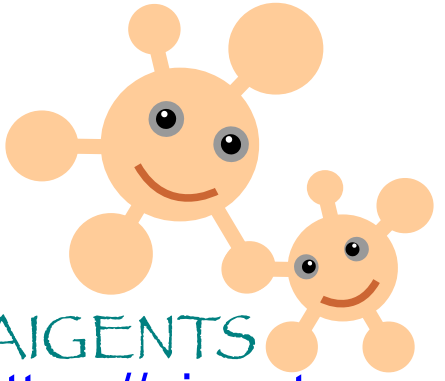
On probability and uncertainty in unsupervised language learning

Anton Kolonin
akolonin@aigents.com
Facebook: akolonin
Telegram: akolonin

https://agirussia.org

Novosibirsk State University
*THE REAL SCIENCE



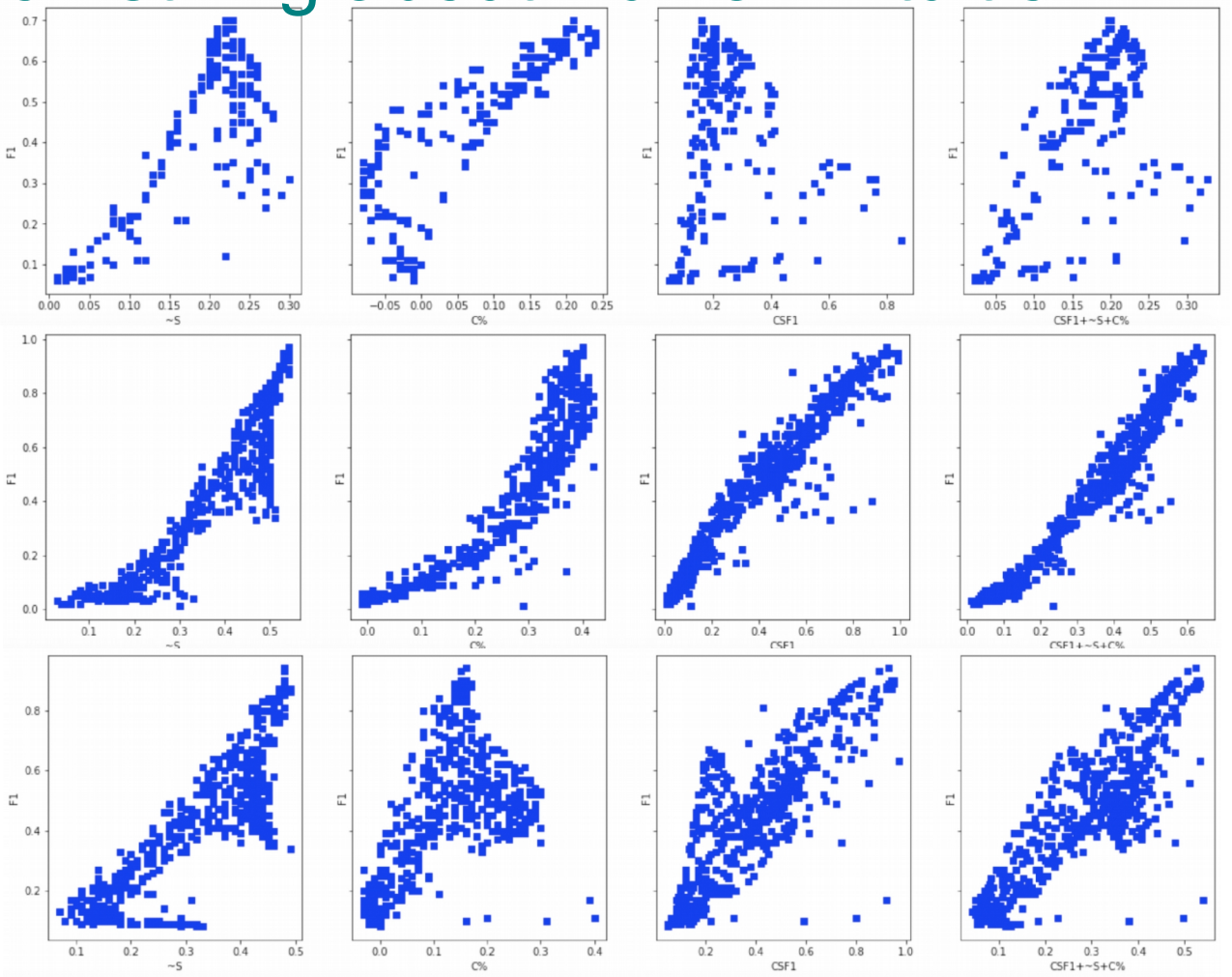
https://aigents.com

Something about Human Intuition?

