Language Acquisition

Andreas van Cranenburgh
Arjan Nusselder
Nadya Peek
Carsten van Weelden

Universiteit van Amsterdam

June 21, 2007

Overview

- Background
- Hypothesis
- Simulation overview
- Schedule
- Corpus Annotation
- One Word Implementation
- Two Word Specification
- Two Word Implementation and Future

Background

How do we acquire language?

- Competence vs. Performance
  - Innate grammar (Chomsky et. al.)
  - Analogy-making process (Scha et. al.)

Innate grammar does not account for many phenomena such as the comprehension of ungrammatical sentences. Innate (generative) grammars do however produce grammatically sound incomprehensible constructions such as *I told him to tell him to tell him to tell him to say*.

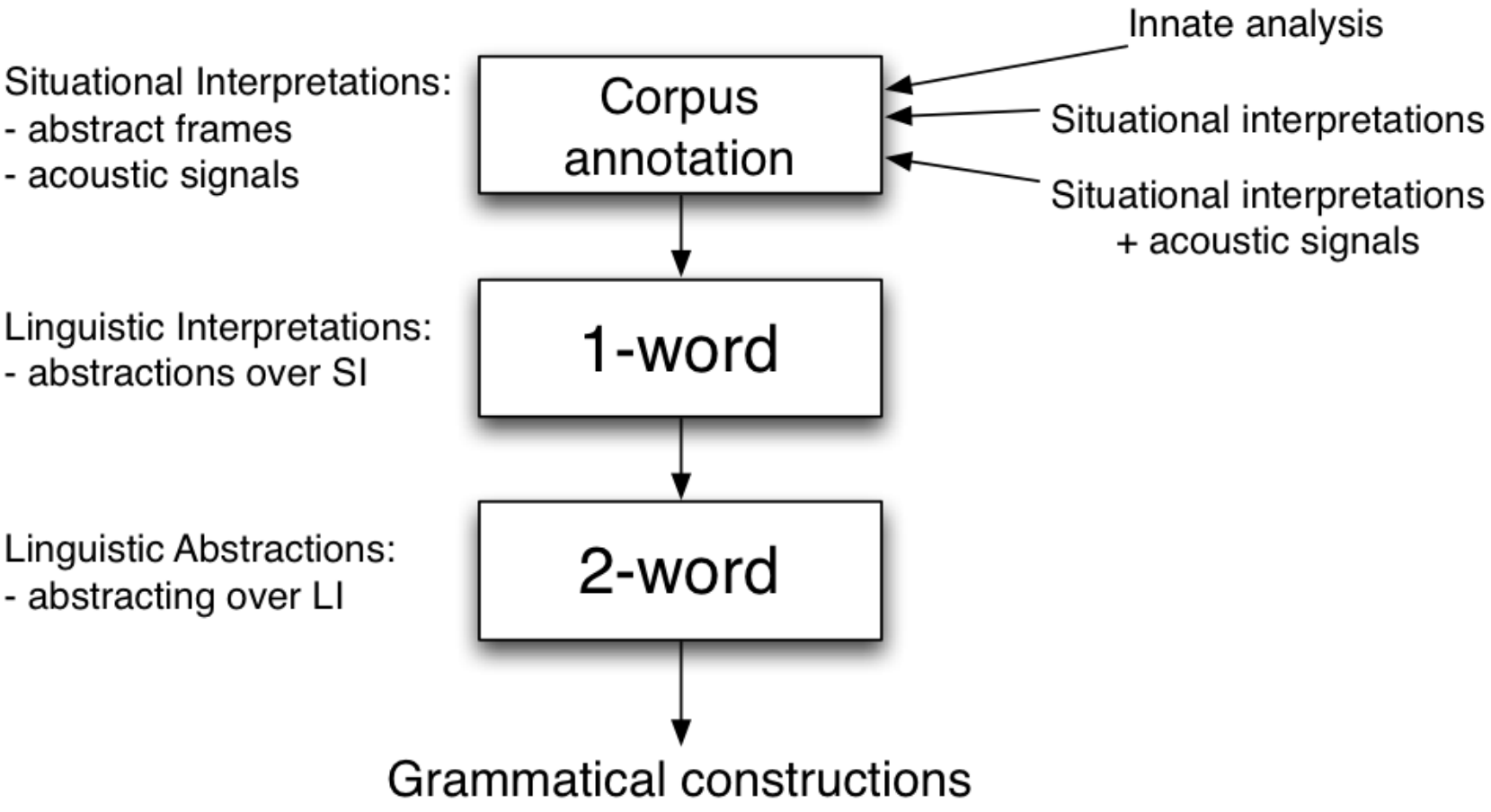
Hypothesis

If a child learns language by basing new grammatical structures on analogy with structures he has seen before,
How does a child acquire his first corpus?

Stages of Language Acquisition

- Pre-linguistic stage: categorizing specific situations into abstract situations.
- One-word stage: associations of words with instances of Humans, Objects, Actions, Locations, etc.
- Two-word stage: formation of first grammatical constructions.
- Multi-word stage: formation of three+ word sentences.

Cycle overview



Schedule

Until now, we have remained on schedule. We foresee that annotating the corpus may turn out to be more work than we hoped. Nevertheless we expect to be able to remain on schedule.

deadline	activity	remarks
Week 1	Background research	done
	Planning	done
Week 2	Implementation One-word stage	done
	Specification Two-word stage	done
Week 3	Implementation Two-word Stage	
	Annotating Corpus	
Week 4	Debugging	
	Report and Final presentation	

Corpus Annotation

To be able to simulate a child extracting structure from its surroundings, we need a corpus which simulates a child's interactions with its surroundings.

CHILDES (Child Language Data Exchange System) is an online corpus of child-adult conversations.

*FAT: <show me> [/] show me a few things in this picture .
*FAT: what's happening here ?
*CHI: happening there .
*CHI: gettin(g) on the other side .

The corpus should be annotated with more complete situational descriptions, the frames. Frames should have rolls and contain an ID, abstractions, other frames and properties.

