# Understandable Unsupervised
## Language Learning

Alex Glushchenko, Andres Suarez, Anton Kolonin,
Ben Goertzel, Matt Iklé, Sergey Shalyapin, Oleg Baskov

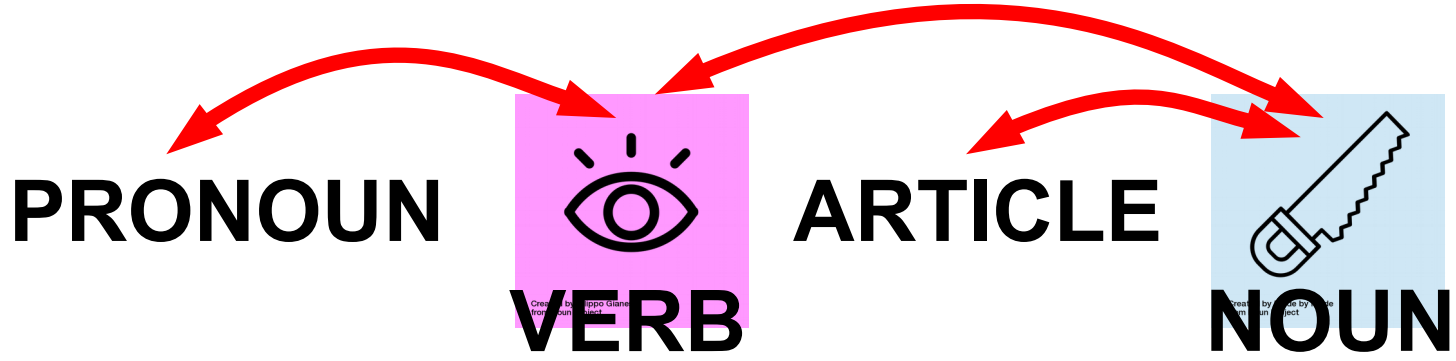
Presenter: Anton Kolonin
akolonin@aigents.com

OpenCog
https://opencog.org/

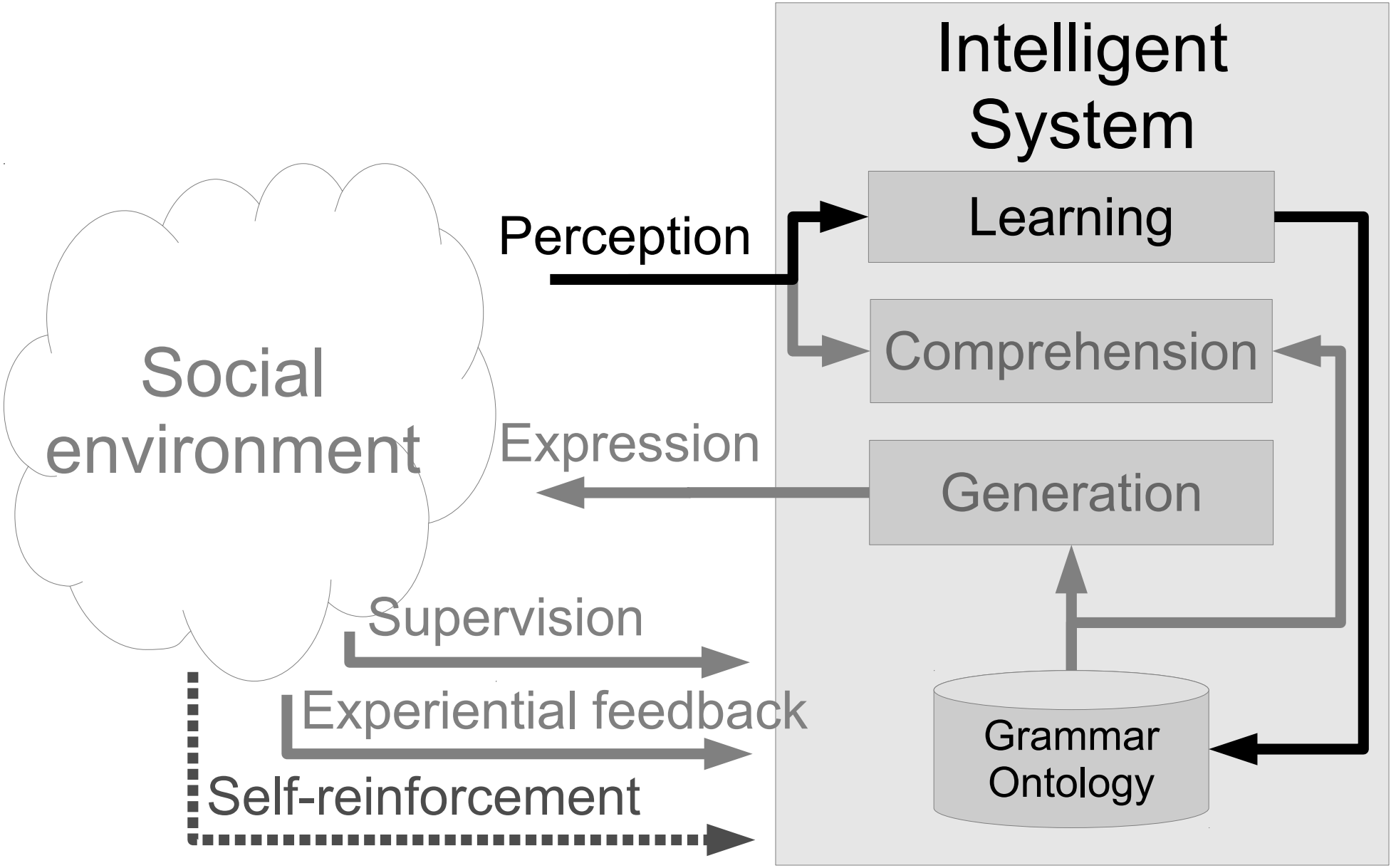
SingularityNET
https://singularitynet.io

http://www.hansonrobotics.com/

# Grammar Learning from Scratch - Programmatically

# Language Learning Environment



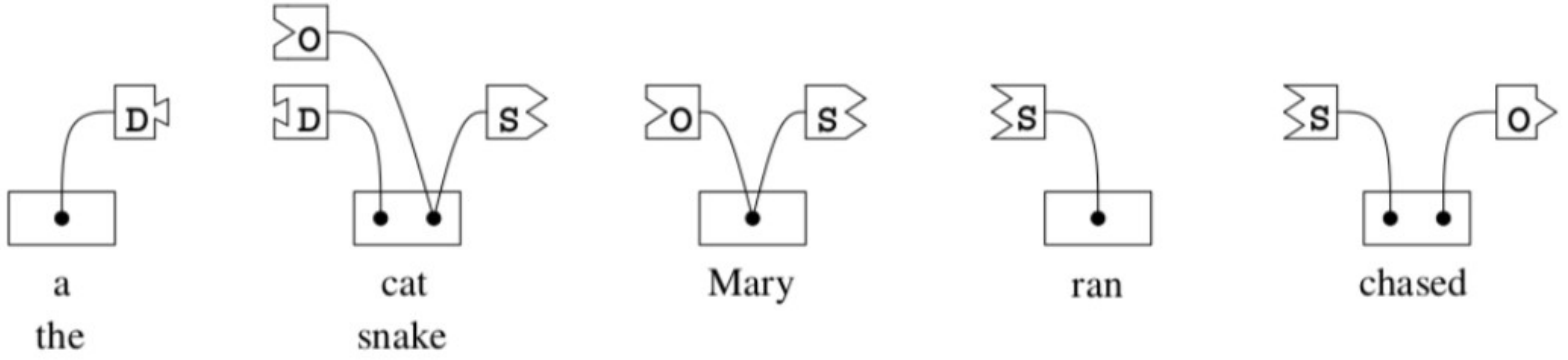
# Project goal and applications

- Grammar learning from scratch - programmatically
- Grammar extension/customization for specific domains
- Building dictionaries and patterns for NLP applications
- Parsing texts for NLP applications
- Grammar checking (more than spell checking)

# Constraints of the currently explored approach

- Controlled corpora
- Using Link Grammar formalism
- Relying on MST parses
- No account for morphology
- Self-reinforcement with F1 on parses
- Test against training data