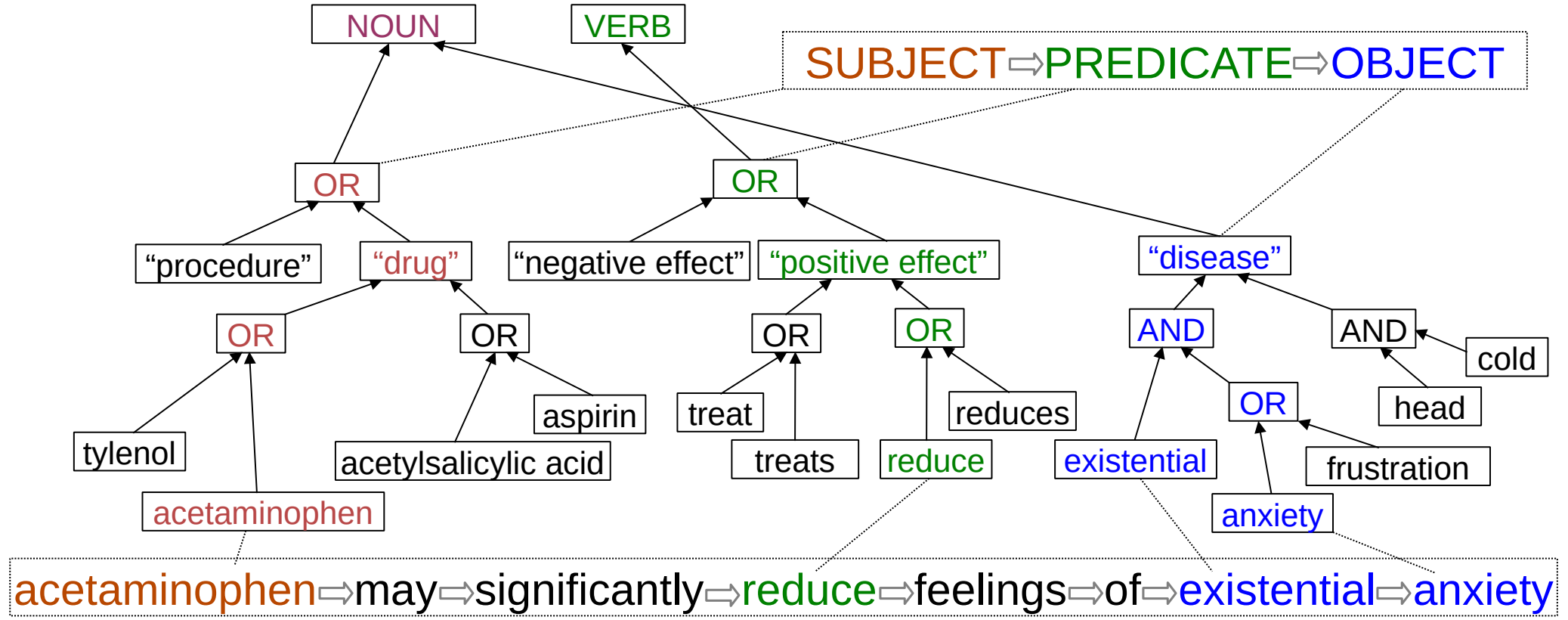
Language 1

Language 2

Language 3



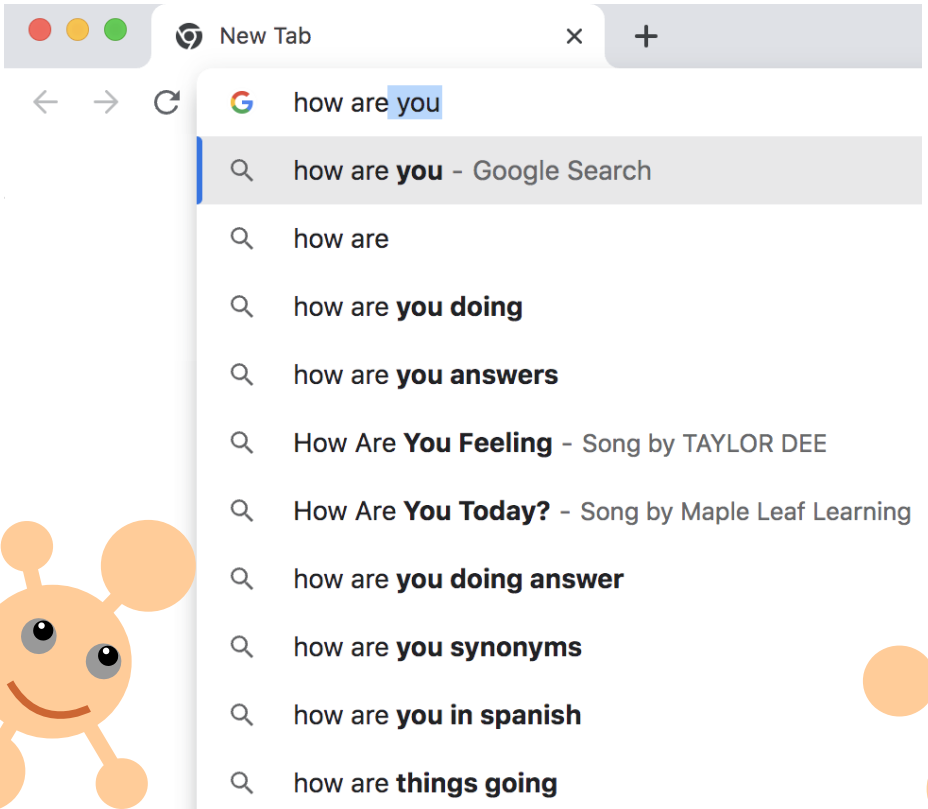
Goal: Discovering NLP patterns (words, punctuation, phrases) for unsupervised language learning (Aigents® “Deep Patterns”)

