Göttingen Symposium on
The Semantics of Action

Towards a model for learning senso-motorically grounded action constructions

Prof. Dr. Philipp Cimiano

Language Acquisition

Understanding language acquisition

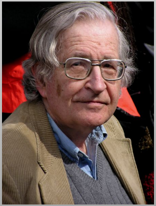
Understanding language acquisition is a holistic problem requiring to understand:

- representation of concepts
- generalization from observed input
- incremental learning
- timing of learning (critical periods),
- ...

Understanding language acquisition requires to understand how our cognitive systems works!

Is language inborn?

Dichotomy between inborn, universal language skills and fully constructed (a.k.a. acquired) grammar



Hypotheses and research questions more nuanced

- Mechanisms for acquiring a language available, but what exactly is inborn and what is acquired?
- How do domain-specific and domain-independent mechanisms interact?
- How are concepts acquired and represented in order to ground meaning of words?
- How is grammar represented in the cognitive / neural system?

Cognitive Grammar (Langacker)

Criticism: Generative Grammar neglects the function of language, i.e meaning, pragmatics etc.
Cognitive Grammar sets this right by considering both: form *and* function

Construction Grammar (Goldberg et al.)

Construction Grammar (Goldberg and others) was inspired by Cognitive Grammar and puts functional aspects of language at the center.

- Focuses on **constructions** as pairings of form and meaning/function.
- They are **signs,** i.e. arbitrary pairings between a form and a meaning or function.
- Constructions are primitive in the sense that they are **non-compositional.**
- Constructions are represented in a network of constructions (**constructicon**) of interrelated constructions.

Example constructions from Goldberg

Morpheme	e.g. pre-, ing
Word	avocado, anaconda, and
Complex Words	daredevil
Inflectional Constructions	[N-s]
Idiom	kick the bucket
Covariational Conditional	The more X the more Y
Ditransitive verb	Subj V Obj_1 Obj_2
Passive	Subj aux VPP (PP_{by})

Empirical Findings within Language Acquisition

- learning is item-based: word islands that are not fully productive (verbs, e.g. Tomassello, determiners, e.g. Pine et al.)
- language acquisition has an important statistical component, i.e. children accumulate evidence across many contexts:
  - nouns (Yu and Smith, 2006)
  - verbs (Scott and Fisher, 2012)
- theory of mind and inferring intentions is crucial (Tomassello et al.
- ...

Towards a model for learning action constructions

Focus on lexicalized constructions involving verbs.
Such a model needs to explain:

- How constructions are represented: form and meaning
- How they are acquired: form and meaning
- Generalization: form and meaning
- Compositionality
- Incrementality

Properties of our model

- It represents linguistic knowledge as a **construction network** containing linguistic constructions at different levels of abstraction
- It assumes **no pre-coded linguistic knowledge**
- It performs **unsupervised** learning, i.e. it requires no explicit tutoring
- It learns **online** in the sense that each example directly causes a change in the network structure
- It learns **incrementally** in the sense that it first learns the meanings of simpler linguistic structures and then bootstraps on these to acquire more complex constructions
- It is capable of both language **understanding** and **generation**

Input

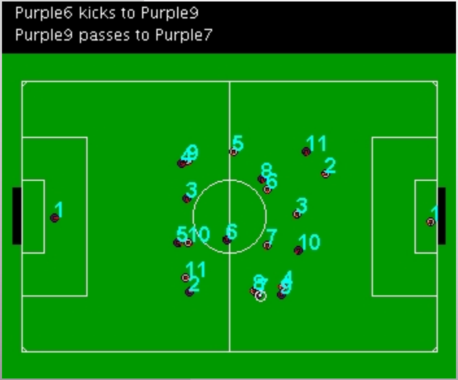
Pairings of observed action/situation and utterance:

Elena is passing the salt to Steven.

But: referential uncertainty!!!

