

# Self-tuning hyper-parameters for unsupervised cross-lingual tokenization



<https://agirussia.org>

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

Facebook: akolonin  
Telegram: akolonin

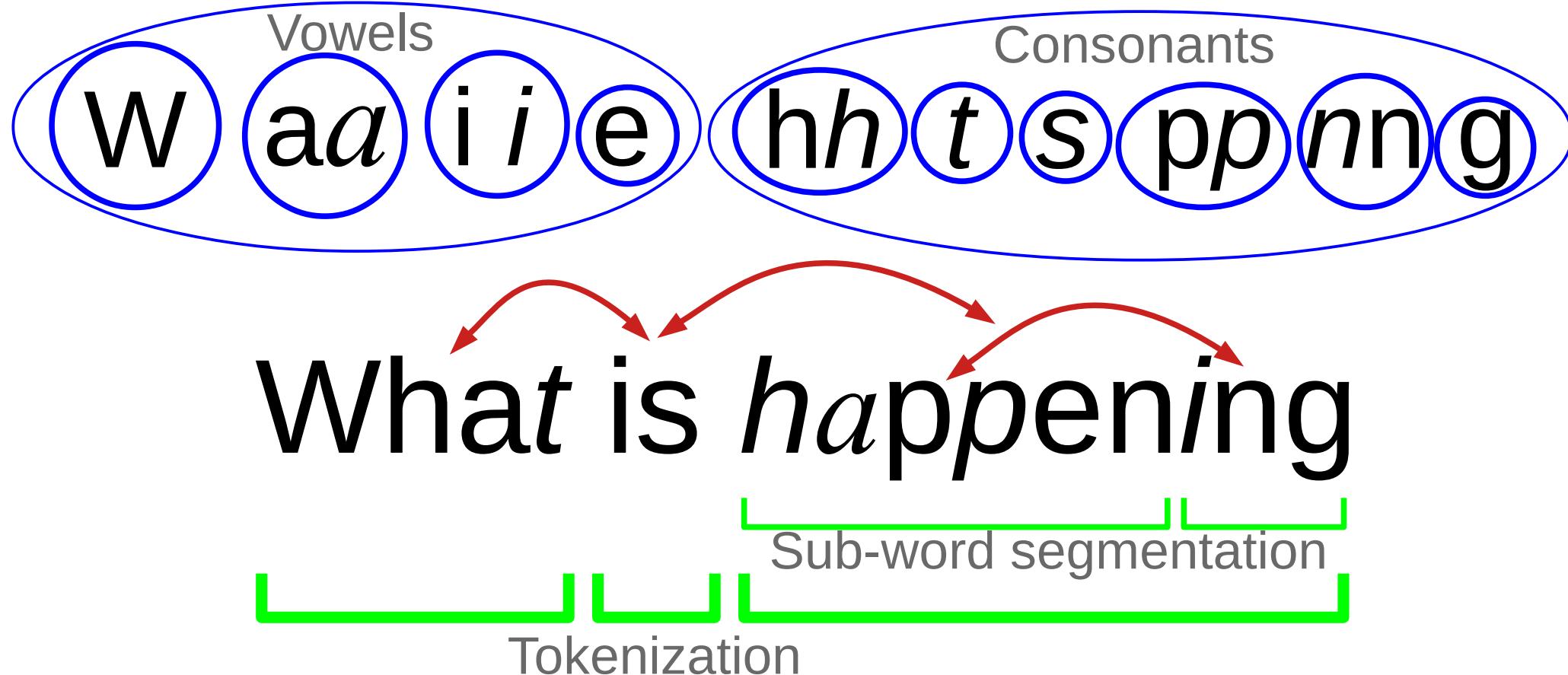


**N\*** Novosibirsk  
State  
University  
\*THE REAL SCIENCE

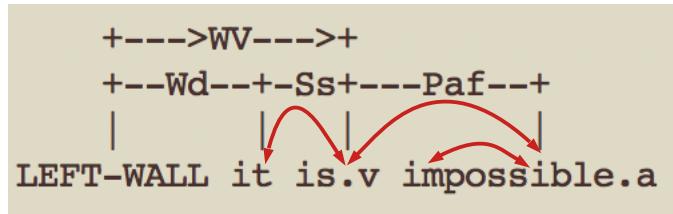


