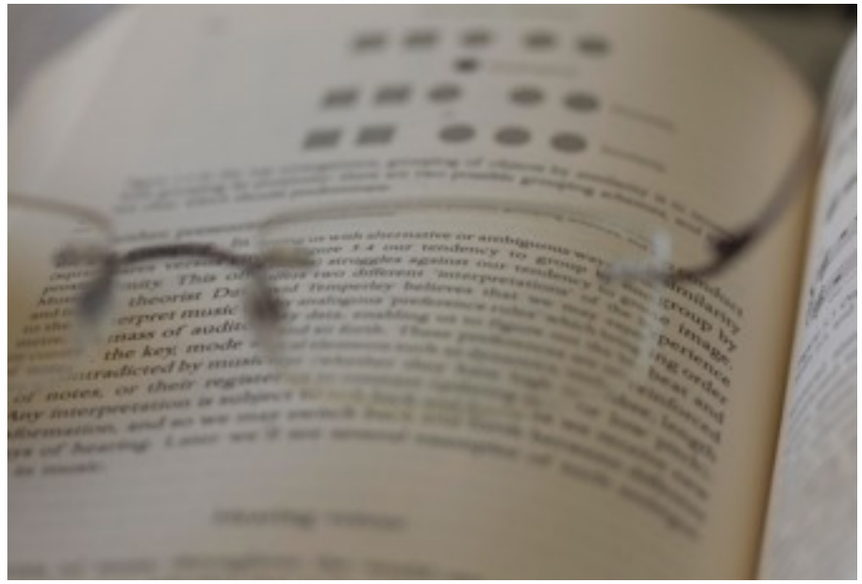
What can we do about it?



One possible way to go about it:

Cognitive Processing Data

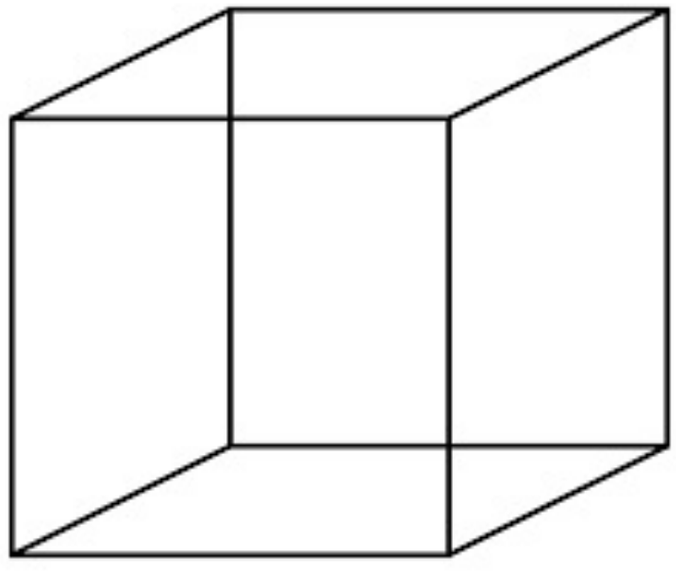
Processing data: What?



- When people **read** or **produce** texts they *unconsciously produce loads of cognitive by-product* (e.g., gaze patterns, keystroke logs)
- Additional side benefit!

Processing data: Why?

- NLP so far **mostly on textual input alone**
- Idea: harvest & use additional signal from **non-obvious** sources **(fortuitous data)**



Fortuitous data

(Plank, 2016 KONVENS)

Define fortuitous!

fortuitous

/fɔːˈtjuːɪtəs/

adjective

happening by chance rather than intention.
"the similarity between the paintings may not be simply fortuitous"
synonyms: chance, unexpected, unanticipated, unpredictable, unforeseen, unlooked-for, serendipitous, casual, incidental, coincidental, haphazard, random, accidental, inadvertent, unintentional, unintended, unplanned, unpremeditated
"his success depended on entirely fortuitous events"

- happening by a lucky chance; fortunate.
  "the ball went into the goal by a fortuitous ricochet"
  synonyms: lucky, fortunate, providential, advantageous, timely, opportune, serendipitous, expedient, heaven-sent, auspicious, propitious, felicitous, convenient, apt; More

Fortuitous data

- Data out there,

that waits to be harvested (**availability**),

and can be used (relatively) easily (**readiness**)

How can fortuitous data help?

- Reuse **data that was not explicitly annotated**
- **Gather data from new varieties quickly** to build more robust models

Examples of fortuitous data

Type / Side benefit of	Examples	Availability	Readiness
user-generated data	hyperlinks, HTML markup, unlabeled data	+	+
annotation	annotator disagreement	-	+
behavior	cognitive processing data	+	-

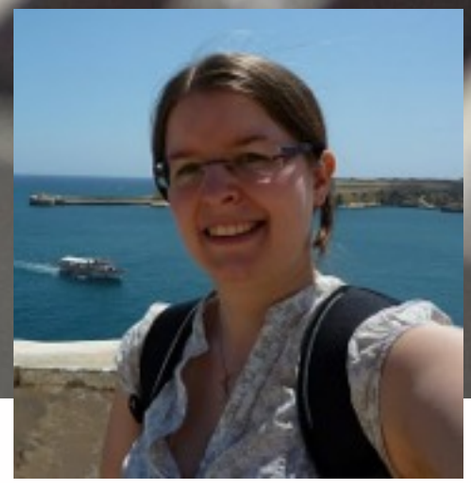
Keystroke dynamics as source for shallow syntactic parsing

Barbara Plank
University of Groningen
The Netherlands
b.plank@rug.nl

COLING 2016

timepress,timerelease,keycode,keyname
1304433167859,1304433168307,16,shift
1304433168227,1304433168371,67,c
1304433168291,1304433168451,79,o
1304433170051,1304433170179,69,e
1304433170451,1304433170531,70,f
1304433170579,1304433170675,70,f
1304433170675,1304433170851,73,i
1304433171171,1304433171299,6