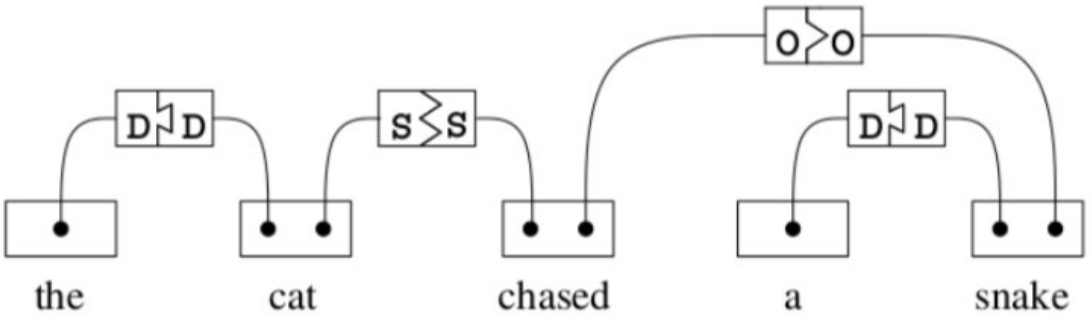
# OpenCog Link Grammar Disjuncts & Connectors




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

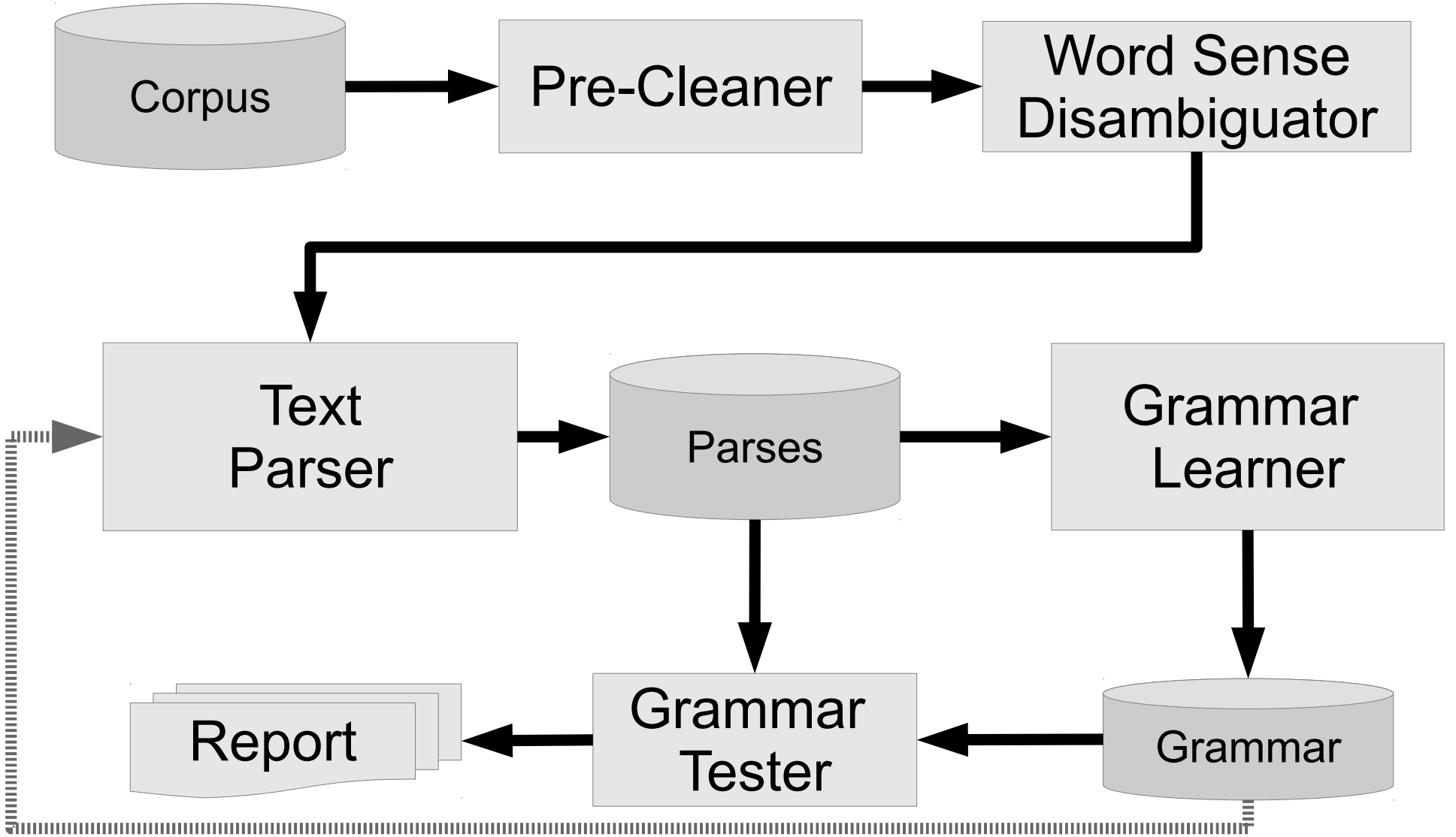
```
a the: D+;
cat snake: D- & (S+ or O-);
Mary: O- or S+;
ran: S-;
chased S- & O+;
```

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:

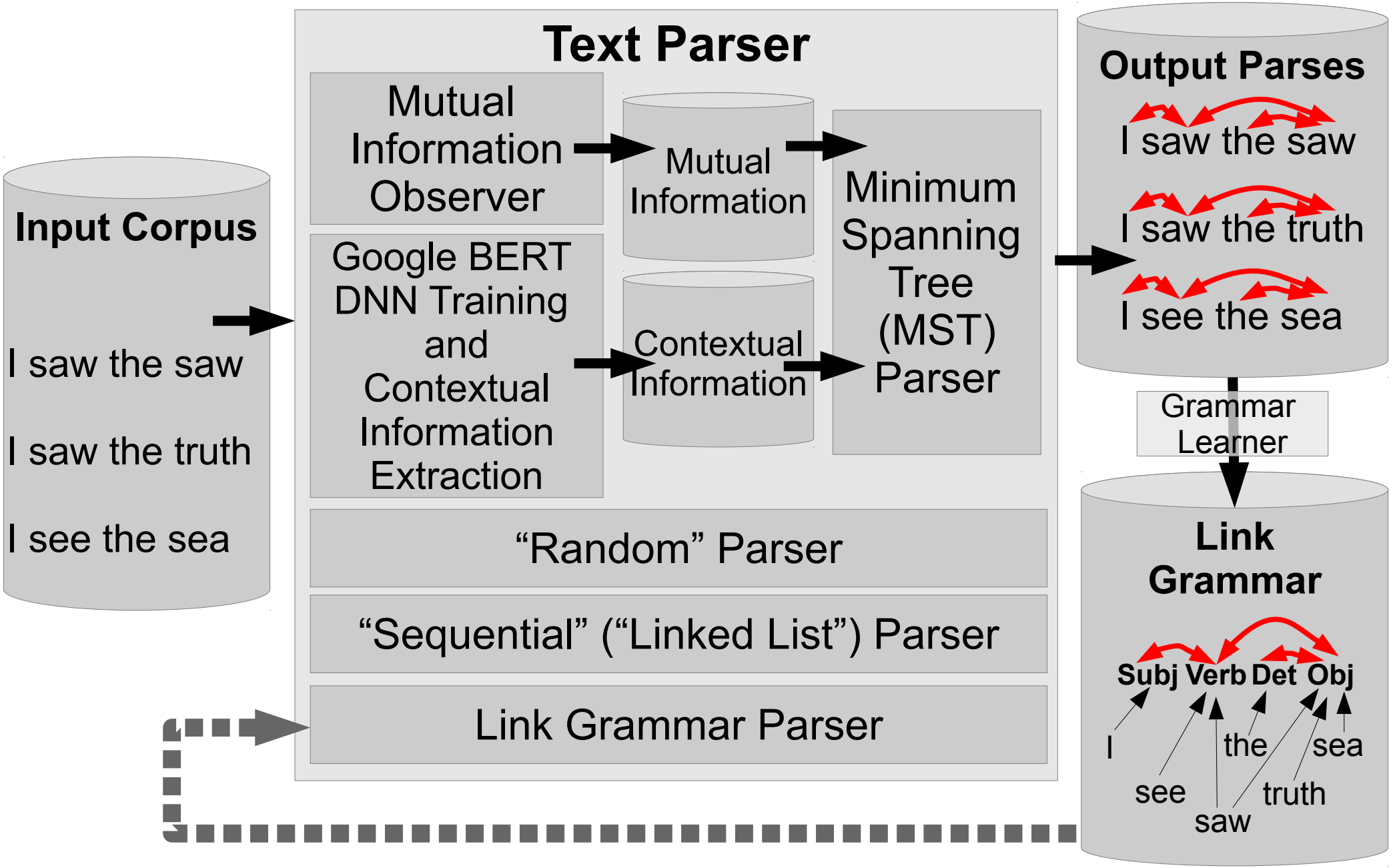



B. Goertzel,
L. Vepstas,
2014

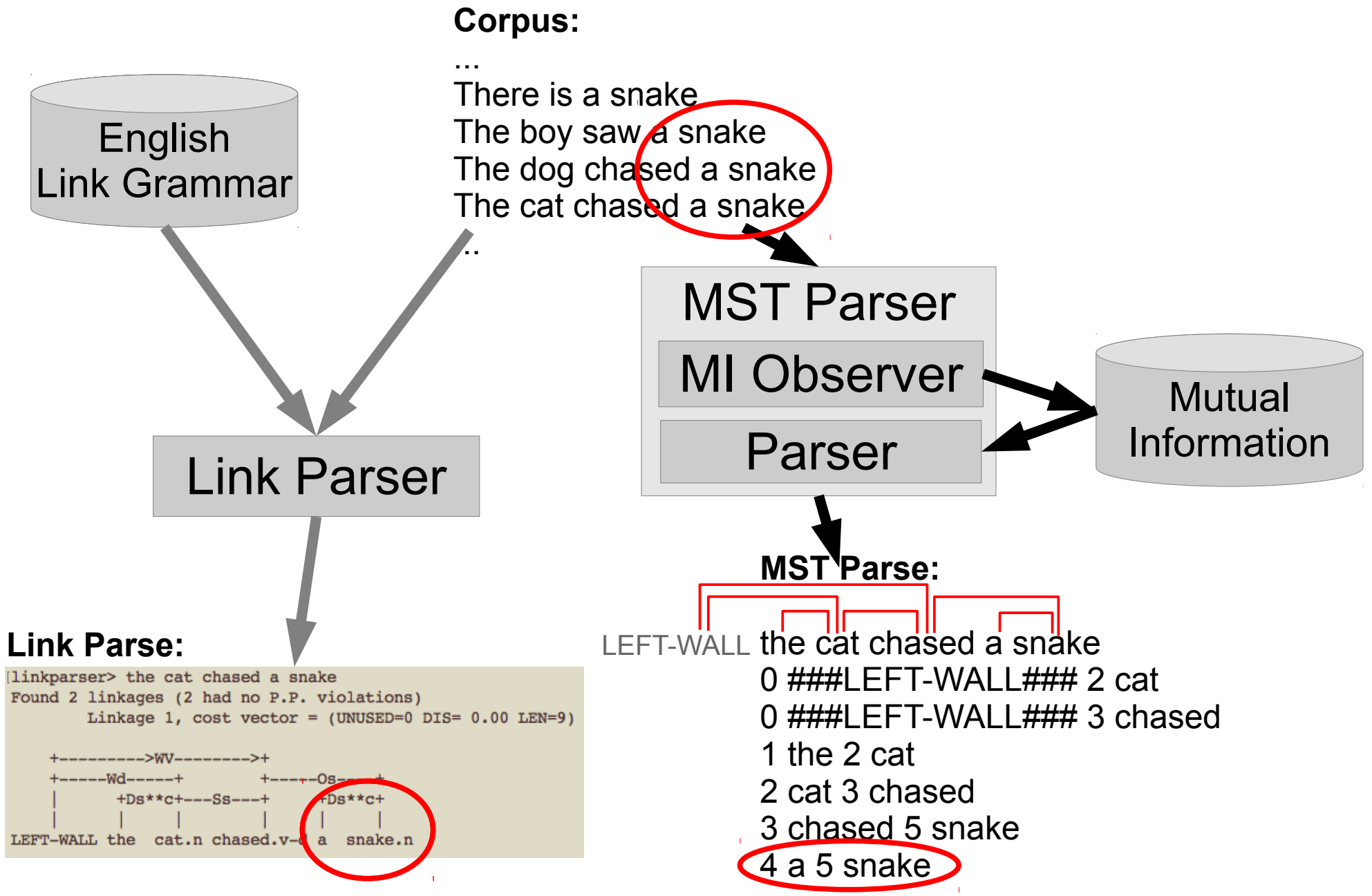
# Unsupervised language learning pipeline with OpenCog