# Learning Lexicon, Punctuation, Morphology and Grammar

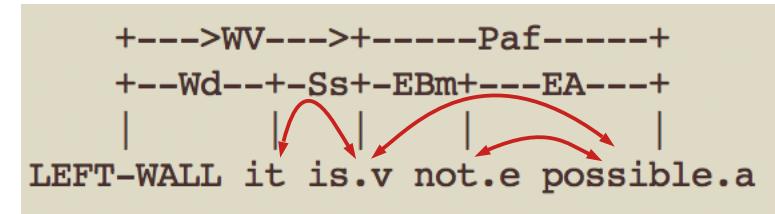
## Clustering, Segmentation and Parsing



# Blurring boundaries between grammatical parsing and morphological parsing causes tokenization ambiguity

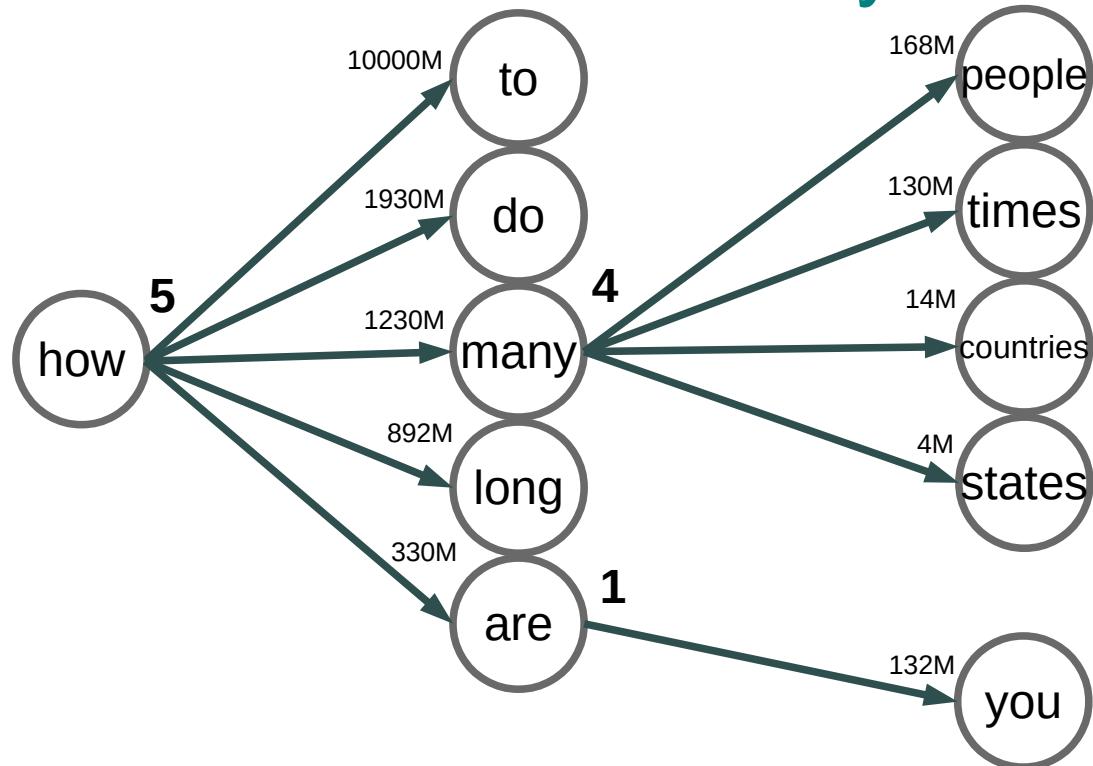


这不可能  
(Google)



这 不 可 能  
(Jieba)