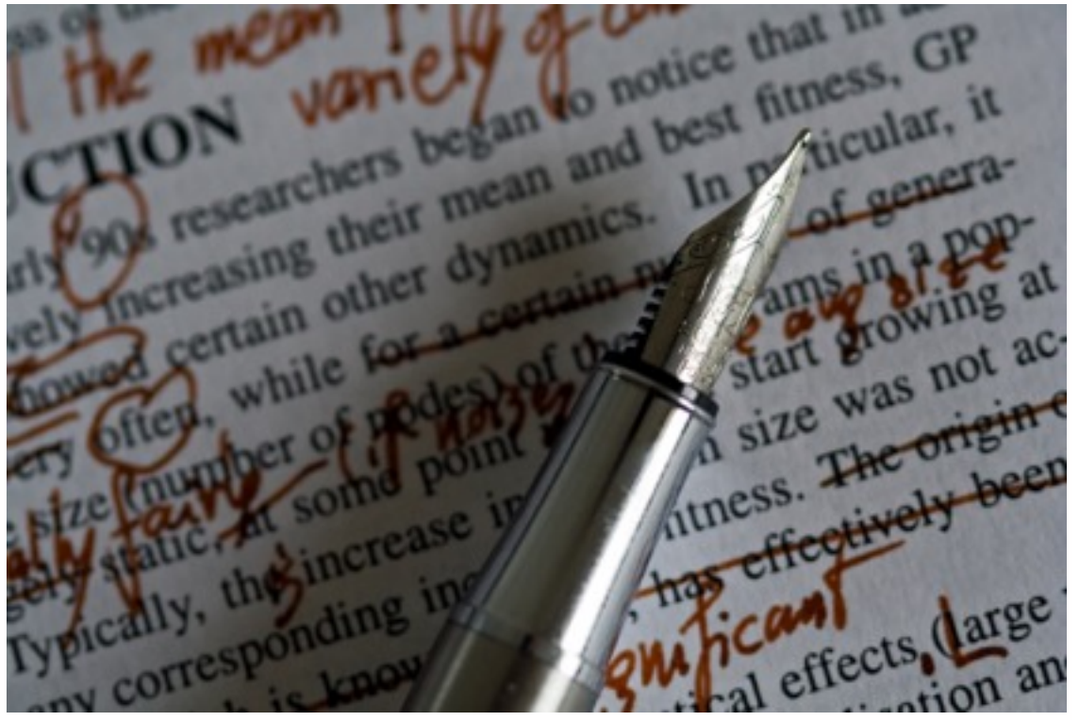
LOADS of
(noisy) data



Keystrokes logs

Cognitive writing research:

- How did you write that essay?



non-intrusive!

Keystrokes logs

Security/user authentication/profiling:

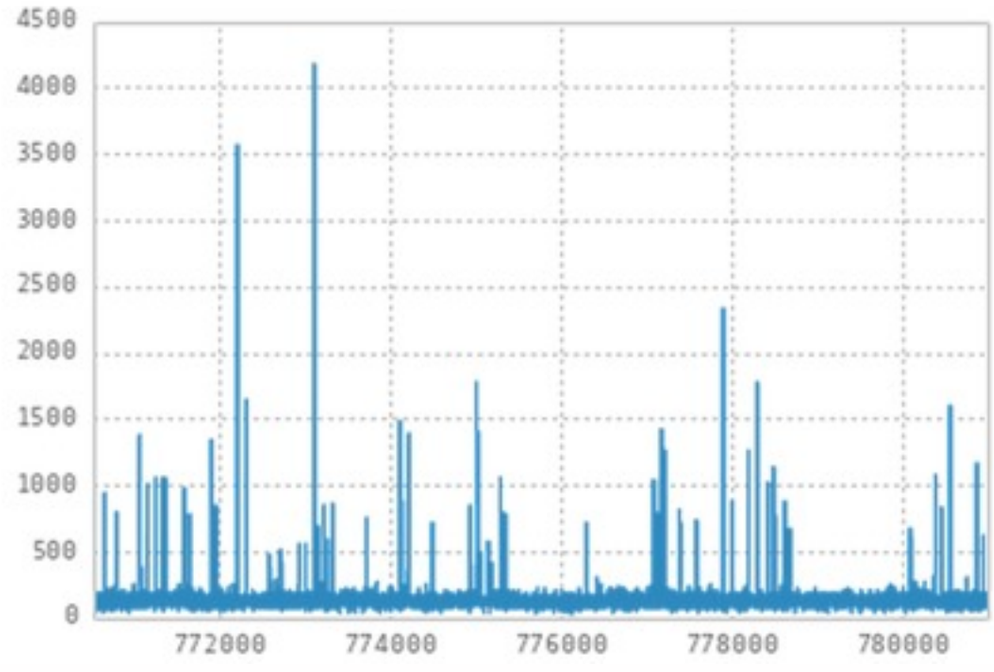
- Did you actually write that essay?

