Research Proposal

'Towards an experience-based model of early syntax acquisition'

4. Previous and Future Submissions

None

5. Institutional Setting

The research program involves the cooperation between the Institute for Logic, Language and Computation (ILLC, UvA) and the Utrecht Institute of Linguistics OTS (UiL OTS, UU).

6. Period of Funding

Postdoctoral research: 3 years
PhD project 1: 4 years
PhD project 2: 3 years
Date: 01-01-2007 till 31-12-2010

7. Composition of the Research Team

Name	Specialization/function	Affiliation
a. Main applicant		
Prof. dr. ir. R. Scha	computational linguistics	ILLC, UvA
b. Other applicants		
Prof. dr. P.H.A. Coopmans	first language acquisition	UiL OTS, UU
c. Postdoctoral researcher		
dr. N.J. van Kampen	first language acquisition	UiL OTS, UU
d. Advisors		
Prof. dr. M.J. Moortgat	computational linguistics	UiL OTS, UU
Prof. dr. E.J. Reuland	theoretical linguistics	UiL OTS, UU
dr. K. Sima'an	data-oriented parsing	ILLC, UvA

8. Structure of the Proposed research

1. Postdoctoral project
Title: 'The steps of first language acquisition'
Institutional environment: Utrecht Institute of Linguistics OTS (UiL OTS, UU)
Postdoctoral researcher: dr. N.J. van Kampen

2. PhD project 1
Title: 'The development of context structure for lexical items in first language acquisition'
Institutional environment: Utrecht Institute of Linguistics OTS (UiL OTS, UU)
Promotors: Prof. dr. M.J. Moortgat, Prof. dr. ir. R. Scha
Co-promotor: dr. N.J. van Kampen

3. PhD project 2
Title: 'A computational model of step-wise language acquisition'
Institutional environment: Institute for Logic, Language and Computation (ILLC, UvA)
Promotor: Prof. dr. ir. R. Scha
Co-promotor: dr. N.J. van Kampen

9. Description of the Proposed Research

A. Theoretical background

Language acquisition theories

During first language acquisition the child builds up a grammar on the basis of a finite number of child-directed sentences. Present-day research offers two fundamentally different views on this process. The principle-based approach of most generative grammarians maintains that grammars are too abstract to be induced by toddlers from a few simple sentences. Grammar should rather follow from a genetically given innate matrix (Universal Grammar, Chomsky 1981). In opposition to the generative tradition, there is the "usage-based" approach. It assumes that the abstractness of grammar is an invention of linguists, devoid of psychological reality. For instance, Tomasello (2003) argues that our knowledge of language is less abstract, but more situation-oriented and "construction-based".

The present proposal will demonstrate the viability of a third point of view. We recognize that early child language is indeed more construction-based than rule-governed, but we also notice that it does not take long before the grammatical categories and principles postulated by generative grammar start to manifest themselves. Yet, the shift towards abstract grammar does not need the postulation of a grammar-specific genetic endowment. It is possible to analyze child language as a series of successive intermediate grammars G_o….G_i….G_n. Abstract principles grow within the succession of grammars. In this view, an acquisition theory should not try to explain directly how the input data could motivate all the abstract ins and outs of the final outcome G_n. Instead, it should explain how the input data give rise to each of the intermediate steps in an incremental development sequence.

Before we describe our research proposal, we summarize some important findings from Van Kampen (1996, 1997) which constitute our point of departure.

Language acquisition steps

The series G_o….G_i….G_n starts with a syntax-less language: the "one-word stage". The transition to the "two-word stage" can be best explained if we assume that the child ignores almost everything in its input sentences except the words it already knows, and new words that are explicitly stressed. This leads to the first "syntax", which combines two content words in a topic-comment structure. In the next stage, these two-word structures are nested: *"beer slapen"* + *"moet slapen"* → *"beer [moet slapen]"* or *"moet [beer slapen]"* (*"bear sleep"* + *"must sleep"* → *"bear [must sleep]"* or *"must [bear sleep]"*). The first nestings are probably triggered when the child extracts similar sequences from adult sentences by ignoring all words and suffixes she doesn't understand. The structures in this early multi-word stage have a property in common. They can be analyzed as topic-comment structures where the comment is marked by an element that in the adult language is a copula, modal or finite auxiliary (for short <+fin>-marking). Note that at this stage two (semantically equivalent) syntactic structures occur side by side in the child's language: marking the predicate (*"bear [must sleep]"*), and marking the predication as a whole (*"must [bear sleep]"*). Both are seen here as "predicate marking". The tension between these two alternatives may lead later on to syntactic movement.

After predicate marking, the next stage sets in. Most "quasi-names", used as arguments of the predicates, now get adorned by articles (D^o<±definite>) or get replaced by personal pronouns (D^o<+pro>). After predicate marking (<+finite>-marking) and argument marking (D^o-marking) have been mastered in principle, around the third birthday, new constructions and grammatical concepts start to arise at a much faster rate. Child language acquisition has been slow so far, but now it finally begins to show its proverbial speed in expanding phrases (in Pinker's 1994: 269 description "all hell breaks loose").

