

# Interpretable Natural Language Processing

## Applied results and fundamental studies



<https://agirussia.org>

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

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

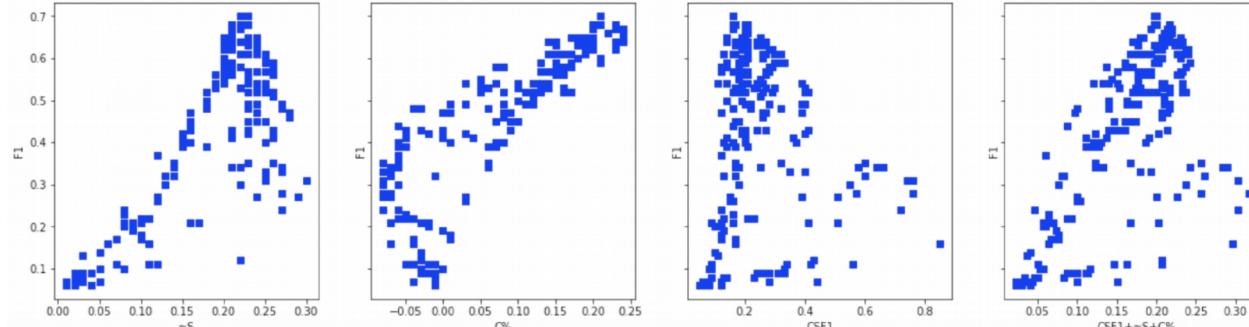


 SingularityNET

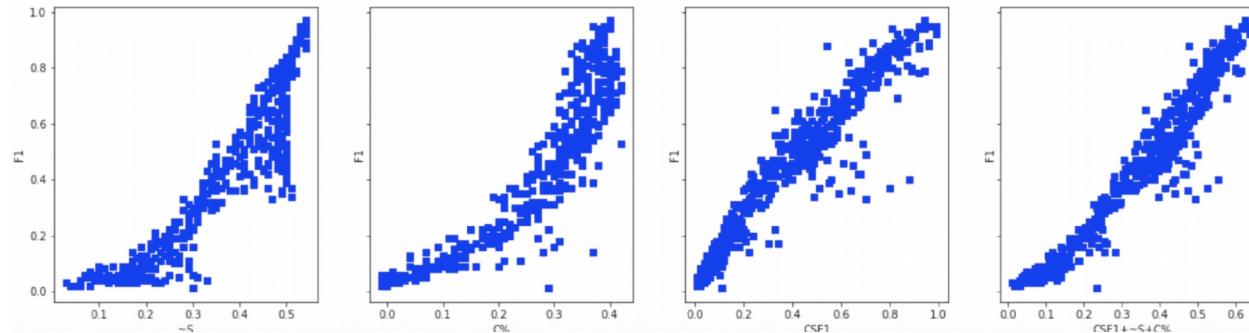
 SingularityDAO

# The riddle: find English, Chinese and Russian!

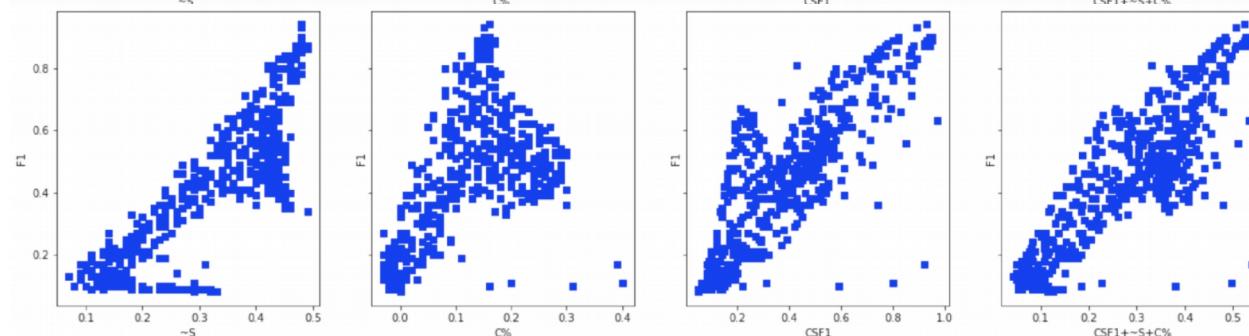
Language 1



Language 2



Language 3



# The plan

## Introduction

## Applications

Entity extraction and attribution

Mining for sentiment and cognitive distortions

Language parsing and segmentation

Language generation and question answering

## Fundamental studies

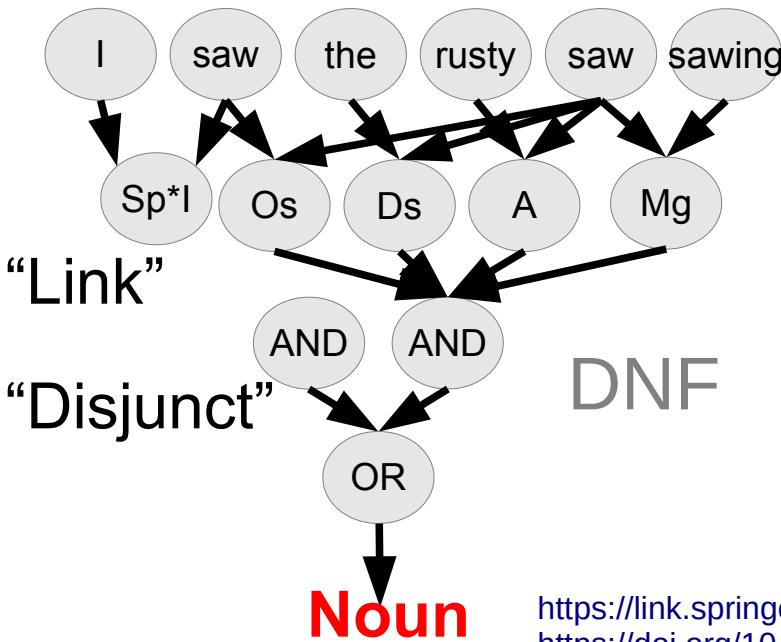
Language Learning for Grammar

Language Learning for Lexicon, Punctuation and Morphology

# Introduction

# Bridging the Symbolic-Subsymbolic gap in NLP

Formal  
Link Grammar



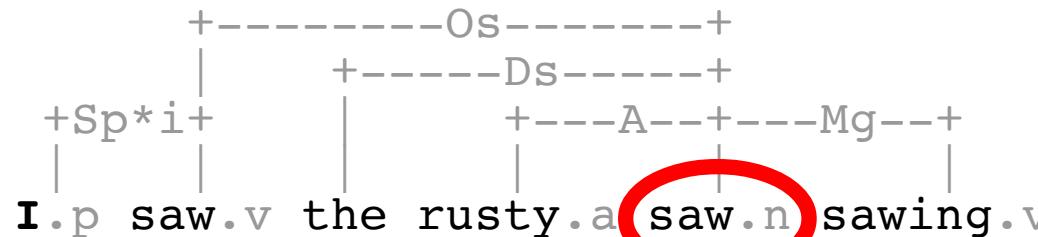
"Link"

"Disjunct"

DNF



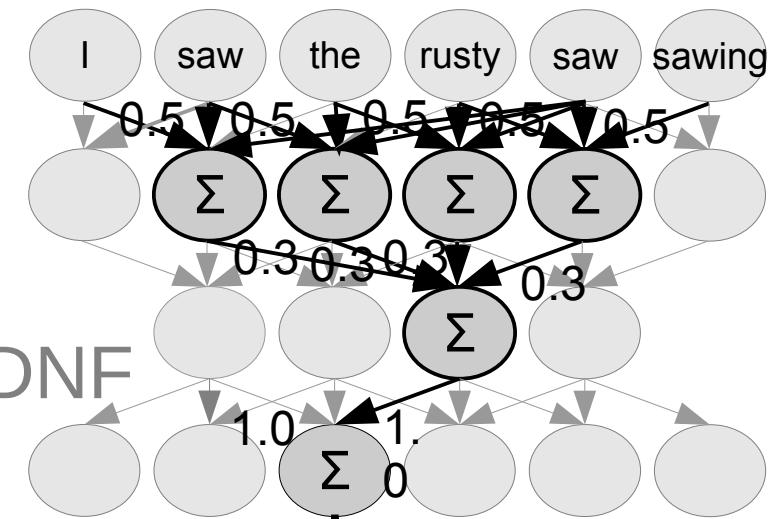
SingularityNET



Deep  
Language Model

← Explain  
Transfer →

Soft-DNF

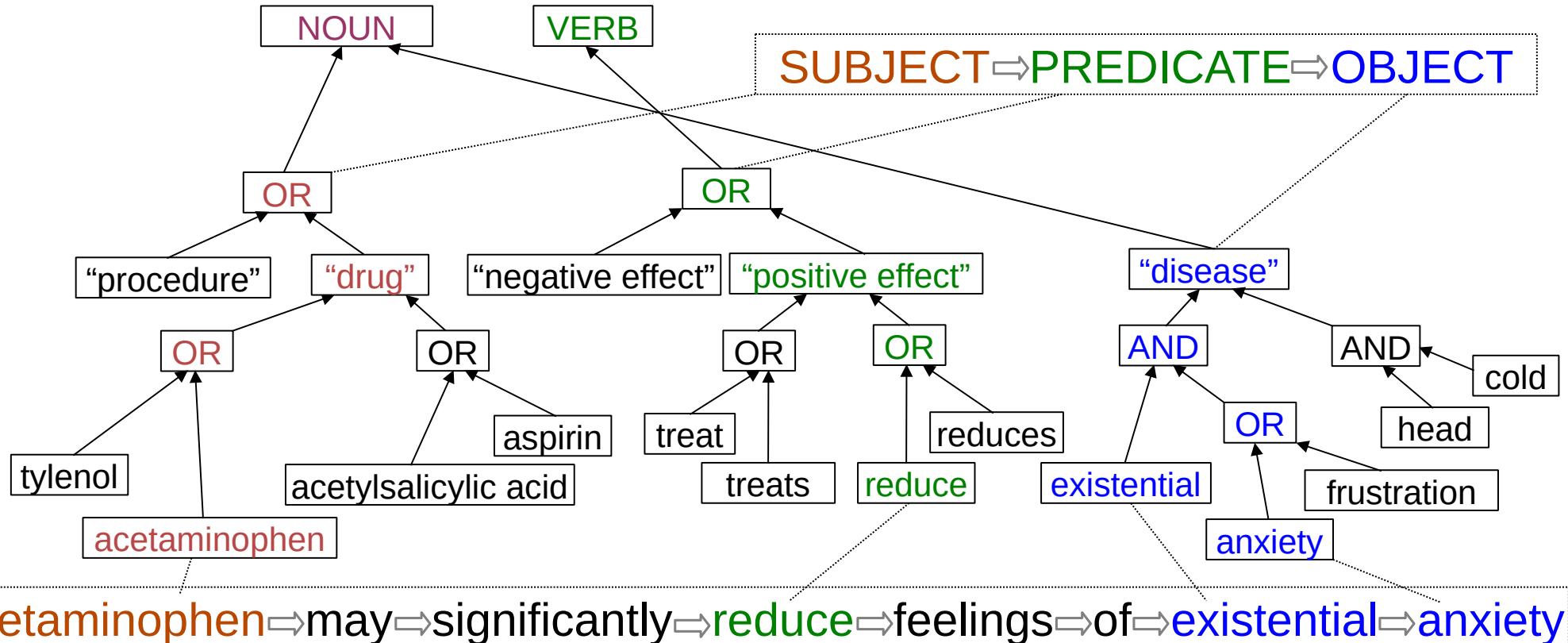


Noun

[https://link.springer.com/chapter/10.1007/978-3-319-97676-1\\_11](https://link.springer.com/chapter/10.1007/978-3-319-97676-1_11)  
[https://doi.org/10.1007/978-3-030-27005-6\\_11](https://doi.org/10.1007/978-3-030-27005-6_11)

<https://github.com/singnet/language-learning/>  
Copyright © 2023 Anton Kolonin, Aigents®

# 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>

# Applications

# Aigents® “Deep Patterns” - Generalized Text Mining

## Classification

Category:

“Healthcare”

tylenol  
acetaminophen  
placebo

/S

Here's the Tylenol twist: Before they began writing, half of each group received acetaminophen while the other half swallowed a placebo. Even among those people who wrote about death, the Tylenol takers set bail at roughly \$300—a sign that acetaminophen may significantly reduce feelings of existential anxiety, explains study lead author Daniel Randles, a PhD candidate in UBC's department of... psychology.

## Case/Relationship Extraction

Entity (Case): “Treatment:  
Healing anxiety with Tylenol”

significantly  
reduce  
feelings  
study

HAS

“acetaminophen may  
significantly reduce  
feelings of existential  
anxiety, explains study  
lead author Daniel  
Randles”

## Property Attribution Entity Extraction

Brand: Tylenol

Substance: acetaminophen

Reliability: medium

Effect: positive

Diagnosis: Anxiety

Reporter: Daniel Randles

acetaminophen  
may  
reduce  
anxiety  
explains

acetaminophen may  
significantly reduce  
feelings of existential  
anxiety, explains study  
lead author Daniel  
Randles.

# Aigents® “Deep Patterns” - Property Attribution

```
<set> := <disjunctive-set> | <conjunctive-set> | <M-skip-N-gram>
<disjunctive-set> := { <pattern> * }
<conjunctive-set> := ( <pattern> * )
<N-gram> := [ <pattern> * ]
<pattern> := <token> | <regexp> | <variable> | <set>
```

Variables may have domain restrictions  
in ontology and/or refer to other  
patterns as subgraphs

## Example:

```
{[$description catheter] [$coating coating] [$inner-diameter
    {diameter inner-diameter}] [$tip tip] [$pattern pattern]}
```

X

“Convey Guiding Catheter. Unique hydrophilic coating.

Smallatraumatic soft tip. Ultra-thin 1 × 2 flat wire braid pattern”

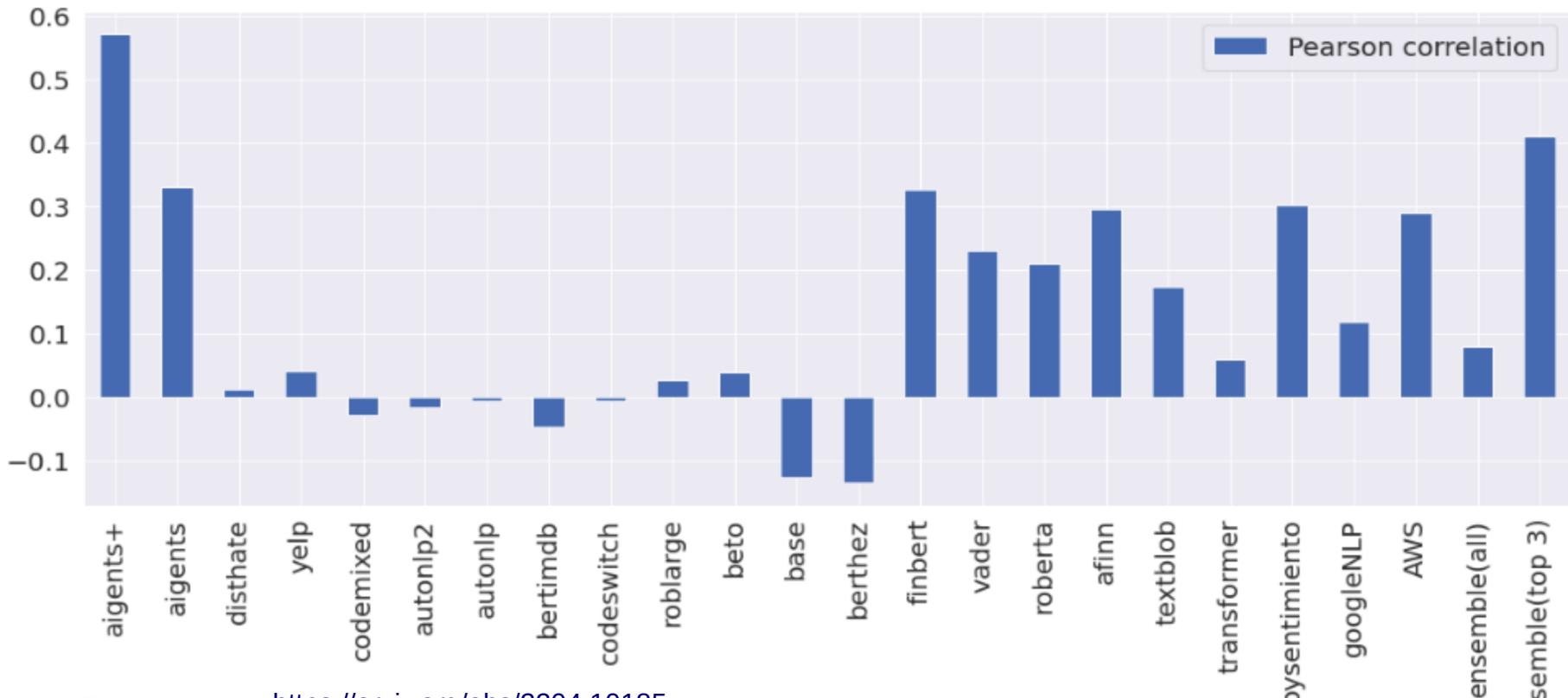
=

```
{ coating : "hydrophilic", description : "convey guiding",
    pattern : "ultra-thin 1 × 2 flat wire braid", tip : "soft" }
```