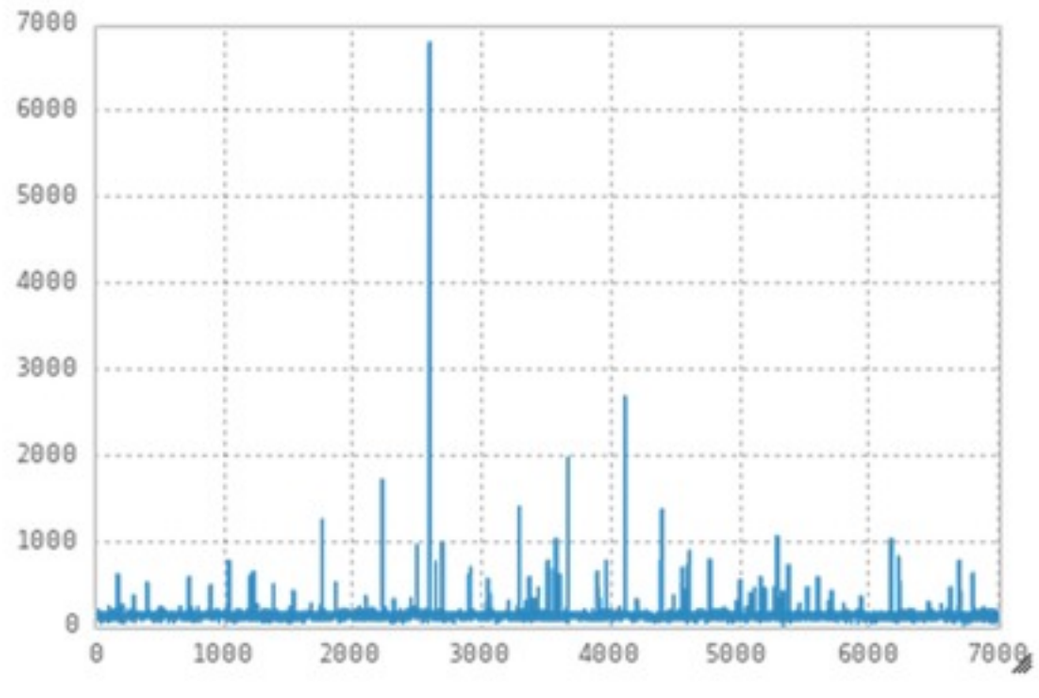
Keystrokes logs

In experimental research (MT research)

- Do glossaries help machine translation (MT) post-editing?

Do pre-word pauses carry syntactic information?

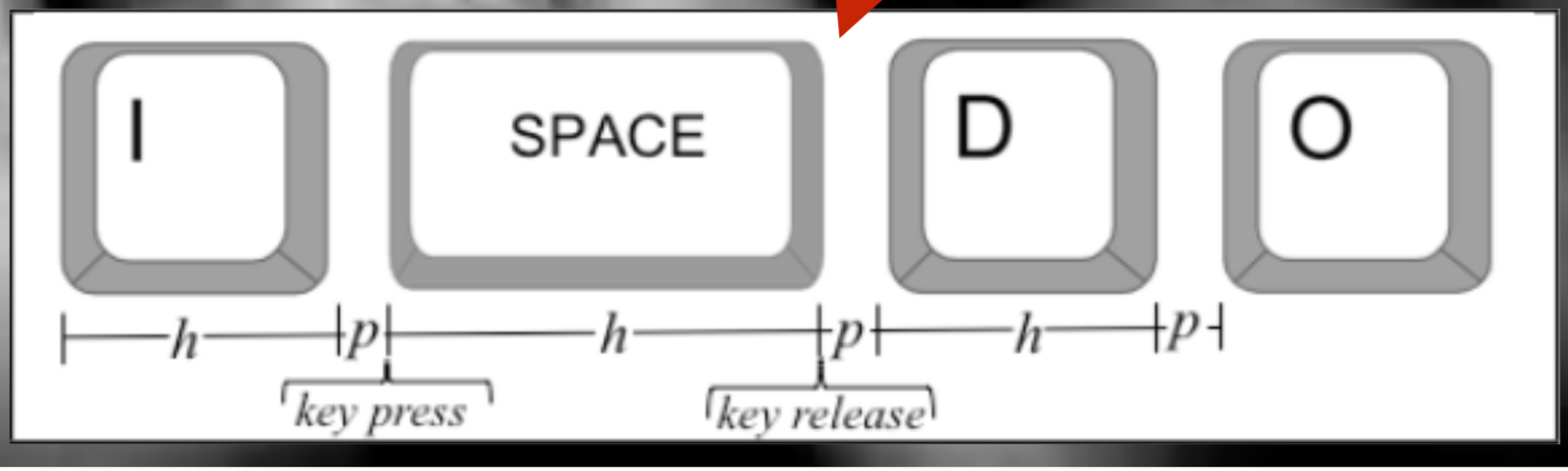




Hypothesis:

keystroke dynamics contain information about **syntactic structure** that can inform shallow syntactic parsing

pauses before words (ms)