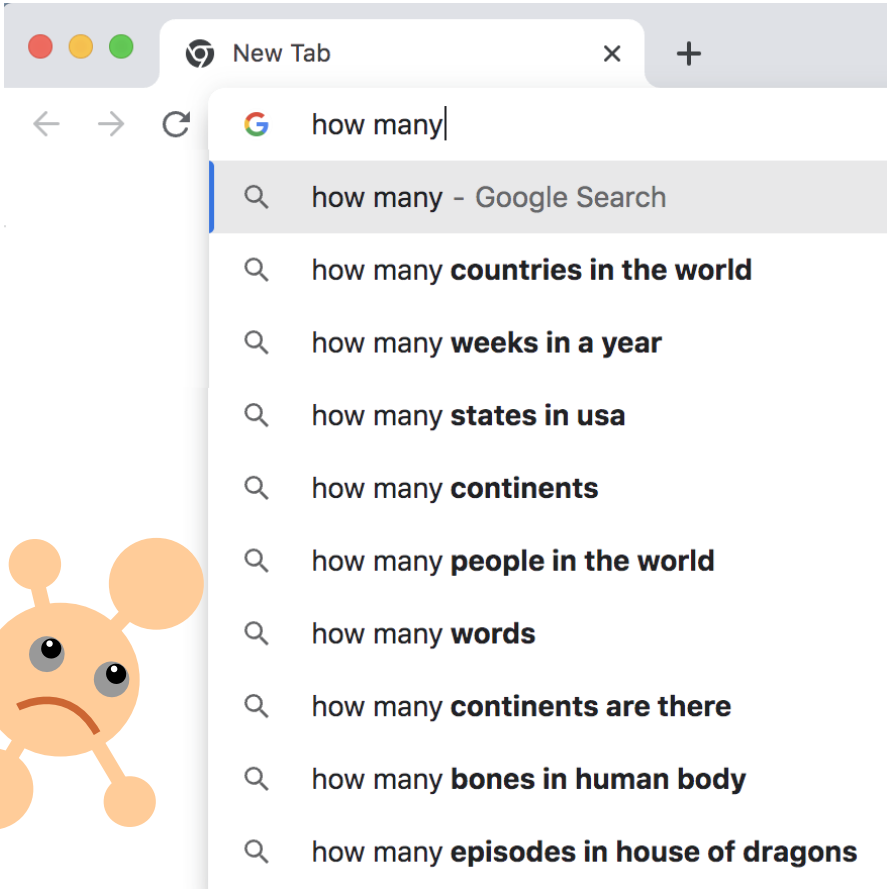


https://ieeexplore.ieee.org/document/7361868
https://github.com/aigents/aigents-java

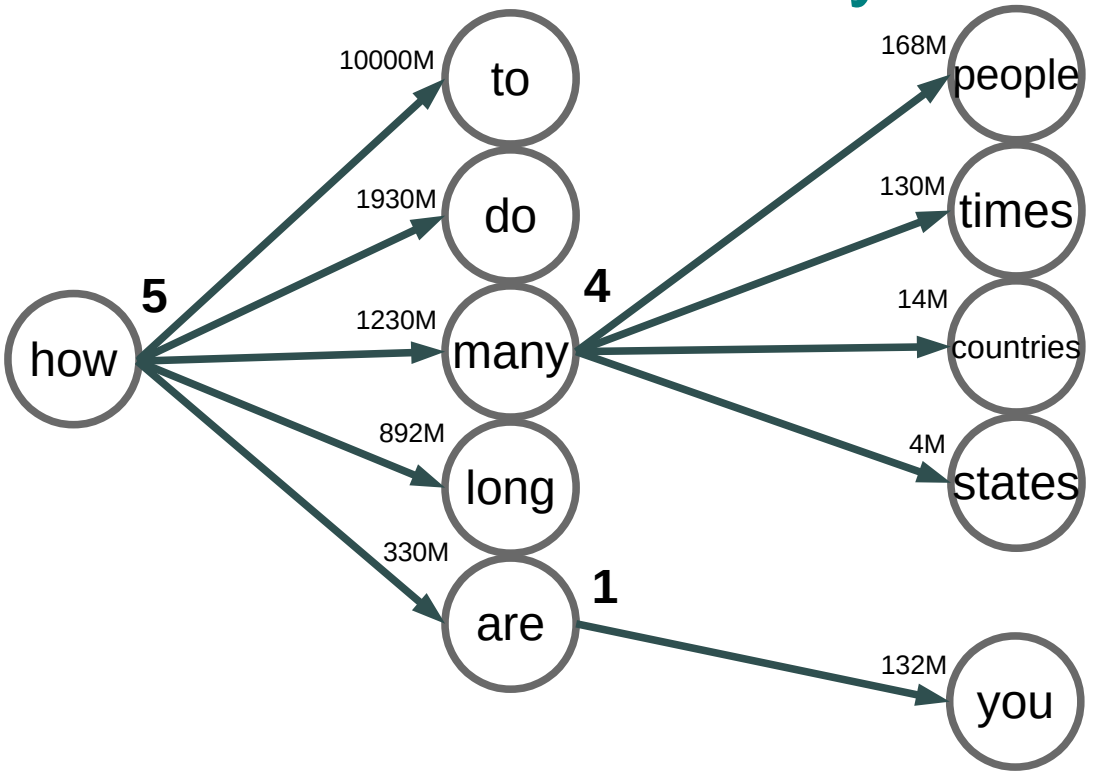
https://www.springerprofessional.de/unsupervised-language-learning-in-opencog/15995030
https://www.springerprofessional.de/en/programmatic-link-grammar-induction-for-unsupervised-language-le/17020348
https://github.com/singnet/language-learning/

Minimizing Uncertainty





Unsupervised Learning for Text Segmentation based on Probability and Uncertainty Measures



Metrics/Indicators:

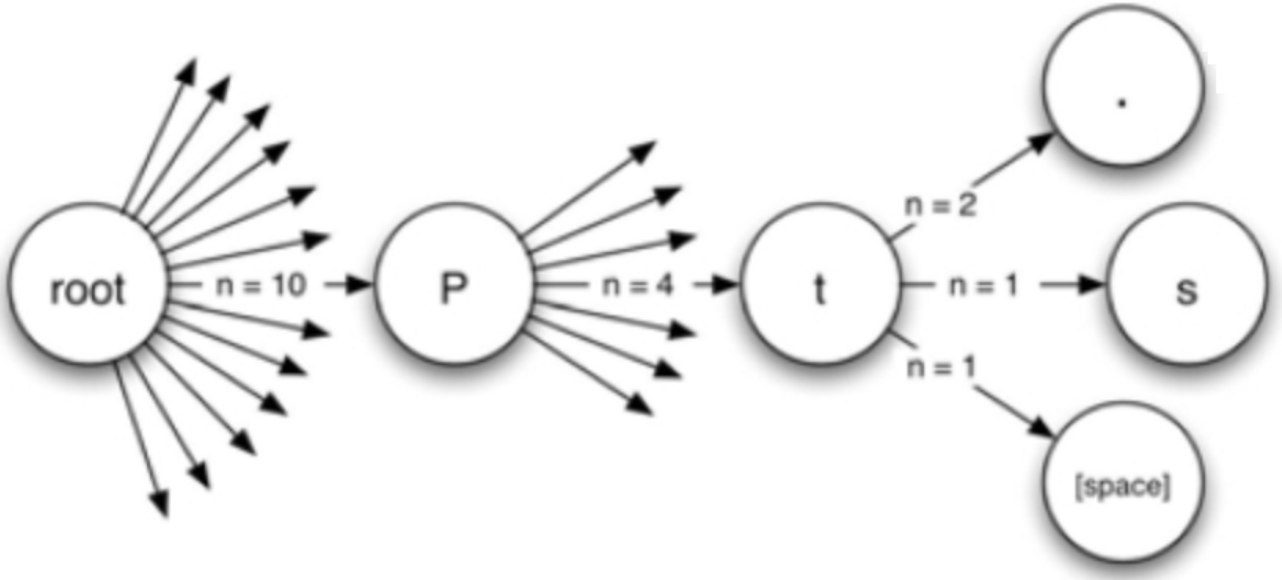
Mutual Information[1]
Conditional Probability[1,2]
Transition Freedom[2,3]

[1] https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=6983&context=etd

[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655800/

[3] **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 ¼, or 25%. The freedom following “pt” is 3.