An acquisitional problem

Many human languages make use of grammatically marked words and phrases that have a double dependency. Examples from Dutch and English are question words (<+wh>), finite verbs (<+finite> and negation (<+neg>) elements. They function as sentence-level operators, but at the same time they fill a particular sentence-internal (predicate or argument) position. For instance, question words occur overtly in operator position, whereas the sentence-internal position is filled with a co-indexed "trace" ($t_{<+wh>}$) left behind by movement transformation (1)a. Negation elements, by contrast, occur overtly in the sentence-internal position, whereas the operator position is empty, see (1)b.

(1) a. *welk* boek heb jij $t_{<+wh>}$ gelezen?
(which book have you read?)
b. $operator_{<+Neg>}$ jij hebt dit boek *niet* gelezen
(you have not read this book)

In other languages, the same structure is sometimes expressed by a double occurrence ("concord") of the scope-bearing word. Interestingly, such syntactic doublings may also appear spontaneously in child language, even when the adult language makes no use of doubling. Syntactic doublings, like other distance markers in grammar, may be seen as manifestations of underlying structure. Examples are given in (2). The dependencies in (2) have been analyzed as precursors of the movement analyses in (3) (Jordens 1990). Three child language cases of explicit doubling are given in (4).

(2) a. ik *doe/ga* [ook praten] (overuse of *doen/gaan*) (S.2;5)
(I do/go also talk)
b. *welke* ga jij [liedje zingen] (partial wh-scope) (S.3;7)
(which willyou song sing?)

(3) a. ik praat ook $t_{<+fin>}$

b. welk liedje wil jij $t_{<+wh>}$ zingen.

(4) a. die is *niet* [ei *niet*] (scope of the neg-word) (S.2;5.22)
(this is not egg not)
b. *niemand* wil [*niet* met mama spelen] (scope of the neg-word) (L.4;3)
(nobody wants not with mummy play
c. jij moet zien *heb* ik [een nieuwe fiets *heeft*] (scope of the auxiliary) (E.3;1)
(you must see $have_{<+agr>}$ I a new bicycle $have_{<+agr>}$)
d. *in welk huis* denk je [*waar* jij woont]? (scope of the wh-word) (S.4;10)
(in which house think you where you live)

The <+neg>, <+finite> and <+wh> positions within the brackets are interpreted as more sentence-internal, those outside the brackets as operators.

The temporary syntactic doublings in Dutch child language offer a welcome challenge to modeling language acquisition as a stepwise procedure that follows from a preceding data-reduction (Van Kampen 2006). Where do the doublings come from? We propose to study this question by simulating the acquisition of syntax in a computer program. The program must start modeling the child's natural reduction of the adult input. At first all grammatical markings are left out. Fusing the

remnants will yield certain doublings. Subsequently, there is a natural order of acquisition steps in which these massive reductions and mismatches are successively overcome.

This hierarchy of acquisition steps is not only an issue relevant for the study of language acquisition. It is also relevant for typological issues and the general design of grammatical forms. Many typological properties are known to have implicational relations (Jakobson 1942, Greenberg 1963). In our view such implicational relations are natural consequences of the stepwise acquisition procedure. The earlier acquisition steps set the stage for later ones that now follow by a (conditioned) application. It is nevertheless remarkable, be it "natural", that grammatical systems allow an acquisition procedure due to a decoding procedure based on systematic simplifications. In the background of our proposal stands the following. It has been claimed in Koster (1987), Neeleman & Van de Koot (2002), Zwart (2006) that grammatical relations are based on a local and binary structure in all human grammars, essentially Koster's Configurational Matrix. This invites the conjecture that the configurational matrix survives in all grammars because it is a formal condition that guarantees the natural learnability of the system. It needs no control from genetic endowment specific for language.

B. The research proposal

General description

In the language acquisition process as we conceptualize it, the child invents a sequence of grammars. She starts with rigorous, but systematic simplifications. Subsequently, and in a stepwise fashion, the child approaches the grammar of the adult environment. Each acquisition step is characterized by a) a new intermediate grammar and b) a corresponding input reduction strategy that selects the new data for the next acquisition step. We will demonstrate the viability of this perspective by a computational model that simulates our observations about actual child language. The research program will therefore intertwine three strands of inquiry: (i) empirical data analysis, (ii) grammatical system analysis and (iii) computational modeling. T

The starting point of the research program will be the identification of stages in the development of Dutch child grammar on the basis of CHILDES files (MacWhinney 1991). For each intermediate stage, the new grammar and the new data-set will be precisely described. The computational model should then show how a new intermediate grammar arises due to the new data selection. The algorithm that develops syntactic hypotheses about new input will be based on general machine learning principles (Michalski et al. 1983, Langley 1996, Mitchell, 1997) instead of invoking innate linguistic knowledge. A pilot study (Obdeijn 2004) has shown the feasibility of this approach.

The system's syntactic analysis component will be based on the Data-Oriented Parsing (DOP) approach (Scha et al. 1999), a probabilistic model that represents its syntactic information by storing observed input-constructions as well as abstractions from them. DOP thus embodies a synthesis between the usage-based and the rule-based approaches to language. Although it is often presented as a purely usage-based approach (e.g. Scha, 1992), the development of the probability distributions in a DOP system can in principle model how new rules start their life as collections of specific constructions, from which more abstract representations gradually emerge.

In the final phase of the research program, we will consider the implications of our input-based computational model for some fundamental theoretical questions. We will be able to observe to what extent the child's grammar, in the various stages of its development, consists of specific constructions, and to what extent it employs abstract rules. As a result, we hope to show how the surface reduction simplifies the processing costs of the scope-bearing items and how the doublings in (4) are a temporary concession to the construction-based approach.

