Programmatic Link Grammar Induction for Unsupervised Language Learning

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

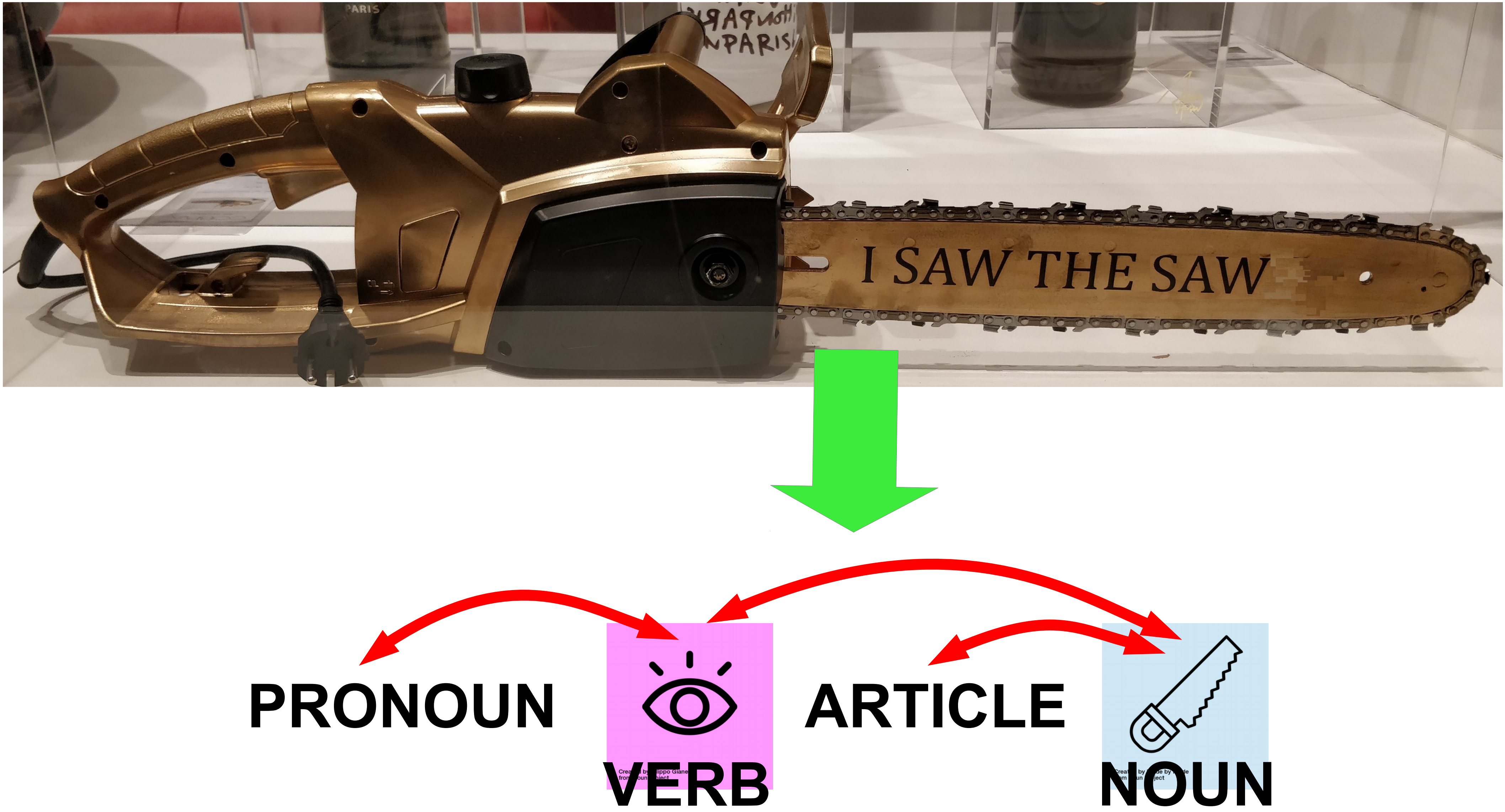
Presenters: Ben Goertzel, Anton Kolonin
ben@singularitynet.io anton@singularitynet.io