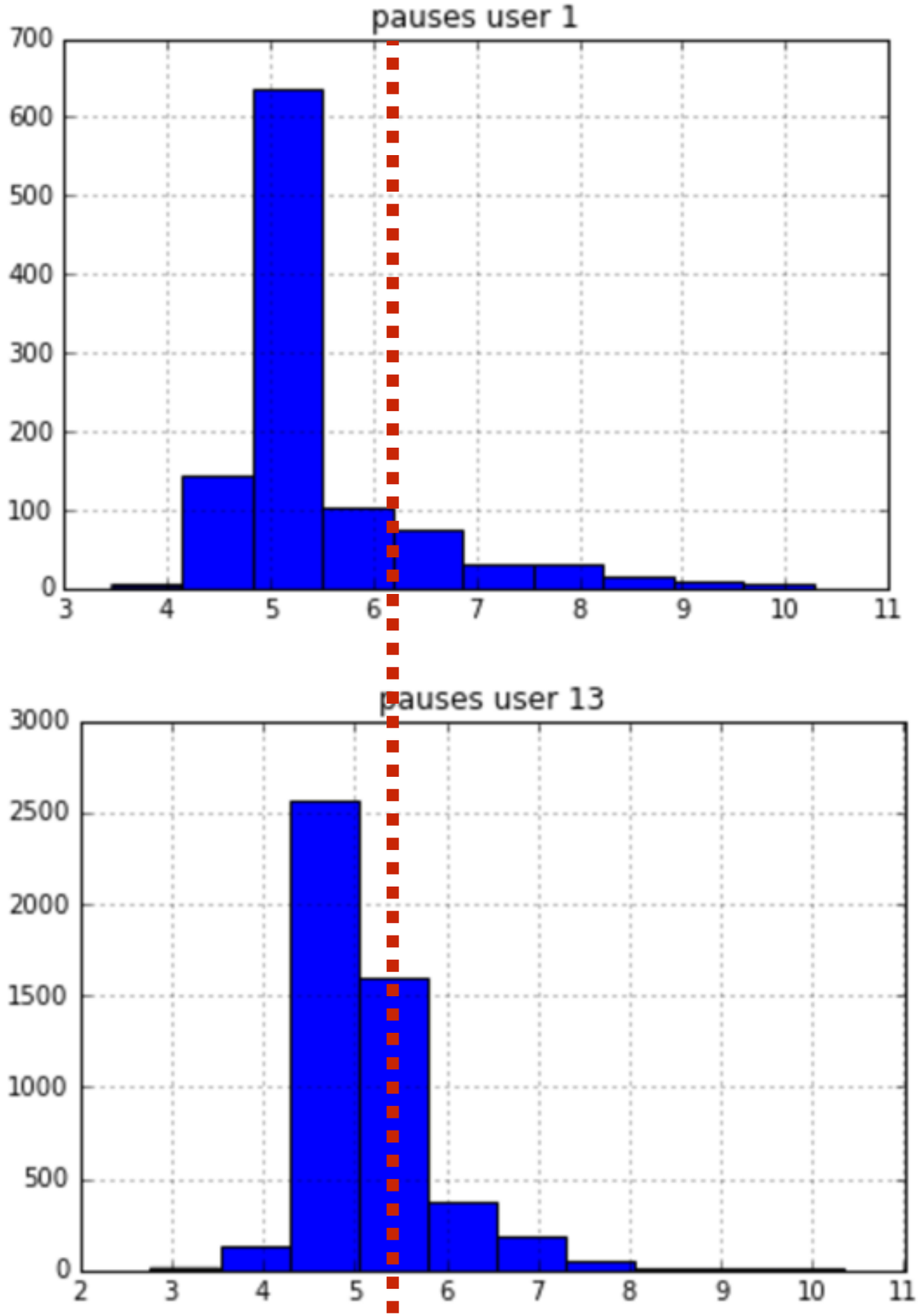


>500ms pauses

==========	
'is'	**[is a]**
'a'	
==========	
'measure'	**[measure**
'used'	**used**
'in'	**in]**
==========	
'statisitcal'	**[statisitcal** *
'model'	**model]**
==========	
'analysis'	**[analysis]**

