1 Simulating Language Games of the Two Word Stage

Bachelor project proposal, Andreas van Cranenburgh (0440949), March 2009

1.1 Introduction

General linguistics has been dominated by Chomskian generative linguistics for several decades. The focus is on rules and their creativity, viz. systematicity and productivity. The central dogma is that an in-born, Universal Grammar is necessary to adequately explain these phenomena. It holds on to the continuity assumption, which states that language as used and understood by children is qualitatively equal to that of adults (Tomasello 2003).

However, from a developmental psychology angle, several empirical findings (Tomasello 2000, 2003) shed doubt on whether this approach is applicable to language acquisition by children. It rather appears that language learning is bootstrapped in a haphazard fashion, learning constructions here and there, which can only later be synthesized to form a coherent grammar.

Rather than trying to resolve this age-old debate between rationalism and empiricism along theoretical lines, it might be fruitful to try to model the behavior of early language users, and demonstrate in this way that a universal grammar is in fact not necessary to explain the phenomena observed. This strategy echoes a suggestion made by Turing (1950):

> "Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? [...] Presumably the child-brain is something like a notebook as one buys it from the stationers. Rather little mechanism, and lots of blank sheets."

1.2 Previous work

One of the foremost proponents of the developmental take on language acqusition is Tomasello (2003). He argues that linguistic abilities are acquired gradually, in an incremental fashion. Linguistic forms are memorized in conjunction with their communicative functions or meanings. These constructions are then generalized so that language use becomes ever more expressive and productive. Aspects which distinguish this approach from that of generative linguistics is the rejection of the autonomoy of syntax and the consequential focus on semantic and pragmatic influences on learning. Aside from that the idiomatic dimension of language presents problems for purely formal accounts of semantics and syntax, so a certain informality should be embraced by models of language.

In a previous project three students and I (van Cranenburgh 2007) attempted to model the two word stage of early child language. The model used a corpus of utterances spoken to children, annotated with semantic representations of the context. The aim was for this informal model to be able to generalize over the sentences to discover the correct associations between words and their semantic representations, and to be able to combine sentence fragments into novel utterances. This model did not consider syntax and semantics separately, in the style of construction grammar (Tomasello 2000, 2003). Although indeed correct associations were found, and novel utterances could be recognized, most of the former were incorrect, and most of the latter non-sensical (although in part this was due to the first issue worsening the second). Here is an example of an utterance as it was interpreted by our model:

```
1. "ball gone"  la score = 1
LINGUISTIC ABSTRACTION:
        WORDORDER: VAR:gone
        FRAME: action
                ID: action:move
                FRAME: object
                        ID: VAR
                        ABSTR: object:toy
```

In this sentence the construction "X gone" was applied to "ball", because it matched the condition of being a toy. The construction was apparently previously encountered when a toy was being moved.

The problem was that sentences were being learned as isolated fragments, without any notion of discourse or pragmatics. Also, the semantic representation did not fit well with all the words to be learned: it was good at representing actions and objects, so prepositions and demonstratives and other abstract words were not being learned. Instead of merely focusing on semantically describing a situation, the learner should consider the total communicative function of an utterance. The learning was implemented as making associations between words and all possible parts of the semantic representation, and counting how often these associations were made. This meant that a lot of incorrect associations were made. Unfortunately the model did not make use of pruning, so these incorrect associations were being retained.

Last year another project (Odolphi 2008) developed a formal grammar for the two word stage, based on empirical work on child language (eg. van Kampen 2003). This grammar does not make use of adult-like syntactic categories such as verb and noun, but groups expressions as topics, comments and operators. Using this grammar it is possible to produce plausible child utterances, because it turns out that the almost all of the two word utterances follow the pattern of this formal grammar.

The work of van Kampen (eg. 2003) on children's' use of languages in the two word stage indicates that their (proto-)grammar employs pragmatic operators and content signs, instead of distinguishing all the syntactic categories present in adult language. Verbs are not yet inflected, and determiners are absent.

Chang & Gurevich (2004) demonstrate a computational model of Embodied Construction Grammar that combines constructions to interpret new constructions. Their semantic representation could serve as an inspiration for how to improve the semantic representation. Also, the use of Minimum Description Length learning provides a good way to prune the database of learned constructions.