OpenCog
https://opencog.org/

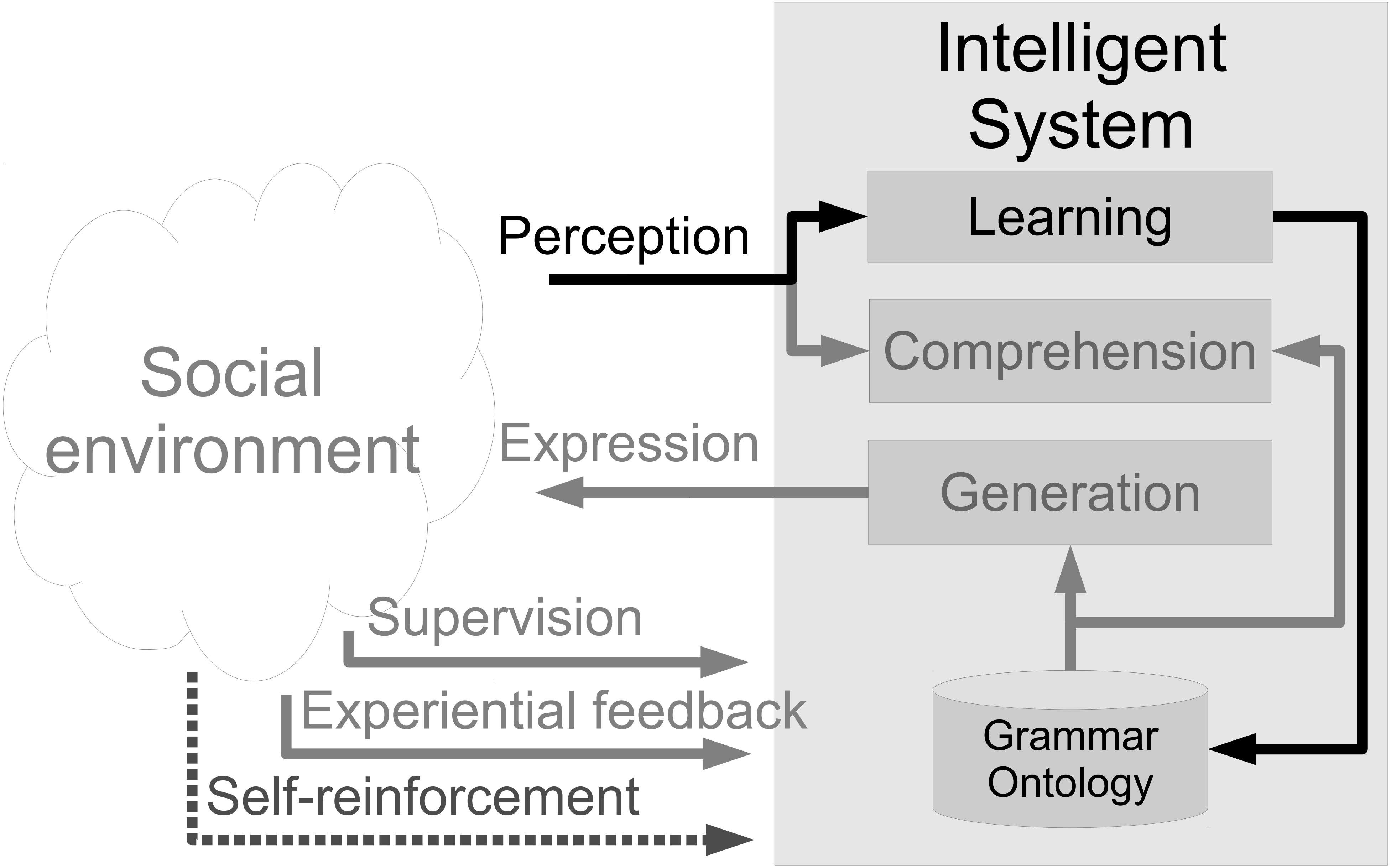
SingularityNET
https://singularitynet.io