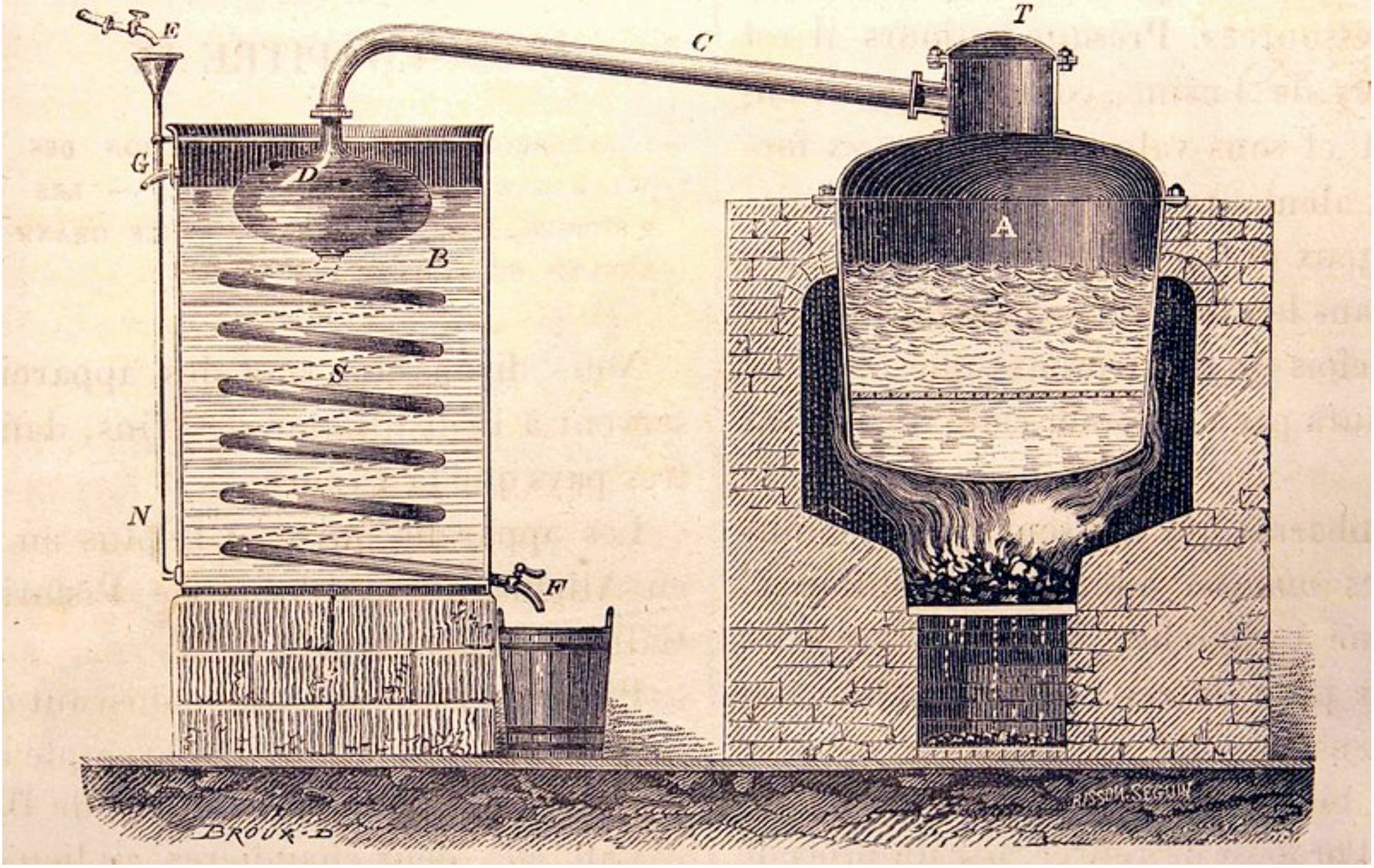
* typo in actual data

500ms
seems
arbitrary



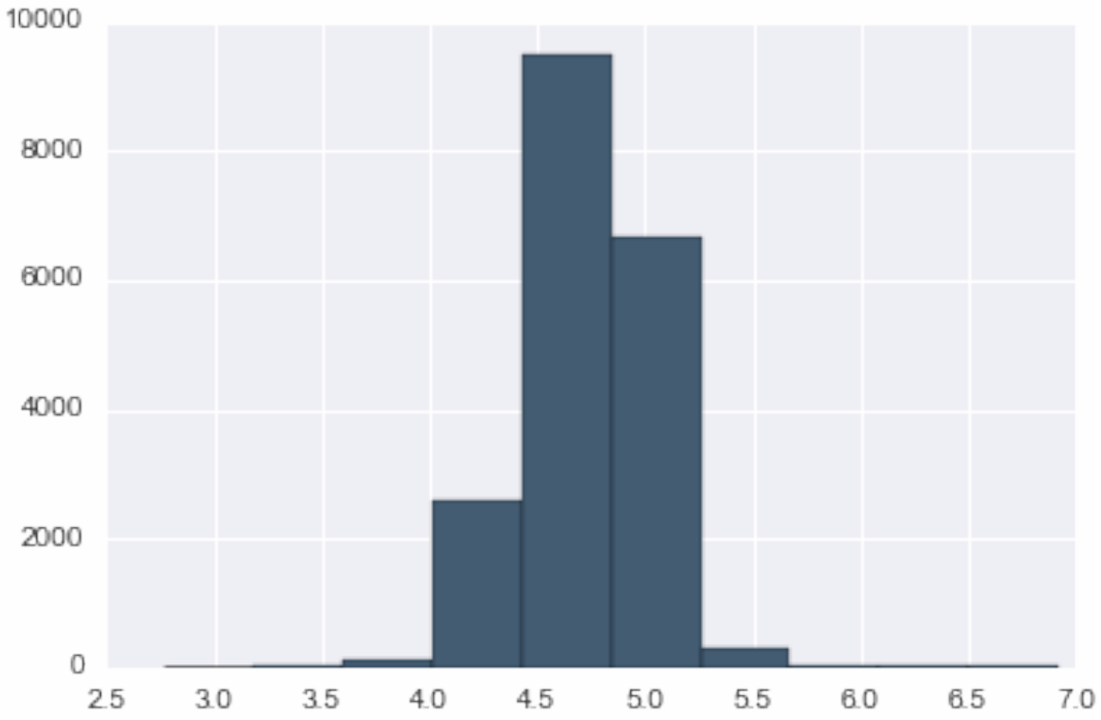
Can we use keystroke logs to inform NLP?

https://commons.wikimedia.org/wiki/File:Alambicdouble_fond_pour_la_distillation_des_marcs_de_raisin.jpg

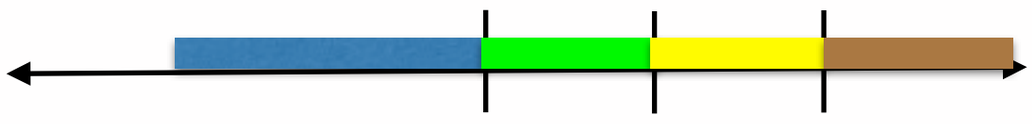


Step 1: *refine* the data

From keystrokes to auxiliary labels



short middle long



m *+mad*

the closer the number is to 1

Derive Labels

[the]
[closer]
[the number]
[is to]
[1]