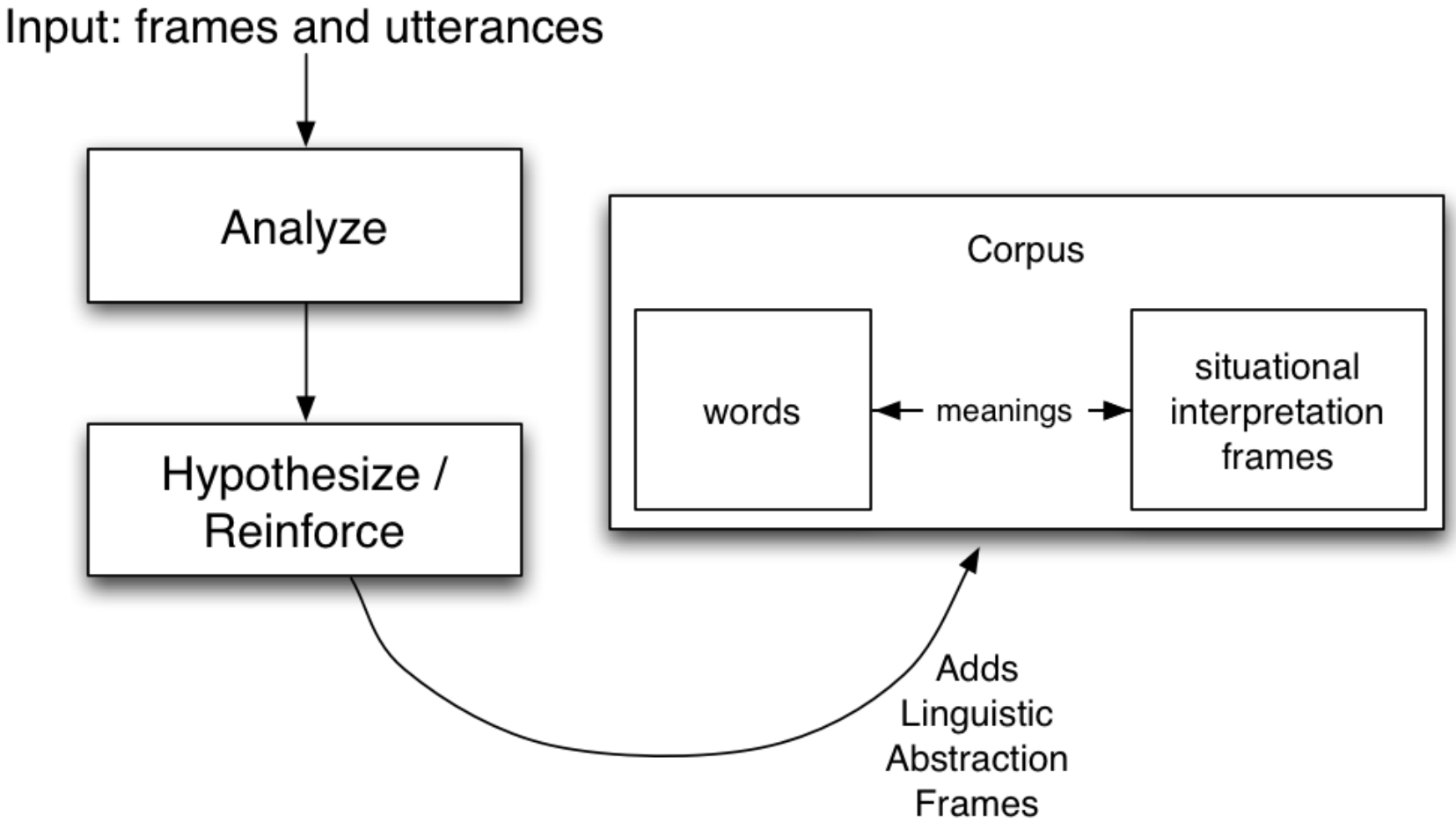
One-word Implementation

- *language*: Python
- *datastructure*: XML
- *scoring*: todo
  - associate frames with words using statistics
  - cut of words like “the,” “that” etc.

oneword.py example output

```
$ python oneword.py
Talk to me:  throw the block
throw
MEANING:
     FRAME: who
          ABSTR: object:human
          ID: child
          PROP: family = yes
-------------------------------------------------------------------
the
MEANING:
     PROP: family = yes
-------------------------------------------------------------------
block
MEANING:
     PROP: shape = square
```

Two-word Specification
Overview



Two-word Specification Cont.
Algorithm specification

Analyze: receives a situational interpretation with an utterance. It searches for the most probable meaning compatible with the SI. It then searches for a linguistic abstraction.

- Reinforce: if an appropriate LA is found, its occurrence is reinforced in the corpus.
- Hypothesize: if no appropriate LA exists, we must create a new one.

Two-word Specification Cont.
Creating a new LA

The Hypothesize part of the 2-word stage is a little tricky. However once a child *expects* to learn, he will have more ease in making more Linguistic Abstractions.

Hypothesize (Utterance, Word-Meaning-List):

- Find connected words, where one meaning is a subframe of another.
- Make an abstraction of the subframe. (e.g. balls can be red, yellow or blue)
- Add word-order.

Output: [throw BEFORE Var]

Two-word Implementation

Next week we will work on annotating the corpus and implementing the two-word stage.

The two-word specifications have already been written in pseudo code, so we hope we can implement the two-word stage in the third week, and will have time to run through multiple corpora for both the one- and two-word stage in the fourth week.

This of course will require a lot of corpus annotation.