Unsupervised Text Segmentation (Tokenization)

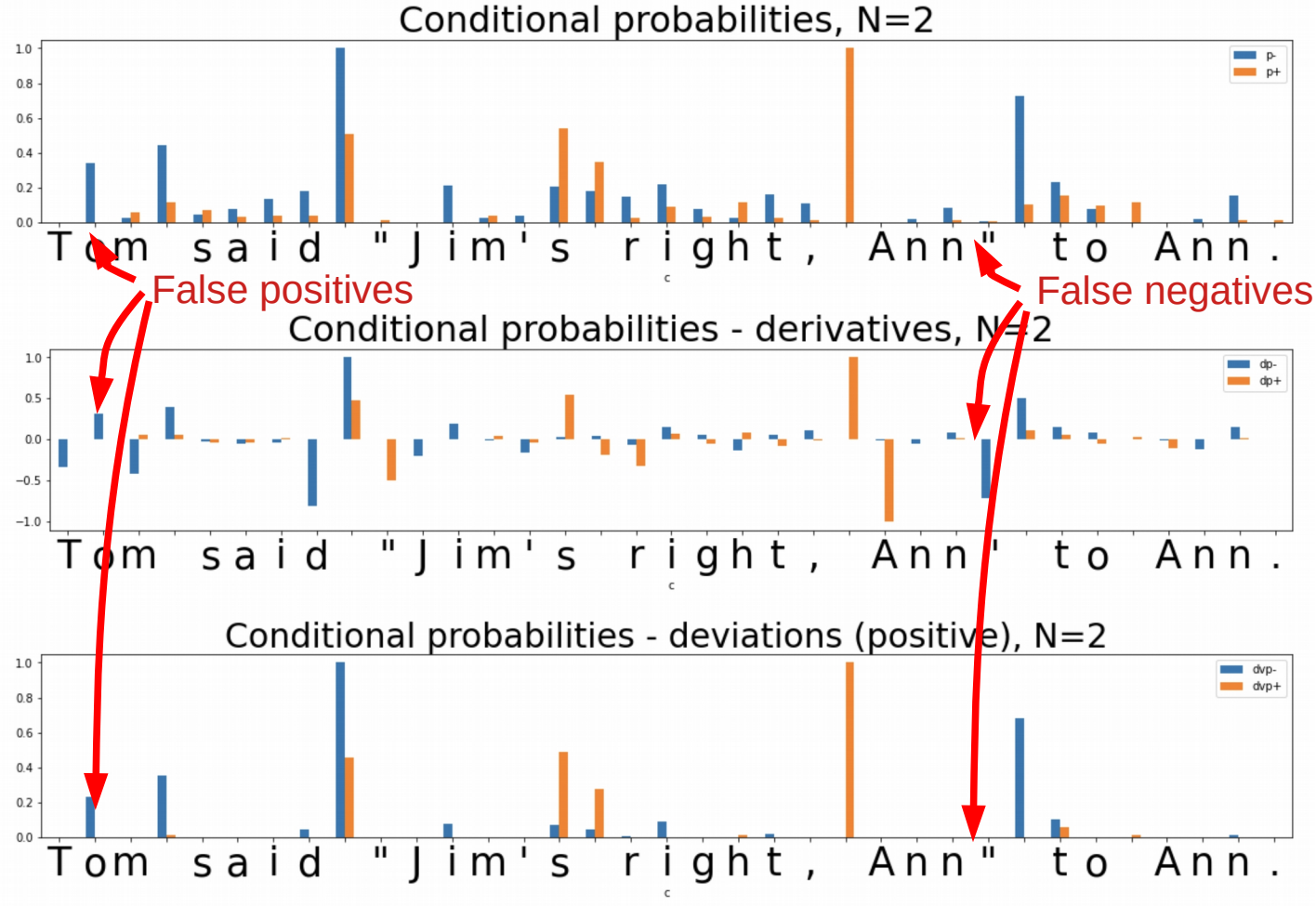
Metrics/Indicators:

Ngram (Character) Conditional Probability

(of Transition)

$P(Ngram_{n+1})/P(Ngram_n)$

P(“m ”)/P(“m”)