# 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>

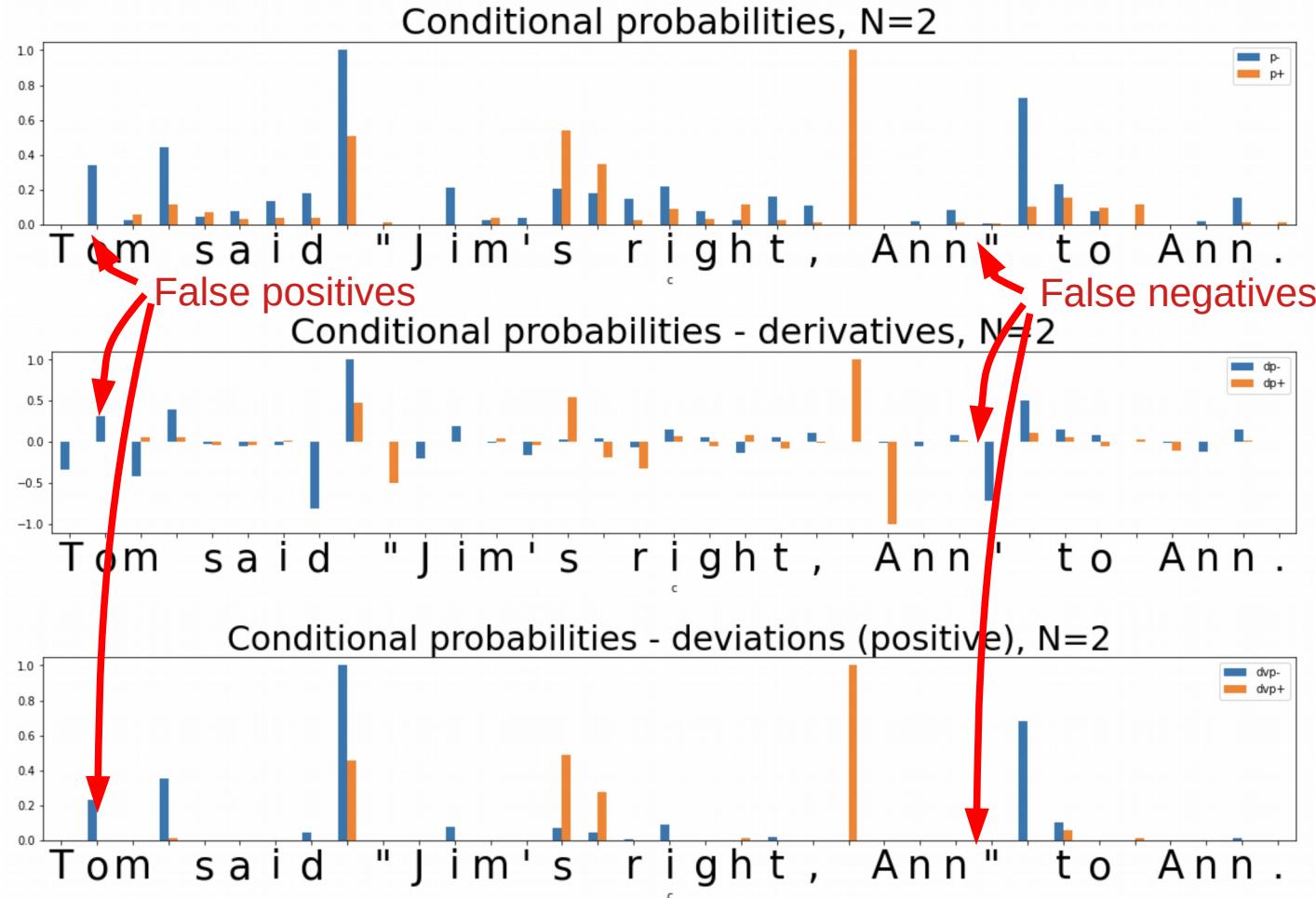
# Unsupervised Text Segmentation (Tokenization)

## Metrics/Indicators:

Ngram (Character)  
Conditional  
Probability  
(of Transition)