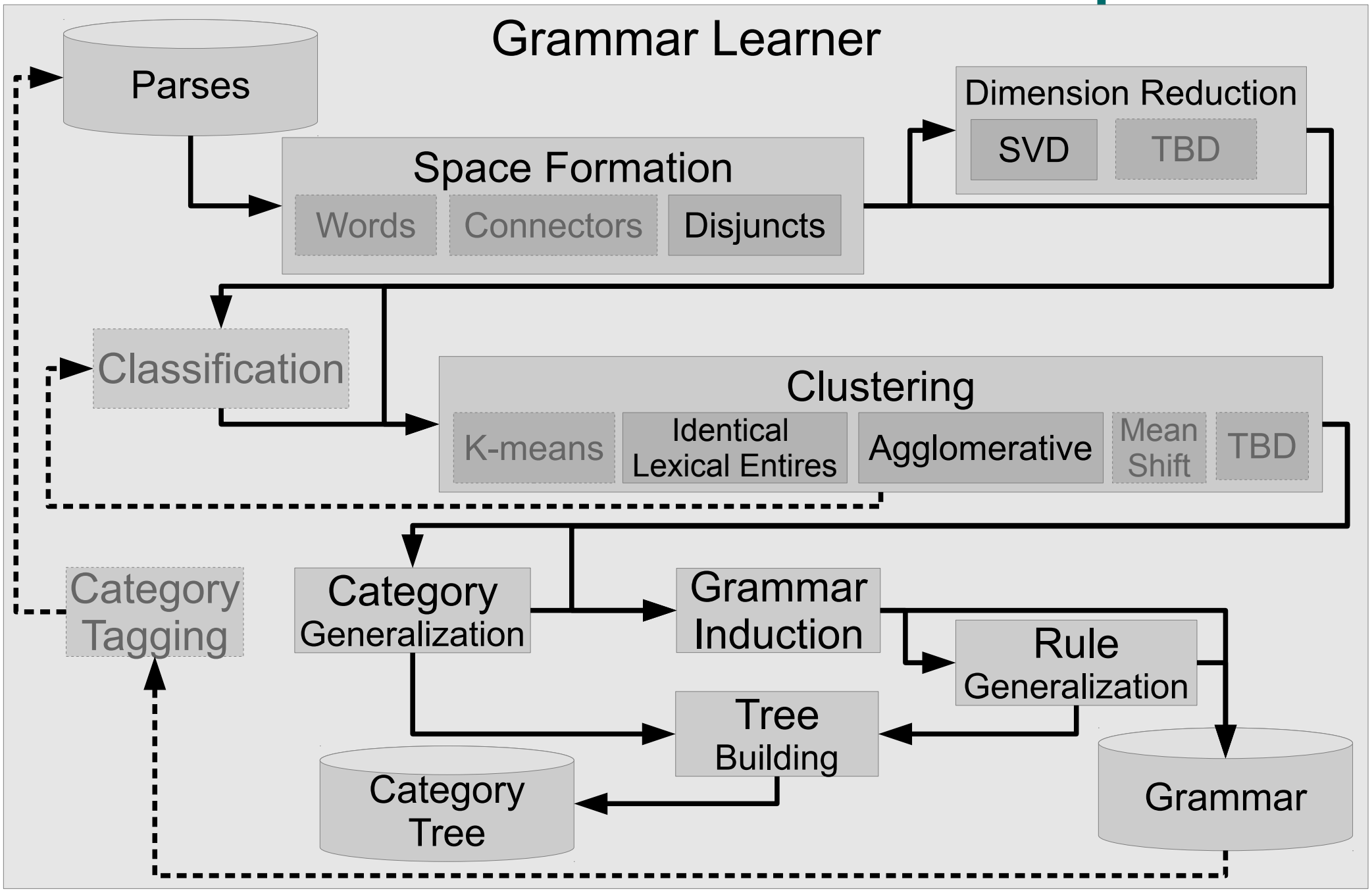
# Text Parsing for Link Grammar



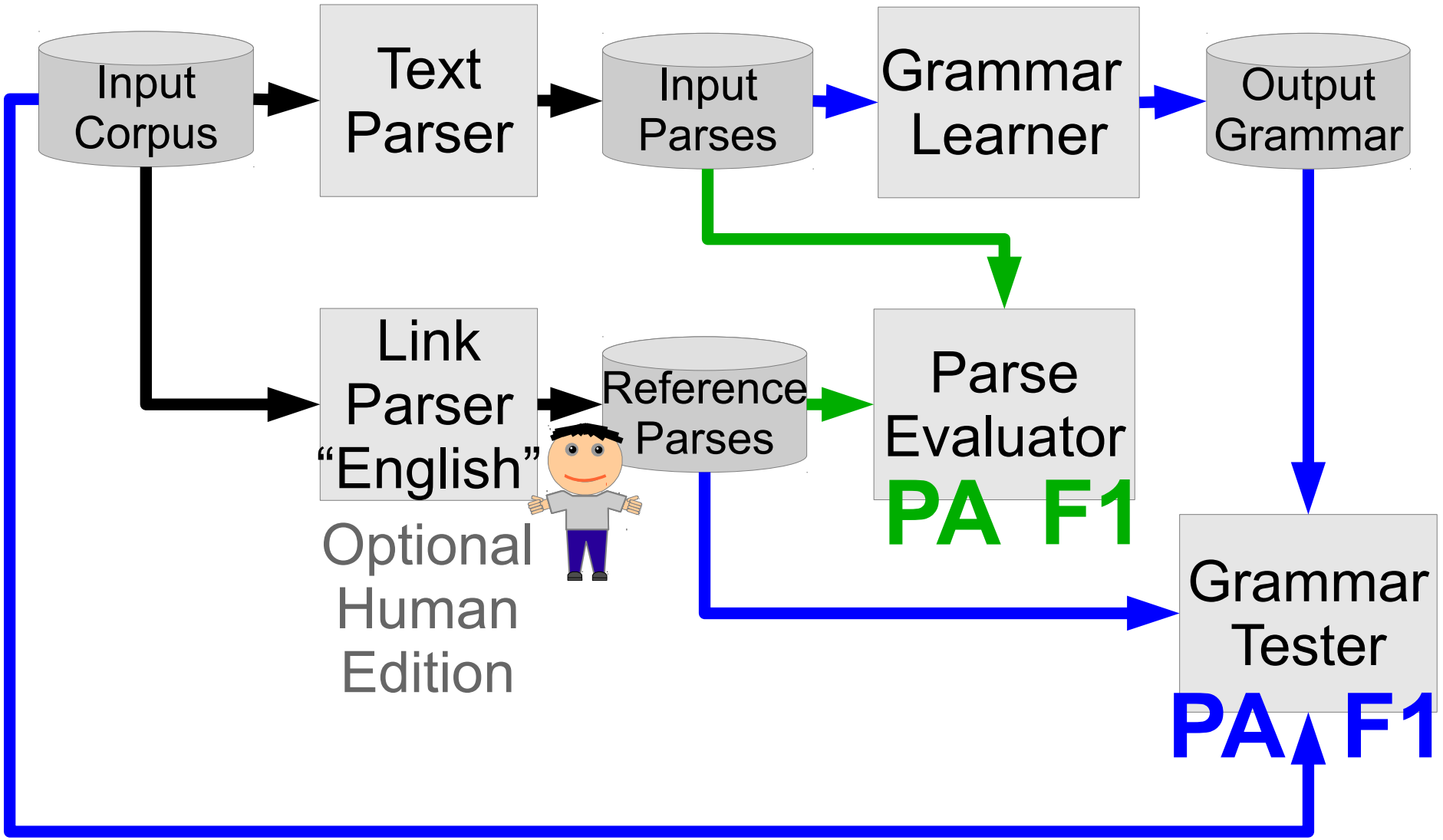
# MST Parses vs. Link Parses



# Link Grammar Learner Pipeline



# Quality-Assessment with on Parses and Grammar



# Corpora in Use

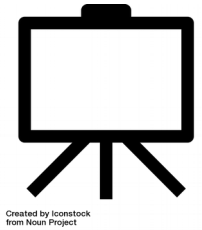
| Corpus | Total words | Unique words | Occurrences per word | Total sentences | Average sentence length |
|---|---|---|---|---|---|
| POC-English | 388 | 55 | 7 | 88 | 4 |
| Child-Directed Speech | 124185 | 3399 | 37 | 38181 | 4 |
| Gutenberg Children | 2695151 | 54054 | 50 | 207130 | 13 |

- POC-English – Proof-of-Concept corpus made of artificially selected sentences on limited number of topics (“small world”).
- Child Directed Speech (CDS) – corpus obtained from subsets of the CHILDES corpus – a collection of English communications directed to children with limited lexicon and grammar complexity (https://childes.talkbank.org/derived/)
- 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/)

# Word-Sense Disambiguation

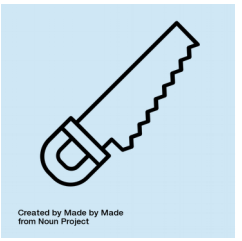
Using AdaGram[1] we disambiguate our POC-English corpus without supervision.

Two ambiguous words in corpus, with only two senses each:





board



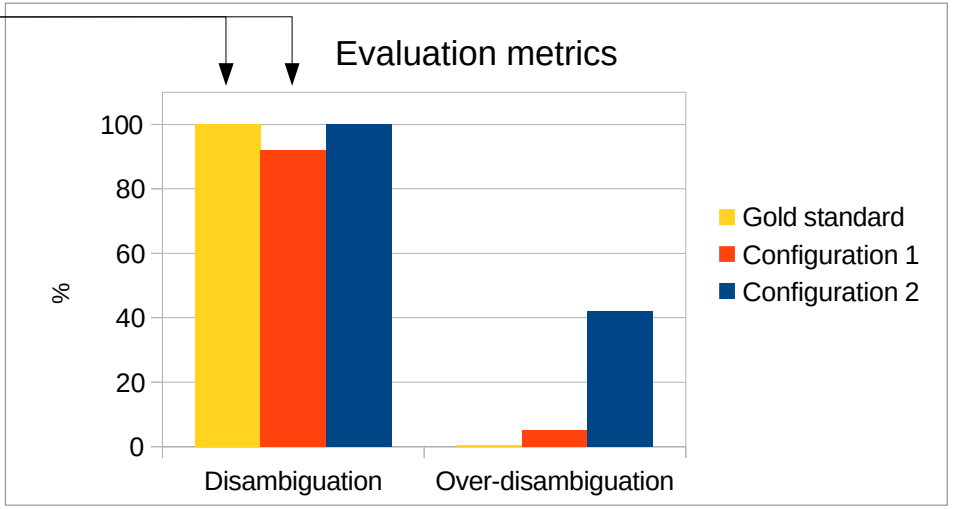


saw

After parameter tuning, we found two promising results:

mom saw@a dad with a saw@b .

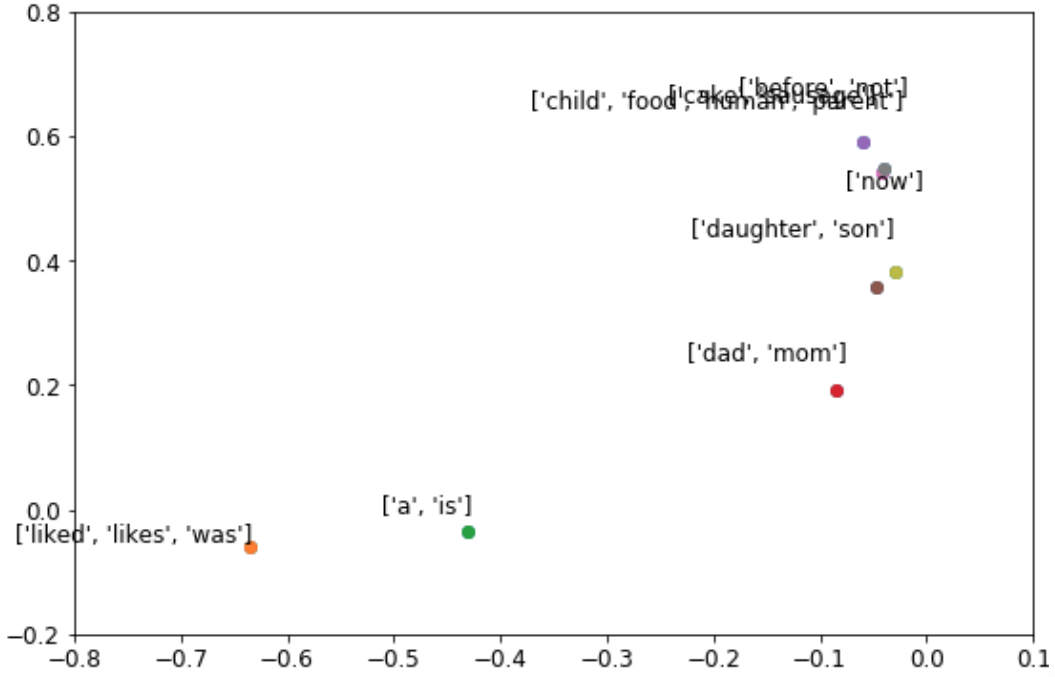
mom@a saw@a dad@b with a@c saw@b .