Innovation

In his recent work, Chomsky (2005) has suggested that Universal Grammar may be equated with the human capacity to deal with recursive structures in general. This reopens the language acquisition question. If UG merely provides our capacity for recursion, all further syntactic complexities must be due to the way our learning procedure deals with the actual input. Adequate models of this process are dearly lacking.

The model we intend to develop will differ from earlier attempts (Berwick 1985, Sakas & Fodor 2001, Daelemans & Van den Bosch 2005) in several ways. Our model

- emphasizes the step-wise nature of the language acquisition process.
- emphasizes the role of the child's input reductions, and models them explicitly.
- does not induce a grammar from "flat text", but projects its initial structures from semantic properties of the child-directed speech.
- exploits the Data-Oriented Parsing approach to account for the complete continuity between usage-based and rule-based knowledge of language.

Note that all these innovative properties of our model are strongly motivated by empirical observations concerning actual child language development.

Quality of the research group and (inter)national cooperation

The research program involves a cooperation between the UiL OTS and the ILLC. They bring complementary expertise to an area of common concern.

The UiL OTS is an internationally recognized center of excellence in the area of generative grammar and language acquisition. In particular, our research program will build on innovative empirical investigations of language acquisition by Van Kampen in the generative framework, and on Moortgat's expertise in Categorial Grammar.

The ILLC is an internationally recognized center of excellence in logical semantics and computational linguistics. Our research will crucially build on the expertise about probabilistic language processing in Scha's "Language and Computation" program. The ILLC expertise in logical semantics and formal pragmatics will be brought to bear on corpus annotation and the study of sentential operators in early child language.

At the international level contacts have been established with other researchers that use computer simulations to model language acquisition, such as Culicover (Ohio University, Culicover & Nowak 2003) and the CUNY group around Fodor (Fodor 2001). We also will take into account the findings of the Edisyn project (Barbiers, Meertens Instituut), concerning the typological spread of doubling constructions.

C. Description of the projects

General description

The program as a whole is concerned with the explanatory value of the successive acquisition steps in language acquisition. It will consist of a linguistically oriented post-doctoral project and two PhD projects. Their relation can be described as descriptive (postdoctoral project), formal theoretical (PhD project 1) and computational (PhD project 2). The overall aim is to articulate a theory of first language acquisition that will be implemented as a computational model and tested on Dutch child language. The acquisition theory we have in mind emphasizes the step-wise nature of language acquisition. Our model will specify an acquisition procedure that initially applies drastic reductions to the child-directed speech and derives the order of acquisition steps by which the child overcomes these reductions.

The *postdoctoral project* (Van Kampen, UiL OTS) will be mainly concerned with the grammatical characterization of the successive reductions and acquisition steps and their quantification in the files of spontaneous child language conversations. This part of the project will be supported by much work from the previous five years. The postdoctoral project should guarantee the empirical basis in actual child language of the more formally and computationally directed PhD studies. The model-building part of the two PhD studies must reproduce the succession of the child's acquisition steps.

PhD project 1 (supervision Moortgat UiL OTS and Scha ILLC) will describe the successive languages of the Dutch child by means of lexicalized grammars using the framework of Categorial Grammar. Child-utterances in the CHILDES corpus will be annotated in terms of these grammars, thus testing and demonstrating their adequacy. (Child-directed speech will also be annotated, in largely semantic terms.) The outcome of this project is a description of the language development process as the growth of a system for lexical categories and their context conditions.

PhD project 2 (supervision Scha ILLC) builds a computational model of language acquisition as a parser that updates its grammar on the basis of (reduced) child-directed speech. This model will be tested on the annotated part of the CHILDES corpus. The existing implementations of DOP are based on Stochastic Tree Substitution Grammar. For the first stages of language acquisition, this formalism seems immediately adequate. It may also be able to deal with 'movement-phenomena' by employing slash category systems as in Categorial Grammar (Moortgat 1997, 2001); see also the restatement of Minimalism by Neeleman & Van de Koot (2002). It is also possible to extend the DOP-formalism to allow movement of constituents (as suggested by Scha 1992)).

All projects will use the same dataset from CHILDES (Kampen/Groningen/Utrecht corpus) and apply the same longitudinal method and the same grammatical analyses. Interaction between the linguistic and computational sides of the research program will be crucial. For example, data selections from the adult input postulated to be possible by the linguistic branch might run into actual problems when implemented by the computational model. The reverse type of problem is expected as well. Steps quite easy for the computational model might turn out to be delayed or avoided in the real world. Both types of discrepancies appeared in the pilot project (Obdeijn 2004) and they necessitated revisions in one of the research components.

PhD project 2 (Scha) will start a year later than the postdoctoral project (Van Kampen) and PhD project 1 (Moortgat/Scha), which will develop a first ordering and annotation of the CHILDES data.

Postdoctoral Project **(UiL OTS)**
'The steps of first language acquisition'

The postdoctoral research develops the grammatical data arrays and analyses that will pave the way for the computational modeling of successive Dutch grammars. This will in part be based on previous work (Van Kampen 1997, forthcoming). It implies the following tasks.

