

# On probability and uncertainty in unsupervised language learning

Anton Kolonin

[akolonin@aigents.com](mailto:akolonin@aigents.com)

Facebook: akolonin

Telegram: akolonin



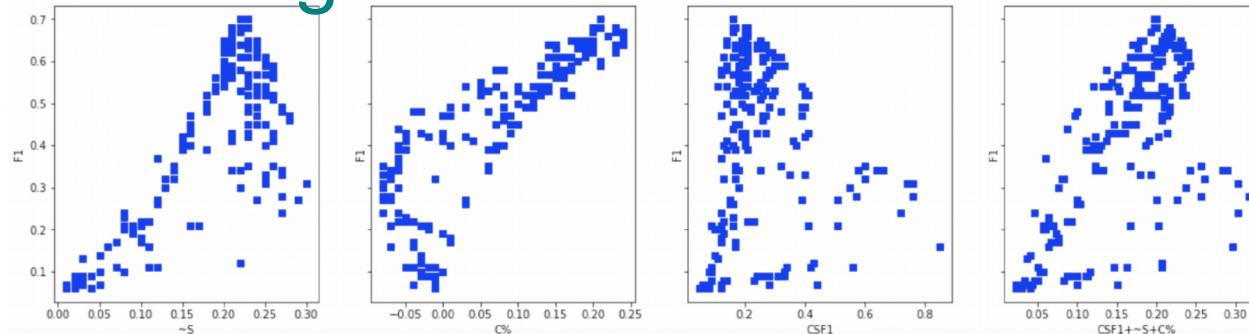
<https://agirussia.org>

N\* Novosibirsk  
State  
University  
\*THE REAL SCIENCE

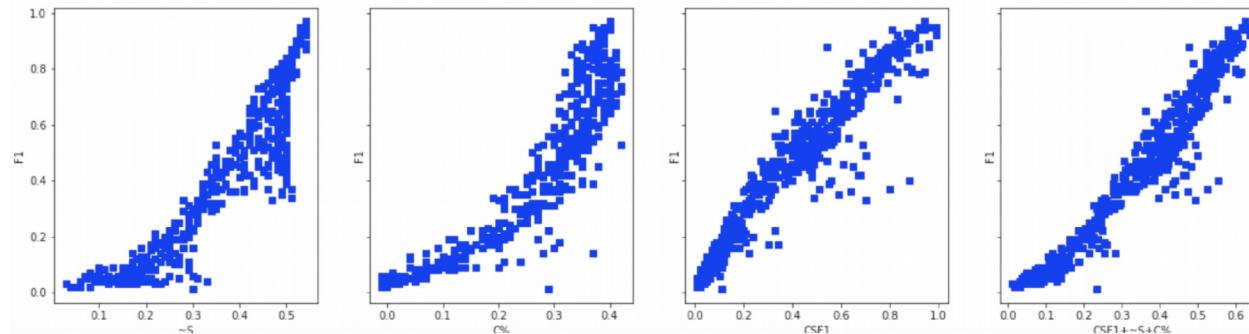


# Something about Human Intuition?

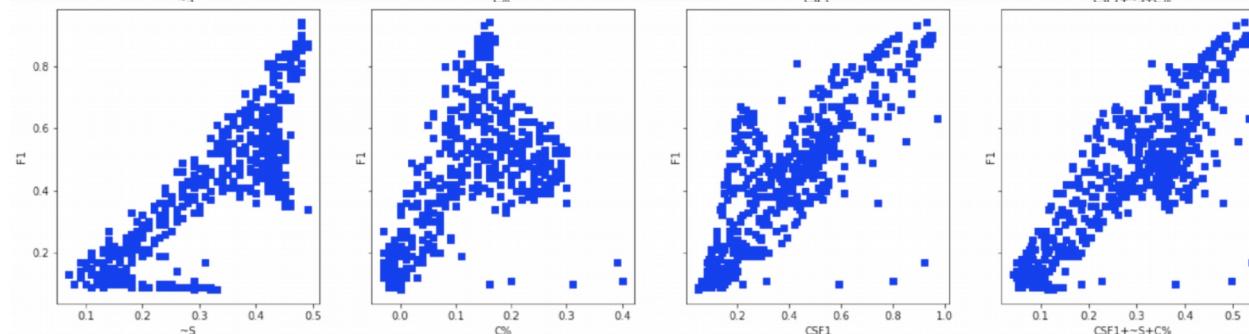
Language 1



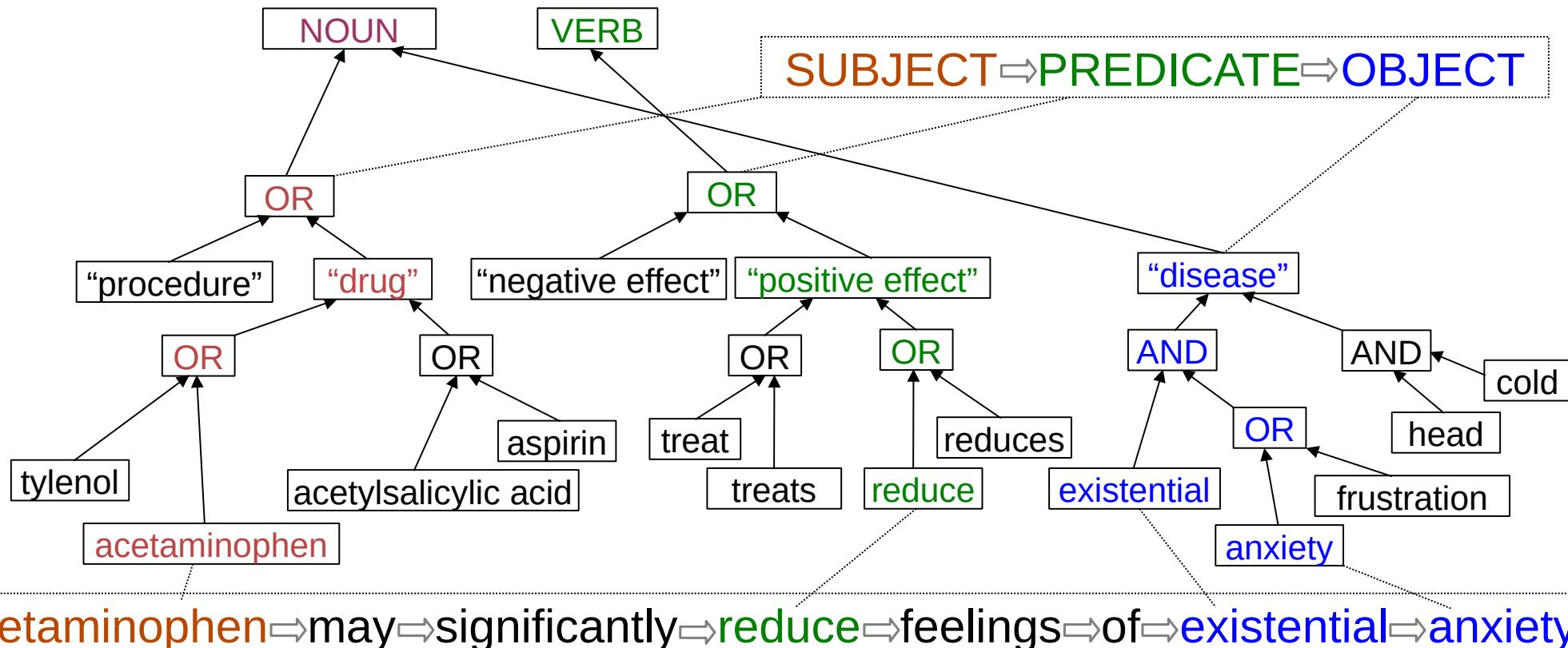
Language 2



Language 3



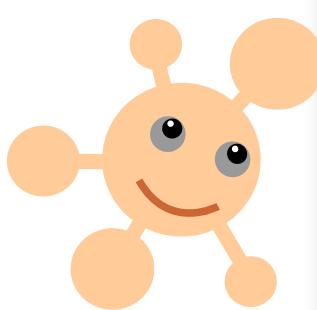
# Goal: Discovering NLP patterns (words, punctuation, phrases) for unsupervised language learning (Aigents® “Deep Patterns”)



<https://ieeexplore.ieee.org/document/7361868>  
<https://github.com/aigents/aigents-java>

<https://www.springerprofessional.de/unsupervised-language-learning-in-opencog/15995030>  