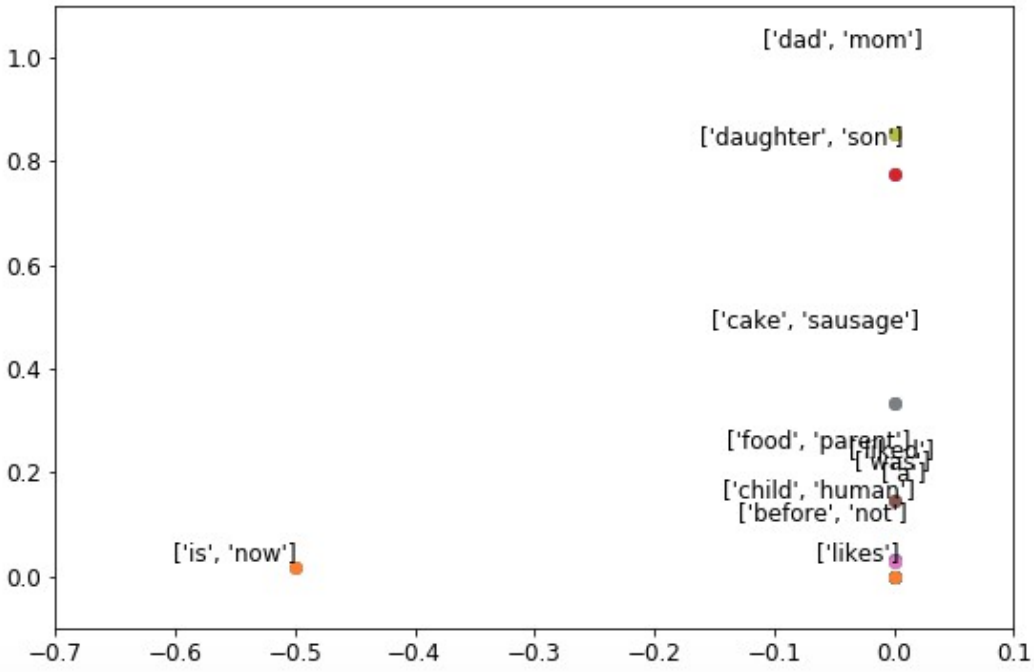

[1] https://github.com/glicerico/AdaGram/tree/take_sentences

# OpenCog Unsupervised Language Learning of Grammatical Categories and Link Grammar Dictionaries

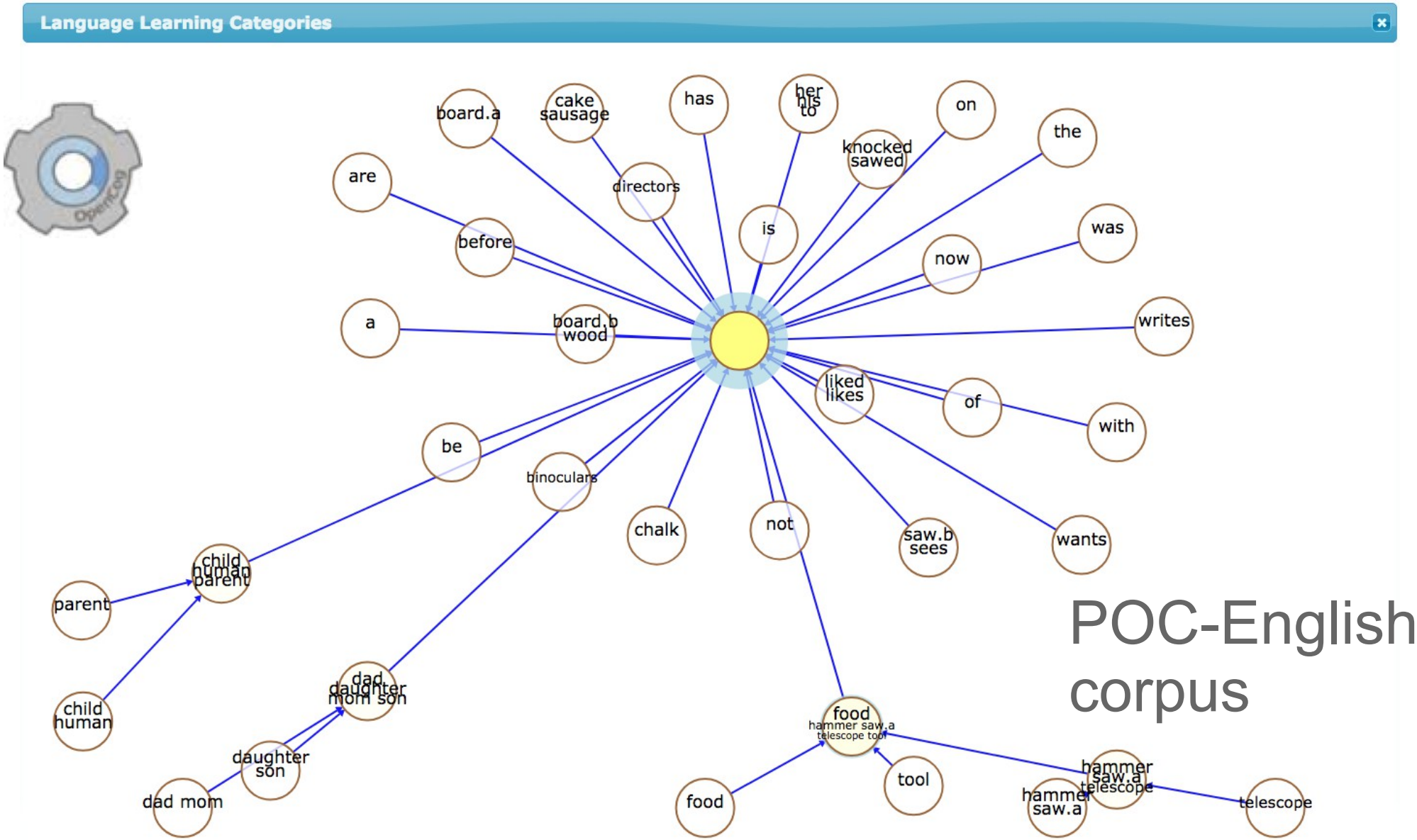
POC-English (Connectors)