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

# Sentiment Analysis – Models' Fight

Aigents® “interpretable” model vs. Bert fine-tuned on financial data  
Average correlation across all models



# N-M-skip-grams for Aigents® Sentiment Analysis

A screenshot of the Aigents news feed. The interface includes a top navigation bar with links for Aigents, Topics, Sites, News, Friends, Graph, Chat, and a user profile for 'Madmind Siberian'. Below the navigation is a search bar labeled 'Input search text' with a '+' button. The main area displays a list of news items with their publication date (today, yesterday), a small thumbnail image, a brief summary, and a link. Each news item has a colored sentiment bar above it, ranging from blue (positive) to red (negative). The news items include:

- today: putin says he has noted joe biden's sharp anti-Russian rhetoric ([https://www.reddit.com/r/JoeBiden/comments/j6zeex/putin\\_says\\_he\\_has\\_note](https://www.reddit.com/r/JoeBiden/comments/j6zeex/putin_says_he_has_note))
- today: the united states of america is set for a pivotal election where joe bide current president donald trump (<https://moderndiplomacy.eu/2020/09/12/kamala-harris-as-vice-president-attractive-for>)
- yesterday: \*\*learn more about blockchain domains:\*\* introduction video see you on thursday ([https://www.reddit.com/r/CryptoCurrency/comments/j6rjh/ama\\_with Brad\\_k](https://www.reddit.com/r/CryptoCurrency/comments/j6rjh/ama_with Brad_k))
- yesterday: The Russians may know more than we do : cia analyst says putin I doctors amid covid secrecy ([https://www.reddit.com/r/politics/comments/j6t6yp/the\\_russians\\_may\\_know\\_more](https://www.reddit.com/r/politics/comments/j6t6yp/the_russians_may_know_more))
- yesterday: alexa-derived chatbot tech allegedly understands the british accent ([https://www.theregister.co.uk/2020/09/17/amazon\\_alex\\_can\\_speak\\_british\\_english](https://www.theregister.co.uk/2020/09/17/amazon_alex_can_speak_british_english))
- yesterday: but not strongly enough to make a difference published: 16 mar 2 bernie sanders lost his last chance to take joe biden down (<https://www.theguardian.com/commentisfree/2020/mar/16/bernie-sanders-lost-h>)
- yesterday: it must also support in-space assembly manufacturing of hardware be tested and evaluated for future use ([https://www.theregister.co.uk/2020/07/18/sierra\\_nevada\\_space\\_station](https://www.theregister.co.uk/2020/07/18/sierra_nevada_space_station))

Copyright 2020 IP Anton Kolonin, Aigents®, Privacy Policy

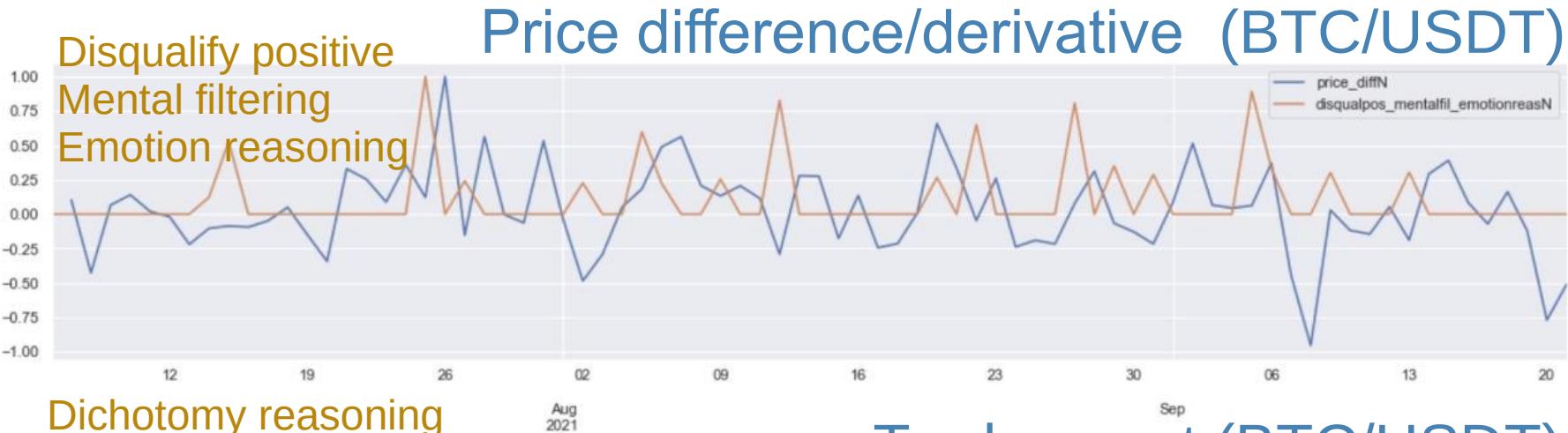
A screenshot of the Aigents search results page. The interface is similar to the news feed, with a top navigation bar and a search bar labeled 'Input new thing name or template' with a '+' button. The main area displays a list of search terms with colored bars above them. The search terms include:

- biden (red)
- jinpingle (green)
- online social help (blue)
- personal artificial intelligence (red)
- putin (red)
- sanders (red)
- trump (red)
- {agi hlai [strong ai] [strong artificial intelligence] [artificial general intelligence]} (blue)
- {chatbot chat-bot [chat bot]} (green)
- {published putin vladimir} (blue)
- {bitcoin btc btc/usd btcusd} (blue)
- {dollar usd usd/xau usdxau} (blue)
- {ethereum eth eth/usd ethusd} (blue)
- {gold xau xau/usd xauusd} (blue)

Copyright 2020 IP Anton Kolonin, Aigents®, Privacy Policy

<https://blog.singularitynet.io/aigents-sentiment-detection-personal-and-social-relevant-news-be989d73b381>

# Mining Cognitive Distortions for FinTech



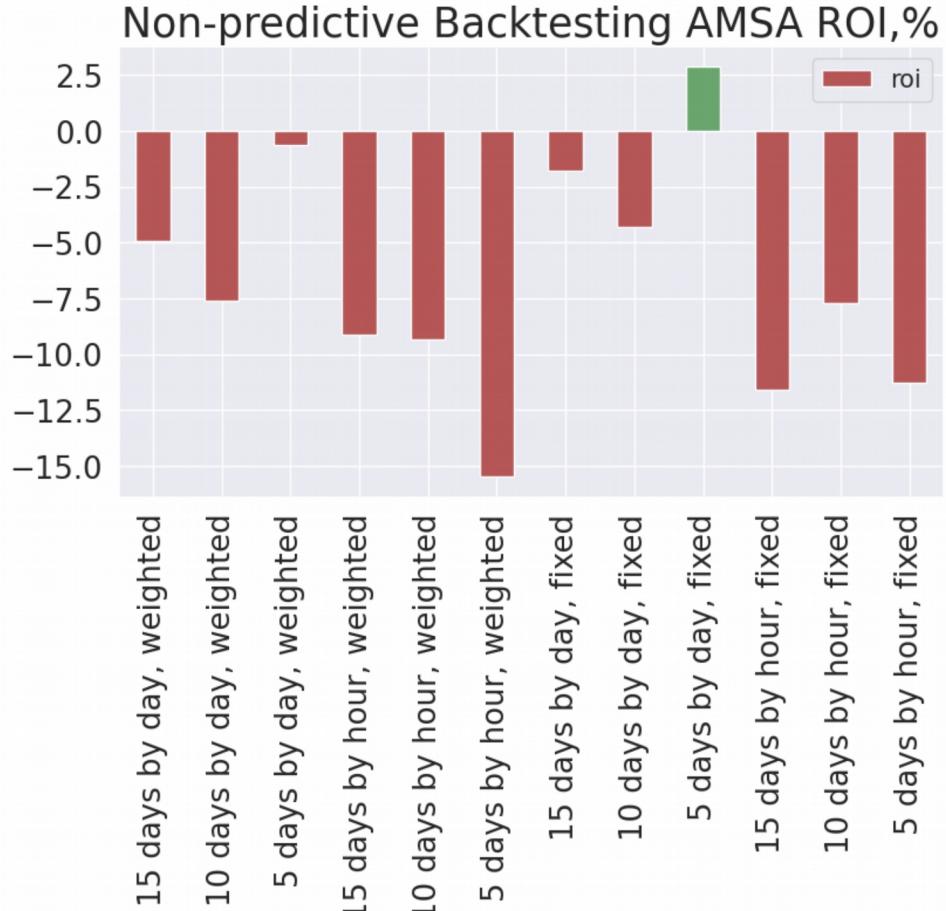
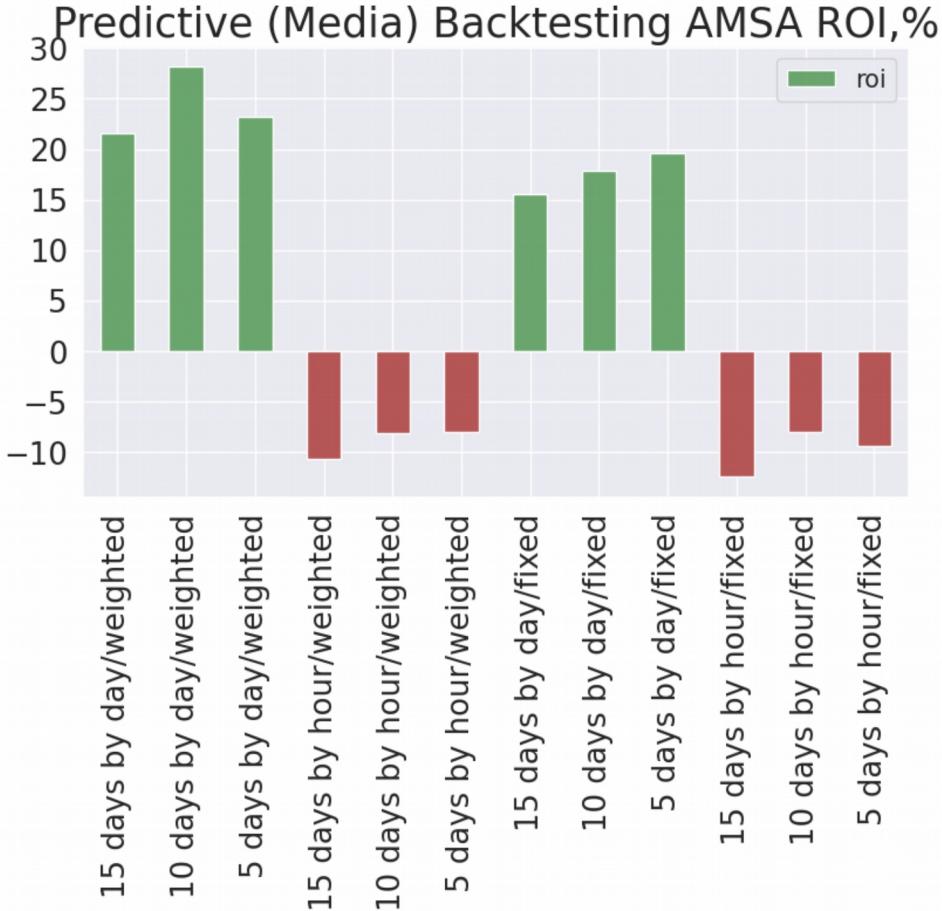
<https://arxiv.org/pdf/2303.02342.pdf>

[https://link.springer.com/chapter/10.1007/978-3-031-19907-3\\_4](https://link.springer.com/chapter/10.1007/978-3-031-19907-3_4)