HANSON ROBOTICS
http://www.hansonrobotics.com/

Grammar Learning from Scratch - Programmatically



Language Learning Environment



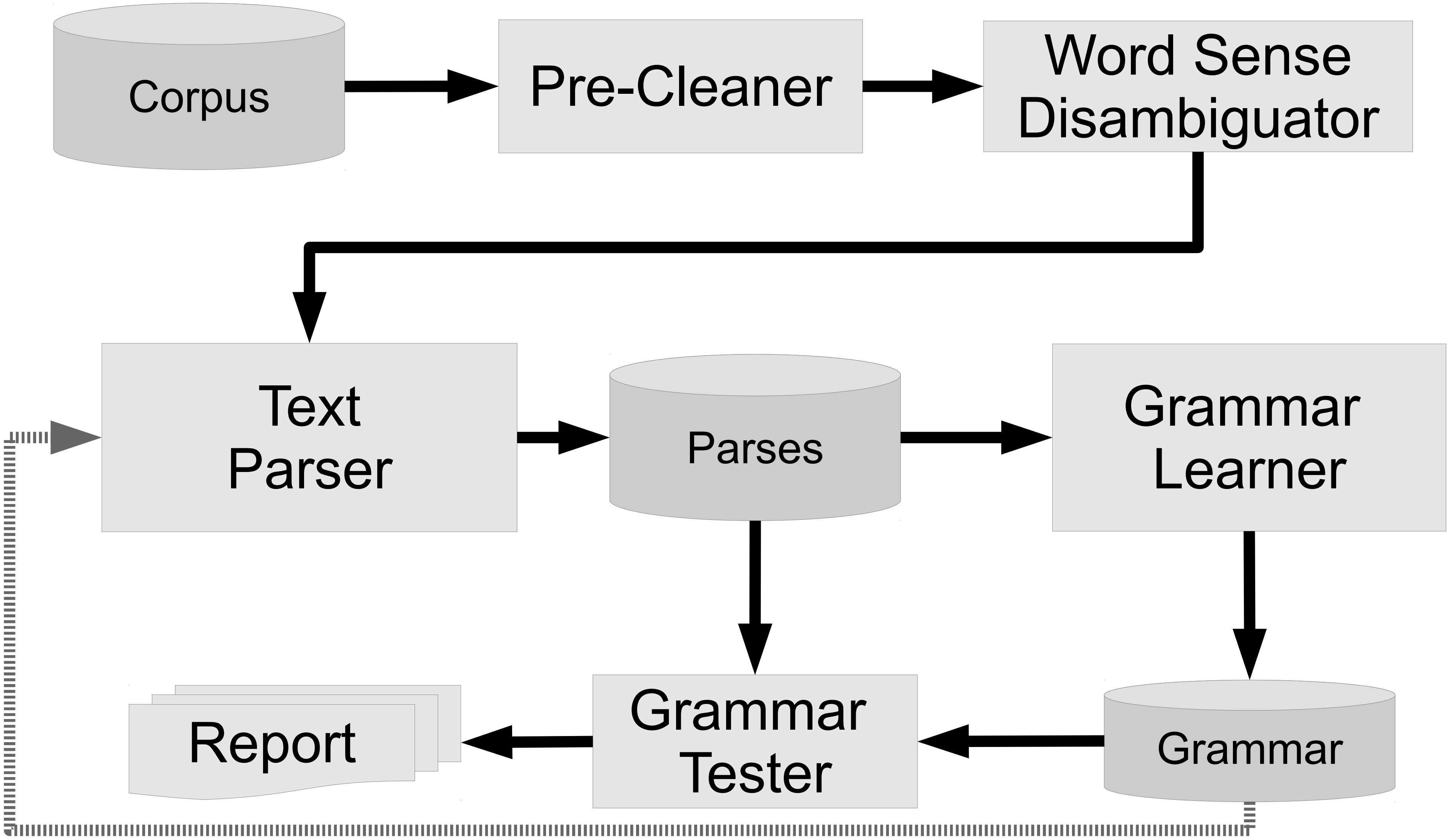
Project goal and applications

- 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)

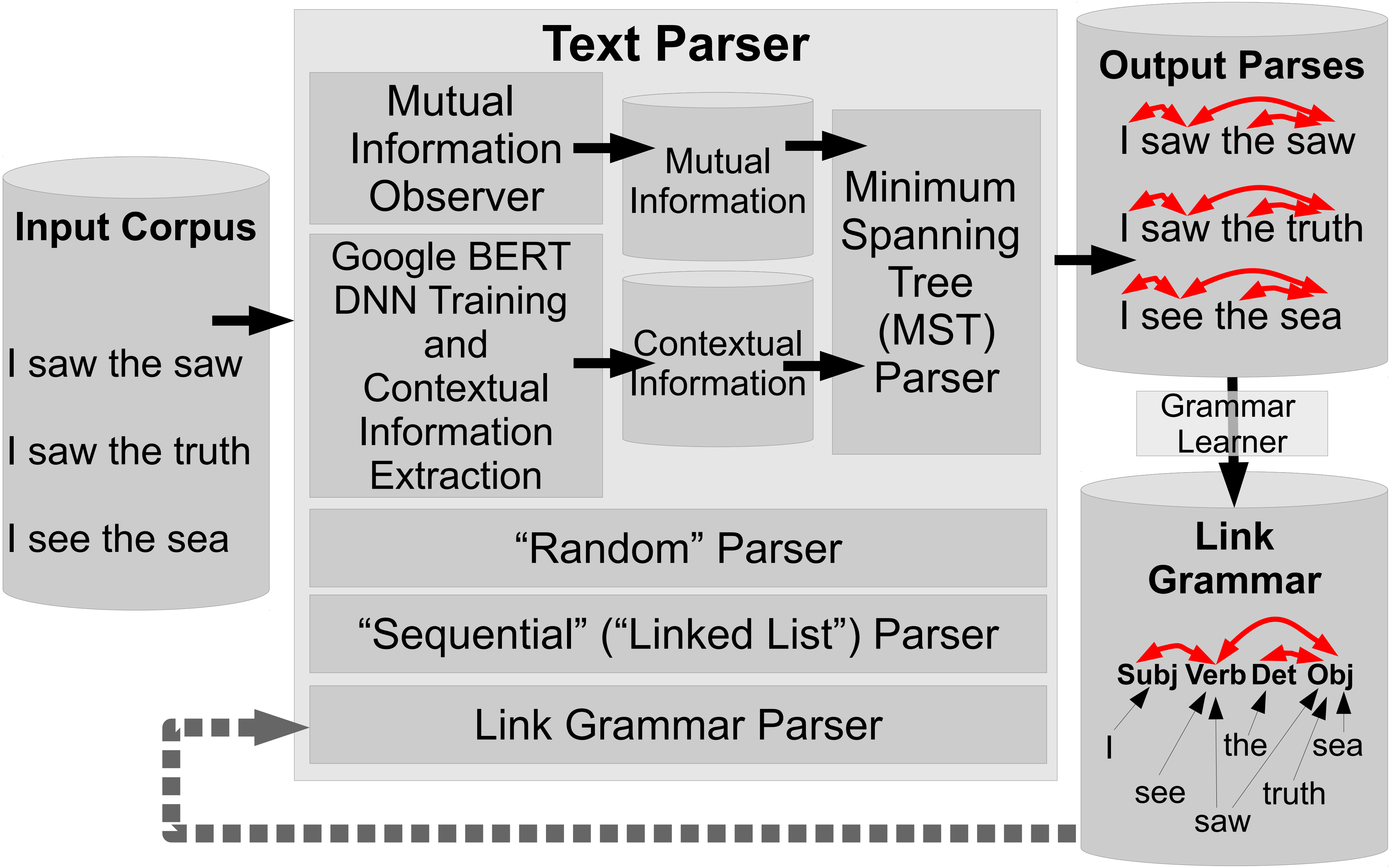
Constraints of the currently explored approach