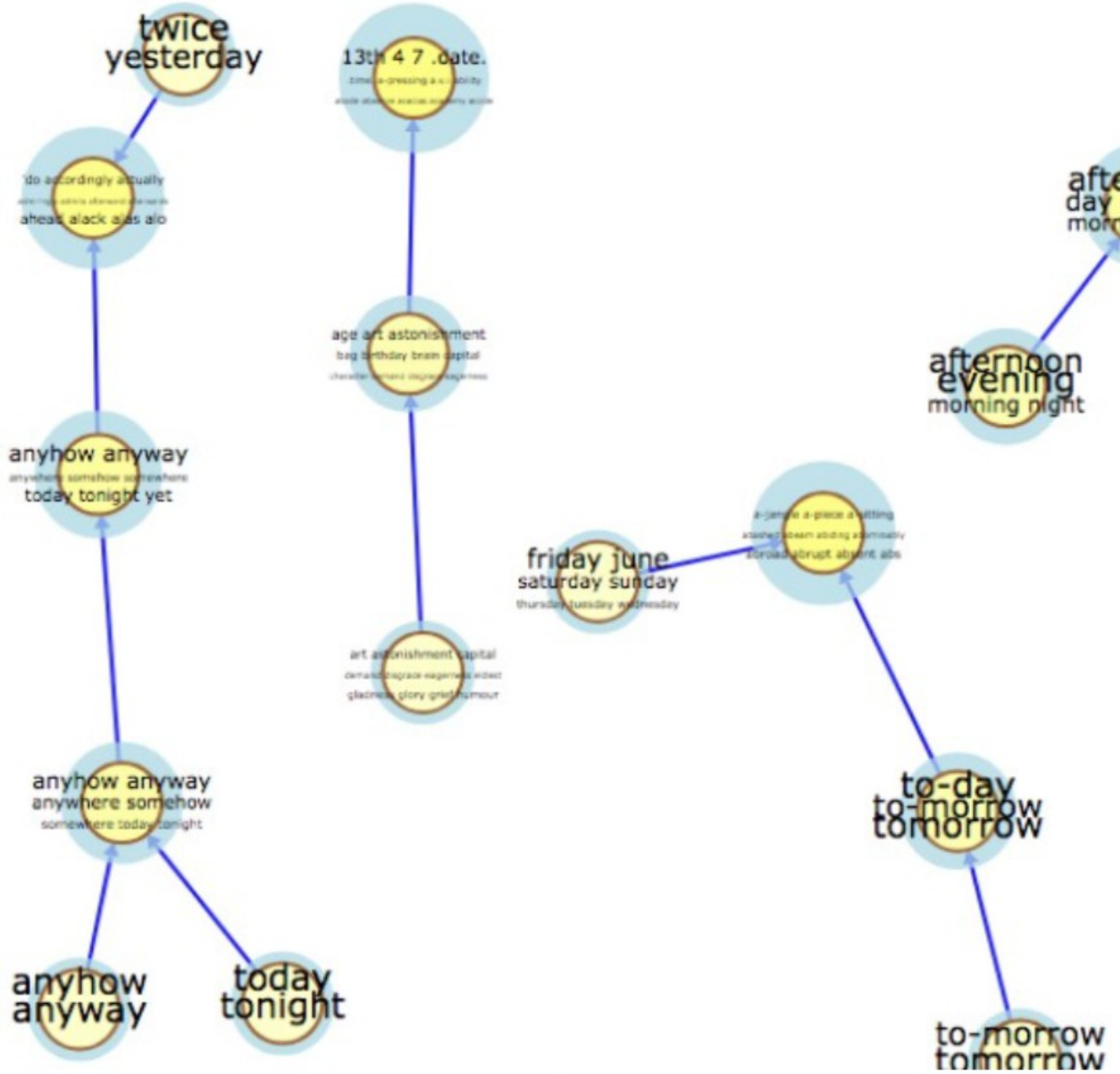

POC-English (Disjuncts)



# OpenCog Unsupervised Language Learning for Grammatical and Semantic Categories



# Grammar Ontology from Parses



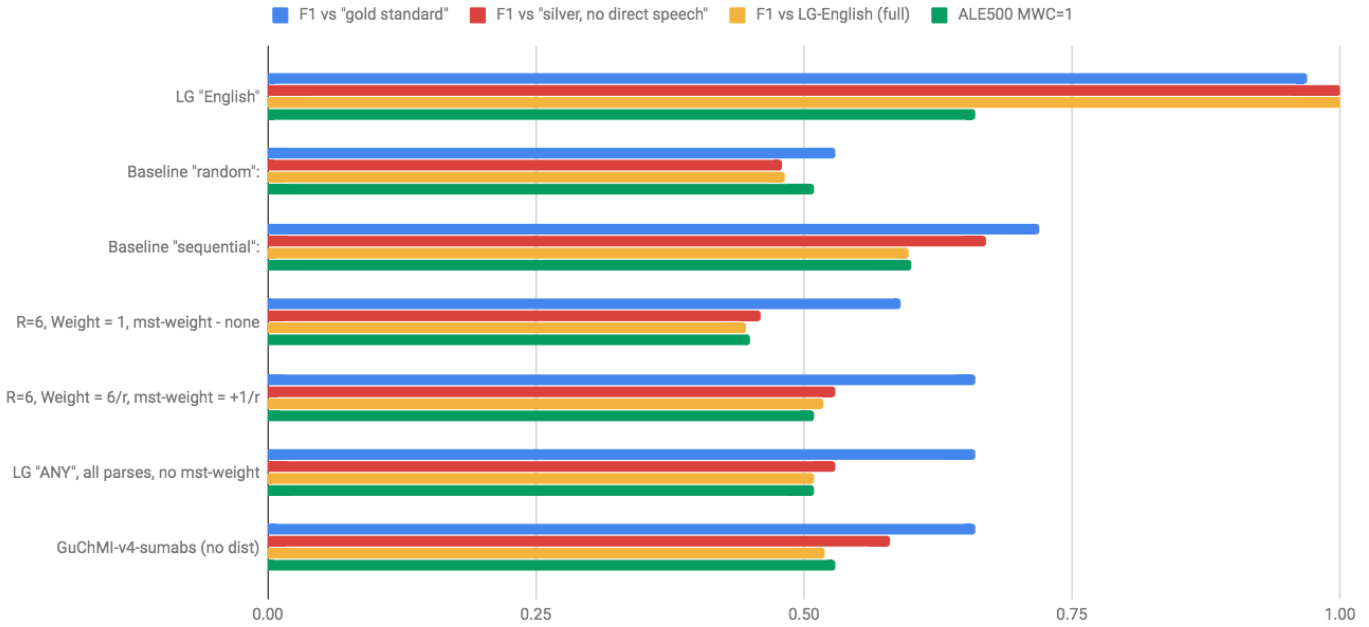

Gutenberg Children corpus

# F1 Results 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 |

# F1 Results Across the Parsers

| **Gutenberg-Children, GL on full corpus, max_unparsed_words=99, MWC(GL/GT) (test with full corpus "bronze standard")** | | F1 vs "gold standard" | F1 vs "silver, no direct speech" | F1 vs LG-English (full) | ALE500 MWC=1 | ALE500 MWC=2 | ALE500 MWC=3 | ALE500 MWC=4 | ALE500 MWC=5 |
|---|---|---|---|---|---|---|---|---|---|
| **Gutenber-Children** | LG "English" | 0.97 | 1.00 | 1.00 | 0.66 | 0.66 | 0.66 | 0.65 | 0.65 |
| **Gutenber-Children** | Baseline "random": | 0.53 | 0.48 | 0.48 | 0.51 | 0.51 | 0.51 | 0.51 | 0.51 |
| **Gutenber-Children** | Baseline "sequential": | 0.72 | 0.67 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 | 0.60 |
| **Gutenber-Children** | R=6, Weight = 1, mst-weight - none | 0.59 | 0.46 | 0.45 | 0.45 | 0.45 | 0.46 | 0.46 | 0.46 |
| **Gutenber-Children** | R=6, Weight = 6/r, mst-weight = +1/r | 0.66 | 0.53 | 0.52 | 0.51 | 0.52 | 0.53 | 0.53 | 0.53 |
| **Gutenber-Children** | LG "ANY", all parses, no mst-weight | 0.66 | 0.53 | 0.51 | 0.51 | 0.51 | 0.51 | 0.52 | 0.52 |
| **Gutenber-Children** | GuChMI-v4-sumabs (no dist) | 0.66 | 0.58 | 0.52 | 0.53 | 0.54 | 0.54 | 0.54 | 0.54 |



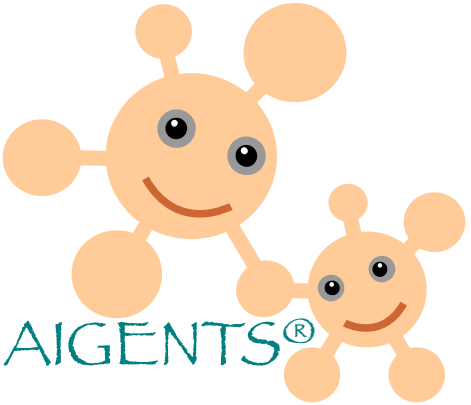
# Conclusions and Next Steps

- Grammars can be induced from parses
- Better parses => better grammars
  (Pearson between F1 on parses and F1 on grammar ≥ 0.9)
- MST-Parsing can't get better than "sequential" ("linked list") parsing
- Curriculum learning is a next try for:
  - Parses better than "sequential"
  - Better grammars for larger corpora