Step 2: *train* model

multi-task learning

X	Y
i	B-NP
luv	B-VP
jaxx	B-NP

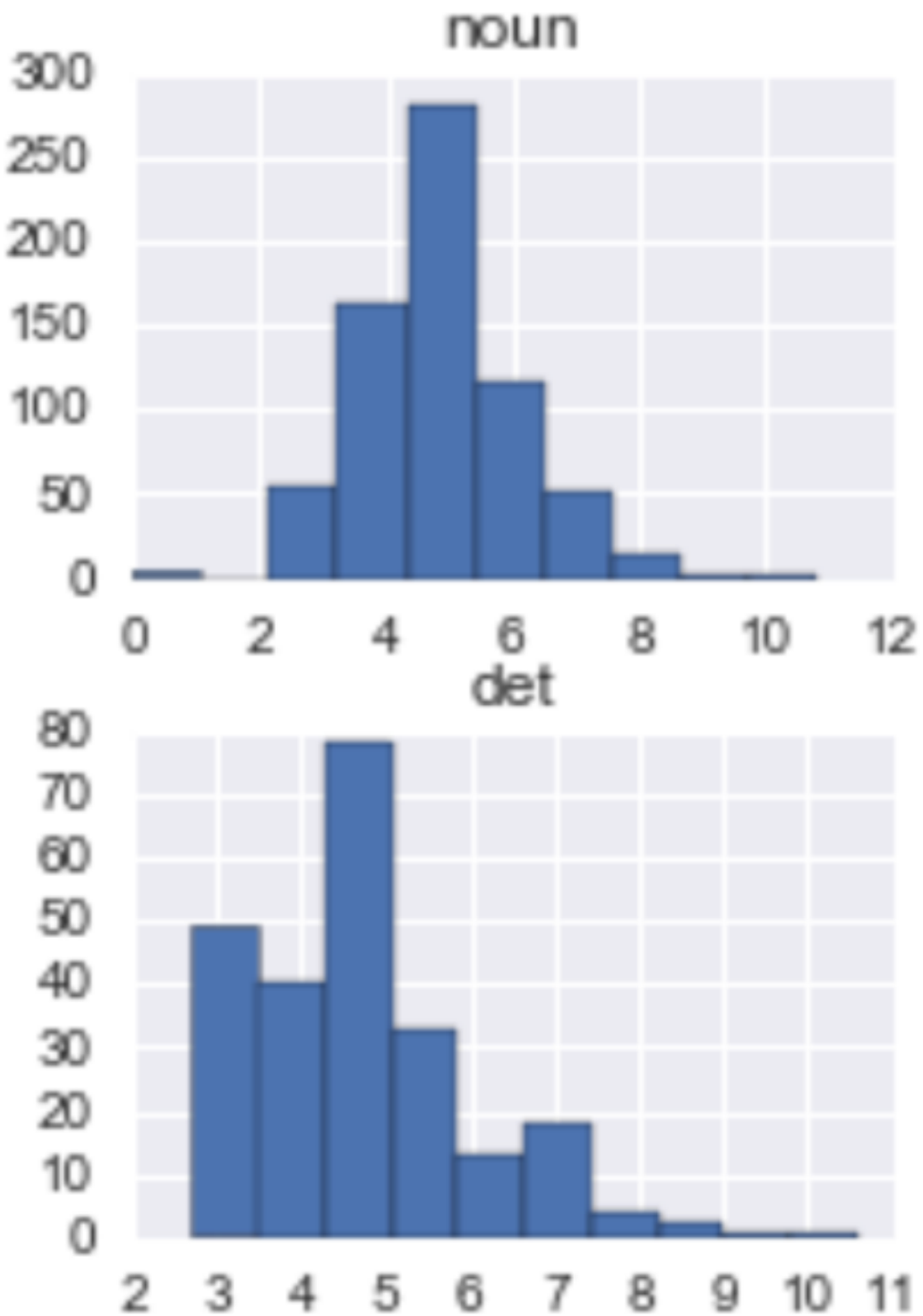
learn from BOTH

X	Y
a	B-m1
pop	B-m2
quiz	I-m2

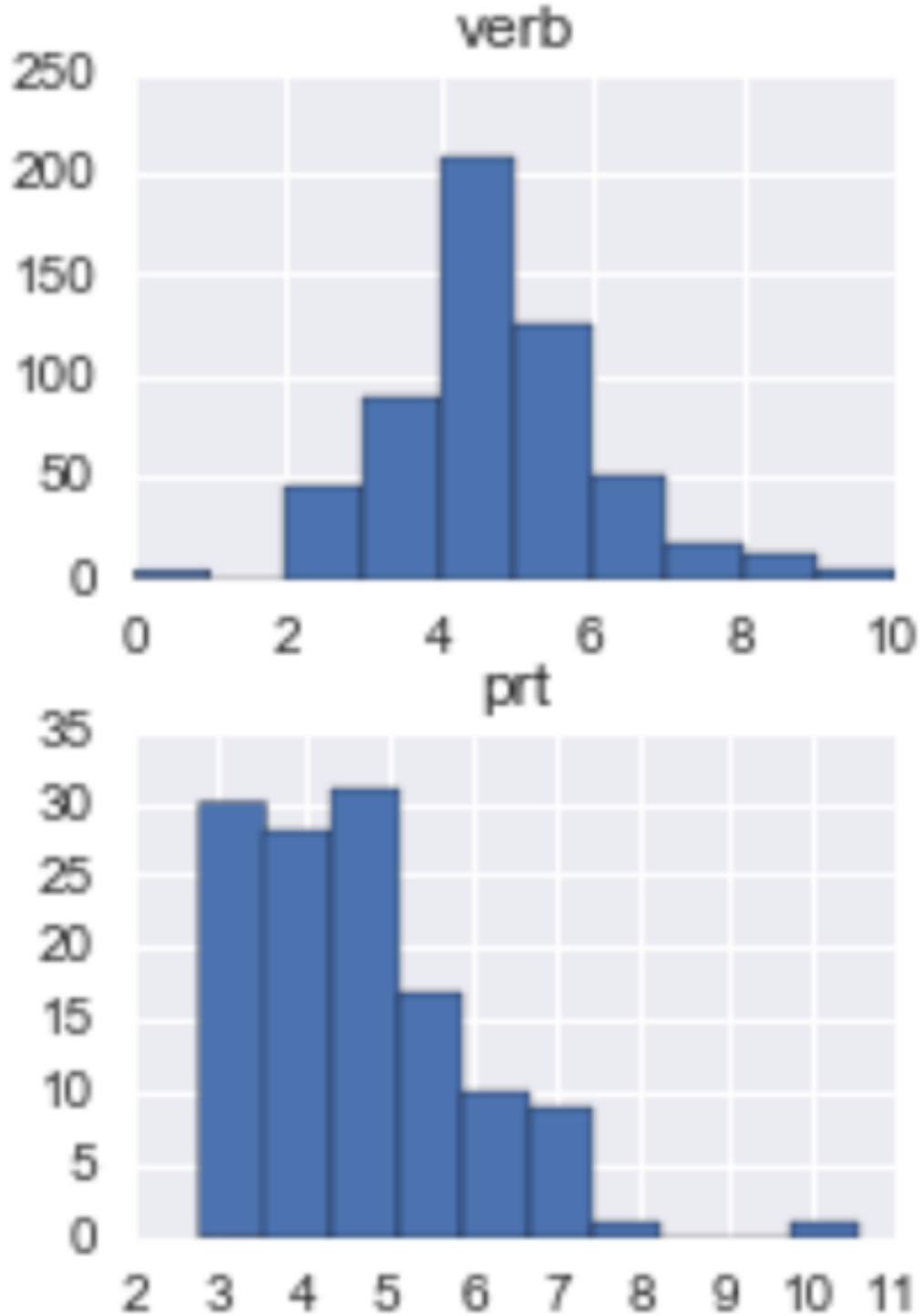
auxiliary data (distinct source!)

Does it help?

First, a look at keystroke data..