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

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

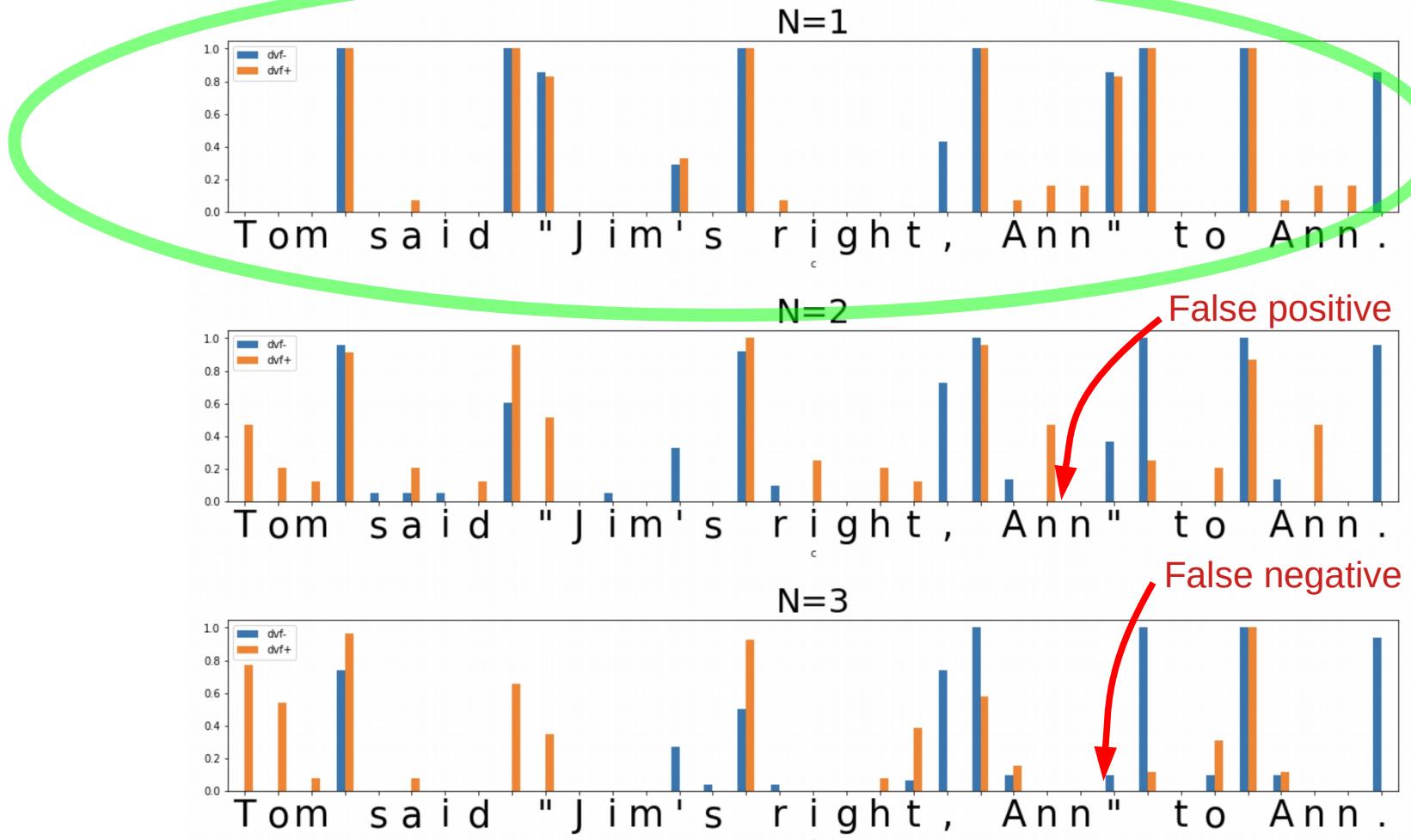


# Unsupervised Text Segmentation (Tokenization)

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	-

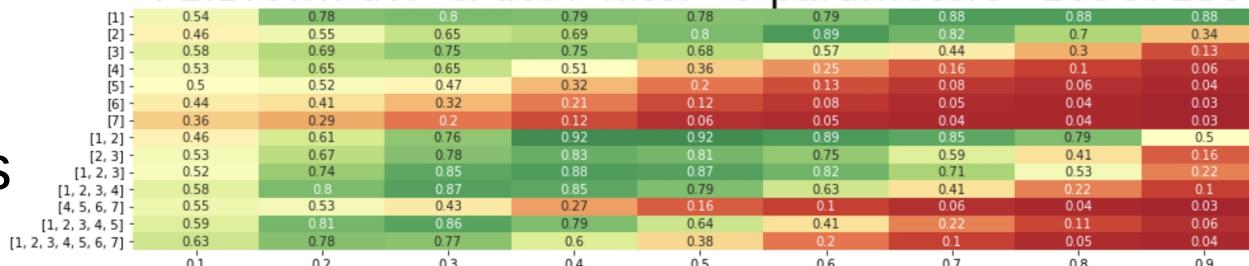
\*Lexicon-based Tokenization - greedy/beam search on word length (optimal) or frequency

# Hyper-parameters: F1 vs. language-agnostic metrics

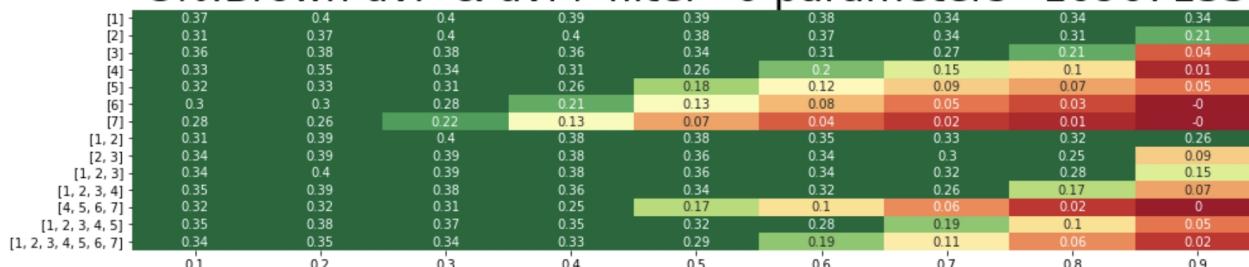
English,  
Brown corpus

Maximizing F1,  
compression  
factor (C%)  
and  
normalized  
anti-entropy  
(~S) in the  
space of hyper-  
parameters

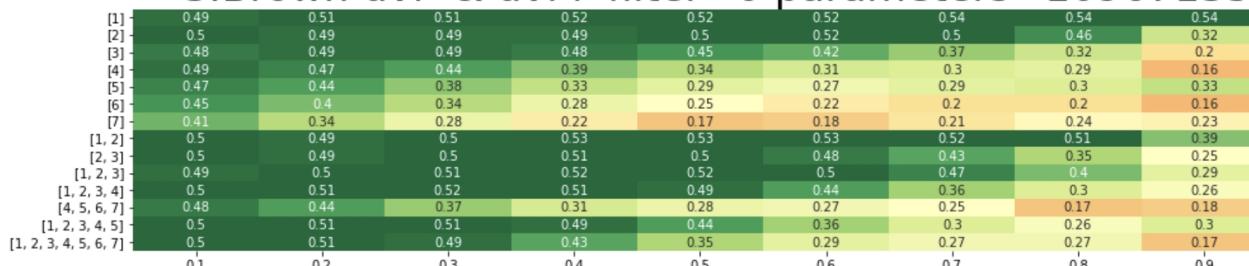
F1:Brown dvf- & dvf+ filter=0 parameters=10967135