(i) Propose grammatical rules that reduce adult input sentences into actual Dutch child language. These rules must be computationally implementable.
(ii) Measure the growth of grammatical markings and lexical categories by longitudinal corpus methods.
(iii) Indicate successive steps of the acquisition procedure by plotting the emergence of new grammatical markings in the reduced child language, such as the appearance of finite verbs in predicates, the rise of articles on syntactic arguments, and the appearance of dummy subjects.
(iv) Find linguistically plausible reasons for the actual order of the acquisition steps
(v) Consider the quantitative conditions for the learnability of underlying order. Part of the early constructions in child language (wh-movement, finite verb placement) imply underlying structure and movement (in a transformational model) c.q. construction-specific category assignments (in non-movement Categorial analyses).
(vi) Analyze temporary doublings in child language as linguistically plausible precursors of dependency relations in the adult language and see later how the computational model could simulate the same doublings.

The timing of these tasks over three years will be as follows.

Timing of the Postdoctoral Project
1. In the first year, a preliminary analysis of the acquisition of finite verb placement, wh-movement and negation in Dutch and its implementation will be given. Using the earlier ideas of Van Kampen (1997), we will automatically be confronted with the six points listed above.
2. In the second year the focus will lie on dependency relations and the doubling phenomena in child language. They represent a challenge for the mental reality of the computational implementation.
3. The third year will be devoted to an integrated presentation of the results.

Method
The order of the acquisition steps can be shown by longitudinal graphs. The method of comparing longitudinal graphs was developed in Van Kampen (1997, forthcoming). It has shown to be effective and will be used again in the present research. It will be applied to the Dutch child-corpora in CHILDES (MacWhinney 1991). CHILDES contains extensive Dutch corpora (Groningen/Kampen/Utrecht corpus).

The postdoctoral line of research will have parallels in two PhD projects. Both will be directed at a formal simulation of findings in the postdoctoral project.

PhD Project 1 **(UiL OTS)**
'The development of context structure for lexical items in first language acquisition'

The first PhD project will develop a system of semantic/syntactic frames and features that can be assigned to lexical items, defining the grammatical category of the lexical item. The extensions of the category system must be able to reflect the child's progress in language acquisition.

It has already been argued by Gleitman from an acquisitional point of view that cognitive-semantic oppositions develop due to oppositions within the same category frame (Gleitman 1990 "syntactic bootstrapping"). That same perspective (lexical context frames) returns in Borer's (2005) "exo-skeleton" for lexical meanings and in Briscoe's (2001) "valency frames" in a computational approach to automatic parsing. Once the semantic/syntactic frame for *told* *me that* has been acquired, the lexical items *assured*, *promised*, *reported, warned* , etc. *me that* are in the same dimension and easily learnable. The project will spell out how the grammatical properties of locality and inclusiveness, characteristic of Categorial Grammar and of Minimalist conceptions in the generative enterprise, enable the child to build up a category context system that allows its lexical expansion. This implies the following general questions about the structure of the lexical item:

a) What do the initial and the final context specifications of a lexical item look like in a DOP environment?
b) How does the acquisition step from the less to the more advanced specification take place?

The initial distinctions in child Dutch will be tested by applying them to the data files in CHILDES. The PhD will reproduce and extend for Dutch the results reached in Briscoe (2001). The annotation for semantic/syntactic frames must reveal how language acquisition can be represented as an expansion of lexical categories.

The PhD researcher will start by a semantic/pragmatic annotation of child language and child-directed speech. An appropriate part of CHILDES will be annotated in this way. This will be done in cooperation with the postdoctoral researcher.

PhD thesis outline

1. 'The structure of lexical items'. This chapter introduces the syntactic, semantic and pragmatic dimension of lexical items in Categorial and Unification-based grammars.
2. 'The growth of lexical category specification'. Lexical categories in child language and the child's initial reduction of the input are cast in a formal representation.
3. 'Lexical extension based on a common theta-frame'. Borer's (2005) exo-skeletal verb meaning and the quantification of valency frames in Briscoe (2001) will be tested for the acquisition of Dutch.
4. 'Light verbs and fixed prepositions in child and adult language'. This chapter treats the child's strong preference for light verbs as explicit syntactic parts of theta frames.
5. 'Using CHILDES in a computational model'. This chapter investigates which amount of semantic/pragmatic annotation of child language and of child-directed speech as needed for DOP in an acquisition simulation.
6. 'The syntactic basis of the child's lexical growth' (quantitatively and qualitatively). This chapter returns to the two general questions about the structure of the lexical item.

PhD Project 2 **(ILLC)**
'A computational model of step-wise language acquisition'

The second PhD project will develop an implemented and tested computational model. The computational model encompasses:

(i) a syntactic analysis algorithm (parser) based on a generalization of the Data-Oriented Parsing approach.
(ii) an algorithm which determines, given a grammar and an input of adult sentences, the corresponding data selection for each new acquisition step.
(iii) a machine learning component which extends and adapts the given grammar to deal with new input.

The model will be tested on data from the CHILDES corpus to determine whether it predicts the transitions between subsequent intermediate grammars as they actually occur.

The PhD researcher will start by extending and improving upon Obdeijn's (2004) pilot study of data reduction and make use of a system for lexical categories.