<https://www.pnas.org/content/118/30/e2102061118>

Copyright © 2023 Anton Kolonin, Aigents®

# Applying Cognitive Distortions in Crypto-Trading

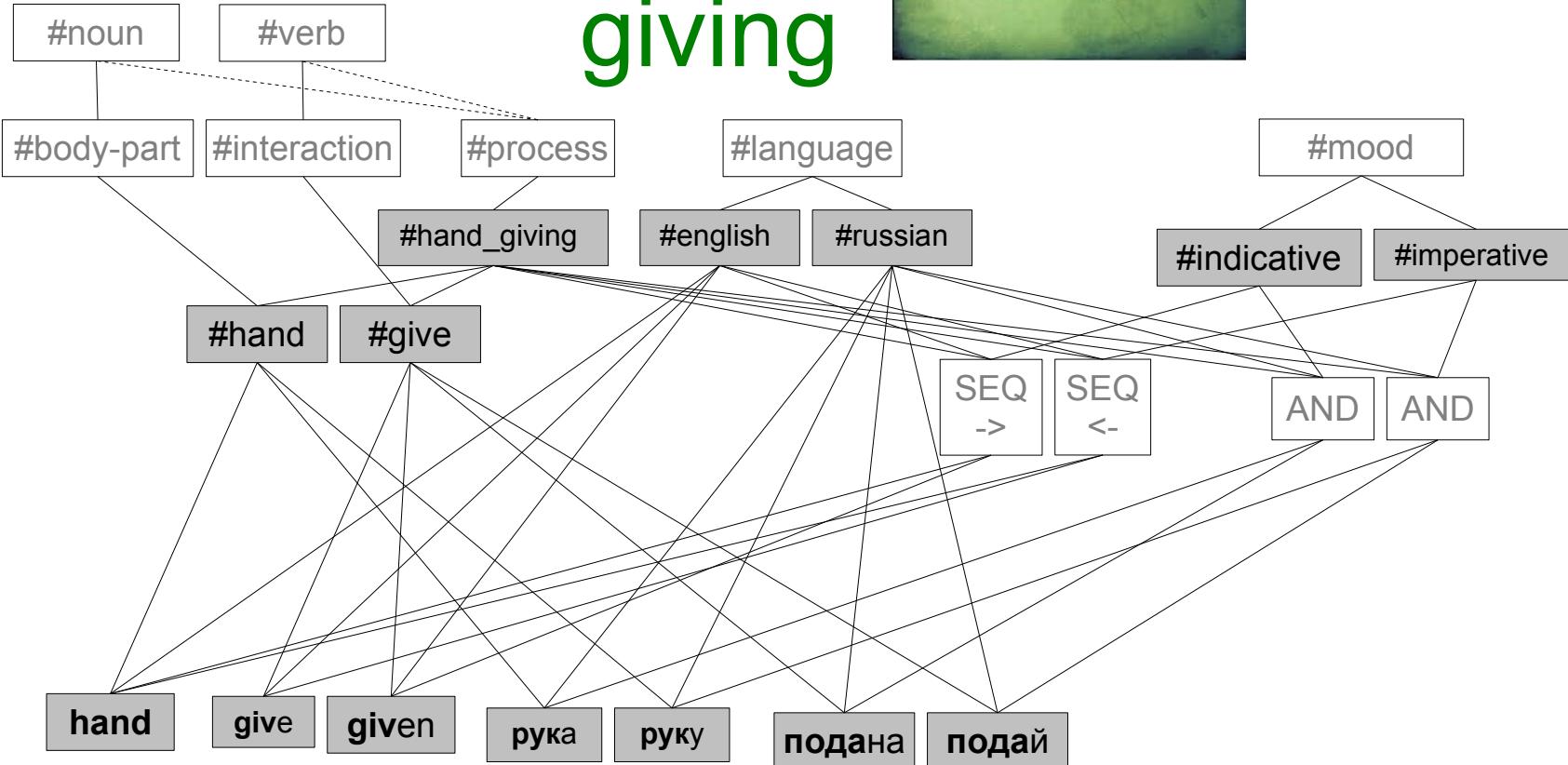


<https://arxiv.org/abs/2303.02342>

Copyright © 2023 Anton Kolonin, Agents®

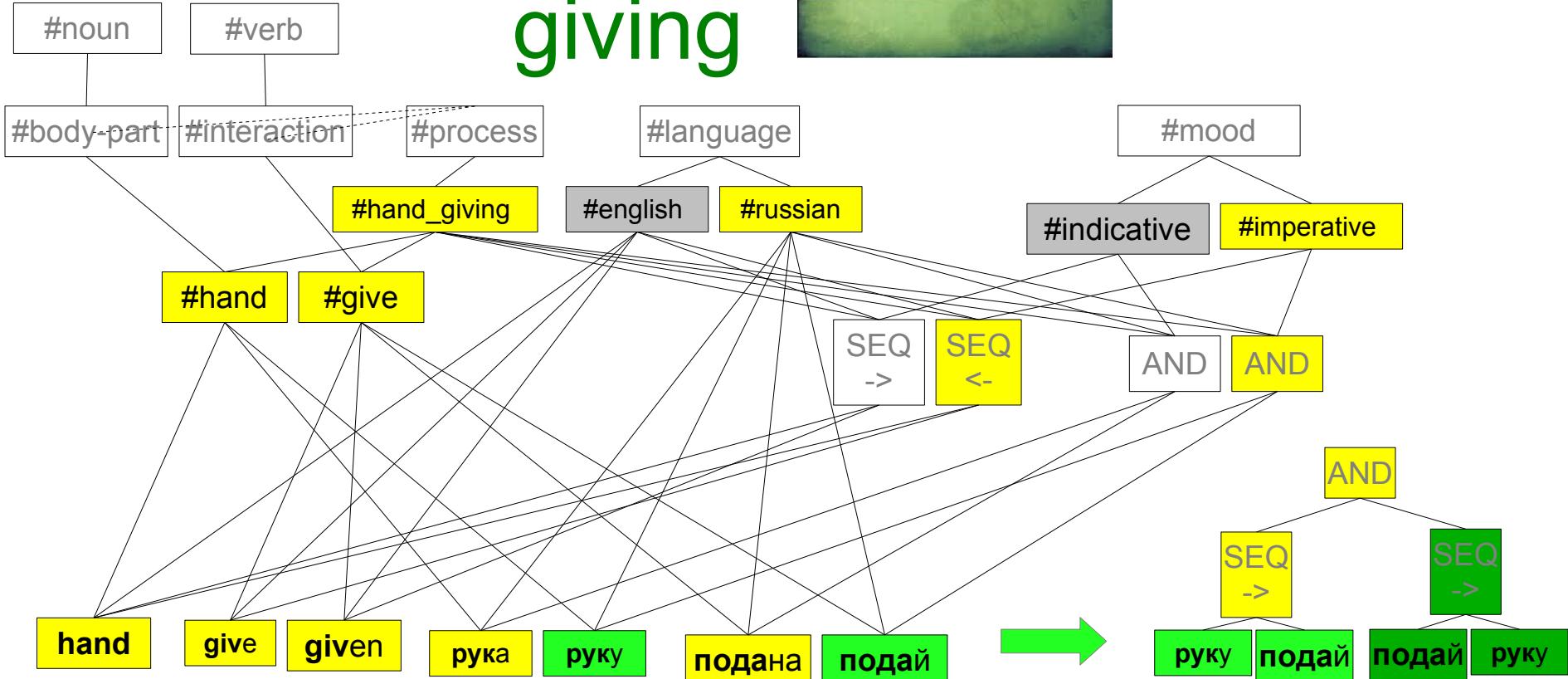
# Grammar & Ontology Graph - Structure

## Hand giving

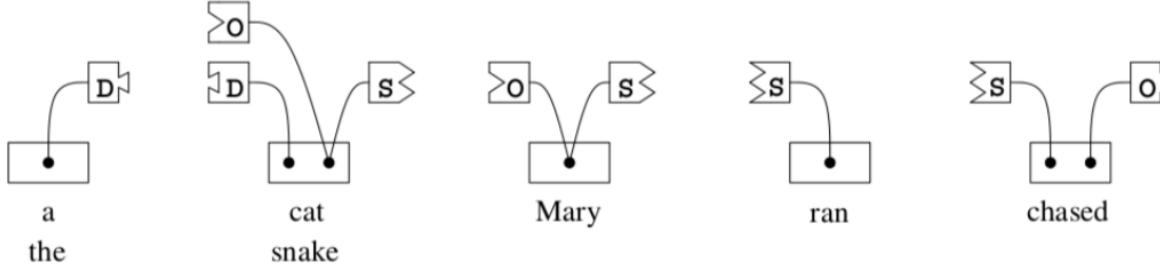


# Grammar & Ontology Graph - Generation

## Hand giving



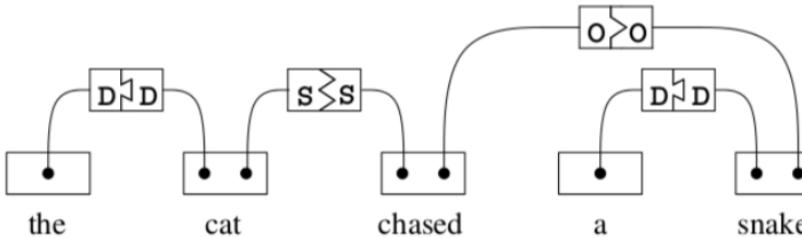
# Link Grammar Connectors & Disjuncts



An illustration of Link Grammar connectors and disjuncts. The connectors are the jigsaw-puzzle-shaped pieces; connectors are allowed to connect only when the tabs fit together. A disjunct is the entire (ordered) set of connectors for a word. As lexical entries appearing in a dictionary, the above would be written as

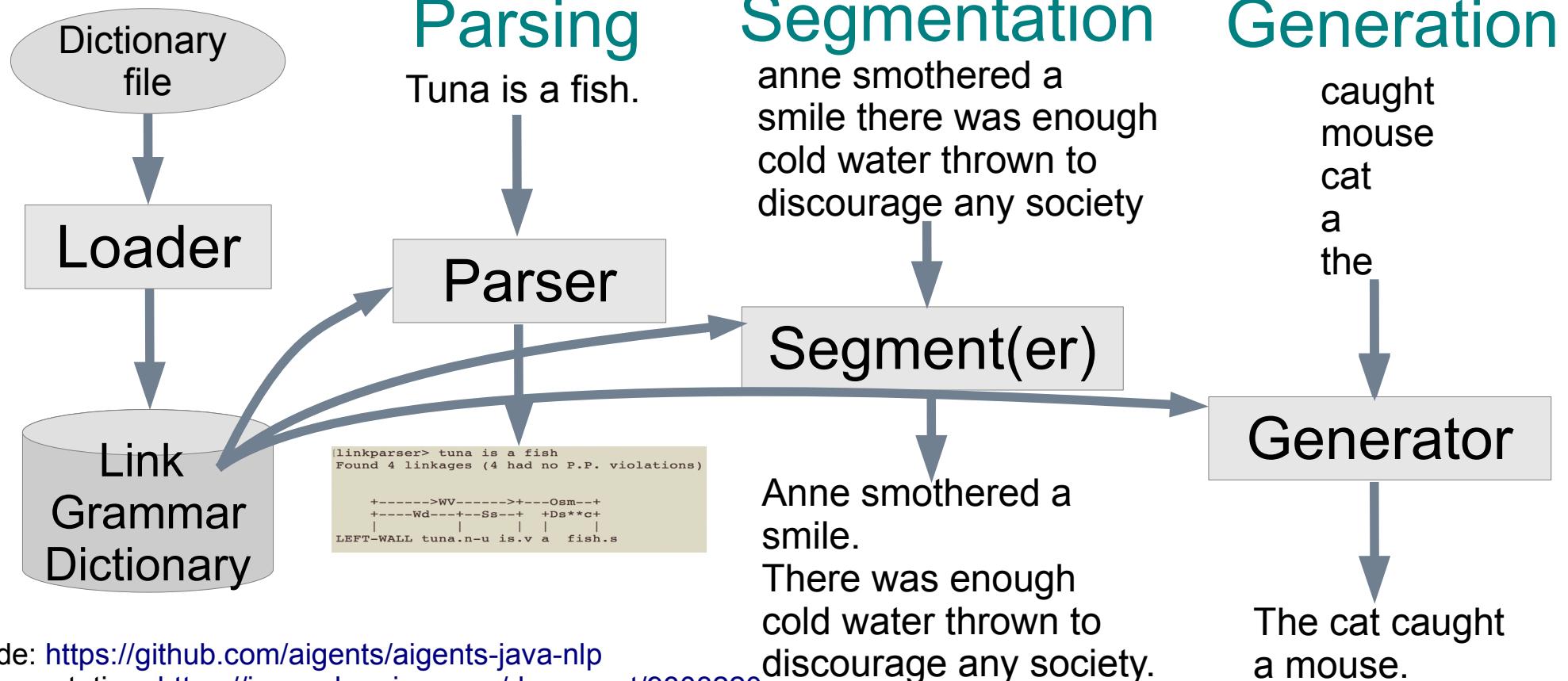
```
a the: D+;  
cat snake: D- & (S+ or O-);  
Mary: O- or S+;  
ran: S-;  
chased S- & O+;
```

Note that although the symbols ‘‘&’’ and ‘‘or’’ are used to write down disjuncts, these are **not** Boolean operators, and do **not** form a Boolean algebra. They do form a non-symmetric compact closed monoidal algebra. The diagram below illustrates puzzle pieces, assembled to form a parse:



<https://arxiv.org/abs/1401.3372>  
B. Goertzel, L. Vepstas, 2014

# Link Grammar – not just for parsing only



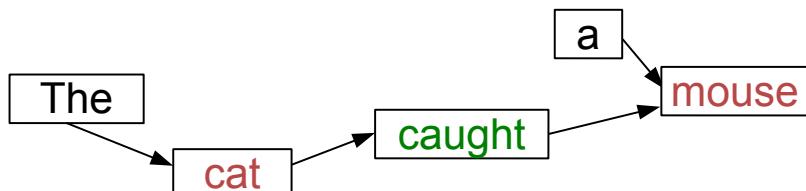
# Grammatical Language Generation

Generator determines what sentences can be formed from a given list of words via valid Link Grammar rules:

- 1) Given a list of words, the Generator determines a subset of all orderings of those words that satisfies initial checks of the planarity and connectivity metarules.
- 2) For each ordering in the subset, the Generator determines if that ordering is valid; specifically, it ensures that every pair of consecutive words can be connected via links part of the Dictionary objects. To do so, the Generator uses the `connects()` function, which returns a boolean value indicating whether its two parameters left and right can be linked together.

**Planarity metarule:** links do not cross

**Connectivity metarule:** links and words of a sentence must form a connected graph that can be completely traversed via one path



---

## Algorithm 2: CONNECTS

---

**Input :** A pair of strings *left* and *right*, representing the two words to potentially be connected

**Output:** An boolean value indicating whether *left* and *right* can be connected via valid Link Grammar rules

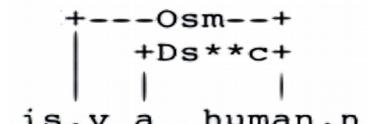
Obtain *leftList*, the list of rules corresponding with *left* (i.e. the rule when *left* is a verb, the rule when *left* is a gerund, etc.), from the global Dictionary variables *dict* and *hyphenated*

Obtain *rightList* in a similar manner

```
for leftRule in leftList do
    for rightRule in rightList do
        Split leftRule and rightRule into lists of Disjuncts ld and rd
        for l in ld do
            for r in rd do
                Replace all instances of ‘-’ in l with ‘+’ and vice versa
                if l = r then
                    | return true
                else
                    | continue
                end
            end
        end
    end
return false
```

---

`connects()` is not always applicable. For instance, when the determiner “a” is present in the phrase “is a human,” the links are not “is” → “a” and “a” → “human” but rather “is” → “human” and “a” → “human” as shown in this Link Grammar parse:



<https://arxiv.org/abs/2105.00830> V. Ramesh, A. Kolonin, 2021

# Grammatical Language Generation - Results

Our algorithm was primarily tested on 92 sentences with words all part of SingularityNET's "small world" POC-English corpus. For this purpose, we have used the Link Grammar dictionary (automatically inferred from high quality Link Grammar parses created by SingularityNET's ULL pipeline) containing 42 total words and 5 total word clusters.

When tested on the same 92 sentences while using the complete Link Grammar dictionary for English, the algorithm achieved the following results. The decrease in "Single correct generated sentence" and increase in "Multiple sentences with one correct" is a direct result of the increased grammatical and semantic ambiguity from using Link Grammar instead of "small world" grammar. Since the "small world" grammar was created from the "small world" corpus itself, each of the words in the corpus contains only a subset of the grammatical or semantic contexts that Link Grammar does.

Our NLG architecture was also tested on 54 sentences part of Charles Keller's production of Lucy Maud Montgomery's "Anne's House of Dreams" as found in the Gutenberg Children corpus and performed as follows

- POC-English – Proof-of-Concept corpus made of artificially selected sentences on limited number of topics ("small world").
- Gutenberg Children (GC) - compendium of books for children contained within Project Gutenberg (<https://www.gutenberg.org>), following the selection used for the Children's Book Test of the Babi CBT corpus <https://research.fb.com/down-loads/babi/>

<https://github.com/aigents/aigents-java-nlp>

<https://arxiv.org/abs/2105.00830> V. Ramesh, A. Kolonin, 2021

Metric	Result
Single correct generated sentence	62/92
Multiple sentences with one correct	30/92
Multiple sentences with none correct	0/92
No generated sentences	0/92
Too many results*	0/92
Accuracy	1.000
Total runtime	18 min, 46 sec
Average runtime per sentence	0 min, 12 sec

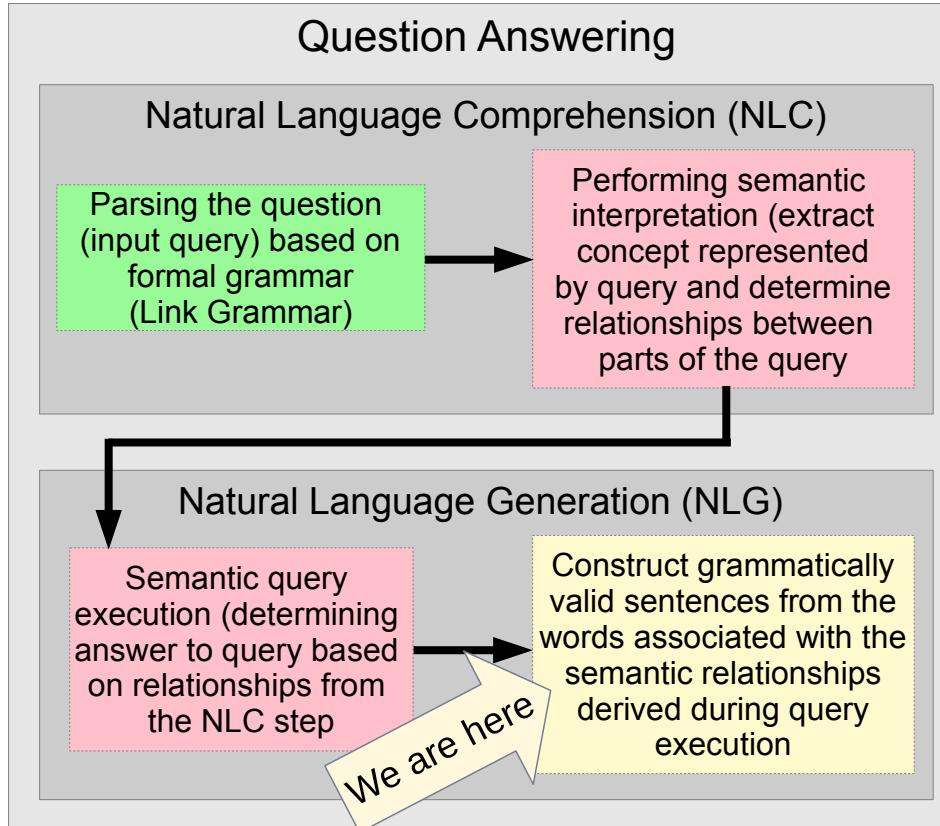
Metric	Result
Single correct generated sentence	8/92
Multiple sentences with one correct	57/92
Multiple sentences with none correct	0/92
No generated sentences	0/92
Too many results*	27/92
Accuracy	0.707
Total runtime	115 min, 6 sec
Average runtime per sentence	1 min, 15 sec

\* "Too many results" is defined as over 25 generated sentences.

Metric	Result
Single correct generated sentence	1/54
Multiple sentences with one correct	53/54
Multiple sentences with none correct	0/54
No generated sentences	0/54
Accuracy	1.000
Total runtime	141 min, 51 sec
Average runtime per sentence	2 min, 37 sec

# Grammatical Generation for Question Answering Applications

## Limitations



**Grammatical ambiguity:** same word may have different roles in a sentence

"I saw the saw."