Steels (2004) describes his experiments with situated agents that employ language games as a learning strategy. An example of a language game is the description game: one agent describes an event that has just happened, and the other responds by agreeing if the description matches its own experience.

Van Kampen & Scha (2007) discuss the modeling of early syntax acquisition using the Data Oriented Parsing framework. This means that all input is stored in memory, and made available for recombination in the recognition of novel utterances.

1.3 Research question

Can an examplar-based model of language acquisition account for the discrepancy between language comprehension and production of children in the two word stage? Can this model facilitate the simulation of simple language games of parent and child?

These questions will be addressed by attempting to implement a simple model of linguistic comprehension and production using an exmplar-based model of language.

1.4 The model

To answer this research question, a model will be devised. The model will start with a collection of concepts and interpreted constructions. Concepts are grouped as referents and predicates. Constructions are multi-word utterances with a (possibly partial) interpretation of their meaning. Allow user of program to act as parent by specifying a situation with an action and attention focusing, and making an utterance. Then the program produces a child reaction, possibly using the mentioned two word grammar. Eg. "throw the ball" and child reacts by acknowledging or refusing, or by picking up the ball. The focus of the model will be to model interactions, not the understanding and production of single utterances.

The model will make use of a database of examplars, storing all linguistic input and associated situations. The discrepancy of speech comprehension

and production abilities can be simulated by employing differing algorithms for comprehension and production. Comprehension should attempt to find the meaning of an utterance by combining any possibly relevant examplars together. Production can tuned to proceed conservatively. Responses will be generated using a very limited form of imitative creativity, perhaps informed by the grammar of two word stage. This discrepancy makes sense because the child might lose the attention of its parent by saying incomprehensible or irrelevant things, and on the other hand it is clear that before starting to speak children have been listening to language for some time. Language has been trickling in, but constructions that can be reproduced presumably need a certain critical mass.

An idea for indexing the examplars is to use perceptual hashes. These are like normal hashes in that they compress the input with a high loss of information, but different in that similar inputs yield similar hashes, so that there is an effective way to compare the similarity of exemplars.

The last part will be to evaluate the model. Simple ideas are to judge whether the model performs better than parrot behavior, or better than through simple conditioning. A more elaborate way could involve a kind of Turing Test: presenting real and simulated parent-child dialogues to people and establishing the recognition rate.

1.5 Plan

Breakdown of work to be done (12 weeks, the last 4 of which will be full time):

- 2 weeks: A corpus collected from selected fragments of available corpora (eg., CHILDES)
- 2 week: Devise frame-based semantics formalism, and define speech act operators / language games
- 3 weeks: Annotate the corpus using this formalism
- 4 weeks: Implement analyses of adult utterances, implement response generation
- 1 week: Evaluate cognitive plausibility of these responses.

1.6 References

Chang, Nancy & Gurevich, Olya (2004) "Context-Driven Construction Learning." Proceedings of the 26th Annual Meeting of the Cognitive Science Society. Chicago.

van Cranenburgh, Andreas, Arjan Nusselder, Nadya Peek & Carsten van Weelden (2007): Towards a Computational Model for Early Language Acquisition, 2nd year bachelor of AI project. `https://unstable.nl/andreas/ai/2p/laac/verslag.pdf`

Tomasello, Michael (2000) "The item-based nature of children's earl syntactic development," Trends in Cognitive Science, Vol. 4, No. 4 (April 2000), pp. 156-163

Tomasello, Michael (2003), "Constructing a Language. A Usage-Based Theory of Language Acquisition," Cambridge MA: Harvard University Press.

van Kampen, Jacqueline (2003) "The learnability of Syntactic Categories," Proceedings of GALA.

van Kampen, Jacqueline & Scha, Remko (2007): "Modelling the Steps of Early Syntax Acquisition," Proceedings of the Workshop Examplar-Based Models of Language, Dublin.

Steels, Luc (2004): "Constructivist Development of Grounded Construction Grammars," Proceedings of the 42nd Annual Meeting of the ACL, Barcelona.

Odolphi, Ernst (2008): "Formal Grammars for Early Child Language," BSc. thesis, Artificial Intelligence, University of Amsterdam.

Turing, Alan (1950): "Computing Machinery and Intelligence," Mind LIX, no. 2236 (Oct. 1950): 433-60