<https://www.springerprofessional.de/en/programmatic-link-grammar-induction-for-unsupervised-language-le/17020348>  
<https://github.com/singnet/language-learning/>

# Minimizing Uncertainty



New Tab    ×    +

how are you

- how are you - Google Search
- how are
- how are you doing
- how are you answers
- How Are You Feeling - Song by TAYLOR DEE
- How Are You Today? - Song by Maple Leaf Learning
- how are you doing answer
- how are you synonyms
- how are you in spanish
- how are things going

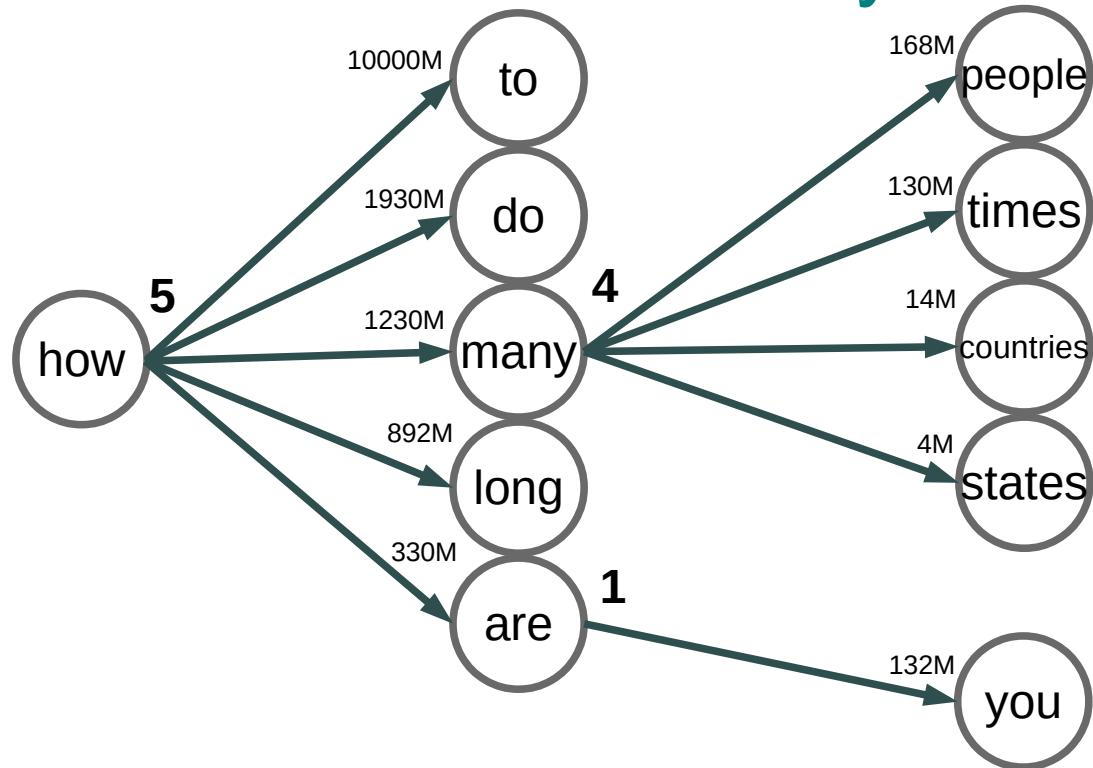


New Tab    ×    +

how many

- how many - Google Search
- how many countries in the world
- how many weeks in a year
- how many states in usa
- how many continents
- how many people in the world
- how many words
- how many continents are there
- how many bones in human body
- how many episodes in house of dragons