First "saw" – verb, second "saw" – noun (different sets of grammar rules for each instance of "saw" – semantic/word sense disambiguation)

**Subject-object ambiguity:** a specific case of grammatical ambiguity which refers to the potential interchangeability of the subject and object in a sentence

[ "mouse", "a", "the", "caught", "cat" ]

Result 1: "The cat caught a mouse."

Result 2: "The mouse caught a cat."

Both results are grammatically valid, but "The cat caught a mouse" is more contextually valid. Implementing grammatical and semantic disambiguation to solve these issues will be a product of our future work, along with extending the algorithm's generation capabilities to languages other than English (including those that require heavy morphology usage, such as Russian).

# Question Answering with Link Grammar

## Context: Identity and Relationships

A **mom** is a **human**. A **dad is a human**. A **mom** is a parent. A **dad** is a parent. A son is a child. A **daughter** is a child. A son is a **human**. A **daughter** is a **human**. **Mom** is a **human** now. **Dad** is a **human** now. **Mom** is a parent now. **Dad** is a parent now. Son is a child now. **Daughter** is a child now. Son is a **human** now. **Daughter** is a **human** now. **Mom was a daughter before.** **Dad** was a son before. **Mom** was not a parent before. **Dad** was not a parent before.

**Question:** mom daughter

**Answer:** [Mom was a daughter before., Daughter was a mom before.]

**Question:** dad human

**Answer:** [Dad is a human.]

**Table 1.** Results when tested on 60 queries from SingularityNET's POC-English corpus.

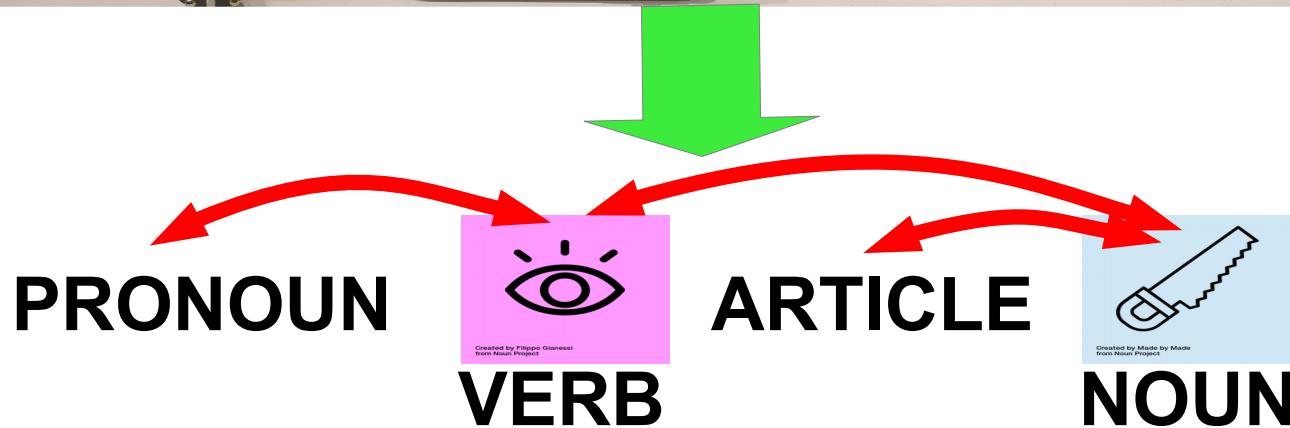
Metric	Ours	BERT	ELECTRA	DistilBERT	RoBERTa
BLEU	<b>0.878</b>	0.639	0.712	0.604	0.767
WVCS	<b>0.944</b>	0.606	0.741	0.595	0.799
WER	0.645	0.924	0.550	1.095	<b>0.150</b>
TER	<b>0.166</b>	0.381	0.342	0.457	0.245

<https://www.youtube.com/watch?v=MKIOqO9FRq0>

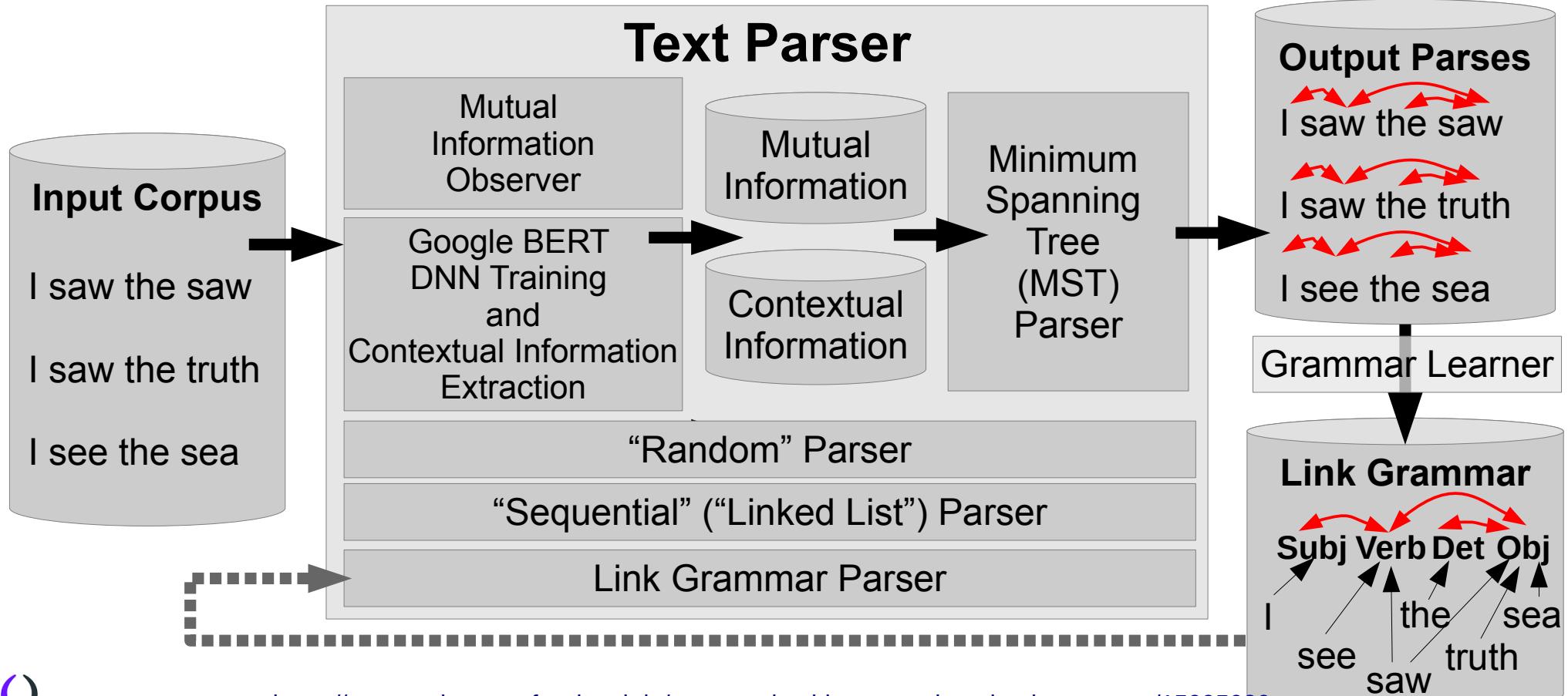
[https://link.springer.com/chapter/10.1007/978-3-030-93758-4\\_22](https://link.springer.com/chapter/10.1007/978-3-030-93758-4_22) V. Ramesh, A. Kolonin, 2021

# Fundamental studies

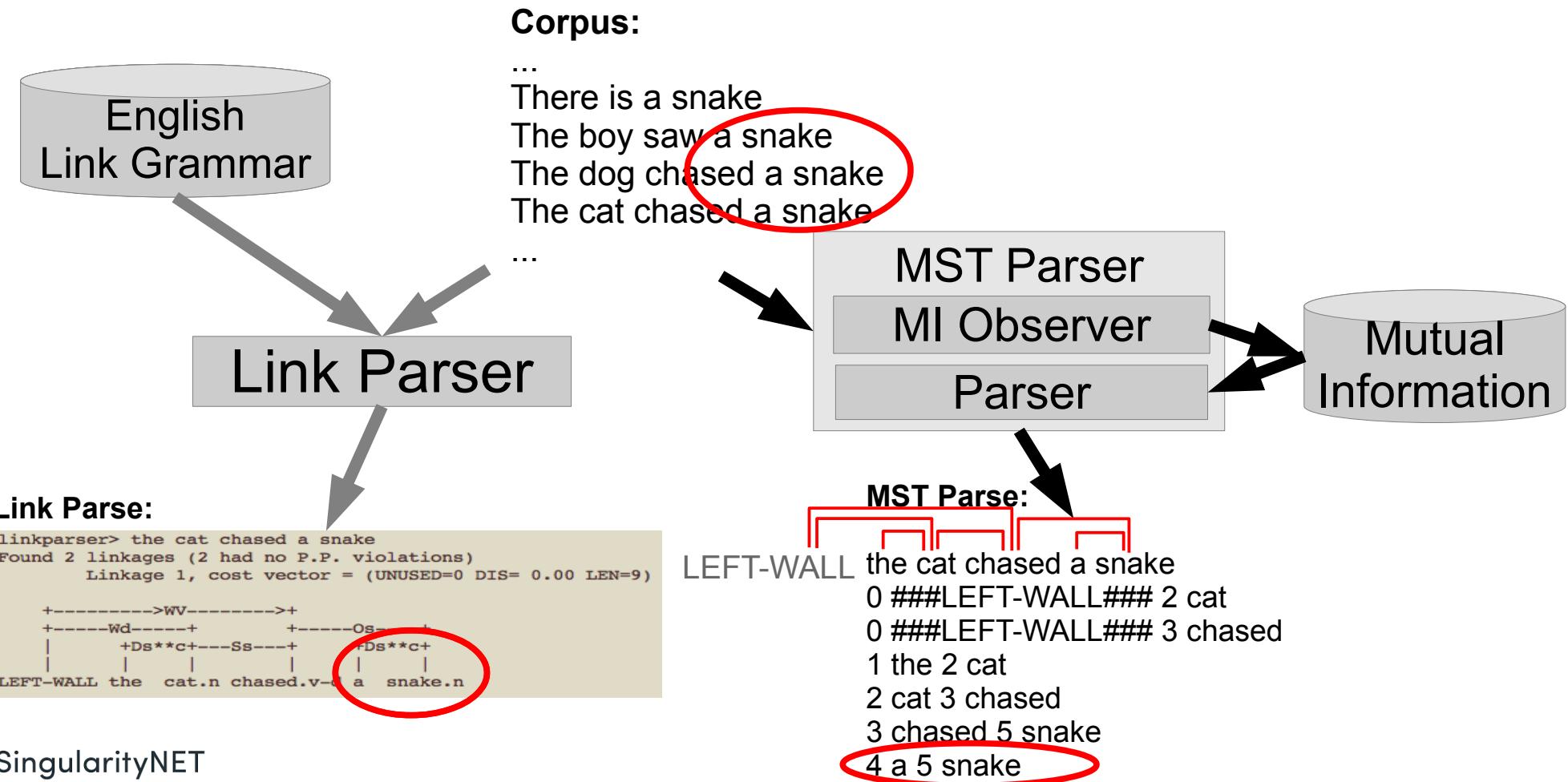
# Unsupervised Link Grammar Learning



# Text Parsing for Link Grammar



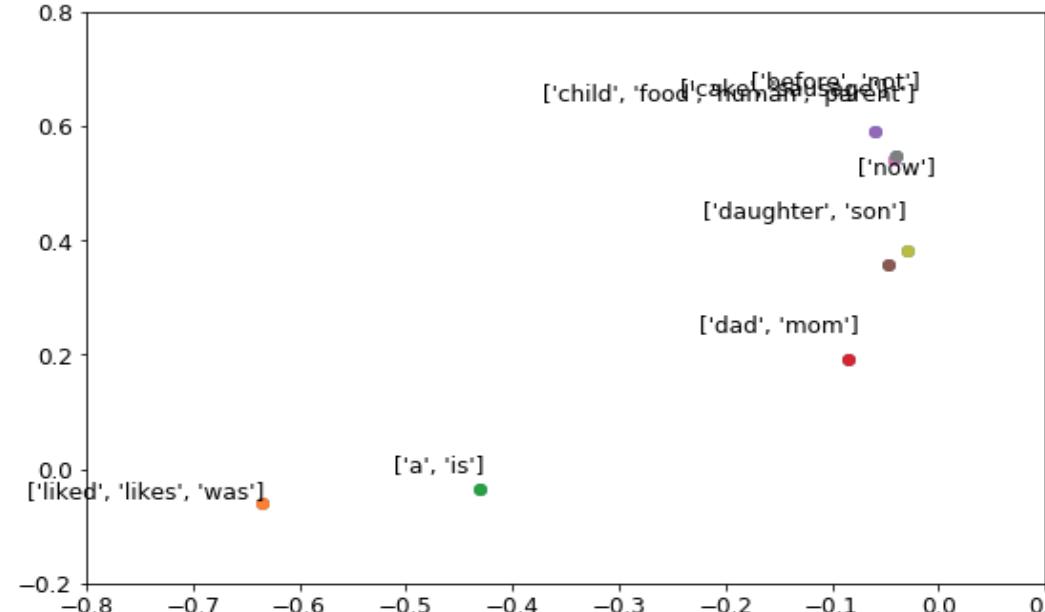
# MST Parses vs. Link Parses



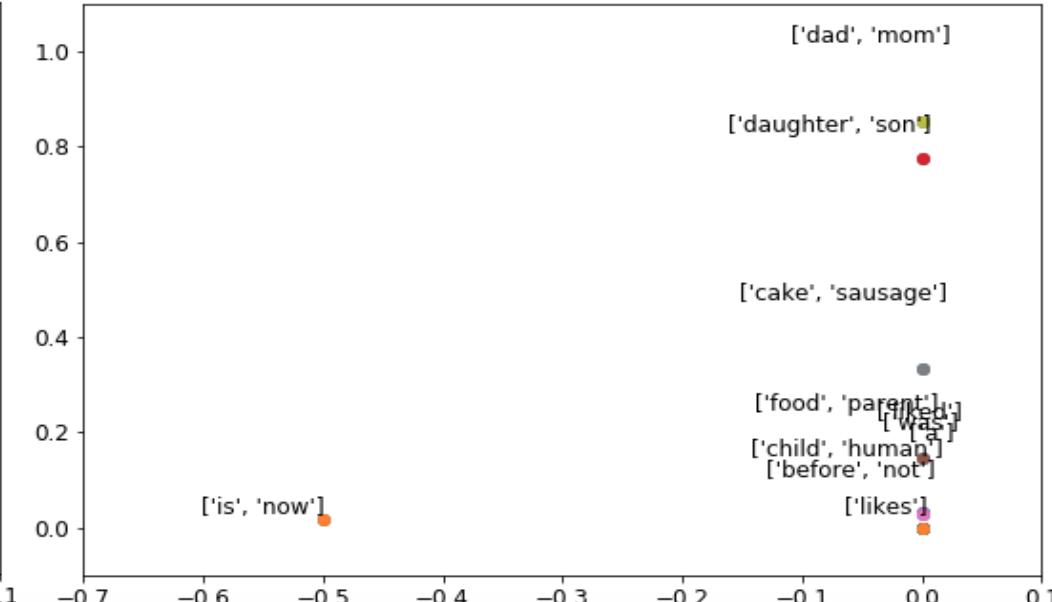
SingularityNET

# Learning Grammatical and Semantic Categories in Vector Spaces of Connectors and Disjuncts

POC-English  
(Connectors)



POC-English  
(Disjuncts)