PhD thesis outline

1. 'The architecture of a computational model of first language acquisition'. This chapter describes the overall structure of the model, consisting of the components (i),(ii), (iii) above.
2. 'Data-Oriented Parsing for first-language acquisition'. The standard DOP models are Stochastic Tree Substitution Grammars. Such models seem perfect for the earliest stages of grammar acquisition, but it is not obvious that they would deal adequately with double dependencies. This chapter investigates this issue, and develops, if necessary, a richer DOP model.
3. 'Input reduction strategies'. Van Kampen's insights concerning the child's input reduction strategies are reviewed in this chapter, and implemented in a computational model. Obdeijn's (2004) pilot model will be taken into account.
4. 'Grammar extension strategies'. This chapter extends the parser as designed in Chapter 1 so that it assigns structures to utterances containing previously unseen words, either by projection from a semantic annotation or by construing an analogy with other utterances in its database. This new "construction" is then added to the grammar. When several similar constructions appear, the "abstraction" which represents what they have in common, automatically emerges in the DOP grammar.
5. 'Tests with the CHILDES corpus'. This chapter treats the design of tests with the CHILDES corpus, the results of the tests, and an analysis of the results
6. 'Conclusion'. This chapter may touch on important theoretical issues such as: rule-based vs. usage-based description of grammar; child language vs adult language; movement-transformations in data-oriented parsing.

10. Work program

Postdoctoral Project

Year	Deliverables	Activities
1	• Article 'Simulating child language'	• Adding an annotation to a part of CHILDES (with PhD researcher 1) • Formulating a computational approach to the acquisition of finite verbs, wh-questions and negation
2	• Article 'The learnability of early syntax in stages' • Article 'The learnability and implementation of dependency relations'	• Analyzing the doubling phenomena in acquisition and the learnability of dependency relations • Designing tests with CHILDES (with PhD 2) • Writing of articles
3	• Chapters for monograph	• Organizing the workshop • Writing of chapters monograph (with members of the full research group)

PhD Project 1

Year	Deliverables	Activities
1	• Literature review • Corpus annotation	• Literature study on language acquisition. • Sketch of chapters 1-2 based on a set of examples • Adding a first annotation to a part of CHILDES. • Courses: following PhD courses
2	• Corpus annotation • Internal paper on 'Theta frames and lexical expansion'	• Adding an annotation to a part of CHILDES. • Reproducing Briscoe (2001) for Dutch • Writing pre-version chapters 1 and 2 • Courses: following PhD courses
3	• Internal paper on 'Light verbs in child language' • Chapter for monograph	• CHILDES quantifications for chapters 3, 4 and 5 • Writing of internal paper • Writing of chapter monograph (with Moortgat/Scha) • Courses: following PhD courses
4	• Thesis	• Thesis revisions chapters 1 and 2 • Writing chapters 3, 4 and 5 of the thesis • Final editing thesis

PhD Project 2

Year	Deliverables	Activities
1	• Literature review • Parser	• Literature study on language acquisition • Decide on a version of DOP to use; implementing required extensions • Design and implement input reduction strategy • Design and implement grammar extension strategy • Courses: following PhD courses
2	• CHILDES test results • Internal paper on 'DOP for first language acquisition'	• Design, carry out and evaluate tests with CHILDES • Writing pre-version of chapters 2 and 3 of the thesis • Courses: following PhD courses
3	• Chapter for monograph • Thesis	• Writing pre-versions of chapters 4 and 5 of the thesis • Thesis revisions • Writing chapters 1 and 6 of the thesis • Final editing thesis

11. Word Count

- General description proposed research: 1993 words
- Description of the three projects: 1832 words

12. Planned Deliverables and Knowledge Dissemination

Scientific

The results from the PhD projects will result in two doctoral dissertations, one at the ILLC and one at the UiL OTS. The results from the postdoctoral project and the combined results from the PhD and postdoctoral projects will be reported in a synthesizing monograph (see below) and in articles in international journals. Journals are available that focus on various aspects of this research:

- language acquisition (e.g. *Language Acquisition*, *Journal of Child Language*)
- cognitive science (e.g. *Cognition*, *Cognitive Scienc)*
- computational linguistics and A.I. (e.g. *Computational Linguistics, Jornal of Experimental and Theoretical A.I., Journal of Machine Learning Research*)
- theoretical linguistics (e.g. *Lingua*, *Linguistics, Syntax*)

The results will also be reported at various international conferences, that constitute scientific platforms for presenting and discussing the research topics (GALA, Conference on Computational Psycholinguistics, Meetings of the ACL)

A workshop will be organized in the third year in order to focus on the issues and make the results available to national and international researchers.

We will conclude the research with a synthesizing monograph that will discuss the general issue of language learning and its models and their application to constructions in early child language. The monograph will be co-authored by different participants in the research program. A tentative table of contents:

1. Introduction: An experience-based account of grammar acquisition. (Van Kampen/Scha)
2. The facts: The early stages of Dutch child language. (Van Kampen)
3. Formal grammars for the successive stages of Dutch child language. (Moortgat/PhD-student 1)
4. A computational model of early grammar development. (Scha/PhD-student 2)
5. Constructions or abstractions or both? (A theoretical reflection on the properties of the grammars that describe the different stages of Dutch child language.) (Van Kampen/Scha).
6. Describing double dependencies: slash-categories or movement? (Van Kampen/Moortgat/ Scha)
7. Implications for the description of adult grammar. (Van Kampen/Scha)

Software

The computational model developed at the ILLC will be implemented on a widely used platform and be provided with reasonably user-friendly interfaces and an adequate description of its operation. It will be made available online to the international research community.