# Unsupervised Learning for Text Segmentation based on Probability and Uncertainty Measures



## Metrics/Indicators:

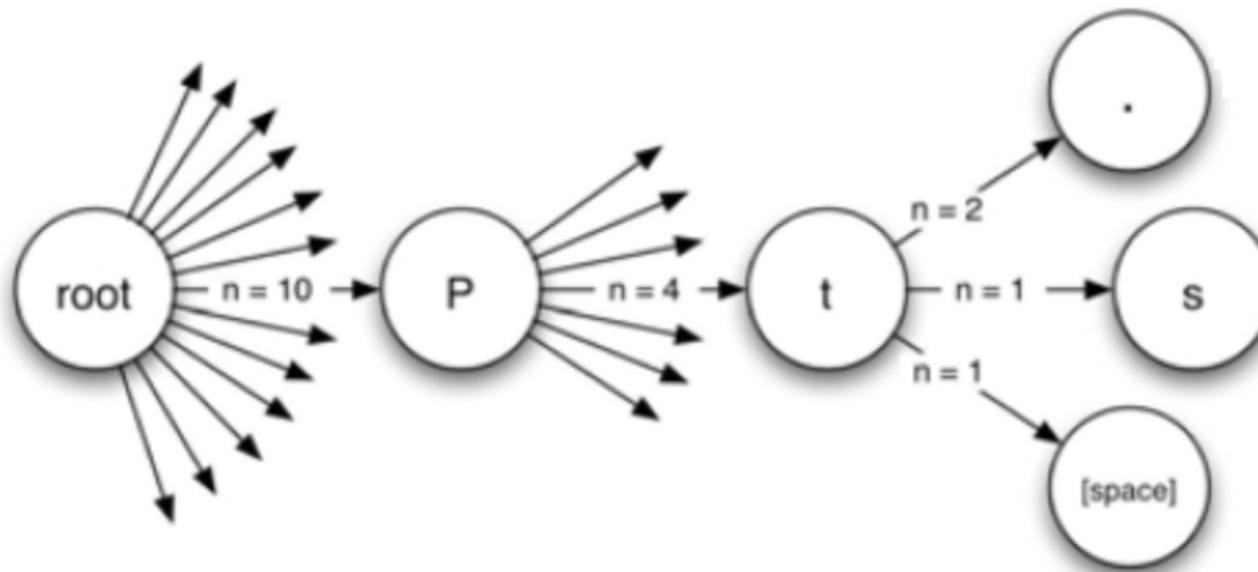
Mutual Information<sup>1</sup>  
Conditional Probability<sup>1,2</sup>  
Transition Freedom<sup>2,3</sup>

<sup>1</sup> <https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=6983&context=etd>

<sup>2</sup> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655800/>

<sup>3</sup> Karl Friston. The free-energy principle: a unified brain theory? <https://www.nature.com/articles/nrn2787>

# Background – Tokenization as Language Modeling



## Metrics/Indicators:

Conditional Probability  
“Transition Freedom”

Trie data structure. The probability of observing an ‘s’ given the preceding string “Pt” is  $\frac{1}{4}$ , or 25%. The freedom following “pt” is 3.

*Copyright ©2007 AMIA - All rights reserved. Jesse O. Wrenn, Peter D. Stetson, and Stephen B. Johnson. 2007. An unsupervised machine learning approach to segmentation of clinician-entered free text. AMIA Annu Symp Proc. 2007; 2007: 811–815.*