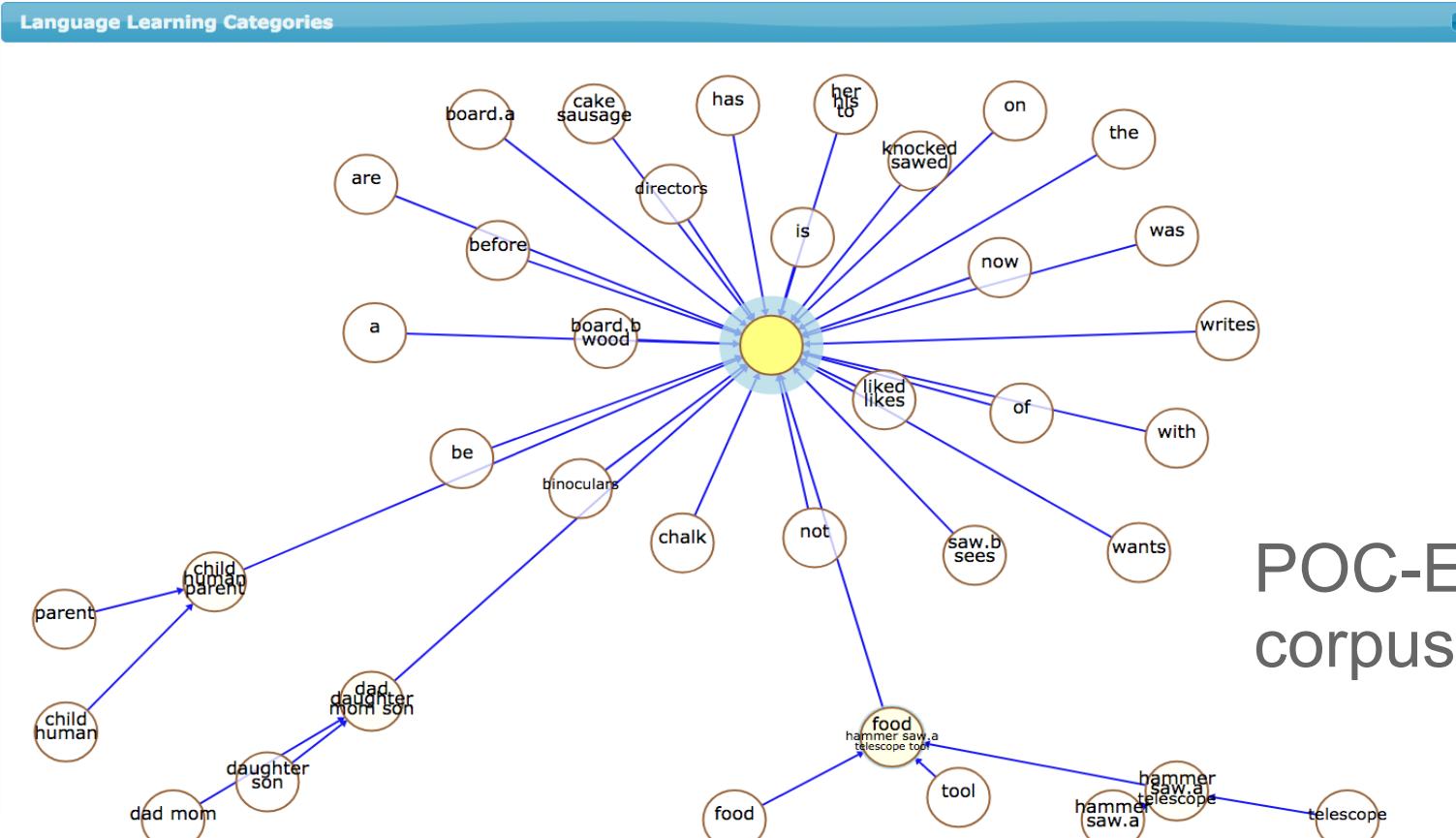
SingularityNET

<https://github.com/singnet/language-learning>

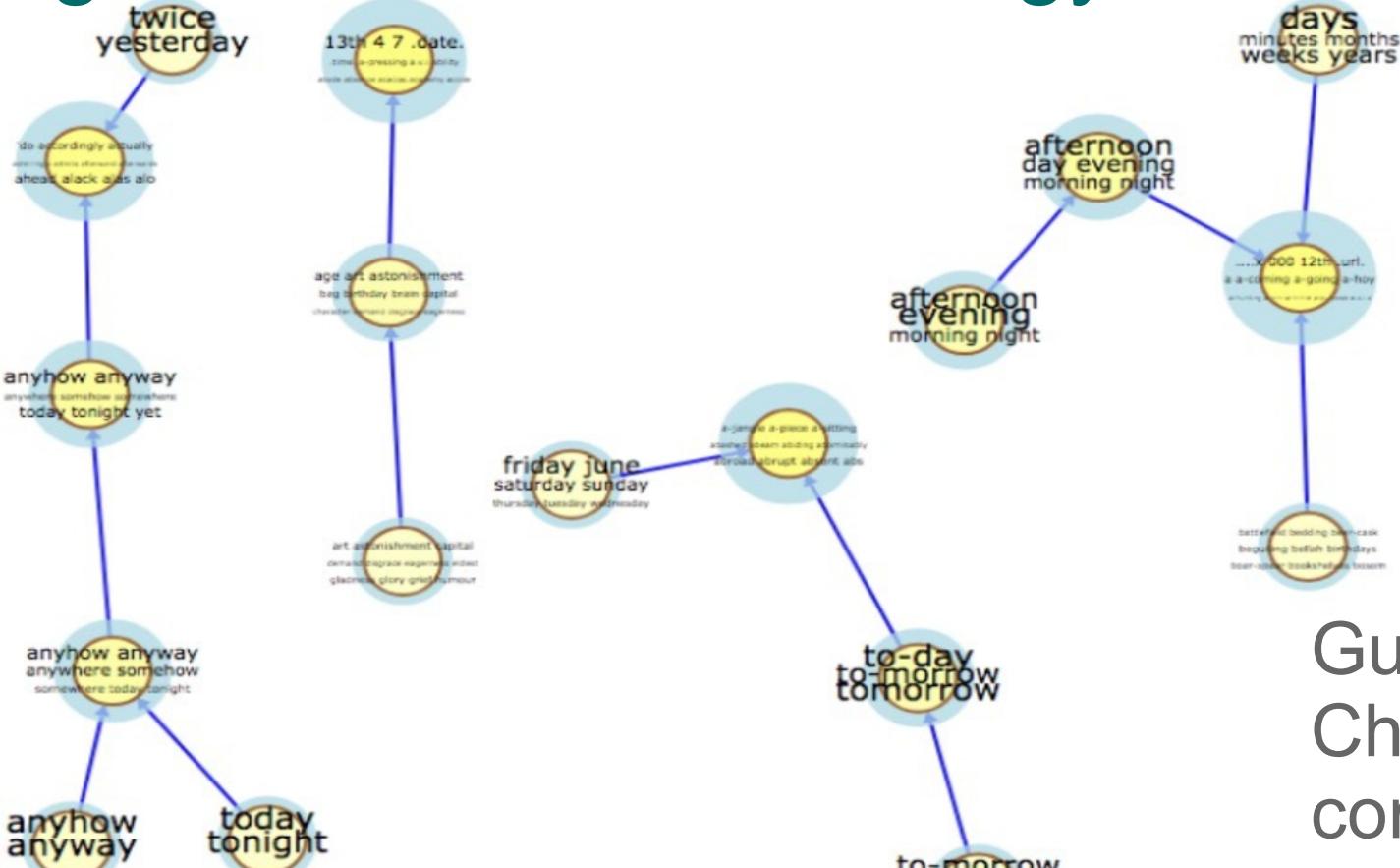
<https://www.youtube.com/watch?v=cwgtcOfA3KI>

<https://www.springerprofessional.de/en/programmatic-link-grammar-induction-for-unsupervised-language-le/17020348>

# Learning Grammar and Ontology from Parses



# Learning Grammar and Ontology from Parses



Gutenberg  
Children  
corpus



SingularityNET

<https://www.springerprofessional.de/unsupervised-language-learning-in-opencog/15995030>

<https://www.springerprofessional.de/en/programmatic-link-grammar-induction-for-unsupervised-language-le/17020348>

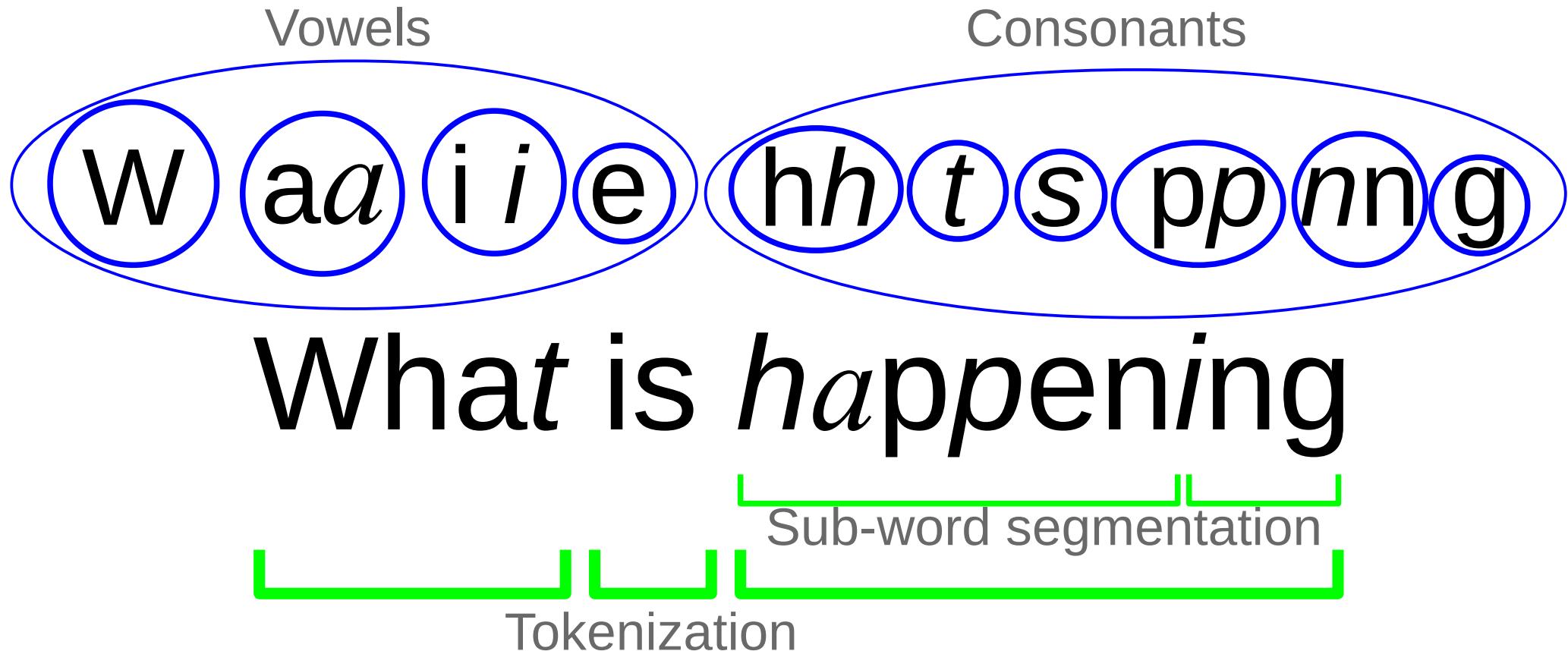
# Grammar Learning F1 Across the Corpora

Corpus	Parses	Parses F1	Clustering	Parse-Ability	Grammar F1
POC-English	Manual	1.00	ILE	100%	1.00
POC-English	Manual	1.00	ALE-400	100%	1.00
POC-English	MST	0.71	ILE	100%	0.72
POC-English	MST	0.71	ALE-400	100%	0.73
Child-Directed Speech	LG-English	1.00	ILE	99%	0.98
Child-Directed Speech	LG-English	1.00	ALE-400	99%	0.97
Child-Directed Speech	MST	0.68	ILE	71%	0.45
Child-Directed Speech	MST	0.68	ALE-400	82%	0.50
Gutenberg Children	LG-English	1.00	ILE	63%	0.65
Gutenberg Children	LG-English	1.00	ALE-500	69%	0.66
Gutenberg Children	MST	0.52	ILE	93%	0.50
Gutenberg Children	MST	0.52	ALE-500	99%	0.53

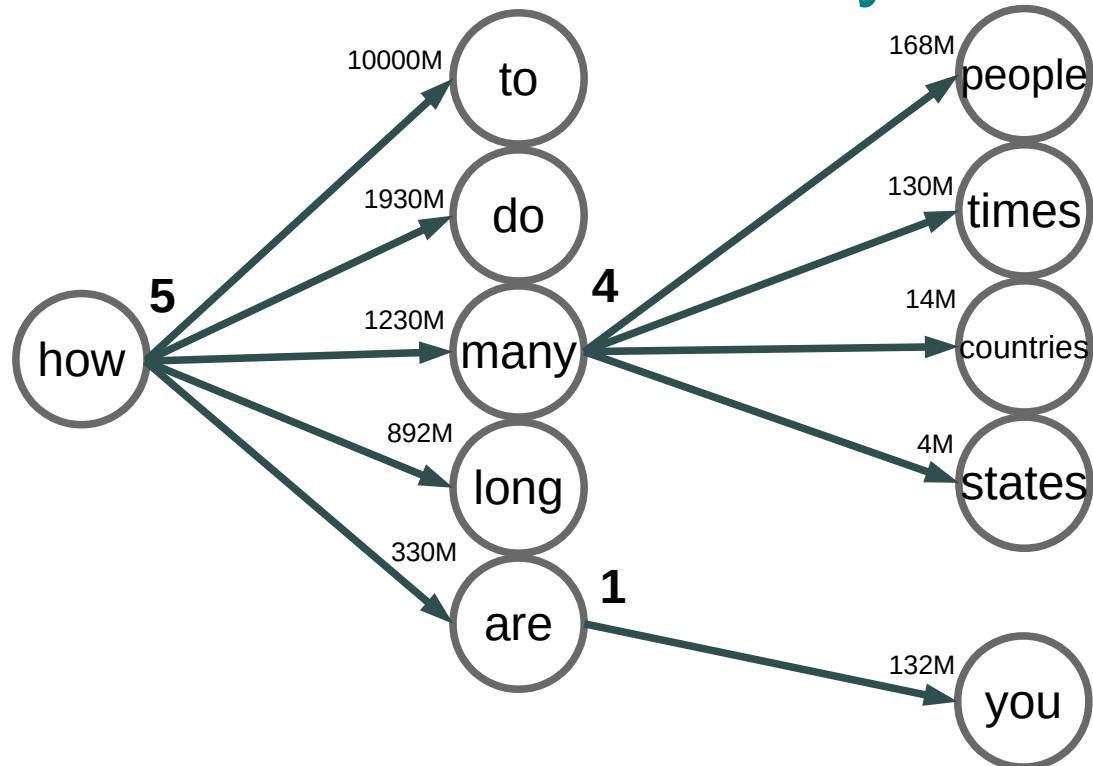


# Learning for Lexicon, Punctuation and Morphology

## Clustering and Segmentation



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



## Metrics/Indicators:

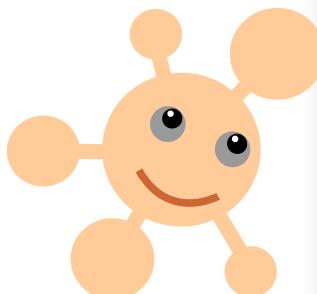
Mutual Information<sup>1</sup>  
Conditional Probability<sup>1,2</sup>  
Transition Freedom<sup>2,3</sup>

<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>

# 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

<https://aclanthology.org/2022.emnlp-main.239/>

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

Copyright © 2023 Anton Kolonin, Aigents®

32

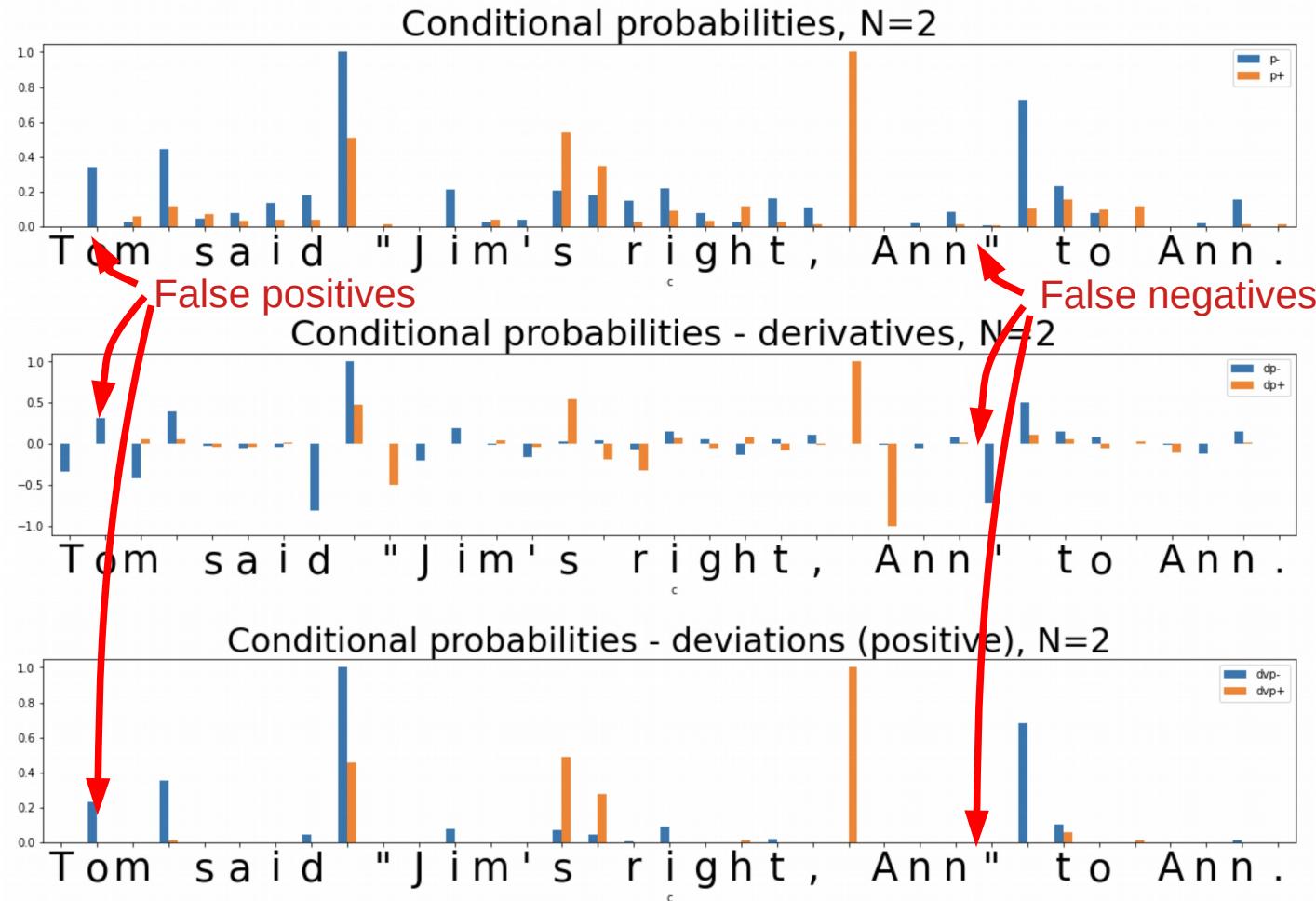
# Unsupervised Text Segmentation (Tokenization)