The corpus annotation created at the UiL OTS will also be made available online to the international research community.

13. Short Curriculum Vitae Principal Applicant and Postdoctoral Researcher

CV Principal Applicant

Personal details

- Name: R.J.H. (Remko) Scha
- Name and place of birth: 15-09-1945, Eindhoven
- Nationality: Dutch
- Address: Institute for Logic, Language and Computation (ILLC)
  Plantage Muidergracht 24, 1018 TV Amsterdam
- Telephone: 020 5252075/5235
  E-mail: scha@uva.nl

University education

- Ph.D. in Computational Linguistics
  - Faculty of Letters, University of Groningen, the Netherlands, 1983
  - Promotors: Joyce Friedman and Frank Heny
- Engineering Degree in Physics
  - Technological University Eindhoven, the Netherlands, 1970
  - Specialization areas: computer science, information theory, auditory perception

Professional Experience

1988-present	Professor of Computational Linguistics (Faculty of Humanities) and Program Leader *Language and Computation* (Institute for Logic, Language and Computation). *University of Amsterdam,* the Netherlands.
1990	Visiting Professor. Linguistics Department, *Tel Aviv University,* Israel
1985-1988	Manager of Natural Language Group. Speech and Signal Processing Department / Artificial Intelligence Department, *BBN Laboratories,* Cambridge, Massachusetts, USA.
1984	Senior Scientist. Artificial Intelligence Department, *BBN Laboratories,* Cambridge, Massachusetts, USA.
1983-1984	Visiting Professor. Artificial Intelligence Department (SWI), Faculty of Social Sciences, *University of Amsterdam,* the Netherlands.
1975-1983	Research Scientist. Computer Science Department, *Philips Research Laboratories,* Eindhoven, the Netherlands.
1970-1975	Research Scientist. Advanced Systems Development Department, *Philips Electrologica,* Apeldoorn, the Netherlands.

Relevant research experience

- Remko Scha has extensive experience in applied and theoretical computational linguistics. He made internationally recognized contributions in the areas of logical semantics, discourse processing, and probabilistic syntax.
- At Philips' Research laboratories, Remko Scha designed the question-answering system PHLIQA1 in collaboration with Jan Landsbergen. At BBN Laboratories (Cambridge, Mass.), he developed the Linguistic Discourse Model in collaboration with Livia Polanyi. At the University of Amsterdam, he developed the Data-Oriented Parsing model, in collaboration with his Ph.D. students Rens Bod, Khalil Sima'an, and Remko Bonnema.

Short list of core publications

Bod, Rens, Remko Scha & Khalil Sima'an (2003) (eds.) *Data-Oriented Parsing* Stanford: CSLI Publications.

Scha, Remko, Rens Bod and Khalil Sima'an (1999) 'A memory-based model of syntactic analysis: Data-Oriented Parsing' *Journal of Experimental and Theoretical Artificial Intelligence* 11 (3), 409-440. (Special Issue on Memory-Based Language Processing, edited by Walter Daelemans).

Scha, Remko (1992) 'Virtuele grammatica's en creatieve algoritmes' *Gramma/TTT* 1 (1), 57-77. [English translation: http://iaaa.nl/rs/inaugureE.html]

Scha, Remko, Livia Polanyi & Bertram Bruce (1987) 'Discourse understanding', in: S. C. Shapiro (ed.) *Encyclopedia of Artificial Intelligence,* Vol. 1. New York: John Wiley and Sons, 233-245.

Scha, Remko (1981) 'Distributive, collective and cumulative quantification', in: J.A.G. Groenendijk, T.M.V. Janssen & M.B.J. Stokhof (eds.): *Formal Methods in the Study of Language* Part 2. Amsterdam: Mathematisch Centrum, 483-512. [Reprinted in: Javier Gutierrez-Rexach (ed.) *Semantics: Critical Concepts in Linguistics.* New York: Routledge, 2003.]

CV Postdoctoral Researcher

Jacqueline van Kampen is a postdoctoral researcher at Utrecht University. Her primary research interest is the acquisition of syntax and morphology of Dutch, French, German and English learning children. Her research combines the field of linguistic theory and learnability theory. This has brought about collaboration with people who work on formal learnability and on computer simulations (ILLC Amsterdam, CUBY New York). She is currently working on a forthcoming monograph 'Hierarchies of Learning Steps in First Language Acquisition'.

Peer-reviewed publications last three years

Evers, A. & J. van Kampen (accepted) 'Parameter setting and input reduction', in: M.T Biberauer and A. Holmberg (eds.) *The Structure of Parametric Variation*, 25 pages. Amsterdam/ Philadelphia: John Benjamins.

Kampen, J. van (to appear) 'An acquisitional view on Germanic finite verb agreement' *International Journal of Morphology*, 20 pages.

Kampen, J. van (2006) 'Early operators and late topic-drop/pro-drop', in: V. Torrens & L. Escobar (eds.) *The Acquisition of Syntax in Romance Languages*, 203-223. Amsterdam/Philadelphia: John Benjamins.

Kampen, J. van (2006) 'The acquisition of the standard EPP in Dutch and French', in: J. Costa & M.C. Figueiredo Silva (eds.) *Studies on Agreement*. 99-119. Amsterdam/Philadelphia: John Benjamins.