C%:Brown dvf- & dvf+ filter=0 parameters=10967135



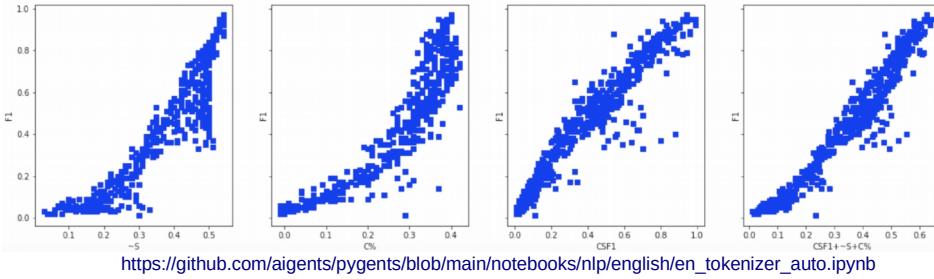
~S:Brown dvf- & dvf+ filter=0 parameters=10967135



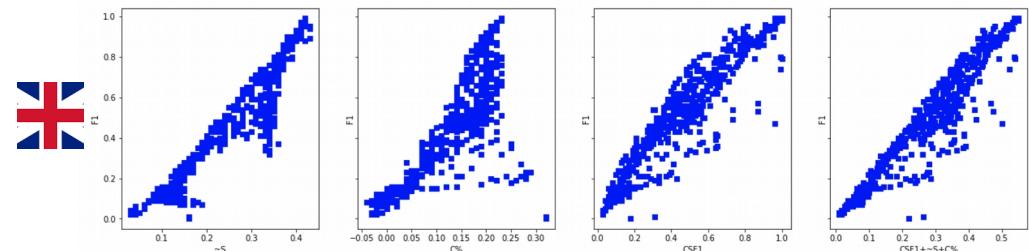
<https://arxiv.org/abs/2303.02427>

# Different corpora, different sizes

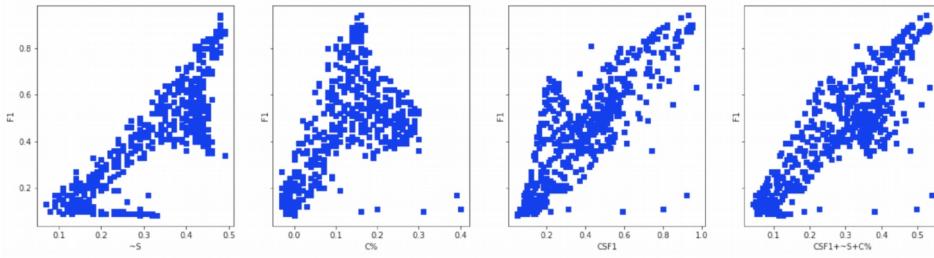
English, Train: Brown, Test: Brown 1000



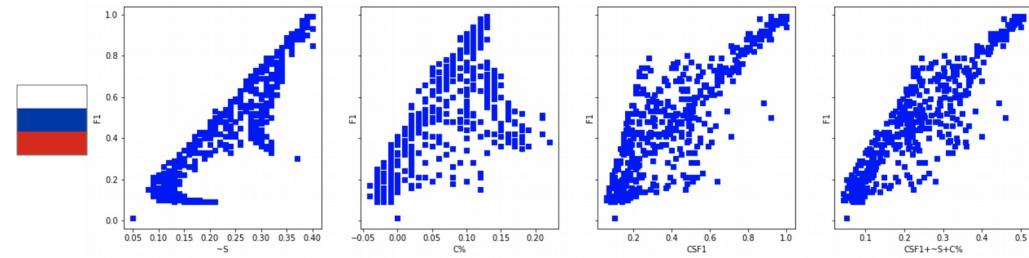
English, Train: Brown, Test: MagicData 100



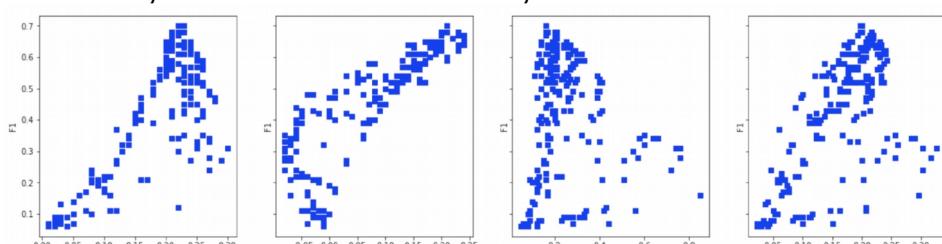
Russian, Train: RusAge, Test: RusAge 1000