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

Unsupervised language learning pipeline with OpenCog



Text Parsing for Link Grammar



Input Corpus

I saw the saw

I saw the truth

I see the sea

Text Parser

Mutual Information Observer

Mutual Information

Google BERT DNN Training and Contextual Information Extraction

Contextual Information

Minimum Spanning Tree (MST) Parser

“Random” Parser

“Sequential” (“Linked List”) Parser

Link Grammar Parser

Output Parses

I saw the saw

I saw the truth

I see the sea

Grammar Learner

Link Grammar

Subj Verb Det Obj

I

see

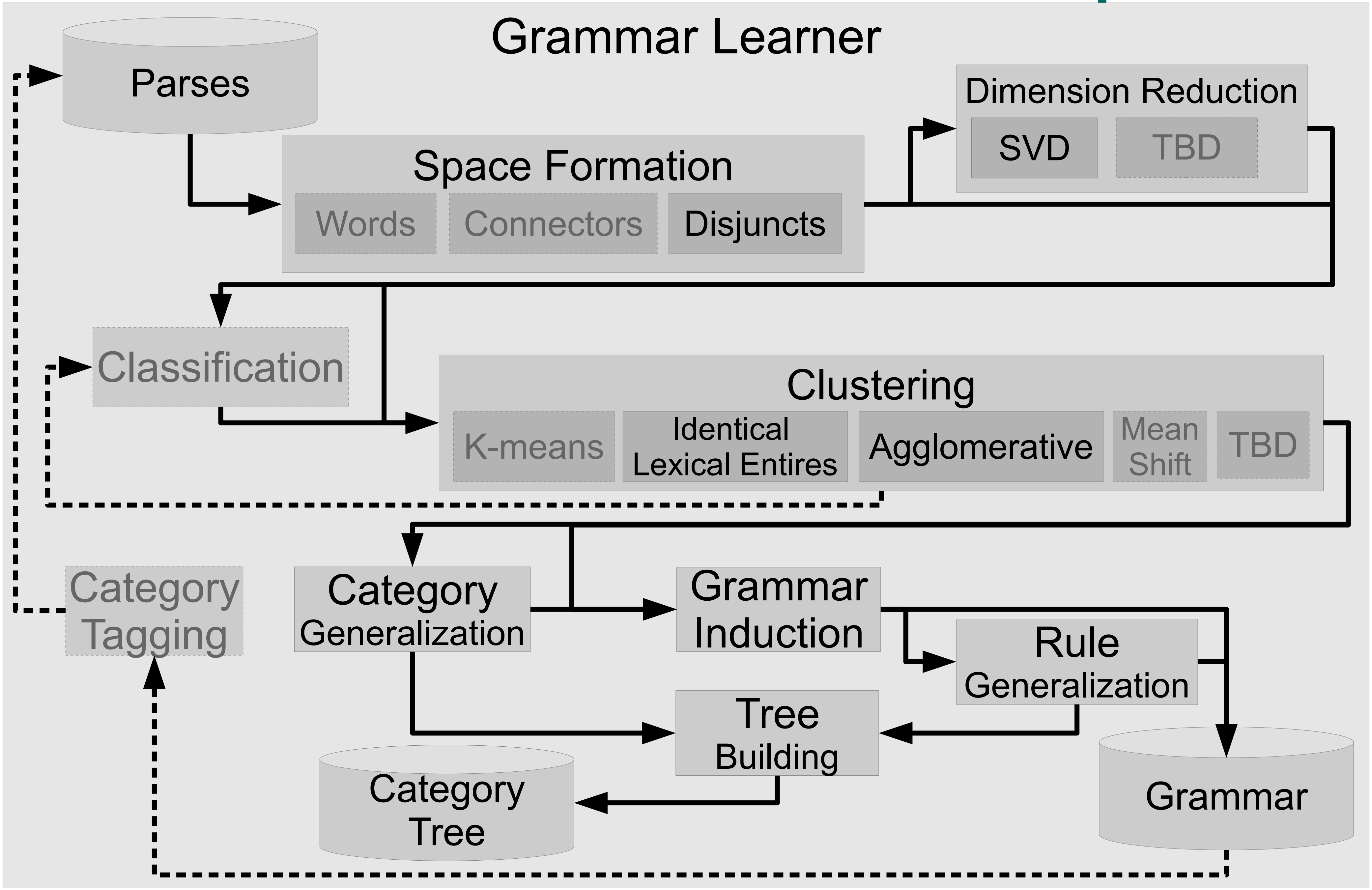
saw

the

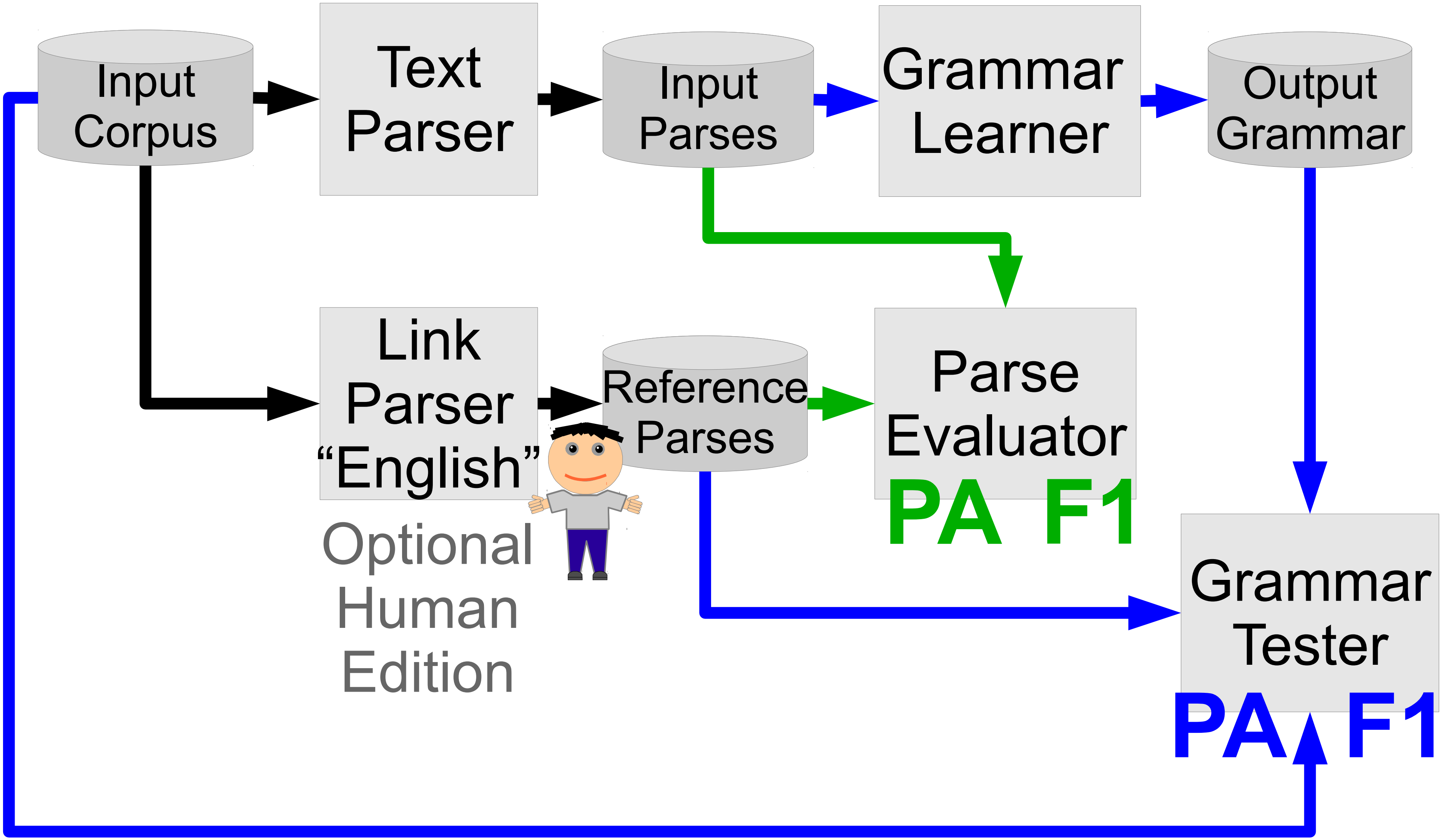
truth