## Metrics/Indicators:

Ngram (Character)  
Conditional  
Probability  
(of Transition)

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

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

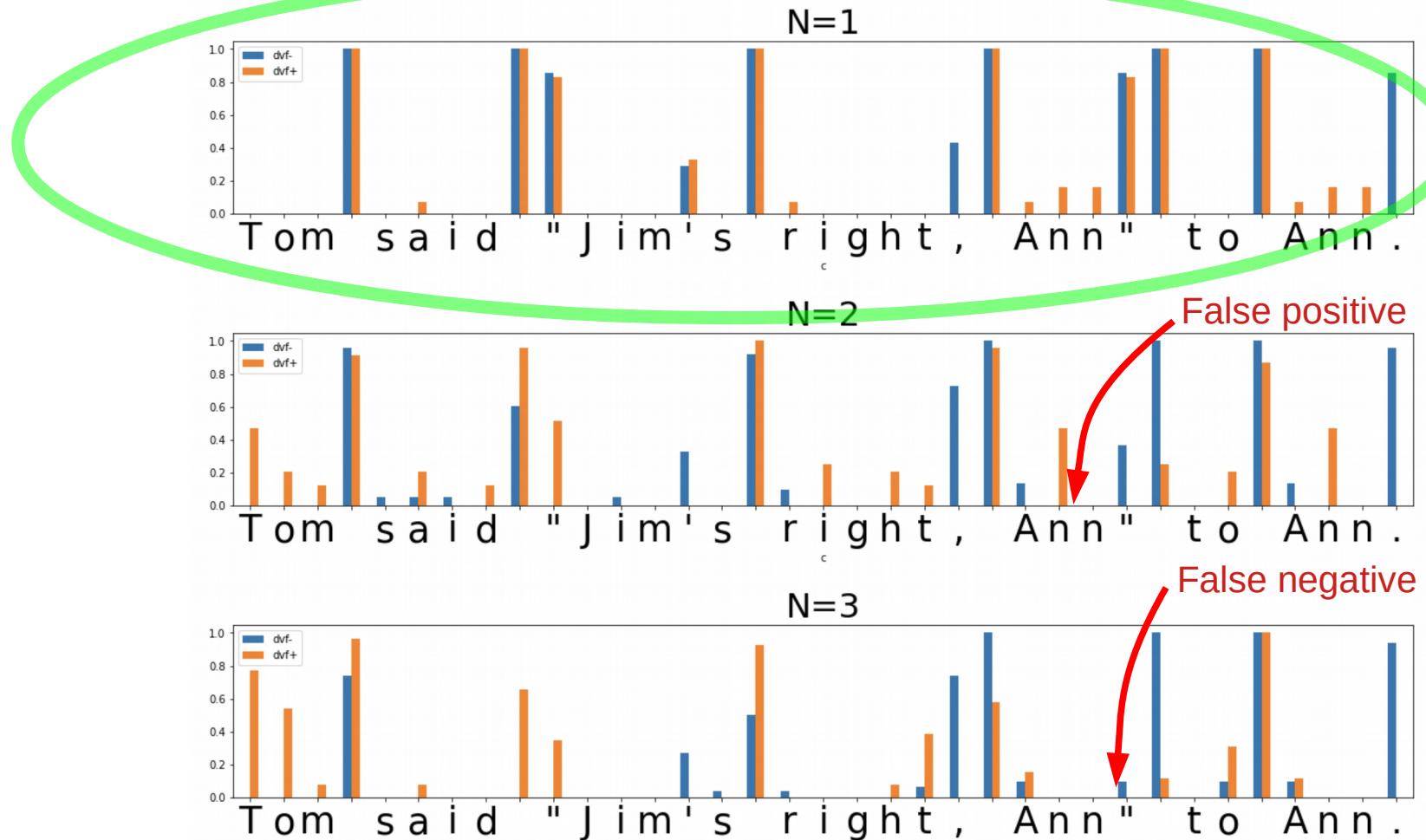


# Unsupervised Text Segmentation (Tokenization)

Metrics/  
Indicators:

Transition  
Freedom  
Deviation

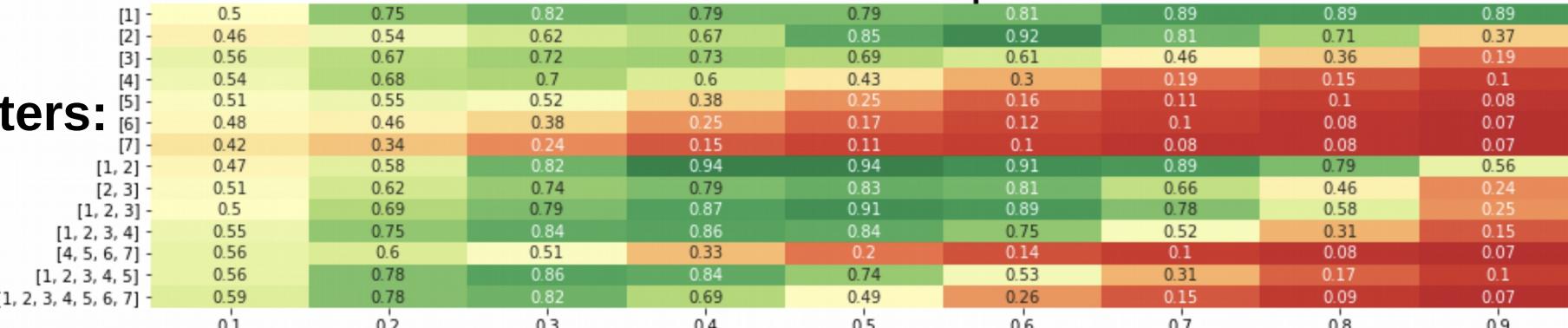
(varying “N”  
of N-gram)



# Unsupervised Text Segmentation (Tokenization)

**English**

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



**Hyper-Parameters:**

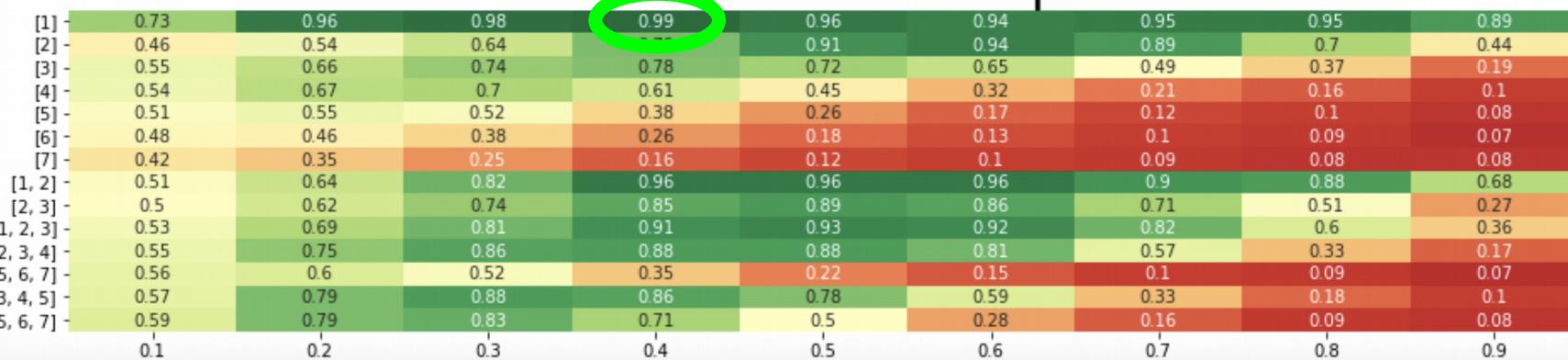
Metric:  
Transition  
Freedom

Threshold  
for model  
compression

Combination  
of Ngram N-s

Threshold for  
segmentation

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



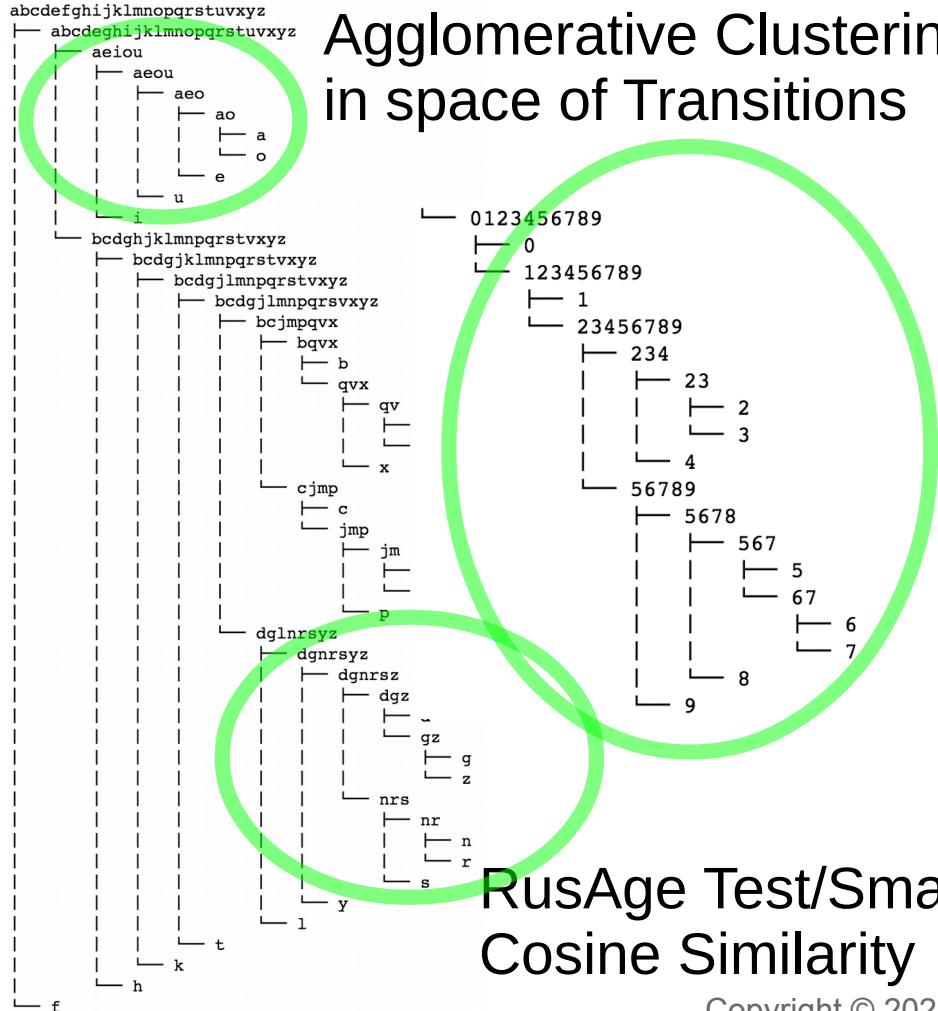
# Results – Freedom-based Tokenization against Lexicon-based one (referring to Rule-based)

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	-

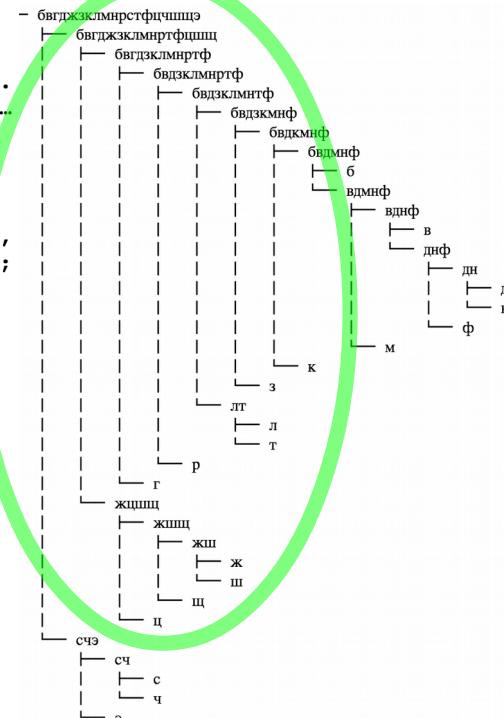
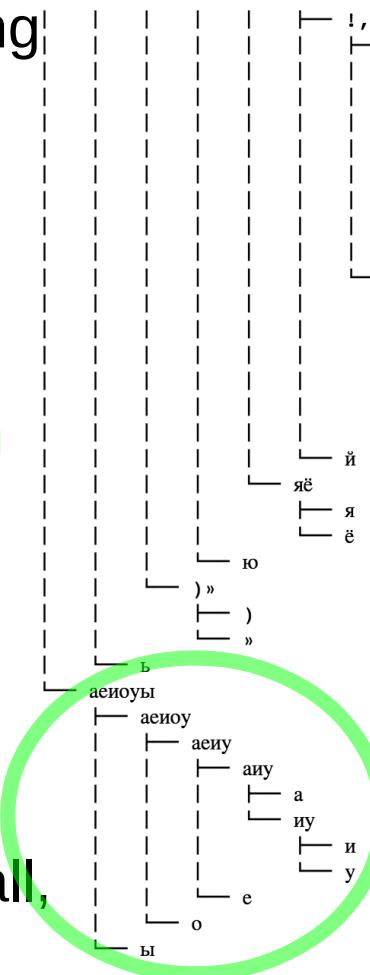
\*Lexicon-based Tokenization - greedy/beam search on word length (optimal) or frequency