Kampen, J. van (2005) 'Subjects and the (Extended) Projection Principle', in: Keith Brown (editor-in-chief) *Encyclopedia of Language and Linguistics*, 2nd ed. volume 12, 242-248. Oxford: Elsevier Science.

Kampen, J. van (2005) 'Language specific bootstraps for UG categories', *International Journal of Bilingualism* 9-2, 253-277.

Kampen, J. van (2004) 'An acquisitional view on optionality', *Lingua* 114, 1133-1146.

Kampen, J. van (2004) 'Learnability order in the French pronominal system', in: R. Bok-Bennema, B. Hollebrandse, B. Kampers-Manhe & P. Sleeman (eds.) Selected Papers from *Going Romance* 2002, 163-183. Amsterdam/Philadelphia: John Benjamins (series: Romance Languages and Linguistic Theory).

14. Summary for Non-specialists (in Dutch)

Het is een intrigerende vraag hoe het mogelijk is dat heel jonge kinderen in relatief korte tijd een zo complex systeem als taal kunnen leren. Vaak wordt daarom gedacht dat taalverwerving alleen maar verklaard kan worden als het kind over aangeboren linguïstische kennis beschikt.

Het hier voorgestelde onderzoek probeert die veronderstelling te vermijden. Het wil een model ontwikkelen van taalverwerving. Het model gaat ervan uit dat de kinderen eerst de volwassen taal op een uniforme wijze vereenvoudigen en vervolgens in het leerproces die vereenvoudiging weer ongedaan maken in een vaste voorspelbare volgorde. Zowel de vereenvoudingen zelf als de volgorde waarin ze weer verdwijnen, kunnen gekarakteriseerd worden door een voorspelbaar algorithmisch proces dat met een computerprogramma gesimuleerd kan worden. Het programma kan laten zien in welke mate algemene leerstrategieën en taalkundige a prioris elhkaar kunnen steunen of vervangen. Het onderzoek richt zich op de verwerving van het Nederlands en is kwantitatief en computationeel geörienteerd.

Kinderen negeren aanvankelijk heel veel details van de uitingen die ze horen. Ze proberen vooral de woorden en structuren eruit te pikken die ze al kennen. Afhankelijk van het stadium van hun grammaticaontwikkeling krijgen ze daardoor telkens nieuwe samenhangen tussen de al bekende eenheden in het vizier, dankzij nieuwe grammaticale woordjes die zulke eenheden modificeren of markeren. Het model richt zich op 'selectieve aandacht' tijdens de taalverwerving. Een pilot project uitgevoerd door Obdeijn (2004) heeft laten zien dat fases van kindertaal reproduceerbaar zijn door een 'selective aandacht' te zien als een reductieprogramma toe te passen op (tot de kinderen gerichte) volwassentaal. De volgorde van leerstappen is vervolgens afleidbaar als die reeks minimale toevoegingen die maximaal effect sorteren in het benaderen van de volwassen taal.

Bij sommige constructies maakt kindertaal steevast een tijdelijk omweggetje. In de Nederlandse kindertaal is dat te zien bij ontkenningen, vragen en het gebruik van de 'persoonsvorm'. Er treden syntactische verdubbelingen op. Zie de voorbeelden in (1)-(2) (Van Kampen 1996, 1997).

(1)	a.	ik *doe* [ook praten]	(overuse of *doen*)	(S. 2;5)
	b.	*welke* wil jij [liedje zingen]	(partial wh-scope)	(S. 3;7)

(2)	a..	ik heb *niemand* [*niet* gezien]	(neg-word)	(S. 3;2)
	b.	die is *niet* [ei *niet*]	(neg-word)	(S. 2;5.22)
	c.	(jij moet zien) *heb* ik [een nieuwe fiets *heeft*]	(auxiliary)	(E. 3;1)
	d.	(ik heb voor jou wat gemaakt) *ben* jij [bijna jarig *is*]	(auxiliary)	(E. 3;1)
	e.	*in welk huis* denk je [*waar* jij woont]?	(wh-word)	(S. 4;10)
	f.	*hoe* denk je [*hoe* ik eet]?	(wh-word)	(S. 5;11)

Hier is steeds sprake van woorden die een dubbele afhankelijkheid hebben. Enerzijds karakteriseren ze de zin als geheel (ontkenning, vraag, hoofdzin), anderzijds is er een sterk plaatselijke afhankelijkheid binnen een gewoon zinsdeel (binnen de rechte haken). Bij dubbele afhankelijkheid is er sprake van een speciaal taalverwervingsprobleem dat ontweken wordt met een omweggetje. De omweggetjes zijn voorspelbaar. Het beoogde verwervingsmodel moet die tijdelijke verdubbelingen kunnen reproduceren.

Het onderzoeksvoorstel heeft drie oriëntatiemiddelen om de aard van de vereenvoudigingen en de volgorde van leerstappen in een verklarend model te brengen. In eerste instantie is er de *descriptieve component* die de vereenvoudiging en de volgorde van leerstappen documenteert in de files van spontane Nederlandse kindertaal (beschikbaar in de CHILDES databank). De reeks van grammaticale vorderingen moet vervolgens vertaald worden in een stelsel van categorieën voor lexicale eenheden waarbij elke categorie zijn eigen combinatieregels vastlegt. De ontwikkeling van kindergrammatica's als een ontwikkeling van lexicale categorieën en hun combinatie-mogelijkheden vormt het aandachtsgebied van de *theoretische component*. Tenslotte moet een

computationele component laten zien hoe de kindergrammatica's een automatische zinsontleder kunnen sturen die minder of meer aan kan naar gelang het categorieënstelsel in zijn combinatie-mogelijkheden verder ontwikkeld is.