Output: a grammar that can be used to parse input (no direct correspondence to formal grammars)

Robocup



Learning

Hebbian-style learning in a dynamically and incrementally growing network with different levels of representational granularity and complexity.

Problem Definition

NL:	purple10 kicks to purple7
mr_1:	ballstopped
mr_2:	badPass(pink1, purple10)
mr_3:	turnover(pink1, purple10)
mr_4:	playmode(play_on)
mr_5:	kick(purple10)
mr_6:	pass(purple10, purple7)

Constructions at different levels

- Word constructions
- Slot&Frame Patterns

Example

NL:	purple10 kicks to purple7
mr_1:	ballstopped
mr_2:	badPass(pink1, purple10)
mr_3:	turnover(pink1, purple10)
mr_4:	playmode(play_on)
mr_5:	kick(purple10)
mr_6:	pass(purple10, purple7)

NL:	pink goalie kicks to pink5
mr_1:	pass(pink1, pink5)
mr_2:	badPass(pink1, purple10)

Word Constructions

$\hat{NL}$	purple10
$\hat{mr}$	purple10

$\hat{NL}$	pink goalie
$\hat{mr}$	pink1

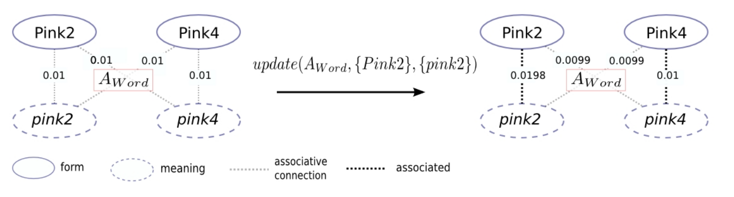
$\hat{NL}$	pink5
$\hat{mr}$	pink5

$\hat{NL}$	purple7
$\hat{mr}$	purple7

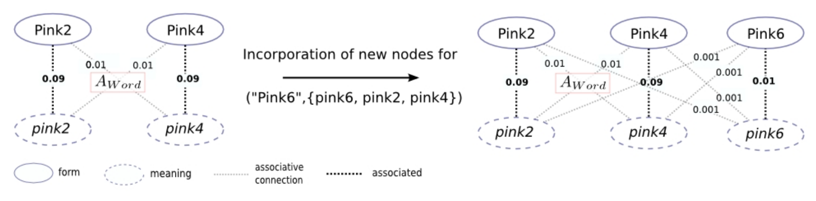
Slot&Frame Constructions

$\hat{NL}$	SE_1 kicks to SE_2
$\hat{mr}$	$pass(ARG_1, ARG_2)$
Φ	$SE_1 \longrightarrow ARG_1$ $SE_2 \longrightarrow ARG_2$

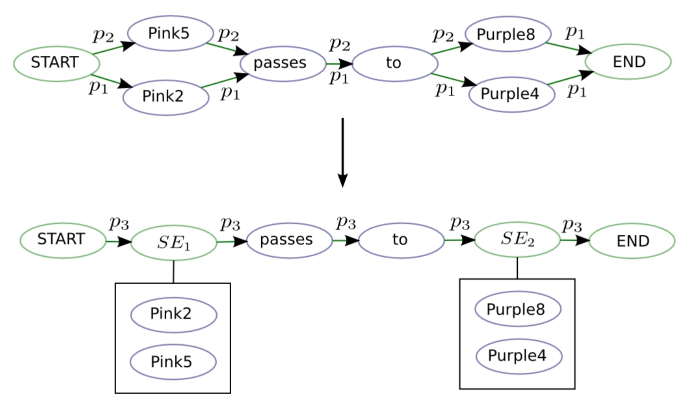
Associations between form and meaning at the word level