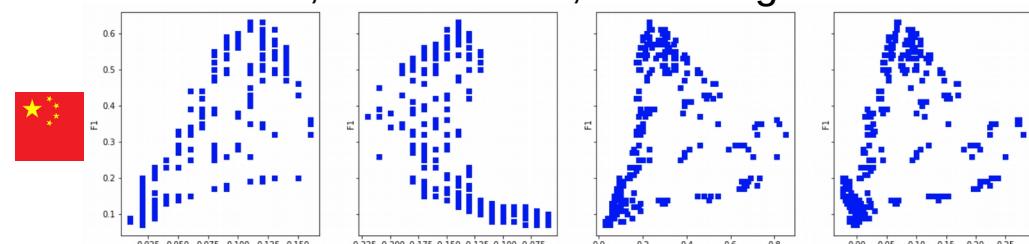
Russian, Train: Brown, Test: MagicData 100



Chinese, Train: CLUE News, Test: CLUE News 1000



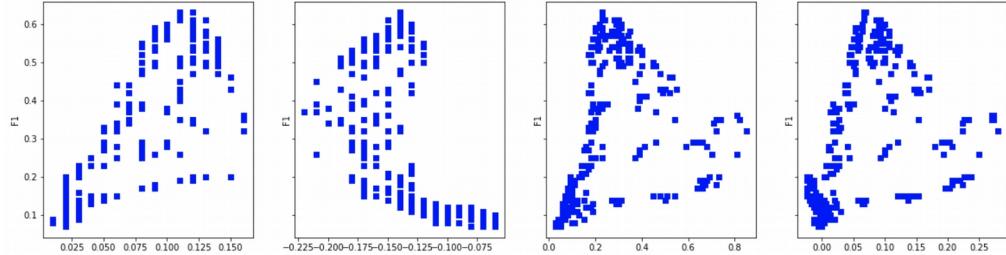
Chinese, Train: Brown, Test: MagicData 100



# Chinese corpora, different sizes

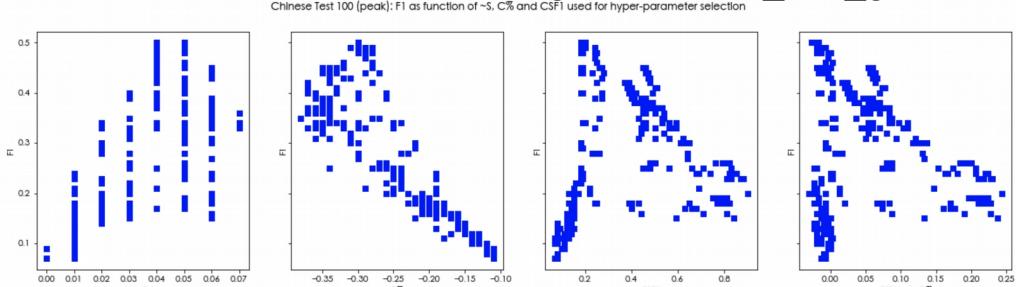


Chinese, Train: Brown, Test: MagicData 100

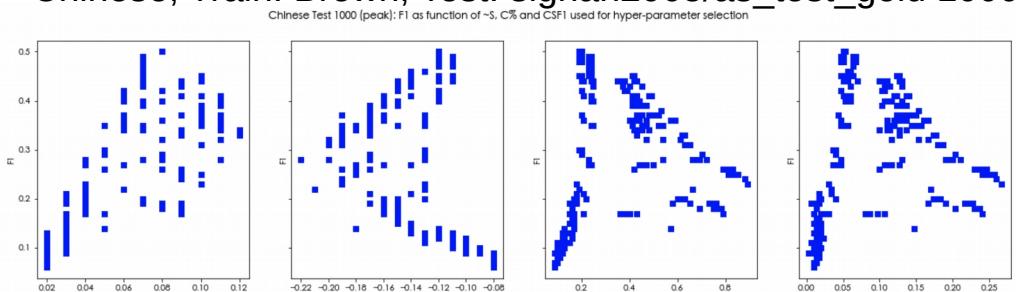


[https://github.com/agents/pygents/blob/main/notebooks/nlp/tokenization/brown/tokenization\\_brown\\_en\\_ru\\_zh.ipynb](https://github.com/agents/pygents/blob/main/notebooks/nlp/tokenization/brown/tokenization_brown_en_ru_zh.ipynb)