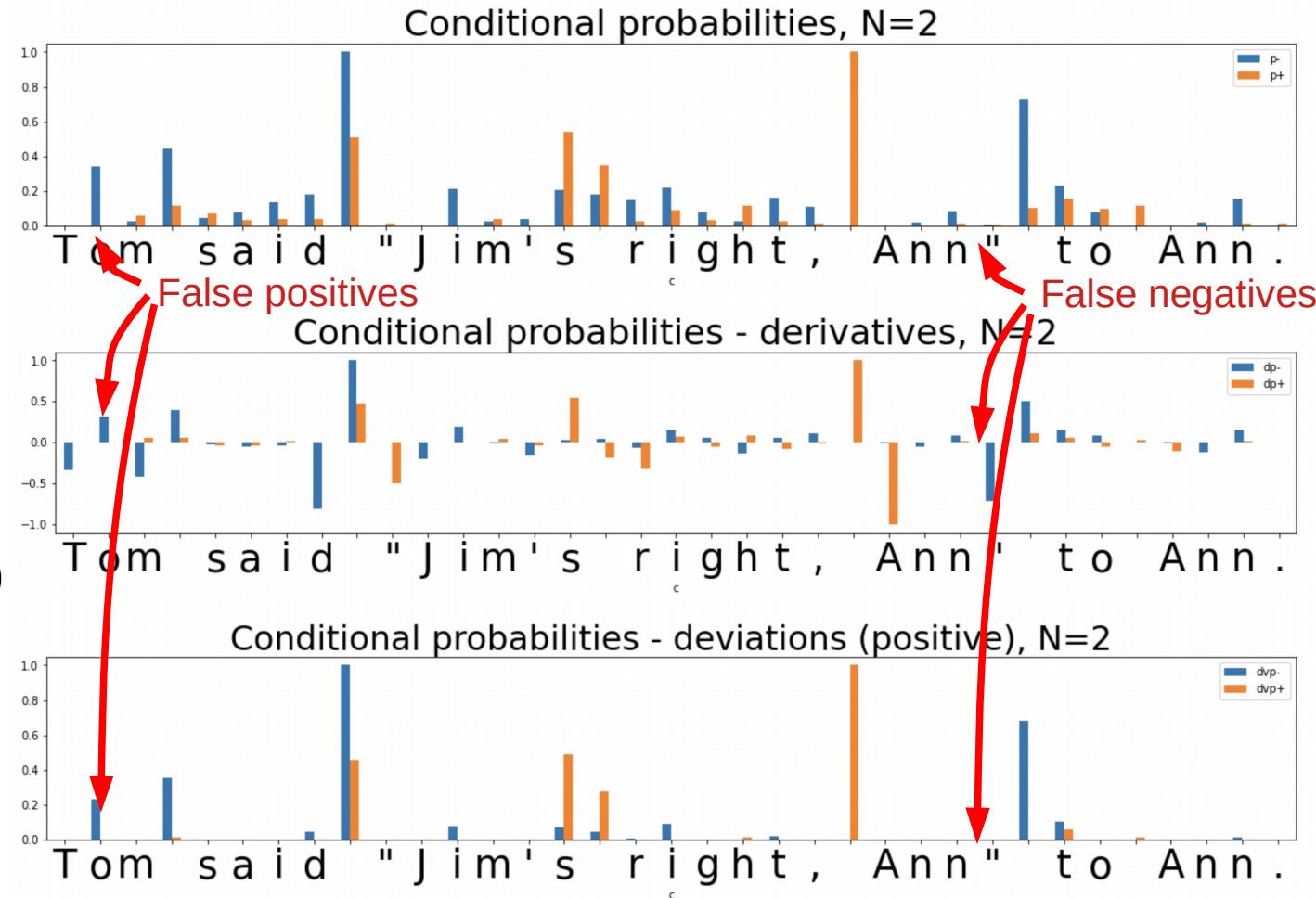
# Unsupervised Text Segmentation (Tokenization)

## Metrics/Indicators:

Ngram (Character)  
Conditional  
Probability  
(of Transition)

$P(\text{Ngram}_{n+1})/P(\text{Ngram}_n)$

$P("m")/P(m")$



# Unsupervised Text Segmentation (Tokenization)

Metrics/  
Indicators:

Transition  
Freedom  
Deviation

(varying “N”  
of N-gram)



# Unsupervised Text Segmentation (Tokenization)

**English**

F1 - Brown ddf- & ddf+ filter=0 parameters=10967135

**Hyper-  
Parameters:**

Metric:  
Transition  
Freedom

[1]	0.5	0.75	0.82	0.79	0.79	0.81	0.89	0.89	0.89
[2]	0.46	0.54	0.62	0.67	0.85	0.92	0.81	0.71	0.37
[3]	0.56	0.67	0.72	0.73	0.69	0.61	0.46	0.36	0.19
[4]	0.54	0.68	0.7	0.6	0.43	0.3	0.19	0.15	0.1
[5]	0.51	0.55	0.52	0.38	0.25	0.16	0.11	0.1	0.08
[6]	0.48	0.46	0.38	0.25	0.17	0.12	0.1	0.08	0.07
[7]	0.42	0.34	0.24	0.15	0.11	0.1	0.08	0.08	0.07
[1, 2]	0.47	0.58	0.82	0.94	0.94	0.91	0.89	0.79	0.56
[2, 3]	0.51	0.62	0.74	0.79	0.83	0.81	0.66	0.46	0.24
[1, 2, 3]	0.5	0.69	0.79	0.87	0.91	0.89	0.78	0.58	0.25
[1, 2, 3, 4]	0.55	0.75	0.84	0.86	0.84	0.75	0.52	0.31	0.15
[4, 5, 6, 7]	0.56	0.6	0.51	0.33	0.2	0.14	0.1	0.08	0.07
[1, 2, 3, 4, 5]	0.56	0.78	0.86	0.84	0.74	0.53	0.31	0.17	0.1
[1, 2, 3, 4, 5, 6, 7]	0.59	0.78	0.82	0.69	0.49	0.26	0.15	0.09	0.07
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

Threshold  
for model  
compression

F1 - Brown ddf- & ddf+ filter=0.0001 parameters=8643703

Combination  
of Ngram N-s

[1]	0.73	0.96	0.98	0.99	0.96	0.94	0.95	0.95	0.89
[2]	0.46	0.54	0.64		0.91	0.94	0.89	0.7	0.44
[3]	0.55	0.66	0.74	0.78	0.72	0.65	0.49	0.37	0.19
[4]	0.54	0.67	0.7	0.61	0.45	0.32	0.21	0.16	0.1
[5]	0.51	0.55	0.52	0.38	0.26	0.17	0.12	0.1	0.08
[6]	0.48	0.46	0.38	0.26	0.18	0.13	0.1	0.09	0.07
[7]	0.42	0.35	0.25	0.16	0.12	0.1	0.09	0.08	0.08
[1, 2]	0.51	0.64	0.82	0.96	0.96	0.96	0.9	0.88	0.68
[2, 3]	0.5	0.62	0.74	0.85	0.89	0.86	0.71	0.51	0.27
[1, 2, 3]	0.53	0.69	0.81	0.91	0.93	0.92	0.82	0.6	0.36
[1, 2, 3, 4]	0.55	0.75	0.86	0.88	0.88	0.81	0.57	0.33	0.17
[4, 5, 6, 7]	0.56	0.6	0.52	0.35	0.22	0.15	0.1	0.09	0.07
[1, 2, 3, 4, 5]	0.57	0.79	0.88	0.86	0.78	0.59	0.33	0.18	0.1
[1, 2, 3, 4, 5, 6, 7]	0.59	0.79	0.83	0.71	0.5	0.28	0.16	0.09	0.08
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