15. Research budget

	2007	2008	2009	2010	TOTAL
Staff costs: (k€)					
Postdoc 3 years (1.0 fte)	55	57	59		171
PhD 3 years (1.0 fte)		38	42	45	125
PhD 4 years (1.0 fte)	38	42	45	47	172
Non staff costs: (k€)					
Experimental costs	2	2	1		5
Bench fees	3	4.5	4.5	3	15
Workshop			8		8
TOTAL	98	143.5	159.5	95	496

Experiments: The experimental costs cover software.

Workshop: In the third year an international workshop will be organized. The estimated costs include hotel and travel costs for invited speakers, costs for use of conference site, reception, program booklet, and student assistants.

References

Berwick, R. (1985) *The Acquisition of Syntactic Knowledge* Cambridge MA: MIT Press.

Borer, H. (2005) *Structuring Sense* Oxford: Oxford University Press.

Briscoe, E.J. (2001) 'From dictionary to corpus to self-organizing dictionary: learning valency associations in the face of variation and change', in: *Proceedings of Corpus Linguistics 2001*, Lancaster University, 79-89.

Chomsky, N. (1981) *Lectures on Government and Binding* Dordrecht: Foris.

Chomsky, N. (2005) 'Three factors in language design', *Linguistic Inquiry* 36 (1), 1-22.

Culicover, P.W. & A. Nowak (2003) *Dynamical Syntax* Oxford: Oxford University Press.

Daelemans, W. & A. van den Bosch (2005) *Memory-based Language Processing* Cambridge: Cambridge University Press. Series in Natural Language Processing.

Fodor, J.D. (2001) 'Setting syntactic parameters', in: M. Baltin & C. Collins (eds.) *Handbook of Contemporary Syntactic Theory*, 730-767. Oxford: Blackwell Publishers.

Jakobson, R. (1942) *Kindersprache, Aphasie und Allgemeine Lautgesätze* Upsala: Uppsala Universitets årsskrift 9.

Gleitman, L. (1990) 'The structural source of verb meaning', *Language Acquisition* 1, 3-55.

Greenberg, J.H. (1963) 'Some universals of grammar with particular reference to the order of meaningful elements', in: J.H. Greenberg (ed.) *Universals of Language*, 73-113, Cambridge: MIT Press.

Jordens, P. (1990) 'The acquisition of verb placement in Dutch and German', *Linguistics* 28, 1407-1448.

Kampen, J. van (1996) 'PF/LF conversion in acquisition', *Proceedings of NELS 1995*

Kampen, J. van (1997) *First Steps in Wh-movement* Delft: Eburon.

Kampen, J. van (2006) 'Concord phenomena in first language acquisition', in: *Proceedings of the Workshop on Concord*, ESSLLI 2006.

Kampen, J. van (forthcoming) *Hierarchies of Learning Steps in First Language Acquisition* Monograph in preparation, ms Utrecht University.

Koster, J. (1987) *Domains and Dynasties*. Dordrecht: Foris.

Langley, P. (1996) *Elements of Machine Learning.* San Francisco: Morgan Kaufmann.

MacWhinney, B. (1991) *The CHILDES Project: Tools for Analyzing Talk* Hillsdale New York: Lawrence Erlbaum.

Michalski, R.S., J.G. Carbonell & T.M. Mitchell (1983) (eds.) *Machine Learning. An Artificial Intelligence Approach.* PaloAlto, CA: Tioga.

Mitchell, T.M. (1997) *Machine Learning* New York: McGraw-Hill.

Moortgat, M. (1997). 'Categorial type logics', in: J. van Benthem & A. ter Meulen (eds,) *Handbook of Logic and Language*, 93–177. Amsterdam & Cambridge MA: Elsevier & MIT Press.

Moortgat, M. (2001) *Structural Equations in language Learning* LNCS volume 2099. Berlin: Springer.

Neeleman, A. & H. van de Koot (2002) 'The configurational matrix', *Linguistic Inquiry* 33-4, 529-574.

Obdeijn, A. (2004) *Taalverwerving door Kinderen en Machines* Master thesis University of Amsterdam, Institute for Logic, Language and Computation, Supervisors P. Adriaans (ILLC Amsterdam) and J. van Kampen (UiL OTS Utrecht).

Pinker, S. (1994) The Language Instinct, New York: William Morrow.

Scha, R. (1992) 'Virtuele grammatica's en creatieve algoritmes' *Gramma/TTT* 1 (1), 57-77.

Scha, R., R. Bod & K. Sima'an (1999) 'A memory-based model of syntactic analysis: Data-Oriented Parsing' *Journal of Experimental and Theoretical Artificial Intelligence* 11 (3), 409-440. (Special Issue on Memory-Based Language Processing, edited by Walter Daelemans).

Sakas, W.G & J.D. Fodor (2001) 'The structural triggers learner', in: S. Bertolo (ed.) *Language Acquisition and Learnability* Cambridge: Cambridge University Press.

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

Zwart, C. J.-W. (2006) 'Local agreement', in: C. Boeckx (ed.) *Agreement Systems*, 317-339. Amsterdam: John Benjamins.