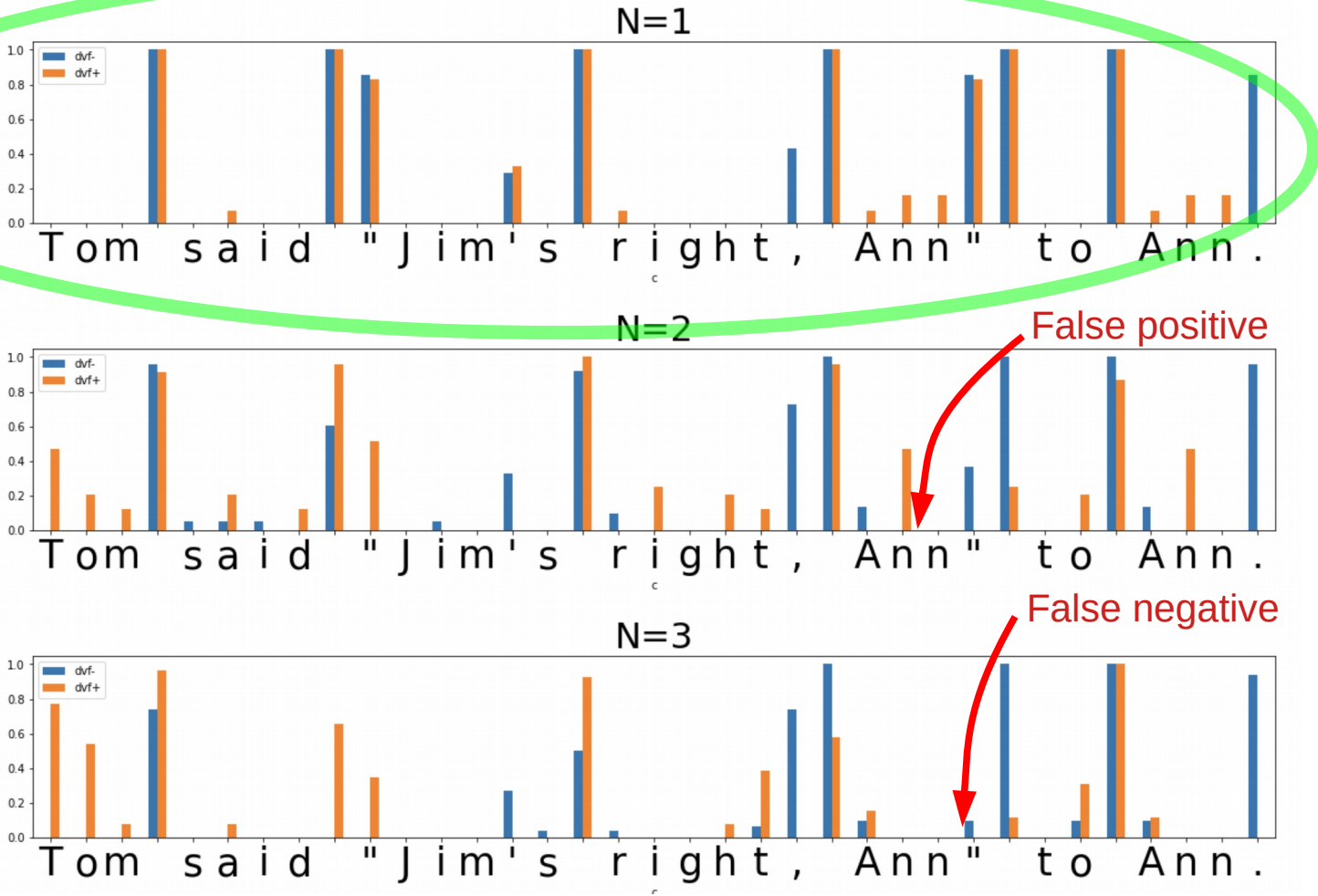


Unsupervised Text Segmentation (Tokenization)

Metrics/ Indicators:

Transition
Freedom
Deviation

(varying “N” of N-gram)



Unsupervised Text Segmentation (Tokenization)

English

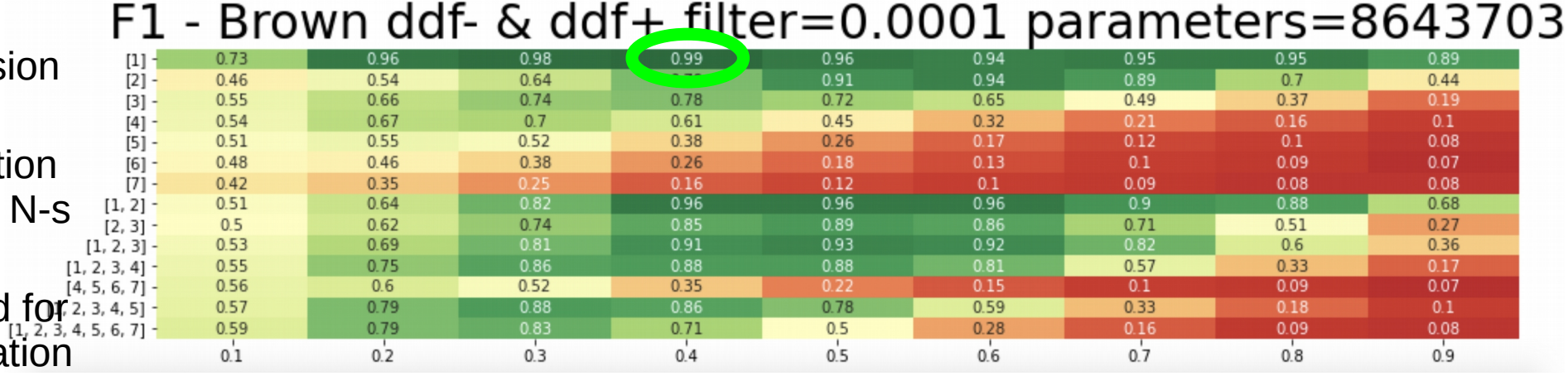
Hyper-Parameters:

Metric:
Transition Freedom

Threshold for model compression

Combination of Ngram N-s

Threshold for segmentation



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

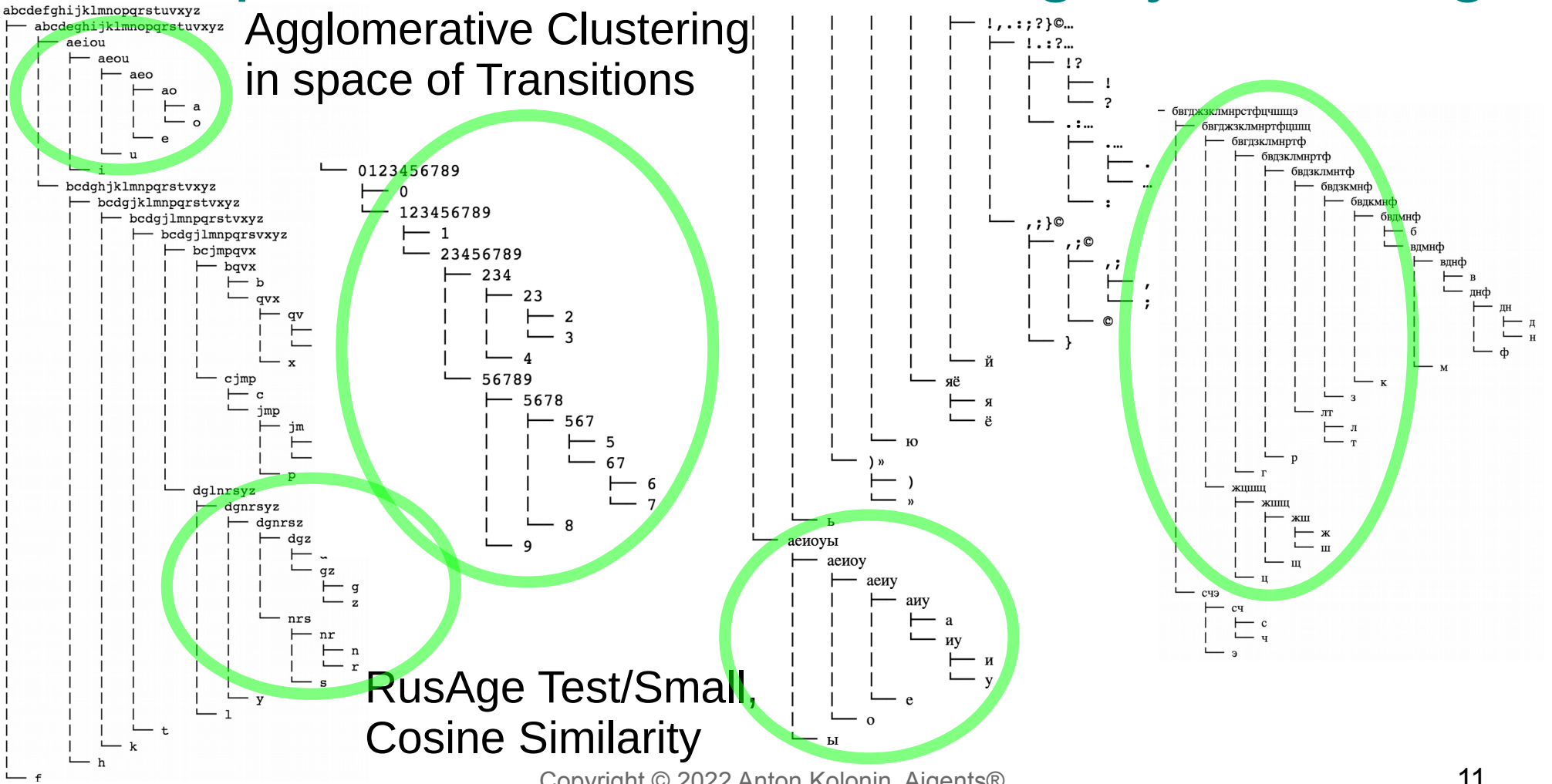
Language	Tokenizer	Tokenization F_1	Lexicon Discovery Precision
English	Freedom-based	**0.99**	**0.99** (vs. 1.0)
English	Lexicon-based*	0.99	-
Russian	Freedom-based	**1.0**	**1.0** (vs. 1.0)
Russian	Lexicon-based*	0.94	-
Chinese	Freedom-based	**0.71**	**0.92** (vs. 0.94)
Chinese	Lexicon-based*	0.83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

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

Preprint: https://arxiv.org/abs/2205.11443 **EMNLP-2022**

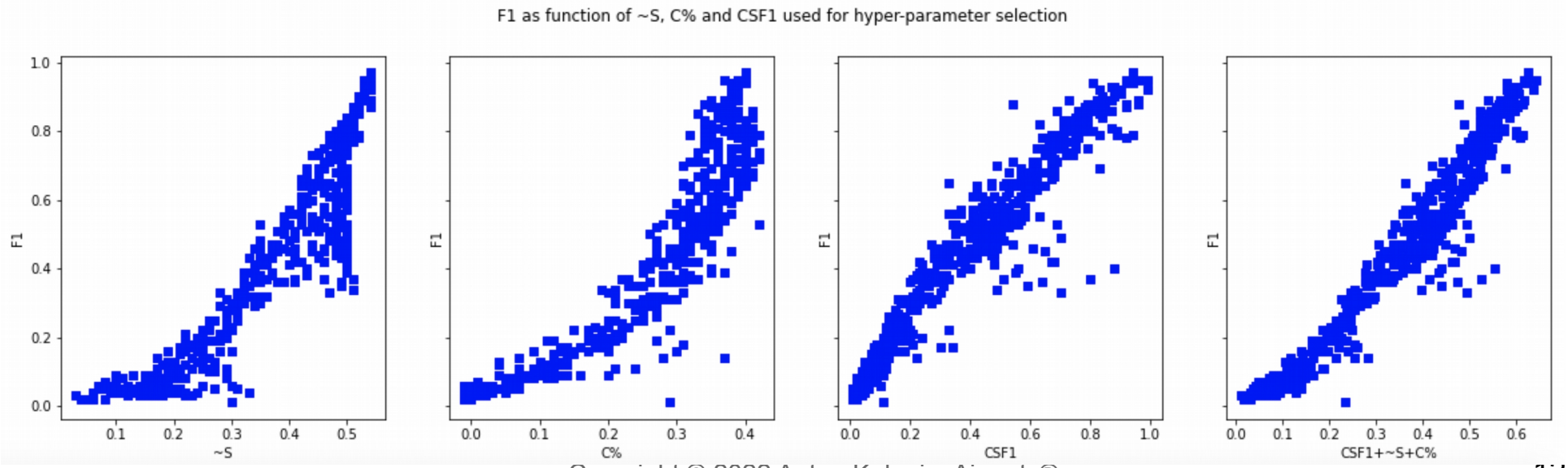
Unsupervised Character Category Learning