# Unsupervised Character Category Learning

# Agglomerative Clustering in space of Transitions



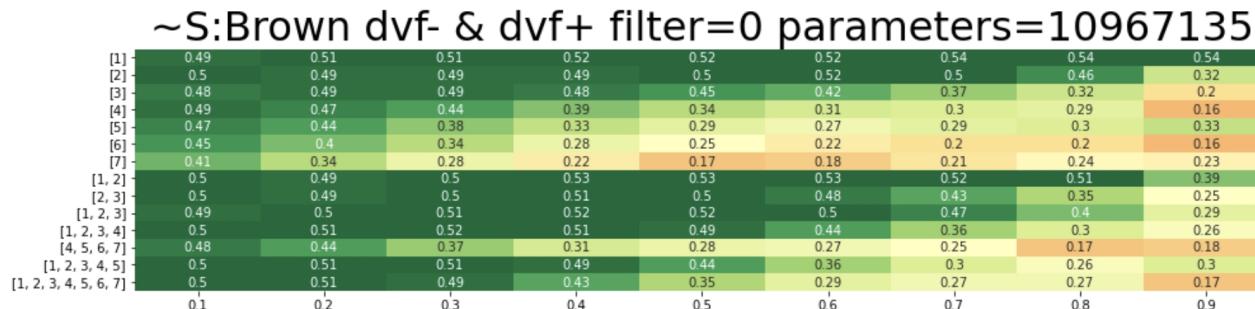
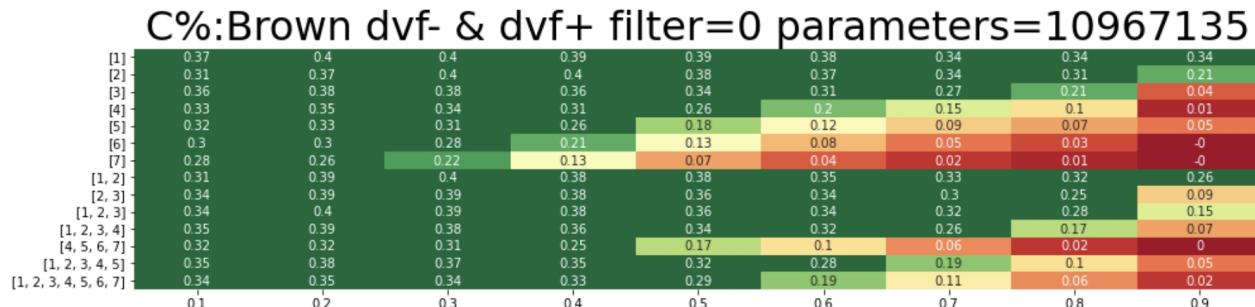
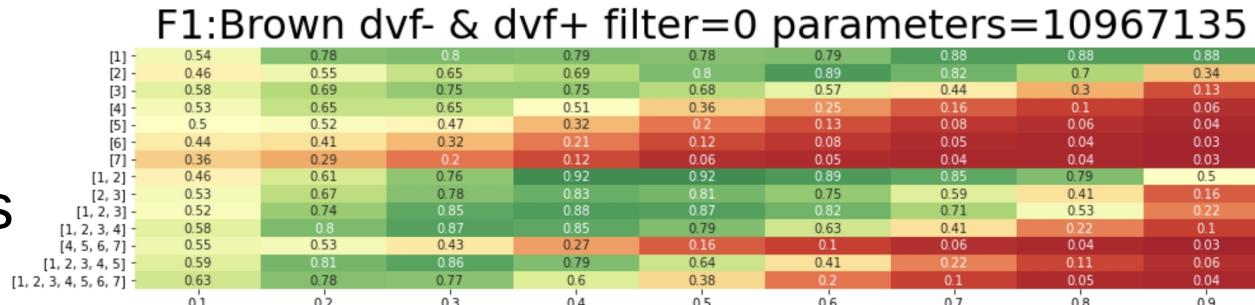
# RusAge Test/Small Cosine Similarity



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

English,  
Brown corpus

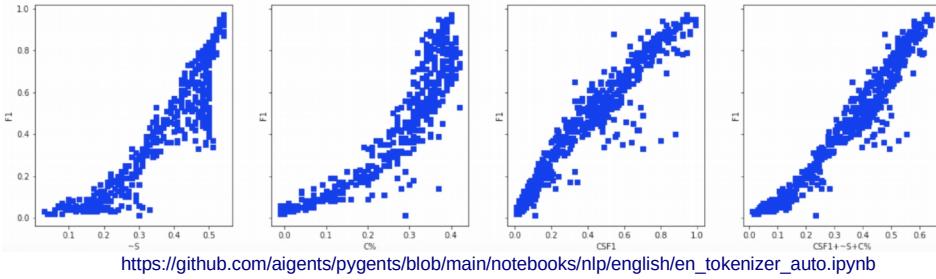
Maximizing F1,  
compression  
factor (C%)  
and  
normalized  
anti-entropy  
(~S) in the  
space of hyper-  
parameters



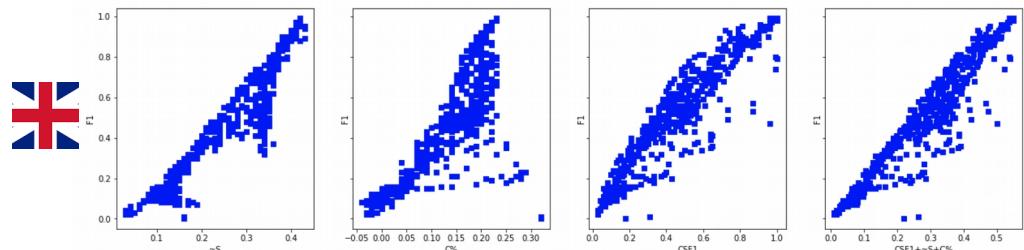
<https://arxiv.org/abs/2303.02427>

# Different corpora, different sizes

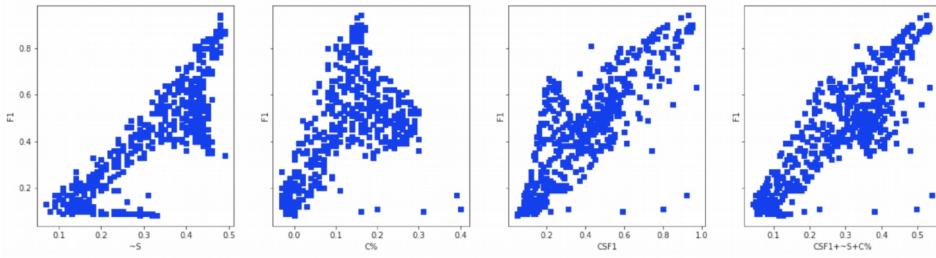
English, Train: Brown, Test: Brown 1000



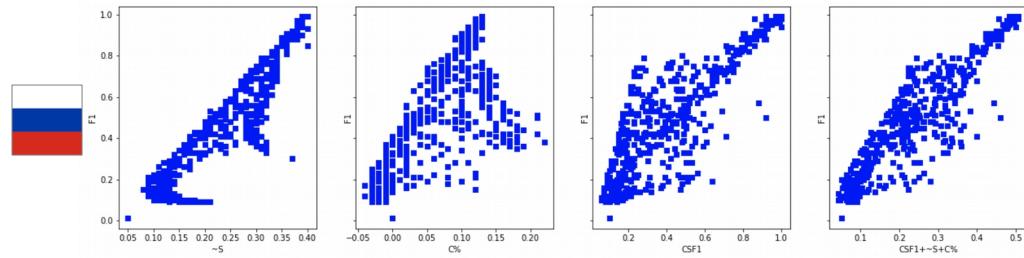
English, Train: Brown, Test: MagicData 100



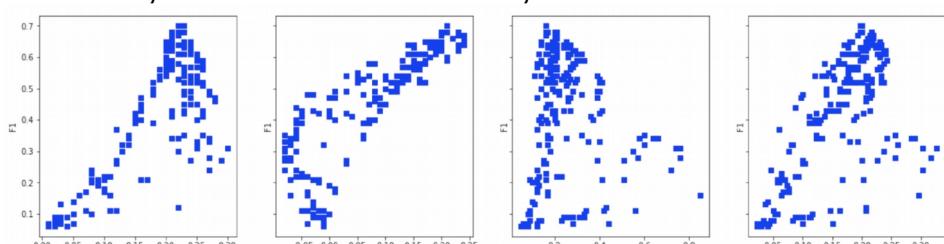
Russian, Train: RusAge, Test: RusAge 1000



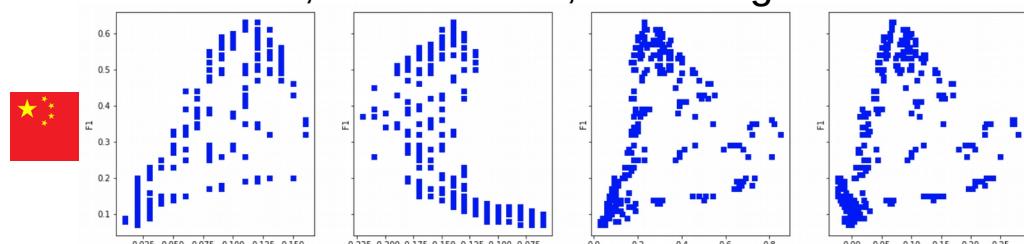
Russian, Train: Brown, Test: MagicData 100



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



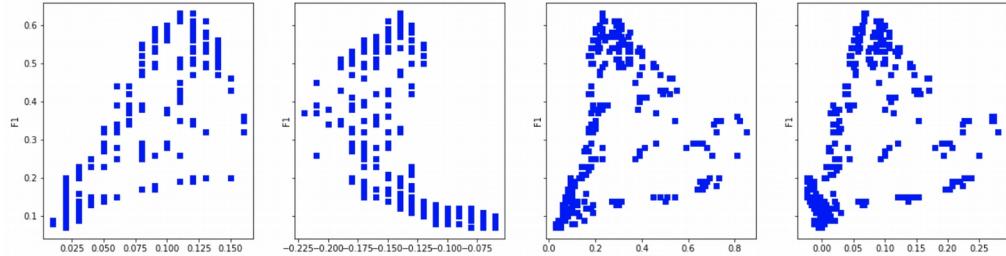
Chinese, Train: Brown, Test: MagicData 100



# Chinese corpora, different sizes

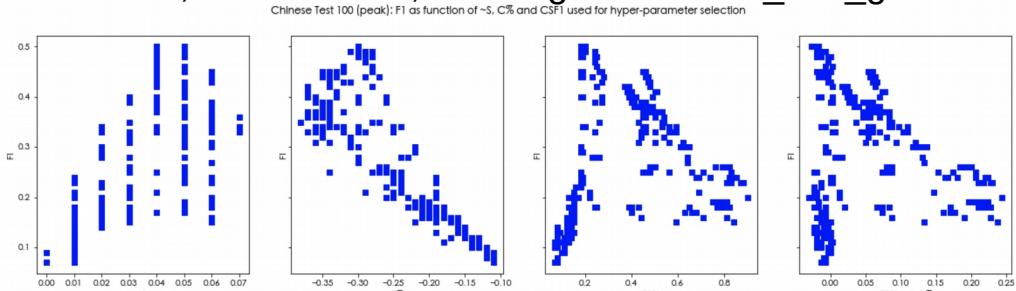


Chinese, Train: Brown, Test: MagicData 100

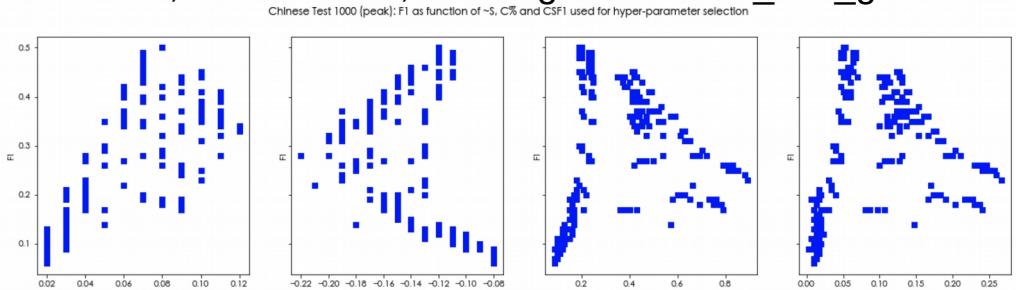


[https://github.com/agents/pygents/blob/main/notebooks/nlp/tokenization/brown/tokenization\\_brown\\_en\\_ru\\_zh.ipynb](https://github.com/agents/pygents/blob/main/notebooks/nlp/tokenization/brown/tokenization_brown_en_ru_zh.ipynb)

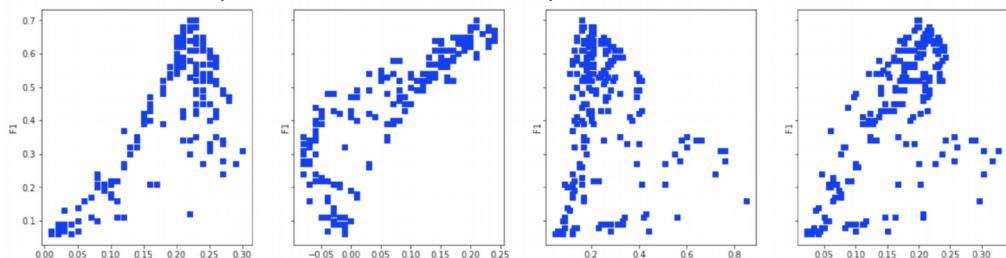
Chinese, Train: Brown, Test: sighan2005/as\_test\_gold 100



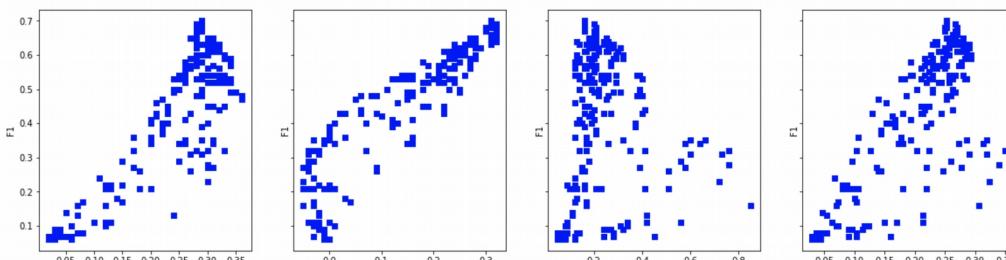
Chinese, Train: Brown, Test: sighan2005/as\_test\_gold 1000



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

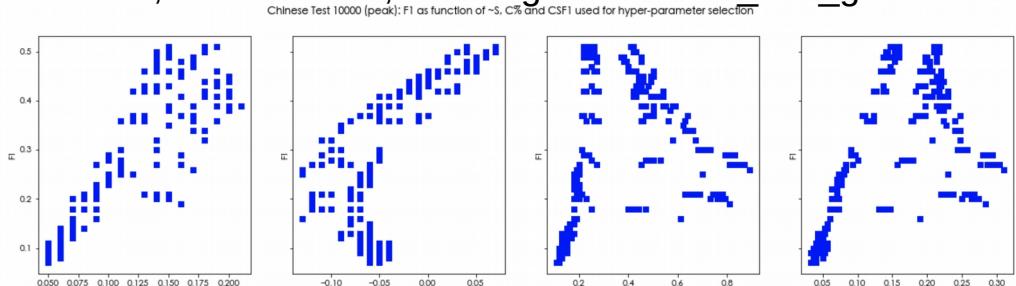


Chinese, Train: CLUE News, Test: CLUE News 10000



[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)

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



[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)

# Chinese corpora, different “ground truth” tokenizers

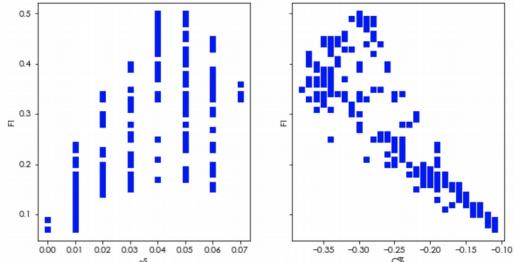


Ground truth: sighan2005

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

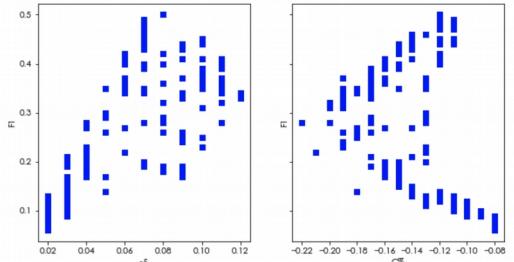
Ground truth: Jieba

Chinese Test 100 (peak): F1 as function of  $\sim S$ , C% and CSF1 used for hyper-parameter selection



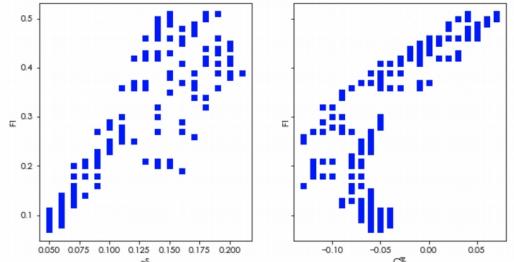
Chinese, Train: Brown, Test: sighan2005/as\_test\_gold 1000

Chinese Test 1000 (peak): F1 as function of  $\sim S$ , C% and CSF1 used for hyper-parameter selection