Threshold for  
segmentation

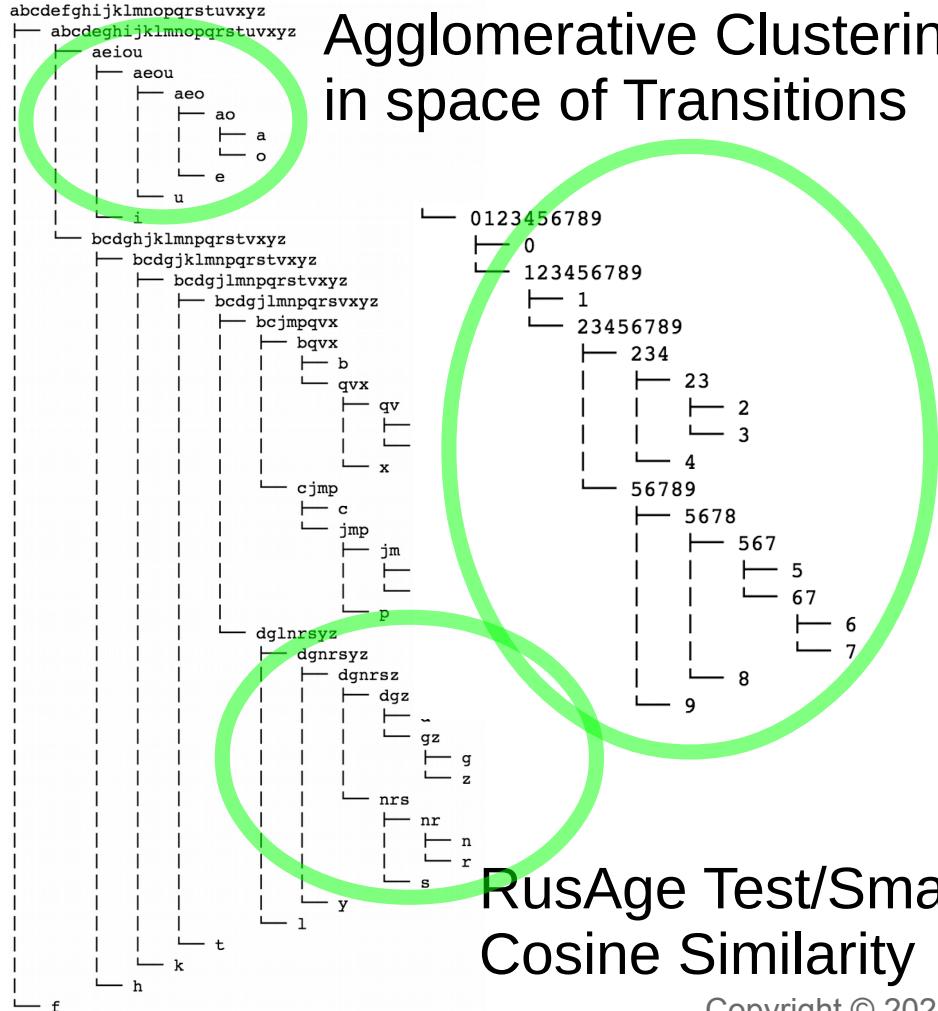
# Results – Freedom-based Tokenization against Lexicon-based one (referring to Rule-based)

Language	Tokenizer	Tokenization $F_1$	Lexicon Discovery Precision
English	Freedom-based	<b>0.99</b>	<b>0.99</b> (vs. 1.0)
English	Lexicon-based*	0.99	-
Russian	Freedom-based	<b>1.0</b>	<b>1.0</b> (vs. 1.0)
Russian	Lexicon-based*	0.94	-
Chinese	Freedom-based	<b>0.71</b>	<b>0.92</b> (vs. 0.94)
Chinese	Lexicon-based*	0.83	-