sea

Link Grammar Learner Pipeline

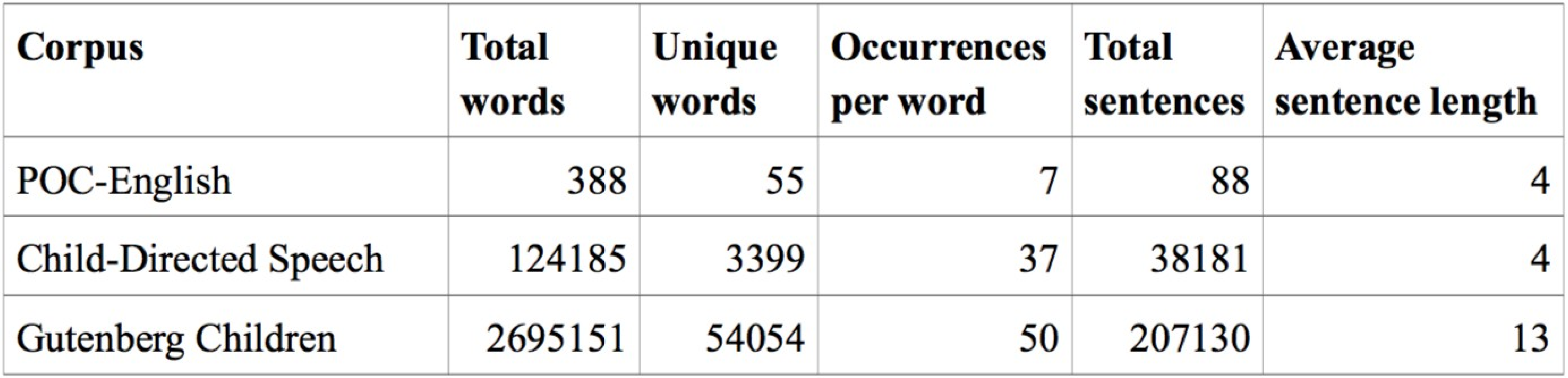


Quality-Assessment with on Parses and Grammar



Copyright © 2019 SingulariytNET Foundation

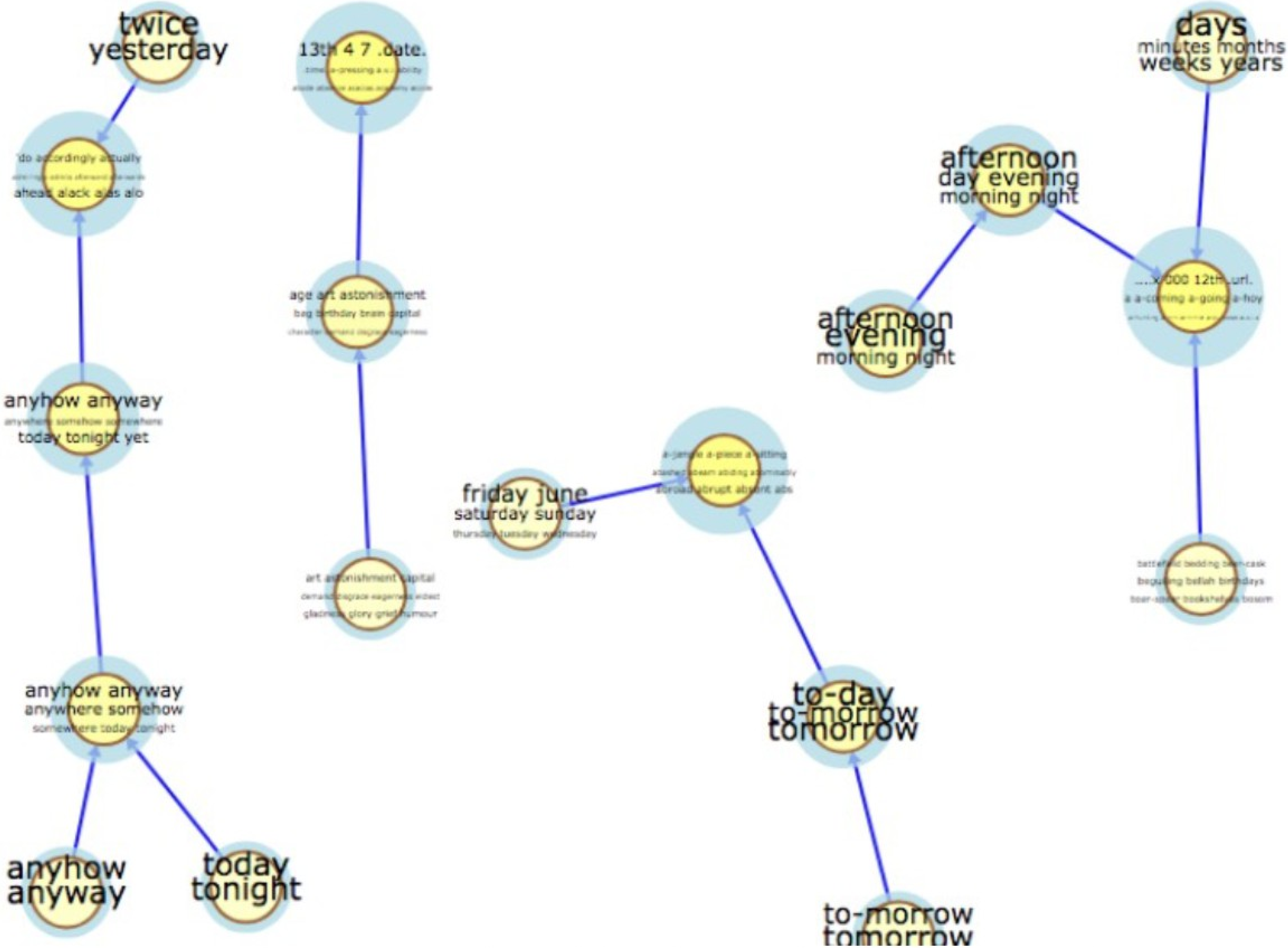
Corpora in Use



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

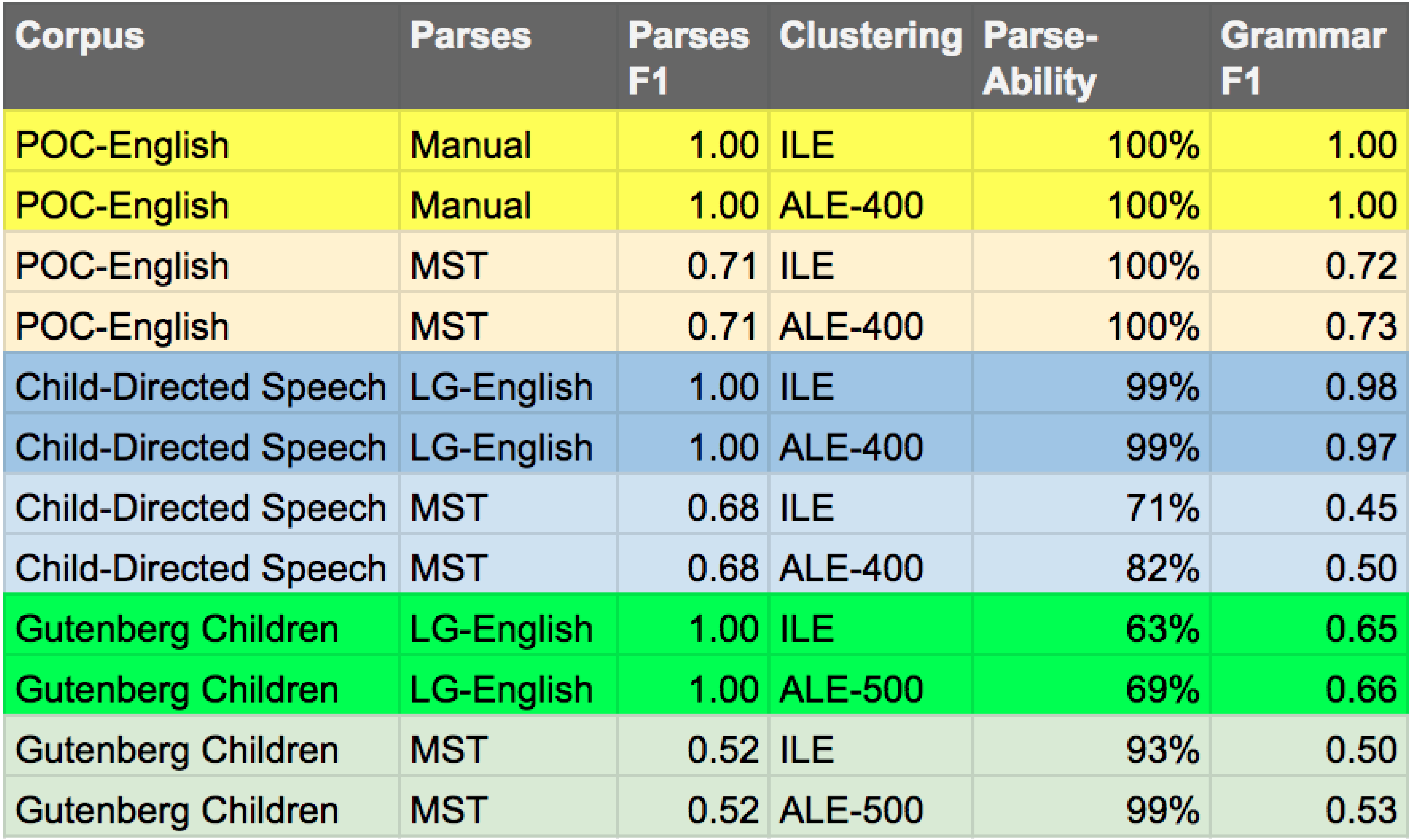
- POC-English – Proof-of-Concept corpus made of artificially selected sentences on limited number of topics (“small world”).
- Child Directed Speech (CDS) – corpus obtained from subsets of the CHILDES corpus – a collection of English communications directed to children with limited lexicon and grammar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