Dynamicity: Incorporating new nodes



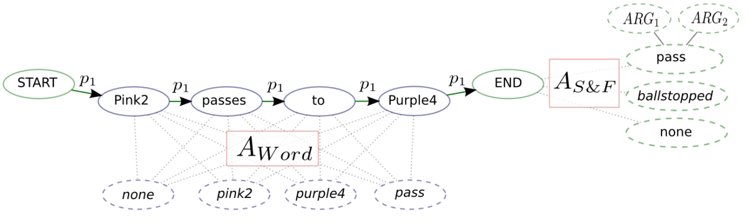
Word Order Graph / S&F Constuctions



Example: Pink2 passes to Purple4

NL:	Pink2 passes to Purple4
mr_1:	pass(pink2, purple4)
mr_2:	ballstopped

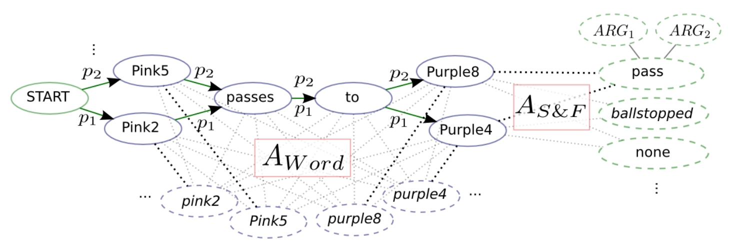
Result of incorporating: Pink2 passes to Purple4



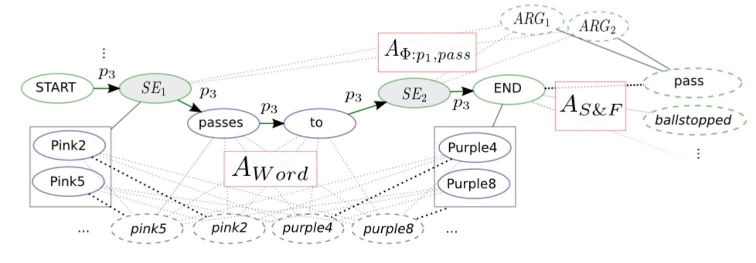
Pink5 passes to Purple4

NL:	Pink5 passes to Purple4
mr_1:	pass(pink5, purple4)
mr_2:	ballstopped

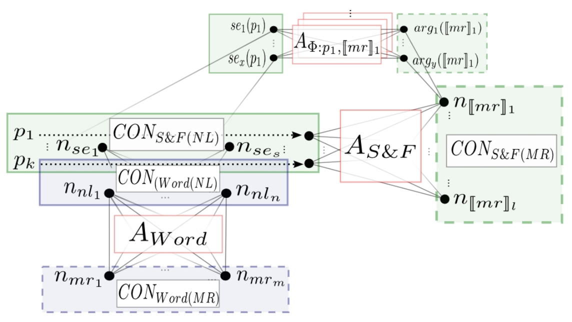
Result of incorporating: Pink5 passes to Purple4



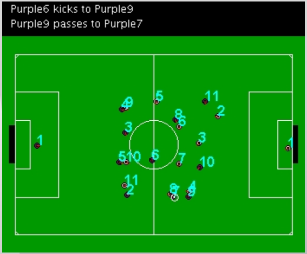
Adding argument mappings



Network Architecture



Robocup dataset (Mooney et al.)



Robocup dataset: 4 games generated by a Robocup simulator with human comments

Evaluation method: 4-fold cross-validation over 4 games

Semantic parsing task: Precision/Recall/F-Measure

Statistics about Robocup dataset

Total number of comments	1,872
Comments having correct mr	1,539
Average number of events per comment	2.5
Maximum number of events per comment	12
SD in number of events per comment	1.8
Mean utterance length	5.7
Number of tokens	10,700
Vocabulary size	443