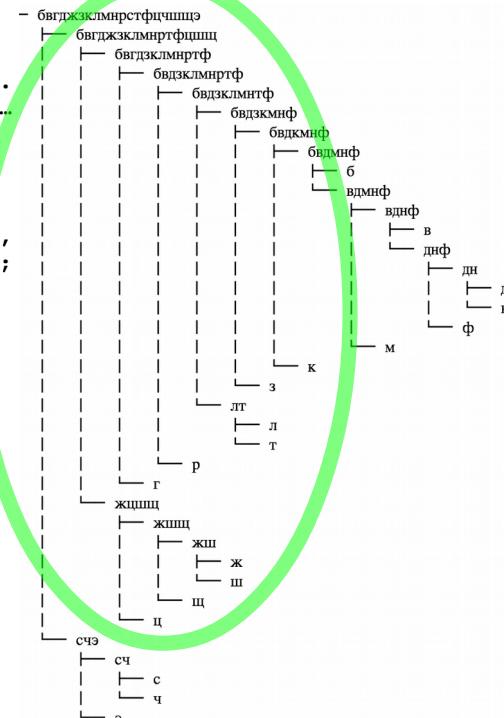
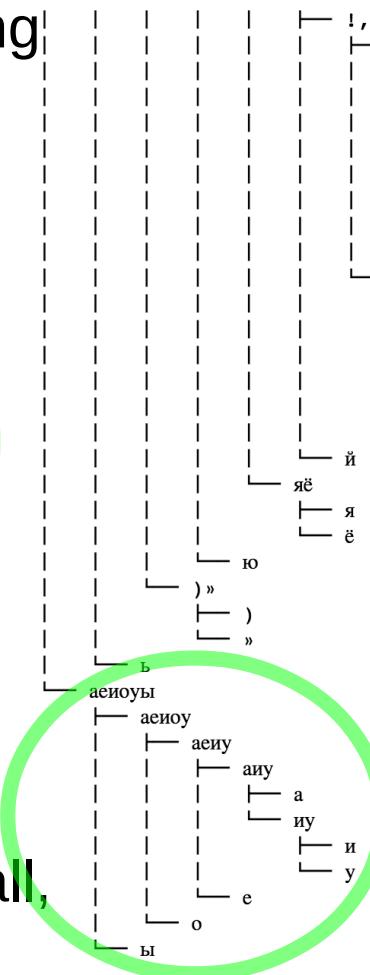
\*Lexicon-based Tokenization - greedy/beam search on word length (optimal) or frequency