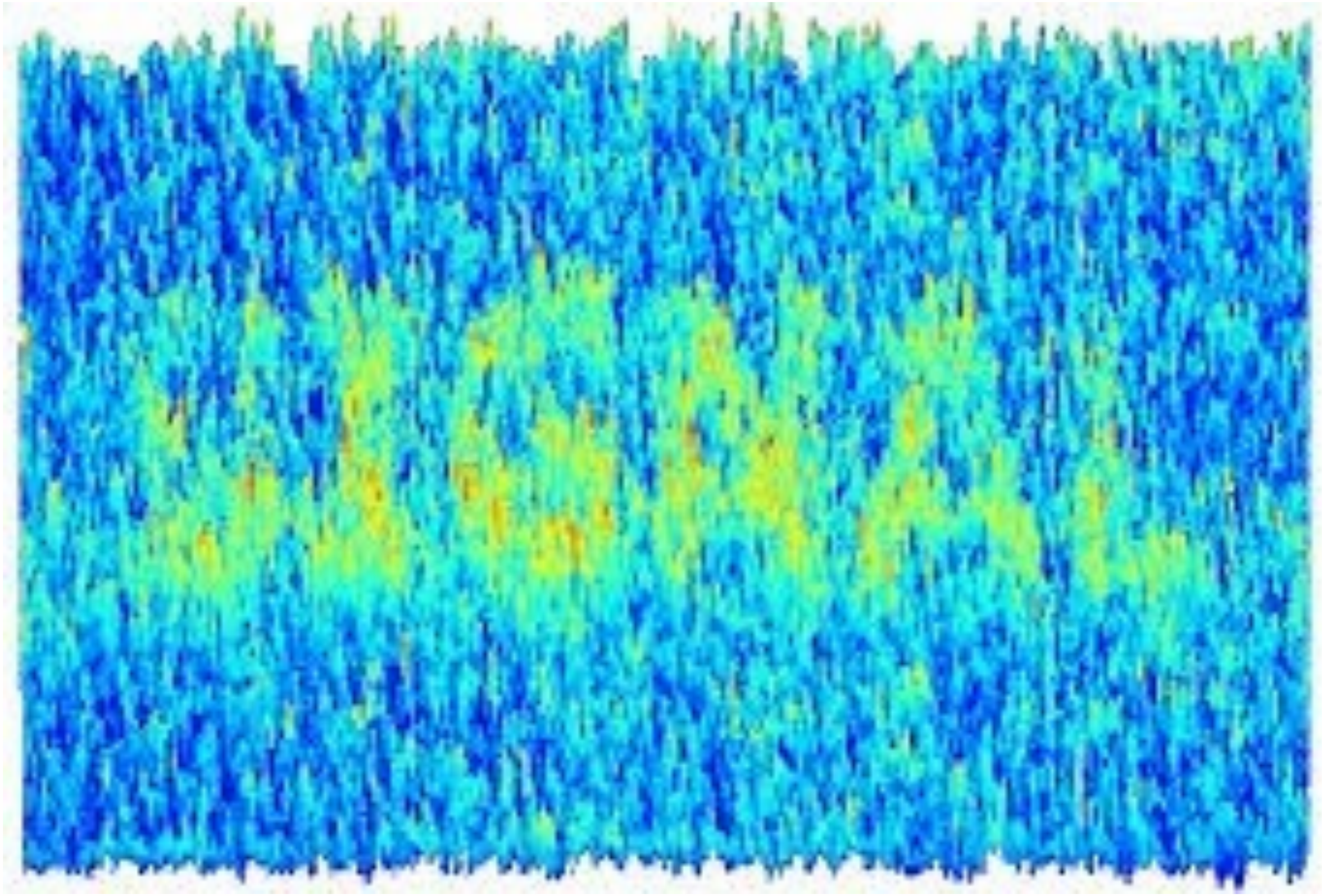


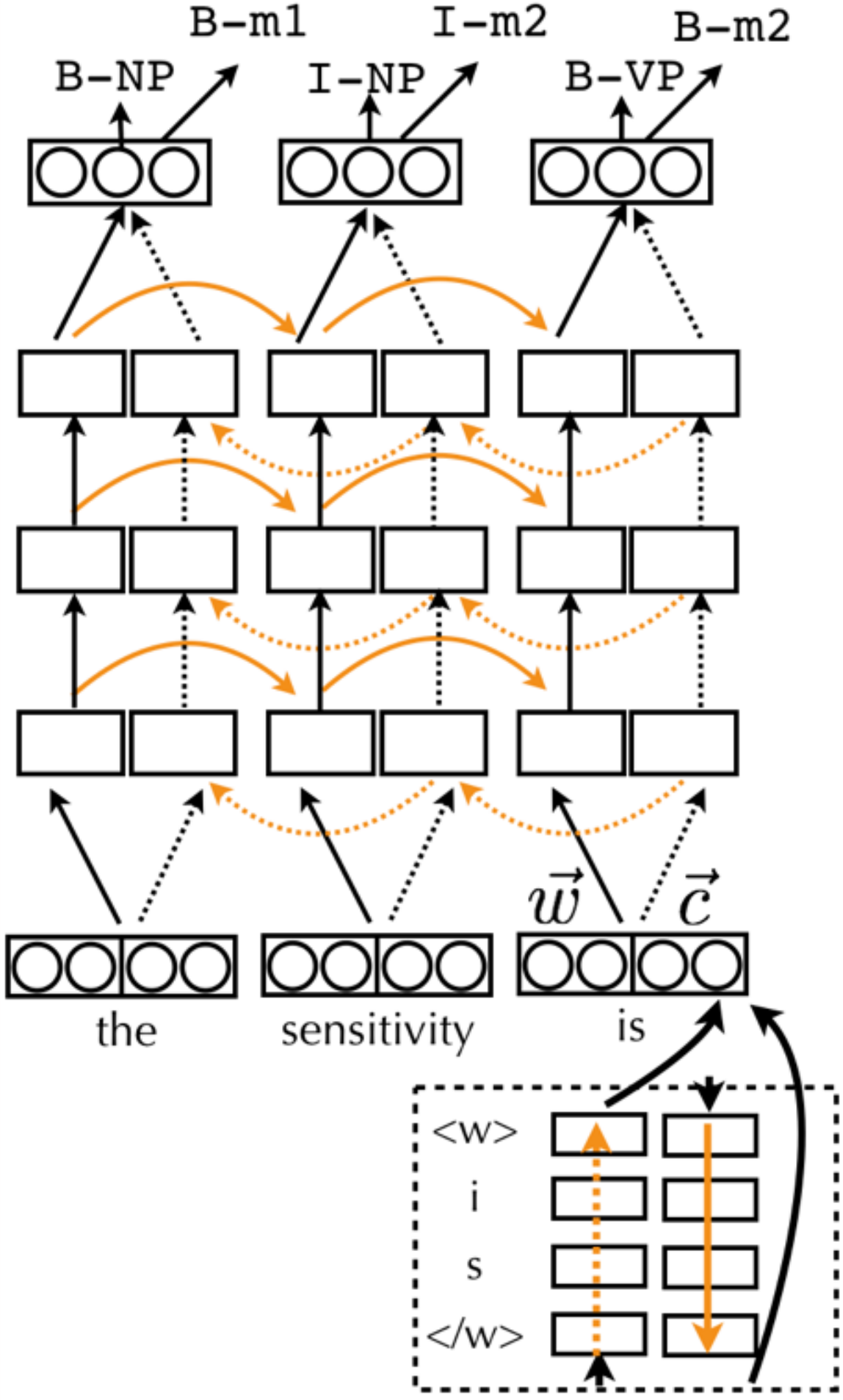
log(pause)