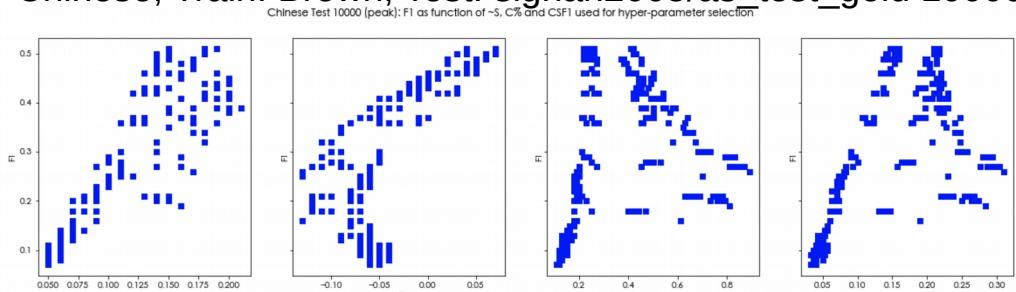
Chinese, Train: Brown, Test: sighan2005/as\_test\_gold 100



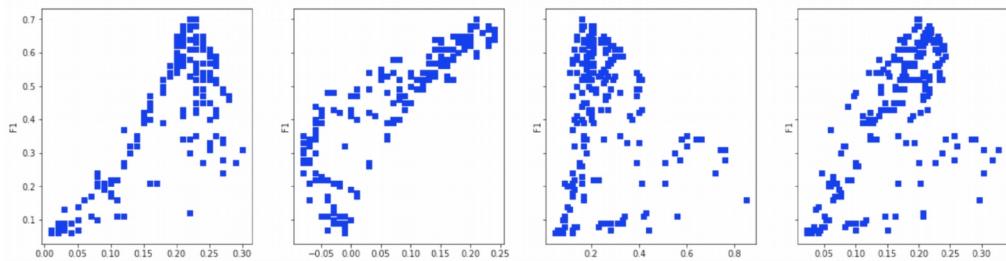
Chinese, Train: Brown, Test: sighan2005/as\_test\_gold 1000



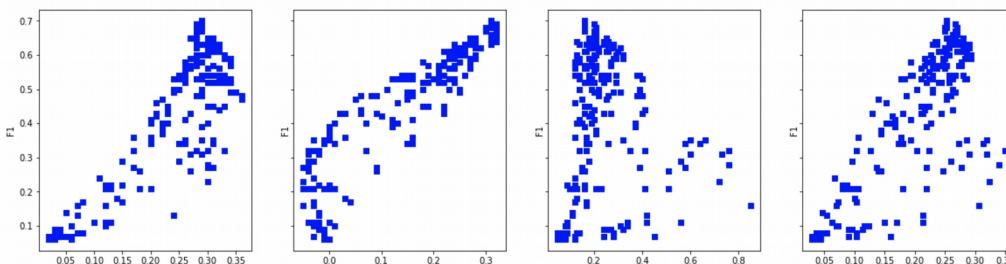
Chinese, Train: Brown, Test: sighan2005/as\_test\_gold 10000



Chinese, Train: CLUE News, Test: CLUE News 1000



Chinese, Train: CLUE News, Test: CLUE News 10000



[https://github.com/agents/pygents/blob/main/notebooks/nlp/chinese/zh\\_tokenizer\\_auto.ipynb](https://github.com/agents/pygents/blob/main/notebooks/nlp/chinese/zh_tokenizer_auto.ipynb)

[https://github.com/agents/pygents/blob/main/notebooks/nlp/chinese/zh\\_tokenizer\\_multi-criteria-cws.ipynb](https://github.com/agents/pygents/blob/main/notebooks/nlp/chinese/zh_tokenizer_multi-criteria-cws.ipynb)

# Unsupervised learning for subword segmentation



Reference	Morphology-based	BPE	DPE	Transition-freedom-based
['euro', 'zone']	['eu', 'rozone']	['eurozone']	['euro', 'zone']	['euro', 'z', 'one']
['entrepreneur', 'ship']	['ent', 're', 'pre', 'neur', 'ship']	['entrepreneurship']	['entrepreneur', 'ship']	['entre', 'preneur', 'sh', 'ip']
['pre', 'sent', 'ly']	['pre', 's', 'ent', 'ly']	['pres', 'ently']	['present', 'ly']	['pre', 's', 'ently']
['bloc']	['bloc']	['blo', 'c']	['bl', 'oc']	['b', 'lo', 'c']
['tree', 's']	['tr', 'ee', 's']	['tre', 'es']	['tr', 'ees']	['tre', 'es']
['multi', 'lateral', 'ism']	['multi', 'lat', 'er', 'al', 'ism']	['multilater', 'alism']	['multilateral', 'ism']	['multi', 'later', 'al', 'ism']
['motive', 's']	['mot', 'ive', 's']	['mo', 'tives']	['motiv', 'es']	['mo', 'tiv', 'es']
['progress', 'ive', 'ly']	['pro', 'gr', 'ess', 'ive', 'ly']	['pro', 'gressively']	['progressive', 'ly']	['pro', 'gressiv', 'ely']
['de', 'cent', 'ral', 'isation']	['dec', 'ent', 'r', 'al', 'isation']	['decent', 'ralisation']	['decent', 'ral', 'isation']	['de', 'centralis', 'ation']
['margin', 'al', 'is', 'ed']	['marginali', 's', 'ed']	['margin', 'alised']	['marginal', 'ised']	['mar', 'ginal', 'is', 'ed']
['re', 'cast']	['re', 'cast']	['rec', 'ast']	['re', 'cast']	['re', 'c', 'ast']
['out', 'line', 's']	['out', 'l', 'ine', 's']	['out', 'lines']	['outline', 's']	['out', 'l', 'ines']
['pre', 'vent', 'at', 'ive']	['pre', 'v', 'ent', 'ative']	['preven', 'tative']	['prevent', 'ative']	['pre', 'vent', 'ative']
['en', 'danger', 'ed']	['end', 'an', 'g', 'er', 'ed']	['endang', 'ered']	['endanger', 'ed']	['en', 'dang', 'ered']
['vulner', 'abil', 'ity']	['vulnerabil', 'ity']	['vul', 'n', 'era', 'bility']	['vul', 'ner', 'ability']	['vul', 'ner', 'ability']