# Unsupervised Character Category Learning

# Agglomerative Clustering in space of Transitions

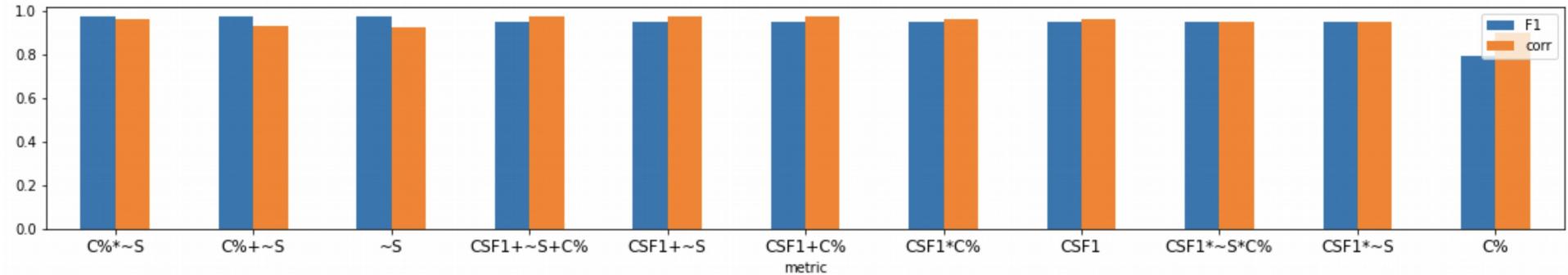


# RusAge Test/Small Cosine Similarity

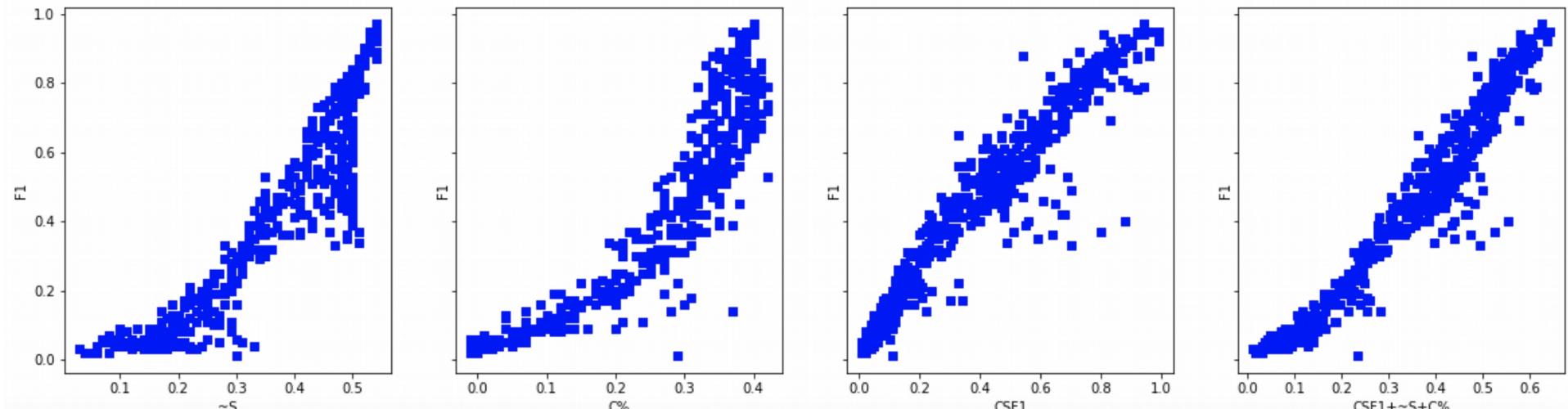


# Self-tuning Hyperparameters – English (TF variance)

Test 1000

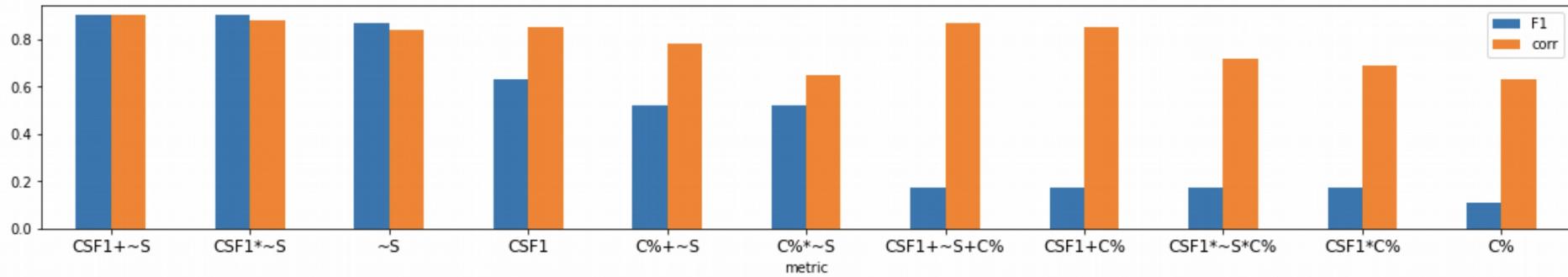


F1 as function of  $\sim S$ , C% and CSF1 used for hyper-parameter selection

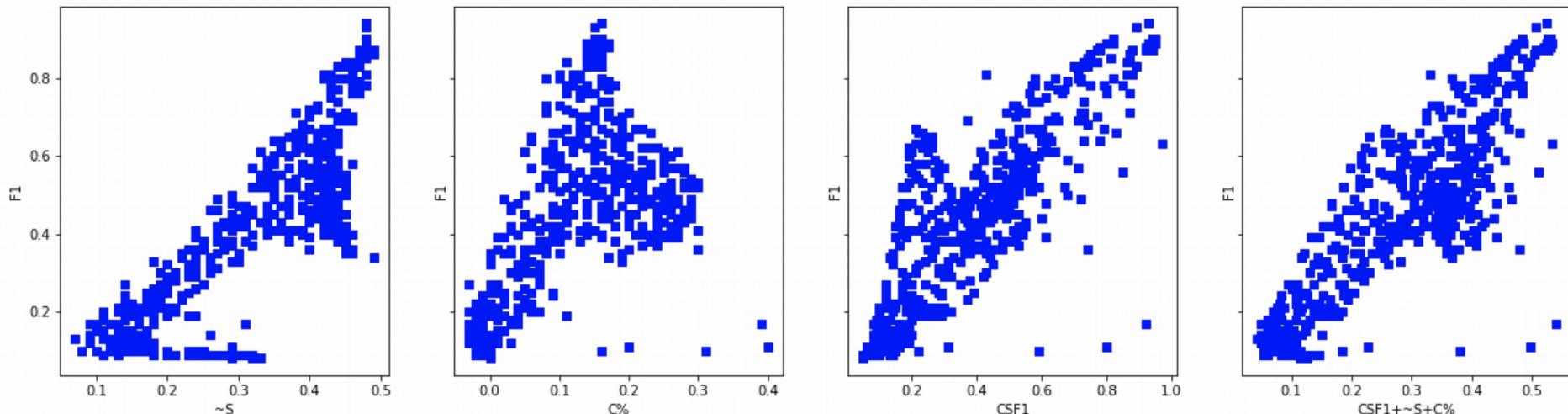


# Self-tuning Hyperparameters – Russian (TF variance)

Test 1000

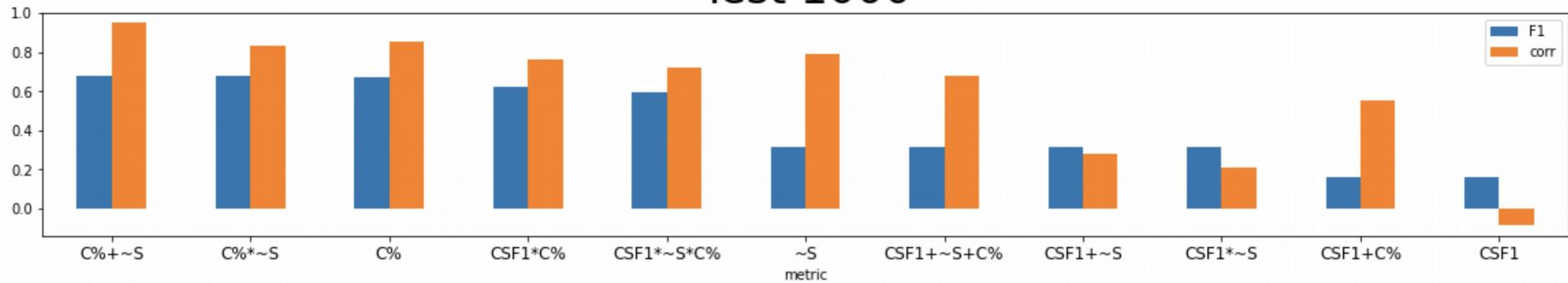


F1 as function of ~S, C% and CSF1 used for hyper-parameter selection

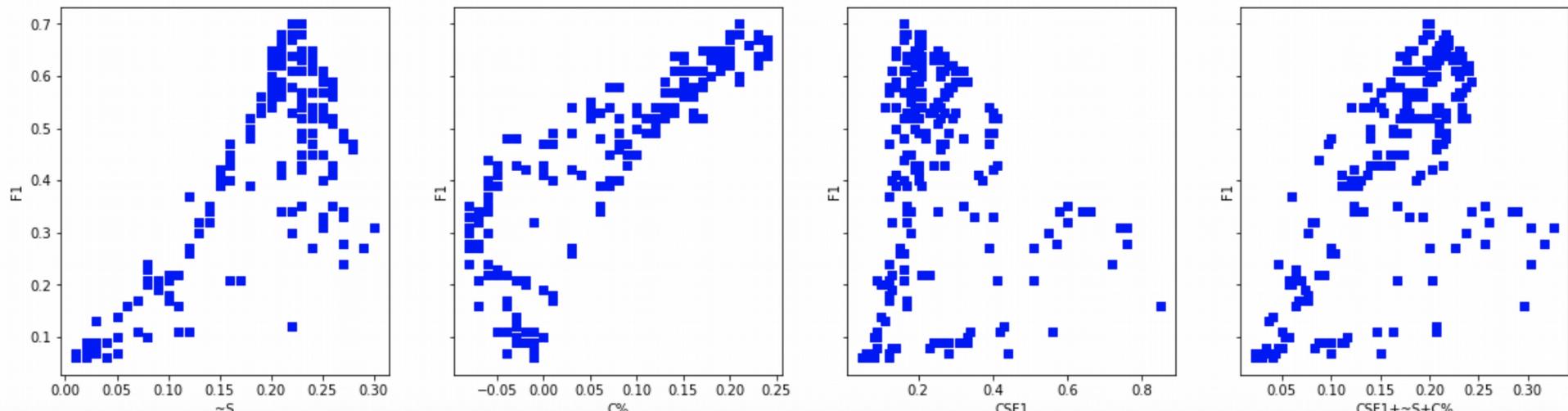


# Self-tuning Hyperparameters – Chinese (TF “peak”)

Test 1000



F1 as function of  $\sim S$ , C% and CSF1 used for hyper-parameter selection



# Something about Human Intuition!

Screen Shot 2022-06-16 at 11.08.54.png  
247.8 KB

OPEN WITH

Language 1 11:22 ✓

Screen Shot 2022-06-16 at 11.09.45.png  
256.8 KB

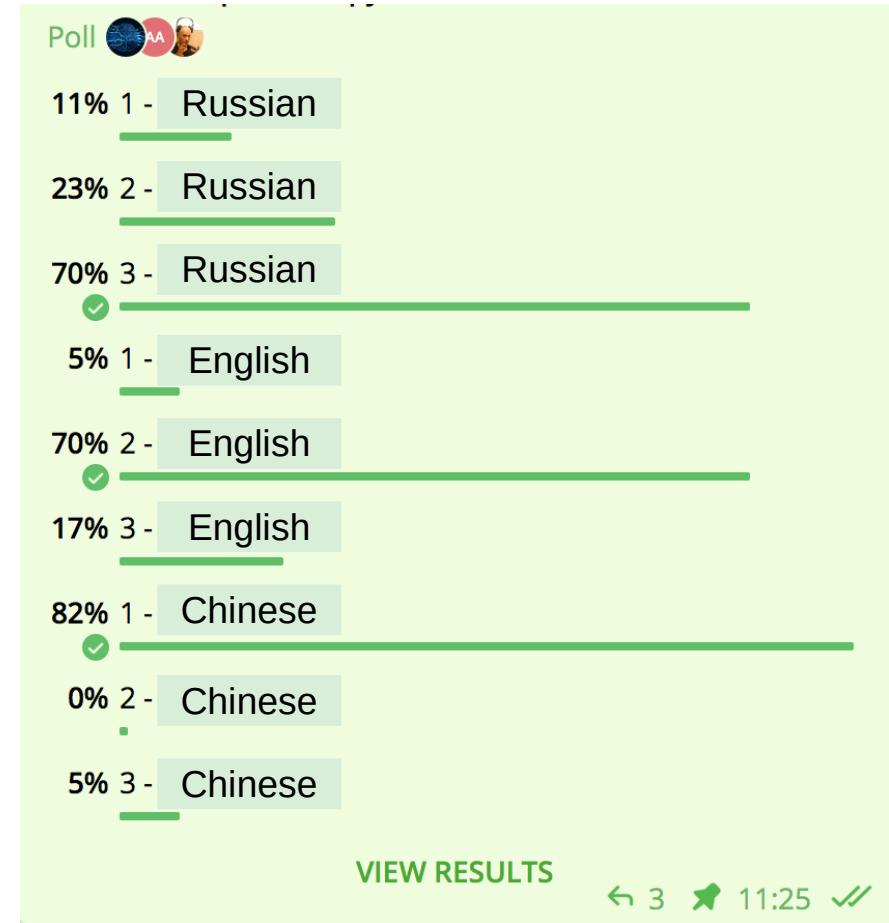
OPEN WITH

Language 2 11:23 ✓

Screen Shot 2022-06-16 at 11.09.59.png  
276.4 KB

OPEN WITH

Language 3 11:23 ✓



# Conclusion and Further Work

Unsupervised Tokenization based on Transition Freedom (TF) recall and precision appears good enough as initial approximation for further applications of self-reinforcement learning as part of interpretable unsupervised learning of natural language.

Optimal thresholds and specific TF-based metrics are specific to language. The process and policy of their discovery and adjustment should be further explored.

Clustering or parts of speech on space of transition graphs may provide some insights on morphology and punctuation structure of low-resource and domain-specific languages.

Hybridization of TF-based tokenization approach with lexicon-based one might be efficient for low-resource and domain-specific languages.

Further unsupervised grammar learning experiments can be run on the basis of suggested unsupervised tokenization approach.

Applications for other Experiential Learning environments, including the ones with delayed/sparse feedback.

Using Reinforcement Learning techniques with self-reinforcement on historical data under Unsupervised Learning setup.

<https://arxiv.org/abs/2205.11443>  
<https://github.com/aigents/pygents>

# Thank You and Welcome!



<https://agirussia.org>

Anton Kolonin  
[akolonin@aigents.com](mailto:akolonin@aigents.com)

Facebook: akolonin  
Telegram: akolonin

N\* Novosibirsk  
State  
University  
\*THE REAL SCIENCE