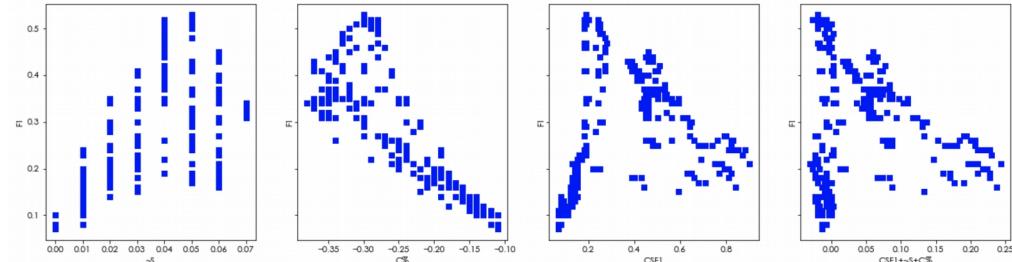


Chinese, Train: Brown, Test: sighan2005/as\_test\_gold 10000

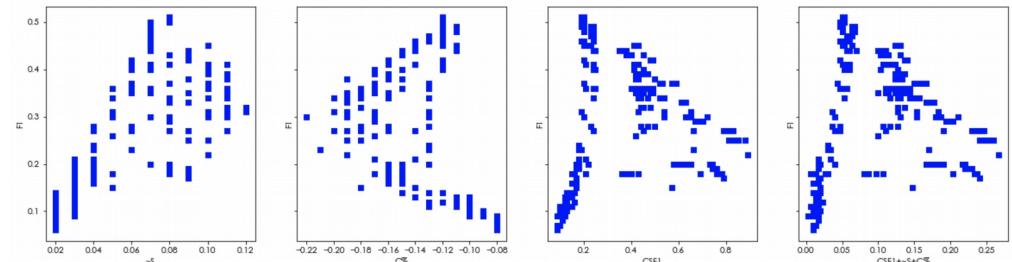
Chinese Test 10000 (peak): F1 as function of  $\sim S$ , C% and CSF1 used for hyper-parameter selection



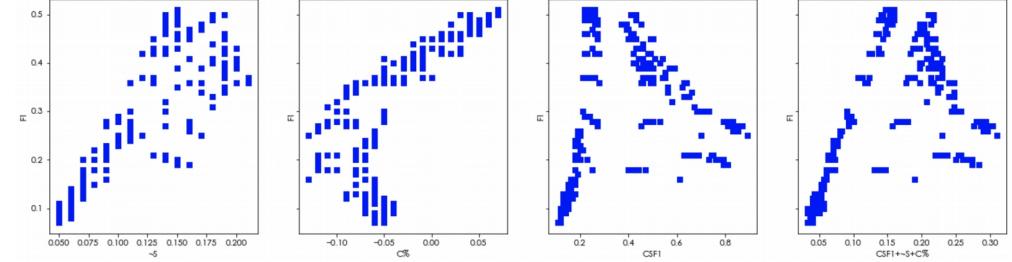
Chinese Test 100 (peak) - sighan2005/as\_test\_gold: F1 as function of  $\sim S$ , C% and CSF1 used for hyper-parameter selection



Chinese Test 1000 (peak) - sighan2005/as\_test\_gold: F1 as function of  $\sim S$ , C% and CSF1 used for hyper-parameter selection



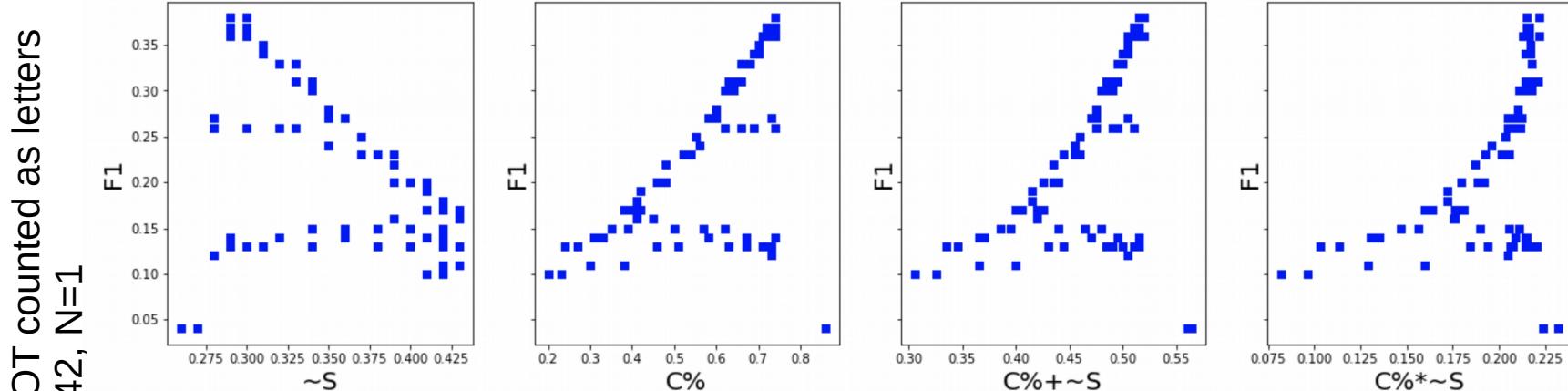
Chinese Test 10000 (peak) - sighan2005/as\_test\_gold: F1 as function of  $\sim S$ , C% and CSF1 used for hyper-parameter selection



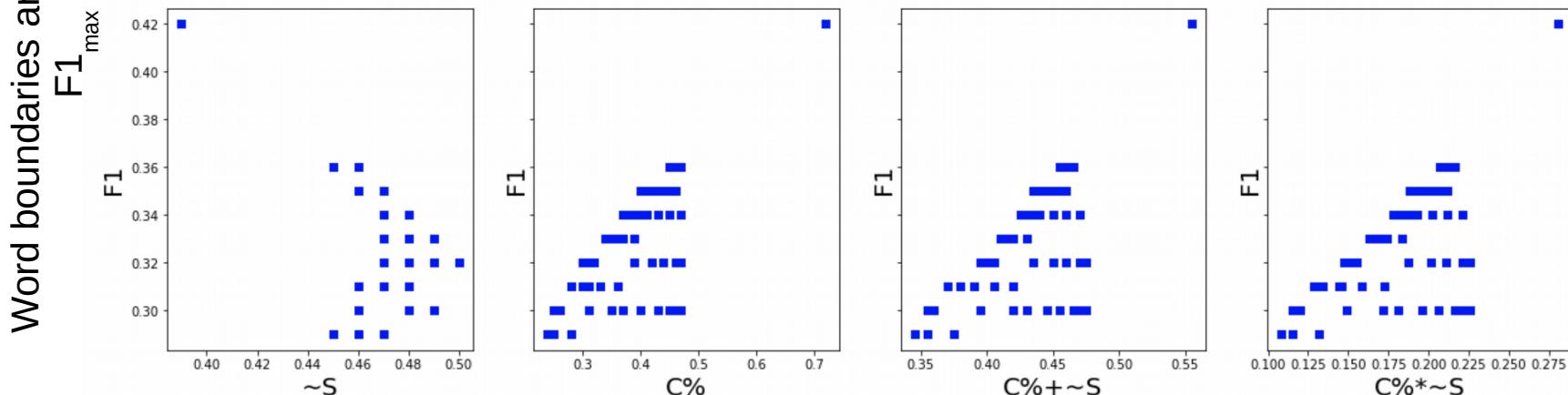
# Unsupervised learning for subword segmentation



F1 as function of  $\sim S$  and C% used for hyper-parameter selection, Nchars > 10



F1 as function of  $\sim S$  and C% used for hyper-parameter selection, Nchars > 0

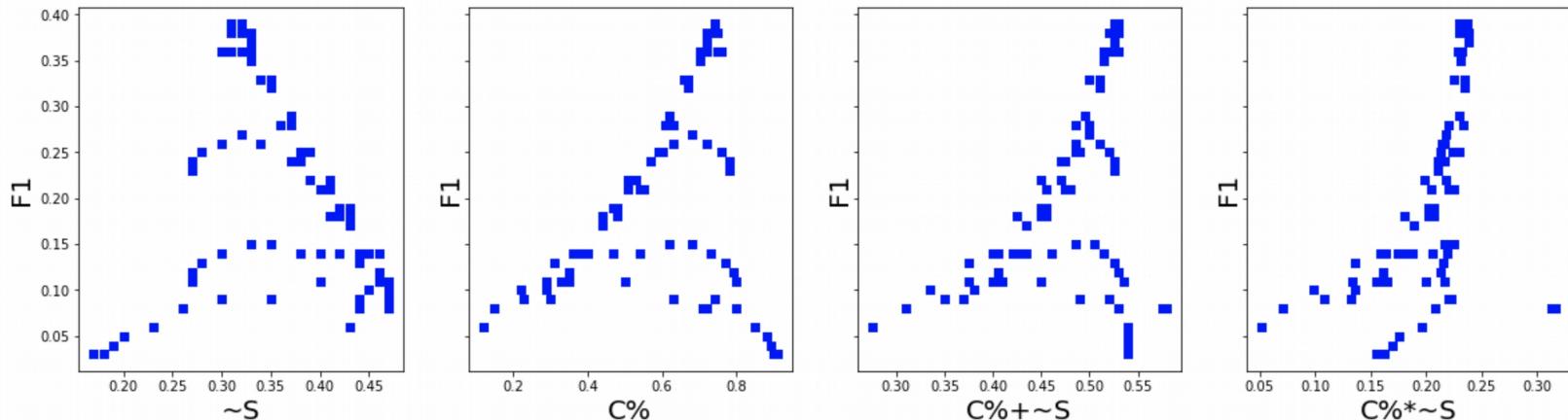


# Unsupervised learning for subword segmentation

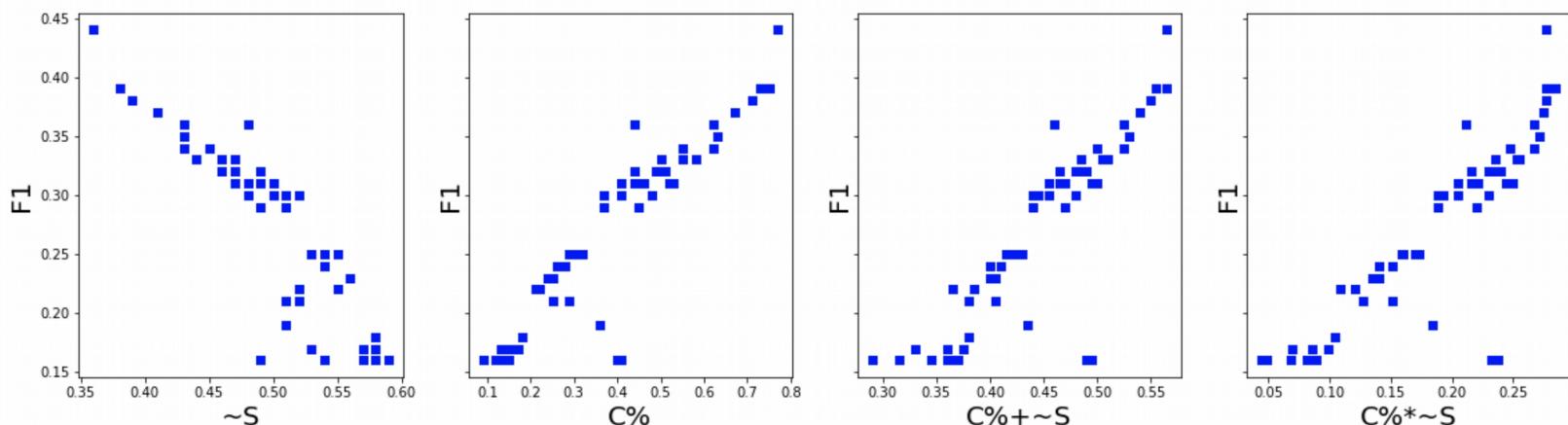


F1 as function of  $\sim S$  and C% used for hyper-parameter selection, Nchars > 10

Word boundaries are NOT counted as letters,  
 $F1_{max} = 0.44, N = 2$



F1 as function of  $\sim S$  and C% used for hyper-parameter selection, Nchars > 0



# 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

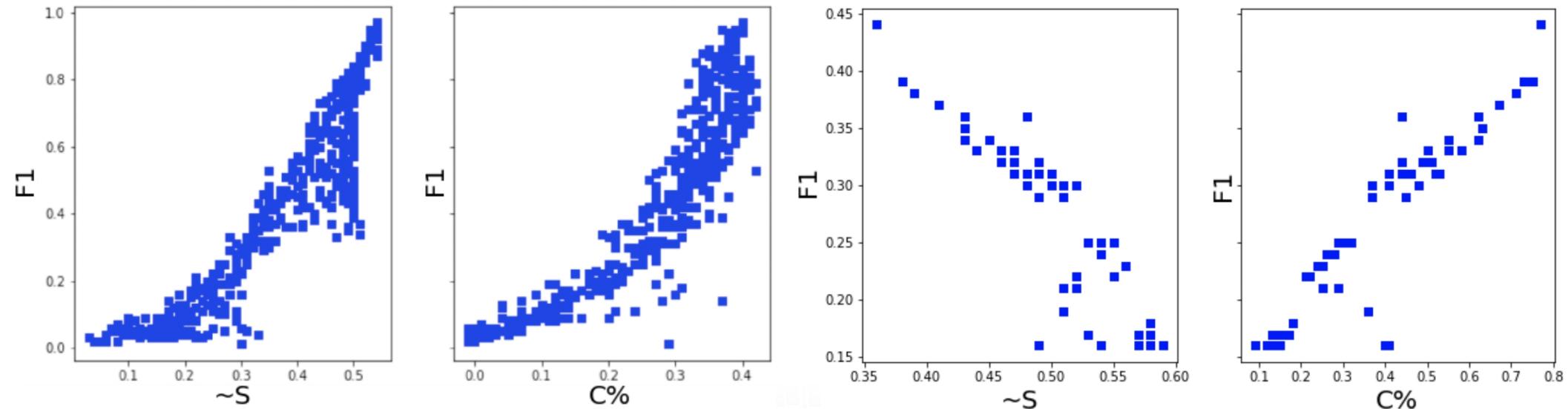
0.25

<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

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



Tokenization (word-level segmentation)

Morpho-parsing (sub-word segmentation)

# The riddle - answer: you have got it!

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

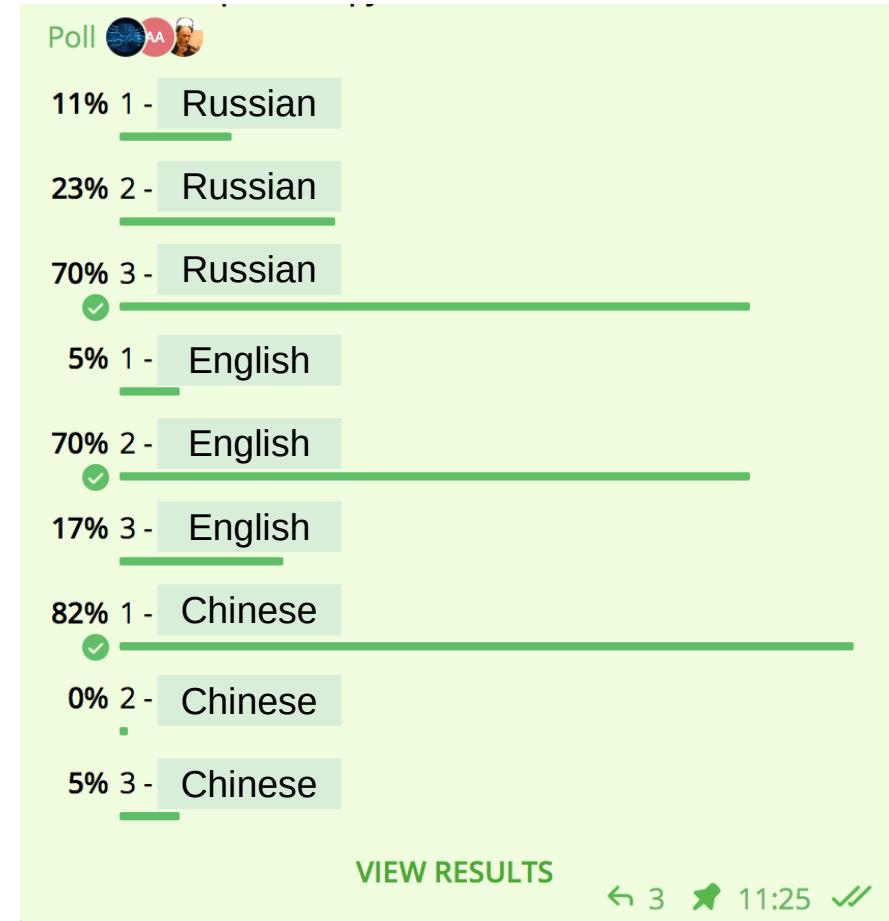
Language 1 11:22 ✓

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 ✓



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

Telegram English  
<https://t.me/internlp>

Telegram Russian  
<https://t.me/agibots>

Workshop 2023 (Stockholm, Sweden & Virtual)  
<https://aigents.github.io/inlp/2023/>

Workshop 2022  
<https://aigents.github.io/inlp/2022/>

# Thank You and Welcome!



<https://agirussia.org>



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