Agglomerative Clustering
in space of Transitions

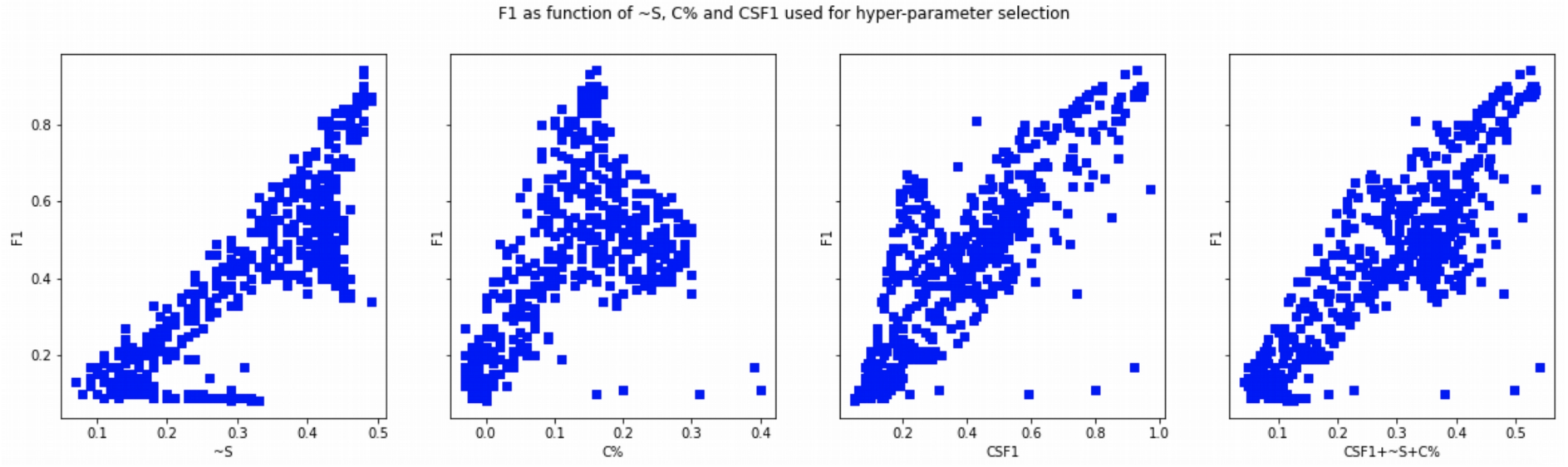


RusAge Test/Small,
Cosine Similarity

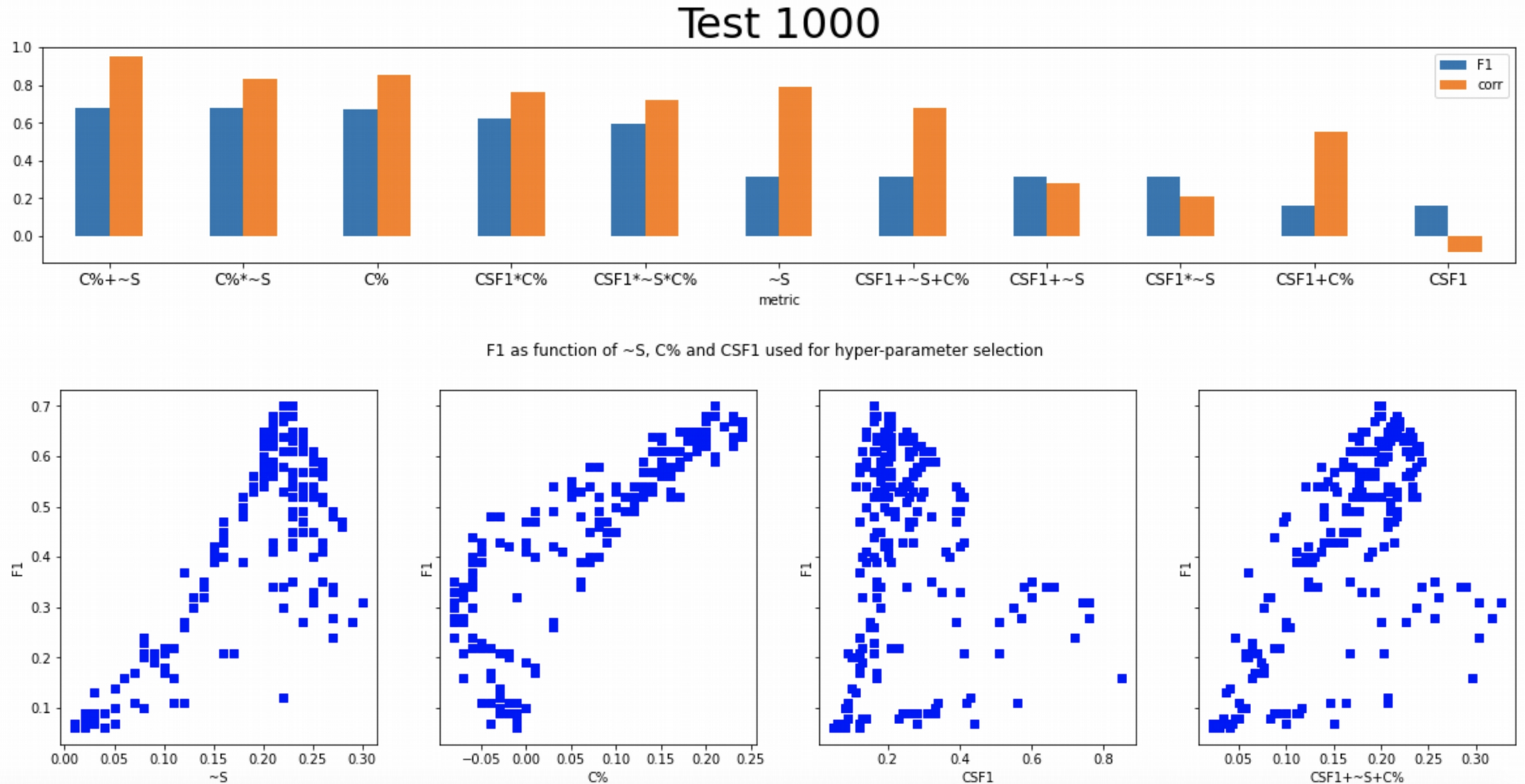
Self-tuning Hyperparameters – English (TF variance)