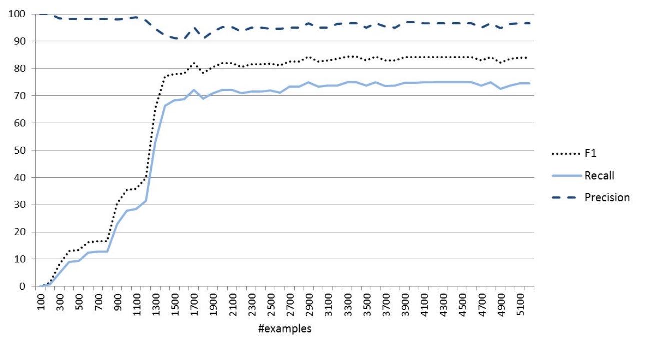
Results: Examples

Predicate	Avg #patterns	Example of an extracted $\hat{NL}$
pass	77.25 (SD: 22.4)	SE fires a pass to SE
kick	41 (SD: 7.4)	SE dribbles the ball
badPass	40.25 (SD: 11.6)	SE makes a bad pass that was intercepted by SE
turnover	20.25 (SD: 4.0)	SE turns the ball over to SE
steal	7.75 (SD:2.2)	SE steals the ball
block	5.5 (SD: 2.2)	SE blocked the ball
playmode	3 (SD: 1.7)	SE team scores
defense	0	--
ballstopped	0	--
$\perp$	182.5 (SD: 40.1)	The shot was just a bit wide of the goal

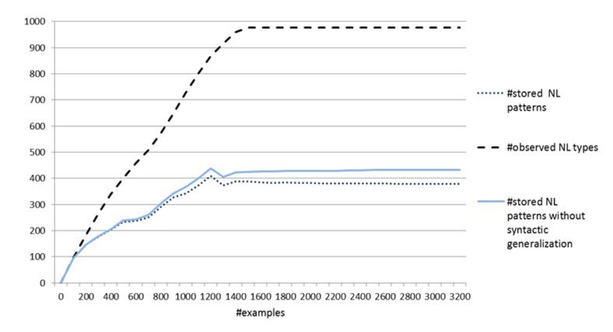
Evaluation

Grammar	#times training data was seen	F_1 (%)
Rote Learning	1, 2 or 3	16.3
Our model	1	77.5
Our model	2 or 3	84.3

Learning over time (1)



Grammar Size



Limitations

Nice model, explains many empirical phenomena (fast mapping, cross-situational learning, item-based nature of learning, etc.)

Main problem: symbolic input

Relax this assumption:

- relying on (perfect) phoneme sequences (it works!)
- relying on imperfect phoneme sequences (hmmm...)
- This talk: from symbolic actions (logical predicates) to subsymbolic representations of action!

Representation of actions

A semantic representation of 'to pass' needs to include:

- Participants and their roles
- Pre- and post conditions of the action
- Spatio-temporal signature of the action (rel. position of objects, arms, hands, ...) over time
- Purpose of the action

Research questions

- How can we represent actions so that we can calculate how similar two actions are?
- Are these representations suitable in order to discriminate between actions?

Conceptual Spaces

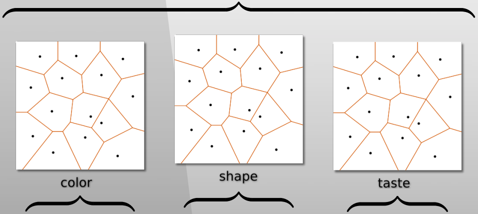
Inspired by 'Conceptual Spaces' (Gärdenfors, 2004)

- Geometric framework
- Concepts are convex regions in a vector space
- Cognitively plausible (prototype effects)
- Lends itself to computational implementation

Conceptual Spaces (2)

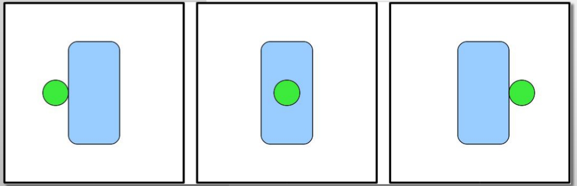
Concept:

Domains:



Properties:

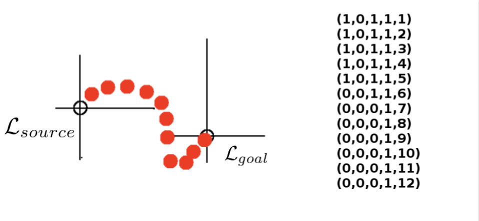
Actions in Conceptual Spaces



A trajector is moved across a landmark.