# Agents® "Deep Patterns" for Text Mining and Production

Anton Kolonin
akolonin@aigents.com

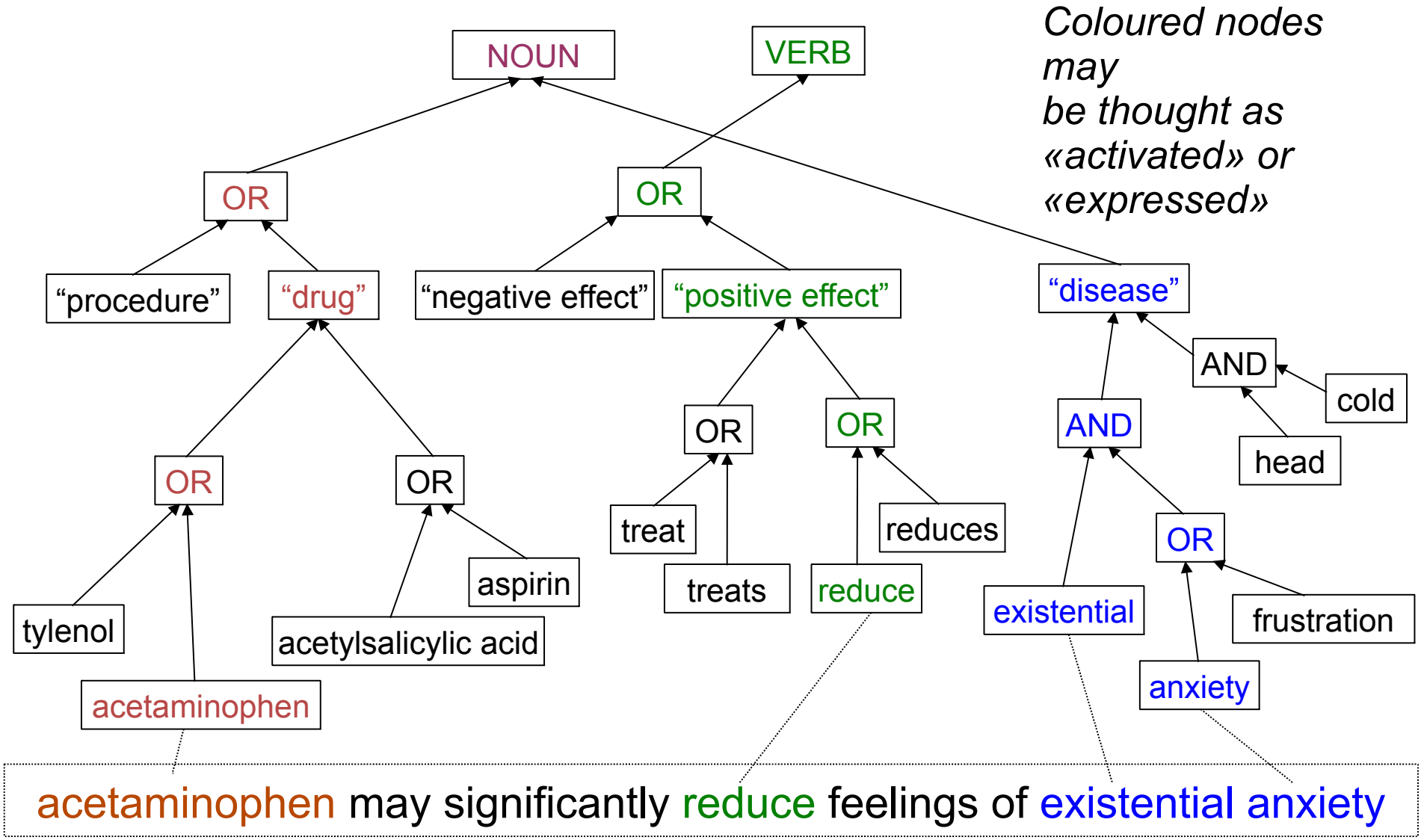
Novosibirsk State University
*THE REAL SCIENCE




SingularityNET
https://singularitynet.io

https://aigents.com

# Aigents® “Deep Patterns” - Language Model

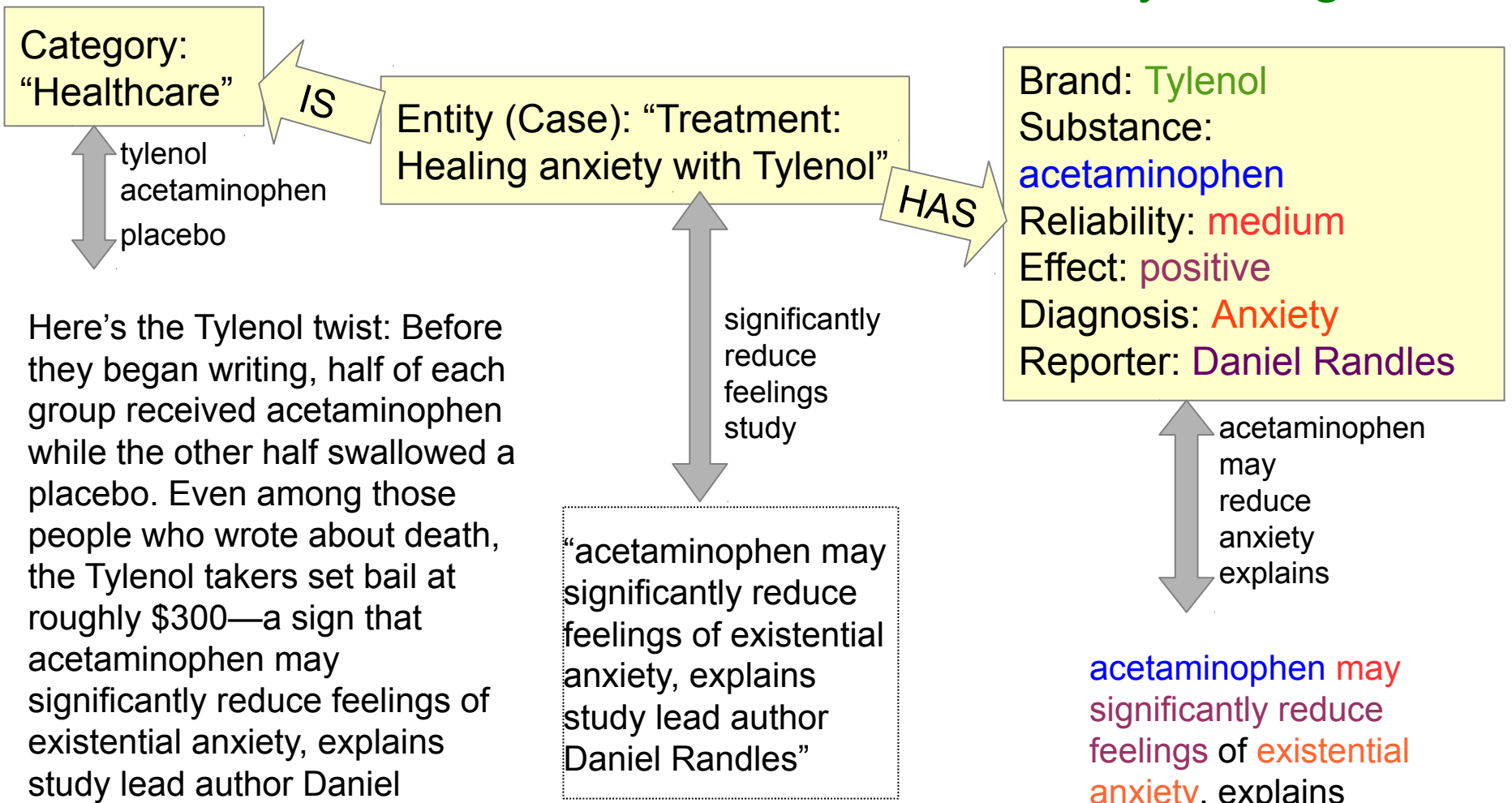


# Aigents® "Deep Patterns" - Text Mining

## Classification

## Case/Relationship Extraction

## Attribution and Entity Recognition




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.