Grammar Ontology from Parses



Copyright © 2019 SingulariytNET Foundation

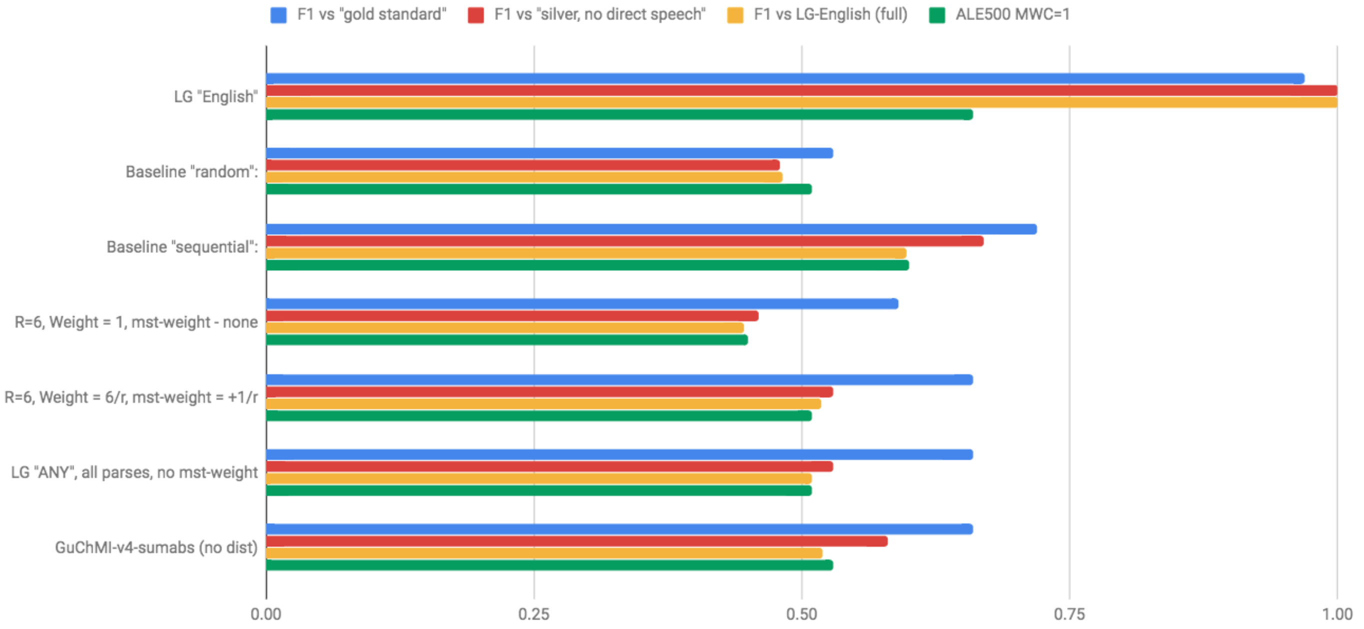
F1 Results Across the Corpora



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

Thank you and visit us at:

http://langlearn.singularitynet.io/

Stay in touch:

Ben Goertzel
ben@singularitynet.io
Anton Kolonin
anton@singularitynet.io

OpenCog
https://opencog.org/

SingularityNET
https://singularitynet.io

HANSON ROBOTICS
http://www.hansonrobotics.com/