Ref:

<https://arxiv.org/pdf/2005.06606.pdf>

F1

0.46

0.05

0.55

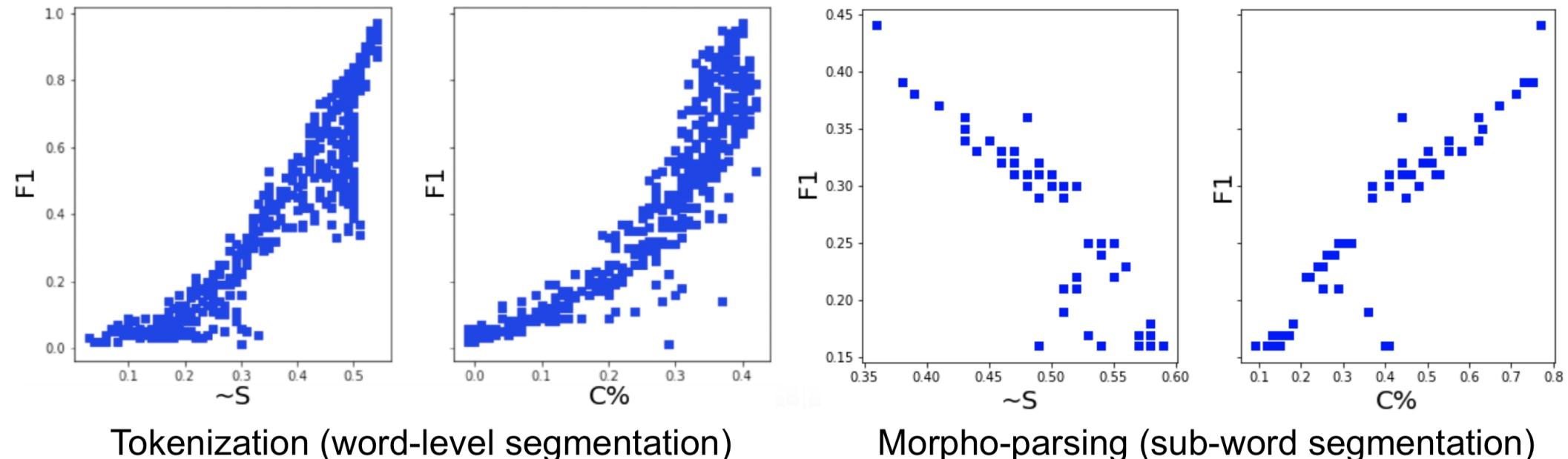
0.25

<https://github.com/aigents/pygents/blob/main/data/corpora/English/morphology/prefixes.txt>  
<https://github.com/aigents/pygents/blob/main/data/corpora/English/morphology/suffixes.txt>

# Tokenization vs. sub-word segmentation ✪

## F1 connection to anti-entropy and compression factor (English)

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



# Takeaways

Grammar and syntactic/semantic word categories **can** be learned, given we can learn tokenization and parses unsupervisedly.

Tokenization and character categories **can** be learned unsupervisedly based on transition freedom metric.

Hyper-parameters for unsupervised text segmentation learning **can** be found based on culture-agnostic metrics such as compression factor (and anti-entropy in case of tokenization).

Unsupervised learning for morphological units and morpho-parsing **may** be possible but remains non-trivial and needs more study.

# Welcome to the Interpretable Natural Language Processing Community and Series of Workshops

AGI-in-Russian on Telegram

<https://t.me/agirussia>

INLP on Telegram

<https://t.me/internlp>

INLP Workshops

<https://agents.github.io/inlp/>