# Aigents® “Deep Patterns” - Text Mining

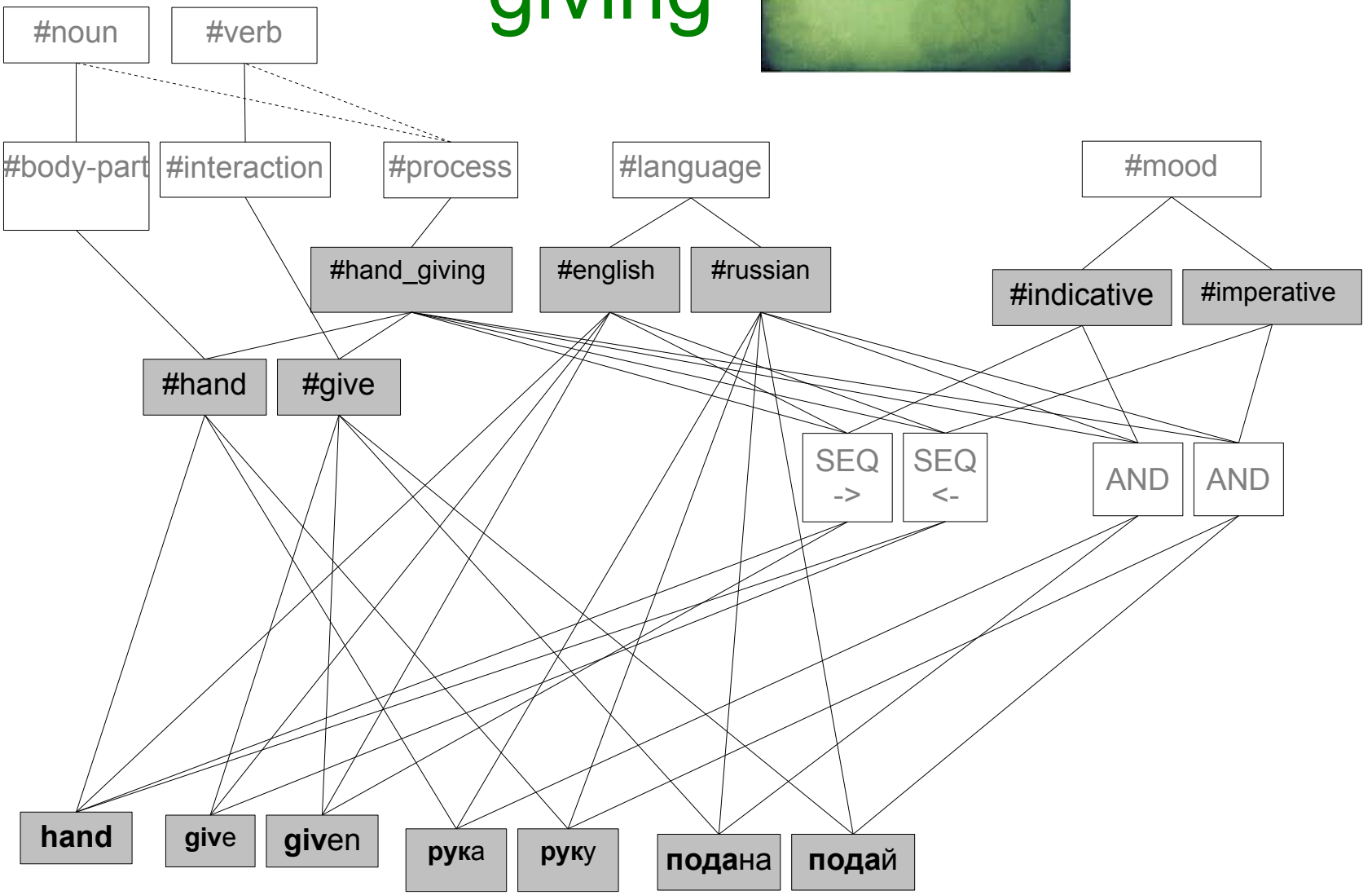
<pattern> := <token> | <regexp> | <variable> | <set>
<set> := <conjunctive-set> | <N-gram> | <disjunctive-set>
<disjunctive-set> := { <pattern> * }
<conjunctive-set> := ( <pattern> * )
<N-gram> := [ <pattern> * ]

**Example:**

{[$description catheter] [$coating coating] [$inner-diameter
{diameter inner-diameter}] [$tip tip] [$pattern pattern]}
X
Convey Guiding Catheter. Unique hydrophilic coating.
Small atraumatic soft tip. Ultra-thin 1 × 2 flat wire braid pattern
=
{ coating : 'hydrophillic', description : 'convey guiding',
pattern : 'ultra-thin 1 × 2 flat wire braid', tip : ‘sof’t }

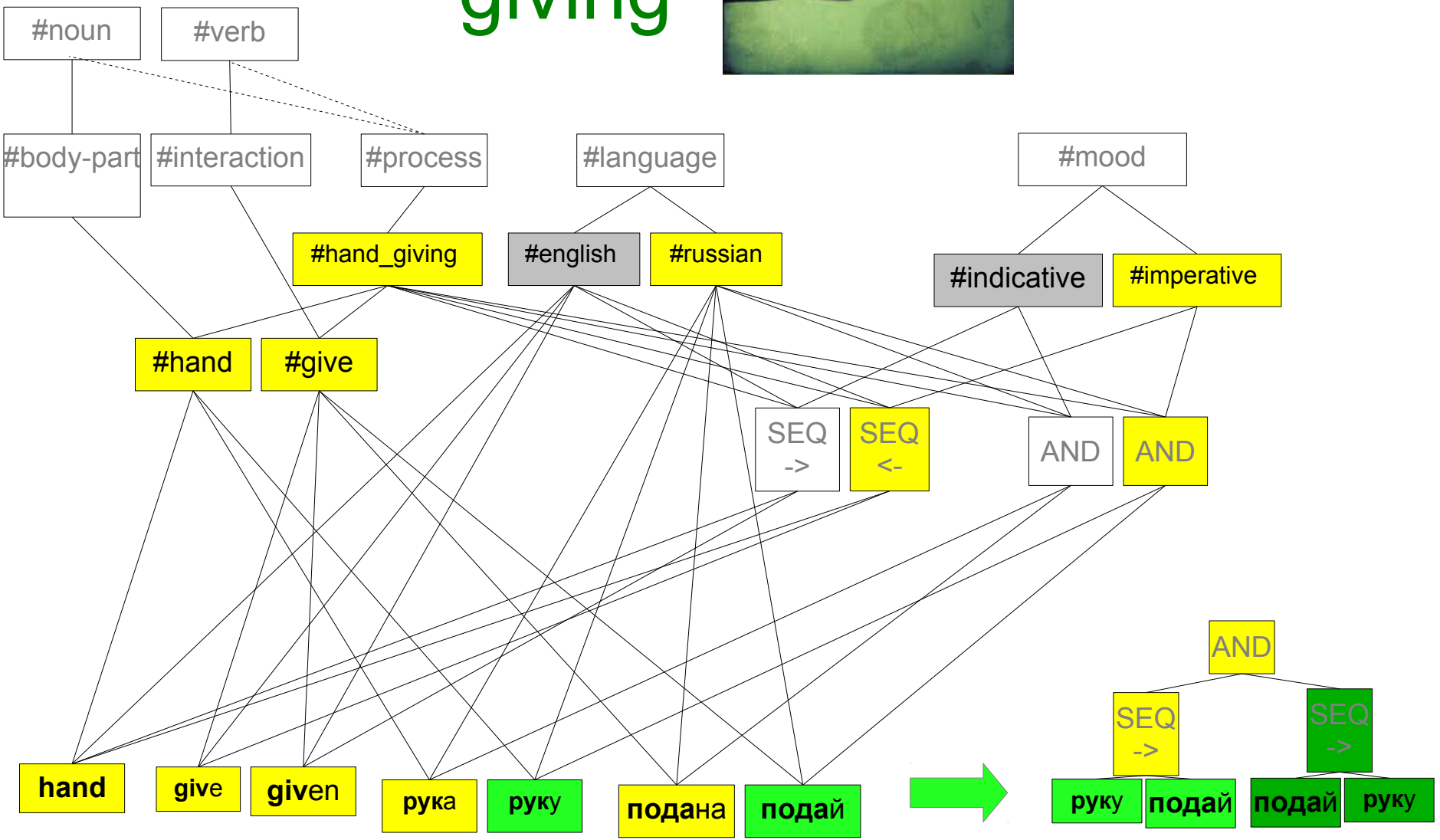
# Grammar & Ontology Graph - Structure

## Hand giving



# Grammar & Ontology Graph - Production

## Hand giving



# Challenge – Integration of Syntactic (tokens and “word-pieces”) and Semantic (“Knowledge Graphs”) Representations for Context-based Word Sense Disambiguation

Какой (свойство зрения)?
Какой (состояние опьянения)?
Кто (профессия)?
Кто (имя, кличка)?
С чем?
Чем?
Что делал?
Где?
Как?

Косой косой косарь Косой с косой косой косил на косе косо.

Drunk oblique mower Kosoy with a slanting spit was mowing on a bar obliquely.

# Thank you for attention!
# Questions?

Anton Kolonin
akolonin@aigents.com

Novosibirsk State University
*THE REAL SCIENCE


https://aigents.com

SingularityNET
https://singularitynet.io