log(pause)


word length
pause length
0
2
4
6
8
10
12
14
16
18
2
4
6
8
10
12





bi-LSTM

with auxiliary loss

(Plank, Søgaard,Goldberg, 2016)

Tasks

Chunking:

B-NP	B-VP	B-NP	I-NP
We	love	deep	learning

CCG:

NP	(S[dcl]\NP)/NP	N/N	N
We	love	deep	learning

Results

	FOSTER.DEV	FOSTER.TEST	RITTER	CCG
Baseline	73.93	73.61	66.65	92.41
+PAUSE	**74.63**†	**74.32**†	**66.91**†	**92.62**†

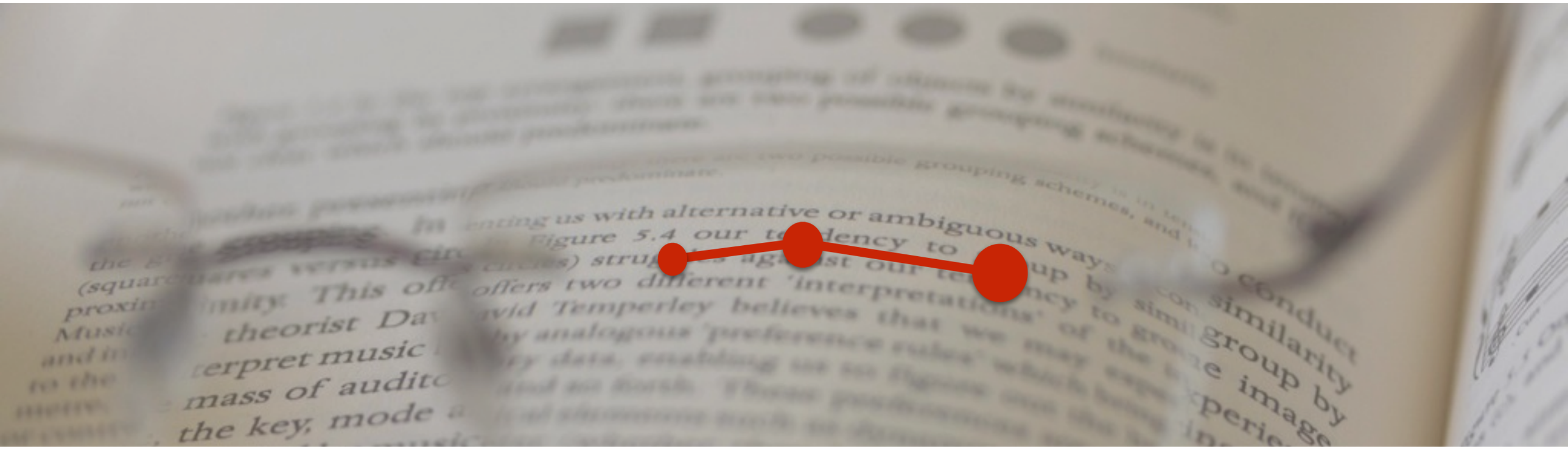
Results: Chunking

		FOSTER.DEV	FOSTER.TEST	RITTER
Baseline	NP	72.18	71.41	61.76
	VP	70.25	73.44	75.13
	PP	93.25	91.85	89.05
+PAUSE	NP	**73.99**	**72.77**	**62.60**
	VP	69.88	**74.93**	75.05
	PP	93.24	90.82	88.87

Table 6: Chunking results per label.

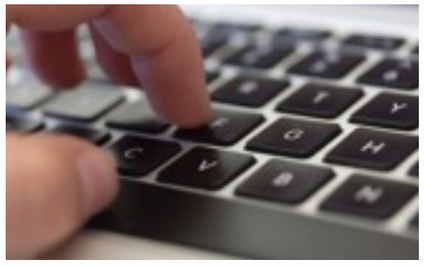
Token	Gold	Baseline	Model
Auburn	B-NP	I-NP	B-NP
party	I-NP	I-NP	I-NP
at	B-PP	B-PP	B-PP
man	B-NP	B-NP	B-NP
utd	I-NP	B-VP	I-NP
sounds	B-VP	B-VP	B-VP
bithcy	B-ADJP	B-VP	B-ADJP
)	O	I-NP	O

Related work: Eye tracking

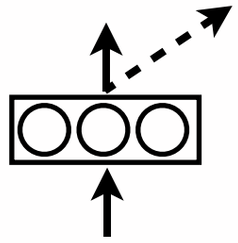


- In NLP very recently used to:
  - discriminate POS & syntactic relations (cf. Barrett et al, 2015; Barrett & Søgaard, 2015)
  - aid sentence compression (Klerke et al., 2016 NAACL);

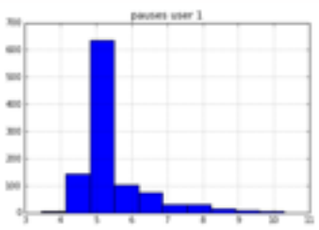
Take-home message



- Keystroke dynamics aid shallow parsing (promising initial results)



- Multi-task learning (promising; distinct sources)



- Lots of more to be done (representation, revisions, relation to speech..)

Thanks!

b.plank@rug.nl