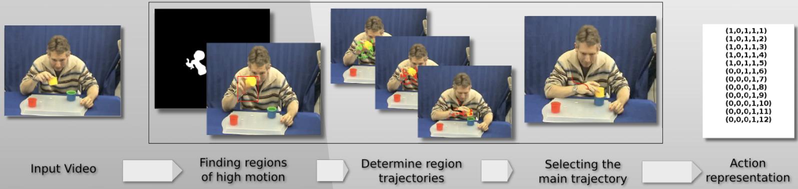
Idea: represent actions by their prototypical spatio-temporal signature!

Action representations



Discretized representation that is very far from a symbolic predicate, at the same time abstracting from a lot of low-level variation!

Trajectory Extraction



Dataset

Subset of Motionese video dataset (Rohlfing et al. 2006)

- Explicit tutoring in adult-child interaction
- 8 actions categories ('put', 'pull', 'open', 'shut', 'switch', 'place', 'close', 'push')
- 19 examples of each category (152 total)

Data manually annotated

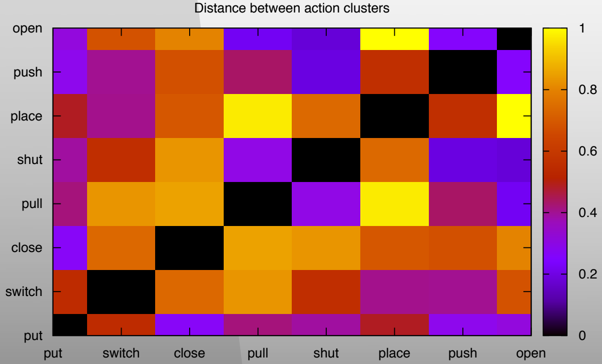
Experiments

1. Experiment: How well are actions represented in our action space?
   - Clustering with k-Median
   - DTW for sequence comparisons
2. Experiment: Are our representations reasonable in order to discriminate actions?
   - Clustering Purity
   - Classification Accuracy (1-NN, 10-fold cross validation)

Results of Classification

k	Action Categories	Accuracy of Classification	Purity of Clustering
2	pull, place	100.00	94.74
3	pull, place, close	95.00	78.95
4	pull, place, close, open	66.67	65.79
5	pull, place, close, open, switch	57.14	53.68
6	pull, place, close, open, switch, shut	44.44	45.61
7	pull, place, close, open, switch, shut, push	40.00	42.86
8	pull, place, close, open, switch, shut, push, put	35.71	36.84

Distances between action clusters



Summary

- Prototype-based representation of actions based on main trajectory of moved object
- Reliable discrimination of a few action categories
- Issues of our approach (auto. trajectory extraction, annotations, limited representation)
- Clearly: this representation is not enough for the purposes of representing the grounded semantics of action verbs
- Btw. we could have made it very easy for ourselves ;-)

Conclusions

- Promising model for language acquisition
- Moving now to subsymbolic input at speech (difficult!)
- Moving now to subsymbolic input of observed actions (more difficult)
- Representation / grounded semantics of actions should encompass:
  - spatio-temporal signature (we have provided a first try!)
  - pre- and post-conditions (discretization, image schemas?)
  - participants and their roles, also to anchor linguistic participants
  - teleological, intentional structure of actions

Puhhh...

Long-term goals

- Understanding language acquisition better
- Understanding representations (of action) better
- Developing machines / robots that learn to understand us (developmental approach rather than blueprinting)

Is Steve passing the salt to Elena?

CITEC