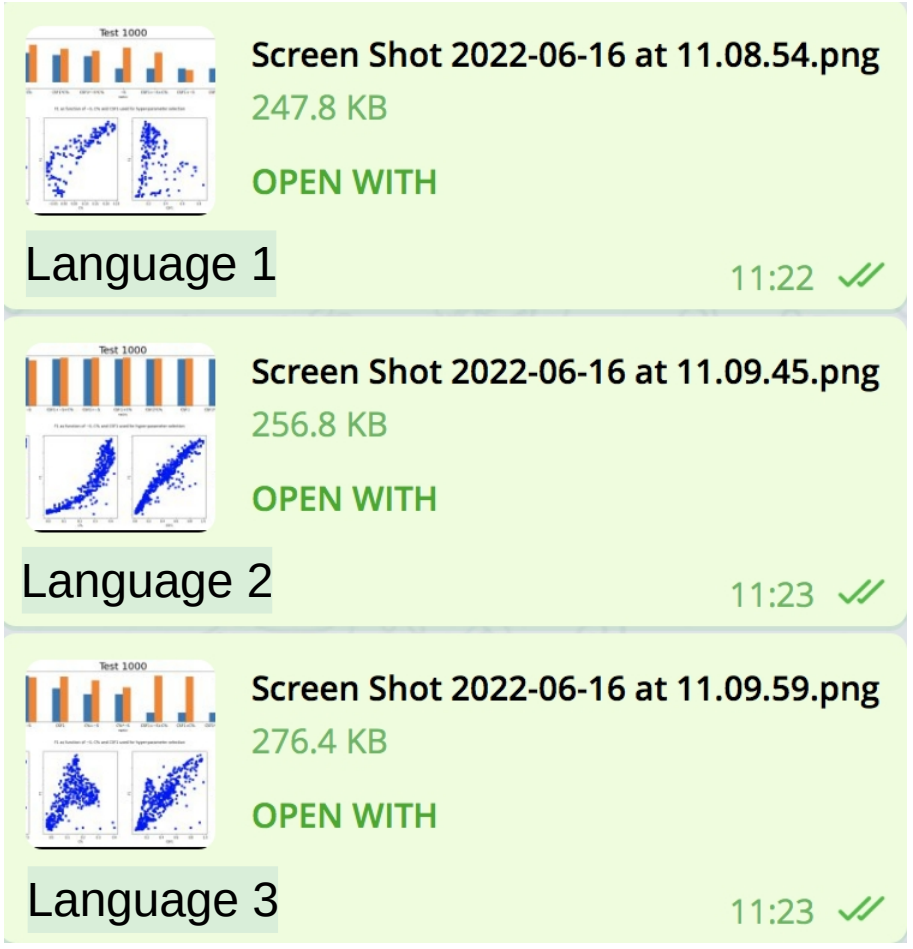
Self-tuning Hyperparameters – Russian (TF variance)

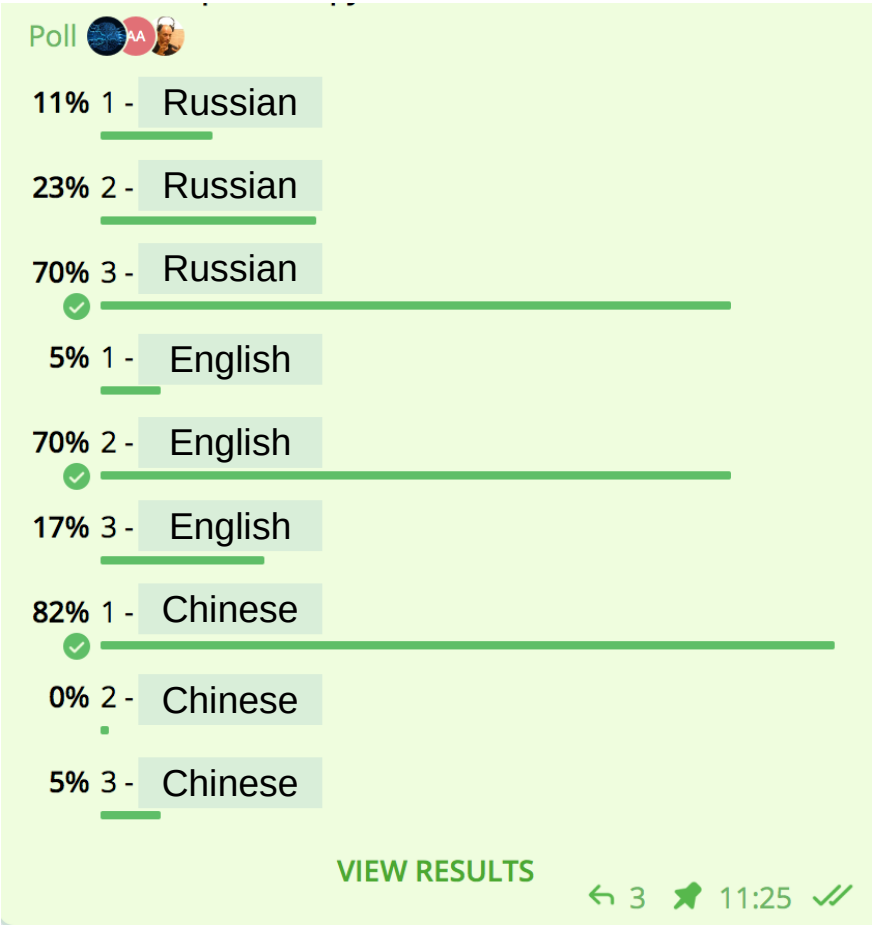


Self-tuning Hyperparameters – Chinese (TF “peak”)



Something about Human Intuition!





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!

Anton Kolonin

akolonin@aigents.com

Facebook: akolonin

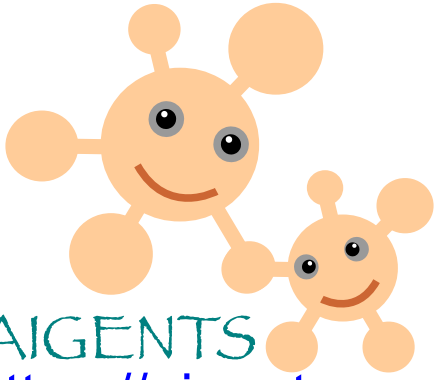
Telegram: akolonin

AGI IN RUSSIAN

https://agirussia.org

Novosibirsk State University

*THE REAL SCIENCE



https://aigents.com